International Journal of

Info & Com Sciences for Decision Making

ISSN:1265-499X

1e trimestre 2003

CONTENTS

All correspondences about I.S.D.M or submission should be sent to: Dr Y.Bertacchini
Université de Toulon, LePont, C205, BP 132, 83957 La Garde Cedex, France
e-mail :bertacchini@univ-tln.fr www server :http://www.isdm.org

INFORMATION ET THEORIE MATHEMATIQUE: UNE IMPASSE EN SCIENCE DE L'INFORMATION ?

LE CAS DE L'INFOMETRIE.

Thierry Lafouge
Université Claude Bernard Lyon1 Laboratoire RECODOC
Bâtiment OMEGA
43, Boulevard du11novembre1918
69622 Villeurbanne Cedex
Tel 04 72 44 58 34
lafouge@enssib.fr

Résumé : La théorie statistique de l'information de C. Shannon, appelée souvent à tort théorie de l'information ou théorie mathématique de la communication, est souvent réduite et connue en SIC (Sciences de l'Information et de la Communication) au travers du schéma du système général de la communication : source, émetteur, signal…bruit….. La théorie de Shannon est connue en statistique par sa célèbre formule de l'entropie. La formule de Shannon est isomorphe à la formule de l'entropie de Boltzmann en physique. Cette théorie est importante car elle est à la jonction de la théorie du signal et de la statistique. Les mesures de l'entropie sont utilisées comme indicateurs en statistique unidimensionnelle et bidimensionnelle. Nous essaierons au travers de cet article de donner le point de vue de l'infométrie.

Mot clefs : entropie/ théorie probabiliste de l'information/ statistique

Abstract : Shannon's theory is commonly aboarded in the narrow statement of the general communication scheme : signal, noise, ... The entropy formula in statistics is a caracteristic of shannon's theory which is isomorphic to Boltzman formula in physic. It's an important issue in that way this theory is between signal and statistical theory. By using entropy measures in unidmentional and bidimentional statistics, we'll try to point out this issue in a infometric approach.

Keywords : entropy/ probabilistic theory of information/ statistic

Information et théorie mathématique: une impasse en science de l'information ? Le cas de l'infométrie.

INTRODUCTION

Le mot information est utilisé dans des contextes très variés, dans des sens totalement différents suivants les disciplines scientifiques : on peut à titre d'exemples citer la thermodynamique avec le concept d'entropie, la physique avec la théorie du signal, la biologie avec la théorie du génome. Se pose alors la question, s'il est possible de construire une théorie de l'information, et si elle est unique. Notre démarche dans cet article vise non pas l'information en tant que telle, mais la quantité d'information. Lorsque l'on parle de quantité d'information et de mesure on pense à la notion de contenu ou de valeur de l'information. La science de l'information de par son objet doit se sentir concernée par ce questionnement. Si on définit l'infométrie comme l'ensemble des techniques de mesures mathématiques et statistiques de l'information, on souhaiterait avoir une définition suffisamment claire du concept de quantité d'information qui puisse nous amener à définir une mesure, c'est à dire un ensemble d'opérations parfaitement définies, nous amenant à des axiomes clairs et dont le résultat est un nombre. La synthèse que nous développons ici n'est pas si ambitieuse. De toute façon à l'heure actuelle, faute de connaissances, ou pire parce que on ne saurait vraiment pas formuler le problème une approche générale du concept de quantité d'information serait vouée à l'échec. Nous nous intéressons ici à la théorie probabiliste de l'information, connue sous le nom de théorie de Shannon, qui est la plus utilisée en science de l'information et de la communication. Ce travail qui à première vue peut paraître « risqué, prétentieux ou obsolète» en science de l'information, nous a semblé nécessaire au vue de prises de position souvent extrêmes de certains chercheurs :

- un rejet de cette théorie, souvent par ignorance et /ou par des présupposés épistémologiques : restriction de la théorie de Shannon au célèbre schéma émetteur, canal, récepteur par exemple,
- une utilisation abusive de cette théorie pour valider des résultats,
- une utilisation naïve de cette dernière.

Nous essaierons de donner au lecteur quelques repères pour lui donner l'envie d'approfondir cette théorie et de se forger sa propre opinion. Nous aborderons principalement dans cet article les relations multiples qu'entretiennent la théorie probabiliste de l'information (travaux d'Hartley, Shannon, Reyni..) et les statistiques en général. Nous n'apporterons pas de résultats théoriques nouveaux mais nous mettrons en parallèle différentes approches utilisant cette théorie et donnerons quelques exemples.

1 - LA MESURE DE L'INFORMATION : DE HARTLEY A SHANNON

1.1. Information d'un ensemble : la formule de Hartley en 1928

Etant donné un ensemble E de k éléments, où l'on suppose $k = 2^n$: si à chaque élément de E on associe un numéro écrit en base 2 qui permet de le coder, il est trivial de dire que n digit suffisent pour le repérer . Le nombre n est dit mesurer la quantité d'information nécessaire pour repérer un élément de E. On définit alors la quantité d'information de E, noté $I(E)$ par la même valeur :

$$I(E) = \log_2 2^n = n \ .$$

Hartley en 1928 généralise la quantité d'information pour un ensemble E ayant un nombre quelconque d'éléments par:

$$I(E) = \log_2 (|E|)$$

où $|E|$ désigne le nombre d'éléments de l'ensemble E.

Notation

Par la suite on notera log au lieu de log_2 le logarithme en base 2, Ln le logarithme népérien, Log le logarithme lorsque l'on ne précise pas.

Exemple

Soient les quatre chaînes de caractères, «islamiste, religieux, abcdefghi, xqzerdfk » : elles ont toutes la même quantité d'information, à savoir : $\log 26^9 = 9 \log 26 =$ 42,3 bit. Ici l'ensemble *E* est constitué de tous les arrangements possibles avec répétitions de 9 caractères choisis parmi les 26 lettres de l'alphabet soit 26^9 éléments; l'unité d'information est le bit, information élémentaire pour repérer les éléments d'un ensemble de cardinal 2. Cet exemple est significatif de ce qu'on appelle quantité d'information d'une chaîne : on prend uniquement en compte sa forme, réduite ici au nombre de caractères. Non seulement la signification est totalement absente mais en plus on ne tient pas compte par exemple d'informations statistiques sur les fréquences des caractères dans la langue, seul le nombre de symboles de l'alphabet est retenu. Si ce nombre de symboles est réduit à deux on retrouve les unités classiques en informatique.

Remarques

Nous parlons de mesure de quantité d'information sans avoir défini avec précision ce qu'on appelle information. Ici *n* est le nombre de cases élémentaires, mémoires, pour coder l'information qui est le cardinal de l'ensemble *E*. L'exemple ci-dessus justifie en partie l'appellation théorie du codage utilisée pour désigner la théorie de Hartley et ses prolongements que l'on développera par la suite.

1.2. Information d'un événement : la formule de Wiener en 1948

On se place ici dans le cadre de la théorie des probabilités qui repose sur la théorie des ensembles et de la mesure en mathématique. Parmi toutes les mesures d'information d'un événement *O,* dont la probabilité de réalisation est notée $p = P(O)$, comment mesurer l'information apportée par cet événement ? On parlera désormais d'entropie d'un événement (on justifiera par la suite cette appellation qui est selon nous abusive et trompeuse). On suppose que cette mesure ne dépend que de *p*. On note H cette fonction sur $]0\ \ 1]$. Construire une théorie dans le cadre des probabilités nous amène à poser les propriétés (1) et (2) :

(1) H est non négative $\quad H >= 0$

En effet la probabilité d'un événement est un nombre positif,

(2) $\quad H$ est additive $\quad H(pq) = H(p) + H(q) \qquad p, q \in]0\ 1]$,

En effet si deux événements A et B sont indépendants on a la relation : $p(A \cap B) = p(A).p(B)$..

(3) Enfin pour des questions de normalisation (propriété non obligatoire) on suppose $H(\frac{1}{2}) = 1$.

On montre que la fonction vérifiant ces trois axiomes(J. Aczel J., Z. Daroczy chapitre 1) est nécessairement de la forme :

$$H(p) = -\log(p) \text{ ou } H(O) = -\log(P(O))$$

La condition (2) signifie que l'entropie apportée par la conjonction de deux évènements indépendants est la somme des entropies apportées par chaque événement. Comme on vient de le voir cette condition est naturelle[1]. La troisième condition qui n'est pas essentielle assigne l'unité d'information (appelé bit) à l'événement de probabilité ½ équivalent quant à sa mesure à son opposé .

Exemple

Soit une distribution de descripteurs caractérisant un domaine scientifique. Les mots très fréquents (*p* voisin de 1) ont une entropie très faible (*H* voisin de 0), c'est ce qu'on appelle les mots triviaux , que connaît parfaitement l'expert du domaine. Les mots ayant une basse fréquence qui sont très nombreux ont une entropie très forte, on parle alors de bruit ou de marginalité. C'est parmi eux qu'on trouve ce qu'on appelle les signaux faibles en veille. Le problème bien connu dans les analyses bibliométriques de références bibliographiques est que ces mots sont très nombreux et ont des fréquences identiques très faibles.

1.3. Information d'une suite d'évènements : la formule de Shannon en 1948

1.3.1 Les différentes approches

[1] Naturelle par rapport à la notion d'indépendance. Rappelons que cette question d'indépendance n'a pas d'équivalent dans la théorie de la mesure en mathématique comme en ont les notions d'espérance, de variable aléatoire……

L'entropie de Shannon, notée *H* ou H_n, mesure la quantité d'information moyenne, elle peut être introduite de plusieurs façons, elle généralise les mesures précédentes.

- Le point de vue ensembliste : information apportée par un caractère

Soit $E_i \ i \in I$, une partition de l'ensemble E par le caractère I où l'on note

$p_i = \frac{|E_i|}{|E|}$, on montre facilement (M. Volle) en se servant de la formule de Hartley que la connaissance de *I* permet d'économiser pour repérer les éléments de *E* la quantité d'information suivante :

$$H(I) = -\sum_i p_i . \log(p_i)$$

H(I) est aussi appelée entropie de la partition de *E* définie par *I*.

On remarquera que $H(I)$ ne dépend pas du caractère *I*, ni même du type de ces modalités, mais uniquement de la distribution des fréquence p_i.

- La démarche de Shannon

Voici rapidement les hypothèses que formule Shannon. (W. Weaver, CE. Shannon) Supposons que nous ayons un ensemble *n* d'événement possibles dont les probabilités d'occurrence sont $p_1, p_2 ... p_i .... p_n$. Comment trouver une mesure de l'incertitude du résultat, c'est à dire du nombre de choix possibles? Les probabilités sont connues a priori et c'est tout ce que nous connaissons sur le futur. Si tous les évènements sont équiprobables il est raisonnable de considérer qu'il est souhaitable que l'incertitude soit maximum. Shannon impose à cette mesure *H* trois conditions:

-*H* est une fonction continue des p_i,

-si tous les p_i sont égaux alors *H* est une fonction monotone croissante de n,

-si un choix se décompose en deux choix successifs le *H* original devra être la somme pondérée des valeurs individuelles (Cette propriété est appelée par la suite récursivité ou "branching process"). Shannon montre que la seule fonction *H* satisfaisant aux trois hypothèses ci dessous est de la forme :

$H = k.\sum_{i=1}^{n} p_i . Log(p_i)$ où *k* est une constante dépendant des unités.

Donnons une forme non normalisée: soit p_i une suite de *n* nombres positifs ou nul quantifiant la probabilité de *n* évènements et vérifiant la relation : $\sum_{i=1}^{n} p_i <= 1$ l'entropie de cette suite est définie par la quantité :

$$H_n(p_1, p_2 ... p_i, ... p_n) = -\sum_{i=1}^{n} p_{i.} . Log(p_i) / \sum_{i=1}^{n} p_i$$

Si la suite définit une distribution de probabilité on retrouve la formule bien connue de Shannon qui est à l'origine, rappelons le, une théorie élaborée en vue de modéliser la transmission des signaux électriques.

L'approche de Shannon généralise celle de Hartley et de Wiener. Si la suite se réduit à un seul élément on retrouve la formule précédente de Wiener. $H(p) = -\log(p)$

Enfin si $p_i = \frac{1}{n}$, c'est à dire si tous les évènements sont équiprobables on obtient la formule précédente de Hartley $I(E) = \log(|E|) = \log(n) = H_n(\frac{1}{n}, \frac{1}{n} ..... \frac{1}{n})$. On montre que $\log(n)$ est la valeur maximum

de l'entropie : on définit alors la quantité d'information relative H_r variant entre 0 et 1 qui est le rapport de l'entropie sur l'entropie maximum.

$$H_r = \frac{H_n}{\log(n)}$$

- Le point de vue informationnel pragmatique

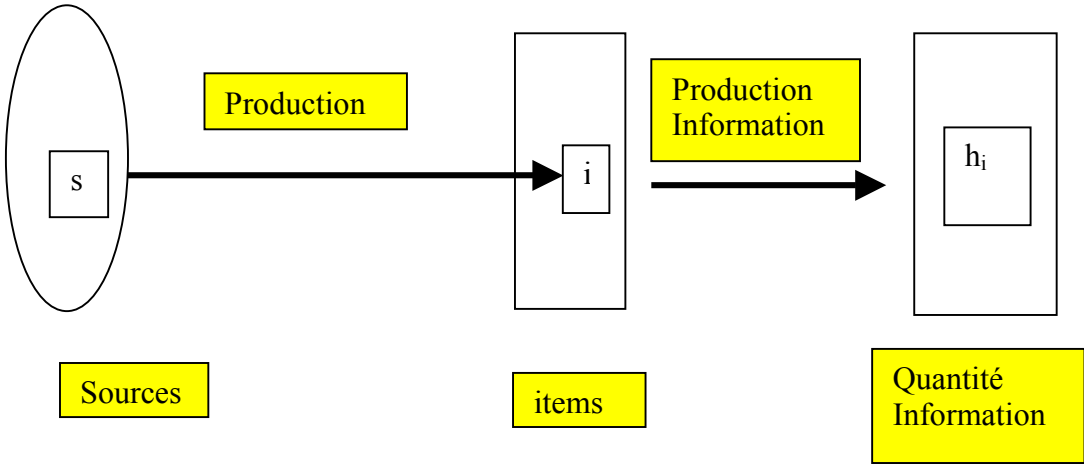
Si l'on veut définir une fonction (notée aussi *H)* qui mesure l'entropie d'une suite d'événements on peut (Un système de production bibliographique (Egghe 1990) est un ensemble de sources qui produisent des items) utiliser le formalisme de la bibliométrie distributionnelle (Lafouge 1991).

- les chercheurs produisent des articles,
- les formes graphiques[2] d'un texte produisent des occurrences,
- les mots clefs de références bibliographiques produisent des occurrences,

…….

On écrit *H* sous la forme $H = \sum_i h_i p_i$ où p_i est la probabilité qu'un événement élémentaire i se produise (un chercheur produit *i* articles……) et h_i l'information apportée par cet événement *i*. On définit alors h_i par $h_i = -\log(p_i)$ (On s'inscrit dans les hypothèses de Wiener)

H est l'espérance de la variable aléatoire $i \to \log(p_i)$. Le schéma ci-dessous résume le problème.



$$H = \sum_i h_i \log(p_i)$$

H mesure une quantité moyenne d'information. Nous obtenons le même résultat que Shannon. Ceci n'est pas un hasard car nous nous situons dans le cadre de la théorie des probabilités. Alors que dans la théorie de Shannon on parle de transmission de quantité d'information, ici on parle de production de quantité d'information.

Exemple

Les phénomènes précédents de production bibliométrique peuvent être décrits suivant différentes formes. Nous utilisons la représentation générale des fonctions zipfiennes étudiées par Haitum (S. D Haïtum) où une telle distribution est définie par la fonction de densité hyperbolique suivante :

$v(t) = \frac{C}{t^{\alpha+1}}$ où C est une constante, α un nombre positif, et où $t \in [1 \quad \infty]$

Il est possible de généraliser la définition de l'entropie lorsque on a une distribution continue :

[2] Ensemble de caractères délimités par un séparateur.

$$H(v) = -\int v(t).Log(v(t))dt$$

Les entropies des distributions continues héritent de la plupart des propriétés du cas discret défini précédemment. Si on calcule l'entropie d'une zipfienne on a :

$H(\alpha) = -Log(\alpha) + \frac{1}{\alpha} + 1$. Il est aisé de montrer que l'entropie est une fonction décroissante de α . On retrouve l'interprétation classique de la loi de Lotka qui stipule que plus α est élevé, plus grand est le fossé entre le petit nombre de chercheurs qui produisent beaucoup et le grand nombre de chercheurs qui produisent très peu, et donc plus grand est la quantité d'information. Yablonsky (A. L Yablonsky) montre qu'il existe un lien entre ce type de distribution et le principe du maximum d'entropie (notée MEP), lui même lié au principe de la loi du moindre effort (PLE), résultat que nous avons prolongés (T. Lafouge, C. Michel) pour le cas de la distribution binomiale négative. Nous pensons que cette voie de recherche encore peu explorée peut être féconde.

- Le rêve idéaliste : entropie et information

Soit un langage constitué de *n* symboles chacun ayant une fréquence N_i le nombre total *W* de messages possibles de *N* symboles respectant les fréquences précédentes (c'est à dire l'égalité : $N = \sum_{i=1}^{n} N_i$) est :

$$W = \frac{N!}{N_1!.N_2!.N_3!.....N_n!} \text{ (formule de Brioullin)}$$

Si l'on pose $p_i = \frac{N_i}{N}$ $i = 1..n$ et que l'on utilise la formule de Stirling pour calculer factoriel on obtient : $\frac{\log(W)}{N} = k.H_n(p_1.....p_i...p_n)$: sous cette forme la mesure de Shannon est bien la quantité moyenne d'information apportée par un symbole .Son analogie avec la fonction entropie (Voir ci-dessous) de la thermodynamique est patente (D. Parrochia).

Les notions d'entropie et d'information découlent de la thermodynamique. C'est Sadi Carnot qui en formulant le premier principe va initier ces travaux. Enfin Boltzman en étudiant la mécanique va obtenir une formule identique en calculant l'énergie d'un gaz comme la moyenne des énergies correspondantes . Si on désigne par *S* la fonction entropie, cette dernière peut s'écrire :

$S = K.Ln(\Omega(E))$ où $\Omega(E)$ désigne le nombre d'états possibles d'un système ayant une énergie donnée *E* et où *K* est une constante . Tous ces résultats permettent de montrer qu'il existe une isomorphie entre l'entropie de Bolzmann et l'entropie de Shannon.

Il est alors tentant de postuler une équivalence entre énergie et information. L'exploitation qu'on peut faire d'une telle analogie montre maintenant ses limites. L'analogie entre entropie et information est séduisante mais inopérante pour notre discipline. Nous avons cependant gardé ici comme dans beaucoup d'ouvrages et d'articles le terme entropie.

Remarque

Lorsque que l'on travaille avec les fréquences il est souvent intéressant de mettre l'entropie de Shannon sous la forme :

$$H = \log(F) - \frac{1}{F}.\sum_{i=1}^{n} f_i.\log(f_i) \qquad F = \sum_{i=1}^{n} f_i$$

Exemple

On parle en linguistique quantitative de l'entropie des lettres de l'alphabet d'une langue. Ainsi en français la fréquence moyenne de la lettre E est 0,175 ……. On définit alors l'entropie moyenne d'une lettre par $H(p_1, p_2, ...p_{26})$; en langue française le calcul nous donne 3,98 bit. On peut alors étudier l'utilisation des caractères dans plusieurs langues et faire des comparaisons. Il n'en est pas de même pour les formes graphiques ou les mots d'une langue car on ne travaille pas dans un univers fermé.

1.3.2 Les propriétés de la mesure de Shannon

Nous rappelons sans les démontrer les principales propriétés algébriques
de cette mesure :

(a) Symétrie $H_n(p_{1,}...p_i,....p_n)=H_n(p_{k(1)}.....p_{k(i)}.... p_{k(n)})$ où k est une permutation arbitraire sur l'ensemble {1….n}

(b) Normalité $H_2(\frac{1}{2},\frac{1}{2})=1$

(c) Décision $H_2(1,0) = 0$

(d) Linéarité

Soient deux distributions de probabilité : $p_i \quad i=1..n \quad q_j \quad j=1..m$

$$H_{nm}(p_1.q_{1,}.....p_i.q_j.....p_nq_1, p_nq_2.....p_nq_m) = H_n(p_1..p_i..p_n)+H_m(q_1...q_{ij}.....q_m)$$

(e) Linéarité forte

Soient une distribution de probabilité $p_i \quad i=1..n$ et m distributions de probabilités : $q_{jk} \quad k=1..n \quad j=1..m$

$$H_{nm}(p_1q_{11}, p_1q_{12}.....p_1.q_{1m}.,.....p_iq_{i1}, p_iq_{i2}.....p_iq_{im},.....p_nq_{n1}, p_nq_{n2},....p_nq_{nm}) =$$

$$H_n(p_{1,}...p_i....p_n)+\sum_{j=1}^{n} p_j.H_n(q_{j1}, q_{j2}.....q_{jn})$$

(f)Récursivité $H_n(p_{1,}...p_i,....p_n)= H_{n-1}(p_1+p_2, p_3,.....p_n)+\frac{p_1}{p_1+p_2}H_2(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2})$

Les propriétés (a) (b) (c) sont communes à toutes les mesures statistiques de l'information (Voir les mesures de Reyni au paragraphe 1.4, et les mesures d'ordre supérieure dans la conclusion).

- La symétrie signifie que c'est une mesure globale d'un ensemble d'événements,
- la normalité traduit que l'incertitude est maximum lorsque tous les événements sont équiprobables. Elle croît en fonction du nombre d'événements.
- la propriété de décision signifie qu 'il n'y a pas d'incertitude si un événement est sûr.
- la linéarité est l'équivalent de la propriété de Wiener pour des évènements indépendants,
- la linéarité forte joue un rôle important : elle va nous permettre de donner un sens à la notion d'entropie conditionnelle ; elle est à la base de la construction d'une série d'indicateurs (cf. 3.1.1).

1.4 Une autre approche : le gain d'information

A première vue la notion de gain d'information dans la vie courante semble plus intuitive que la notion de quantité d'information. Dans la théorie de l'information un gain d'information sera matérialisé par une quantité d'information.

Pour introduire cette notion on peut procéder comme pour le point de vue ensembliste précédent (Voir paragraphe 1.3.1) en utilisant les mêmes notations. On définit F un sous ensemble de E et on construit la partition de F : $F_i = F \cap E_i$ avec $q_i = \frac{|F_i|}{|F|}$. La question est alors : pour repérer un élément de E, qu'apporte le fait de savoir que cet élément appartienne à F ? *Cette quantité* $\Delta(J)$ *est égale est égale à l'information pour repérer un élément suivant I si on ignore que l'élément est dans F moins l'information si l'on sait que l'élément est dans F(Voir M. Volle p58).*

$$\Delta(J)=\sum_i q_i \log(\frac{q_i}{p_i})$$ *(Cette information est dite information de Kullback.)*

Le lecteur peut penser que nous jouons avec les mots (ou avec les notations en mathématiques) Il vérifiera cependant que cette quantité est différente de $H_n(p)-H_n(q)$, et donc que le gain d'information n'est pas une différence de quantité d'information, c'est bien un autre concept.

De la même façon soient deux caractères *X* et *Y* définis chacun par une distribution de probabilité, $p_i \quad i=1,n$ et $q_j \quad j=1,m$, on définit $\Delta(X,Y)$ le gain d'information lorsque l'on passe des distributions marginales de *X* et de *Y* à la distribution des fréquences des évènements *(i j)* . On montre alors que le gain d'information est égal à :

$$\Delta(X,Y)=\sum_{i\,j} p(i,j)\log\frac{p(i,j)}{p_i.p_j}.$$

On remarque que cette expression est symétrique en *i* et *j* et qu'elle est nulle si les deux distributions sont indépendantes.

Lorsque l'on élabore l'axiomatique de la théorie probabiliste de l'information deux approches sont possibles.

La première consiste à suivre les étapes suivantes :

Définition de la quantité d'information -> axiomatisation ->définition du gain d'information.

La deuxième :

Définition du gain d'information -> axiomatisation - > définition de la mesure d'information

Dans (A. Renyi) ce dernier développe la deuxième approche qui lui permet d'axiomatiser la notion de gain d'information et de construire dans un deuxième temps un ensemble de nouvelles mesures dont on trouvera la formule ci-dessous :

Soit une distribution de probabilité $p_i \quad i=1,n$ et un nombre $\alpha \neq 1$ il pose :

$$H_a = \frac{1}{1-a} Log\,(\sum_{i=1}^{n}(p_i)^a)$$

Attention

On vient de voir dans tout ce qui précède que l'entropie de Shannon est une mesure de l'incertitude. Au sens de Shannon, c'est à dire la théorie mathématique de la communication l'entropie mesure l'imprévisibilité moyenne que l'observateur a sur le message émis. Plus cette valeur est élevée plus l'incertitude augmente (en physique on dit que le désordre augmente), ce que nous traduisons en disant que l'information diminue. Quantité d'information et entropie varient en sens contraire. Si un événement est sûr, son entropie est nulle, par contre la quantité d'information est maximum. Les physiciens parlent de néguentropie pour désigner la quantité d'information..

2 - STATISTIQUES ET MESURE DE L'INFORMATION

L'objectif de ce paragraphe est de montrer les liaisons qui existent entre les statistiques classiques utilisées en infométrie et les mesures précédentes. Certains ouvrages de statistique ont un chapitre entier consacré à la théorie statistique de l'information (Voir par exemple M. Volle, A. Reyni dans la bibliographie ci dessous)

2.1. Statistique unidimensionnelle et mesures informationnelles

2.1.1 Indicateurs de concentration et de diversité

Pour caractériser les distributions de fréquences les statistiques unidimensionnelles utilisent trois types de mesure que sont les indicateurs de tendance centrale, de dispersion et de concentration. Les distributions rencontrées en science de l'information sont connues sous le nom de Zipfienne et ont été largement étudiées dans la littérature (S. D Haitum). Elles sont en général de forme hyperbolique et décroissantes (cf 1.3.1) et possèdent une longue queue avec un écart type supérieur à la moyenne. En général on les oppose aux distributions gaussiennes que l'on rencontre fréquemment lorsque l'on étudie des populations physiques. Aussi les indicateurs classiques (moyenne, variance, coefficient de variation) issues de la théorie des moments ne sont pas toujours adaptés pour résumer ces distributions. Les chercheurs en biologie, économie, infométrie ont développé de nombreux autres indicateurs.

Nous allons nous intéresser aux indicateurs de concentration ou de diversité. Donnons rapidement une présentation des indices de ce type d'indice. Soit un ensemble de *n* sources (ouvrages, auteurs, mots, …) qui produisent des items (prêts, articles, occurrences…). Nous désignons par f_i le nombre d'items produits par la i^{ème} source ; un indice de concentration (respectivement de diversité) de cette distribution doit vérifier les trois propriétés :

$D(f_1, f_2, ...f_i .....f_n) = D(f_{k(1)}\, f_{k(2)} .....f_{k(n)})$: invariance par permutation, cela signifie que c'est bien un indice global.

$D(f_1, f_2, ...f_i .....x_n) = D(af_1, af_2 .....af_i .....af_n)$: invariance si on modifie l'échelle de mesure.

Si ces deux propriétés sont classiques pour de nombreux indicateurs il n'en n'est pas de même pour celles qui vont suivre que L. Egghe a longuement étudié (L. Egghe, R. Rousseau 1990) et appelle principe de transfert si on a :

Pour tout $f_i \leq f_j$ et $0 < f \leq f_i$:

un indice de concentration vérifie:

$D(f_1, f_2, ...f_i .....f_n) \quad < \quad D(f_1, f_2 .....f_i - f .....f_j + f ....f_n)$

Cela signifie qu'un transfert d'une source pauvre vers une source riche augmente l'indice de concentration.

Un indice de diversité vérifie :

$D(f_1, f_2, ...f_i .....f_n) \quad > D(f_1, f_2 .....f_i - f .. + ...f_j + f ....f_n)$

Cela signifie qu'un transfert d'une source pauvre vers une source riche diminue l'indice de diversité.

La mesure de Teil (provenant de l'entropie qui est une mesure de diversité) est une « bonne mesure de concentration » au sens qu'elle vérifie toutes les propriétés souhaitées (L. Egghe, R. Rousseau) ; elle est souvent écrite sous la forme : $Th = \frac{1}{n}.\sum_{i=1}^{n} \frac{f_i}{\mu}.\log(\frac{f_i}{\mu}) \quad où \quad \mu = \frac{1}{n}\sum_{i=1}^{n} f_i$ on montre aisément que l'on a l'égalité :

$$H = Th - Logn$$

Pielou a généralisé ce type d'indice. Ce dernier plus connu sous le nom d'entropie généralisée a été développé par Reyni, dans un tout autre contexte (Voir paragraphe 1.4). On définit l'entropie généralisée d'ordre α par :

$H_a = \frac{1}{1-a} Log(\sum_{i=1}^{n} (p_i)_a) \quad a \neq 1$ on montre que $H = Lim\ (a \to 1) H_a$. Hill propose une approche (M. Hill) qui va définir un ensemble de mesures de diversité : $D_a = \exp(H_a)$. Lorsque $\alpha = 2$ on retrouve l'indice bien connu de Simson : $D_2 = \sum_{i=1}^{n} p^2{}_i$.

Il est intéressant de noter que deux approches différentes, une en statistique unidimensionnelle et l'autre liée à la notion de gain d'information aboutissent à des résultats identiques.

2.1.2 Variance et Entropie

Une des propriétés selon nous la plus intéressante de l'entropie pour notre discipline est le caractère agrégatif de cette mesure semblable à celui de la variance. Comme la variance l'entropie peut être découpée en fractions après un regroupement en classes. Une classe est un regroupement de sources, on peut citer :

en scientométrie :un laboratoire regroupement des chercheurs,
en bibliométrie :un journal regroupant des articles,
en infométrie :un thème regroupant des mots spécifiques

On notera :

n	nombre de sources,
m	nombre de classes,
f_i	nombre d'items produit par la i^{eme} source,
F_k	nombre d'items produits par la k^{eme} classe,
F	nombre total d'items, $F = \sum_{i=1}^{n} f_i$
H_k	entropie de la k^{eme} classe.

On a alors le résultat suivant : $H = H\,\text{int}\,er \;+\; H\,\text{int}\,ra$

Où comme pour la variance (voir variance interclasse, intraclasse) :

$$H_{\text{int}\,er} = \sum_{j=1}^{m} \frac{F_k}{F}.H_k \qquad\qquad H_{\text{int}\,ra} = -\sum_{j=1}^{m} \frac{F_k}{F}\log(\frac{F_k}{F})$$

Cette propriété est intéressante car la quantité d'information ainsi décomposée donnera une vision de la quantité d'information (ou plutôt diversité) apportée par chaque classe et par leur différenciation au sein du groupe. Cette propriété est intéressante quand il est possible de regrouper un ensemble de codes, de mots (Voir Thésaurus, Plan de classement…), en unités hiérarchiquement supérieures. C'est dans cet esprit que nous avons préconisé l'emploi de nouveaux indicateurs pour résumer les distributions Zipfiennes (J. Lhen,), l'entropie de Shannon étant la diversité d'ordre 1.

2.2 Statistique bidimensionnelle et mesures informationnelles

La statistique unidimensionnelle résume avec des indicateurs une distribution de valeurs. La statistique bidimensionnelle a pour objectif de découvrir des liens éventuels qui existent entre deux variables.

2.2.1 Entropie conditionnelle et information mutuelle

Nous allons tirer les conséquences de la propriété de linéarité forte de la mesure de Shannon. Supposons deux évènements *X* et *Y* chacun étant défini par une suite de *n* (respectivement de m) éventualités définis par une suite p_i (respectivement p_j).

On note :

$p(i,j)$ probabilité d'avoir les évènements *i* et *j* simultanément,

$p(j/i)$ probabilité d'avoir *j* sachant *i* réalisé (probabilité conditionnelle),

$$p(j/i) = \frac{p(i,j)}{\sum_{i=1}^{n} p(i,j)}$$

On a $H(X,Y)$ entropie conjointe de *X* et *Y* :

$$H(X,Y) = -\sum_{i\ j} p(i,j).\log(p(i\ j)$$

Si les évènements sont indépendants ($p(i,j) = p_{i.}p_{j}$) on obtient la relation : $H(X,Y) = H(X) + H(Y)$ (Voir propriété (d) de linéarité)

On définit l'entropie conditionnelle de Y sachant *X* comme la moyenne de l'entropie *Y* pondérée pour chaque valeur de *X* pondérée.

$$H(Y/X) = \sum_{i\ j} p(i,j).\log(p(j/i))$$

On montre alors facilement que les propriétés suivantes (Voir propriété de linéarité forte) :

$$H(X,Y) = H(Y/X) + H(X)$$

D'où on tire : $H(X) + H(Y) \geq H(X,\ Y) = H(X) + H(Y/X)$

On en déduit le résultat : si *X* et *Y* dépendent l'un de l'autre alors la connaissance de *X* entraîne une diminution de l'entropie c'est à dire de l'imprévisibilité qu'on a sur *Y* : $H(Y/X) \leq H(Y)$. De même on montre : $H(X/Y) \leq H(X)$.

Ce résultat nous permet de construire les indicateurs souvent utilisés en théorie mathématique de la transmission du signal et par certains chercheurs en scientométrie que l'on verra par la suite lors d'un exemple :

Information sur *X* contenu dans *Y* : $T(X,Y) = H(X) - H(X/Y)$

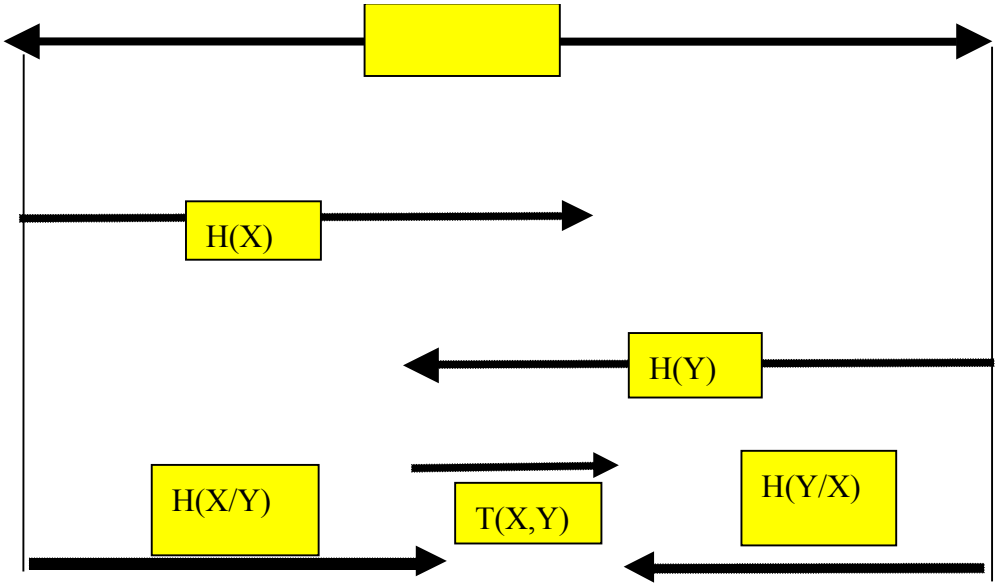
Information sur *Y* contenu dans *X* : $T(Y,X) = H(Y) - H(Y/X)$

Ces indicateurs sont souvent appelés liens en statistique, ils ne sont pas symétriques.

Comme pour l'entropie relative on utilise de préférence des indicateurs quantifiant les informations relatives :

$$\frac{T(XY)}{H(X)};\frac{T(YX)}{H(Y)}$$

On résume traditionnellement toutes ces propriétés dans le schéma ci-dessous reproduit dans des contextes différents. On va développer un exemple en s'inspirant des travaux de Loet Leydesdorf (L. Leydesdorff) qui fait partie des chercheurs utilisant cette théorie dans ses nombreux travaux.



Exemple

Soit un ensemble *A* de *m* articles concernant un thème scientifique. Chaque article est caractérisé par un ensemble *M* de *n* de mots (mots clés, mots du résumé, du titre….) caractérisant son contenu. La fréquence de chaque mot (appelée également nombre d'occurrences) est connue. Soit f_{ij} la fréquence du mot *i* dans l'article *j* on alors un tableau croisé (Articles * Mots) de n*m valeurs qu'on va traiter dans un premier temps classiquement :

$N = \sum_{i\,j} f_{ij}$ nombre total d'occurrences dans les *m* articles,

$p(i,j) = \frac{f_{ij}}{N}$ probabilité d'occurrence du mot *i* dans l'article *j*,

$a_j = \frac{\sum_{i=1}^{n} f_{ij}}{N}$ probabilité d'occurrence d'un mot dans l'article *j*

$m_i = \frac{\sum_{j=1}^{m} f_{ij}}{N}$ probabilité d'occurrence du mot i dans les articles.

On utilise la formalisation précédente. On calcule en premier lieu les entropies des trois distributions :

$H(A) = -\sum_{j=1}^{m} a_j.\log(a_j)$ distribution des articles au travers des mots,

$H(M) = -\sum_{i=1}^{n} m_i.\log(mi)$ distributions des mots au travers des articles,

$H(A,M) = -\sum_{i\,j} p(i,j).\log(p_{ij})$ distribution conjointe.

On en déduit les entropies conditionnelles :

$H(A/M)=H(A,M)-H(M)$ $H(M/A)=H(A,M)-H(A)$

Puis ce qu'on a appelé les informations mutuelles

$T(A,M)=H(A)-H(A/M)$ $T(M,A)=H(M)-H(M/A)$

On construit alors les indicateurs (compris entre 0 et 1) qui vont résumer l'information de ces mots distribués dans les articles :

$Hr(A,M)=\frac{H(A,M)}{\log(n.m)}$: pourcentage d'entropie de la distribution des valeurs du tableau par rapport à l'entropie maximale .

$Tr(A,M)=\frac{T(A,M)}{H(A)}$: réduction de l'entropie concernant la distribution des mots dans les articles.

$Tr(M,A)=\frac{T(M,A)}{H(M)}$: réduction de l'entropie concernant la distribution des articles à travers les mots.

C'est en partie en utilisant ces trois indicateurs que L. Leydesdorff analyse les résultats d'une étude menée par l'Ecole des Mines qui est à l'origine de la méthode des mots associés. Le lecteur trouvera en bibliographie la référence de cet article (L. Leydesdorf(1992)) ainsi que la réponse de J. P Courtial un des nombreux chercheurs à l'origine de cette méthode. (J. P Courtial). Leydesdorff calcule les paramètres ci-dessous sur des données avant classification et après classification (après formation de ce qu'on appelle généralement des agrégats); dans ce cas il fait les mêmes calculs sur ces données en regroupant les mots de chaque agrégats. Il obtient un tableau Articles* Agrégat qu'il est obligé de compléter par les mots de fréquence faible qui ont été éliminés lors de classification. Cette étude est intéressante car elle manipule les mêmes données et permet de faire des comparaisons. Cette démarche est malheureusement bien rare dans notre discipline. Sans vouloir donner raison à l'un ou à l 'autre nous voulons faire une ou deux remarques sur l'utilisation de ces indicateurs comme outil d'évaluation de la méthode des mots associés.

Les mesures faite avec l'entropie sont des moyennes. Dans la pratique l'écart type calculé sur la distribution informationnelle (cf 1.3.1 approche informationnelle pragmatique) est souvent supérieur à l'entropie et donc relativise les conclusions car il existe une dispersion tres forte autour de la moyenne. Nous avons conscience qu'il faudrait développer cette affirmation d'un point de vue théorique. Cette étude a été entreprise.

La méthode des mots associés utilise un indice de proximité appelé coefficient d'équivalence qui est la mesure du cosinus de Salton au carré. L'évaluation de la méthode des mots associés est faite avec l'entropie d'ordre 1 qui est une mesure lié à la distance du khi2. Pourquoi tester uniquement l'ordre 1 ?

2.2.2 Distance du $\chi 2$ et mesure du gain d'information

Explicitons les relations entre la distance du χ^2 et de la mesure du gain d'information vue au paragraphe 1.4.

Rappel : la métrique du χ^2

Si l'on considère les trois distributions *p*, *q*, *r* (p_i q_i r_i $i=1,n$) la distance entre *p* et *q*, calculée avec la métrique du χ^2 centrée sur *r* par :

$D^2{}_r(p,q)=\sum_{i=1}^{n}\frac{(p_i-q_i)^2}{r_i}$. L'utilisation de cette métrique est justifiée par le test du χ^2 (Voir la loi de probabilité du χ^2) et ses propriétés qui en en font une distance adaptée à de nombreux problèmes en Statistique lorsque l'on croise deux variables nominales ou que l'on veuille ajuster une distribution par exemple. Elle est également utilisée en analyse factorielle où la distance entre les points est celle du χ^2 centrée sur le centre du nuage de points.. Nous allons voir que cette distance est liée avec la notion d'entropie. Nous utilisons par la suite les notations du paragraphe 3.1.1 . On définit le *Lien* entre deux caractères par :

- $p(i,j)$ distribution du caractère conjoint (*X* ,*Y*),

- $p_i . p_j$ distribution des caractères conjoints *(X, Y)* supposés indépendants,

$$Lien(X,Y) = \sum_{i\ j} \frac{(p(i,j) - p_i . p_j)^2}{p_i . p_j}$$

D'après ce qui précède, le *Lien* mesure donc la distance au sens du χ^2 entre les deux distributions ci dessus Ce dernier mesure la dépendance entre les deux caractères, plus la valeur sera élevée, plus les caractères seront dépendants. Si *X* et *Y* sont indépendants le *Lien* entre *X* et *Y* est nul. Cette mesure symétrique est équivalente à une constante près au gain d'information mutuelle $\Delta(X,Y)$ défini au paragraphe 1.4 (Voir M. Volle p 65)

on a $\Delta(X,Y) \approx k.Lien(X,Y)$

Cette approximation est valable uniquement pour des faibles valeurs du Lien c'est à dire si les caractères ne sont pas « trop indépendants ».
On a des résultas identiques entre la meure du cosinus distributionnel de Salton et la mesure du *Lien* de Reyni d'ordre ½..

CONCLUSION

Toutes ces mesures de quantité d'information sont basées sur la notion d'évènement et sont des mesures de type probabiliste qui sont construites suivant le schéma mental suivant :

à un événement on fait correspondre sa probabilité *p,* puis un nombre qui est fonction de *p* caractérisant une mesure de sa quantité d'information. Nous avons choisi cette approche pour introduire la mesure de Shannon de façon pragmatique. Cette dernière est mathématisée dans (J. Aczel J., Z. Daroczy chapitre 3) où on définit ce qu'on appelle une fonction d'information (fonction de Shannon), puis une fonction d'information d'ordre α. Cette dernière permet de définir les mesures d'information d'ordre supérieur. Pour $\alpha \neq 1$ on pose :

$$H_n{}^{\alpha}(p_1, p_2, ...p_n) = \frac{1}{2^{1-\alpha} - 1} . (\sum_{k=1}^{n} p^{a}{}_{k} - 1)$$

Ces mesures qui sont des mesures de type entropie généralisent celle de Shannon. Comme pour la mesure de Reyni on montre le résultat suivant : $H_n = Lim\ (a \to 1) H_n{}^{a}$ .Cet ensemble de mesures au vu des propriétés mathématiques (J. Aczel J., Z. Daroczy chapitre 6) (Voir linéarité, linéarité forte, récursivité) est vraiment la généralisation de la mesure de Shannon... Nous ne connaissons pas d'utilisation de ces mesures en infométrie, encore faudrait t-il donner des raisons d'utiliser ces mesures entropiques généralisées et savoir donner une interprétation au coefficient α. Les deux approches, celle de Reyni et celle de Shannon ne sont pas indépendantes, il existe une relation mathématique entre ces deux formes.
Nous espérons avoir ouvert des directions de Recherche qui peuvent encore selon nous révêler des résultats nous permettant de mieux comprendre et utiliser les mathématiques mesurant les quantités d'information.

REMERCIEMENTS

Je remercie Sylvie Lainé Cruzel , maitre de conférence de Recodoc, qui a bien voulu critiqué ce travail.

RÉFÉRENCES BIBLIOGRAPHIQUES

J. Aczel Z. Daroczy On Measures of Information and their Characterisations. Mathematics in Science and Engineering Vol 115.

J. P. Courtial (1992) Comments on Leydesdorf's a validation study of Leximappe. Scientometrics 25 p 313-316.

L. Egghe, R. Rousseau Sensitivity Aspects of inequality measures(preprint)

L. Egghe (1990) *The duality of informetric systems with applications to the empirical law.*Journal of Information Science, Vol 16, 1990, p 17-27

L. Egghe Development of hiearchy theory for digraphs using concentration theory based on a new type of Lorentz curve (preprint)

L. Egghe R. Rousseau (1990), Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, 1990, 450 pages, Amsterdam.

S. D. Haitum (1982) *Stationary Scientometric Distributions.* Scientometrics n°4, 1982, Part I p.5-25, Part II p.89-104, Part III p.181-194.

M. O. Hill (1973) *Diversity and Eveness: a unifying notation and its consequences.* Ecology, 1973, Vol 54, N°2, p. 427-433.

T. Lafouge, C. Michel (2001). *Links between information construction and information gain. Entropy and bibliometric distributions*. Journal of Information Science, 27 (1) 2001, p 39-49.

T. Lafouge, L. Quoniam (1991). *Les distributions bibliométriques.* Revue française de bibliométrie N° 9, 1991, p. 128-138.

L. Leydesdorff (2001) The challenge of Scientometrics . The developpment, Measurement, and Self-Organisation of Scientific Communications Published by Universal Publishers

L. Leydesdorf(1992) *A validation study of Leximappe Scientometrics* 25 (1992) p295-312.

J. Lhen, T. Lafouge, Y. Zilsken, L .Quoniam, H. Dou (1995) *La " Statistique" des lois de Zipf.* Revue Française de Bibliométrie N°14, 1995, p. 135-146.

D. Parrochia (1994) Cosmologie de l'Information Chapitre 1, 2, 3. Hermès 282 pages

A. Renyi (1966) Calcul des probabilités, Edition Jacques Gabay 1992, 618 pages

M. Volle (1985) Analyse des données Collection "Economie et Statistiques avancées ", Economica.323 pages.

W.Weaver, C.E Shanon (1975) Théorie mathématique de la Communication. Les classiques de sciences humaines La bibliothèque du CEPL 188 pages.

A. L. Yablonsky (1980) On fundamental regularities of the distribution of scientific productivity. Scientometrics 2(1) 1980 p3-34.

MATHEMATIQUE ET STATISTIQUE EN SCIENCE DE L'INFORMATION
INFOMETRIE MATHEMATIQUE ET INFOMETRIE STATISTIQUE

LE COADIC Yves F.
CNAM - ICST
2 rue Conté - 75141 PARIS Cedex 03
Téléphone/télécopie - (0)140272866
lecoadic@cnam.fr

Résumé : L'application de la mathématique et de la statistique à l'étude des phénomènes informationnels a entraîné la naissance en science de l'information d'un nouvel axes de recherche et de développement, l'infométrie. Après avoir montré l'intérêt de cette application mais aussi avoir mis en garde contre certains abus et contre certains mauvais usages, nous présentons quelques exemples d'infométrie mathématique et d'infométrie statistique. Ils illustrent l'étendue et l'efficacité des analyses qui peuvent être faites sur une ou plusieurs variables informationnelles.

Abstract : Informetrics, the mathematical and statistical study of information processes, is a new promising field of research in information science. Advantages but also pitfalls and misuses of mathematics and statistics in social sciences are presented. A selection of applications (mono and multidimensionnal) coming from mathematical informetrics and statistical informetrics illustrate the efficiency of these methods.

Mots-clés : Mathématique, statistique, infométrie mathématique, infométrie statistique, bibliométrie, scientométrie, médiamétrie, muséométrie, webométrie, nombre, mots, documents, cartes, ZIPF, BOOLE.

Keywords : Mathematics, statistics, mathematical informetrics, statistical informetrics, bibliometrics, scientometrics, mediametrics, museometrics, webometrics, number, words, documents, maps, ZIPF, BOOLE.

« Or, je soutiens que dans toute théorie particulière de la nature, il n'y a de science proprement dite qu'autant qu'il s'y trouve de mathématique »

E. KANT – Premiers principes métaphysiques de la science de la nature

- INTRODUCTION

L'étude des phénomènes informationnels a révélé l'existence de régularités, de rapports mesurables, de distributions qui ne peuvent être mis à jour que par l'application de la mathématique et de la statistique. Cela a donné naissance à un nouveau champ de recherches en science de l'information appelé INFOMÉTRIE. À l'intérieur de l'infométrie sont regroupés les sous-champs de recherches formés sur des secteurs informationnels spécialisés comme celui du livre, la bibliométrie (la première née), de la R&D (recherche-développement), la scientométrie, des mass-médias, la médiamétrie, des musées, la muséométrie et du WorldWideWeb, la webométrie (la dernière née).

Mathématique et statistique s'appliquent donc en science de l'information et ont, si l'on en juge par le panorama des applications que nous avons choisi de présenter, une incroyable efficacité. Mais elles peuvent aussi se révéler nocives si on n'en fait pas bon usage.

- I - LA MATHÉMATIQUE S'APPLIQUE

Traditionnellement, pour beaucoup, la mathématique s'applique pour construire des ponts, des machines; elle s'applique aussi en physique, discipline particulièrement mathématisée, en chimie, en biologie. De plus en plus aux sciences sociales comme l'économie, la psychologie, la sociologie et …la science de l'information. Mais dans l'esprit des professionnels de ce secteur, cela ne va pas forcément de soi.
Les succès de la physique classique, puis de la relativité et de la mécanique quantique ont mis en lumière sa pleine fécondité. Mais ce sont les beaux travaux de sociologie mathématique (R.BOUDON, J.S. COLEMAN) qui nous ont révélé son incroyable efficacité.

Qu'est-ce que cette efficacité ? Elle apparaît au travers de trois capacités : une capacité prédictive, une capacité rétrodictive et une capacité explicative.

Une capacité prédictive

La mathématique est efficace dans la mesure où elle suggère la réalisation d'observations ou d'expérimentations et fournit des résultats numériques qui, à une certaine marge d'erreur près, rejoignent les résultats empiriques issus de ces observations ou de ces expérimentations.

Une capacité rétrodictive

La mathématique est efficace parce qu'elle reproduit des résultats déjà connus en les organisant dans un formalisme concis. La mathématique fournit ici des outils servant seulement à « sauver les phénomènes ». Par exemple, grâce à la méthode des moindres carrés, on recherche des courbes passant au plus près des points expérimentaux.

Une capacité explicative

Pour qu'une théorie mathématique soit vraiment efficace en science, il faut qu'elle rende manifeste une explication des phénomènes, c'est-à-dire une suite d'inférences reliant leurs descriptions à des principes reconnus comme fondamentaux. Cette capacité explicative va de pair avec une capacité unificatrice (expliquer, c'est ramener la diversité des phénomènes à un très petit nombre de principes) et une capacité générative (suggérer des concepts nouveaux, des stratégies nouvelles).

En résumé, une mathématique efficace est un formalisme doué de capacités prédictives, rétrodictives et explicatives; autrement dit un langage permettant de décrire, d'expliquer et de maîtriser les phénomènes.

ATTENTION !

Si nous avons l'espoir que cette incroyable efficacité, que nos qualités de logique, de clarté devraient aider la science de l'information, il peut aussi avoir une contamination en sens inverse. Dans la mesure où la culture mathématique est imposée de façon artificielle, de l'extérieur, sans qu'il y ait – comme ce fut le cas en physique – de véritable exigence interne, les mathématiques perdent de leur caractère de sûreté puisqu'elles s'appliquent en définitive sur n'importe quoi et n'importe comment[3]. L'exigence en physique impose de repérer des régularités qu'on représente par des fonctions analytiques simples et d'exiger de bons ajustements. Alors qu'en bibliologie, discipline avatar de la bibliométrie, la tendance est plutôt la recherche de la corrélation même faible en s'en tenant au minimum de maths nécessaires.

Plus que partout ailleurs peuvent jouer l'esbroufe, la manière de faire croire que l'on comprend mieux que l'autre, les connivences entre initiés (les matheux) qui comprennent par-dessus la tête de ceux qui ne comprennent pas (les non-matheux).

Quelles sont alors les mathématiques efficaces pour décrire, expliquer et maîtriser les phénomènes informationnels ? Que représente la branche mathématique de l'infométrie et quelles sont les principales applications de mathématique infométrique ? Ce sera l'objet de notre première partie.

---

[3] XIRDAL Zéphirin – Mathématiques et sciences humaines – Union libre ou mariage forcé – *Impascience*, 4/5, printemps 1976.

- II - L'INFOMÉTRIE MATHÉMATIQUE

Quelles sont les premières applications des mathématiques à l'étude des phénomènes informationnels? Elles vont constituer la branche mathématique de l'infométrie, branche que nous appelons infométrie mathématique. Ferons partie de cette branche les applications de ces mêmes mathématiques aux bibliothèques (bibliométrie mathématique), aux médias (médiamétrie mathématique), au WEB (webométrie mathématique), à la recherche-développement (scientométrie mathématique) et aux musées (muséométrie mathématique).

Les applications mathématiques peuvent prendre en compte une information ou un ensemble d'informations.

2.1. - une information :

- La fonction puissance et la mesure de la fréquence des mots dans un texte (loi de Zipf)

Les fonctions polynomiales simples sont bien connues :

$$y = x^m$$

où l'exposant *m* est un nombre entier positif ou négatif.

x^m signifie que l'on fait :

- ***m* fois le produit de *x* si *m* est un entier positif : c'est la fonction puissance,**
- ***m* fois l'inverse de ce produit si *m* est un entier négatif : c'est la fonction hyperbolique[4]. Quel que soit *m* entier positif, on a :**

$$y = x^{-m} = \frac{1}{x^m}$$

<u>Application</u> :

Ce qui caractérise un certain nombre de phénomènes informationnels, ce sont des comportements de nature hyperbolique[5], c'est-à-dire que le produit de puissances fixes des variables est constant :

$$F(x).x^n = constante$$

[4] G.K. Zipf, *Human behavior and the principle of least effort*, Cambridge, Addison-Wesley, 1949 (Reprinted Hafner, New York, 1965).

[5] En science de l'information, on a l'habitude d'appeler fonction hyperbolique toute fonction puissance ayant un exposant négatif, qu'il soit entier ou non.

Dans leurs manifestations discrètes, cela se traduit par le fait qu'à une cause croissant de façon géométrique correspond un effet croissant de façon arithmétique.
Ainsi, le nombre d'occurrences de tout objet dans un ensemble, par exemple un livre dans une collection ou un mot dans un texte, obtenu par comptage, est appelé fréquence. Si on ordonne les objets en fonction de leur fréquence décroissante, on peut leur attribuer un rang. Plusieurs objets ayant la même fréquence auront des numéros d'ordre consécutifs. Les propriétés des courbes (rang/fréquence) ont été observées et étudiées dans des domaines très variés. Dans les années 50, George Zipf s'est intéressé à la fréquence des mots dans les textes. Il a observé une relation constante, de type hyperbolique, entre la fréquence et le rang des mots :

Rang. Fréquence = constante (notée k)

La relation entre rang et fréquence est de type puissance inverse d'exposant $b \geq 0$:

$$U(r) = \frac{k}{r^b}$$

où *U* représente la fréquence et *r* le rang.

- La fonction exponentielle et l'obsolescence de l'information :

La fonction exponentielle est parfois appelée « fonction de croissance naturelle » car de nombreux processus naturels, comme la croissance d'une forêt, d'une population ou du nombre des publications scientifiques, varient de façon exponentielle.
La fonction exponentielle dite de base *e*(*e*=2,72828…, constante d'Euler) est notée :

$$exp(x) = e^x$$

Application :

Corollaire de la croissance rapide du nombre de publications, il existe une obsolescence également rapide du stock d'informations disponibles. Ce qui veut dire que si les références à la littérature passée sont distribuées de façon aléatoire, sans rapport avec la date de publication, une majorité d'entre elles renvoie à des travaux récents, puisqu'il y a plus d'articles disponibles pouvant être cités :

$$C(t) = C(0)e^{-at}$$

où *a* est un nombre positif supérieur à 1 (figure 1).

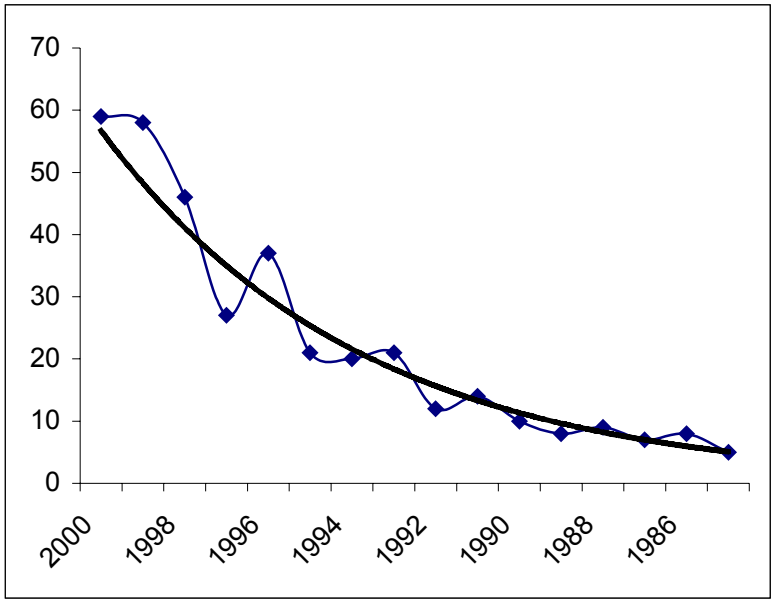


Figure 1 –Obsolescence de l'information

Les recherches sur la demi-vie des littératures scientifiques fournissent des éléments permettant d'éclairer ce type d'interrogation. La demi-vie d'une littérature est le temps pendant lequel la moitié de la littérature active a été citée. Les études d'obsolescence des différentes littératures ont montré des variations importantes de cette caractéristique : 4,6 années en physique, 7,2 années en psychologie, 10,5 années en mathématiques. De façon identique, connaissant le nombre total de citations reçues par une revue, la demi-vie de cette revue mesure le nombre d'années pendant lesquelles elle a reçu 50 % de ces citations. À titre d'exemple, voici les valeurs de ces demi-vies pour quelques revues de science de l'information :

Revues	Demi-vies (années)
J AM SOC INFORM SCI	6,8
SOC STUD SCI	9,6
SCIENTOMETRICS	5,1
INFORM PROCESS MANAG	6,8
J INFORM SCI	6,2

Tableau 1 : Demi-vie des revues en science de l'information (année 1999) (source JCR)

2.2 - un ensemble d'informations:

- La logique classique booléenne et le repérage de l'information:

La logique classique booléenne du nom du mathématicien George Boole (1815-1864) (encore appelée logique mathématique) identifie, sur des ensembles finis, trois relations de dépendance grâce aux opérateurs booléens ET, OU et NON. Ces trois opérateurs permettent d'effectuer les importantes opérations ensemblistes (figure 3) que sont respectivement l'intersection, l'union et le complémentaire.

ET **(produit logique) relie les composantes d'une phrase,**

OU **(somme logique) relie les termes synonymes ou quasi synonymes,**

NON **(négation logique) élimine les termes.**

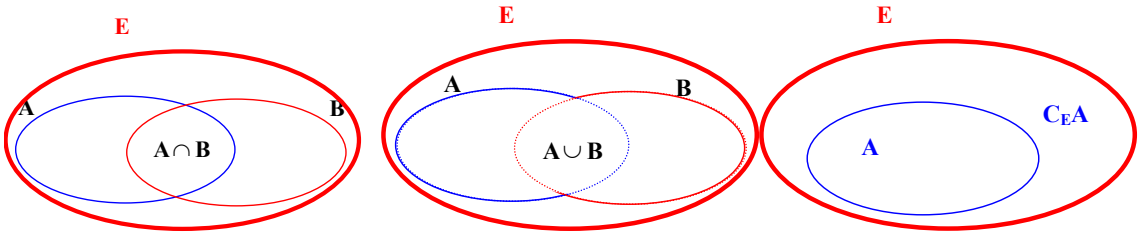


Figure1 – Opérations ensemblistes

ATTENTION, le OU utilisé ici est le « ou » logique et non pas le « ou » exclusif utilisé dans le langage courant.

Application :

Un exemple d'équations de recherche booléenne lors d'une interaction informationnelle personne-ordinateur (P-O) (U représente l'usager et O l'ordinateur)

U - question 1 = "Qu'avez-vous sur l'esclavage aux Etats-Unis?"
interrogation = (slave?) and (United(w)States) or America?)
O - réponse 1 = 2504 références

U - question 2 = " et sur les soulèvements des esclaves dans le Sud avant la guerre de sécession?"
interrogation = (slave?) and (rebellion? or uprising?) and (south?) and HP=1800h)
O - réponse 2 = 21 références

U - question 3 = "plus précisément, sur l'effet de la rébellion de Nat Turner en Virginie?"
interrogation = Nat(w)Turner and Virginia

O = réponse 3 = 13 références[6].

- Les vecteurs et la similitude entre questions et réponses :

Dans l'espace à trois dimensions de la géométrie euclidienne, on appelle vecteur un segment de droite orienté. Si ($a_1, a_{1,...}$ a_1) est un point dans cet espace, alors la ligne qui va de l'origine (0,0,…0) à ce point est le vecteur. Il est représenté par une flèche.

Application :

Comment peut-on mesurer la proximité de deux ensembles informationnels qui sont définis selon plusieurs critères ? Un des modèles de description possible des ensembles est celui des espaces vectoriels, développé par Salton[7].

Soit un ensemble *D* de documents et *M* l'ensemble des *m* mots {M_1, M_2…. M_i,…M_m} présents dans les documents. Chaque document sera représenté sous la forme d'un vecteur ayant *m* composantes :

Document A : $\vec{A} = [a_1 \quad a_2 \quad \ldots \quad a_m]$

Document B : $\vec{B} = [b_1 \quad b_2 \quad \ldots \quad b_m]$.

Dans un espace à trois dimensions, les documents seront donc représentés de la façon suivante :

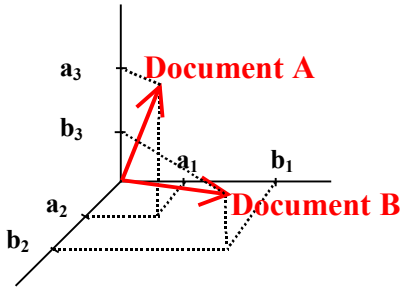


Figure 4 – Représentation vectorielle des documents A et B dans un espace à trois dimensions

Les valeurs a_i et b_j sont les « poids » des mots M_i et M_j présents dans les documents *A* et *B*. Ils quantifient la manière dont *A* et *B* sont représentés par ces deux mots.

Ce type de modèle a été utilisé pour calculer la proximité d'une question (composée de *m* mots) et d'un document, et pour calculer la proximité de deux documents.

Pour déterminer cette proximité, on calcule le cosinus de l'angle que forment les deux vecteurs documents entre eux :

---

[6]KENNEDY L., COLE C., CARTER S. - Connecting on-line strategies and information needs: a user-centered focus labeling approach - RQ, 36, 4, 1997.

[7] G. Salton and M.J. McGill, *Introduction to modern information retrieval,* New York, McGraw-Hill, 1984.

Le cosinus ou coefficient de Salton : $Cos(\vec{A},\vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\left\|\vec{A}\right\|.\left\|\vec{B}\right\|} = \frac{\sum_{k=1}^{m} a_k b_k}{\sqrt{\sum_{k=1}^{m} (a_k)^2} \sqrt{\sum_{k=1}^{m} (b_k)^2}}$

$\vec{A} \bullet \vec{B}$ **est le produit scalaire des vecteurs** $\vec{A}\ et\ \vec{B}$ **et** $\left\|\vec{A}\right\|\ \ et\ \left\|\vec{B}\right\|$ **désignent la norme euclidienne des vecteurs** $\vec{A}\ et\ \vec{B}$.

- III - LA STATISTIQUE S'APPLIQUE

La statistique, une branche de la mathématique, s'applique à l'analyse des valeurs numériques ; en particulier, celles pour lesquelles une étude exhaustive est impossible, à cause de leur grand nombre et de leur complexité. La valeur statistique obtenue pour une variable est une estimation de la valeur vraie de cette variable. Une fois collectées, les valeurs numériques devront être analysées de façon à les mettre en ordre, à leur donner un sens :

- **l'analyse peut être simplement descriptive, donnant par exemple un état des usages faits de l'information ou du système d'information par les usagers. On fera alors appel à la statistique descriptive.**
- **l'analyse peut être aussi interprétative, permettant de dire ce que signifient ces valeurs. C'est alors la statistique bidimensionnelle qui décrit et mesure la liaison entre deux variables informationnelles et à la statistique multidimensionnelle qui décrit les relations existant entre trois et plus de trois variables informationnelles.**

Le dimensionnement de ces analyses sera différent selon que l'on a en vue un travail consistant, c'est-à-dire de recherche approfondie, ou une évaluation rapide. Dans le premier cas, recherchant dans les valeurs des relations qui permettront d'infirmer ou de confirmer les hypothèses formulées, il sera nécessaire de travailler avec un grand nombre de variables informationnelles. Dans le second cas, on aura seulement besoin d'une analyse à deux ou trois dimensions. La démarche traditionnelle statistique qui consiste à confirmer les hypothèses formulées a considérablement évolué avec la généralisation d'outils d'analyse statistique multidimensionnelle (encore appelés en France analyse de données) qui, en particulier grâce aux outils infographiques, permettent de formuler des hypothèses que l'on vérifiera ensuite en utilisant d'autres méthodes, comme les statistiques exploratrices ou « fouilles de données » (texte mining, data mining, Web mining).

En résumé, une statistique efficace fournit des méthodes descriptives, interprétatives et exploratrices permettant d'évaluer la validité des modélisations des phénomènes informationnels qu'elle propose.

ATTENTION, ce peut être un moyen de mentir ! Stade suprême de l'impérialisme mathématique, la statistique prétend formaliser la démarche scientifique en proposant des règles pour évaluer la validité

d'un modèle. Il est, bien entendu, que l'on peut développer toutes sortes de modèles statistiques autour des phénomènes sociaux et en particulier des phénomènes informationnels. Mais ce qui est suspect, c'est cette tendance à la complication non nécessaire. C'est aussi la pénombre discrète où on laisse l'évaluation des limites d'un modèle.

Pourtant un des mérites de l'attitude scientifique classique est de connaître ses propres limites. Ici, les insuffisances, quand elles sont reconnues, sont justifiées par le fait qu'il s'agit des débuts d'une nouvelle science[8]. Prédiction et analyses sont faites dans le flou[9].

Quelles sont alors les statistiques efficaces pour décrire, expliquer et maîtriser les phénomènes informationnels ? Et que représente la branche statistique de l'infométrie et quelles sont les principales applications de statistique infométrique ? Ce sera l'objet de notre deuxième partie.

- IV - L'INFOMÉTRIE STATISTIQUE

Quelles sont les premières applications des statistiques à l'étude des phénomènes informationnels? Elles vont constituer la branche statistique de l'infométrie, branche que nous appelons infométrie statistique. Ferons partie de cette branche les applications de ces mêmes statistiques aux bibliothèques (bibliométrie statistique), aux médias (médiamétrie statistique), au WEB (webométrie statistique), à la recherche-développement (scientométrie statistique) et aux musées (muséométrie statistique).

Les applications statistiques peuvent prendre en compte une variable informationnelle, deux variables informationnelles ou une multiplicité de variables informationnelles.

- *- une variable informationnelle :*

La statistique unidimensionnelle fournit des méthodes et des procédures permettant de résumer des grands ensembles de valeurs numériques d'une variable afin de les rendre intelligibles, de communiquer l'essence de ces valeurs.

- Les taux et l'évaluation des produits et des services d'information :

Le taux de croissance (ou de décroissance) est une catégorie de taux particulièrement intéressante. Il est calculé en déterminant la différence entre la valeur d'une variable au début d'une période donnée et sa valeur à la fin de cette période et en divisant cette quantité par la valeur de la variable au début de la période.

---

[8] XIRDAL Zéphirin, op. cité.

[9] Exception notoire: les fourchettes des pronostics électoraux, un des grands jeux de la télévision technocratique ! Les experts se portent bien mais s'en tirent mal comme on l'a vu en 2002. Du fait même qu'ils sont des experts, il y a des choses que les experts ne peuvent pas prévoir. Ce qui n'empêche pas qu'ils peuvent aussi causer des dégâts.

Application :

Le taux de croissance d'un service en ligne qui est passé de 5 000 connexions en 1997 à 15 000 en 2002 est de :

$$Taux\,de\,croissance = \frac{15000 - 5000}{5000} = 2$$

En pourcentage, le nombre de connexions s'est accru de 200 % en 5 ans, soit 40 % par an. Le nombre de connexions a été multiplié par 3. Mais attention, il n'y a pas 300 % d'augmentation !.

- *- deux variables informationnelles :*

La statistique bidimensionnelle est plus audacieuse et donc plus risquée. Elle permet de découvrir les liens qui existent entre deux de ces variables.

- La co-occurrence et les cartographies informationnelles

Considérons un ensemble d'articles scientifiques où chacun est caractérisé par différents mots. Nous ne connaissons a priori ni ces mots, ni leur nombre. Les premiers traitements simples que l'on peut faire sont d'établir la liste des mots utilisés et de calculer leurs fréquences (nombre d'occurrences), puis de s'intéresser à la co-occurrence de deux mots, c'est-à-dire au nombre de fois qu'ils apparaissent ensemble dans un texte. Si les mots sont ainsi associés, les intérêts des auteurs des articles le sont aussi.

Le rôle des mots en tant qu'opérateurs de l'auto-structuration des domaines scientifiques et techniques a été en effet mis en évidence. Les mots indiquent quels sont les sujets intéressants dans un domaine de recherche donné à un moment donné. Lorsque deux mots apparaissent simultanément dans un ensemble d'articles, les sujets qu'ils représentent sont associés. Les schémas d'association des mots permettent donc de mettre en évidence les tendances de la recherche, ainsi que les principaux centres d'intérêt des chercheurs (figure 5).

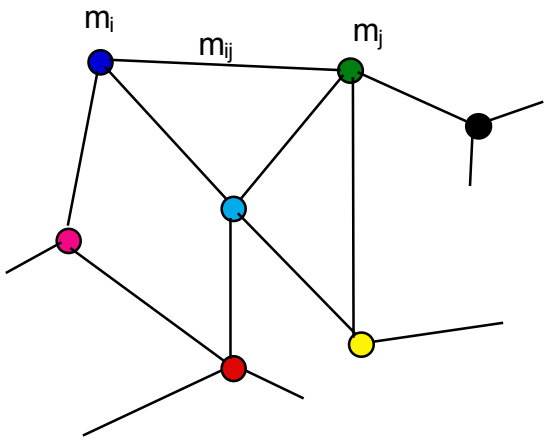


Figure 5 – Réseau d'associations de mots

Pour construire le réseau des associations de mots, la première étape consiste à calculer le nombre d'occurrences m_i de chaque mot i dans l'ensemble d'articles et le nombre de co-occurrences m_{ij} de chaque paire de mots m_i et m_j . Cependant, la co-occurrence ne permet pas à elle seule de mesurer la force des associations entre les mots, car elle avantage les mots apparaissant un grand nombre de fois par rapport aux autres. On calcule donc un coefficient normalisé (c'est-à-dire dont les valeurs sont comprises entre 0 et 1), croissant avec le nombre de co-occurrences, appelé le coefficient d'association noté E_{ij} . Un coefficient d'association égal à 1 signifie que les mots i et j sont systématiquement trouvés ensemble ; un coefficient à 0 signifie au contraire qu'ils ne sont jamais ensemble dans un document. Il existe plusieurs méthodes de calcul d'un coefficient d'association. Le coefficient utilisé dans la méthode des mots associés est :

$$E_{ij} = \frac{m_{ij}^{2}}{m_i . m_j}$$

Ce coefficient varie entre 0 et 1. Il vaut 0 si les mots i et j n'apparaissent jamais simultanément et 1 dans le cas inverse.

On trouvera sur la figure 6 un graphe des mots portant sur les revêtements céramiques ; les textes analysés proviennent d'une banque de brevets et sont constitués des titres et des résumés de 16 000 brevets extraits de cette banque.

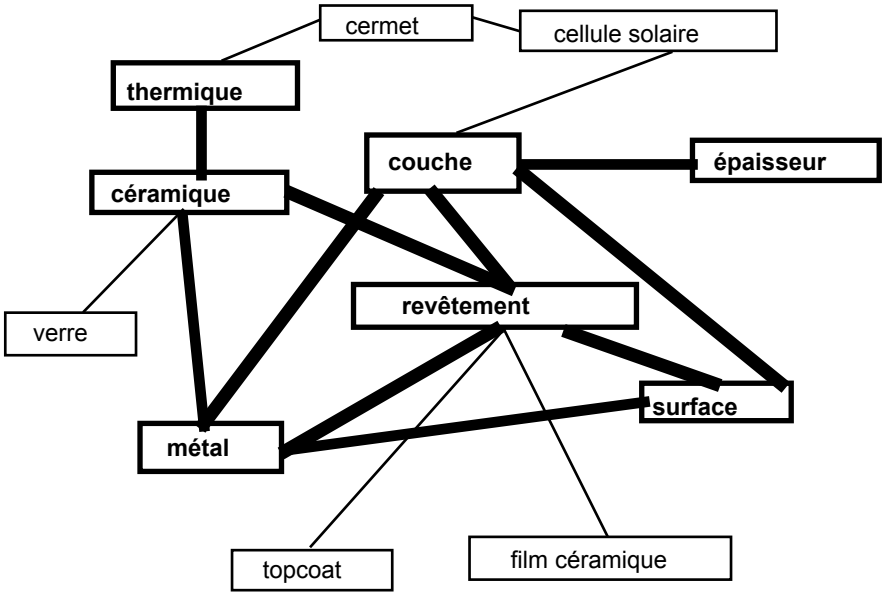


Figure 6 – Graphe « revêtement céramique »

Dans ce graphe, l'épaisseur des traits entre les mots est proportionnelle à l'intensité de liaison des associations. Chaque amas ou cluster est alors caractérisé par les mots du thème et ses associations internes et externes.

Cette classification permet ensuite de construire une représentation cartographique originale appelée diagramme stratégique. La figure 7 représente le diagramme stratégique « revêtement céramique ».

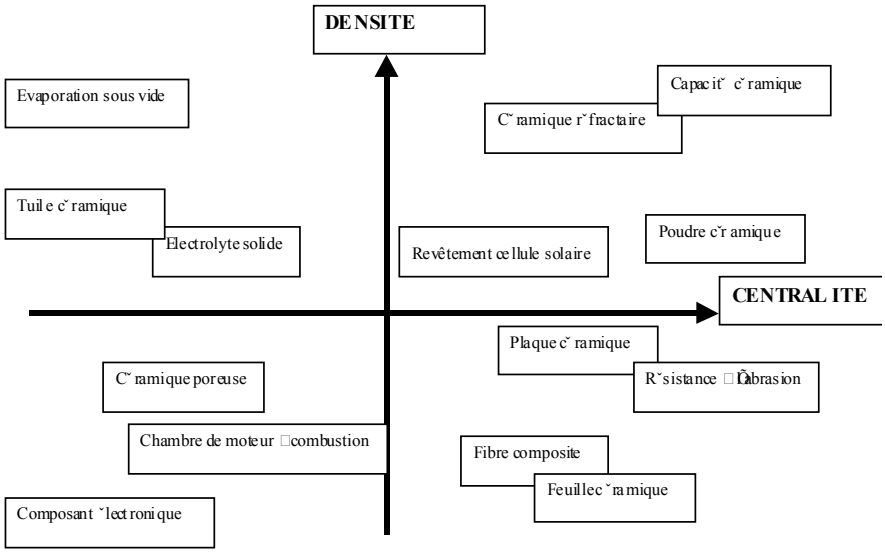


Figure 7 – Diagramme stratégique « revêtement céramique »

Les thèmes de recherche sont ainsi représentés dans un diagramme bidimensionnel défini par deux indicateurs : la densité et la centralité.

- La densité d'un amas est la moyenne de ses associations internes : c'est un indicateur de sa cohérence.

- La centralité d'un amas est la moyenne de ses associations externes : elle indique si l'amas est plus ou moins isolé.

Les sujets qui apparaissent dans le quadrant en haut à droite sont les sujets qui sont dans le front de la recherche (très structurés, très développés). Par contre, ceux qui se situent dans le quadrant en bas à gauche sont des sujets périphériques et faiblement structurés.

- *– de multiples variables informationnelles :*

La statistique multidimensionnelle permet de rechercher les relations existant entre plusieurs variables. Elle fournit des outils dont on attend d'abord qu'ils aient une efficacité pratique, la justification théorique n'étant recherchée qu'en second. On distingue trois grands types de méthodes : la classification, l'analyse factorielle des correspondances et l'analyse relationnelle. Par exemple :

- **L'analyse factorielle des correspondances et la typologie des productions d'informations scientifiques techniques selon les disciplines :**

L'analyse factorielle consiste à traiter des grands tableaux de nombres, difficiles à lire, en les remplaçant par des tableaux plus simples qui soient une bonne approximation des premiers. Les diverses méthodes d'analyse factorielle (en particulier l'analyse factorielle des correspondances, que nous étudions ci-dessous) emploient le même procédé : étant donné un nuage de points (les individus), munis de masse (les effectifs), dans un espace dont le grand nombre de dimensions interdit la visualisation du nuage, espace muni d'une métrique (qui mesure la distance entre les individus), il s'agit de trouver les axes d'inertie du nuage et d'obtenir des visualisations sur des plans formés par les couples d'axes. On pourrait résumer ceci en disant que ces méthodes d'analyse permettent de « géométriser des tableaux de nombres ».

<u>Application :</u>

La production de littérature scientifique et technique des chercheurs d'une université est ventilée dans les quatre grandes disciplines de la façon suivante :

	Article	Livre	Brevet	*Total*
Sciences de la matière	**13**	**2**	**5**	**20**
Sciences de la vie	**20**	**2**	**8**	**30**
Sciences sociales	**10**	**5**	**5**	**20**
Sciences de l'ingénieur	**7**	**1**	**22**	**30**
Total	**50**	**10**	**40**	**100**

Tableau 2 – Production de littérature scientifique et technique

La lecture de ce tableau nous apprend que 50 % de la production scientifique des chercheurs est faite d'articles *(ligne 5, colonne 1)*. Si on applique ce pourcentage aux sciences de la matière, on constate que, sur une production totale de 20, il ne devrait y avoir que 10 articles :

$$\frac{20.50}{100} = \mathbf{10}$$

Or, il y en a 13. Les scientifiques des sciences de la matière produisent en priorité des articles. Donc, il n'y a pas indépendance entre la discipline et le type de production choisi : le respect de la proportion moyenne correspond à ce que l'on appelle la situation d'indépendance.

Grâce *à l'analyse factorielle des correspondances*, on obtient une représentation de la typologie de la production de littérature scientifique et technique selon les disciplines :

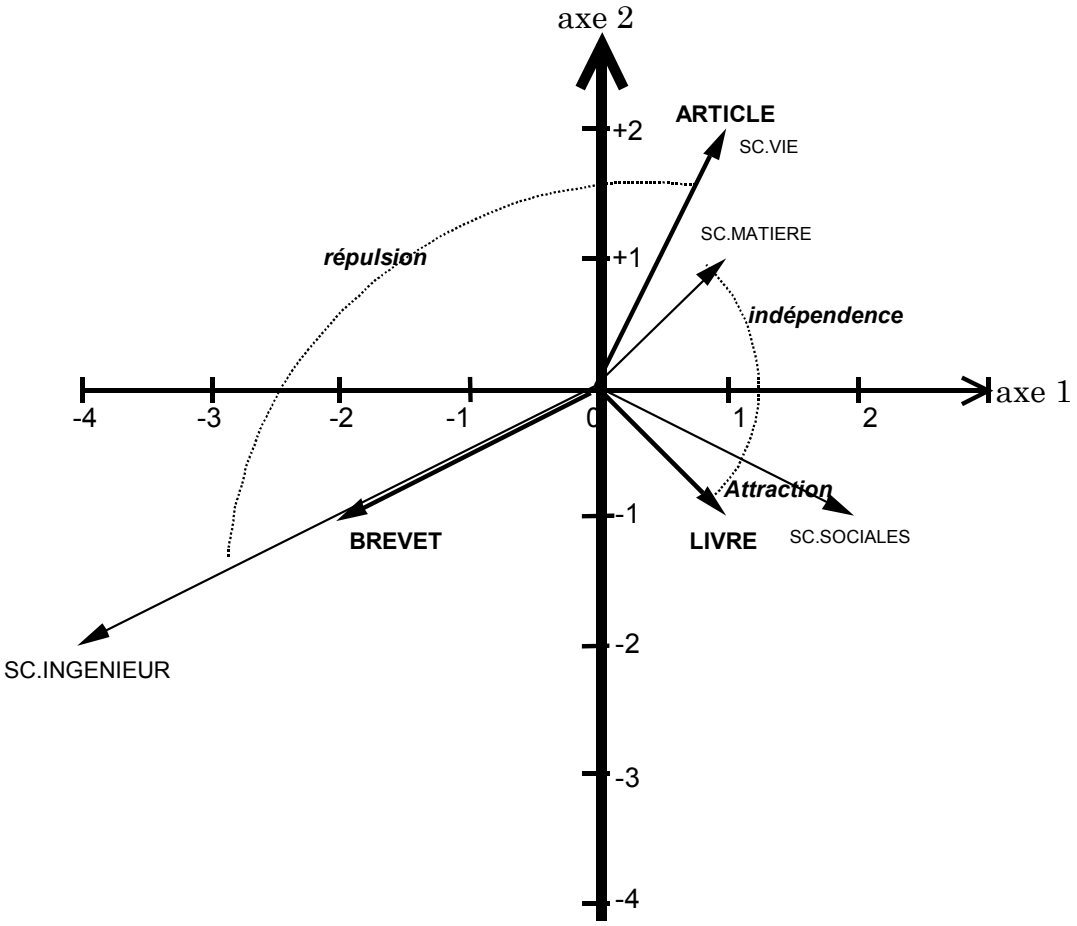


Figure 8 – La littérature scientifique et technique selon les disciplines
(Plan des axes 1 et 2)

- CONCLUSION

Alors que les techniques qui produisent de l'information, les techniques qui la mémorisent, les techniques qui la véhiculent frôlent tous les jours les limites de l'infiniment petit et les limites de l'infiniment grand, mathématique et statistique nous permettent d'explorer plus facilement ces univers inconnus de l'information. Elles nous aident aussi à mieux saisir cette information devenue de ce fait infiniment

croissante, infiniment rapide et infiniment complexe pour mieux maîtriser sa production, sa communication et son usage. Les exemples que nous avons présentés ne sont qu'un premier pas dans le sens d'un engagement plus profond de l'outil mathématique en Science de l'information.

Les développements actuels des activités scientifiques, techniques et industrielles dans les différents secteurs de l'information et de la culture laissent présager un usage plus intensif de cet outil mais aussi, il faut le souhaiter, la découverte de nouvelles méthodes, de nouvelles lois et de nouvelles techniques mathématiques et statistiques encore mieux adaptées à l'objet information.

A côté des diverses cultures qu'elle incorpore jusqu'à maintenant, la Science de l'information ajoute une culture que peu attendaient peut-être, la culture mathématique.

REFERENCES

BOUDON Raymond – L'analyse mathématique des faits sociaux – Plon, Paris,1970.

BORGMAN C.L. (ed.) – Scholarly communication and bibliometrics – Sage Publications, London, 1990.

BOYCE B.R., MEADOW C.T., KRAFT D.H. – Measurement in information science – Academic Press, San Diego, 1994

CALLON M., COURTIAL J.P., PENAN H. – La scientométrie – Que sais-je ?, PUF, Paris, 1993.

COLEMAN James S. – Introduction to mathematical sociology – The Free Press, New York, 1964

EGGHE L., ROUSSEAU R. - Introduction to informetrics : quantitative methods in library, documentation and information science, Elsevier, Amsterdam, 1990.

ELKANA Y. (ed.) – Towards a metric of science – John Wiley & sons, New York, 1978.

LAFOUGE T, LE COADIC Y.F., MICHEL C. – Eléments de Statistique et de Mathématique de l'information : infométrie, bibliométrie, médiamétrie, scientométrie, muséométrie, webométrie. – Les Presses de l'ENSSIB, Lyon, 2001.

MISE EN PLACE D'UN SYSTEME DYNAMIQUE ET INTERACTIF DE GESTION D'ACTIVITE ET DE CONNAISSANCES D'UN LABORATOIRE (PROJET GACO LAB)

Leitzelman Mylène
Intelligence Process SAS
Espace Rossignol, Bd Abbadie, 13730 Saint Victoret

Kister Jacky
UMR 6171 S.C.C
Faculté des Sciences et Techniques de St Jérôme - Boite 561 - 13397 Marseille Cedex 20

Résumé : Il s'agit de mettre en place de façon expérimentale et pour le compte de l'UMR 6171 associé au CRRM, un système interconnecté de gestion d'activité et de connaissances pour gérer l'activité scientifique d'une unité de recherche. Ce système sera doté de modules de visualisation synthétique, statistiques et cartographiques s'appuyant sur des méthodologies de datamining et de bibliométrie. Le point clé de ce système sera de proposer en même temps un outil de gestion stratégique et d'organisation d'un laboratoire et un outil permettant la compilation interlaboratoires pour en faire un outil d'analyse ou de stratégie à une plus grande échelle en laissant des accès plus ou moins libres pour que des agents extérieurs puissent à partir des données générer des indicateurs de performance, de valorisation, de qualité des productions scientifiques et de relations laboratoire/entreprises.

Abstract : The goal of this experimental project in association with UMR 6171 and the CRRM, is to implement an interconnected Knowledge Management System in order to manage the scientific production of a laboratory. This dynamic system will operate with datamining and bibliometric modules. The gist of this project is on the one hand to create a strategic and organizational management tool and on the other hand to generate a system that can gather interlaboratories informations to help external actors to generate different indicators of relevance comparing scientific production, relation with industries, …

Mots-clés : Gestion des connaissances, KM, système dynamique de bases de données, bibliométrie, génération d'indicateurs, évaluation de la production scientifique

Keywords : Knowledge Management, Dynamic Database System, bibliometry, indicator of relevance

Mise en place d'un système dynamique et interactif de gestion d'activité et de connaissances d'un laboratoire (projet GACO LAB)

INTRODUCTION

Les laboratoires sont soumis à des demandes nombreuses sur la rédaction de rapports d'activité et de production scientifique tant de la part de leur propre hiérarchie, des écoles doctorales, ... que de la part des partenaires privés ou des financeurs publics. Ils doivent aussi répondre à des appels d'offre nationaux ou internationaux pluridisciplinaires ou pluri-états.

Cette pression sur un compte rendu permanent de leurs activités de recherche les oblige à une organisation des connaissances, de leur gestion et d'une interopérabilité avec l'existant (autres laboratoires, partenaires,...).

A ce jour, il existe des systèmes isolés répondant à des besoins ponctuels, développés spécifiquement dans quelques unités, le site internet ou intranet des universités (cf. l'intranet de l'Université Aix Marseille II) ou encore le site du CNRS, très généraliste et qui n'a pas vocation à aider un laboratoire à gérer sa production.

Il n'existe donc pas à proprement parler de système intégrant les potentiels des différents laboratoires, pour la gestion des différents documents propres à l'activité d'un laboratoire. Actuellement la perception des équipes de recherche et des compétences est très floue et la multiplication des projets incomplets ou non actualisés nuit au développement régional ou même à la perception d'une institution.

Suite à ces constatations, l'UMR6171 projette de se doter d'un système interconnecté de gestion d'activité et de connaissances pour gérer l'ensemble de son activité scientifique avec comme spécificités des modules de visualisation synthétique, statistiques et cartographiques s'appuyant sur des méthodologies de datamining et de bibliométrie.

Nous verrons dans une première partie l'importance théorique d'évaluer la recherche scientifique, aux vues des évolutions technologiques et économiques actuelles où les sphères de l'économie, du politique et de la science doivent collaborer étroitement pour favoriser l'innovation et la croissance économique. Nous aborderons dans une seconde partie les points pertinents du système GACO LAB et dans une dernière partie nous dresserons le plan d'action d'un tel projet.

1 - LA QUESTION DE L'EVALUATION DE LA PRODUCTION SCIENTIFIQUE

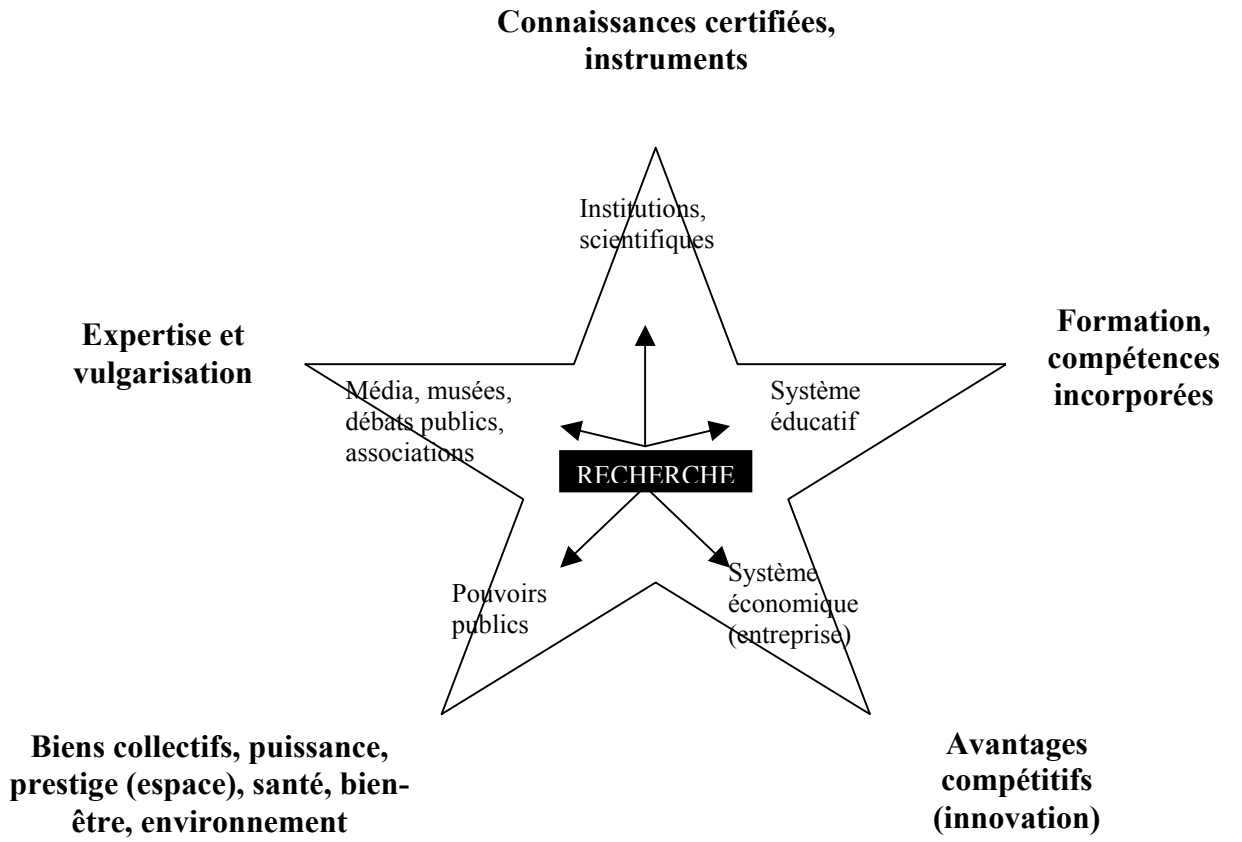
Force est de constater que la Science et la Technologie sont aujourd'hui les ressources clés du développement économique [Mansfield 91]. Les incertitudes actuelles et les problèmes devenus structurels de la société de l'immatériel et de l'information ne peuvent se résoudre dans la sphère économique que par un nouveau cycle d'innovation qui lie plus étroitement l'industrie, le politique et la science. D'où l'inspiration du projet GACO LAB de mettre enfin en place un outil de gestion stratégique et d'organisation d'un laboratoire pour en faire un outil d'analyse ou de stratégie à une plus grande échelle en laissant des accès plus ou moins libres à des agents extérieurs pour qu'ils puissent à leur tour générer des indicateurs de performance et s'approprier les retombées scientifiques du laboratoire.

1.1 - Evaluer la production scientifique : pourquoi faire ?

Les récents travaux de Jevons [Jevons 93] et Ziman [Ziman 92] montrent que, de plus en plus, il va être nécessaire au plan régional (comme au plan national) de s'approprier les savoirs et les compétences locales, pour d'une part bénéficier des retombées des travaux universitaires de recherche et d'autre part pour créer une médiation implicite vers la population locale.

Pour reprendre les termes de Michel Callon, "la recherche devient une activité 'totale' et ouvertement 'multidimensionnelle' qui doit contribuer simultanément à la production de connaissances certifiées, de biens collectifs, d'avantages compétitifs, de compétences professionnelles, mais aussi d'une culture et de décisions partagées par le plus grand nombre" [Callon 95].

L'idéal serait que la recherche irrigue les cinq facettes de la société, le monde de la science, le monde de l'éducation, le monde des entreprises, le monde des citoyens et le monde des pouvoirs publics. C'est ce que reprend le schéma de "la rose des vents de la recherche" qui rend compte des relations entre sciences, techniques et sociétés.



1.2 - Objectif visé : la mise en relation des acteurs économiques

Le projet GACO LAB vise à mettre en relation les divers acteurs économiques présents régionalement, voire à l'échelle nationale, fédérés autour d'un système permettant la lisibilité de la production scientifique d'un laboratoire et dont les retombées scientifiques peuvent être appropriées par les autres acteurs économiques que ce soit le politique et l'industriel.

Retombées pour les laboratoires

Aujourd'hui, les universités ont besoin de transférer leur savoir et leurs compétences vers l'industrie pour s'assurer une marge financière de sécurité, elles peuvent d'un autre côté approfondir leurs activités de recherche fondamentales. Comme les fonds publics deviennent moins importants avec la crise, ces dernières doivent trouver des sources de revenus alternatives. Même si ce débat sur l'ouverture de la recherche au monde de l'entreprise reste encore tabou en France, il n'en est pas moins actuel.

Par exemple, à l'étranger (Angleterre, Hollande, USA, Allemagne,...) comme dans certaines universités françaises, les transferts de technologies vers l'industrie prennent plusieurs formes, communications (conférences et publications des recherches), activités de consultant, ouverture de formations aux industriels, transfert direct de technologie grâce à la vente de licences de brevets, de copyright ou d'autres types de propriétés intellectuelles.

Le fait aussi que des acteurs de la sphère publique accèdent aux productions scientifiques d'un laboratoire peut aussi engendrer plusieurs avantages : possibilité de contrats de collaboration avec les services publics, réalisation d'études conjointes, valorisation de travaux de recherche, placement d'étudiants,...

Retombées pour les industriels

Les entreprises qui proposent des produits ou des services dans les domaines scientifiques ou technologiques doivent affronter une dure compétition. Ces entreprises doivent entretenir leurs avantages concurrentiels à grands frais de R&D et d'innovations successives car l'obsolescence dans ces domaines est rapide. Les liens avec l'université permettent aux entreprises de développer de nouveaux produits et de nouveaux procédés, qu'elles n'auraient pas pu créer en interne par manque de temps, d'expertise ou d'argent.

Retombées pour les acteurs des collectivités publiques

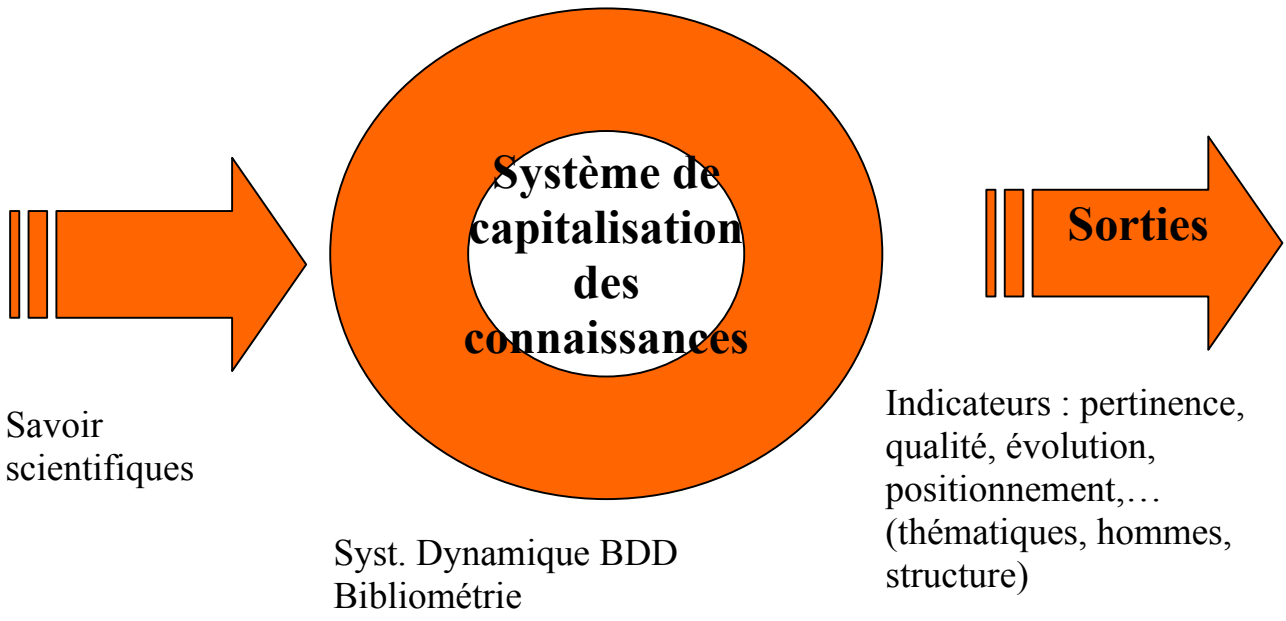
Les collectivités locales trouvent des avantages intéressants dans la collaboration université/entreprise, car cela permet le développement de la compétitivité régionale et nationale face à la concurrence internationale. En effet, les transferts de technologie de la recherche vers l'industrie représentent un important ferment pour l'innovation [Retourna 95], soit en donnant de nouvelles idées d'applications pour l'industrie locale, soit en trouvant des solutions à des besoins techniques précis. De plus, l'amélioration des techniques de transferts de technologie permet d'ouvrir le pays sur l'international, par la mise en valeur de nouveaux débouchés sur les pays demandeurs de technologies comme les pays en développement.

De ce fait, si une collectivité est capable de connaître quelles sont les forces et les faiblesses de sa localité en terme de demande et d'offre de science et de technologies, alors elle peut développer une politique de formation adéquate, ou importer utilement ce qui lui manque. Ce même constat peut être fait à l'échelle nationale. L'Angleterre, dans une directive ministérielle soulignait en 1993 que « les capacités d'innovation du Royaume Uni ne peuvent se développer uniquement si les relations entre la recherche universitaire, les organismes publics de recherche et l'industrie se renforcent » [Preston 93].

2 - AVANTAGES ET PERTINENCE DU PROJET GACO LAB

Le projet GACO LAB apporte plusieurs solutions aux problèmes de gestion et d'animation d'un laboratoire scientifique devant utiliser les Technologies de l'Information et de la Communication.

2.1 - Présentation générale du projet



Il s'agit d'un système d'informations où les entrants sont les connaissances scientifiques des membres du laboratoire, c'est à dire, les références complètes de leur bibliographie (publications scientifiques, conférences, posters, …), les rapports techniques, les contrats de recherche.

Le système repose ensuite sur une architecture logicielle de langage internet (php, mysql, Linux) associé à des gestions de bases de données. Le choix de cette architecture est dicté par le fait que ces logiciels sont libres sans licence d'exploitation et que les universités françaises ont massivement choisi l'environnement Linux pour faire tourner leurs serveurs internet.

Toutes les données étant stockées sous forme de base, il est alors aisé de manipuler ces données avec des modules de traitements bibliométriques et cartographiques afin de mettre en perspective ces informations sous divers angles.

Enfin, les outputs de ce système sont la génération dynamique d'indicateurs, de pertinence du travail de recherche, de qualité, de positionnement, de valorisation industrielle, …, qu'il est possible de fabriquer avec les informations brutes entrées dans les bases de données.

Le tout sera visualisable via une interface Internet, un simple navigateur Internet et accessible de n'importe quel endroit.

2.2 - Système de gestion des connaissances adapté aux besoins de gestion d'un laboratoire

Tout d'abord, le système dynamique et intégré de connaissances scientifique d'un laboratoire GACO LAB offre un gain de temps dans l'élaboration de rapports d'activités et de rapports techniques puisque les sorties envisagées du systèmes génèreront automatiquement de tels rapports d'activités.

Ensuite, ce qu'offre un site internet pour gérer l'activité d'un laboratoire c'est évidemment une visibilité de la production scientifique et technique pour les partenaires publics et privés qui peuvent accéder de manière plus ou moins directe (accès protégés par login et mots de passe) à tout ou partie des informations générales et stratégiques du laboratoire.

Le projet GACO LAB reposant sur une architecture de forme « portail intégré » rend alors possible une lisibilité de l'activité et de la valeur scientifique du laboratoire renforcée par la possibilité de générer des indicateurs propres intégrés dans les rapports d'activités.

2.3 - Interopérabilité du système

Dans la mesure où le système GACO LAB est accessible via Internet, il est envisageable d'interconnecter d'autres plates-formes de gestion, ce qui entraîne la possibilité de mutualisation d'acquisition de moyens coûteux et de partage des recherches scientifiques interdisciplinaires et inter laboratoires.
GACO LAB est un système pour la gestion d'un laboratoire permettant l'extension à tout autre laboratoire le désirant et utilisant la même base.
Enfin, puisque le point fort du système est la génération de statistiques et d'analyses agrégées, il permettra de regrouper les données d'un laboratoire avec, par exemple, les laboratoires d'un même site universitaire pour générer le rapport de la fédération, de l'UFR recherche ou d'une école doctorale dans une discipline ou même de manière plus générale pour l'ensemble des disciplines.

2.4 - Responsabilisation des acteurs

Les chefs d'équipes et les individus sont concernés par l'actualisation en continu de leur propre production ou activité avec le risque de ne pas être bien perçu si cela n'est pas fait

3 - PLAN D'ACTION

3.1 - Présentation des partenaires

Une UMR interdisciplinaire pilote (Chimie et Environnement) : UMR CNRS-CEA 6171, porteur du projet et maître d'œuvre de la solution. Cette UMR est en outre très ouverte aux entreprises, ce qui pourra montrer l'intérêt d'un tel projet pour des acteurs persuadés de la relation entreprises et de la force d'un outil de perception actualisée.

Un laboratoire spécialisé CRRM : apporte son soutien méthodologique en terme de pertinence des indicateurs bibliométriques et de datamining

3.2 - Répartition des rôles

Les phases ci-après ne sont pas disjointes. Certaines se recouvrent et impliquent des personnels communs. De plus, le projet prévoit l'implémentation progressive auprès des labos volontaires au départ. En revanche, les phases ultimes d'extension aux autres labos ne sont pas prises en compte ici : l'objectif est de livrer une solution généralisable, opérationnelle pour un laboratoire test, pour la diffuser ultérieurement.

Acteurs	Rôle phase étude pour l'appel à proposition	Budget temps total	Rôles phase analyse initiale	Rôles phase développement
UMR 6171	Expression de besoins Synthèse rédactionnelle Implémentation	Direction projet	- Maître d'œuvre du projet Régler les répartitions budgétaires (rentrées respectives) et contributives (sorties) - Rédaction du cahier des charges et des spécifications techniques et fonctionnelles Etude terrains des existants	- Contrôle planning, qualité et budget Régler les problèmes conceptuels Validation des spécifications et test grandeur réel dans le laboratoire - Développement, intégration des technologies et modules de traitements, paramétrage,

			et des besoins	intégration des contenus, formation des utilisateurs
CRRM	Donner le cadre théorique	< 10 jours	Préciser le cadre théorique	Fournir un « comité scientifique » tout au long du projet.
Phirama	Expression de besoins des laboratoires	15 jours	Contribuer à l'analyse des besoins et l'inventaire de l'existant	Validation des spécifications et test

CONCLUSION

Le projet GACO LAB s'inscrit en droite ligne dans les besoins désormais urgents de lisibilité et de gestion de l'activité scientifique au regard de la communauté, qu'elle soit scientifique, mais aussi, industrielle et politique.

Les pouvoirs publiques doivent pouvoir savoir à tout moment où se trouve les pôles d'expertises scientifiques de leur région, afin de mieux piloter les systèmes d'aides, de formations et de fertilisation croisée avec le monde de l'entreprise.

De leur côté, les entreprises doivent aussi avoir accès aux avancées scientifiques dont les applications peuvent renforcer leur compétitivité vis à vis de la compétition mondiale.

Enfin, les laboratoires entre eux et en interne doivent pouvoir connaître à l'instant t leur position…

BIBLIOGRAPHIE

- Callon M et all "Réseaux technico-économiques et analyse des effets structuraux " chap 20 dans "La gestion stratégique de la recherche et de la technologie", Economica 1995

- Callon M. et all "La gestion stratégique de la recherche et de la technologie", Economica, 1995

- Jevons F. "The co-location assumption and models of innovation", Science and Public Policy, vol 20 p 51-56, 1993

- Jevons F. "Who wins from innovation?", Technology Analysis and Strategic Management, p 399-412, 1993

- Leitzelman M. "Mise en place d'un Système d'Informations stratégiques multicritères facilitant l'intégration des ressources régionales et la prise de décision dans le domaine de l'Environnement. Application à la Ville de Marseille" Thèse doctorale – CRRM Aix Marseille III, Novembre 98

- Mansfield E. « Academic research and industrial innovation », Research Policy 1991

- Preston JT. "Success factors in Technology Development", TII European Technology Transfert conf, 4 may 1993, ICC, Ghent

- Retourna C. "Analyse de cas concrets d'innovation dans les PME/PMI : problématiques et discussions",Thèse Univ. Aix-Marseille III - CRRM, 1995

- Ziman J. "A neural net model of innovation", Science and Public Policy, vol 18 p 65-67, 1992

INFORMATION, MANAGEMENT ET EVOLUTION SOCIETALE : UNE APPROCHE PAR LA METHODE TRIZ

Cécile Loubet
Maître de Conférences en Génie des Systèmes Industriels
Cecile.loubet@spi-chim.u-3mrs.fr, ✆ (33) 4 91 28 82 73

Joëlle Gazérian
Maître de Conférences en Génie des Systèmes Industriels
gazerian@spi-chim.u-3mrs.fr, ✆ (33) 4 91 28 82 86

Jean-Michel Ruiz
Professeur en Génie des Systèmes Industriels
JM.RUIZ@wanadoo.fr, ✆ (33) 4 91 28 82 46

Henri Dou
Professeur en Sciences de l'information
dou@crrm.u-3mrs.fr, ✆ (33) 4 91 28 80 50

Adresse professionnelle
Université d'Aix-Marseille III, Avenue Escadrille Normandie Niemen, 13397 Marseille Cedex 20

Résumé : Le management d'entreprise doit faire face en ce début du 21ème siècle à l'utilisation conjointe des différents supports de l'information qui ont tout à la fois des conséquences sur le plan des technologies, des organisations et de la culture. La plus ou moins grande aptitude à les utiliser et à les laisser pénétrer dans l'entreprise constitue l'une des contraintes fortes de l'innovation. L'information constitue par ailleurs l'un des sujets les plus captivants d'applicabilité de la méthode TRIZ, méthode relativement proche de l'analyse de la valeur sur le plan de la démarche et qui se singularise par une volonté de recherche systématique des axes créatifs.
Une application de la matrice TRIZ des conflits et contradictions à l'information ouvre des portes de recherche intéressantes en démontrant la possibilité d'une telle transposition.

Summary : At the beginning of this 21st century, companies managment has to face the joint use of various information supports which, at the same time, have consequences on technologies, organisations and culture. The more or less important capacity to use them and to let them penetrate the company is one of the major innovation constraints.
On an other side, information is one of the most attractive TRIZ method applicability, a method fairly near of value analysis in the field of its various steps and which attracts attention by its will of systematic research of creative axis.

An application of the TRIZ matrix for conflicts and contradictions to information open interesting research gates as it shows the possibility of such a transposition.

Mots clés : Systèmes d'information, supports d'information, innovation, créativité, méthode TRIZ, analyse de la valeur.

Key words : Information System, Information Supports, Creativity, TRIZ Approach, Value Analysis.

Information, management et évolution sociétale :

une approche par la méthode TRIZ

Le management d'entreprise doit faire face en ce début du 21ème siècle à l'utilisation conjointe des différents supports de l'information qui ont tout à la fois des conséquences sur le plan des technologies, des organisations et de la culture. La plus ou moins grande aptitude à les utiliser et à les laisser pénétrer dans l'entreprise constitue l'une des contraintes fortes de l'innovation. Sur le plan sociétal, les évolutions se situeront autour de l'intégration progressive des systèmes analogiques dans le monde numérique. Pour autant le support papier demeure, dans la prise de décision, un acte incontournable. La gestion de ces antagonismes ne pourra pas longtemps échapper à une prise de conscience que l'information interagit dans tous les actes du management.
L'information constitue l'un des sujets les plus captivants d'applicabilité de la méthode TRIZ qui depuis environ cinq ans s'avère être l'une des plus puissantes méthodes d'innovation dans le contexte des entreprises privées (tout particulièrement celles américaines et asiatiques). Cette méthode, relativement proche de l'analyse de la valeur sur le plan de la démarche, se singularise par une volonté de recherche systématique des axes créatifs.

Cet article se propose de montrer comment l'histoire de l'information permet de découvrir les principales fonctions afférentes à l'information et aux trois vecteurs qui ont traversé ce dernier siècle : les supports papiers, le média analogique et les technologies numériques.
Puis après une description comparative mais sommaire de l'analyse de la valeur (AV) et de TRIZ, les auteurs s'attacheront à appliquer sur l'information la complémentarité de ces deux approches. Les notions d'information, de communication et de système auront été auparavant débattues. Ces premières réflexions semblent ouvrir des portes sur des voies d'investigation intéressantes.

1 - UN BREF HISTORIQUE POUR SITUER LE CONTEXTE ET LES ENJEUX

Depuis la plus haute antiquité, l'échange d'informations constitue un acte fondamental de l'humanité. Avec l'invention de l'écriture en Mésopotamie, l'information connaît une évolution capitale, celle de pouvoir trouver un support assurant à la fois une conservation de l'information et une possibilité de transmission.

Le Moyen Age a pérennisé cette notion, par le biais des ordres monastiques qui ont donné une qualité à l'information alors essentiellement scientifique. On peut déjà, à cette époque, mettre en exergue deux caractéristiques fondamentales liées à l'information, il s'agit de l'accessibilité et de la disponibilité.

C'est avec l'invention de l'imprimerie que l'on découvre de nouvelles possibilités de diffusion qui s'accentueront au fil des siècles avec notamment :
- la littérature à grande diffusion,
- la conservation de la connaissance par le biais des encyclopédies et,
- les premiers journaux à support papier.

C'est vers la fin du 19ème siècle qu'apparurent les premières grandes évolutions qui toucheront les possibilités d'utiliser de nouveaux média en terme de stockage, de transmission et de diffusion de l'information. La voix et son support ont été les premiers à en bénéficier, grâce notamment à Edison au niveau du support et Graham Bell pour la transmission. Le son, support traditionnel de la voix, change pour la première fois sous la forme d'un support électrique qui autorise une transmission à la fois plus rapide et plus lointaine. L'apparition de la radio et plus tard de la télévision, ne font que confirmer ces premières découvertes en utilisant un support analogique pour la voix et l'image. Les supports magnétiques sont encore loin d'être vulgarisés.

A l'issue de la 2ème Guerre Mondiale, la découverte du transistor, qui semble être par certains aspects une évolution du support analogique, permet par son rôle amplificateur de se substituer aux lampes et de contribuer au transport analogique. Mais il va très rapidement acquérir une seconde fonction, celle de pouvoir stocker une information sous forme binaire. Le monde de l'entreprise s'empare rapidement de cette nouvelle possibilité dans deux domaines, celui de la maîtrise des technologies et celui d'une nouvelle démarche dans le domaine de la gestion.

A l'issue des années 70, l'ordinateur fait partie du panorama de l'entreprise. La fonction de stockage, notamment dans le domaine de la gestion, apparaît évidente et l'informatique de gestion connaît un essor considérable. C'est pourtant la problématique du transport qui va donner une nouvelle dimension à l'informatique.

L'invention et surtout la vulgarisation du Modem changent complètement le contexte de l'informatique dans les années 90. En peu de temps, une véritable toile informatique apparaît reliant plusieurs centaines de millions d'ordinateurs à travers le monde. Rapidement, tout le contexte de la vie sociale et de l'entreprise est changé, le numérique est partout. Il continue cependant à cohabiter avec les deux autres supports que sont le papier et l'analogie.

Cet historique nous a montré que les évolutions successives se faisaient dans quatre axes qui sont : la Quantité d'informations, la vitesse de Transmission, la possibilité de Stockage, une Diffusion à un plus grand nombre.

2 - INFORMATION, COMMUNICATION ET SYSTEME D'INFORMATION ET DE COMMUNICATION (SIC)

L'Encyclopédie Universalis définit la communication comme étant :

La transmission, supposée au moins réciproque, de messages et de leur signification.

Cet énoncé a le mérite de bien différencier au sein de l'information, l'idée et sa matérialité. Elle souligne également que la communication ne peut pas se schématiser par un simple processus physique de transmission dont la seule valeur ajoutée serait de transporter une information d'un point à un autre. Bien que depuis Shannon et col. [SHA, 49] l'information se trouve intégrée à la théorie de la communication, cette dernière ne doit pas se réduire au traitement du signal et encore moins au caractère probabiliste ou non probabiliste de tel ou tel élément porteur d'information.

La figure 1a, schématisant le système d'information et de communication selon les conventions de la systémique, montre que la finalité d'un tel système est l'échange d'idées entre individus ou groupes d'individus, voire organismes. Ainsi, on imagine bien vers quoi devrait tendre le système : une amélioration dans tous les secteurs :

- la Quantité d'information
- la vitesse de Transmission
- la possibilité de Stockage
- une Diffusion à un plus grand nombre

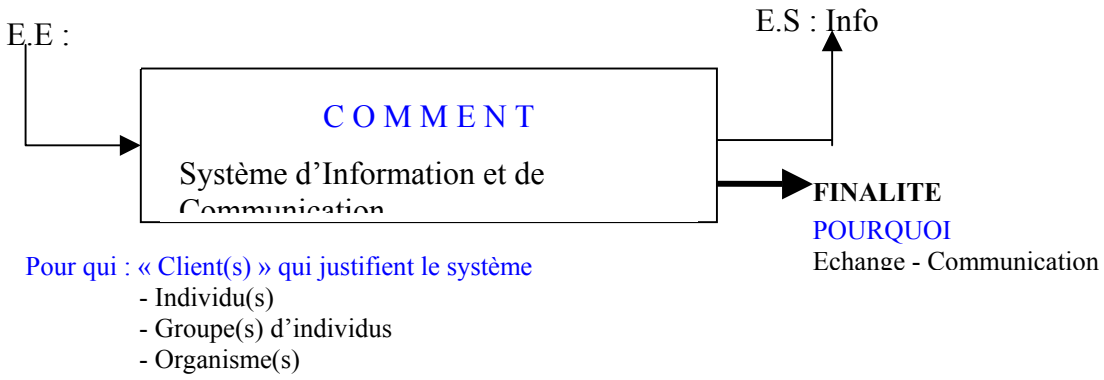


Figure 1a : représentation systémique du SIC

Cette vision systémique permet d'identifier les 3 objectifs que sont : l'atteinte des finalités, la réaction face aux contraintes et l'optimisation de la transformation des entrées en sorties. [MIN, 00]

1- Atteinte de la finalité : Le message transmis doit être :

- reçu (appropriation physique du message),
- compris (décodage du sens) et,
- déclencher une action (échange, stockage, ordre).

On voit ici apparaître des difficultés liées aux choix du mode de codage/décodage car les conventions choisies sont étroitement liées au type de support. Par ailleurs, ce schéma met en évidence ce que la cybernétique avait énoncé auparavant : la transmission d'information prend un sens organisationnel puisqu'un programme porteur d'information peut ordonner un certain nombre d'opérations. [MOR, 91]

2- Réaction face aux contraintes : Le choix du support de transmission peut impliquer ses propres contraintes, contraintes purement techniques. Les intégrer ne voulant pas dire s'y soumettre ! La vision systémique permet de comprendre comment TRIZ peut trouver des solutions d'innovation en changeant "d'angle d'attaque" : super système – système – sous-système (figure 1b).

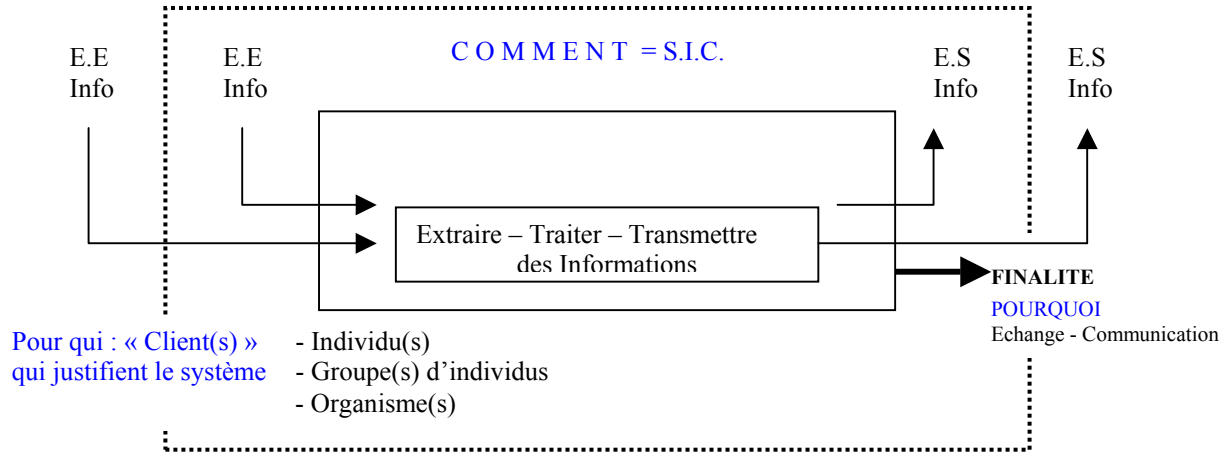


Figure 1b : représentation systémique du SIC

3- Optimisation de la transformation des entrées en sorties : figure 1b
Selon la définition proposée par l'AFITEP [AFI, 00] qui dit que :

> **L'information est un élément de connaissance susceptible d'être représenté à l'aide de conventions pour être conservé, traité ou communiqué**

nous pouvons mieux imaginer sur ce schéma les ressources nécessaires et mieux appréhender la notion d'efficience qui consiste à optimiser l'utilisation de ces ressources.

En résumé et comme le montrent les développements biologiques récents en matière de génétique, il faut considérer l'information organisationnelle comme étant à la fois : Mémoire – Message - Programme.
Cependant, « l'Information n'est ni matière, ni énergie. Elle ne peut être manipulée comme un objet mais doit l'être comme un concept…un concept plein de lacunes et de richesse » [MOR, 91].
L'information présente des enjeux mais aussi difficulté et risque (figure 2).

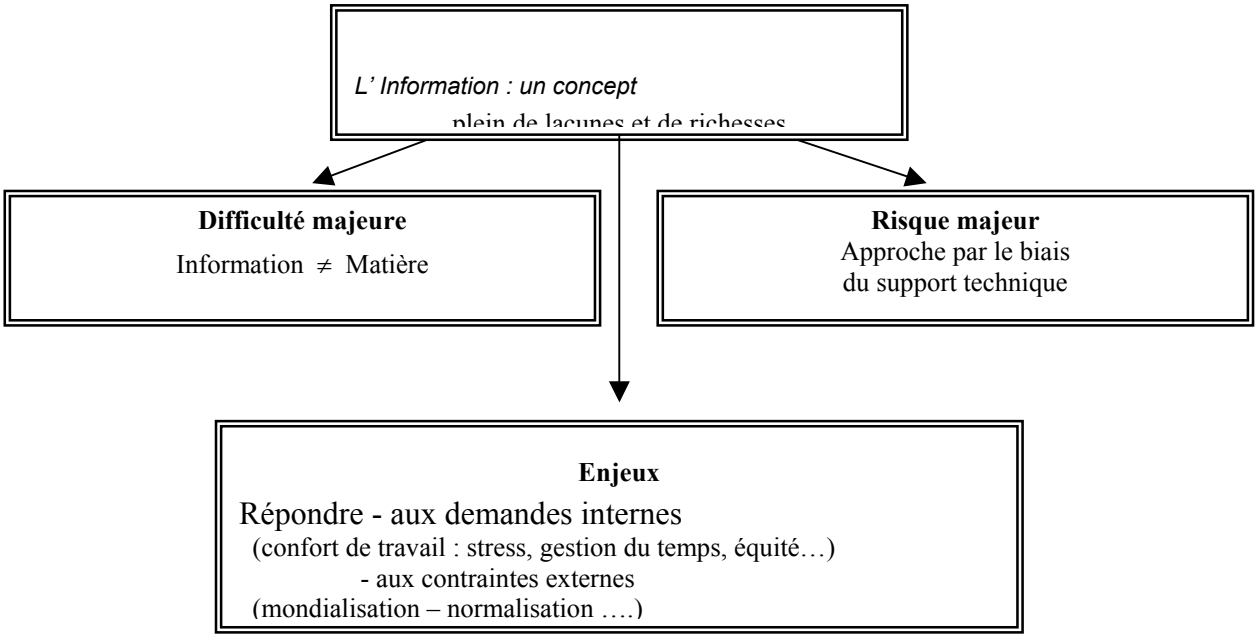


Figure 2 : Enjeux, difficulté et risque liés à l'information

En effet, selon une enquête réalisée sur quinze entreprises de la région Aquitaine [PIN, 00], les arguments en faveur de l'introduction de l'intranet sont par ordre d'importance les suivants :

Argument majoritaire : **Echange et diffusion d'informations ciblées**

Viennent ensuite :

Gain de temps par une mise à jour quasi immédiate

Facilitation dans l'exécution des missions du personnel
Passer de « papier » en documents numériques
Dispersion géographique

Satisfaction du client
Avant-garde technologique

3 - ANALYSE DE LA VALEUR ET TRIZ : DES APPROCHES METHODOLOGIQUES COMPLEMENTAIRES.

L'Analyse de la Valeur, apparue au début des années 60 pour améliorer la compétitivité des entreprises, permet de concevoir ou de re-concevoir un produit ou un procédé. Elle est basée sur l'optimisation de la satisfaction des besoins client tout en minimisant les coûts. Sa mise en œuvre se déroule en sept étapes (tableau 1). [AFN, 90]

N°	**Phases**	**Objectifs**
1	Orientation de l'A.V.	Objet de l'étude, objectifs, enjeux, moyens, échéances, contraintes…
2	Recherche de l'information	Inventaire des données nécessaires.
3	Analyse Fonctionnelle	**Phase la plus caractéristique** : Recherche, analyse et hiérarchisation des fonctions, élaboration du cahier des charges et de l'anti-cahier des charges. Orientation des efforts vers des domaines fructueux générant des gains (Qualité-Coût).
	Analyse des coûts	
	Validation des besoins et des objectifs	
4	Recherche d'idées et de voies de solution	Phase de créativité génératrice d'idées.
5	Etude et évaluation des solutions	Etudes de faisabilité technico-économiques, analyse de risques.
6	Bilan prévisionnel	Dossier avantages – inconvénients – coûts. Argumentaire de sélection et précisions sur la mise en œuvre.
	Présentation des solutions retenues	
	Décision	
7	Réalisation de la ou des solutions choisies	Planification et pilotage de la mise en œuvre, bilan définitif.
	Suivi	
	Bilan définitif	

Tableau 1 : Déroulement des 7 phases de l'Analyse de la Valeur

TRIZ[10] ou Théorie de Résolution des Problèmes Inventifs, est un corpus de connaissances traitant des mécanismes d'invention, des lois d'évolution des systèmes et de la résolution de problèmes technologiques de toute nature.
A partir de ses premiers travaux, G. Altshuller a su développer avec ses collaborateurs une véritable science expérimentale de l'innovation qui ne sera connue en Occident qu'au début des années 90. [ALT, 84] [ALT, 88].
Ses principaux champs d'application sont :

- la résolution de problèmes inventifs (liés à un produit ou un système),
- l'analyse des défaillances (prévention et élimination) et,
- la prospective technologique (propositions de pistes pour la veille).

On peut souligner l'existence sur le marché d'aujourd'hui de deux grands progiciels d'application de TRIZ, TO[11] et IWB[12].[BRA, 95]

[10] Acronyme d'une expression russe inventée en 1946 par un jeune ingénieur du service des brevets de la marine soviétique, G.Altshuller.

[11] TO : Tech Optimizer de la société « Invention Machine »

[12] IWB : Innovation Work Bench de la société « Idéation »

La méthode TRIZ est basée sur des concepts et propose des modèles que nous avons résumés dans les tableaux 2 et 3 suivants.

• **Les concepts fondamentaux et Leur signification**	
Le problème inventif (ou la notion de contradiction)	Un problème dont la solution amène un autre problème conduit finalement à une contradiction.
Le cycle de vie (ou courbe en S)	Corrélation entre les étapes de développement du système, le nombre d'inventions, leur niveau et le taux de rentabilité technologique à l'étape du cycle.
Les niveaux d'invention	Les cinq degrés d'inventivité : 1-la solution apparente ; 2-l'amélioration mineure ; 3-l'amélioration majeure ; 4-le nouveau concept ; 5-la découverte. [BER, 98]
Les mécanismes d'invention	Principes intellectuels récurrents menant à l'innovation, indépendamment des disciplines concernées et de tout outil psychologique.
Les mécanismes d'évolution	On admet que les systèmes techniques peuvent suivre 8 lois d'évolution (statiques, dynamiques et cinématiques) [CAV, 99]
L'accroissement vers l'idéalité	Un système idéal n'existe pas mais remplit une fonction utile. Toute solution innovante augmente le degré d'idéalité D défini comme suit : D = (Σfonctions utiles) / (Σfonctions néfastes + Σcoûts) Le système doit résoudre son problème avec ses propres ressources.

Tableau 2 : Les concepts fondamentaux de la méthode TRIZ

• **Les modèles**	
La matrice des contradictions ou **les principes de séparation**	Résolution de contradictions physiques : application des principes de séparation : dans le temps ; dans l'espace et sous condition. Résolution de contradictions techniques : application des 40 principes d'innovation.
Le modèle substance – champs (souvent associé au précédent)	Recherche d'interactions entre les composants du système et liste les fonctions utiles et néfastes pour tendre vers l'idéalité.
Afin de vaincre **l'inertie psychologique**	Méthode des hommes miniatures : résoudre le problème dans un état imaginaire, zone de conflit d'hommes miniatures, pour transposer la solution dans une réalité technique. Les opérateurs TTC Taille – Temps – Coût : changement d'échelle pour aborder le problème (minuscule/immense ; instantané/infini ; coût élévé). La méthode des 9 écrans : modèle systémique proposant d'aborder le problème sous l'angle des sous-systèmes ou du super-système.

Tableau 3 : Les modèles et méthodes préconisés par TRIZ

La complémentarité et non la compétitivité des deux approches peut être source de progrès. En effet Il est relativement aisé d'établir de nombreuses similitudes entre les deux et de mettre en évidence des apports réciproques. On peut citer pour exemple :
- la phase 5 de l'AV est tout à fait en accord avec les cinq degrés d'inventivité de TRIZ mais les niveaux 2 à 4 peuvent être plus rapidement atteints avec ce dernier [BER, 98].
- Le principe d'idéalité décrit par TRIZ revêt la notion de Valeur développée par l'AV qui offre une recherche plus pragmatique dans ce domaine.

L'AV présente deux points forts : celui de bien positionner les enjeux et les contraintes du problème dans sa première étape et celui d'analyser en phase 6 la mise en œuvre des solutions.
Si dans les phases 2 à 4 de l'AV, TRIZ peut apporter des réponses aux questions en matière de concepts à suivre ou de résolution des problèmes autrement que par des méthodes aléatoires où seul l'homme reste détenteur du moment et de la pertinence de l'idée émise, c'est dans la phase 5 de recherche d'idées que TRIZ montrera tout son intérêt. En effet à ce stade, TRIZ peut dépasser la simple résolution de problème, génératrice d'idées, en orientant les propositions vers un idéal.

Nous proposons, à partir de ces premières réflexions, de schématiser la complémentarité des deux approches par la figure suivante (figure 3):

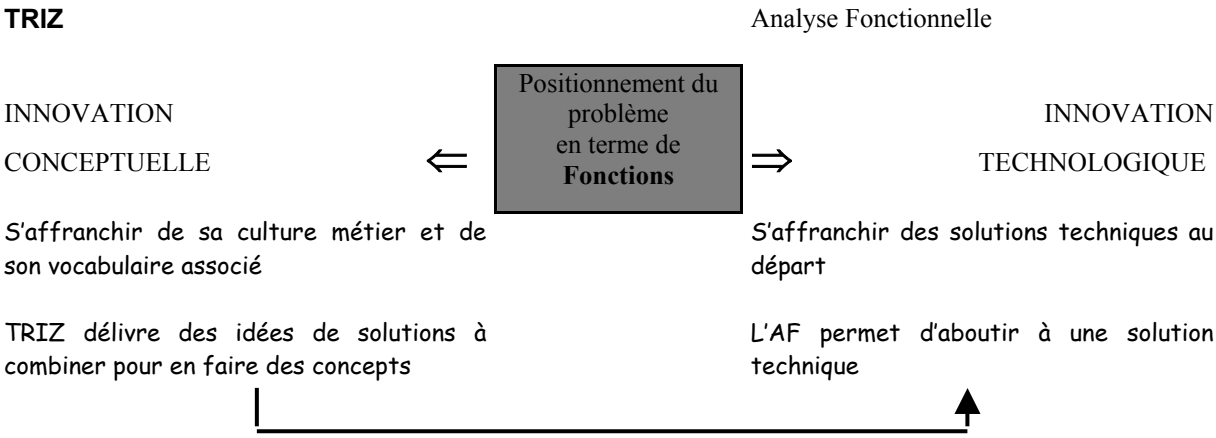


Figure 3 : Complémentarité TRIZ – Analyse fonctionnelle

4 - APPROCHES AV ET TRIZ APPLIQUEES A L'INFORMATION : DES BASES DE REFLEXION ENCOURAGEANTES

Sans entrer ici dans le détail de l'analyse fonctionnelle qui a consisté à qualifier l'ensemble des relations existantes entre l'information et les éléments en contact (source, émetteur, récepteur, support et transport), nous avons pu dégager 9 caractéristiques fonctionnelles de l'information qui sont :

COMPREHENSION
DISPONIBILITE
ACCESSIBILITE
TRANSMISSION
DIFFUSION
SECURISATION
CONTROLE
CORRECTION
CONSERVATION

A partir des 39 paramètres d'innovation énoncés par TRIZ, nous avons dégagé des transpositions possibles et applicables à l'information en essayant de trouver des correspondances entre l'analyse fonctionnelle de l'AV et TRIZ (voir tableau 4 en annexe).

Les résultats compilés dans le tableau 4 montrent que l'ensemble des fonctions trouvées se positionne dans le tableau des correspondances. Par contre, il est à noter que les paramètres TRIZ détectent deux notions supplémentaires à approfondir : la **valeur** et la **productivité**. En effet, l'information par elle-même doit avoir une qualité intrinsèque, être intéressante, amener une valeur ajoutée. Si on dispose d'un ensemble conséquent d'informations de qualité, pouvant être transmises facilement et rapidement vers un public ciblé alors on peut parler de productivité (c'est une notion plus globale).

Afin de poursuivre l'application TRIZ sur l'information, nous nous proposons d'utiliser la matrice des conflits et contradictions. Ceci afin de pouvoir dégager et exploiter un ou plusieurs principes significatifs, sur les 40 principes énoncés par la méthode.

La finalité d'utilisation de ce tableau à double entrée est de proposer des voies d'investigation possibles appelées principes d'innovation en croisant chacun des 39 paramètres de la méthode (donnés dans le tableau 4). Ces principes, au nombre de quarante au total et inscrits dans la cellule d'intersection, veillent à améliorer le paramètre pointé sur la première colonne sans détériorer celui pointé sur la première ligne.

Résultat non désiré ⇒	1 masse d'un objet mobile	…	39 productivité
à améliorer ⇓			
1 – masse d'un objet mobile	__	…	35, 3, 24, 37
…	…	…	…
39 – productivité	35, 26, 24, 37	…	__

Figure 4 : Extrait de la matrice TRIZ des contradictions

Compte tenu des multiples possibilités qui s'offrent à nous, nous avons décidé de focaliser notre attention sur le paramètre 24, c'est-à-dire d'informer sans qu'il y ait « perte d'information », ce qui paraît être la fonction principale de l'information. La matrice des contradictions va permettre de recenser et d'analyser l'ensemble des principes donnés par la méthode, principes qui améliorent tous les paramètres sans détériorer la perte d'information. Concrètement, seuls les résultats de la colonne 24 seront considérés.

Au terme de cette analyse, il apparaît clairement que le principe n°10 se détache des autres. En effet, il est cité dans 25% des cas alors que les autres sont en moyenne cités entre 0.5 et 8% des cas. Il nous faut encore compléter ce premier travail par une analyse de ce que nous indique ce principe émergeant en nous posant la question :

Que nous indique le principe n°10 ?

- **ANTICIPER.**
- Réaliser par anticipation tout ou partie d'une action requise.
- Réaliser à l'avance un changement requis plus tard.
- Prédisposer les objets de telle façon qu'ils soient prêts à entrer en action efficacement et sans perte de temps.
- Réorganiser l'objet pour optimiser une action future.

La matrice TRIZ nous dit qu'en appliquant ce principe, les fonctions suivantes se verront améliorées sans perte d'information :

- Compréhension
- Accessibilité
- Transmission
- Diffusion
- Sécurisation
- Contrôle
- Conservation

Il faut donc que l'information, dès sa réalisation, soit mise en forme à travers un canevas prédéfini, indépendant du support choisi, destiné à remplir correctement toutes ses fonctions.
Il est logique de ne pas retrouver les fonctions « disponibilité » et « correction », en effet, une telle anticipation ne les améliorera pas.

Cette toute première application a démontré qu'il était possible d'utiliser TRIZ pour traiter des problèmes liés à l'information. Cela laisse donc présager que la transposition de tous les outils TRIZ disponibles pourrait nous éclairer sur de nombreux points. Il faudra cependant développer une capacité à penser « information » à partir de la lecture des termes très technologiques de TRIZ. La porte est en tous les cas ouverte à de futures recherches.

CONCLUSION

Les entreprises se sont longtemps focalisées sur les trois notions essentielles que sont la qualité, les coûts et les délais puis, pour conserver un avantage concurrentiel, pris en compte une quatrième notion, tout aussi essentielle, l'INNOVATION. L'innovation ne s'improvisant pas, le Management de l'Innovation Technologique a fait son apparition pour se développer à grande échelle. On observe deux types d'innovation, l'innovation incrémentale qui consiste à améliorer l'existant (l'appareil photo jetable en est un bon exemple) et l'innovation radicale, caractérisée par une rupture avec les technologies antérieures comme dans le cas du disque compact [ROB, 96].

Pour les accompagner dans leur démarche, les entreprises disposent d'un ensemble de méthodes qui ont déjà fait leurs preuves (brainstorming, benchmarking, synectique, Quality Function Deployment ou QFD ou encore l'analyse de la valeur) et d'autres plus récentes quant à leur application comme la méthode TRIZ.
Cette dernière présente, comme nous l'avons vu, de grandes similitudes avec l'analyse de la valeur. En effet, TRIZ comme l'analyse de la valeur s'intéresse aux fonctions mais là où l'analyse de la valeur propose des solutions technologiques, TRIZ délivre des idées de solutions à combiner pour en faire des concepts. Dans ce sens, il semble que le développement de TRIZ pourrait bien, en terme de compétitivité et de management de l'information, bouleverser les données actuelles.

REFERENCES BIBLIOGRAPHIQUES

[AFI, 00] AFITEP, Dictionnaire de management de projet, 4ème édition, AFNOR, 2000

[AFN, 90] AFNOR, Analyse de la valeur, caractéristiques fondamentales, Normes NF x 50-152, 1990

[ALT, 84] G. ALTSHULLER, And suddenly the inventor appeard, Technical Innovation Center, Second edition, INC. Worcester MA, may 1984

[ALT, 88] G. ALTSHULLER, Creativity as an exact science, Gordon and Breach, New York, 1988

[BER, 98] G. BERTOLUCI, M. LECOQ, R. CANONNE et Y. LE MEUR, TRIZ, une aide à la décision encore mal connue des praticiens, Revue française de gestion industrielle, Vol.18 n°4, 1998

[BRA, 95] J. BRAHAM, Inventive ideas grow on TRIZ, Machine Design, october 12, 1995

[CAV, 99] D. CAVALLUCCI, TRIZ : l'approche altshullérienne de la créativité, Les Techniques de l'ingénieur, traité de génie industriel, A 5 211, 1-18, 1999

[MIN, 00] M. MINY, Projets systèmes d'information : une approche d'identification des risques spécifiques, Montpellier, France, 2000

[MOR, 91] E. MORIN, Communication et complexité : Introduction à la pensée complexe, ESF Editeur, France,1991

[PIN, 00] N.W. PINEDE et A SCHOTT, Quelle dynamique organisationnelle pour le déploiement d'un système intranet ? l'exemple d'entreprises en aquitaine, Montpellier, France, 2000

[ROB, 96] M. ROBERT, Stratégie pour innover, Ed Dunod, 1996

[SHA, 49] C. SHANNON et W. WEAVER, The mathematical theory of communications, University of Illinois Press, Chicago, USA,1949

site utilisé www.olats.org/schoffer/definfo.htm

ANNEXE

Tableau 4 : Transposition des paramètres TRIZ applicables à l'information

Paramètres TRIZ	**Transposition possible applicable à l'information**	**Correspondance avec l'analyse fonctionnelle**
1. masse d'un objet mobile	Valeur du contenu	
2. masse d'un objet immobile		
3. longueur d'un objet mobile		
4. longueur d'un objet immobile		
5. surface d'un objet mobile		
6. surface d'un objet immobile		
7. volume d'un objet mobile	Volume d'informations disponibles	DISPONIBILITE
8. volume d'un objet immobile		

9. vitesse	Rapidité de transmission	TRANSMISSION
10. force	Impact	DIFFUSION
11. tension, pression		
12. forme	Mise en forme	COMPREHENSION
13. stabilité de l'objet	Durée de vie	CONSERVATION
14. résistance		
15. longévité d'un objet mobile	Durée de vie	CONSERVATION
16. longévité d'un objet immobile		
17. température		
18. brillance		
19. énergie dépensée par l'objet mobile	Facilité ou non de transmission	TRANSMISSION
20. énergie dépensée par l'objet immobile		
21. puissance	Impact	DIFFUSION
22. perte d'énergie	Transmission restreinte	TRANSMISSION
23. perte de substance	Perte d'information	INFORMATION
24. perte d'information	Perte d'information	INFORMATION
25. perte de temps	Perte de temps	ACCESSIBILITE
26. quantité de substance	Valeur du contenu	
27. fidélité	Fidélité	SECURISATION
28. précision de la mesure		
29. précision de l'usinage		
30. facteur nuisible agissant sur l'objet	Facteur nuisible agissant sur l'information	CONTROLE
31. facteurs nuisibles annexes	Bruit de fond	ACCESSIBILITE
32. usinabilité		
33. facilité d'utilisation	Facilité d'utilisation	COMPREHENSION
34. aptitude à la réparation	Aptitude à la correction	CORRECTION
35. adaptabilité	Adaptabilité	CORRECTION
36. complexité de l'appareil		
37. complexité de contrôle	Complexité de contrôle	CONTROLE
38. degré d'automatisation		
39. productivité	Productivité	

VEILLES , INTELLIGENCE COMPETITIVE ET DEVELOPPEMENT REGIONAL DANS LE CADRE DE L'AUTONOMIE EN INDONESIE.

Sri Manullang
CRRM,
Centre Scientifique de St Jérôme, Université Aix-Marseille III
13397 Marseille cedex 20

Résumé : Introduire les méthodologies de l'Intelligence Compétitive et des Veilles dans des pays où la culture et le niveau technologique sont très différents des nôtres ne peut pas se faire sans tenir compte de l'approche culturelle et des savoirs tacites locaux. Pour ce faire deux régions de l'Indonésie ont été choisies : Célèbes du Nord et Sumatra Nord. Est appliquée au niveau de ces territoires la méthodologie de la Turbulence Positive, couplée avec l'Intelligence Compétitive et la veille pour une exploitation plus directe des informations ; Après avoir analysé les freins et les leviers culturels, deux approches sont exposés . L'une concerne une zone de petites et moyennes industries située près de la ville de Manado dans les Célèbes du Nord, l'autre une zone rurale situés dans le région de Toba Samosir (Sumatra Nord).
On met en évidence les différences entre les deux régions, et les approches proposées sont exposées avec le plan d'action qui a été mis en place durant une année et demie de travail. Ce plan d'action, basé sur l'enseignement, les cottages industries et l'information informelle débute à la fin du mois d 'Octobre 2002.

Abstract: To introduce the methodologies of Competitive Intelligence and Competitive Technical Intelligence in Countries where the knowledge and the culture are very different of ours cannot be achieved without the consideration of tacit knowledge and cultural behaviour. To do so, we have choose two indonesian's regions : North Sulawesi and North Sumatra. The methodology of Positive Turbulence has been applied, coupled with Competitive Intelligence and Competitve Technical Intelligence for a most direct usage of the information.
This work emphizises the differences between the two regions, and the various approaches are described. They have been worked out during one year and half of work. This action plan is grounded to teaching, cottage industries clusters and informal information. It will begin at the end of October 2002.

Mots-clés : développement local, culture, intelligence compétitive, indonésie, enseignement, autonomie, CI cluster, veille technologique

Key-Words : local development, culture, competitive intelligence, indonesia, teaching, autonomy, CI cluster, technology watch

Veilles , Intelligence compétitive et développement régional dans le cadre de l'autonomie en Indonésie.

INTRODUCTION

Depuis 1965 à 1997 c'est à dire 32 ans, le Gouvernement Indonésien avait développé l'« Era orde baru » c'est à dire «la nouvelle ère » mise en place par le Président Suharto. Cette gouvernance était basée sur une très forte centralisation et un système autocratique. Ces systèmes ont fonctionné pendant 32 ans. Mais la crise économique a conduit la population a entamer en 1998 une révolution pour lutter contre ce système de gouvernement.

Durant les 32 premières années d'existence des ces systèmes, il n'existait pas d'autonomie possible pour les différentes régions Indonésiennes. La seule région ayant de ce fait un niveau de développement acceptable a été l'île de Java (qui concentre plus de la moitié de la population indonésienne) où des investissements « énormes » ont été réalisés. Mais ces investissements concentrés seulement sur cette zone géographique (Java) ont conduit les autres régions et par contrecoup les populations de ces régions a rester dans un état de sous – développement. Cette état de fait pouvant conduire à un éclatement de l'Etat Indonésien et à une quasi disparition de celui-ci.

Pour éviter un éclatement du pays, l'autonomie qui a été mise en place, basée sur la loi No. 22 en 1999 concernant le gouvernement local et la loi No. 25 concernant de l'équilibre de la finance entre le gouvernement central et les gouvernements locaux, doit permettre à chaque région de bénéficier de la mise en place de programmes de développement équitables acceptés par la population et destinés au développement local réel. Ceci par opposition aux pratiques du « era orde baru » où le fond de développement régional était uniquement destiné à satisfaire des réalisations de groupes de pressions ou de groupes d'intérêt privés dont les activités n'étaient pas nécessairement en phase avec le développement souhaitable au niveau local.

L'autonomie a pour but de développer à court terme et à moyen terme des processus de gestion et de création d'initiatives locales. Par exemple la zone Asie du Sud Est va, dès 2003, entrer dans l'AFTA (2). De ce fait , pour participer efficacement à ce nouvel ordre régional, l'autonomie devient un moteur, facilitant les projets et raccourcissant les temps de prise de décision.

Mais, l'autonomie introduit au sein des gouvernements régionaux des contraintes nouvelles. En effet, la promotion de nouveaux projets, la nécessité de sélectionner des projets à investissements moyens et des projets à investissements lourds (avec l'aide de l'Etat comme dispensateur de moyens), vont nécessiter de la par des autorités locales un pouvoir d'analyse (pour déterminer les meilleurs choix), de proposition et de planification nouveaux. Ceci n'étant pas nécessaire du temps de la centralisation.

Ainsi, l'autonomie en rendant le pouvoir politique plus près des citoyens, devrait faciliter un développement du pays plus harmonieux et permettant d'utiliser au mieux les richesses locales, qu'elles soit au plan des matières premières, de l'agriculture, mais aussi du savoir tacite des populations.

Le travail que nous menons en Indonésie s'inscrit dans cette perspective. Après avoir effectué deux missions sur place, l'une pour faire connaître nos compétences (CRRM http://crrm.u-3mrs.fr), l'autre pour les utiliser dans un développement local accepté, nous sommes entrés dans une phase active que nous allons décrire maintenant.

Les Catalyseurs du développement local

La globalisation, les changements technologiques rapides, le développement d'une concurrence mondiale, ont précipité le monde dans un tout capitalisme libéral dont les limites semblent cependant atteintes. Mais, jusqu'ici on n'a pas trouvé mieux comme moteur général. Cependant, des voix s'élèvent (UNESCO, OCDE (3) ..) pour dire que le développement doit aussi être soutenable d'une part, et que d'autre part on ne pourra pas avec les méthodes occidentales arriver en un temps très court à mettre à notre niveaux l'ensemble des pays du globe.
Nous avons donc choisi de focaliser notre action sur la mise en place de systèmes de Veille Technologique et d'Intelligence Compétitive adaptés au plan local, c'est à dire tenant compte des niveaux technologiques actuels, des équipements accessibles (en accès à l'information), de la culture locale, et aussi des zones géographiques où les applications pourraient s'effectuer.
L'objectif n'est pas de transposer directement un enseignement et des savoir occidentaux en Indonésie, mais de modeler ces derniers afin qu'ils soient le mieux adaptés possibles aux condition locales et à la culture locale.

1 - ORGANISATION ADMINISTRATIVE, STRUCTURE DE LA POPULATION ET DU TERRITOIRE

Comme le reste de l'archipel, la province est découpée administrativement en 4 niveaux : kabupaten (département), municipalité (kotamadya), canton (kecamatan), et enfin commune, ou village. Les prérogatives de chacune de ces zones sont les suivantes :

Etat Central
Défense, Politique étrangère, religion, justice, fiscalité, monnaie-finances, éducation (enseignement supérieur).

Province
Administration des différentes Régions de la Province, planification régionale, Education, ..infrastructure générale (avec l'Etat central pour les grands projets)

Niveau Local (Région)
Développement local, travaux publics, santé, éducation, culture.

Il est à noter que cette partition n'est pas figée du fait de la toute récente autonomie, ce qui va nécessairement conduire à des compléments et des ajustements au niveau législatif.

Sur le plan des territoires, l'Indonésie est formée de 18.000 îles et îlots, comme on a l'habitude de le dire. La carte montre qu'en fait elle est composée de 5 parties majeures : Java, Sumatra, les Célèbes (Sulawesi), le Kalimantan (Bornéo), Irian Jaya (Papouasie). Sur le plan du développement, les villes les plus importantes sont localisées à Java (Jakarta, Bandung, Surabaya), puis à Sumatra (Medan) et dans les Célèbes du Sud (Ujung Pandang – Macassar).
On doit noter que la population de plus de 206 millions d'habitants (2001), est concentrée pour 60% dans l'île de Java (la plus peuplée du monde). Sur le plan des populations, il existe différentes ethnies ayant des caractéristiques précises, et des cultures différentes. L'unité linguistique, bien qu'il existe de nombreux dialectes locaux est réalisée via le Bahasa Indonesia, (la langue Malaise) enseignée (taux d'alphabétisation proche de 90%) dans tout le pays.

Lorsqu'on analyse le territoire au niveau local, on est frappé par le fait que si Java est relativement développée au plan industriel, du fait de sa concentration en habitants et du fait des investissements concentrés sur cette île durant la période passée, il n'en va pas de même des autres territoires. Là, la densité de population est faible, et l'industrialisation très peu importante. C'est surtout le développement de zones agricoles qui est courant, avec peu de moyens modernes et une utilisation importante de la population locale. Ceci n'est pas un mal, compte tenu du fait que cette population serait sans doute au chômage si elle allait vers les villes. Le faible nombre de villes importantes, hors Java, est donc une caractéristique qui permet de maintenir dans les zones rurales une population importante.

En ce qui concerne les ressources, l'Indonésie est un pays riche, il possède du pétrole, du gaz, de l'étain, de l'or. Son agriculture (tropicale) permet le développement de nombreux fruits, légumes, et céréales. C'est un des grands producteurs mondiaux d'huile de palme par exemple. Cependant, le riz, consommé par coutume par la population pose un problème car le pays est loin (malgré les dires officiels) d'être self suffisant dans le domaine et un déficit important est de ce fait créé par l'importation de cette céréale.

Par sa diversité géographique, l'Indonésie est un archipel volcanique, étendue sur plus de 5000 kilomètres le long de l'équateur, l'Indonésie est constituée d'une myriade d'îles - on en compte près de 17, 677 dont 3000 sont habitées. Les principales sont Sumatra, Java, Kalimantan, les Célèbes, Irian Jaya (Papou) - cela lui a valu le surnom de " continent maritime ". Sa superficie est de 1,919 millions de Km2.
Sur la carte, on peut suivre d'Est en Ouest Sumatra d'abord, riveraine du détroit de Malacca face à la fédération de Malaisie puis Java, séparée de Bornéo par la mer de Java, mer intérieure indonésienne, puis les Célèbes, et enfin la Papouasie dont la partie occidentale est indonésienne. Deux cent six millions de personnes sont ainsi dispersées inégalement dans les provinces qui composent cet immense archipel. Java abrite notamment la capitale, Jakarta, et concentre plus de la moitié de cette population ainsi que la majeure partie des richesses et des infrastructures du pays.

carte de l'Indonésie

2 - LA CULTURE INDONESIENNE

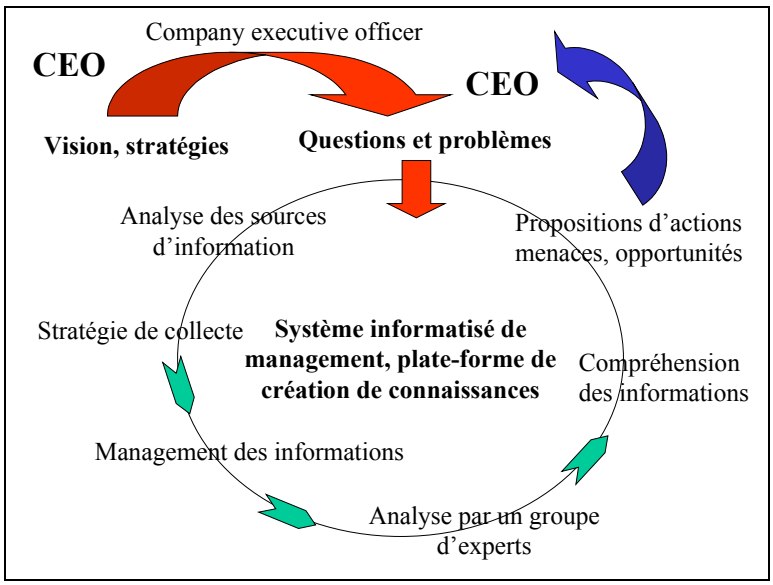
Apprécier l'impact de la culture indonésienne sur le développement d'une action de Veille ou d'Intelligence Compétitive, ne peut se concevoir que dans une analyse comparée des freins et des leviers que celle-ci va pouvoir générer. Nous avons donc choisi comme méthodologie de présenter en regard des principales étapes du cycle de l'intelligence compétitive l'impact positif ou négatif induit par la culture indonésienne.
La notion de culture devient de plus en plus importante dés lors que l'on quitte le développement technologique classique et la mentalité occidentale pour aller vers des horizons nouveaux où la perception des « choses est très différentes ».

a – Rappel du cycle de l'intelligence compétitive

Le cycle de l'intelligence compétitive est relativement simple à mettre en place, puisqu'il est quasi identique au cycle du renseignement, c'est à dire :

En fonction d'une question posée, d'un problème précis on va analyser les sources d'informations formelle ou informelle qui permettront de répondre à la question ou d'analyser les implications d'un problème. On mettra ensuite en place une stratégie de collecte, puis le management des informations et enfin l'étape de compréhension de ces informations en fonction du problème ou de la question. L'étape finale consistant à présenter aux décideurs les conclusions, généralement en terme de menaces ou d'opportunités.

Nous avons en premier lieu indiqué que nous avions à répondre à des questions, ou à mettre en évidence les implications d'un problème précis (on peut appeler cela les Facteurs Critiques de Succès). Cela ne peut se faire que si on a au préalable une vision de l'entreprise, de son futur . On aura alors à l'esprit des stratégies pour atteindre les objectifs fixés, et le système d'intelligence compétitive conduira à éclairer les choix, à donner des pistes d'action.
On présente généralement cela sous forme d'un cycle simple qui doit faire l'objet de recherche et d'analyses permanentes(4).



b - Les impacts culturels

L'attention :
En Indonésie la population a des préoccupations souvent très éloignées de l'actualité, et elle est souvent attirée et focalisée sur des occupations secondaires. En outre il y avait aussi dans le pays une tendance à focaliser l'attention (souvent par l'intermédiaire de la télévision), sur des problèmes éloignés des événements importants pour le pays, économiques, sociaux ou politiques. Dans ce cadre par exemple, l'équipement des villages en postes de télévision, pour contribuer à la maîtrise de la natalité (deux enfants par famille), a conduit à donner à ce média une importance particulière et à diminuer de ce fait les réunions de discussion, la lecture, etc.
Cette situation aura, au bout du compte pour résultat de diminuer le pouvoir d'analyse et de continuité dans la compréhension.

La prépondérance du chef :
Les personnes âgées (expérience), ou les chefs, qu'ils soient tribaux ou d'entreprises ne sont que rarement contestés au niveau de leur décision. De ce fait, l'esprit critique reste faible d'une part, et d'autre part, les dirigeants n'auront pas naturellement la propension à poser des questions, à organiser des réponses concertées et étayées par un mécanisme reconnu. Ceci se retrouve aussi dans les structures de direction qui sont très hiérarchisées.

L'esprit général :
La notion de prépondérance du Chef, le défaut d'attention conduisent souvent à fonctionner au niveau du travail sans réelle incitation, avec un enthousiasme faible. L'esprit de compétitivité est limité du fait même que l'indonésien n'a pas volonté d'être le meilleur ou parmi les meilleurs, mais de réussir avec l'autre, en groupe. Ce manque de compétitivité se reflétera aussi dans l'accès aux technologies : on préfère acheter des technologies au lieu de les développer soi même.

La prise en compte du temps :
En Indonésie le temps n'a pas la même valeur qu'en occident. L'expression « Jam Karet » le temps élastique, veut bien dire cela. Puisque le temps n'introduit pas une pression forte, cela veut dire aussi que la réactivité des institutions ou des entreprises va rester faible. La notion de « dead line » n'aura pas le même sens pour un indonésien et pour un occidental. Ainsi, les projets technologiques à cycles courts ne sont pas très bien adaptés (du moins actuellement), et la réactivité des entreprises à l'Intelligence Compétitive et aux Veilles restera faible.

Le modèle mental :
Les indonésiens ont une pensée multichrome, alors que les occidentaux avons une pensée monochrome. Cela se traduit par le fait que la pensée indonésienne n'est pas linéaire. Expliquons nous :

La **pensée linéaire** peut s'exprimer de la manière suivante :

Si a=b et si b=c alors a=c

Cette pensée est celle des occidentaux. La grande différence avec la pensée linéaire est que si a=b et si b=c cela ne veut pas dire que a=c . Cela reste une possibilité, mais il en existe aussi d'autres, liées aux mondes (air, eau , de la forêt …), aux dieux, etc…
Nous verrons, au niveau des leviers que cette pensée non linéaire est intéressante, car elle conduit à des solutions ou des points de vue innovants, mais elle est cependant freinée par la tradition.

La tradition :
Elle a un poids important, surtout dans les zones rurales et au niveau de l'artisanat. En effet, on peut avec habileté rester dans la tradition, c'est à dire en fait faire toujours la même chose et donc diminuer fortement ses capacités à innover.

Les voyages :
L'indonésien, est curieux et comme tous les peuples des îles, il aime voyager, voir des gens différents, apprécier des cultures différentes. C'est un atout important quand il est bien canalisé.

La géographie :
L'Indonésie est le pays des 18.000 îles et îlots, et ce n'est pas à Ile Rousse que nous allons expliquer ce que cela veut dire au niveau de la diversité, des ethnies, etc… Mais, si cette diversité renforce en un certain sens les traditions, elle introduit aussi la curiosité liée aux voyages.

Le lendemain :
Il existe un dicton qui dit « vivons aujourd'hui, demain on se débrouillera toujours». On mesure ici toute la différence avec notre approche souvent prévisionnelle ou prospective, entre autre au niveau de la vision de l'entreprise et de son devenir. Cette même attitude va constituer un frein au niveau par exemple du développement soutenable, ou de l'introduction de contraintes pour lutter contre la pollution ou pour préserver l'environnement dans son ensemble. Ceci met en évidence la difficulté à mettre en place des projets à cycles courts, en fait la notion de temps n'est pas fondamentale, on vie dans le moment présent.

L'habileté
L'indonésien est habile, on retrouve cela au niveau de l'artisanat, du tissage, de la peinture (batik), de la broderie, …Nous utiliserons cette habileté, en essayant de la canaliser vers un système productif différent de celui de la tradition qui contribue souvent à figer les idées.

3 - LES LEVIERS ET DES ORIENTATIONS POUR L'ACTION

Agir, pour le développement de l'Intelligence Compétitive et de la Veille, c'est avant tout essayer non pas d'apprendre aux autres directement ce que nous savons, mais de développer en s'appuyant sur les savoirs tacites des gens et sur leur culture des actions permettant de changer le **statu quo**.

Nous nous rapprochons ainsi de la notion de turbulence positive au sens de Stanley S. Gryskiewicz (5), avec comme idée d'utiliser les turbulences de l'environnement (entre autre celles générées par l'autonomie) en les rendant positives et mettant ainsi les gens "en action".

L'Indonésie est un pays qui possède d'importantes ressources au niveau du gaz naturel, du pétrole, du bois, et de certains minéraux comme l'or, l'étain … Le climat et la fertilité du sol (volcanique) permettent de nombreuses et abondantes récoltes. Le riz étant à mettre à part, car sa consommation augmente rapidement, et sa production décroît. De ce fait, malgré les dires officiels, le déficit de la balance commerciale vis à vis de cette céréale augmente fortement.

a – Comment canaliser l'habileté vers l'innovation

Avant de mettre en évidence la méthodologie que nous avons choisie, nous devons aborder rapidement la structure du territoire indonésien. Il est logique, que le développement d'un territoire soit le plus harmonieux possible, c'est à dire que les zones "défavorisées" soient aussi prises en considération. Sur le plan de l'Indonésie, il existe des zones (entre autre près des Villes) où existent de petites et moyennes industries (6), mais dans la majeure partie des cas, comme nous l'avons expliqué au début de cette présentation ce sont les zones rurales (faiblement ou très faiblement mécanisées) qui sont les plus importantes. De ce fait, nous avons choisi deux approches qui vont conduire à deux traitements différents:

Une zone de moyenne industrie (la zone de Manado, Ville des Sulawesis du Nord) et une zone rurale, celle de Toba Samosir (Tapanuli Utara) située dans l'Ile de Sumatra (Sumatra Nord).

Nous allons en page suivante indiquer la situation géographique de ces deux régions, ainsi que leurs caractéristiques physiques.

Localisation des zones de Manado et Toba Samosir

Région de Toba Samosir (Sumatra Nord)

On peut constater que si la région de Sumatra Nord est plus développée (au sens de petites localités), elle reste néanmoins agricole, principalement autour du Lac Toba.
En ce qui concerne la Région de Manado, on constate que si celle-ci à un développement de petites industries, spécialement autour du Port, pour le traitement des noix de coco et pour l'aquaculture et la pêche, elle reste principalement agricole.

Région North Sulawesi, Manado

b – Deux méthodes de travail

Zone de petites industries Manado	**Zone rurale, Toba Samosir, Tapanuli Utara**
Il faut envoyer vers les industries des personnes formées, susceptibles de détecter des problèmes et d'analyser ces derniers pour proposer des réponses possibles. Pour ce faire il faut former ces personnes en "quantité" et les introduire où cela est nécessaire. Nous avons choisi d'opérer avec deux partenaires: l'Université UNIMA et la KAPET. L'Université UNIMA, en y implantant une session du DEA Veilles, Intelligence Compétitive, et d'autre part la KAPET, institution de gestion du développement local qui placera et suivra les étudiants au cours de leur stage en entreprise. De ce fait on réuni localement et avec une formation spécialisée adéquate les étudiants, le potentiel de l'Université, les entreprises. Il faut certes ajuster ces différentes parties du puzzle, mais si on regarde l'ensemble, on voit que l'on a intégré positivement les impacts de la culture. On crée une motivation, on donne une échéance de temps, on pose des questions et des problèmes réels, on fait agir l'étudiant seul. (7)	Le travail en zone rurale est plus difficile. Il n'y a pas en effet d'interlocuteurs potentiels susceptibles de "créer le mouvement". Pour cela, nous avons choisi d'analyser soigneusement la zone en terme de production agricole et de potentialités artisanales. On retrouve ainsi la notion de Cottage Industries (productions réalisées à la maison). Ce que nous souhaitons faire c'est mettre en place des clusters de CI (8)comme cela avait été commencé en 1988, puis abandonné. Mais en utilisant de meilleures potentialités technologiques, en trouvant des niches, c'est à dire en connaissant mieux que les personnes locales les potentialités de certains marchés, les valeurs nutritives de certains fruits et légumes, en en combinant contenu et contenant (par exemple boîte artisanale pour vendre le café, séchage de tomates, traitement du styrax benzoin (9), ….). L'objectif est d'utiliser les connaissances tacites locales, pour passer du stade de l'individuel au stade du collectif, donc d'augmenter le niveau de l'offre pour permettre l'ouverture de certains marchés et pour montrer que l'on peut gagner plus de cette façon. Cela introduit les notions de gestion, de qualité, d'innovation dans les produits, et de recherche de marchés.

Cette présentation, dans le cadre de ce travail reste sommaire, mais cela a été concrétisé à la suite d'une première mission sur place en fin 2000, suivi d'une période de socialisation (10) des projets pendant un an environ, puis d'une autre mission sur le terrain en Mai 2002, suivi de l'implantation du DEA et d'un colloque qui se réalisera à Toba Samosir en Mai-Juin 2003.

4 - CE QUE NOUS TRANSPOSONS DES VEILLES, DE L'INTELLIGENCE COMPETITIVE ET DU CYCLE DU RENSEIGNEMENT

a – Vers les petites industries

Nous allons garder pour la Région concernée (Sulawesis du Nord et Manado), trois aspects principaux:

- Mettre en évidence les turbulences créées par la nécessité de réaliser sur cette zone un développement avec les ressources locales, mais avec les "clients" externes à la Région et sans doute pour une grande part à l'Indonésie elle même. Cela concerne entre autre le tourisme, la pêche, l'organisation urbaine (urbanisme, traitement des déchets, aspects généraux entre autre du Port, etc…). Cette mise en évidence des turbulences (qu'il nous faudra ensuite rendre positives), nécessite d'avoir localement une crédibilité. C'est pour cela que nous avons préparé la délocalisation du DEA Veilles, Intelligence Compétitive sur ce site (Université UNIMA) et ceci avec une demande venant directement du Gouverneur des Sulawesis du Nord et l'accord du Ministère de l'Education Central. Ceci permet d'avoir le consensus politique et une forme de reconnaissance.
- Impliquer les acteurs principaux au niveau de la politique du développement. Ceci va nécessiter de travailler sur la Province d'une part et sur la Région (Municipalité de l'autre). Cela va se faire par l'intermédiaire d'un groupe d'orientation du DEA et ensuite des étudiants qui entreprendront des thèses sur ces sujets. Seront impliqués à ce niveau les décideurs locaux, politiques, industriels et sans doute sociaux.
- Faire descendre les concepts auprès des entreprises locales. Pour ce faire nous allons utiliser les étudiants du DEA qui devront aller en stage dans le entreprises et qui de ce fait vont aller au contact des réalités. De par l'implication politique précédente du projet, ils auront le contact avec le chef d'entreprise. Ces étudiants seront "managés" sur le plan des choix des stages et des contacts par la KAPET, organisme de développement local, qui ensuite centralisera les résultats. Les étudiants auront pour charge de détecter des problèmes et d'analyser qu'elles pourraient (en fonction des possibilités locales et de celles que nous apporteront sur place) être les sources d'information les plus pertinentes.
- Analyser les sources d'information disponibles en Indonésie, amener nous même entre autre via l'internet l'accès à des informations du type techniques: réglementations, brevets, méthodologie de prise de contacts via ce média, etc… Ceci sera traité dans le cours du DEA, à la fois par des indonésiens, mais aussi par des enseignants français spécialistes de la question. Il est évident que dans cette approche, il n'est pas utile de mettre en place des processus de traitements sophistiqués de l'information. L'accent sera mis principalement sur la gestion des informations et sur les ou le groupes d'experts qui analyseront les informations en termes d'opportunités ou de menaces.

b - Développement des zones rurales

Pour les zones rurales, ici la zone de Toba Samosir, le développement est tout autre. On se trouve en effet en face de personnes qui ne peuvent pas intégrer les démarches précédentes. Il faut alors intégrer des méthodes nouvelles tenant compte de la nature des acteurs locaux. Nous distinguerons:

- Les travailleurs locaux, généralement des paysans ou de petits artisans,
- Les enseignements qui se font par l'intermédiaire de lycées techniques agricoles. La zone ne comprend pas d'Universités, la plus proche étant à Medan, capitale de Sumatra.
- Les acteurs politiques. Nous sommes là au niveau de la Région (les Municipalités) mais pas au niveau de la Province.
- Une fondation particulièrement active pour former des Lycéens, la fondation Soposorung.
- Des conseillers, envoyés sur le terrain par des pays de l'Ouest. Nous avons rencontré à ce propos un conseiller pour l'agriculture dépendant de l'Ambassade d'Allemagne.

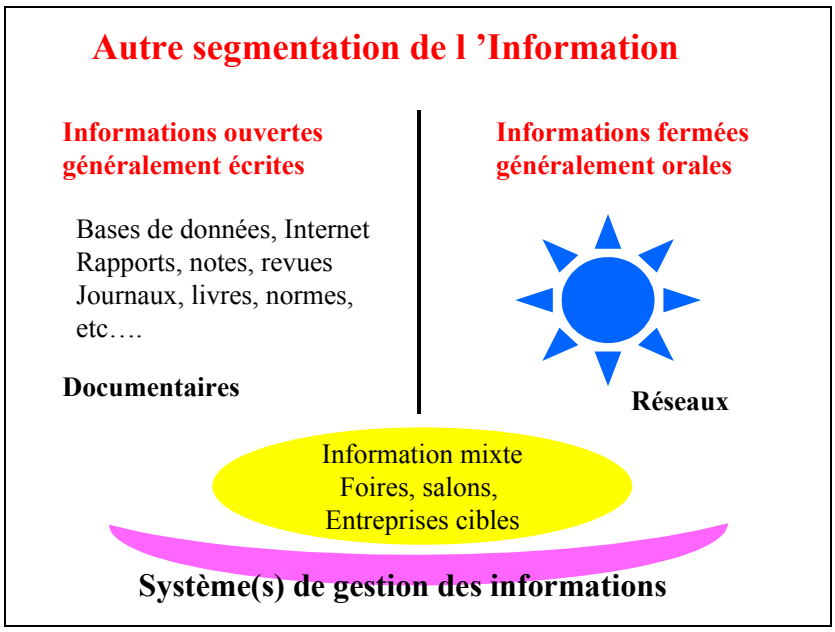
Sur le plan des informations : il n'existe pratiquement pas ou très peu d'informations concernant les aspects scientifiques et techniques. Des informations de base concernant l'agriculture peuvent être obtenues via les Lycées Techniques ou des groupements (un peu analogues à des syndicats professionnels). Sur le plan du tourisme, il existait un début d'infrastructure, mais sans proximité d'aéroport, il est difficile de mener des touristes sur place. Cela peut faire sur la base de tours (Sumatra Nord par exemple, incluant le lac Toba, etc…) ou sur la base de l'ethnie dominante de la province : les Bataks. Ce groupe de population, très soudé, avec des migrations constantes vers l'extérieur de la Province comprend des personnalités importantes. Ils reviennent souvent sur place, berceau de la famille. On rencontre localement (Toba Samosir) les mêmes pratiques que dans d'autres îles, entre autre la Corse : l'enterrement des morts dans des tombeaux locaux situés dans les propriétés, etc.

Les moyens d'action :

- Les observations antérieures : Sur le plan de la recherche, des essais avaient été développés localement, avec l'aide du Gouvernement Central (BIPIK). Selon les études qui ont suivi ce développement, on a pu remarquer que « if CI workers in Idonesia were seen as remarkably passive in skills, product range, marketing and raising of capital (ce que nous avons signalé plus haut dans l'aspect manque d'initiatives) : they sticks to known products and wait for buyers to come to their doors. This has, however, not prevented an ancouraging success in exporting furniture, clothing, baskets, carvings, etc. (11). This observation ran quite contrary to the generally belief that CI goods could be sold only in small quantities to markets. » (si les travailleurs dans les cottages industries en Indonésie semblent remarquablement passifs dans leur habileté, gamme de produits, marketing et recherche de capitaux : ils restent liés à des produits connus et ils attendent que les acheteurs viennent à leurs portes. Cela cependant n'empêche pas des succès au niveau de l'exportation de mobilier, vêtements, paniers, sculptures, etc.. Cette observation va tout à fait à l'encontre de la croyance générale selon laquelle les produits des cottages industries peuvent être seulement vendus en petites quantités sur les marchés locaux.

- Propositions d'actions locales : pour arriver à « lancer un mouvement innovant » nous allons utiliser une méthode de prospective à court terme . Différentes méthodes de ce type ont été décrites par l'Université des Nations Unies (12). Nous avons choisi de travailler avec « the participatory methods » (les méthodes participatives) (13), en impliquant la majorité des acteurs locaux. Ceci a pour but de « socialiser » notre entreprise et de créer les relais d'opinion vers les travailleurs locaux, paysans et artisans. Cela sera lancé durant le mois d'Avril 2003, et se terminera par un colloque. Nous espérons, pour avoir les relais sur place, faire travailler dans ce cadre les lycéens de dernière année de la fondation Soposorung. (14)

- Les informations : on peut essayer de relayer des informations touristiques ou agricoles, mais compte tenu du milieu et de la culture, nous avons choisi de développer plus spécialement le secteur des informations informelles. Rappelons que dans la typologie des informations il existe deux types d'informations, les informations formelles et les informations informelles. On peut représenter la typologie de la manière suivante :



On conçoit bien, que dans le contexte, ce sont les informations à caractère informel qui seront les mieux perçues. Pour atteindre un tel objectif, nous allons nous baser sur l'ethnie dominante de la région, c'est à dire les Bataks. Largement dispersés en Indonésie, voire dans le monde, ils restent cependant une communauté unie. Mais, s'ils retournent au pays, dans des circonstances familiales (vacances, mariages, naissances, deuils …), ils constatent souvent que rien ne « bouge », mais malgré leur position sociale ils ne font pas vraiment œuvre locale de développement. Des efforts en ce sens ont déjà été accompli par la fondation Soposurung. A signaler aussi le début d'initiatives locales allant dans ce sens. Les placer dans un système d'information aura pour nous un double but : leur apporter une réflexion générale sur le rôle qu'ils pourraient jouer dans le développement local, en restant aux aguets et en ayant toujours à l'esprit que ce n'est pas nécessairement en apportant localement de l'argent qu'ils feront changer les mentalités et donneront une impulsion au développement (15).

CONCLUSION

Cette présentation montre que même lorsqu'on domine les méthodes et les techniques de Veille et d'Intelligence Compétitive, il n'est pas possible de réaliser la transposition de celles-ci directement dans des pays très différents, tant au niveau culturel que des approches scientifiques. Ceci pose en quelque sorte le problème de l'enseignement de certaines disciplines, où il est illusoire de penser que la simple venue en présentiel des étudiants dans le pays dispensateur des savoirs est suffisante. Le CRRM, après de nombreuses années de travail avec l'Indonésie a du, pour arriver à un développement plus rapide et harmonieux, aller vers une intégration locale plus complète qui ne peut se réaliser que dans le pays. (16)L'aspect culturel a en fait une très grande importance. Il doit être pris en compte, mais pas au hasard. C'est en mettant en relations les étapes des veilles, de l'intelligence compétitive et en analysant les impacts de la culture sur les différents points clés de ces méthodologies que l'on parviendra à faciliter la mise en place de ces systèmes au niveau local.
Cela bien sûr va requérir de la persévérance, mais aussi un travail qui ne doit pas être centré uniquement sur les sciences dures (en fait sur les aspects théoriques, informatiques et de réseau) de la Société de l'Information. Les aspects culturels et humains conduisant aux contenus acceptables devront être pris en compte. On quitte ainsi la simple prévision technologique pour aller vers les impacts sociaux c'est à dire vers la prospective technologique (17).

REMERCIEMENTS

Je remercie le Professeur Henri Dou, avec qui au cours de la mission que nous avons effectué conjointement en Indonésie en 2002 (Henri Dou, Eric Thouvenin, Sri Manullang, Philotheus Tuerah et Yanto Santosa - Conseiller Scientifique et Culturel), a permis par analyses et discussions successives la réalisation de ce travail.

BIBLIOGRAPHIE

2 - ASEAN FREE TRADE AREA http://www.aseansec.org/economic/afta/afta.htm

3 - Le développement durable, Projet d'action Marie Paule Verlaeten Unit for forward and strategic studies Administration of economic relations Ministery of economy Industry street 6 B 1040 Brussels

4 - Passer de la représentation du présent à la vision prospective du futur – « Technology Foresight » **Henri Dou , Jin Zhouying** Humanisme et Entreprise, Décembre 2002

5 - Positive Turbulence Stanley S Gryskiewicz Center for creative Leadership Jossey-Bass Publishers, San Francisco 1999

6 - Nous ne parlons pas ici de grandes industries, qui ont les moyens de développer elles mêmes les processus de veille et d'intelligence. Il semble plus important que globalement ce soit une fraction plus large du territoire qui soit concernée. En effet, nous avions réalisé un travail sur le développement de l'Algérie et entre autre sur le développement des villages pour constater que le développement volontaire passant par la grande industrialisation, puis revenant au niveau agricole et ensuite de la petite industrie avait été un échec. (thèse de Chabani, CRRM, Université Aix-Marseille III, 1987)

7 - Mismanagement of Tacit Knowledge the importance of tacit knowledge, the danger of information technology and what to do about it. Jon-Arild Johannessen!, Johan Olaisen, Bjorn Olsen International Journal of Information Management 21 (2001) 3}20

8 - Microenterprise Clusters in Rural indonesia: Industrial Seedbed and Policy Target. Hermine Weijland World Development, vol 27, n°9, pp.1515-1530, 1999

9 - http://www.styrax-benzoin.com/pricelist.htm

10 - On entend par socialisation le fait que le projet, après qu'il ait été exposé, doit être compris, examiné, débattu dans le temps. Cette période "de gestation sociale" est très importante, car lorsqu'on retourne sur place on fait alors preuve de civilité, de suivi et de prise en considération réciproque de la valeur de l'autre. Cette attitude, qui entre autre a été

exposé dans un travail réalisé au Collège d'Europe, le Collège de Bruges doit être ici prise en compte au "pied de la lettre". Il n'y a pas d'égalité au sens propre du terme, mais une réciprocité des savoirs, des coutumes, des comportement des uns et des autres. Tout le monde en fait à des droits, mais aussi des devoirs.

11 - UNDP, DGIS, ILO and UNIDO 1988, Dévelopment of Rural Small Industrial Enterprises. UNIDO, Vienna, page 146

12 - The United nations University, The millenium project, version 1.0 ISBN 0-9657362-2-9 CD-ROM Future Projects, Future Research Methodology American Council for the United Nations University 4421 Garrison Street, NW Washington, D.C. 20016-4055 USA http://millennium-project.org Voice & Fax 202-686-5179

13 - Voir l'article in extenso dans le CD-ROM cité ci-dessus. Jerôme C Glenn, How to do participatory methodology, p. 11.

14 - La fondation Soposorung est citée en exemple en Indonésie. Elle permet d'aider des enfants après l'Ecole primaire, en prenant les plus méritants et en leur donnant pendant tout le collège et le Lycée une instruction gratuite, tant au niveau des cours que de leur hébergement et nourriture. Sri Manullang, auteur de ce travail a été pendant deux ans (1987-1989) employée à la fondation.

15 - Creativity in contexte T.M. Amabile Boulder Colorado. Westview Press, 1996

16 - Réflexions de Professeur Henri Dou: il est remarquable que l'Indonésie, et entre autre le Ministère de l'Education du Gouvernement Central, et le Gouverneur des Sulawesis du Nord aient intégré le fait que pour gagner du temps et avoir des transferts de connaissances adaptés aux besoins du pays, il fallait favoriser l'intégration sur place des enseignements aux besoins locaux et à la culture locale. Certains pays d'Amérique du Sud, par protectionisme n'acceptent pas ce point de vue. Ils ont tort, car en fait ils auront un savoir superficiel des "choses" apprises, mais ils ne pourront pas developper rapidement "un savoir pour l'action".

17 - Technology Foresight for Latin America (Prospective Technologique pour l'Amérique Latine (et les Caraïbes) » où l'on trouve le maximum d'informations : http://www.foresight.ics.trieste.it

ORDRE, AGREGATION ET REPETITION : DES PARAMETRES FONDAMENTAUX DANS LES COMPARAISONS D'OBJETS INFORMATIONNELS

Michel Christine
CEM-GRESIC
MSHA - Université Bx3
10, Esplanade des Antilles
33607 PESSAC Cedex
Tel : 05 56 84 68 13
Christine.Michel@montaigne.u-bordeaux.fr

Resumé : On désigne généralement les objets informationnels comme les supports, les médiateurs permettant d'améliorer la perception dans un contexte de travail avec des ordinateurs (Auziol, 2001). Concrètement ces objets peuvent être de simples textes, des images fixes ou animées produites manuellement par un ou plusieurs auteurs; ou bien des constructions dynamiques, réponses de systèmes, générées par le besoin informationnel ou la définition du profil particulier d'une personne. Ces objets informationnels sont généralement des agrégats d'éléments indexés et stockés dans des bases de données. Les stratégies de construction varient en fonction des objectifs spécifiques du système, nous ne les détaillerons pas car notre propos n'est pas ici de faire une typologie exhaustive de ces objets. Nous nous contenterons de les regarder au travers de trois caractéristiques : *l'agrégation, l'ordonnancement et la répétition (ou l'unicité) des éléments.* Ces trois caractéristiques sont très souvent prises en compte par les ergonomes et développeurs de systèmes pour en améliorer l'utilisation, la pertinence et l'efficacité. Paradoxalement, parce qu'aucun formalisme n'a été défini, ils sont rarement pris en compte pour évaluer ou comparer les objets informationnels une fois construit. Nous proposons dans un premier temps de présenter les concepts de semi ordre (de classe et d'éléments) et de disjonction, nécessaires pour définir les différents types d'objets informationnels. Dans un second temps nous présentons une démarche utilisée pour résoudre complètement le problème de la comparaison d'objets de type semi-ordonné disjoint de classe. Pour conclure nous ouvrons sur les perspectives de travaux à mener dans les autres contextes et pour les autres objets informationnels.

Abstract : Informational objects are support or mediator used to improve perception of information when people are working with computer. They must be simple texts, combined or not with fixed or animated picture. They must be produced manually by one or many authors, or dynamically by information retrieval systems like answers are. These objects must be viewed as aggregates of indexed elements, generally called fragments, stocked in databases. Strategies of constructions are varying from systems to systems so objects are very different from each others but we can describe formerly them by three characteristics; there degree of aggregation, scheduling and repetition of fragments. Developers and ergonomicists very often use these three characteristics to improve use, relevance and efficacy of systems. Paradoxically, because any formalism as been defined, they are not taken into account to

evaluate or compare results constructed. In this paper we propose to present a mathematical formalism based on the concept of semi-order and possible to use in the case of IR results. Then we present how it can be write with the fuzzy set theory and used to construct similarity measure taken into account the presence and decrease rank of documents. To finish we present other contexts of use and research perspectives.

Mots clés : mesure de similarité, ordre, ordonnancement, agrégation, recherche d'information, évaluation,

Key Words : similarity measures, order, scheduling, aggregation, information retrieval, evaluation

Ordre, agrégation et répétition : des paramètres fondamentaux dans les comparaisons d'objets informationnels

1 - SEMI-ORDRE ET DISJONCTION

Les réponses proposées par les systèmes de recherche d'information peuvent se présenter a l'extrême sous la forme d'éléments non ordonnés (figure 1) ou d'éléments totalement ordonnés (figure 2).

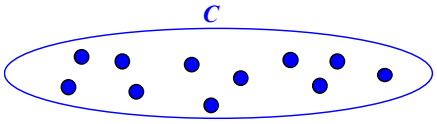


Figure 1

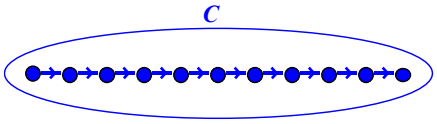


Figure 2

La recherche de nouveaux modes de présentation et de nouvelles interfaces a fait apparaître d'autres formes de présentation comme les sous-listes thématique ou les amas cartographiques, caractérisés globalement par des regroupements de réponses dans des classes construites selon une unité sémantique. L'ordre, s'il y a, n'est généralement pas total, nous appellerons cette forme hybride **le semi-ordre**. Deux cas de figure peuvent se présenter :

- Les classes sont ordonnées les unes par rapport aux autres mais les éléments à l'intérieur des classes ne le sont pas (figure 3), on parlera de *semi ordre de classe*.

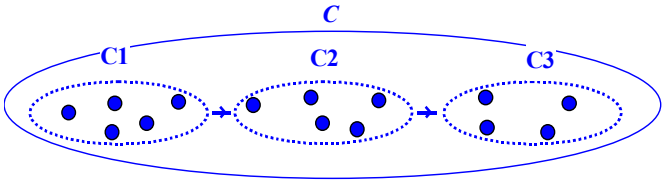


Figure 3

C'est typiquement le cas de figure qui se présente dans un système comme Spirit (Fluhr, 97). Les documents sont regroupés dans des classes selon qu'ils contiennent ou non une combinaison des mots informationnels de la question, les documents ont la même importance dans la classe car ils contiennent tous la même combinaison de mots, les classes par contre ont plus ou moins d'importance selon qu'elles sont caractérisées par plus ou moins de mots informationnels.

- Les classes ne sont pas ordonnées les unes par rapport aux autres mais les éléments à l'intérieur des classes le sont (figure 4), on parlera de *semi ordre d'éléments*.

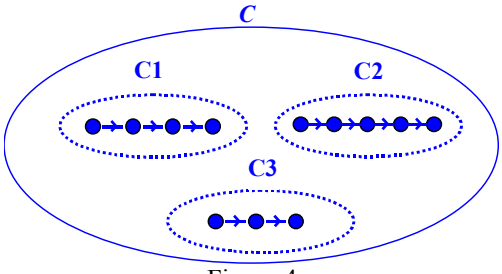


Figure 4

C'est typiquement le cas de figure qui se présente dans un système cartographique de type réseau sémantique, les classes sont formées, agrégeant les documents selon leur degré de proximité, les documents sont ensuite ordonnés dans les classes, le plus représentatif de la classe étant présenté en premier, ou bien les système qui regroupe les documents selon des méta critères comme un type de support éditorial (résumé, article de vulgarisation) ou une source (Site Web commercial, pages personnels, magazine, …) différent du contenu sémantique (Zamir, 99).

Dans tous les cas de figure les éléments ne sont présentés qu'une fois, dans le cas des figures 3 et 4 on dira que le semi ordre est **disjoint**. Cette caractéristique est propre aux réponses de systèmes en recherche d'information ; dans bon nombre d'autres contextes, les objets informationnels sont construits à partir d'éléments qui peuvent se répéter : un texte est composé de mots qui peuvent se répéter, une page Web dynamique est composé d'éléments, image ou lien qui, pour des raisons d'ergonomie, peuvent se répéter …Nous ne traiterons pas ces cas de figure et

nous restreignons au contexte où les éléments ne sont présentés qu'une fois. Il est souvent nécessaire de pouvoir comparer de tels objets, par exemple il est nécessaire de pouvoir comparer les réponses de plusieurs systèmes lorsque l'on cherche à les évaluer. Comment le faire au mieux?

2 - CAS D'ENSEMBLES NON ORDONNES

2-1- Les mesures de similarité fortes et faibles

La similitude entre les réponses proposées par les système se base généralement sur le nombre d'éléments qu'elles peuvent avoir en commun, calculé, si A et B sont les ensembles à comparer, par le cardinal de $A \cap B$ (noté $|A \cap B|$). La similitude est quantifiée par un nombre compris entre 0 et 1, 0 signifiant que les ensembles A et B comparés n'ont aucun éléments en commun c'est à dire que $A \cap B = \varnothing$ et 1 signifiant classiquement qu'ils sont strictement identiques c'est à dire $A = B$. En fait cette condition se révèle fausse lorsque des rôles spécifiques sont attribués à A et B comme c'est le cas pour le rappel et la précision. En effet, le rappel (R) est la proportion de documents pertinents trouvés par rapport au nombre de documents pertinents. Considérons que A est l'ensemble des documents retrouvés et B l'ensemble des documents pertinents alors le Rappel R s'écrira :

$$R(A,B) = \frac{|A \cap B|}{|B|}$$ (Grossman, 1998) **(Équation 1)**

Un rappel R=1 signifie que tous les documents pertinents sont retrouvés c'est à dire que $B \subset A$ et non pas que *seuls* les documents pertinents sont retrouvés c'est à dire que $B = A$.

De la même manière, une précision P=1 signifie que tous les documents retrouvés sont pertinents c'est à dire $A \subset B$ et non pas que seuls les documents retrouvés sont pertinents.

Cette observation a permis de séparer les mesures de similarité en deux groupes : **les mesures fortes et les mesures faibles** formellement définies dans (Egghe, 2002). Très simplement, **les mesures fortes sont caractérisées par une stricte identité des ensembles comparés si la proximité est de 1, les mesures faibles le sont par une inclusion des ensembles dans pareil cas**.

2-2- Principales mesures de similarité fortes

Le coefficient de Jaccard : $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ **(Équation 2)**

Le coefficient de Dice : $Dice(A,B) = \frac{2 \times |A \cap B|}{|A| + |B|}$ **(Équation 3)**

Le cosinus : $Cos(A,B) = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$ **(Équation 4)**

La mesure N $N(A,B) = \sqrt{2}\frac{|A \cap B|}{\sqrt{|A|^2 + |B|^2}}$ **(Équation 5)**

Le coefficient de débordement 2 (overlap 2) : $O_2(A,B) = \frac{|A \cap B|}{\max(|A|,|B|)}$. **(Équation 6)**

Les mesures d'efficacité : construites comme combinaison ou moyenne du rappel R et de la précision P et dont la plus générale est :

Le coefficient de Dice généralisé[13] construit par Van Rijsbergen[14] : $E_\alpha = 1 - \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}}$ **(Équation 7)**

---

[13] α étant le poids relatif assigné à la valeur de précision par l'utilisateur. α est une valeur comprise entre 0 et 1. Par hypothèse, le poids accordé au rappel est le complémentaire de celui accordé à la précision.

[14] VAN RIJSBERGEN C. J. - Retrieval effectiveness - Sparck Lones K ed Information retrieval experiments ; London : Butterworth - 1981 - pp32-43.

2-3- Principales mesures de similarité faible

Le coefficient de débordement 1 (overlap1) : $O_1(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$. **(Équation 8)**

Le rappel : $R(A,B) = \frac{|A \cap B|}{|B|}$ **(Équation 9)**

La précision: $P(A,B) = \frac{|A \cap B|}{|A|}$ **(Équation 10)**

Leurs mesures dérivées :

Le bruit (Noise) : $Noise = 1 - P$ **(Équation 11)**

Le facteur d'omission (O) ou le silence : $O = 1 - R$ **(Équation 12)**

Dans ces quatre derniers cas, A est l'ensemble des documents retrouvés et B l'ensemble des documents pertinents.

3 - CAS D'ENSEMBLES ORDONNES

L'arrivée de systèmes ordonnant totalement ou partiellement les documents selon une valeur de pertinence a rendu nécessaire la construction de mesures de similarités reflétant ce paramètre. On trouve dans la littérature deux manière de le faire :

- en calculant le coefficient de similarité sur les 10, 20 ou 30 premiers éléments considérés comme des ensembles sans ordre. C'est la solution choisie dans le cadre de TREC ou l'on retrouve des adaptations du Rappel et de la Précision à des listes ordonnées. Par exemple P(10) est la précision donnée par les 10 premiers documents, R-Prec est la précision donnée par les R premiers documents, R étant le nombre de document pertinents pour le sujet donné. On retrouvera plus de 80 mesures du même type et leur analyse comparative dans (Voorhees 98).

- en pondérant la mesure de similarité de deux ensembles par le coefficient de corrélation de rang R représentatif du nombre de permutations à effectuer pour replacer deux les éléments communs à deux ensembles dans le même ordre. Une expérimentation de cette combinaison est présentée dans (Tague-Sutcliffe, J, 1995).

Ces deux types de construction ne sont pas assez précises ; la première car elle ne prend en compte qu'une fraction de la réponse, la seconde car elle ne prend pas en compte l'intérêt de retrouver des éléments communs dans les premiers rangs plutôt que dans les derniers. De plus, ces deux mesures ne sont pas applicables dans le cas d'un semi-ordre. Notre propos a donc été de chercher comment construire des mesures de similarité prenant en compte le semi-ordre, avantageant les éléments présentés tôt et ayant un lien avec les mesures originelles de similarité. Notre travail initial (Michel, 1999), (Michel, 2000) (Michel, 2001) a été amélioré par une collaboration avec Leo Egghe, les résultats publié dans (Egghe, 2002) et (Egghe, 2003) permettent de construire des mesures de proximités ordonnées valides sur des ensembles en semi-ordre disjoint de classe. Bien entendu, **l'ordonnancement total et le non-ordonnancement n'étant que des cas particuliers du semi-ordre de classe, ces mesures sont applicables aussi dans ces contextes**. Nous nous proposons ici de les présenter brièvement.

3-1- Formalisation du problème pour le semi-ordre de classe

Considérons que deux ensembles C et C' à comparer sont construits selon un *semi-ordre de classes disjointes* comme le représente la figure 3. Leurs classes respectives seront notées C_i et C'_j, i variant de 1 à *m*, j variant de 1 à *m'*. L'idée initiale (Michel, 1999), (Michel, 2000) a été de construire des mesures de similarité en combinant une mesure de proximité classique calculée sur chaque couple de classes Ci et C'j puis pondérée par un coefficient représentatif des rangs i et j de chacune. Ainsi, une mesure de proximité ordonnée Q peut être construite à partir de *toute mesure de proximité classique de type faible et forte* (noté D) de la manière suivante :

$$Q(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} D(C_i, C'_j) \times \varphi(i,j)$$ **(Équation 13)**

$\varphi(i,i)$ étant une fonction vérifiant 5 conditions définies (pour les détails on se reportera à (Michel, 2001).

Nous avons proposé 3 mesures concrètes : une basée sur le Jaccard, une sur le rappel et une sur la précision. Dans une expérimentation nous avons comparé une mesure de Jaccard classique avec la mesure de Jaccard ordonnée. Les résultats ont montré la plus grande précision et donc l'intérêt d'utiliser une mesure de proximité ordonnée.

3-2- Les mesures de proximité ordonnées à pondération : une deuxième méthode de construction basée sur la pondération

Leo Egghe intéressé par cette problématique y a contribué en proposant une autre méthode de construction qui consiste à appliquer une fonction f sur la mesure puis à la pondérer par une fonction φ de la manière suivante[15] :

$$Q(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} f\left(D_{strong}(C_i, C'_j)\right) \times \varphi(i,j)$$ **(Équation 14)**

L'avantage de cette méthode de construction par rapport à la précédente porte principalement sur l'étendue du nombre de mesures qu'il est possible de construire. En effet, l'une des problématiques à résoudre dans l'équation 13 consiste à trouver une fonction de pondération φ respectant la normalisation de la mesure Q entre 0 et 1. Dans le cas de l'équation 14, les contraintes posées sur la fonction de pondération sont moins fortes du fait de l'application de la fonction f. Les démonstrations et définitions sont publiées dans (Egghe, 2002), il faut préciser que cette méthode n'est applicable *qu'aux mesures de type fort*.

Nous présentons les 6 mesures (Jaccard (J^{Ω}), Dice (E^{Ω}), Cosinus (Cos^{Ω}), Dice Généralisé ($E_{\alpha}{}^{\Omega}$), la mesure N (N^{Ω}), l'overlap de type 2 ($O^{2\Omega}$), construites à partir des indicateurs de type fort présentés précédemment (équation 2 à 6). On remarquera que les mesures de proximité ordonnées basées sur le Jaccard et le Dice sont identiques.

Jaccard et Dice ordonné à pondération :

$$J^{\Omega}(C,C') = E^{\Omega} = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j) \times \varphi(i,j)$$ **(Équation 15)**

Cosinus ordonné à pondération :

$$Cos^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|^2}{|C_i| \cdot |C'_j|} \times \varphi(i,j)$$ **(Équation 16)**

La mesure N ordonnée à pondération :

$$N^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|}{\sqrt{|C_i|^2 + |C'_j|^2 - |C_i \cap C'_j|^2}} \times \varphi(i,j)$$ **(Équation 17)**

Le coefficient de débordement O_2 (overlap 2) ordonné à pondération :

$$O^{2\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\left|C_i \cap C'_j\right|}{\max\left(|C_i|, |C'_j|\right)} \times \varphi(i,j)$$ **(Équation 18)**

Le coefficient de Dice généralisé ordonné à pondération :

$$E_{\alpha}{}^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{\alpha\left|C_i \cap C'_j\right|}{\alpha|C_i| + (1-\alpha)\left|C'_j \setminus C_i\right|} \times \varphi(i,j) \quad si\ 0 < \alpha < \frac{1}{2}$$ **(Équation 19)**

[15] f et φ ne sont pas fixées mais définies comme devant respecter un certain nombre de conditions (IPM2002).

$$E_{\alpha}^{\Omega}(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} \frac{(1-\alpha)\left|C_i \cap C'_j\right|}{\alpha\left|C_i \setminus C'_j\right| + (1-\alpha)\left|C'_j\right|} \times \varphi(i,j) \quad si\ \frac{1}{2} < \alpha < 1$$ **(Équation 20)**

Nous avons fait une comparaison expérimentale de 6 mesures : Cos, J, Cos_{l}^{Ω}, Cos_{p}^{Ω}, J_{l}^{Ω} et J_{p}^{Ω}. Cos_{l}^{Ω} et J_{l}^{Ω} étant construites avec une fonction φ de type linéaire, Cos_{p}^{Ω} et J_{p}^{Ω} étant construites avec une fonction φ de type puissance. Nous avons pu observer que :

- la fonction de pondération atténue considérablement l'effet particulier de chaque mesure; ainsi $Cos_{p}^{\Omega} \approx J_{p}^{\Omega}$ et $Cos_{l}^{\Omega} \approx J_{l}^{\Omega}$.
- les fonctions de pondération agissent avec plus ou moins d'influence en fonction des contextes; la pondération puissance est plus précise quand les ensembles à comparer sont très différents, à l'inverse lorsque les ensembles sont très similaires la fonction linéaire est plus précise.

3-3- Les mesures structurelles de proximité ordonnées : Construction à l'aide de la théorie de la logique floue

Le problème de construction de mesures ordonnées de type faible et fort a été résolu par Leo Egghe (Egghe, 2003) en utilisant le **formalisme des ensembles flous** (Zadeh, 1979). Le principe consiste à reformuler la définition de l'ensemble semi-ordonné C grâce à la fonction d'appartenance de la logique floue (équation 21), ensuite de réécrire les intersections et unions ensemblistes grâce à ce formalisme puis de les appliquer sur les formules classiques des mesures.

Les ensembles sont définis selon le formalisme suivant : $U_C = \bigcup_{i=1}^{n} C_i$ est un ensemble équipé de la fonction d'appartenance $P_{U_C} = \varphi(i) \Leftrightarrow x \in C_i$ où φ est une fonction strictement décroissante **(Équation 21).**

En utilisant la fonction $\varphi(i) = \frac{1}{2^{i-1}}$, les quatre résultats suivants sont démontrés (IPM2003):

$$\left|U_C \cap U_{C'}\right| = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left|C_i \cap C_j'\right| \frac{1}{2^{\max(i,j)-1}}$$ **(Équation 22)**

$$\left|U_C\right| = \sum_{i=1}^{\infty} \left|C_i\right| \frac{1}{2^{i-1}}$$ **(Équation 23)**

$$\left|U_{C'}\right| = \sum_{j=1}^{\infty} \left|C_j'\right| \frac{1}{2^{j-1}}$$ **(Équation 24)**

$$\left|U_C \cup U_{C'}\right| = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left|C_i \cap C_j'\right| \frac{1}{2^{i-1}} + \sum_{i=1}^{\infty} \sum_{j=1}^{j-1} \left|C_i \cap C_j'\right| \frac{1}{2^{j-1}} + \sum_{i=1}^{\infty} \left|C_i \setminus \bigcup_{j=1}^{\infty} C_j'\right| \frac{1}{2^{i-1}} + \sum_{j=1}^{\infty} \left|C_j' \setminus \bigcup_{i=1}^{\infty} C_i\right| \frac{1}{2^{j-1}}$$ **(Équation 25)**

En réécrivant les mesures de similarité avec l'intersection, l'union et le cardinal formulé en logique flou il est possible de construire nombre de nouvelles mesures de proximité ordonnées. Il est très intéressant de noter qu'elles peuvent se construire **à partir de mesures de type faible et fort** comme nous pouvons le voir ci dessous.

Jaccard ordonné :

$$J_F(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\left|U_C \cup U_{C'}\right|} = \frac{\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \left|C_i \cap C_j'\right| \frac{1}{2^{\max(i,j)-1}}}{\alpha}$$ **(Équation 26)**

avec

$$\alpha = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{i-1}} + \sum_{i=1}^{\infty}\sum_{j=1}^{j-1}\left|C_i \cap C_j'\right|\frac{1}{2^{j-1}} + \sum_{i=1}^{\infty}\left|C_i \setminus \bigcup_{j=1}^{\infty} C_j'\right|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}\left|C_j' \setminus \bigcup_{i=1}^{\infty} C_i\right|\frac{1}{2^{j-1}}$$ **(Équation 27)**

Dice ordonné :

$$D_F(C,C') = \frac{2\left|U_C \cap U_{C'}\right|}{\left|U_C\right| + \left|U_{C'}\right|} = \frac{2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}} + \sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}}$$ **(Équation 28)**

Cosinus ordonné :

$$Cos_F(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\sqrt{\left|U_C\right|\left|U_{C'}\right|}} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}}\right)\left(\sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}\right)}}$$ **(Équation 29)**

La mesure N ordonnée :

$$N_F(C,C') = \frac{\sqrt{2}\left|U_C \cap U_{C'}\right|}{\sqrt{\left|U_C\right|^2 + \left|U_{C'}\right|^2}} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\sqrt{\left(\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}}\right)^2 + \left(\sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}\right)^2}}$$ **(Équation 30)**

Le coefficient de débordement O_2 (overlap 2) ordonnée :

$$O_{2F}(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\max\left(\left|U_C\right|,\left|U_{C'}\right|\right)} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\max\left(\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}}, \sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}\right)}$$ **(Équation 31)**

Le coefficient de débordement 1 (overlap1) :

$$O_{1F}(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\min\left(\left|U_C\right|,\left|U_{C'}\right|\right)} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\min\left(\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}}, \sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}\right)}.$$ **(Équation 32)**

Le rappel :

$$R_F(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\left|U_{C'}\right|} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\sum_{j=1}^{\infty}\left|C_j'\right|\frac{1}{2^{j-1}}}$$ **(Équation 33)**

La précision:

$$P_F(C,C') = \frac{\left|U_C \cap U_{C'}\right|}{\left|U_C\right|} = \frac{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\left|C_i \cap C_j'\right|\frac{1}{2^{\max(i,j)-1}}}{\sum_{i=1}^{\infty}\left|C_i\right|\frac{1}{2^{i-1}}}$$ **(Équation 34)**

Comme précédemment nous avons fait une comparaison expérimentale prenant en compte :

- les mesures J_F , D_F , Cos_F N_F O_{2F} , O_{1F} , R_F , P_F définie ci dessus(équation 26-34)
- les mesures classiques J , D , Cos , N , O_2 , O_1 , R , P définies en début d'article (équation 2-10)

- la mesure de proximité ordonnée à pondération de type puissance J_p^{Ω} construite à partir du Jaccard comme indiquée dans l'équation 15. Rappelons que nous n'avons pas besoin de prendre en compte toutes les mesures ordonnées à pondération, en effet, un des résultats de (Egghe, 2003) est que la fonction de pondération supprime l'effet particulier de la mesure c'est a dire $J_p^{\Omega} \approx D_p^{\Omega} \approx Cos_p^{\Omega} \approx N_p^{\Omega}$

Les résultats ont montré en ce qui concerne les mesures de type fort (J_F, D_F, Cos_F N_F O_{2F}) que ces mesures sont plus **précises et plus représentatives sur les caractéristiques de similarité de rang et de documents communs** que les mesures classiques ou les mesures ordonnées à pondération. De plus, dans le cadre des mesures fortes, nous avons observé une plus grande **sensibilité** des mesures floues. En revanche, de telles conclusions n'ont pu être mises en évidence pour les mesures de type faible (O_{1F}, R_F, P_F), nous avons supposé que cela venait du corpus qui n'était pas une réelle collection test.

CONCLUSION

Dans le contexte de comparaison d'objets informationnels à éléments disjoints, nous avons proposé deux solutions originales permettant de construire des mesures de similarité prenant en compte l'ordre des éléments présentés et applicables aux objets informationnels construits selon un semi-ordre disjoint. Les solutions proposées s'appuient soit sur une fonction de pondération, soit sur une redéfinition structurelle des mesures en utilisant la logique floue. Le travail de recherche théorique est a notre sens complètement finalisé *dans ce contexte précis*, il ne reste à notre sens à réaliser que des expérimentations concrètes pour observer le comportement de certaines mesures, en particulier les mesures de type faible (O_{1F}, R_F, P_F). D'autres contextes sont cependant ouverts et nécessitent une réflexion théorique.
Le formalisme présenté n'est pas applicable dans le contexte de semi-ordre d'éléments disjoints c'est à dire en particulier la comparaison de réponses de systèmes basés sur des interfaces cartographique construites à partir de réseaux sémantiques, cas de figure souvent présenté dans les systèmes d'aide à la navigation. Nous pensons que ce problème peut être résolu grâce au formalisme de la logique floue, en redéfinissant l'appartenance à un ensemble comme il l'a été fait dans l'équation 21.
Un autre contexte reste complètement inexploré, celui des ensembles à éléments répétés comme les textes. Les notions de *semi ordre d'éléments ou de classes non disjointes* pourrait être particulièrement intéressantes pour ces derniers : la phase peut être considérée comme une classe composé d'éléments "mots", la racine lexicale peut permettre de créer des classes composées d'élément "mots", …. A notre connaissance, toutes les méthodes de comparaisons de texte s'appuient actuellement sur le nombre de mots qu'ils ont en commun, la prise en compte non seulement de l'ordre mais ici surtout de la granularité de classe, exprimée sous la forme de semi-ordre, devrait ouvrir considérablement les perspectives dans bon nombre de domaines comme la recherche d'information, la fouille de texte ou la bibliométrie entre autres.

BIBLIOGRAPHIE

Auziol, E. (2001). Problématique d'analyse de situations de communication et contextes multimédias In *Actes du colloque La Communication Médiatisée par Ordinateur : un carrefour de problématiques* - Université de Sherbrooke, 15 et 16 mai 2001.

Boyce, B.R., Meadow, C.T., & Kraft, D.H. (1994). *Measurement in information science*. Academic Press.

Egghe, L., C. Michel C. (2002). Strong similarity measures for ordered sets of documents in information retrieval, In *Information Processing & Management* 38 (6) (2002) pp. 823-848. (novembre 2002)

Egghe, L., C. Michel C. (2003). Construction of weak and strong similarity measures for ordered sets of documents using fuzzy set techniques In *Information Processing & Management* (à paraître)

Grossman, D.A., Frieder, O. (1998). *Information retrieval. Algorithms and heuristics.* Kluwer Academic Publishers, Boston

Fluhr, C. (1997).. SPIRIT.W3 : A distributed Cross.Lingual Indexing and Search Engine. *Proceeding of the INET 97 « The Seventh Annual Conference of the Internet Society »*. June 24-27 1997. Kuala Lumpur, Malaysia.

Lainé-Cruzel, S., Lafouge, T., Lardy, J.P., & Ben Abdallah, N. (1996). Improving information retrieval by combining user profile and document segmentation. *Information Processing and management*.32 (3), 305-315.

Losee, R.M. (1990). *The science of information. Measurement and applications.* Academic Press, Inc.

Michel, C. (1999). *Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et evaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs*. PhD Thesis. University Lyon II. 6 January 1999. 322 p.

Michel, C. (2000). Diagnostic Evaluation of a personalized filtering information retrieval system. Methodology and experimental results. *Proceeding of RIAO 2000 "Content-Based Multimedia Information Access"*. Paris, 12-14 april 2000.

Radasoa, H. (1988). *Méthodes d'amélioration de la pertinence des réponses dans un système de bases de données textuelles.* PhD Thesis. University Paris Sud-Orsay. 28 November 1988. 156 p.

Salton, G. & McGill, M.J. (1983). *Introduction to modern Information Retrieval.* New York : McGraw Hill.

Van Rijsbergen, C. J. (1981). Retrieval effectiveness In *Information retrieval experiments*. London : Butterworth - pp32-43.

Voorhees E.M., Harman D. – Overview of the seventh Text Retrieval Conference TREC 7. In *Proceedings of the seventh Text Retrieval Conference TREC 7. – Gaitherburg 9-11 november 1998*

Tague, J. (1990). Rank and sizes : some complementarities and contrasts. *Journal of information science*. 1990, 16 (1), 29-35.

Tague-Sutcliffe, J. (1995). *Measuring information. An information services perspectives*. Academic Press.

Zadeh, L. (1979). *Fuzzy sets and their applications to cognitive and decision processes.* Academic Press, Ney York.

Zamir O., Etzioni O.(1999) : Grouper : A dynamic clustering interface to web search results – In *Proceeding of the Eighth International World Wide Web Conference* – May 11-14 1999 – Toronto, Canada (http://www8.org)

DE L'UTILITE D'UNE VEILLE PEDAGOGIQUE

PINTE Jean-Paul
Enseignant Université Catholique de Lille - Université d'Artois à Arras
Doctorant - Université Marne La Vallée - Champ Sur Marne - 77454 Marne La Vallée
Tél : 06.80.60.04.35
Mail : pintejp@aol.com

Résumé : L'enseignement universitaire est condamné à se renouveler, à redéfinir ses paradigmes, sinon il se sclérose. Des changements pour ces derniers sont apparus depuis quelques années dans de nombreux domaines comme le téléphone sans fil, la médecine préventive, l'écologie, la mondialisation… Pour ce qui est de l'enseignement nous sommes entrés dans le paradigme de l'apprentissage. Les pressions nous viennent principalement du monde du travail avec entre autres la création de nouveaux environnements de travail, l'apparition de nouvelles caractéristiques de clientèles, une explosion des connaissances et des ressources, le développement fulgurant des Technologies de l'Information et de la Communication, et, surtout, l'arrivée de nouveaux étudiants de tout âge, de toute provenance, avec des motivations et des compétences diversifiées à l'extrême. En dehors de l'enseignement de matières au niveau le plus élevé de la connaissance, de la recherche et de la production de savoirs, l'université assure aujourd'hui un troisième rôle économique et social ayant pour objectif la production de valeur ajoutée et débouchant sur la recherche finalisée. L'économie du savoir supplante l'économie matérielle et les universités sont de plus en plus "entrepreneuriales". Une remise en cause profonde de l'université est déjà en cours. Elle vise à s'ouvrir à ce nouveau rôle par la mise en place de pédagogies " actives ", de formations ouvertes et à distance (E-learning, campus virtuels, numériques, …). Avec les TIC on ne peut plus enseigner comme avant. Des TIC aux TIC, il nous faut maintenant passer à " Technologies pour l'Intelligence et la Connaissance " . La veille pédagogique est une des principales clés de réussite pour accompagner ce changement. Cette article met l'accent sur l'intérêt de la veille pédagogique appliquée à la gestion des connaissances en université. Des exemples portant sur l'ingénierie propre à ce type de veille nous ouvrent les portes des nouveaux agents virtuels et physiques de construction de la connaissance et sur des plates-formes d'enseignement virtuels où étudiant et enseignant construisent et complètent leurs savoirs.

Abstract : University teaching has no choice but to be constantly renewed and to redine its paradigms, otherwise it will get stuck in a rut. In the last few years changes have appeared in many areas : mobile phones, preventative medicine, ecology and globalization for instance.... In the realm of teaching we have entered the learning paradigm. pressure is being put on us, principally from future employers with regard to the new work environment, the appearance of a new type of customer and an explosion of knowledge and resources, the breathtaking development of new communication and information technology and, above all, the arrival of students of all ages from all backgrounds with very varied abilities and motivation. In

addition to teaching different subjects at the highest level, supervising reserch and the production of knowledge, today's university has an economic and social role which aims at producing added value and an end product from its research. Knowledge economics are replacing material economics and universities are beginning to ressemble businesses. The universities are already engaged in profound reflections as to their role with the aim of finding ways of opening themselves up to this new role by setting up interactive teaching methods, courses of open or distance learning (e-learning, virtual campuses etc.). With the new technology and means of communication it is no longer possible to teach as we did before. From communication technology we have to move on to technological intelligence and knowledge. This article aims to show why pedagogy watch can be applied to knowledge management in universities. Examples taken from the appropriate engineering for this type of exercise open the door to new virtual and physical agents in knowledge-construction and virtual teaching platforms where the teacher and the student can advance together and build-up their expertise.

Mots clés : Gestion des connaissances, knowledge management, travail collaboratif, veille pédagogique, formation à distance

Keywords : Knowledge management, collaborative working, pedagogy watch, e-learning

De l'utilité d'une veille pédagogique

1 - ESSAI DE DEFINITION

La veille reste aujourd'hui encore une notion floue pour la plupart d'entre nous : elle est bien souvent comparée par méconnaissance à une forme de surveillance, voire d'espionnage.
Pourtant il est important de constater que parmi la typologie des différents domaines d'action de cette veille toute particulière qu'est la veille pédagogique qu'elle soit culturelle , environnementale, scientifique, sociétale, stratégique, technologique, axée sur le marketing, …), c'est l'éducation et ses acteurs qui auraient probablement le plus à gagner des activités d'une veille efficace et novatrice.
En effet, les ressources didactiques, les pratiques pédagogiques, scénarios et contenus de cours abondent sur Internet et croissent rapidement au fur et à mesure qu'apparaissent de nouveaux sites éducatifs.
La veille pédagogique vise à susciter, à promouvoir et à faire pratiquer les recherches qui sont réalisées dans le domaine de l'éducation et à offrir un lieu virtuel de rencontre pour les professeurs, chercheurs, enseignants à tous les niveaux, formateurs, étudiants et intervenants intéressés par ce sujet:

- en enrichissant les modes d'accès et de transmission de la connaissance et du savoir entre les différents acteurs internes et externes de l'université;
- en soutenant l'étudiant dans son apprentissage, renforçant son activité propre et contribuant à la lutte contre l'échec par des dispositifs d'enseignement sur mesure et de tutoriels d'auto-formation;
- en repérant (par exemple sur les sites web des établissements ou au cours d'animations) les ressources pédagogiques pertinentes proposées par les collègues de façon à les répertorier et à les valoriser via le web de l'établissement ;
- en proposant des pistes et des témoignages d'intégration de ces nouveaux outils en salle de cours dans la discipline ou dans le projet d'établissement ;
- en stimulant l'activité économique dans la valorisation et l'intégration des TIC dans l'enseignement ;

« La veille pédagogique peut donc être définie comme le processus d'intelligence qui consiste à détecter les signaux internes et externes, faibles ou forts susceptibles d'affecter l'université dans sa mission. La veille doit devenir un état second qui nous habite et nous aide à assurer la survie de nos institutions et à consolider nos positions stratégiques »[16].

Les nouveaux enjeux de la formation, les diverses contraintes liées à l'accès d'un nombre croissant d'étudiants à l'enseignement supérieur, l'exigence des méthodes adaptées à des publics diversifiés conduisent l'Université à proposer la création d'un centre de ressources NTE[17] afin de susciter et d'aider l'innovation et l'adaptation des pratiques éducatives, notamment grâce à l'utilisation des possibilités offertes par les « supports multimédias ».

Le grand nombre de sujets (disciplines) et les budgets amaigris de l'éducation ne sont malheureusement pas toujours propices au développement de ce type de service que certains ont commencé en France à qualifier de cellule ou d'observatoire de veille pédagogique.

Des objectifs possibles :

- identifier des ressources tels des scénarios d'apprentissage ou de matériel didactique;
- éviter la duplication des efforts de recherche et favoriser le partenariat;
- susciter et animer un rapprochement entre l'Université et l'entreprise et participer à la création de réseaux scientifiques et de professionnels, experts, chercheurs et praticiens;
- collecter, stocker, structurer et diffuser l'information utile à l'innovation de produits et de process ;
- Créer et développer les outils nécessaires à la communication régulière entre chercheurs et industriels: réseaux informationnels d'interconnexion, rencontres, journées d'étude, … ;
- Valoriser les connaissances scientifiques et techniques présentes et maximiser le flux des échanges avec l'environnement (littérature scientifique et publications internes) ;
- Participer au développement local et régional …

---

[16] Source : Mission pédagogique de la Mission pédagogique de l'Assemblée Nationale du Québec (http://www.assnat.qc.ca/fra/fondationbonenfant/veille/veille.htm)

[17] NTE : Nouvelles Technologies Educatives

Outre le repérage de ressources utiles et pertinentes dans le cadre de la conception d'un contenu ou de son design pédagogique, la veille pédagogique peut également fournir de l'information sur les nouvelles méthodes employées dans le secteur technique (ingénierie pédagogique) et contribuer à la formation continue des enseignants.

2 - LE XXIEME SIECLE : UNE SOCIETE DE LA CONNAISSANCE, DE L'INFORMATION ET DE L'EDUCATION

Le développement de l'Enseignement Supérieur a connu un rythme d'expansion des plus spectaculaires au cours de la seconde moitié du XXème siècle : il a été l'un des facteurs décisifs des avancées effectuées dans le domaine de l'éducation dans son ensemble et de l'extraordinaire progrès du savoir.
Ce dernier a assuré en quelques décennies un renforcement sans précédent des capacités d'avancement et de diffusion des connaissances, de leurs applications pratiques et des innovations technologiques.
Jamais aucune période de l'humanité comme le siècle de Périclès, la Renaissance ou encore le siècle des Lumières n'avaient été marquées comme le déclare Gilbert Paquette[18] par cet effet de masse que nous vivons aujourd'hui .
Cette turbulence avait déjà bien été amorcée dés les années 90 avec l'arrivée des contextes de mondialisation et de changement technologique.
La révolution est culturelle avant tout et touche toutes les sociétés au niveau planétaire.
Gestion des connaissances, outils de traitement de l'information, acquisition du « savoir », Knowledge Management[19], … etc. Pas un jour ne passe sans que l'un de ces termes ne fasse la une ou l'objet d'un article de presse.

Notre début de XXIème siècle marquera, quant à lui, un tournant, voire une révolution en termes de nouveaux paradigmes pédagogiques au sein de nos structures universitaires.
En dehors de l'évidente avancée de l'information et des techniques de communication confortée par l'arrivée massive d'Internet, formidable accélérateur de ce changement, les deux traits marquants de cette révolution sont d'une part l'explosion des connaissances avec les moyens de production, de stockage et de diffusion de ces dernières, et d'autre part, la transformation de l'environnement dans lequel elle se déploie et qui prend figure de mondialisation des échanges économiques et culturels caractérisée par la circulation des biens matériels et immatériels et des personnes à l'échelle de la planète. Ces effets sont aujourd'hui considérables dans de nombreux domaines, mais dans l'éducation, ils n'en sont qu'à leurs débuts…

Ce double mouvement, comme le précise Céline Saint-Pierre[20] dans son discours d'introduction du colloque (Du livre à Internet, quelle(s) université(s) ?)[21], « crée le besoin et la nécessité pour l'université de se redéfinir comme système d'action et de revoir ce qui crée son identité de même que le sens de son action et de ses activités institutionnelles, soit les activités de gestion, d'enseignement et de recherche, et de services aux collectivités.
Tous les acteurs animant cette institution sont interpellés dans leurs raisons d'être et leurs façons de faire. »

Dans un tel contexte, l'apprentissage humain prend une importance nouvelle qui est définie comme le processus par lequel des informations, éparses ou structurées dans les domaines du savoir, deviennent des connaissances et des habiletés intégrées à l'intellect d'un individu, lui permettant d'exercer des compétences nouvelles.
Le télé-apprentissage gagne aussi du terrain et aujourd'hui n'importe quel quidam est à même de trouver un cours disponible sur l'Internet sur le thème, à l'endroit et à l'heure qui lui plairont.
La Formation A Distance (FAD), le e-learning, ou encore le « blended-learning » (système de formation alliant le face à face et la formation à distance) deviennent progressivement des modes de formation de plus en plus incontournables constituant ainsi le fait marquant de la décennie qui commence.

Au XXIème siècle, l'Enseignement Supérieur doit faire face dans ses activités d'enseignement et de recherche aux effets et aux conséquences du processus de mondialisation et de l'internationalisation de la vie des sociétés, du développement des technologies de l'information, de l'évolution rapide de la structure des besoins en matière d'emploi et de l'augmentation continue des besoins en personnels hautement qualifiés.
La nécessité de mise à jour et de perfectionnement des connaissances générales et professionnelles et la reconversion professionnelle de plus en plus pressante de nouveaux publics sont également à son programme.

---

[18] Gilbert Paquette – L'Ingénierie pédagogique – 2002 – Presses de l'Université du Québec

[19] Knowledge Management (gestion du savoir) : Concept qui vise à partager au sein d'un groupe, d'une entreprise, ou d'une quelconque organisation l'ensemble de la connaissance et du savoir de cette entité.

[20] Céline Saint-Pierre est *Présidente du Conseil Supérieur de l'Education du Québec – www.cse.gouv.qc.ca*

[21] Colloque Franco-Québecois - INJEP - *Juin 2002 - http://www.fse.ulaval.ca/ext/cipte/Programme.pdf*

L'importance grandissante du savoir couplée avec l'évolution du nombre des apprenants ou formés par l'enseignement supérieur accroissent comme le rappellent Frédérico Mayor et Sema Tanguiane dans leur ouvrage « l'Enseignement Supérieur au XXIème siècle »[22] la responsabilité et l'influence du savoir dans la société.
Les conférences régionales sur l'Enseignement Supérieur de 1996 à 1998 à La Havane, Dakar, Tokyo, Palerme et Beyrouth ajoutées à celle de Paris en 1998 ont ouvert la marche et ont eu pour objectifs de sensibiliser et de faire prendre conscience du virage qui s'opérait.

Le contexte en évolution rapide de la situation internationale des principaux pays exportateurs de formation par Internet comme les Etats-Unis, la Nouvelle Zélande, l'Australie et le Canada ainsi que les nombreuses initiatives et instances développées en termes de e-learning et de campus virtuels ou numériques par la Grande Bretagne, le Danemark, la Finlande, les Pays-Bas, la Norvège et le Japon préfigurent également de la croissance rapide de ce nouveau marché qu'est la formation en ligne.[23]

En France, les «appels à projet campus numériques français » lancés en 2000 par le Ministère de la Recherche avec le soutien financier et logistique du DATAR[24], de l'AUF [25]et l'implication du CNED[26] ont cherché à faire naître des projets inter-établissements, ouverts à des partenaires internationaux et du monde de l'entreprise. (Dotations financières de 9,3 M€ en 2002).
Cette logique de consortium a entraîné la mutualisation des compétences pour garantir une qualité élevée et donner une visibilité nationale et internationale aux campus numériques.
Pour la première fois, la France avec près de 200 projets déposés, est devenue le deuxième pays porteur de projets du plan e-learning de la Commission Européenne[27].

On a eu en effet trop tendance à considérer jusqu'à nos jours l'éducation comme une branche de l'économie et à ne pas privilégier son soutien en termes de financement.

Sir W.Arthur Lewis[28] conseillait :
« A l'aube du XXIème siècle se fait sentir le besoin urgent «d'éduquer l'économie » et non pas « d'économiser l'éducation »

Dans ce contexte, une question se pose: l'université est-elle encore ce lieu de production et de diffusion de haut savoir ayant le monopole de la formation spécialisée et de pointe et constitue t-elle encore cet espace institutionnel dédié à cette mission première rassemblant dans un même lieu et dans une même unité de temps, ceux qui produisent et transmettent la connaissance et ceux qui sont en processus d'apprentissage et de formation ?

3 - LES NTE[29] A L'UNIVERSITE, LES PROFESSEURS DOIVENT S'EN MELER RAPIDEMENT

En cette période d'accroissement phénoménal de l'information associé à un déclin significatif des ressources financières disponibles, l'Université doit faire preuve d'imagination et de créativité pour réussir à remplir les différents mandats qui lui sont confiés.
L'intégration des NTIC aux activités quotidiennes d'enseignement est, depuis un certain temps déjà, perçue par plusieurs comme la solution aux problèmes pédagogiques qui assaillent l'Université. Cette « solution miracle » risque, si on l'applique sans discernement, d'entraîner rapidement désillusions et frustrations même chez les plus enthousiastes.

---

[22] *L'Enseignement Supérieur au XXIème siècle (Editions Hermes France)*

[23] Campus numériques , enjeux et perspectives pour la formation ouverte et à distance – Rapport de mission sous la direction de Michel Averous et Gilbert Touzot – Avril 2002

[24] DATAR : Délégation à l'Aménagement du Territoire et à l'Action Régionale)

[25] AUF : Agence Universitaire de Francophonie

[26] CNED : Centre National de l'Enseignement à Distance

[27] *La Commission Européenne définit le e-learning comme « l'utilisation des nouvelles technologies multimédias et de l'Internet pour améliorer la qualité de l'apprentissage en facilitant l'accès à des ressources et des services, ainsi que les échanges et la collaboration à distance».*

[28] *Sir W.Arthur Lewis (W.A., Economic aspects of quality in education. Qualitatives Aspects of Educational Planning. Unesco. IIPE, 1969, p87)*

[29] NTE : Nouvelles Technologies Educatives

Il est donc essentiel, si l'on veut réussir l'intégration harmonieuse des NTIC aux outils pédagogiques déjà en place, de bien en cerner les forces et les faiblesses, et surtout, de bien identifier les situations pour lesquelles ces nouveaux outils seront les plus prometteurs et adaptés à la « pédagogie universitaire ».

La diffusion des connaissances, par le biais de systèmes multimédias plus ou moins interactifs et à distance, peut permettre d'accroître l'accessibilité au savoir, l'enrichissement des contenus de formation, la dynamique et l'actualisation permanente des connaissances, l'individualisation de l'apprentissage, l'adaptation de l'organisation universitaire aux nouvelles conditions socio-économiques de la clientèle étudiante et une plus grande interactivité entre le professeur et l'étudiant, solution éventuelle à des problèmes d'encadrement de plus en plus grands auxquels nous faisons face depuis longtemps.
Mais cela ne vaut uniquement que si nous nous posons comme objectif de départ d'accroître la qualité de la formation universitaire et son accessibilité.
Qu'on le veuille ou non, l'Université en tant que campus traditionnel est appelée à se transformer et non à disparaître comme certains auraient pu le laisser entendre, il y a encore cinq ou six années.
La peur de disparaître a aujourd'hui fait place à la nécessité de faire face à ce nouvel environnement, tout en développant une vision critique.

L'Université devra néanmoins changer avec tous les enjeux que cela comporte au plan de sa mission, de son organisation, de ses liaisons avec les autres lieux de conception et de diffusion du savoir et au plan du rôle du corps professoral.
L'élaboration d'une nouvelle problématique sur le rôle de l'université fait désormais partie de son agenda. Elle fait ressortir la nécessité non seulement de réaffirmer sa mission première de formation et de recherche, qu'elle devra réactualiser en y intégrant deux nouveaux paradigmes, celui de l'apprentissage et celui de l'éducation tout au long de la vie, en favorisant, dans l'enseignement et dans l'apprentissage, le recours à ces nouveaux outils que sont les TIC[30].

Tous reconnaissent maintenant que cette explosion de connaissances et la place centrale occupée par le savoir et la technologie dans l'économie de nos sociétés situent plus que jamais l'université au cœur du développement économique, social et culturel.
Encore faut-il saisir cette chance de solidification de sa mission dans cette nouvelle perspective qui ne peut faire l'économie de la qualité et de la pertinence de l'enseignement supérieur dans ce nouveau contexte.
Pour le sociologue Michel Serres, la société de l'information donne à l'éducation une place centrale et nouvelle, qu'il qualifiait de « société éducative », lors d'une conférence à la Fédération des cégeps du Québec le 16 octobre 1999.
Le savoir change de nature et les supports informatiques dont Internet multiplient les portes d'entrée à la connaissance.
Les mécanismes de transmission des connaissances se modifient et posent les questions du « quoi enseigner et du comment enseigner ? ».
Dorénavant, l'accès à une information abondante de toute nature et de qualité variable qu'offre le branchement en réseau oblige le système d'éducation à jouer un rôle prédominant dans la formation nécessaire à un usage éclairé de ces informations et à leur transformation éventuelle en savoir maîtrisé.

- mettre à profit les TIC dans une perspective d'intégration réussie dans l'enseignement et l'apprentissage
- réussir l'intégration pédagogique des technologies dans l'apprentissage et l'enseignement
- apprendre autrement, enseigner différemment (méthodes actives, apprentissage par problème, etc…)
- s'assurer un avenir prometteur dans la société de l'information et du savoir
- …

Autant de bouleversements qui avec la révolution « informationnelle » qu'entraînent l'émergence des technologies nouvelles et leur pénétration dans le grand public, montrent bien que c'est le fonctionnement même de l'université qui se pose aujourd'hui avec acuité.
A ce sujet, Jean Claude Guédon [31] constate que « l'université virtuelle affaiblira fort probablement les empires internes des universités que sont les départements et les facultés (…) Cela perturbera les structures disciplinaires et départementales et une pression se fera sentir en faveur d'un décloisonnement ».

---

[30] TIC : Technologies de l'Information et de la Communication

[31] Diplômé en chimie et Docteur en histoire des sciences, Jean-Claude Guédon est professeur de littérature comparée à l'Université de Montréal, il est également l'auteur de La planète Cyber: Internet et cyberespace (Découvertes, Gallimard)

4 - DES CHANGEMENTS CULTURELS, ECONOMIQUES …

Des objectifs de Newman, il y a plus d'un siècle (Université des Aristocrates), nous passons aujourd'hui à une université où l'offre de formation devient un marché lucratif et où l'étudiant est « actif ».
Concilier apprentissage et action devient indispensable. Il y a aujourd'hui des inventions à exploiter dans notre pédagogie et il faut maintenant se résigner à penser que l'on ne pourra plus jamais enseigner comme avant.

Il faut aussi changer notre représentation des TIC :

- 1- en explorant, découvrant et expérimentant des réalisations, services ou sites pas toujours aussi froids et inhumains que l'on pourrait se l'imaginer au premier abord ;
- 2- en s'intéressant aux transformations cognitives par les TIC et en découvrant les potentialités au regard du développement cognitif ;
- 3- en expérimentant les capacités de médiation sociale des TIC et en les exploitant comme socio-médias.

D'une posture de spectateur l'apprenant va donc passer progressivement à celle d'utilisateur - créateur, producteur et communicateur.
Quant au Professeur, il n'aura plus d'ici 2010 ce rôle de transmetteur de connaissances et de « pontificateur » à quelques exceptions près de cours magistraux et séminaires se révélant toujours efficaces, mais plutôt celui de «facilitateur d'apprentissage» intervenant ici et là pour questionner, suggérer, tutorer, encourager et guider les étudiants dans leur recherche d'informations sur le Web par exemple…

L'utilisation des nouvelles technologies et la création progressive des « universités virtuelles » vont différencier et alourdir les tâches de ce que l'on appelle encore, rituellement, des «enseignants-chercheurs».
L'université devra donc rapidement composer avec le « marché » de la formation en ligne ou du télé-apprentissage et s'y tailler une place tout en personnalisant ses offres de formation avec une souplesse et flexibilité que son système éducatif devra apprendre à développer.
Le rapport Esperet[32] envisageait ainsi de les transformer en tuteurs, administrateurs, conseillers, démarcheurs, diffuseurs d'information, constructeurs de réseaux et de cours en ligne, et accessoirement (la nuit ou pendant les vacances) un peu chercheurs.
La crainte des remous électoraux a provisoirement écarté la mise en œuvre de ces propositions. Elles impliquaient une contractualisation individuelle des personnels de la création (déjà amorcée de fait avec les personnels à statut dérogatoire dans les universités) d'une université et d'universitaires déliés de l'obligation de recherche pour cause d'utilité sociale ou pédagogique.
La relance du thème de l'autonomie par le nouveau ministre, Mr Luc Ferry, est peut-être le signe avant-coureur de la reprise de cette individualisation des fonctions sous la houlette des présidents managers comme le soulignent Christophe Charle et Daniel Roche dans le monde du 11 juillet 2002.[33]
Cette évolution en filigrane est d'autant plus préoccupante, signalent-ils, qu'on la retrouve au niveau international. Le rêve d'une université virtuelle vendant à des clientèles délocalisées diplômes et formations est un programme largement mis en œuvre par certaines universités anglo-saxonnes. Il suscite des tentatives d'imitation dans certaines universités françaises.
Vantées pour leur modernité et leur rentabilité, ces procédures sont très inégalement adaptées aux différents domaines du savoir. Elles dépendent très largement de la solvabilité (corrélée à l'origine sociale et géographique) des étudiant(e)s - Internet.
On ne risque pas grand-chose à prédire qu'elles renforceront les dominants : les disciplines les plus riches (et les plus rentables pour le placement professionnel) des universités les plus riches pourront attirer les étudiants les plus riches des pays aux universités sinistrées.
Seule une minorité d'équipes et d'universités françaises pourront jouer ce jeu de la mondialisation technologique et professionnelle.

Le plus grand nombre restera hors jeu parce que leurs domaines n'entreront guère dans les critères de rentabilité, par attachement à un autre idéal de leur métier ou parce que les arbitrages entre les ressources à affecter aux divers modes d'enseignement se feront à leur détriment.

---

[32] http://www.education.gouv.fr/rapport/esperet/default.htm

[33] Le Monde – Rubrique Horizons/Débat du jeudi 11 juillet 2002

Charles Roche est professeur d'histoire contemporaine à l'Université Paris I

Daniel Roche est professeur au Collège de France (chaire d'histoire de la France des lumières)

Ils s'expriment en tant que responsables de l'ARESER (Association de Réflexion sur les Enseignements Supérieurs et la Recherche)

Des conflits et débats houleux naissent déjà sur ce sujet au sein de plusieurs universités.

5 - LA GESTION DES CONNAISSANCES, PAS SEULEMENT POUR LES ENTREPRISES …

L'évolution de l'informatique ces vingt dernières années est sans précèdent.
Malheureusement cette extraordinaire évolution n'a pas eu que des effets bénéfiques et l'un des effets secondaires a été une croissance toute aussi exponentielle, non contrôlée, du volume des données publiées électroniquement.
L'ordinateur s'est transformé en une super machine à produire de l'information.
Il y a vingt ans la denrée rare était l'information, aujourd'hui elle est pléthorique et nous devons nous équiper d'outils pour nous repérer dans une vraie jungle informationnelle.
Ainsi « *trop d'infos tue l'info* » entend on souvent dire autour de soi lorsqu'il s'agit de trouver la bonne !!!.
On peut discerner à travers ces remarques le rêve que caresse alors chacun :
« *Apportez moi l'information dont j'ai besoin, au moment où j'ai besoin*. »

Contrairement au système d'information qui s'inscrivait dans une logique d'accumulation, on est ici, dans le cadre du management de la connaissance centrée sur l'utilisateur final et sur trois concepts clés qui régissent son rapport à l'information : le repérage ,la pertinence et la capitalisation.

Repérage :
Le lecteur ne lit plus ce qu'il reçoit mais cherche à repérer l'information que le document contient, dans le but de le classer. (Dans sa mémoire ou dans son système de classement), de façon à pouvoir y revenir lorsqu'il en aura besoin.
Il active donc sa mémoire de repérage et non de contenu.

Pertinence :
C'est le ratio de l'information utile par rapport au bruit[34], c'est en quelque sorte la « productivité relationnelle » du document qui est en cause.
Le problème ne dépend pas alors uniquement de l'auteur mais surtout du lecteur, et notamment de deux facteurs inhérents à ce dernier.
Son niveau de connaissance sur le sujet et son mode de relation avec la source d'information.

Capitalisation :
Thomas Jefferson avait compris le vrai sens du concept de capitalisation des connaissances lorsqu'il disait : « Celui qui apprend quelque chose de moi enrichit son savoir sans réduire le mien, tout comme celui qui allume sa chandelle à la mienne se donne de la lumière sans me plonger dans l'obscurité ».
Le mot capitalisation doit être compris comme « cultiver pour faire germer, fructifier » et non comme « collecter, ranger et conserver dans une armoire ».
Aujourd'hui , pour survivre les entreprises ne doivent pas se contenter uniquement d'investir dans le capital physique, il leur faut également tirer parti de ce capital physique grâce à l'acquisition des connaissances, leur diffusion, leur capitalisation et leur exploitation.

Les universités vont tendre de plus en plus vers ce schéma qui consiste autant pour les producteurs (enseignants) que pour les consommateurs (étudiants) à utiliser ce processus de création, d'enrichissement, et de capitalisation des savoirs.
En effet, grâce aux réseaux, aux techniques de traitement informatique et à la numérisation, chaque université et chaque étudiant, à son poste de travail ou à domicile, peuvent accéder à une masse impressionnante d'informations. Une telle disponibilité, si elle est bien exploitée, offre à celui ou à celle qui sait la valoriser dans son travail, dans sa formation ou dans ses loisirs, une possibilité d'épanouissement ou un avantage concurrentiel de première importance.
Gérer, organiser et traiter les données disponibles à tout moment et à n'importe quel endroit de la planète pour produire un cycle de formation, construire un parcours culturel, conduire une recherche exigent alors une maîtrise de la gestion des données et une capacité d'innovation qui sont aujourd'hui indispensables à la création de valeurs par l'université ou à l' acquisition de savoirs par l'étudiant.

De même le partage des connaissances n'est plus un concept nouveau convenant au seul monde de l'entreprise. La FAD (Formation A Distance), les systèmes de formation e-learning ont par exemple tout intérêt à adopter les techniques développées pour les grandes enseignes.

[34] Bruit désigne toute réponse non pertinente à une recherche documentaire (AFNOR)

Le croisement entre e-learning et le Knowledge Management a permis de comprendre que la connaissance doit être envisagée sous plusieurs aspects : information, formation et communication, en utilisant toutes les atouts offerts pat les nouvelles technologies.

Comme pour les entreprises, le Knowledge Management a pris le pas dans les universités et désormais l'information est conçue comme un actif productif pour le capital des universités ayant pour nouvel objectif : Diriger les connaissances vers ceux qui en ont besoin.

Nous rentrons dans une logique de flux ou la veille pédagogique va jouer un rôle important.

6 - LES NOUVEAUX AGENTS DE CONSTRUCTION DE LA CONNAISSANCE

L'intégration des agents de construction de la connaissance dans les applications liées aux nouvelles technologies nous confronte à de nouveaux défis.
Avec les TIC la définition des agents[35] est devenue plus complexe et dans le contexte qui prévaut ici il conviendra de définir ces derniers comme des « intermédiaires actifs dans un système de transfert de la connaissance »[36].
Quant à la construction des connaissances, elle est issue d'une nouvelle théorie de l'apprentissage.

Ces agents de construction de la connaissance se divisent en deux parties : les agents physiques et les agents virtuels
Comme il est défini dans le projet VIGIE mené au Québec, dans son système d'agent de le construction de la connaissance (ACC), « les experts et les novices sont des agents physiques qui partagent un champ de connaissance circonscrit par les limites des outils disponibles que sont les agents virtuels ».
Un agent de la construction de la connaissance est donc avant tout un système d'auto-formation en mode collaboratif.
A l'aide d'agents virtuels (robots, engins, formulaires, bases de données) et d'agents physiques (experts, enseignants, professeurs, formateurs, l'apprenant novice crée son propre réseau de connaissance avec ou sans distance entre les intervenants. Certains agents peuvent même être utilisés dans un laboratoire informatique dans un mode de partage et de construction en temps réel.

Un modèle théorique de construction des connaissances ne peut prétendre englober toutes les possibilités de constructions, ni de connaissances. Il convient donc aujourd'hui pour l'université de construire différents modèles selon les probabilités d'utilisation et d'application des champs des connaissances abordés.
Les agents permettent d'emmagasiner une quantité d'informations disparate dans une base de données organisée selon une taxonomie [37]déterminée par les experts.
Une fois cette base construite, elle permet à l'aide d'une interface d'ajouter des éléments de connaissance et d'en extraire le sens dans le but de créer son propre système de connaissance.
Il conviendra donc en fonction des particularités des domaines de connaissance de construire les agents sur mesure avec leurs contraintes et leurs champs de compétence.
Aucun consensus sur l'élaboration des agents n'exista à ce jour et chacun construit son agent avec les outils dont il dispose. Il est fort probable qu'un jour des protocoles de base seront établis pour permettre la création de méta-agents.

De nos jours les agents physiques humains sont les individus enseignants, professeurs , formateurs et leur rôle est de construire les structures des bases de connaissances, mettre en place la taxonomie pertinente à l'agent, participer à la conception du design pédagogique et de l'interface, alimenter de son expertise le processus.
Dans un cadre plus global, comprendre le sens de l'évolution et de ses tendances pour permettre à leurs congénères de survivre dans un monde en perpétuelle mutation.

Les agents virtuels sont avant tout des outils permettant une certaine forme d'intelligence artificielle limitée. Leur durée de vie est relative à l'ampleur de la tâche qui leur est confiée.
Une fois la tâche terminée, il s'éteint jusqu'à la prochaine réactivation. Il ne se reproduit pas, du moins pas encore .

---

[35] Agent : le Petit Robert « être qui agit ou encore ce qui agit, opère ; corps, substance intervenant dans la production de certains phénomènes » ou « personne physique ou morale jouant le rôle d'intermédiaire dans les opérations commerciales, industrielles et financières »

[36] Définition extraite du projet VIGIE mené au Québec et ayant débouché sur le portail Cantic (Regroupement des collèges Performa, Cégep du Vieux Montréal, Université de Sherbrooke)

[37] Taxonomie : Méthode de catégorisation de l'information à l'aide d'un vocabulaire contrôlé et normalisé.

Ces agents sont de conception très complexe mais peu d'élément composent leur fabrication comme les algorithmes[38] et les scripts[39] écrits avec des langages de haut niveau informatique (C++, Visual Basic, Java, Javascript, etc.)

Gerber et Gignoux en 1997 définissent les agents virtuels intelligents comme un « système informatique hardware ou (plus souvent) logiciel qui répond aux propriétés suivantes :

- Autonomie : les agents opèrent sans intervention directe d'être humain ou autre, et on un certain contrôle sur leurs actions et leur état interne.
- Comportement social : les agents interagissent avec d'autres agents (éventuellement inhumains) via une sorte de langage de communication agent.
- Réactivité : les agents perçoivent leur environnement (qui peut être le monde physique, un utilisateur via une interface graphique, une collection d'autres agents, l'Internet ou même tout à la fois) et répondent aux changements qui apparaissent.
- Comportement traditionnel : les agents n'agissent pas simplement en réponse à leur environnement, ils sont capables d'avoir un comportement dirigé vers un but et de prendre des initiatives.

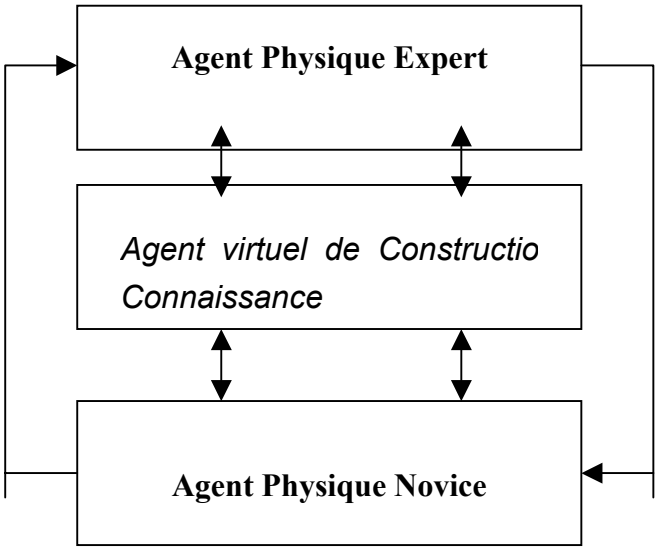
Dans le domaine de l'intelligence artificielle, le terme « agent » a un sens plus fort et plus spécifique . Pour eux, un agent est un système informatique qui, en plus des propriétés citées précédemment , est conceptualisé ou implémenté selon des notions que l'on attribue plus couramment aux humains. (ex : notions mentales comme la connaissance, les convictions, l'intention ou l'obligation).
Certains chercheurs parleront même d'agents émotionnels.
D'autres attributs font également l'objet de discussions concernant les agents :

- La mobilité est la capacité d'une agent à se déplacer dans un réseau informatique.
- La véracité est la conjecture selon laquelle un agent ne communique pas de mauvaises informations sans le savoir.
- Le bénévolat est la conjecture selon laquelle les agents n'ont pas de buts incompatibles, et que chaque agent essaiera de faire ce qu'on attend de lui.
- La rationalité est la conjecture selon laquelle un agent agira de sorte à atteindre ses objectifs, au moins dans la limite des ses convictions.

[38] Algorithmes : *formules mathématiques utilisées par les programmeurs de systèmes et d'applications pour créer des objets, des logiciels, des engins et de moteurs virtuels à l'intérieur desquels circulent les contenus multimédia.*

[39] Scripts : *fondements de l'industrie cybernétique, ils représentent les codes originels (code source) écrits en mode textuel à l'aide des langages de programmation qui intègrent les algorithmes. Ce sont des intermédiaires entre l'humain et la machine. Sans eux il est impossible de faire fonctionner un ordinateur.*

6-1 Schéma relatif aux agents [40]



Le schéma ci-dessus illustre bien les relations fondamentales qui conditionnent la construction de nouvelles connaissances.

L'agent physique expert contrôle le développement de l'agent de construction de connaissance (ACC) et de l'interface représenté par les flèches .

L'agent physique novice construit sa connaissance en ajoutant de l'information et en créant un sens à partir des connaissances intégrées dans l'ACC.

Le rôle de l'expert est de valider le réseau des connaissances acquises par le novice.

Cette activité doit se dérouler à l'extérieur de l'agent par un réseau parallèle de communication verbale ou par une présence physique.
On peut bien sûr utiliser ces agents sans aucune intervention humaine, le rôle de l'intelligence artificielle étant justement de créer des substituts à l'être humain.
Cependant l'être humain devra toujours trouver le moyen de prendre sa place et c'est là le plus grand défi pour les générations futures.

[40] Source : http://www.cvm.qc.ca/cantic/beta/2nouve/13virtuels.htm

6-2 Caractéristiques d'un agent virtuel :

Un agent virtuel peut posséder toutes ou certaines de ces caractéristiques :

- Interface unifiée avec un ou plusieurs formulaires pour permettre l'insertion de l'information dans la base.
- Une ou plusieurs bases de données où l'information insérée est classifiée selon une taxonomie établie par des experts.
- Un ou des formulaires de requêtes pour effectuer des demandes de recherche de l'information.
- Un ou plusieurs types de pages de rapport de résultats de la requête qui vont permettre après une configuration propre à un individu d'établir un profil particulier adapté à la construction de la connaissance d'un apprenant.
- Un système de courriels, de forums ou des communications en mode synchrone (chat).
- Il est aussi possible de greffer d'autres agents ou objets pédagogiques qui viendront renforcer la rétention des connaissances (Questionnaires, tests formatifs, etc.).

Un agent de construction de la connaissance n'est donc pas limité dans le temps mais est un outil de longue portée qui s'inscrit dans un cadre plus vaste de références à long terme.
En résumé, c'est un outil de consultation et de communication dynamique qui permet à des individus experts et novices de participer à la construction d'une base de connaissances commune sur un champ de connaissance circonscrite

6-3 Exemples de la vitrine APO [41] et du portail CANTIC

La vitrine APO :

L'intérêt d'un système de veille automatisée est d'effectuer une surveillance de certains sites et de signaler à l'utilisateur les changements observés.

- identifier des ressources tels des scénarios d'apprentissage ou de matériel didactique ;
- adapter ses pratiques et son rôle à l'usage croissant des nouvelles technologies ;
- éviter la duplication des efforts de recherche et favoriser le partenariat [42]

Le site de veille pédagogique accessible via Internet permet d'accéder à deux services distincts :

1. l'identification de sites pertinents en lien avec certaines disciplines des programmes de formation réseau collégial sans devoir rédiger des requêtes pour les différents moteurs de recherche. Chacun des thèmes sera subdivisé en sous-thèmes et en sujets ;
2. la réception, selon les spécifications de la requérante ou du requérant, d'un courriel avisant des modifications qui ont eu lieu sur le réseau Internet en rapport avec un sujet déterminé. Chaque usager pourra être informé sur un maximum de 10 sujets.

La création du corpus disciplinaire (thèmes, sous-thèmes et sujets) ainsi que l'expérimentation du réseau collégial ont été réalisés par 40 enseignants en 1998-1999 puis le site a été rendu accessible pour l'ensemble de la communauté collégiale à partir du site de la vitrine APO.
En tant qu'agent de construction de la connaissance, La veille automatisée donne un des éléments fondamentaux de l'agent soit celui que le collaborateur puisse créer son profil taxonomique. L'agent conserve le profil des types de recherches courantes de l'usager et effectue une mise à jour des recherches en délestant les éléments des connaissances périmées.

Le portail CANTIC

Dans le cadre de son plan de travail 1998-1999, le cégep du Vieux Montréal s'est doté d'un comité de vigie dont le rôle est d'explorer, colliger et analyser les tendances du marché du travail, les nouveaux emplois qui sont créés et le domaine de l'éducation en général. Cette base de connaissances était à construire et le cégep devait mettre au point un outil qui leur permettrait d'atteindre un certain niveau opérationnel. Le présent projet leur a servi de déclencheur pour construire un agent de construction de connaissance générique. Car une fois construit les composants objets étaient réutilisables dans d'autres agents.

Qualités du processus **(cf : schéma ci dessous)**

[41] La Vitrine APO est un réseau de veille en éducation dont fait partie une centaine de collèges et d'universités. En consultant le site, les élèves peuvent accéder à des actualités, des répertoires de sites éducatifs, une bibliothèque virtuelle de périodiques et un ABC du multimédia. Cf : http://www.ntic.org

[42] Extraits de la présentation du site «Veille technologique automatisée » (Pierre-Julien Guay, 1999)

- Souple
- Efficace
- Robuste
- Esthétique
- Constructif orienté vers les projets
- Simple tout en étant capable d'intégrer la complexité

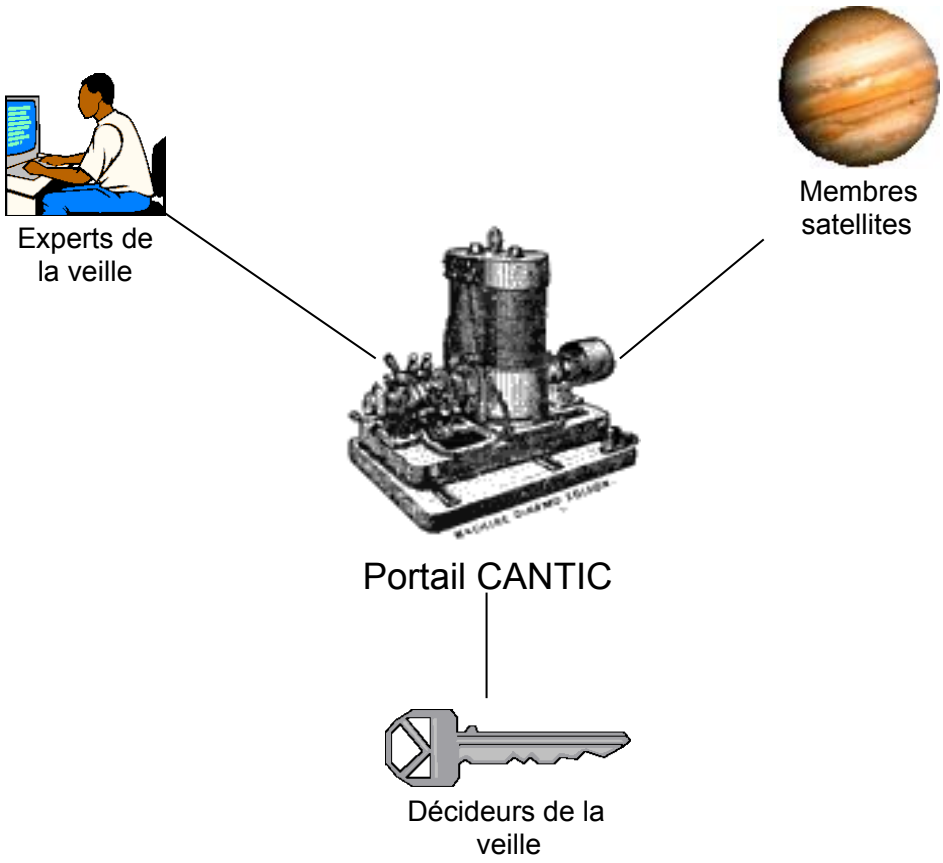
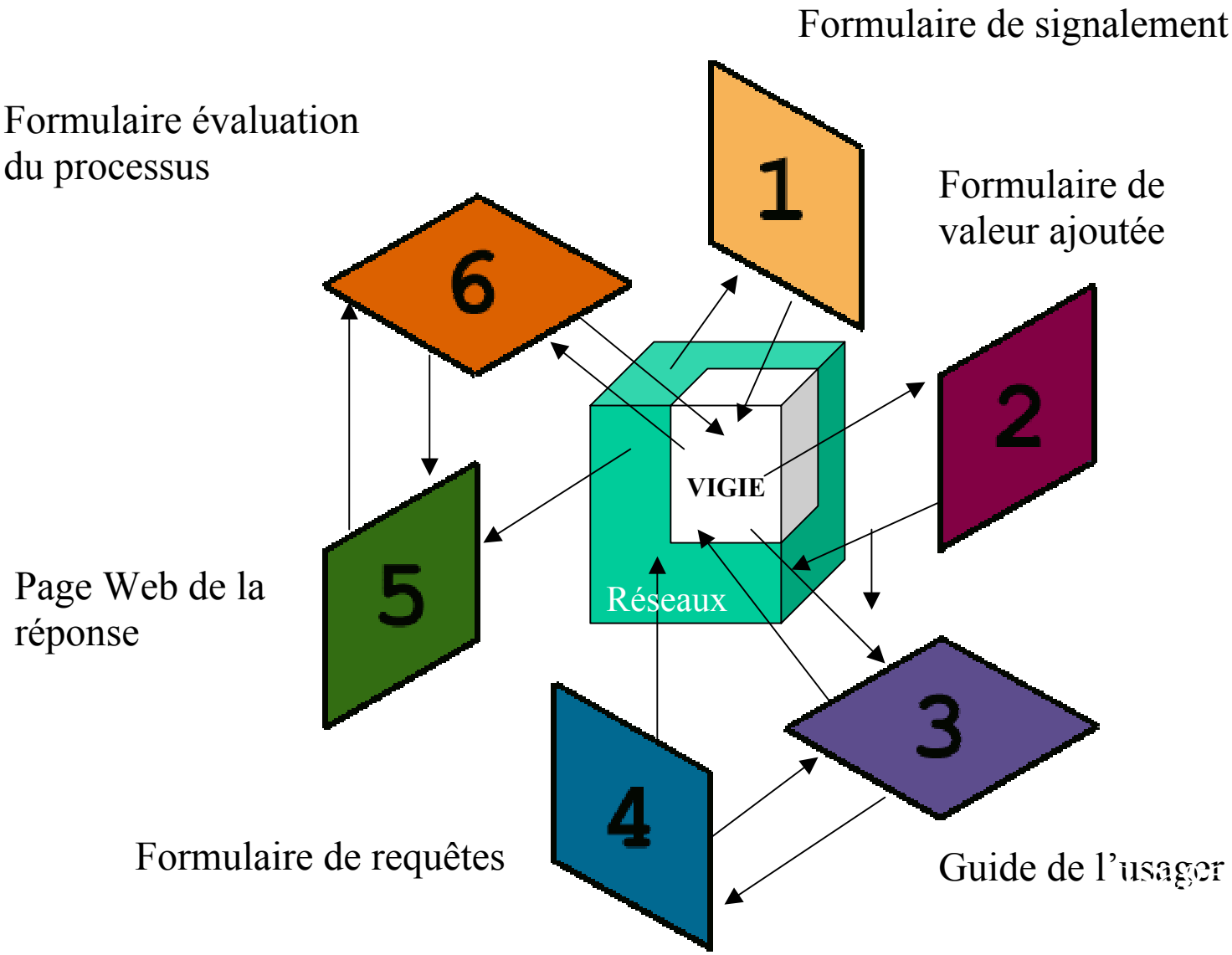


Schéma représentatif du processus de VIGIE[43]



Étapes du processus de la vigie

1. Recherche d'informations pertinentes par les journaux, par le Web, les conversations, les colloques, etc.
2. Mise en forme de cette information avec un formulaire électronique.
3. Acheminement du formulaire à l'expert de la vigie.
4. Vérification des sources et ajouts d'informations complémentaires.
5. Publication de l'information sur le site de la vigie publique et/ou privée.
6. **Communication de la publication aux membres des vigies.**
7. Discussion et évaluation de la pertinence de la poursuite de la recherche dans cette direction.

[43] Source : http://www.cvm.qc.ca/cantic/beta/2nouve/5Protot1.htm

Objectifs du processus et outils associés

- Construire un système opérationnel de veille.

- Élaborer un guide de recherche pour les membres.

- Stimuler la contribution de l'information par le formulaire de signalement

- Construire une BASE DE DONNÉES alimentée via les contributions par formulaire Le classement de l'information respecte une taxonomie consensuelle.

- Expérimenter des ENGINS DE RECHERCHE sur la base

- Faire l'acquisition d'un LANGAGE DE QUESTIONNEMENT de la base.

- Le traitement de l'information est orienté en fonction de l'usager par une interface tableau de bord.

L'interface tableau de bord devra posséder des capacités selon 3 niveaux :

1. Capture des SIGNAUX FORTS
2. Capture des SIGNAUX INTERMÉDIAIRES
3. Capture des SIGNAUX FAIBLES

En lui-même, le traitement de l'information vise à donner du sens à l'information reçue (« contribuée »).

À l'aide de filtres, le bruit doit être éliminé pour laisser place aux tendances les plus probables, classées selon que le groupe de vigie (ou son « V+ », valeur ajoutée) les considéreraient comme associées à :

- de l'information à la mode

- de l'information de tendances où un choix de positionnement ou de non-positionnement est requis

- de l'information prédictrice de tendance où un positionnement sera requis

Pour ce, il faut classer l'information selon des paramètres, une taxonomie entendue de tous, un tableau de bord que tous auront en image. Certains sujets peuvent transpercer différentes strates taxonomiques, exemple le sujet "multimédia" peut se retrouver à l'intérieur de plus d'un type de vigie.

La tentative de normalisation de l'information émanant de ce prototype de système de veille a été réalisée sous la forme de pages HTML et de formulaires repris dans le précédent schéma mais que nous vous proposons de découvrir en détail à l'adresse suivante : http://www.cvm.qc.ca/xcantic/vigie/5formulaire01.htm

7 - VERS DES PLATES-FORMES ET DES PORTAILS DE TELE-APPRENTISSAGE

Il existe actuellement plusieurs plates-formes de télé-formation ou de télé-apprentissage (WebCT, TopClass, VirtualU, LearningSpace, Ingénium, Docent, E-charlemagne, etc.).
Ces plates-formes pour la plupart, sont fondées sur une approche de type 'Didacticiel hypermédia sur l'Internet » auquel on ajoute des moyens de communication fournis ou référencés par la plate-forme. Cela implique notamment que certains modèles techno-pédagogiques (hypermédia de formation autonome, enseignement asynchrone en ligne…) sont privilégiés par rapport à d'autres (communauté de pratique, systèmes de support à la performance…).
Cela implique ainsi que chaque cours sur le web est conçu indépendamment des autres, mais selon un format semblable qui conditionne les approches pédagogiques qu'il est possible de réaliser dans le cadre du système.

Dans le cadre de la conception des systèmes de télé-apprentissage ou portails d'apprentissage (*learning portals*) comportant un ensemble d'outils logiciels, de documents numérisés et de services de communication de plus en plus diversifiée, et de l'évolution rapide des méthodes et des outils de formation, la réalisation de tels systèmes ne peut se faire de manière artisanale.
Cette évolution rapide marque un changement important de paradigme.
Petit à petit, les méthodes de « génie logiciel » s'imposent dans la conception des systèmes d'apprentissage sur Internet.
Les méthodes d'extraction, de formalisation et de traitement des connaissances plus communément appelées l' « ingénierie des connaissances » sont au cœur des processus de gestion des connaissances, et aussi, par voie de conséquence, au cœur de la conception des systèmes d'apprentissage comme le déclare Gilbert Paquette dans son dernier livre.[44]
Il définit l'ingénierie pédagogique comme *« une méthodologie soutenant l'analyse, la conception, la réalisation et la planification de l'utilisation des systèmes d'apprentissage, intégrant les concepts, les processus et les principes du design pédagogique, du génie logiciel et de l'ingénierie cognitive ».*
Autrement dit l'Ingénierie pédagogique est un méta système qui vise à développer d'autres systèmes : les systèmes d'apprentissage.
Actuellement le système Explor@[45] créé par le LICEF [46]permet de construire un centre virtuel de télé-apprentissage accessible du portail d'une institution de formation diffusant une banque de cours et d'évènements d'apprentissage. En ce sens, Explor@ peut-être qualifié d'éditeur de portails de télé-formation.
Son environnement configurable intègre pour chacun des projets de formation, chacun des acteurs impliqués (apprenants, formateurs, concepteurs, experts de contenu, gestionnaires ou autres) et chacun des espaces de travail définis pour ces acteurs (production, information, assistance, collaboration, autogestion ou autres). Explor@ propose des outils intégrés, tels que des outils pour le suivi des apprentissages, calendrier, agenda et autres, mais permet aussi de configurer tous les autres outils ou ressources exécutables à partir d'un PC.
Cette approches vise à constituer des bases de données, des référentiels (sur serveur) de systèmes d'apprentissage et d'outils informatiques.

Cela permet, entre autres : d'alléger les sites Web en cours (qui ne contiennent alors que le contenu) ; le partage des ressources entre les cours ; la réutilisation et la mise à jour centralisée des ressources informatisées.
Un nouvel effort de recherche-développement est d'ailleurs entrepris depuis l'été 2000 au LICEF pour accroître les fonctionnalités du système au delà de ses capacités actuelles.
Il y a fort à parier que la notion de veille pédagogique jouant déjà un grand rôle dans cette structure de recherche soit à l'ordre du jour des fonctionnalités de ce portail de télé-apprentissage.

CONCLUSION

Si parfois les TIC nous dépassent, pour réussir il ne faut pas aller à l'encontre de la richesse de transaction informationnelle offerte aujourd'hui par la relation humaine dans nos universités, mais au contraire de l'outiller pour en tirer toute sa valeur.
Repérer les compétences et les savoirs, repérer les fonctions et les tâches, faciliter les échanges distants et asynchrones, capitaliser sur les questions récurrentes constituent les atouts essentiels pour travailler ensemble à l'intégration des TIC.

---

[44] « L'ingénierie pédagogique – pour construire l'apprentissage en réseau », (Paquette 2002), publié aux presses de l'Université du Québec

[45] Explor@: Centre virtuel de formation créé par le centre de recherche LICEF de la Télé-université (www.licef.teluq.uquebec.ca)

[46] LICEF : Laboratoire en Informatique Cognitive et Environnements de Formation

A une culture de l'enseignement centrée sur l'enseignant s'oppose une culture de l'apprentissage centrée sur l'apprenant.
Il faut ainsi former les étudiants à chercher et trouver l'information pertinente, produire et publier en utilisant les nouvelles technologies. L'enseignant doit enseigner autrement en associant à la pédagogie traditionnelle des techniques multimédias interactives.
L' auto-formation doit être prise en compte pour associer à l'enseignement collectif (de type cours magistral) un enseignement individualisé. Le travail coopératif doit être renforcé, grâce aux réseaux informatiques, entre enseignants, entre enseignants/étudiants et entre étudiants.

J'entendais un jour à un élève que l'ont doit désormais apprendre à marcher dans un train en marche.
Un long bout de chemin reste à faire, et des TIC aux TIC, il nous faut passer à :
«Technologies pour l'Intelligence et la Connaissance ».

Rester que le quai ? c'est impensable. Les institutions éducatives le reconnaissent de plus en plus, et l'on ne peut que s'en réjouir. De plus, la vigilance que les professionnels de la formation doivent exercer ne doit pas ressembler à « une résistance passive mais à une coordination militante d'initiatives». Les partenariats entre les universités et les entreprises participent eux aussi à cette dynamique en marche.

La veille pédagogique est l'une des principales clés de voûte pour accompagner ce changement progressif qui s'opère en ce début de siècle dans nos universités.

BIBLIOGRAPHIE

Education et nouvelles technologies - Pour une intégration réussie dans l'enseignement et l'apprentissage (Rapport annuel 1999-2000 sur l'état et les besoins de l'éducation au Québec) – Internet : www.cse.gouv.qc.ca.

Frédérico Mayor et Sema Tanguiane - L'enseignement supérieur au XXIème siécle – Editions Hermes France.

Multimédia et construction des savoirs – Presses Universitaires Franc-Comtoises- 2000 – ISBN 2-84627-000-7.

Le guide du Knowledge Management – Concepts et pratiques du management de la connaissance – Jean-Yves Prax - Dunod - ISBN 2-10-004701-9.

Benoit Godin, Yves Gingras - The place of universities in the system of knowledge production – Observatoire des Sciences et des Technologies , Cente Interuniversitaire de Recherche sur la Science et la Technologie, Université du Québec à Montréal (UQAM)

Les campus numériques – Enjeux et perspectives – Rapport de mission sous la direction de Michel Averous et Gilbert Touzot – Avril 2002.

Gilbert Paquette - Modélisation des connaissances et des compétences – Un langage graphique pour concevoir et apprendre - 2002, 388 pages, ISBN 2-7605-1163-4, D-1163.

Gilbert Paquette – L'ingénierie pédagogique pour construire l'apprentissage en réseau - 2002, 490 pages, ISBN 2-7605-1162-6, D-1162.

Michel Authier – Pays de connaissance – Editions du Rocher – ISBN 2-268-02969

REMERCIEMENTS

Je tiens à adresser mes vifs remerciements à :

- Madame Céline Saint-Pierre, Présidente du Conseil Supérieur de l'Education au Québec pour son accueil chaleureux lors de mon passage à Montréal et ses éclaircissements sur le rapport annuel 1999-2000 portant sur l'état et les besoins de l'éducation au Québec. Internet : www.cse.gouv.qc.ca

- Monsieur Michel Léonard, Andragogue et Professionnel de recherche au Centre de recherche LICEF, Télé-université du Québec à Montréal pour ses explications et la collaboration qu'il a accepté de mettre en place avec moi dans le cadre de l'adaptation de la veille pédagogique aux outils développés (MOT, MISA, ADISA et Explor@ développés au centre de recherche LICEF de la Télé-université du Québec.
Internet : www.licef.teluq.uquebec.ca/fr/index.htm

- Madame Suzanne Lapointe, aujourd'hui chargée de projets de l'Université du Québec à Montréal (UQAM ESG) pour sa disponibilité et la pertinence des informations communiquées et analysées lors de sont travail dans le domaine de la veille pédagogique sur le portail CANTIC.

- Madame Dominique Chassé, coordination des services TIC à l'École Polytechnique, et monsieur Daniel Oliva, agent de développement pédagogique à l'École de Technologie pour leurs expériences respectives menées dans le cadre du premier plan institutionnel triennal d'intégration des TIC dans l'enseignement (1999-2002)

- Mme Brigitte Gemme, enseignant-chercheur au CIRST (Centre Inter-universitaire de Recherche sur la Science et la Technologie) pour m'avoir permis la rencontre avec Benoît Godin et Yves Gingras tout deux de l'Observatoire des Sciences et des Technologies de Montréal , et ayant travaillé et écrit de nombreux articles sur la place des universités dans le système de production de connaissances et de savoirs.

- Enfin Richard Prégent, Conseiller pédagogique, Directeur du Bureau d'appui pédagogique et consultant de l'École Polytechnique de Montréal pour sa vision éclairée de l'analyse des systèmes d'apprentissage en ligne.

INTEGRER LA CONSULTATION ET LE PARAMETRAGE D'UNE ANALYSE SEMANTIQUE DE DONNEES TEXTUELLES POUR EN FACILITER L'APPROPRIATION

David Roussel
EADS Centre Commun de Recherches
12, rue Pasteur - BP 76 92152 Suresnes Cedex
david.roussel@eads.net

Résumé : Nous présentons une approche reposant sur l'intégration d'un éditeur d'ontologie et d'un moteur d'indexation et de recherche sémantique. Cette méthode en cours d'expérimentation est destinée à guider la construction d'une ontologie, vue à la fois comme une interface de paramétrage et une interface de consultation d'une analyse sémantique facilitant l'appropriation de résultats intermédiaires par des analystes. Un prototype a été développé pour valider la méthode et l'affiner. Il permet 1) de consulter des passages de textes à partir d'une ontologie qui explicite les relations entre certains concepts, acteurs ou projets identifiés 2) de rechercher des passages exprimant une combinaison de concepts et de naviguer depuis ces passages dans les documents d'origine 3) de mettre à jour l'ontologie et reparamétrer la stratégie d'analyse à appliquer sur les données textuelles.

Mots-clés : lecture rapide, moteur d'indexation et de recherche, analyse sémantique, conception d'ontologies orientée par une tâche, synthèse d'information.

Abstract : We investigate the integration of an ontology editor and a semantic indexing and search engine. This approach aims at facilitating the design of an ontology which is seen both as an interface to set parameters, and an interface to consult the results of a semantic analysis still in progress. A prototype has been developed to finalize the process. It enables 1) consulting some texts extracts from an ontology which specifies the relations between some concepts 2) searching the corpus extracts related to specific concepts and navigating from one sentence to a text unit in the original document 3) the upgrading of the ontology of concepts in order to reapply the semantic analysis upon the corpus

Keywords : quick reading, indexing and search engine, semantic analysis, task-oriented ontology design, information synthesis

Intégrer la consultation et le paramétrage d'une analyse sémantique de données textuelles pour en faciliter l'appropriation

INTRODUCTION

L'analyse de l'information par des techniques de fouille de textes est une évolution des services de veille au sein d'EADS destinée à appuyer l'offre d'expertise. Différentes techniques sont en cours d'expérimentation, dont les techniques d'analyse sémantique de différentes sources d'information.
Dans ce cadre, nous présentons un ensemble de problèmes d'appropriation de résultats de fouille de textes pour la planification d'un portefeuille de projets de recherche internes. L'expérience s'est déroulée sur trois niveaux imbriqués impliquant l'aide au positionnement de projets de recherche vis-à-vis de différents axes internes et vis-à-vis de l'extérieur, la présentation et l'évaluation d'un portefeuille de projets, le suivi et recalage d'un portefeuille de projets. Le premier niveau, qui nous intéresse directement ici, se décline en un ensemble de questions auxquels des données de veille apportent *a priori* des éléments de réponse. C'est le cas notamment des deux questions suivantes : quelle est la cartographie des collaborations en rapport avec chaque objectif ?, quels sont les facteurs qui influencent les stratégies de collaboration ?

Après un exposé rapide des techniques et méthodes que nous avons expérimentées pour chacune de ces questions, nous revenons sur deux difficultés: la nécessité de croiser des données textuelles dont le vocabulaire n'est pas immédiatement comparable et la nécessité de mettre en œuvre des hypothèses fortes nécessitant des outils très souples vis-à-vis des méthodologies et des ressources disponibles.

Pour exploiter les résultats obtenus, nous exposons section 2 notre approche, qui consiste à coupler une interface de visualisation d'ontologies et un moteur sémantique d'indexation et de recherche afin de :
- consulter rapidement des propositions linguistiques clés à partir d'une ontologie qui explicite des relations entre concepts, acteurs ou projets. Un retour aux passages correspondant dans les documents d'origine est possible le cas échéant.
- rechercher des propositions linguistiques spécifiques exprimant une combinaison de concepts. Cette recherche se fait au moyen d'un moteur de recherche qui affiche d'abord le « contenu conceptuel » des propositions qui coïncident aux critères de recherche. Ceci permet d'identifier rapidement des propositions linguistiques pertinentes et naviguer depuis ces propositions dans les documents d'origine.
- **finalement, mettre à jour l'ontologie et reparamétrer la stratégie d'analyse à appliquer sur les documents.**

Ainsi les décalages occasionnés par le paramétrage des techniques sont plus simples à relativiser et maîtriser, et il est également plus simple de prendre en compte des hypothèses d'analyse précises formulées par des experts. Une exemple de réanalyse de différents types de données textuelles autour d'un projet de réduction du bruit nous servira d'exemple (section 3).

1 - PREMIERE EXPERIMENTATION

Avant d'expliquer l'approche que nous expérimentons actuellement, il est utile de présenter les résultats d'une première expérimentation. Celle-ci a consisté à cartographier des collaborations existantes et à tenter d'extraire certains facteurs structurant des stratégies de collaboration possibles.

Cartographie des collaborations

Pour établir une cartographie des collaborations, nous avons classiquement utilisé[47] :
- des articles et notices d'articles scientifiques
- des descriptions de brevets

De leur côté, des documents internes à disposition ont permis de mieux cibler les informations à retenir et à synthétiser. Il s'agit en particulier de planches de présentation de projets internes qui présentent l'avantage de concentrer les mots clés en relation avec un projet.

[47] Ce sont notamment les deux sources d'information utilisées dans le *Deuxième Rapport européen sur les Indicateurs en science et technologie* ([1]).

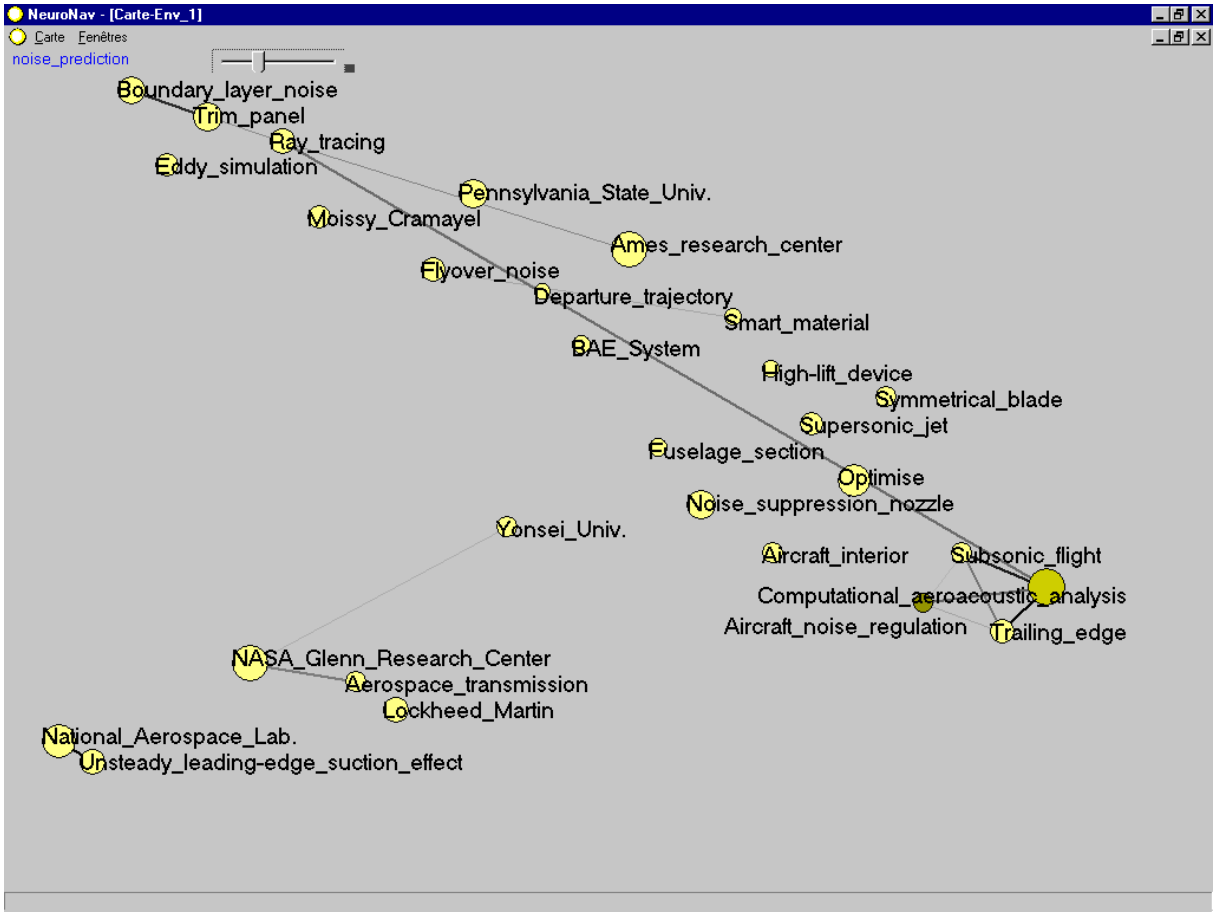
De nombreux outils permettent d'établir une cartographie où la proximité spatiale entre les items de la carte reflète une coarticulation de ces items dans une collection de documents. En pratique, le nombre d'étape à appliquer est très important du fait des précautions techniques nécessaires, du besoin de comprendre la pertinence des tendances à cartographier vis à vis des objectifs des projets de recherche et finalement des difficultés rencontrées pour exploiter les techniques elles-mêmes.

La méthode que nous avons suivie est :

- Identification dans chaque famille de projets internes des références citées sur les acteurs externes.
- Collecte de notices d'article scientifique sur des acteurs identifiés vis-à-vis du thème général d'une famille de projet.
- Analyse cartographique des données. Cette analyse se fonde dans notre cas sur une indexation des noms propres et des groupes nominaux complexes figurant dans les notices d'article.
- Navigation dans les cartographies et identification des mots clés (ex pour le thème de la réduction du bruit des avions : *noise_computation, noise_source_localization, noise_prediction, ...*).
- Collecte d'articles et brevets à partir de ces mots clés (données payantes).
- Indexation contrôlée et analyse cartographique de ces données pour identifier les références les plus pertinentes vis-à-vis des objectifs techniques d'une famille de projets internes (ex : prédire le bruit d'un avion ; diminuer le bruit de 10 décibels par déplacement des moteurs sans augmenter la consommation ; améliorer l'aérodynamique par des corrections de commandes de vol électriques sans compliquer la gestion des commandes, …).
- Extraction des auteurs et co-auteurs des articles identifiés ainsi que les mots clés significatifs vis-à-vis des objectifs techniques de projets internes.

La figure 1 illustre un exemple de cartographie permettant d'analyser plus attentivement des relations de proximité entre un objectif adressé par un projet interne - *noise prediction* - des classes de mots clés traités dans une base de référence de 150 articles scientifiques et d'une vingtaine de brevets. L'importance des classes est indiquée par la dimension du disque qui les représente (le rayon d'un disque représente plus précisément l'inertie d'une classe). La proximité est exprimée sur la carte à deux dimensions par une proximité spatiale ou encore par un lien plus ou moins épais.

Figure 1 : cartographie détaillée sur la base d'une indexation contrôlée d'articles scientifiques et brevets



Cette carte a été calculée par le logiciel Neuronav™ (société Diatopie)[48]. Ce logiciel, issu des travaux de Lelu et *al.* [5] exploite différentes stratégies d'indexation pour classer automatiquement[49] des documents et cartographier le résultat. Il offre de nombreuses fonctionnalités pour l'analyse d'un corpus documentaire: 1) environnement de nettoyage, de modification et de fusion des mots clés de l'indexation, 2) accès lexical flou au vocabulaire indexé, 3) calcul des mots et documents proches de tout mot(s) ou document(s), 4) classification floue des documents et mots associés, représentée par une carte en deux dimensions montrant quelques termes saillants caractéristiques pour chaque classe. Le terme le plus central est utilisé comme une étiquette thématique par défaut, 5) navigation triangulaire entre mots clés, classes et unités documentaires supplée par une cartographie des classes de documents, 6) dérivation de plusieurs environnements de cartographie centrés sur un sous-ensemble de documents, et 7) publication des cartographies interactives pour une consultation sur un intranet.

Ainsi, la carte de la figure 1 montre le résultat de la projection du mot clé « noise prediction » sur un espace de thèmes extraits automatiquement dans notre base de référence. Les thèmes « computational aeroacoutic noise » et « aircraft noise regulation » sont directement concernés par l'objectif « noise prediction » (ce qui se traduit par une couleur plus foncée). Le thème « computational aeroacoustic analysis » est proche des thèmes « subsonic flight » «trailing edge » et « aircraft noise regulation ». En complément, nous pouvons observer sur la carte une relation entre les thèmes « Ray tracing » et « computational aeroacoutic analysis » (matérialisé par la ligne entre les disques représentant respectivement ces thèmes). Dans ce dernier cas, il s'agit certainement d'une relation entre les moyens de calcul, ce que nous pouvons vérifier en comparant les mots clés associés aux deux thèmes.
En prêtant attention aux mots clés associés aux thèmes mentionnés, ou directement à la projection du mot clé « noise prediction » sur l'espace des mots clés, nous identifions les compétences d'organismes comme : NASA_Langley_Research_Center, S._Army_aeroflightdynamics_directorate, Georgia Inst. of Technology, DLR_Inst_fuer_Entwurfsaerodynamik.

L'opération de projection d'un objectif interne peut être répétée, éventuellement sur des cartographies comportant moins de thèmes ou des cartographies d'un sous-ensemble de documents jugés pertinents (par un mécanisme de zoom) afin d'adapter le degré de focalisation. Sur notre base d'articles et de brevets de référence, nous pouvons ainsi profiter des calculs de proximité pour identifier et compiler (manuellement) une synthèse des relations entre certains objectifs techniques et des acteurs du domaine cités dans les documents, eux-mêmes impliqués plus généralement sur certains thèmes, communs à d'autres acteurs du domaine. La nature de ces relations reste à ce stade à déterminer, au besoin par la lecture de certains documents.

Facteurs d'influence des stratégies de collaboration

Répondre à cette question suppose de mettre en évidence les dépendances et les convergences d'intérêt entre différentes organisations ou branches d'organisation.

Les sources d'information mobilisées pour la cartographie des collaborations (essentiellement des articles et brevets) illustrent parfois des coopérations à un niveau scientifique et technique mais pas à un niveau stratégique industriel. Par exemple, nous observons parmi les mots clés associés aux acteurs du domaine aéronautique quelques indications sur les modes de coopération qui distinguent certains compétiteurs. Les mots clés *cadre ISTC*, *cadre COS*, *configuration trilatérale*, *accord de coopération* vont caractériser l'un des compétiteurs, tandis qu'un autre est caractérisé par les mots clés *alliance*, *collaboration scientifique*, *université*, *échange de données*, *coût de développement*, *prototypage virtuel*.
Les mots clés ainsi détectés sont insuffisant pour préciser les facteurs structurant la stratégie de collaboration. Deux compléments nous semblent nécessaire. Le premier est d'utiliser encore d'autres types de donnée, le second consiste à modéliser les concepts qui interviennent dans les stratégies de collaboration afin d'extraire automatiquement des informations précises, et de faciliter les recoupements. Cette approche peut sembler adressée par les analyseurs sémantiques. Plusieurs problèmes d'appropriation des résultats restent cependant à considérer.
L'expérience d'analyse sémantique suivante a permis de réexaminer les problèmes:
- Collecte de discours des dirigeants de sociétés aéronautiques.

[48] Une version de démonstration de l'outil est disponible en ligne. http://www.diatopie.com/install/ (le 28.09.2002)

[49] Ces classes sont définies par :

- la liste des mots les plus caractéristiques, les plus "centraux ", des documents de la classe ;

- la liste des documents qui se "projettent" le plus haut sur l'axe de la classe, que ces documents appartiennent exclusivement ou pas à cette classe.

- Filtrage de propositions linguistiques a priori pertinentes. Ceci suppose de constituer un thesaurus autour des concepts clés des facteurs qui impactent les stratégies de collaborations (attitudes de partenariat, moyen et source de leadership, perception de la performance, ...). Dans les outils d'analyse sémantique testés, la méthodologie de constitution de ce thesaurus reste à l'appréciation de l'administrateur et n'est pas outillée.
- En complément, collecte d'informations de presse (presse spécialisée de l'agence Reuters et communiqués de presse) sur une organisation donnée.
- Analyse générale de ces informations par un moteur d'analyse sémantique disposant d'une ressource linguistique générale.

L'analyse sémantique a été réalisée au moyen du logiciel Tropes™ V5 (Acetic). Ce logiciel, issu des travaux de Ghiglione et Blanchet ([3]) sur l'analyse de contenu, permet notamment de 1) découper les textes en propositions linguistiques[50], 2) extraire ces propositions en catégories homogènes suivant leur contenu thématique[51], 3) exporter le résultat de l'analyse sous une forme XML permettant notamment de générer des liens hypertextes entre une proposition linguistique et son document d'origine, et 4) comptabiliser les fréquences d'apparition et l'interdépendance des concepts.

L'approche et les résultats obtenus sur les discours des dirigeants de sociétés concurrentes sont décrits dans [7]. Un repérage des termes associés à une soixantaine de concepts (issus de [2]) est appliqué pour faire émerger les facteurs de collaboration possibles. Le moyen retenu pour extraire dynamiquement des informations précises suppose que l'analyseur sémantique annote préalablement dans les documents les acteurs, concepts et projets. A partir de cette détection , nous calculons un réseau hypertexte offrant une navigation directe entre les propositions illustrant un concept et les passages des documents en relation avec un concept ou une proposition.
L'analyse des discours des dirigeants des sociétés délivre en particulier des indicateurs prospectifs, par exemple sur le principal attribut de partenariat considéré comme un vecteur de performance, sur la communication autour des partenariats, sur la stratégie de création de valeur, etc. A partir des extraits pertinents identifiés, une première analyse des stratégies de collaboration est décrite dans [7]. Pour notre principale concurrent, la principale source de pouvoir annoncée est la maîtrise des technologies mobiles d'échange d'information à partir de réseaux de satellite.

De son coté, une analyse sémantique générale d'informations de presse permet de repérer des documents ou des ensembles de propositions clés sur les investissements, les nouveaux tests en vol, contrats, accords, projets, partenariats de recherche, et enfin sur les différentes branches d'activité qui sont impliquées dans ou impactées par l'aboutissement d'une innovation technologique donnée. Ces informations présentent l'intérêt d'être actualisées mais le croisement entre ces informations reste une opération intellectuelle laborieuse.

Au final , dans l'expérimentation menée, nous aboutissons principalement à deux résultats intermédiaires pour tenter de situer des réseaux de collaboration vis-à-vis de facteurs stratégiques. Ces résultats intermédiaires sont des résultats comparatifs élaborés semi automatiquement à partir du second résultat intermédiaire : des données textuelles restructurées pour en faciliter la consultation.
Le premier type de résultat obtenus compare des noms propres ou des mots clés représentatifs pour chaque aspect d'un projet interne et renvoie, pour chacun des descripteurs, à des extraits a priori pertinents. Ces comparaisons lexicales sont des résultats bruts qui suggèrent des orientations mais restent à réinterpréter précisément par un ou plusieurs experts à partir d'une navigation dans les données textuelles restructurées.
Ces données restructurées constituent le second type de résultat. Il s'agit essentiellement:

- de représentations graphiques dynamiques des thèmes traités dans les différents ensembles ou sous-ensembles d'information (utilisation des techniques de cartographie par K-Means axiales. *Cf.* [6]). Ces cartographies permettent de naviguer interactivement dans les documents, les thèmes extraits (avec plus ou moins de granularité) et les mots clés associés.
- d'extractions de propositions linguistiques en relation avec une liste de concepts fréquents. Un retour immédiat au passage concerné dans le document d'origine est possible pour chacune de ces propositions. Ce système de navigation préparé pour un expert lui permettra de vérifier le sens des mots clés en contexte

A ce stade, nous rencontrons deux types de difficultés dans l'appropriation des résultats :

- la nécessité de croiser des données dont la terminologie n'est pas a priori comparable. Il existe une différence de langue, niveau de description ou d'arrière plan entre les données textuelles (par exemple entre articles et brevets). Dans [7], nous avons exposé les résultats d'une expérience destinée à faciliter la mise en

[50] Pour Ghiglione et al. [4], la proposition est la forme minimale la plus satisfaisante pour rendre compte d'un micro-univers, en tant « qu'assemblage de mots se rapportant directement ou indirectement à un verbe, base de l'ensemble » qui présente une signification complète et autonome.
[51] Chaque classe est définie par un ensemble de mots fréquents ou de lexicalisations de concepts prédéfinis. Des thesaurus génériques volumineux sont pour cela utilisés.

relation de données textuelles. Une segmentation des documents optimise l'usage des calculs de similarité entre données textuelles, de même que prendre en compte des documents artificiels composés des propositions linguistiques qui illustrent dans un corpus une relation potentielle entre deux concepts. Toutefois, les similarités trouvées entre données textuelles ne concernent essentiellement que des segments d'un même document. Nous observons de même peu de relation de proximité calculées entre brevets et articles. Une « extension » manuelle des réseaux de collaboration scientifiques aux réseaux de co-inventeurs est nécessaire. La prise en compte de concepts est un autre moyen d'améliorer la mise en relation de document qui ne partagent pas le même vocabulaire. Toutefois, il reste très difficile de décider a priori les termes qui lexicalisent les concepts pertinents. Plusieurs boucles sont nécessaires. Il est également difficile de décrire le vocabulaire des différents types de document équitablement, avec un niveau de détail ou généralisation suffisant. Dans le cas contraire, les croisements entre données focalisées sur des points techniques et informations sur les transferts de technologie sont faussés. Un exemple simple concerne l'analyse des différentes dénominations ou divisions d'un même organisme. Dans la figure 1 Ames_research_center et Nasa_Glenn_Research_Center désignent deux département de la NASA. Selon les cas, il est préférable de regrouper ces deux dénominations sous un même concept ou non.

- La seconde difficulté provient de la nécessité de mettre en œuvre des hypothèses fortes à l'aide d'outils très souples. Les premiers résultats obtenus par une technique de synthèse d'information sont surtout un point de départ pour interpréter l'information en appliquant un point de vu formulé par un ou plusieurs analystes. A ce niveau, les divergences de points de vu sont amplifiées par les différences d'appropriation des résultats intermédiaires. Il s'agit donc de proposer une interface qui puisse expliciter (toujours de façon synthétique) quelles ressources linguistiques ont été appliquées, afin de consulter les résultats d'analyse en conséquence, à la façon dont on se sert dynamiquement d'un moteur de recherche par reformulation successive. Plus précisément, l'interface de consultation devrait premièrement proposer une vue sur les différentes ressources linguistiques intégrées, quelque soit la façon dont ces ressources ont été créées au fur et à mesure des besoins (extraction automatique de classes de termes, intégration de thesaurus existant, besoin d'optimiser le comportement d'un algorithme, redéfinition de nouveaux termes pertinents mis en évidence par des modules de traitement,...). Cette vue sur les ressources linguistiques est le moyen d'introduire des points de contrôle qui auront une valeur ajoutée vis-à-vis de la tâche. Dans notre cas, il s'agit principalement d'un point de contrôle de l'adéquation des documents considérées vis-à-vis des stratégies que l'on cherche à connaître. Idéalement, si aucune stratégie ne peut être mise à jour, il faut pouvoir conclure que ces stratégies restent à définir ou pouvoir identifier les termes à renseigner pour la technique d'analyse.

2 - VERS UNE INTERFACE UNIFIEE DE PARAMETRAGE ET DE CONSULTATION D'UNE ANALYSE SEMANTIQUE

Nous avons choisi de faciliter l'appropriation d'une analyse sémantique en utilisant une représentation de haut niveau de type ontologie. De nombreux travaux ont décrit la création d'interfaces de navigation dans l'information à partir d'ontologie pour se familiariser avec un domaine, ou mettre à jour l'ontologie elle-même.
Notre approche consiste à construire une ontologie dans un triple but :
- servir de support/parcours de lecture des informations,
- consulter des informations pertinentes,
- reparamétrer la stratégie d'indexation dynamiquement.

L'approche que nous expérimentons pour faciliter l'appropriation d'une analyse sémantique repose d'une part sur une interface de type moteur de recherche. Ce moteur permet de consulter directement des propositions linguistiques où certains concepts évoqués ou non au sein de ces propositions (mise en relation directe) ou au sein d'un paragraphe (mise en relation indirecte). Un exemple est donné figure 4. Le moteur restitue le contenu conceptuel des propositions linguistiques pour permettre une lecture rapide d'informations pertinentes. Ces informations sont identifiées comme telles parce qu'elles expriment des contenus conceptuels particuliers.
D'autre part, notre approche repose sur une interface de visualisation et de mise à jour d'une ontologie ou sont exprimées (au fur et à mesure) les relations entre des concepts, acteurs et projets identifiés. Cette interface est intégrée au moteur de recherche afin 1) de naviguer directement dans des propositions linguistiques par un simple clic sur un concept, 2) de contrôler l'adéquation entre l'information affichée et la description d'un concept dans l'ontologie et 3) et servir d'interface de reparamétrage de l'indexation utilisée par le moteur de recherche.

Cette double interface permet de :
- Naviguer dans des phrases clés à partir de la représentation d'une ontologie en cliquant méthodiquement sur certains concepts en relation.
- Rechercher rapidement des informations pertinentes en relation avec une combinaison de concepts.

- Mettre à jour facilement le mécanisme de synthèse utilisé en ajustant l'ontologie lorsque sa capacité explicative n'est plus valide[52].

Le schéma suivant illustre l'approche expérimentée.

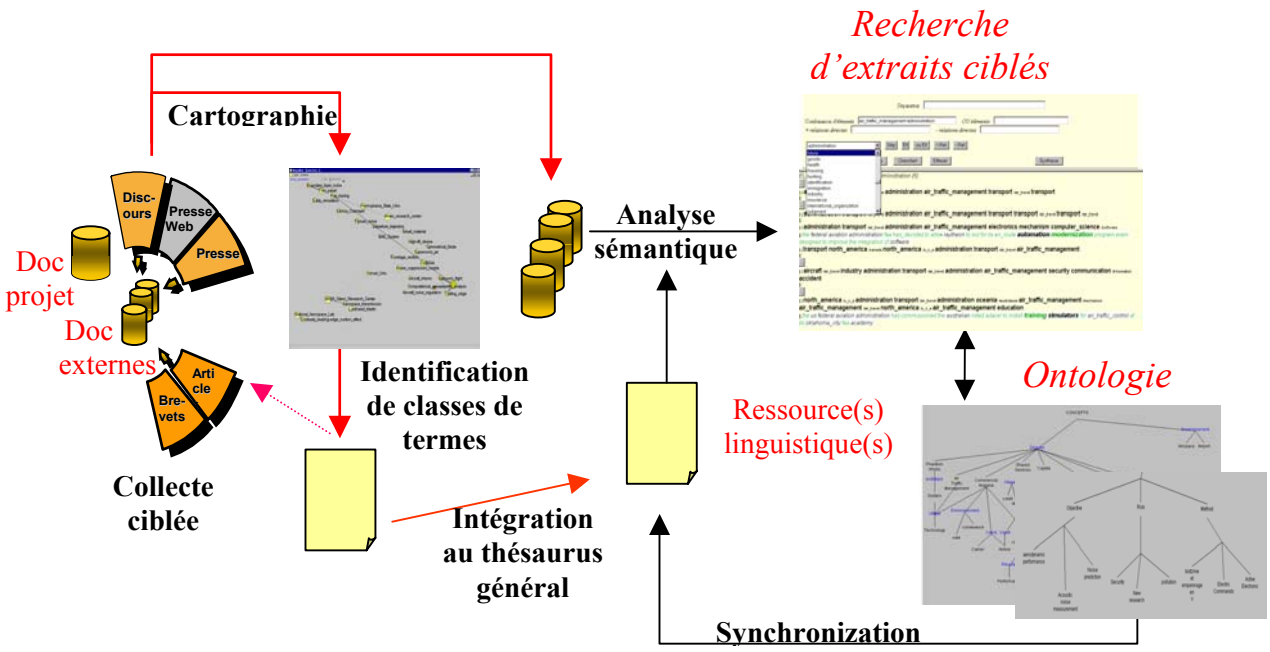


Figure 2 : processus expérimenté

Au minimum, l'objectif de l'interface de définition d'ontologie est de servir d'interface pour des experts qui auraient à interpréter les synthèses d'information produites. Pour cela l'interface de consultation restitue l'information sous forme de chaînes de concepts puis restitue, à l'initiative de l'utilisateur, des propositions linguistiques (voir Figure 3) qui peuvent être lues le cas échéant dans le contexte du document d'origine. Cette gradation facilite, par une lecture rapide des distributions de concepts, le repérage de propositions linguistiques pertinentes. Autre support d'appropriation : un texte libre peut être éventuellement associée à un nœud de l'ontologie pour expliciter la compréhension courante d'un concept ou d'une relation entre deux concepts.
Il est également possible par un simple clic de colorier des concepts ou surligner des concepts, de masquer des propositions, ou encore de collecter des propositions dans une même page. Ces actions graphiques sur le texte simulent l'action de surligner, souligner et barrer des items d'une liste afin de favoriser une lecture sur écran qui profite des possibilités de navigation.

Exemple d'utilisation

Dans cet exemple, le premier rôle de l'éditeur d'ontologie est de définir les concepts qui seront ciblés par l'analyse. Le moteur de recherche permet ensuite de consulter dynamiquement le résultat d'une analyse qui a préalablement indexé les propositions linguistiques comportant des concepts ciblés.
Une visualisation des cooccurrences de chaque concept permet éventuellement de redéfinir l'ontologie à partir des concepts les plus centraux.

[52] Pour des contraintes techniques, l'éditeur d'ontologie utilisé est un développement spécifique compatible avec le système *Protégé-2000* (voir http://protege.stanford.edu/). Plusieurs outils sont dédiés à la création d'ontologies, par exemple OntoEdit, OntoTerm, OilEd, WebOnto, Ontosaurus ou encore Ontolingua. Cependant, ces logiciels utilisent un formalisme spécifiques et ne sont pas couplés à un système d'indexation et de recherche.

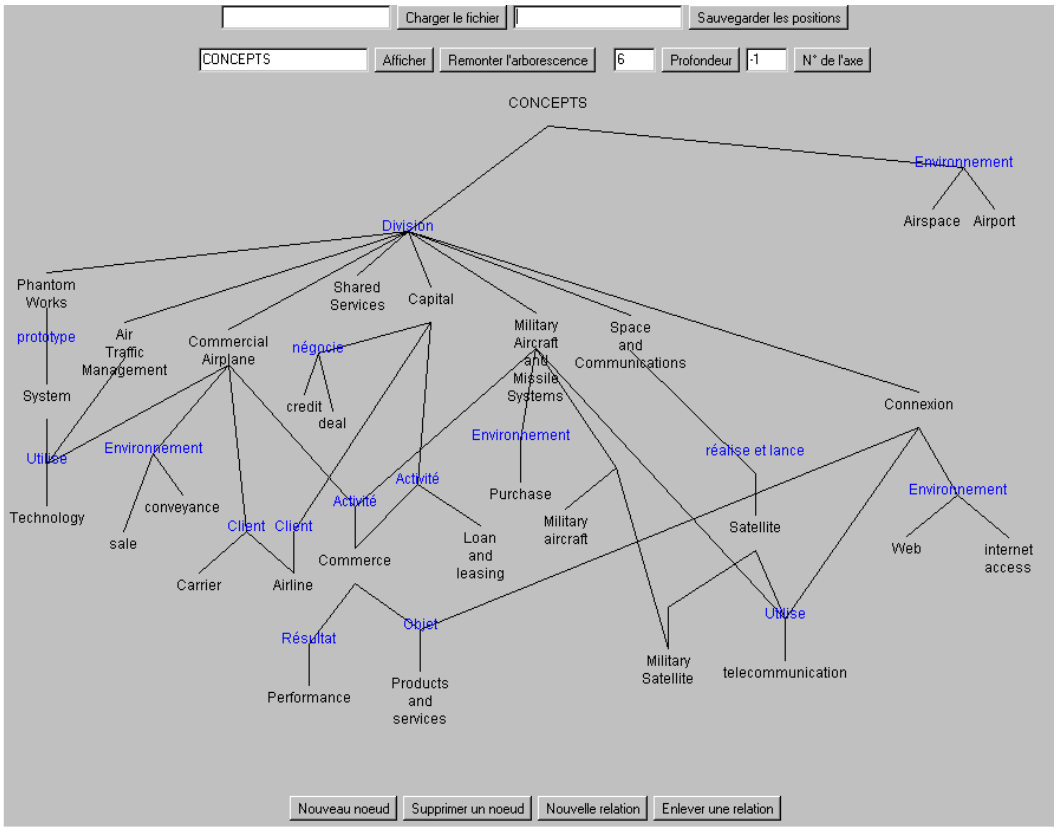


Figure 3 : Interface de visualisation de l'ontologie. Le concept à partir duquel l'ontologie est affiché et le niveau de profondeur sont paramétrables pour conserver une visibilité.

L'ontologie va nous permettre de noter au fur et à mesure des constats afin de réanalyser les données de façon optimale. Pour le projet qui nous concerne (diminution du bruit des avions), l'analyse cartographique des mots clés dans les articles et brevets met en évidence le fait que les recherches menées répondent aux mesures de régulation du bruit (voir figure 1 « noise regulation »). Cette information sera ajoutée dans l'ontologie visualisée figure 3.

Une lecture rapide des extraits concernés et des noms d'acteurs détectés révèle que les recherches répondent plus précisément au besoin de prédire l'effet de nouvelles technologie et de se doter de moyens de mesure pour obtenir une certification, en synergie avec la branche d'activité *gestion du trafic aérien*.

Dans un second temps, la recherche d'extraits sur l'actualité de la branche *gestion du trafic aérien* montre plusieurs cooccurrences avec le concept *administration*. La recherche d'extraits concernant à la fois le concept *administration* et la classe conceptuelle *gestion du trafic aérien* montre les dépendances directes entre les programmes de modernisation et les problèmes de gestion du trafic aérien (voir figure 4).
Plus précisément, la modernisation des systèmes en place de diminution du bruit est étroitement associée aux procédures de gestion du trafic aérien (répartition des vols, optimisation des trajectoires des avions, diminution du temps d'approche, atterrissage et décollage près des aéroports, ...). Ce constat est ajouté dans l'ontologie.

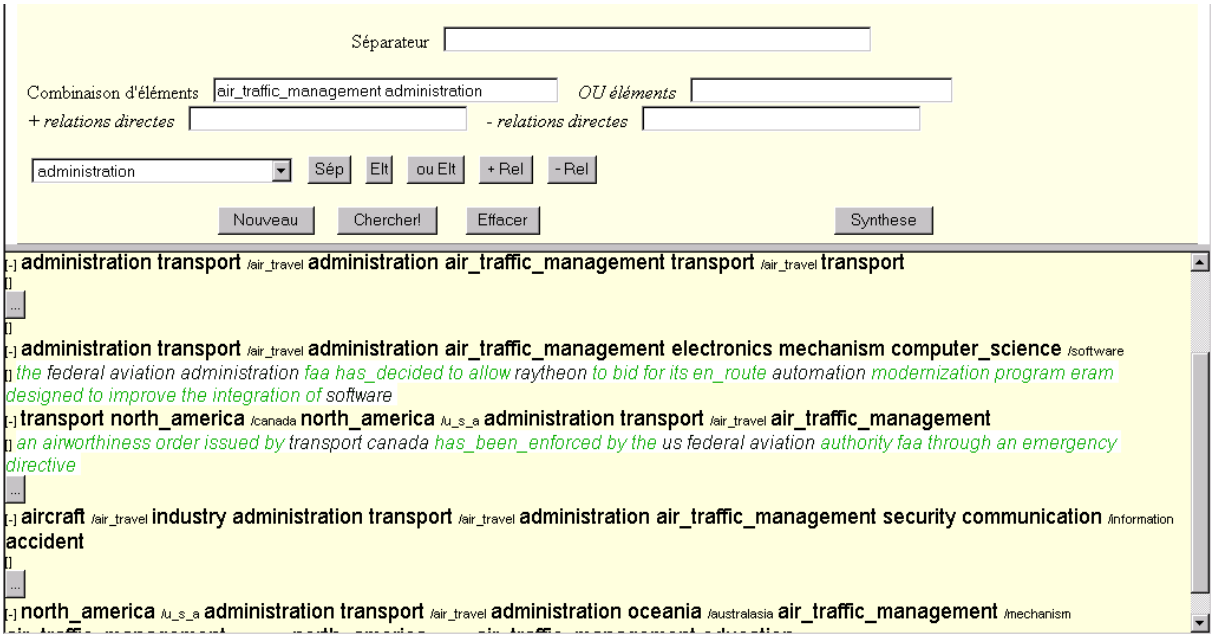


Figure 4 : Interface de recherche et consultation d'information en relation avec une combinaison de concepts. Les concepts pères et fils de l'ontologie qui sont identifiés dans les propositions linguistiques concernées sont d'abord affichés, puis certaines propositions. Un retour au document d'origine est également possible

En améliorant le niveau de détail de l'ontologie (et du même coup de l'analyse), nous pouvons nous concentrer sur le rôle d'un acteur particulier et tenter de réanalyser les facteurs qui influencent la stratégie de collaboration d'un ensemble d'acteur partenaires sur un thème de recherche.
Ainsi, nous pouvons tenter de croiser deux résultats : le fait que les recherches sur la réduction du bruit dépendent de la gestion du trafic aérien et, par exemple, le fait que l'activité d'échange d'information mobile via un réseau de satellites est jugée prioritaire par notre principal concurrent (résultat de l'analyse des discours de dirigeants. *Cf.* [7]). Ceci n'est pas sans conséquence sur le financement des recherches et la stratégie de collaboration, en interne ou avec l'extérieur, pour initier ou finaliser le développement des technologies.
Le moteur de recherche est pour cela utilisé afin de repérer des interactions pertinentes dans les propositions concernant l'application des procédures de réduction du bruit et la branche d'activité spatiale.
Pour ne pas globaliser l'impact des nouveaux dispositifs de gestion du trafic aérien financé par la branche d'activité espace et communication, les termes qui impliquent une relation entre classe conceptuelle *Space and communication* et la gestion du trafic aérien sont cette fois associés à un concept spécifique. C'est le cas par exemple de *space-based air traffic management systems* ou *satellite-based air traffic control systems, satellite-controlled navigation systems*, *telematic systems* ou encore *gps precision landing system.*
Sur un volumineux corpus de textes de presse en aéronautique représentatifs des deux dernières années (durée jugée nécessaire pour obtenir une vision globale), nous dégageons alors deux tendances :

- Une application réussie des procédures de réduction du bruit suppose une gestion des trajectoires aux abords des aéroports qui nécessite des technologies de système de contrôle aérien par satellite afin de transmettre des calculs précis au tableau de bord.
- Pour les aéroports internationaux à forte capacité, la synergie entre les recherches autour des procédures de réduction du bruit particulières (non uniformes) et les systèmes d'assistance à l'application des procédures de réduction du bruit est plus importante.

Partant de ces résultats, l'éditeur d'ontologie va jouer son second rôle à savoir faciliter la compréhension des informations par d'autres utilisateurs en définissant une première mise en relation des concepts fréquents dans ces informations. L'ontologie servira pour des analystes de parcours de lecture d'informations a priori pertinentes afin de les aider à situer des objectifs techniques de réduction du bruit des avions par rapport à des acteurs du domaine mais surtout par rapport à des différentes conditions opérationnelles de niveau de bruit elles-mêmes conditionnées par des technologies de navigation par satellite. Selon que les financements ou la couverture de ces différentes composantes sont assurés ou non, des collaborations originales resteront à monter.

CONCLUSION

Nous avons décrit un cas d'utilisation d'ontologies dans une application d'analyse sémantique d'informations ciblées.
sur un exemple simple, nous avons montré l'importance d'une interface intégrant la consultation et le reparamétrage d'une stratégie d'indexation de données textuelles. Cette intégration nous permet de croiser progressivement des données et d'apprécier certains regroupements d'acteurs selon différentes facettes que nous ne pouvons analyser par une stratégie d'analyse globale. En effet, le bruit en retour serait trop important. En utilisant une interface qui permet d'appliquer itérativement des stratégies d'analyse locales et dérouler certaines hypothèses, nous valorisons par ailleurs l'expertise. Les experts sont mobilisés pour valoriser itérativement des premiers résultats et pas pour valider des ressources linguistiques définies a priori.

L'approche suivie est complémentaire des outils de cartographie statistique. Naviguer dans des synthèses d'information à partir d'une cartographie fait ressortir les termes centraux à prendre en compte dans un ensemble de documents mais ne facilite pas en soi une activité réflexive. En faisant de l'ontologie le passage obligé vers une synthèse d'information, un parcours de lecture est proposé les biais éventuels introduits par une méthode d'analyse « orientée » et la compréhension des modifications à apporter localement à l'ontologie sont plus facile à répercuter.

Le prototype que nous avons testé est une première étape. Son ergonomie doit être revue. Nous n'avons pas non plus mentionné dans ce travail la prise en compte de différents points de vue. Il s'agit pour nous d'une prochaine étape qui permettra d'affiner et de valider éventuellement à plus grande échelle la méthodologie à appliquer.

REMERCIEMENTS

Nous tenons à remercier Jean Yves Fortier du Laria pour son travail sur le développement d'un éditeur d'ontologie à la fois simple et générique que nous avons pu réintégré dans ce projet.

BIBLIOGRAPHIE

[1] *Deuxième Rapport européen sur les Indicateurs en science et technologie 1997.* Publié par la Commission européenne en décembre 1997 (EUR 17639 ISBN 92-828-0271-X, 2 volumes, 729 pp; Annexes 198 pp)

[2] D. Fernandez-Bonet, *Conflit et coopération dans le canal de distribution. L'analyse du discours des acteurs comme révélateur des comportements*, Thèse de doctorat en science de gestion, Université Aix-Marseille II, juin 1999.

[3] R. Ghiglione, A. Blanchet, *Analyse de contenu et contenus d'analyse*, Paris, Dunod, 1991.

[4] R. Ghiglione, C. Kekenbosh C., Landré A., *L'analyse cognitivo-discursive*, P.U.G., 1995.

[5] A. Lelu , S. Aubin, Vers un environnement complet de synthèse statistique de contenus textuels : Neuronav version 2, séminaire ADEST, www.upmf-grenoble.fr/adest/seminaires <http://www.upmf-grenoble.fr/adest/seminaires> (le 28.03.2002).

[6] A. Lelu, A.G. Tisseau-Pirot, A. Adnani, Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation, *Hypertextes et Hypermédias*, vol.1, n°1, p. 23-55, Hermès, Paris, 1997

[7] D. Roussel, Navigation dans l'information par recomposition de documents et cartographie, CIFT02 : Colloque International sur la Fouille de textes, Hammamet, Tunisie, 20-23 octobre 2002.

DETECTION DE CONVERGENCE EN VUE DE L'OPTIMISATION D'UN SYSTEME DE FILTRAGE ADAPTATIF

TMAR Mohamed
tmar@irit.fr

TEBRI Hamid
tebri@irit.fr

BOUGHANEM Mohand
boughane@irit.fr

Adresse professionnelle
IRIT-SIG - Université Paul Sabatier de Toulouse
118, route de Narbonne, F-31062 Toulouse Cedex 4
Tel: (+33) (0)5.61.55.74.16

Résumé : Un système de filtrage adaptatif permet d'extraire, à partir d'une source dynamique de documents, les seuls documents pouvant intéresser un utilisateur ayant des centres d'intérêts relativement stables. Nous avons développé un système de filtrage adaptatif basé sur le principe de renforcement pour apprendre le profil, et la distribution de probabilités des scores des documents pertinents et non pertinents pour adapter la fonction de seuillage. Cet article décrit ce système, et les expérimentations effectuées pour mesurer l'efficacité de notre approche.

Mots-clés : filtrage adaptatif, Utilité, seuillage, convergence, distribution de probabilités.

Abstract : An adaptive filtering system allows to extract, from a stream of documents, the only documents being able to interest a user having relatively stable centers of interests. We developed a model of adaptive filtering system based on the principle of reinforcement to learn the profile. The threshold calibration is performed using the score probability distribution in two samples of documents: a sample of relevant documents and a sample of non-relevant documents. This paper describes our system, and the experimentations carried out to measure the effectiveness of our approach.

Keywords : adaptive filtering, Utility, threshold, convergence, probability distribution.

Détection de convergence en vue de l'optimisation d'un système de filtrage adaptatif

INTRODUCTION

L'avènement Web, contexte dans lequel la recherche d'information est une préoccupation centrale, a réactualisé la problématique de la recherche d'information, particulièrement dans la manière d'accéder aux informations. En effet, si avec un système de recherche d'information on accède volontairement à des informations via des requêtes ou par navigation, on assiste aujourd'hui de plus en plus à la prolifération de services qui ramènent des informations à l'utilisateur. Le processus qui permet de sélectionner l'information désirée dans ces flots d'informations s'appelle le filtrage d'information.

Un système de filtrage d'information permet à partir d'une source dynamique d'information (Internet, E-mail, News,…) de sélectionner et de présenter les seuls documents intéressant un utilisateur ayant un centre d'intérêt relativement stable appelé *profil.*

Le filtrage d'information est un processus dual à la recherche d'information [Belkin & al., 1992]. Ceci traduit qu'un processus de recherche d'information peut simuler un processus de filtrage d'information. Cependant, la plupart des systèmes de filtrage d'information sont basés sur des modèles de recherche d'information. Ainsi, les documents et les profils sont représentés par des listes de mots pondérés. L'appariement document-profil consiste à mesurer une similarité. La décision quant à l'acceptation ou le rejet d'un document est assurée par une fonction de décision souvent de type seuil. Si le score est supérieur au seuil le document est accepté sinon il est rejeté.

Or, en l'absence de base de référence, la détermination de ce seuil et les pondérations adéquates associées aux profils et aux documents sont les problèmes majeurs rencontrés dans ce domaine. En effet, dans un système de recherche d'information, les techniques de pondération et de reformulation automatique de requêtes, basées sur la collection de documents, se sont avérées efficaces, or, dans un système de filtrage d'information, à l'initialisation du processus de filtrage, on ne dispose d'aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision, ni pour bien pondérer les profils et les documents entrants. De plus, l'adaptation des profils aux différents flots pose un problème fondamental lié à l'incomplétude permanente des informations permettant de décrire exhaustivement les profils. Les solutions proposées aujourd'hui sont les suivantes : une solution synchrone ou adaptative, les différents facteurs du système de filtrage sont déduits à partir des documents filtrés cumulés dans le temps, et les solution asynchrone ou différée, les facteurs sont déduits à partir de collections de documents existantes.

La solution que nous proposons est synchrone, les profils et la fonction de décision sont appris d'une façon incrémentale pendant le filtrage. La plupart des systèmes de filtrage d'information actuels respectent très peu la notion d'incrémentalité, car :

- **ils utilisent souvent des bases d'entraînement [Zhai & al., 1998][Robertson & al., 1999]. Ces systèmes se basent sur les statistiques des termes et des documents dans ces bases pour estimer les valeurs de plusieurs paramètres de filtrage, or ces statistiques sont très variables au cours du filtrage et inconnues à l'initialisation du processus. De plus, les documents issus des bases d'entraînement peuvent avoir des caractéristiques des documents provenant de la source, les statistiques utilisées dans l'entraînement peuvent alors être peu crédibles,**
- **ils effectuent l'apprentissage du profil et le seuillage d'une manière quasi-différée [Kwok & al., 1999][Hoashi & al., 1999]. En effet, plusieurs effectuent l'apprentissage du profil et/ou le seuillage à chaque réception de *n* (10, 100, 1000) documents. Avec cette manière périodique de fonctionnement, le système perd beaucoup de qualités incrémentales et tend vers les systèmes de filtrage différé [Malone & al., 1987],**
- **plusieurs informations sont parfois supposées connues avant le démarrage du filtrage (proportion de documents pertinents, quelques documents pertinents d'entraînement, …) [Zhai & al., 1998], ceci n'est pas vrai dans le cas du filtrage, car le système démarre avec une information sauf le profil initial de l'utilisateur.**

Le modèle que nous proposons pour notre part est purement adaptatif et incrémental. Aucune information autre que le profil initial n'est connue au démarrage du processus de filtrage. Les statistiques des documents et des

profils sont actualisées au fur et à mesure que le système reçoit des documents. L'adaptation des profils et le seuillage se font d'une manière adaptative et incrémentale à chaque réception d'un document pertinent.

Nous proposons une méthode d'adaptation du profil basée sur le principe de renforcement. Chaque fois qu'un document est sélectionné et jugé pertinent pour un profil donné, le système doit adapter le profil de sorte à modifier sa représentation. De plus, nous proposons une méthode d'adaptation du seuil basée sur la distribution des scores d'un échantillon de documents pertinents et non pertinents sélectionnés. Dans la première section, nous présentons notre modèle de filtrage. La seconde section décrit notre approche d'adaptation du profil et du seuil. Enfin, nous présentons dans la dernière section les expérimentations et les résultats obtenus.

1 - MODELE DE BASE

Le modèle de filtrage que nous proposons est basé sur une approche vectorielle. Les documents et les profils sont représentés sous forme d'une liste de termes pondérés. Le processus de filtrage consiste à comparer le score résultant de la similarité document-profil à un seuil, si le score est supérieur au seuil le document est sélectionné sinon il est rejeté. Ce seuil, les profils et les caractéristiques liées à la pondération des termes des documents évoluent à chaque arrivée d'un document.

1.1. Représentation des profils

Un profil est un ensemble de termes sans les mots vides. Il est représenté sous une forme vectorielle :

$$p^{(0)} = ((tp_1, \omega_1^{(0)})...(tp_n, \omega_n^{(0)})) \qquad (1)$$

Avec $\omega_i^{(0)}$ le poids du *ième* terme tp_i dans le profil à l'instant $t = 0$. t est incrémenté à chaque arrivée d'un document et représente l'instant où le système reçoit un document. Initialement, le poids du terme dans le profil est calculé comme suit :

$$\omega_i^{(0)} = \frac{tfp_i}{\max_j(tfp_j)} \qquad (2)$$

Où tpf_i est la fréquence du terme tp_i (noté aussi t_i) dans le profil. La formule est à priori simple, car au début du processus de filtrage on ne dispose d'aucune information autre que le profil initial. Cependant, ce poids sera ajusté par apprentissage au fur et à mesure les documents arrivent.

1.2. Représentation des documents

A chaque arrivée d'un document, celui-ci est indexé par un module d'indexation de Mercure **[Boughanem, 2000]. Le résultat de cette opération est une liste de termes. Chaque terme du document est pondéré en utilisant la formule suivante :**

$$d_i^{(t)} = \frac{tf_i^{(t)}}{h_3 + h_4 . \frac{dl^{(t)}}{\Delta l^t} + tf_i^{(t)}} \log\left(\frac{N^{(t)}}{n_i^{(t)}} + 1\right) \qquad (3)$$

$d^{(t)}$: le document reçu à l'instant t ,
$tf_i^{(t)}$: fréquence d'apparition du *ième* terme dans le document $d^{(t)}$,
h_3, h_4 : paramètres constant,
$dl^{(t)}$: longueur ou nombre de termes du document $d^{(t)}$,
$\Delta l^{(t)}$: longueur moyenne d'un document,
$N^{(t)}$: nombre de documents déjà examinés,
$n_i^{(t)}$: nombre de documents contenant le terme tp_i parmi les documents déjà examinés.

Les paramètres $\Delta l^{(t)}, N^{(t)}$ et $n_i^{(t)}$ sont mis à jour à chaque arrivée d'un document.

1.3. Processus de filtrage de documents

Le processus de filtrage consiste à calculer un score, noté $rsv(d^{(t)}, p^{(t)})$ entre le document $d^{(t)}$ et le profil $p^{(t)}$. Ce score est défini par le produit scalaire entre le document et le profil :

$$rsv(d^{(t)}, p^{(t)}) = \sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} \qquad (4)$$

Le score calculé est ensuite comparé à un seuil de filtrage, pour décider si le document est accepté ou non : si $rsv(d^{(t)}, p^{(t)}) \geq seuil^{(t)}$ alors le document $d^{(t)}$ est sélectionné, sinon il est rejeté. Le seuil est appris à chaque arrivée d'un document pertinent. Le processus d'adaptation du seuil sera détaillé ci-dessous.

2 - ADAPTATION DU PROFIL ET DU SEUIL

2.1. Apprentissage du profil

Le processus d'apprentissage du profil est adaptatif et incrémental. Il est effectué à chaque fois un document est jugé comme étant pertinent par l'utilisateur. Il permet de modifier la représentation du profil de l'utilisateur, en ajustant les poids des termes, en ajoutant ou en éliminant les termes du profil. Nous décrivons dans les sections suivantes, les étapes nécessaires pour effectuer cet apprentissage.

2.1.1. Conception du processus d'apprentissage

L'idée de base de notre processus d'apprentissage est basée sur le principe de renforcement [Sutton & al., 1998]. Quand un document est jugé pertinent, il faut pouvoir trouver une nouvelle représentation du profil qui permet de retrouver le document avec un score fort. Autrement dit, on sera amené à améliorer le profil tel que $rsv(d^{(t)}, p^{(t)}) = \beta$ ou β est le score désiré. Le problème à résoudre revient alors à chercher les $\omega_j^{(t)}$ qui satisfont l'équation :

$$\sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} = \beta \qquad (5)$$

Toutefois cette équation admet une infinité de solutions, alors nous proposons d'ajouter une contrainte pour avoir une solution unique. L'intérêt dans tous ça, est de faire tendre le poids de chaque terme vers son poids idéal. Le poids idéal correspond au poids du terme qui permet de discriminer l'ensemble des documents pertinents et celui des non pertinents. Ainsi, si le poids idéal d'un terme t_i est donné par une fonction f donnée par la formule (9) , et si le poids dans le profil est $\omega_i^{(t)}$, alors $\frac{\omega_i^{(t)}}{f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})}$ est une constante, où $r_i^{(t)}$ (resp. $s_i^{(t)}$) représente le nombre de documents pertinents (resp. non pertinents) contenant le terme t_i. Le système à résoudre devient alors :

$$\begin{cases} \sum_{\substack{t_i \in d^{(t)} \\ tp_j \in p^{(t)} \\ t_i = tp_j}} d_i^{(t)} * \omega_j^{(t)} = \beta \\ \forall (t_i, t_j) \in d^{(t)^2}, \ \frac{\omega_i^{(t)}}{f(d_i, r_i^{(t)}, s_i^{(t)})} = \frac{\omega_j^{(t)}}{f(d_j, r_j^{(t)}, s_j^{(t)})} \end{cases} \qquad (6)$$

La solution du système (6) est l'ensemble des poids du profil qui permet de retrouver le document $d^{(t)}$. Elle correspond à des poids provisoires qui vont intervenir dans le calcul du poids global du profil.

Soit n le nombre de termes distincts dans le document à l'instant t, et $f_i^{(t)} = f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})$. Le système (6) peut être réécrit comme suit : $\forall i \in \{1 \ldots n\}$

$$\begin{cases} \frac{\omega_1^{(t)}}{f_1^t} = \frac{\omega_i^{(t)}}{f_i^t} \Leftrightarrow \omega_1^{(t)} d_{j1}^{(t)} = f_1^{(t)} d_{j1}^{(t)} \frac{\omega_i^{(t)}}{f_i^{(t)}} \\ \frac{\omega_2^{(t)}}{f_2^t} = \frac{\omega_i^{(t)}}{f_i^t} \Leftrightarrow \omega_2^{(t)} d_{j2}^{(t)} = f_2^{(t)} d_{j2}^{(t)} \frac{\omega_i^{(t)}}{f_i^{(t)}} \\ \vdots \\ \frac{\omega_n^{(t)}}{f_n^{(t)}} = \frac{\omega_i^{(t)}}{f_i^{(t)}} \Leftrightarrow \omega_n^{(t)} d_{jn}^{(t)} = f_n^{(t)} d_{jn}^{(t)} \frac{\omega_i^{(t)}}{f_i^{(t)}} \end{cases} \qquad (7)$$

Où j_k correspond à l'index dans le document du terme indexé par *k* dans le profil ($t_k = tp_{jk}$). En additionnant le premier opérande de chaque équation et après quelques transformations, on obtient pour chaque terme son poids provisoire $p\omega_i^{(t)}$ qui est donné par :

$$\forall i,\ p\omega_i^{(t)} = \frac{\beta . f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)})}{\sum_j f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}).d_j^{(t)}} \qquad (8)$$

Cette valeur permet d'affecter à chaque terme un poids idéal en tenant compte de son importance dans le document $d^{(t)}$, l'ensemble des documents pertinents et non pertinents ($r_i^{(t)}$ et $s_i^{(t)}$) et le score désiré du document (β).

Le choix de la fonction f dépend de plusieurs paramètres, la fréquence d'apparition du terme dans le document, le nombre de document pertinents et non pertinents contenant ce terme, le nombre total de documents pertinents sélectionnés, etc. Nous avons expérimenté certaines fonctions et avons opté sur une fonction dérivée de la formule de Robertson-Sparck Jones [Robertson & al., 1976] :

$$f(d_i^{(t)}, r_i^{(t)}, s_i^{(t)}) = d_i^{(t)} . \log(1 + \frac{r_i^{(t)}(S^{(t)} - s_i^{(t)})}{(s_i^{(t)} + 1)(R^{(t)} - r_i^{(t)} + 1)}) \qquad (9)$$

Où $R^{(t)}$ (resp. $S^{(t)}$) représente le nombre de documents pertinents (resp. non pertinents) sélectionnés par le système à l'instant t .

2.1.2. Modification du profil

L'adaptation du profil consiste à améliorer sa représentation à chaque fois un document pertinent est sélectionné par le système. Dans notre cas, l'adaptation consiste à utiliser les poids provisoires $p\omega_i^{(t)}$ pour contribuer à l'apprentissage des termes dans le profil. Nous utilisons la formule de distribution de gradient suivante :

$$\omega_i^{(t)} = \omega_i^{(t)} + \log(1 + p\omega_i^{(t)}) \qquad (10)$$

Pour valider ces différentes formules, des expérimentations ont été effectuées sur une base **Reuters** issue de TREC[1]-10. Cette technique d'adaptation permet d'améliorer la représentation du profil et de séparer les documents pertinents des documents non pertinents [Boughanem & al., 2002].

2.2. Adaptation du seuil

Dans un contexte incrémental, l'adaptation du profil entraîne automatiquement une variation à la hausse des documents des scores. Par conséquent, l'adaptation du seuil devient nécessaire pour arriver à sélectionner le maximum de documents pertinents et à rejeter le maximum de documents non pertinents.

Plusieurs techniques ont été envisagées et expérimentées pour le seuillage [Tmar, 2002]. L'approche que nous proposons dans cet article est basée sur la distribution des scores des documents. Nous supposons que la distribution des scores des documents suit une certaine loi de probabilité. En se basant sur cette loi, nous pouvons décider si un document est pertinent ou non, selon sa probabilité de pertinence.

2.2.1. Principe de l'approche de l'adaptation du seuil

L'approche que nous proposons consiste à estimer, pour un échantillon de documents, sa distribution de probabilités discrète. Grâce à une technique de régression linéaire, nous transformons cette distribution de probabilités discrète en une densité de probabilités continue. Ceci va nous permettre de choisir une valeur du seuil dans un intervalle continu de scores.

2.2.2. Modélisation des distributions des scores

L'appariement document-profil permet de fournir un score donné pour ce document. La probabilité qu'un document tiré aléatoirement ait un score donné est par définition

[1] Text REtrieval Conference : *un programme d'évaluation des systèmes de recherche et de filtrage d'information*

égale au nombre de documents ayant eu ce score divisé par le nombre de documents total :

$$p(X = score) = \frac{|\{d \mid rsv(d, p) = score\}|}{|\{d\}|} \qquad (11)$$

Comme les valeurs des scores sont très variées (voir figure. 1), elles ont tendance à être équiprobables ($|\{d \mid rsv(d, p) = score\}| = 1 \; ou \; 0$). La distribution des scores tend alors à être uniforme (voir figure. 2). En effet, dans un échantillon, il est très peu probable de trouver deux ou plusieurs documents ayant exactement le même score.

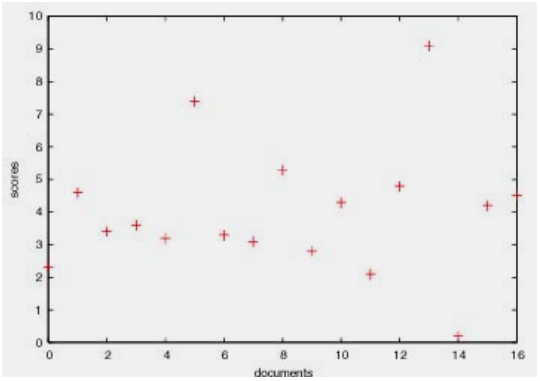


Figure.1 : La distribution des scores est très éparpillée

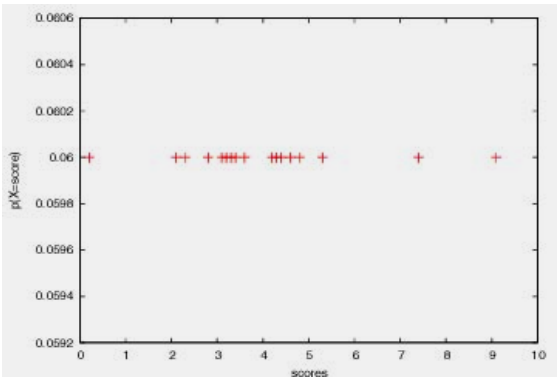


Figure.2 : La distribution uniforme n'est pas significative

Pour donner une distribution de probabilités plus significative, nous proposons qu'au lieu de calculer la probabilité qu'un document ait un score, nous calculons la probabilité que le score d'un document appartienne à un intervalle. Nous choisissons des intervalles réduits pour que les scores des documents appartenant au même intervalle soient réellement presque égaux. Soit n le nombre des ces intervalles, $I_1, I_2, ..., I_n$ de même rayon où :

$$I_i = [score_{i-1}, score_i]$$
$$score_{i-1} = \min_d rsv(d, p) \qquad (12)$$
$$score_i = \max_d rsv(d, p)$$

Le nombre d'intervalles est proportionnel à la taille de l'échantillon, car plus la taille de l'échantillon augmente, plus le domaine de définition des scores des documents s'élargit. Nous choisissons n comme la moitié de la taille de l'échantillon :

$$n = \frac{|\{d\}|}{2} \qquad (13)$$

La probabilité d'appartenance du score d'un document à un intervalle est définie par :

$$p(score_i < X < score_{i+1}) = \frac{|\{d \mid score(d, p) \in]score_i, score_{i+1}[\}|}{|\{d\}|} \qquad (14)$$

La distribution des probabilités par intervalle de scores est plus réaliste. La figure 3 illustre la distribution des probabilités basée sur la formule (14). Elle montre que la distribution des scores des documents admet une allure poisssonnienne.

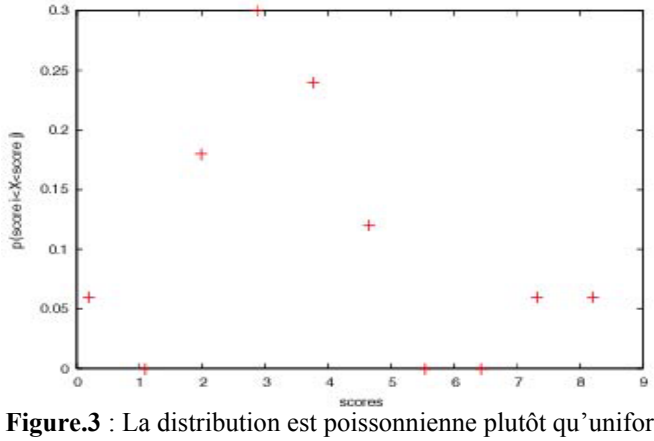


Figure.3 : La distribution est poissonnienne plutôt qu'uniforme

Plusieurs méthodes sont envisageables pour estimer une loi de probabilité suivie par les scores des documents : la régression paramétrique et l'estimation par le maximum de vraisemblance [Saporta, 1996]. L'application de telles méthodes permet d'estimer les paramètres qui permettent de fournir une représentation la plus fiable possible (celle qui passe le plus proche possible par tous les scores).

Dans un contexte expérimental, nous soulignons quelques inconvénients dans l'utilisation de ces méthodes :

- **Si on admet la forme de la fonction à priori, elle peut ne pas être valable pour des conditions expérimentales particulières, car les scores des documents sont aléatoires,**
- **On doit disposer d'un nombre minimum d'échantillons (de documents) pour avoir des estimations non biaisées,**
- **La distribution des scores dépend de la fonction de pondération des documents.**

Pour pallier ces problèmes, nous adoptons une méthode permettant de diviser l'espace probabilisé en plusieurs intervalles, tels que la distribution des scores dans chaque intervalle est linéaire. On applique la linéarisation sur l'ensemble des probabilités des documents pertinents et non pertinents, afin de déterminer le seuil de filtrage.

2.2.3. Linéarisation de la distribution de probabilités

La linéarisation consiste à parcourir le domaine de définition d'une fonction donnée, et de diviser ce domaine en intervalles tels que la courbe représentative de restriction de la fonction sur chaque intervalle peut être assimilée à une courbe linéaire. Nous utilisons cette technique pour linéariser respectivement, la courbe représentative de la densité de probabilités des scores des documents pertinents et non pertinents.

Le processus de détection des intervalles linéaires consiste à chercher le maximum de points adjacents tels que la courbe reliant ces points est linéaire. Nous appliquons la méthode des moindres carrés utilisée pour la régression linéaire [Saporta, 1996], pour déterminer la linéarisation d'un ensemble de points. Le principe est de calculer à chaque fois l'écart quadratique entre les points considérés et la courbe linéaire. Si l'erreur est inférieure à un seuil donné, on ajoute à la courbe le point suivant et on vérifie la linéarité de la nouvelle courbe, sinon on élimine ce point de cet ensemble et on recherche un nouvel ensemble de points à linéariser. La figure.4 illustre un exemple de linéarisation possible. Le processus de linéarisation est le suivant :

1. *c=1 ('c' est indice d'une classe linéaire),*
2. *P=∅,*
3. *seuil_erreur=0.001,*
4. *pour i∈ {0…n-1} (n est le nombre de points (score, probabilité),*
   a. *$P \leftarrow P \cup \{i\}$,*
   b. *Déterminer l'équation de la droite $D_c : y(x) = a + b.x$ par la régression linéaire sur tous les points de coordonnées ($j, p_j = p(score_j < score < score_{j+1})$) $\forall j \in P$,*
   c. *Calculer l'erreur représentée par l'écart quadratique des points ($j, p_j = p(score_j < score < score_{j+1})$) $\forall j \in P$ et la droite D_c :*

$$E = \sum_{j \in P} d^2((j, p_j), D_c)$$

$$d^2((j, p_j), D_c) = \frac{a + b.j - p_j}{\sqrt{a^2 + 1}}$$

   d. *Si E > seuil_erreur,*
      i. *$C_c = (d_c = \min(j \in P), f_c = \max(j \in P), a_c, b_c)$ avec a_c et b_c sont les coefficients de l'équation de la droite $y = a_c + b_c.x$ par la régression linéaire sur tous les points de coordonnées ($j, p_j = p(score_j < score < score_{j+1})$) avec $j \in P \mid \{i\}$,*

ii. $P \leftarrow \{i\}$,
iii. $c \leftarrow c+1$,

Ce processus permet seulement de représenter la distribution de probabilités des documents sous forme de plusieurs segments de droite. Cependant, il est nécessaire de transformer cette représentation pour former une distribution de probabilités continue :

1. **Premièrement, il faut relier les deux extrémités de chaque deux intervalles adjacents. Cette liaison s'effectue comme suit : pour deux classes linéaires adjacentes C_c et C_{c+1}, relier f_c et d_{c+1} par une droite $y = \alpha_c + \beta_c.x$. Cette doit passer par les points $(f_c, a_c + b_c.f_c)$ et $(d_{c+1}, a_{c+1} + b_{c+1}.d_{c+1})$.**

Soit g la fonction définie par :

$$g : [score_0, score_n] \rightarrow R \qquad x \rightarrow \begin{cases} a_c + b_c.x & si\ \exists c, d_c \le x \le f_c \\ \alpha_c + \beta_c.x & sinon,\ avec \end{cases}$$

2. **Deuxièmement, il faut normaliser les coefficients a_c, b_c, α_c et β_c pour que :**

$$\int_{score_0}^{score_n} g(x)\,dx = 1$$

Puisque $\int_{score_0}^{score_n} g(x)\,dx$ représente la surface de l'aire formée par la représentation graphique de g et l'axe des abscisses, il suffit de diviser les coefficients a_c, b_c, α_c et β_c par cette valeur. L'aire est calculée comme la somme des aires de chaque surface d'une classe linéaire.

La figure 5 illustre une linéarisation effectuée sur l'ensemble des documents pertinents et non pertinents de la base **Reuters** dans le cas du profil 1. Elle montre que la linéarisation des probabilités de scores tend à avoir une allure exponentielle pour les documents non pertinents et une allure gaussienne pour les documents pertinents.

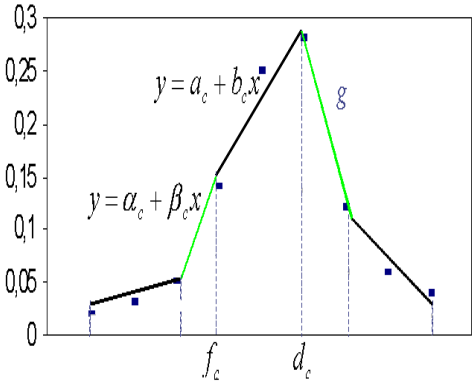


Figure.4 : **linéarisation de la**

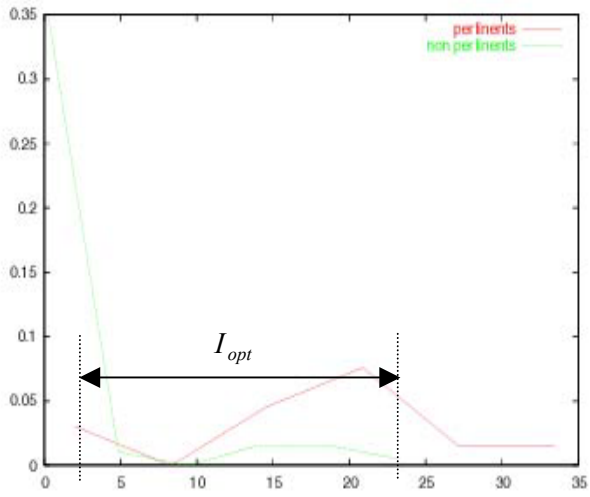


Figure. 5 : Densité de probabilités des scores des documents pertinents et non pertinents

2.2.4. Optimisation de la fonction de seuillage

La méthode que nous proposons pour choisir le meilleur seuil de filtrage est basée sur l'optimisation d'une fonction d'utilité. Une fonction d'utilité permet de mesurer la performance d'un système de filtrage d'information. L'optimisation de la fonction d'utilité consiste à déterminer un seuil de convergence, c'est-à-dire à quel moment il ne devient plus nécessaire de continuer l'adaptation. Ce seuil doit donc permettre au système de sélectionner le maximum de documents pertinents et d'éliminer le maximum de documents non pertinents. Le principe est le suivant :

i. Définir un intervalle de seuils $I_{opt} = [score_{\min p}, score_{\max np}]$ (figure. 5), où $score_{\min p}$ et $score_{\max np}$ représentent respectivement la valeur minimale et maximale des scores des documents pertinents et non pertinents,

ii. Détecter la valeur du seuil dans l'intervalle I_{opt} qui optimise la fonction d'utilité F :.

$$F = a.R_+ - b.S_+ \qquad (15)$$

Alors, notre objectif est de déterminer le seuil qui permet de maximiser théoriquement la fonction F, soit :

$$seuil^* = \arg \max_{seuil \in I_{opt}} F$$

Où :

a, b : deux constantes positives,

R_+ (S_+) : nombre de documents pertinents (non pertinents) sélectionnés,

Les valeurs R_+ et S_+ dépendent évidemment du choix du seuil et sont estimés par :

$$R_+ = p(r|score > seuil) * R$$

$$S_+ = p(s|score > seuil) * S$$

Où R et S représente le nombre total de documents pertinents et documents non pertinents.

Par application de la transformation de la règle de Bayes, nous obtenons :

$$R_+ = \frac{p(score > seuil|r) * p(score > seuil)}{p(r)} * R$$

$$S_+ = \frac{p(score > seuil|s) * p(score > seuil)}{p(s)} * S$$

Avec $p(score > seuil|r)$ (resp. $p(score > seuil|s)$) représente la probabilité qu'un document soit sélectionné sachant qu'il est pertinent (resp. non pertinent) et elle est définie par l'aire de la surface formée par la courbe de g correspondant à la densité de probabilité des scores des documents pertinent à partir du seuil.

3 - EXPERIMENTATION ET RESULTATS

Les expérimentations que nous avons effectuées ont été réalisées sur une collection issue de la compagne TREC-10**. Dans ce paragraphe, les expérimentations sont concentrées sur la dernière base fournie par** TREC **: la base** Reuters**. Cette base est constituée d'un ensemble de documents de 783484 documents en format XML (eXtensible Markup Language). Les tests sur cette base doivent porter sur 84 profils. Le but de ces expérimentations est de mettre en valeur principalement la technique d'adaptation du seuil précitée.**

L'adaptation du seuil exploite directement les scores des documents filtrés, plus les paramètres de la fonction d'utilité *TU10* [Robertson & al., 2000].Ainsi, pour chaque profil, nous appliquons l'algorithme d'adaptation incrémentale du seuil suivant :

1. $echantillonP \leftarrow \Phi$ (un échantillon de documents pertinents)
2. $echantillonN \leftarrow \Phi$ (un échantillon de documents non pertinents)
3. $seuil^{(0)} = 0$
4. *pour chaque document* $d^{(t)}$
   a. *calculer* $rsv(d^{(t)}, p^{(t)})$
   b. *si* $rsv(d^{(t)}, p^{(t)}) > seuil^{(t)}$
      i. *si* $d^{(t)}$ *est jugé pertinent*
         A. $R_+ \leftarrow R_+ + 1$
         B. $echantillonP \leftarrow \{d^{(t)}\}$
         C. Apprendre le profil
         D. Apprendre le seuil en utilisant $echantillonP$ et. $echantillonN$
      ii. *Sinon*
         $S_+ \leftarrow S_+ + 1$

5. *évaluer :* $T10U = 2R_+ - S_+$

Pour mesurer la performance de notre méthode de seuillage, nous avons utilisé les 15 profils de la base **Reuters**, et avons comparé les valeurs d'utilité obtenues, à celles obtenues par les meilleurs systèmes ayant participé à **TREC-10**.

Le tableau 1 montre que sur les 11 profils, on obtient un résultat meilleur que tous les autres participants, et pour la totalité des profils on obtient un résultat qui dépasse la moyenne des résultats des autres participants.

Profil	Utilité obtenue	Utilité maximale	Utilité moyenne
1	0.11	0.10	0.02
2	0.58	0.30	0.13
3	0.03	0.14	0.02
4	0.55	0.24	0.07
5	0.84	0.35	0.06
6	0.63	0.33	0.15
7	0.28	0.37	0.14
8	0.28	0.51	0.26
9	0.36	0.34	0.15
10	0.52	0.77	0.30
11	0.27	0.11	0.01
12	0.45	0.41	0.11
13	0.26	0.21	0.06
14	0.26	0.10	0.03
15	0.52	0.22	0.06

Tableau 1 : Résultats du seuillage selon la distribution des scores des documents

CONCLUSION

Nous nous sommes intéressés dans cet article, plus particulièrement au problème d'adaptation incrémentale du profil et du seuil de décision. L'adaptation est effectuée à chaque sélection d'un document pertinent.

L'adaptation du profil est basée sur la pondération des termes candidats extraits du document en arrivée. Pour chaque terme, un poids provisoire est calculée par un système d'équations sous contraintes. Ce poids provisoire contribue partiellement au poids final de chaque terme dans le profil en utilisant une technique de distribution de gradient.

L'adaptation du seuil, est basée sur une technique probabiliste, elle se base sur le distribution des probabilités des scores des documents d'un échantillon choisi et mis à jour continuellement, à chaque arrivée d'un document.

Des expérimentations ont été réalisées sur une Base **Reuters** issue de TREC-10. Les résultats obtenus montrent l'efficacité de notre processus adaptatif par rapport aux autres systèmes ayant effectués des tests sur la même base.

Nos futurs travaux concernent, l'intégration de la technique de détection de nouveauté durant le filtrage. Autrement dit, chaque fois qu'un document pertinent est sélectionné le système doit détecter si le document est porteur de nouvelles informations ou non. Ainsi, notre processus d'adaptation du profil et du seuil doit agir en fonction du type d'information que le document contient.

REFERENCES

[Belkin & al., 1992] N. J. Belkin, W. B. Croft. "*Information retrieval and information filtering: Two sides of the same coin?*". CACM, pages 29-38, 1992.

[Boughanem, 2000] M. Boughanem. "*Formalisation et spécification des systèmes de recherche et de filtrage d'information*" . HDR de l'université Paul Sabatier de Toulouse, 2000.

[Boughanem & al., 2002] M. Boughanem, M. Tmar. "*Incremental adaptive filtering: Profile learning and threshold calibration*". Proceedings of ACM-SAC, pages 640-644, 2002.

[Hoashi & al., 1999] K. Hoashi, K. Matsumoto, N. Inoue, K. Hashimoto, "*Experiments on the TREC-8 filtering track*". Proceedings of TREC-8, 1999.

[Kwok & al., 1999] K. L. Kwok, L. Grunfeld, M. Chan. "*TREC-8 Adhoc, query and filtering track experiments using PIRCS*". Proceedings of TREC-8, 2000.

[Malone & al., 1987] T. W. Malone, K. R. Grant, F. A. Turbak, S. A. Brobst, M. D. Cohen. "*Intelligent information sharing systems"*. CACM, 30(5), pages 390-402, 1987.

[Robertson & al., 1976] S. Robertson, K. Sparck Jones. "*Relevance weighting of search terms*". JASIS, 27(3), pages 129-146, 1976.

[Robertson & al., 1999] S. E. Robertson, S. Walker. "*Okapi-Keenbow at TREC-8*". Proceedings of TREC-8, 1999.

[Robertson & al., 2000] S.E.Robertson, H. Hull. "*The TREC-9 filtering track final report"*. TREC-P, 2000..

[Saporta, 1996] G. Saporta. *"Probabilités, analyse des données et statistiques*". édition Technip, 1996.

[Sutton & al., 1998] R. S. Sutton, A. G. Barto. *"Reinforcement learning: An introduction*". MIT Press, Cambridge, MA, 1998.

[Tmar, 2002] M. Tmar. *"Modèle auto-adaptatif de Filtrage d'Information: Apprentissage incrémental du profil et de la fonction de décision* ". Thèse de l'Université Paul Sabatier de Toulouse, 2002.

[Zhai & al., 1998] C. Zhai, P. Jansen, E. Stoica, N. Grot, D. Evans. "*Threshold Calibration in Clarit Adaptive Filtering*". Proceedings of TREC-7, pages 149-156, 1998.

L'ANALYSE DES MOTS ASSOCIES POUR L'INFORMATION ECONOMIQUE ET COMMERCIALE

EXEMPLE SUR DES DEPECHES "REUTERS BUSINESS BRIEFING"

B. Delecroix
France Telecom
CESD/ISIS - Université De Marne la Vallée

R.Eppstein
CESD/ISIS - Université De Marne la Vallée

Résumé : L'analyse des mots associés se fonde sur une théorie sociologique principalement développée par le CSI et le SERPIA (Callon, Courtial, Turner) au milieu des années 80. Cette analyse mesure la force d'association entre les termes d'un corpus documentaire pour dégager et mettre en évidence l'évolution d'un domaine scientifique à l'aide de la construction d'agrégats de termes (clusters) et d'un diagramme stratégique. Depuis ces premiers travaux, la méthode des mots associés a été employée avec succès pour aider à comprendre la structure et l'évolution de nombreux champs scientifiques. Aujourd'hui, elle est implantée dans plusieurs logiciels qui sont utilisés par les entreprises dans des systèmes d'aide à la décision leur permettant d'améliorer leur compétitivité et de définir leur stratégie. Cependant, il n'existe pas de travaux permettant de valider cette méthode dans ce type de contexte.
Au travers d'un exemple d'analyse d'un corpus d'information à caractère économique et marketing sur les technologies DSL tiré de "Reuters Business Briefing", cette présentation donne une interprétation aux résultats obtenus par l'analyse des mots associés. Après un rapide aperçu du logiciel utilisé (Sampler), et après avoir clairement défini le protocole expérimental utilisé, nous examinons chaque étape du processus en œuvre lors de l'analyse des mots associés : extraction terminologique, calcul des clusters et du diagramme stratégique. Nous analysons le rôle de chaque paramètre de cette méthode avant de donner une interprétation plus générale de la méthode des mots associés dans un contexte d'information économique tirée de dépêches d'agences. D'autres travaux viendront compléter ceux-ci afin de généraliser les résultats obtenus.

Abstract : Co-word analysis is based on a sociological theory developed by the CSI and the SERPIA (Callon, Courtial, Turner) in the middle of the eighties. It measures association strength between terms in documents to reveal and visualise evolution of science through the construction of clusters and strategic diagram. Since, this method has been successfully applied to investigate the structure of many scientific fields. Nowadays it occurs in many software systems which are used by companies to improve their business and define their strategy but the relevance in this kind of application has not been proved yet.
Through the example of economic and marketing information on DSL technologies from Reuters Business Briefing, this presentation gives an interpretation of co-word analysis for

this kind of information. After an outlook of the software we used (Sampler) and after a survey of the experimental protocol, we investigate and explain each step of the co-word analysis process : terminological extraction, computation of clusters and strategic diagram. In particular, we explain the meaning of every parameter of the method. Finally we try to give global interpretation of the method in an economic context. Further studies will be added to this work in order to allow a generalisation of these results.

Mots Clés : classification, mots associés, intelligence économique

Keywords : clustering, co-word analysis, competitive intelligence

l'Analyse des mots associes pour l'Information economique et commerciale

Exemple sur des dépèches "Reuters Business Briefing"

INTRODUCTION

Beaucoup d'entreprises utilisent la veille et l'intelligence économique pour développer leur activité et mieux appréhender leur environnement. Parmi les logiciels présents sur ce marché, nombre d'entre eux mettent en œuvre la méthode des mots associés. Il existe de nombreuses références sur cette méthode d'analyse (Callon, courtial, Turner, Bauin, 1983; Courtial, Callon & Laville, 1991; Courtial, 1994; Law & Whittaker, 1992). La méthode des mots associés révèle la structure d'un champ scientifique et en dessine l'évolution en mesurant la force d'association entre les expressions représentatives du corpus considéré. Cette méthode n'a donc pas été conçue pour l'analyse de l'information économique ou financière et en particulier des informations provenant de dépêches d'agences de presse.

Le corpus de 700 dépèches extraites de Reuters Business Briefing traite des technologies DSL, secteur économique hautement compétitif durant ces derniers mois. Que devient "l'analyse de l'évolution d'un champ scientifique" sur un tel type de document? Dans cette étude, nous tentons de valider l'utilisation de cette méthode et de fournir une interprétation aux résultats obtenus dans ce contexte.

1 - L'ANALYSE DES MOTS ASSOCIES

La méthode des mots associés réduit un vaste espace d'expressions à de multiples espaces plus petits et donc plus facilement compréhensibles et interprétables qui sont également représentatif des relations entre thèmes abordés dans l'ensemble des documents considérés. Cette analyse nécessite la définition d'une mesure d'association, mais également la mise en œuvre d'un algorithme permettant la mise en évidence de ces thèmes.

Différentes mesures d'association ont été étudiées (Callon, Law, & Rip, 1986; Grivel et François, 1995), mais la plus couramment employée reste l'indice d'équivalence. Deux termes i et j cooccurrent s'ils sont employés simultanément dans un document. Considérons un corpus de N documents. Chaque document est indexé par un ensemble de termes ou d'expressions qui peuvent apparaître dans de nombreux documents. Soit C_k le nombre d'occurrences du terme k, Soit C_{ij} le nombre de cooccurrences des termes i et j, alors l'indice d'équivalence E_{ij} est défini par :

$$E_{ij} = \frac{C_{ij}^2}{C_i * C_j} \text{ , où } 0 \leq E_{ij} \leq 1$$

Cette mesure permet la mise en évidence d'associations fortes, en tenant compte de la fréquence d'apparition de chaque terme dans le corpus considéré. Ainsi deux termes qui apparaissent souvent dans le corpus mais simplement quelques fois ensemble seront représentés par un lien plus faible que deux expressions apparaissant moins souvent mais systématiquement de façon simultanée. Le réseau de termes est constitué de nœuds (les termes) interconnectés par des liens représentant leur force d'association respective. L'algorithme utilisé produit deux types de liens. Les liens internes représentent les relations les plus fortes entre expressions ; les liens externes qui représentent les associations plus faibles et servent à mettre en relation les clusters. Chaque cluster est donc formé par l'ensemble des expressions liées par un lien interne. Les clusters sont liés entre eux par les liens externes.

2 - LE LOGICIEL SAMPLER

Sampler est un logiciel d'analyse lexico-statistique développé par la Cisi (Jouve 1996), aujourd'hui filiale du Groupe CS Communication & Systèmes. Ce logiciel se fonde sur les recherches du *Service d'Etude et de Réalisation de Produits d'Information Avancés (SERPIA)* et du *Centre de Sociologie de l'Innovation (CSI) de l'Ecole des Mines* (Callon&Courtial&Turner,1983) et sur le développement par ces équipes du logiciel Leximappe (Michelet 1988). Dans Sampler, l'analyse des mots associés est complétée par une analyse morpho-syntaxique capable d'extraire des unitermes mais également des multitermes, amenant ainsi une réduction de la polysémie.

Les termes et expressions sont agrégés par la méthode des mots associés pour former des réseaux d'associations au sein desquels il est possible de naviguer graphiquement. Chaque réseau, également appelé "cluster", ne correspond aucunement à une structure sémantique mais plutôt à des associations contextuelles entre expressions. Les clusters ne sont pas orientés et contiennent les deux types de liens décrits dans la méthode des mots associés. Les liens internes représentent les associations où l'occurrence de chaque terme est peu différente de la cooccurrence. Les liens externes représentent les associations de termes qui apparaissent dans un contexte différent. La construction des clusters est effectuée par l'application d'un algorithme de Classification Ascendante Hiérarchique utilisant l'indice d'équivalence comme distance entre expressions.

3 - PROTOCOLE EXPERIMENTAL

Au total, 800 dépêches en langue anglaise concernant les technologies DSL (Digital Subscriber Lines) sur une période de six mois (d'Octobre 2001 à Avril 2002) ont été extraites du service "Reuters Business Briefing". L'équation exacte de recherche utilisée est "dsl OR adsl OR xdsl OR digital subscriber lines".

Pour rendre notre analyse pertinente, un nettoyage complet des données récupérées a été effectué. Nous avons tout d'abord effacé les doublons (le service Reuters Business Briefing collecte les dépêches depuis de nombreux fils de presse sans effectuer ce travail) pour obtenir 700 dépêches uniques. Nous avons ensuite nettoyé ces documents afin d'éliminer les mots, balises ou expressions interférant avec l'analyse tels que le nom de l'auteur, la ville depuis laquelle a été émise la dépêche ou bien encore le nom de l'agence de presse émettrice.

Les paramètres du logiciel Sampler ont été fixés comme suit :

- Nombre minimum de cooccurrences : 3
- Nombre minimum d'occurrences : 3
- Nombre maximum de liens internes : 20
- Nombre maximum de liens externes : 20
- Nombre maximum de mots par cluster : 10

Ces choix ont été effectués en tenant compte des paramètres par défaut ainsi que de notre expérience d'analyse utilisant la méthode des mots associés sur des documents scientifiques mais également des précédentes expériences sur ce sujet (Ding et al., 2000 ; Grivel et al., 1995)

Après une première phase d'extraction terminologique, les descripteurs obtenus ont été standardisés manuellement afin d'éliminer les variantes les plus remarquables. Cette opération a été supervisée par un expert des technologies DSL de France Telecom. On remarquera ici l'importance de cette étape dans le processus d'analyse. Cette expérience a confirmé que l'index final n'était obtenu qu'après plusieurs itérations et dans un délai de plusieurs jours. Cette remarque doit nous conduire à une réflexion sur l'utilisation de ce type d'outil en fonction du problème considéré et des moyens humains à investir pour obtenir un résultat significatif.

4 - ANALYSE

4.1 - examen de l'index

L'index produit par Sampler contient 1394 termes d'occurrence variant entre 3 et 1590. Les termes ayant une occurrence de 1 ou 2 ont été volontairement occultés.

Un examen rapide de l'index permet de situer le thème traité par les documents. Les dix premiers mots sont *Dsl, adsl, broadband, customers, Internet, business, data, users, dsl services, alcatel*. Un peu plus loin, d'autres termes tels que *communication, adsl service, high speed, Internet service, high speed Internet, bandwith*…permettent de situer le contenu des données. D'autre part, un examen rapide permet également d'identifier les principaux acteurs du domaine : *Alcatel* (10^{eme}), *sbc communications* (15^{eme}), *verizon communications* (25^{eme}, *bellsouth* (34^{eme}), *lucent technologies* (36^{eme}), *chunghwa telecom* (46^{eme}), *France telecom* (49^{eme}), *deutsche telekom* (59^{eme}) …

On peut ici remarquer que le nombre d'occurrence (nombre d'apparition du terme dans le corpus) des noms de sociétés est relativement élevé alors que leur fréquence (nombre de documents dans lesquels ces termes apparaissent) est relativement faible. Ce phénomène est facilement explicable par la nature des données à traiter. En effet, les dépêches d'agences qui ont comme sujet le *DSL* sont souvent en relation avec une ou plusieurs sociétés. De plus chaque société est citée plus fréquemment dans chaque dépêche (en moyenne 3 fois par dépêche contre 2 pour des termes plus génériques).
La principale conséquence est la surreprésentation des sociétés dans les clusters produits par l'analyse.

4.2 - Analyse des clusters

Le logiciel Sampler produit 72 clusters à partir de l'index et des paramètres initialement fixés. On peut distinguer deux grandes catégories de clusters :

Les clusters constitués de termes génériques

Les termes génériques sont par définition centraux mais peu denses (par exemple *dsl*). Les clusters constitués de termes génériques permettent de distinguer les grands sous-thèmes du domaine analysé (*dsl equipment*, *telecom*, *dslam*…). Ces clusters apportent peu de valeur ajoutée à l'expert mais sont utiles pour les personnes désirant se familiariser avec le domaine étudié.

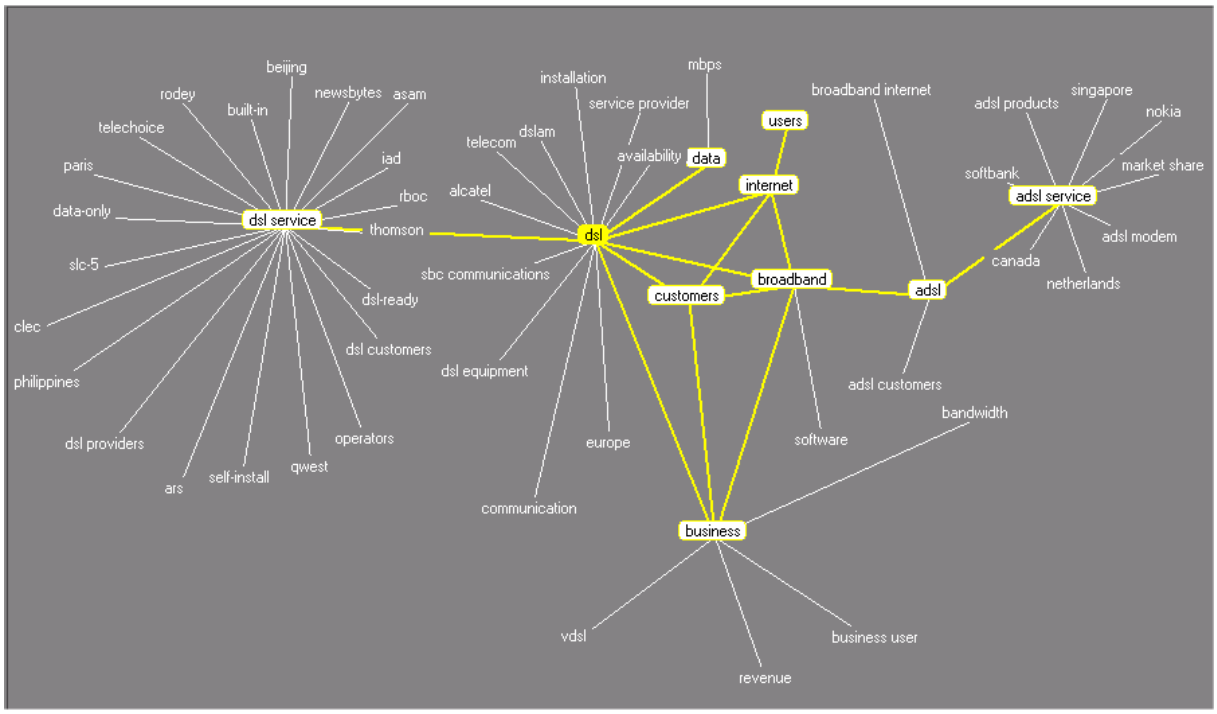


Fig. 1 : Cluster DSL

Dans le deuxième exemple (figure 2), le cluster Alcatel permet d'identifier les cinq acteurs majeurs du marché des équipements DSL. Ces relations ne doivent pas être interprétées comme des accords de partenariats ou tout autre lien qui pourrait exister entre ces sociétés, mais reflètent l'association effectuée dans les communiqués à titre de comparaison ou de référence dans ce secteur.

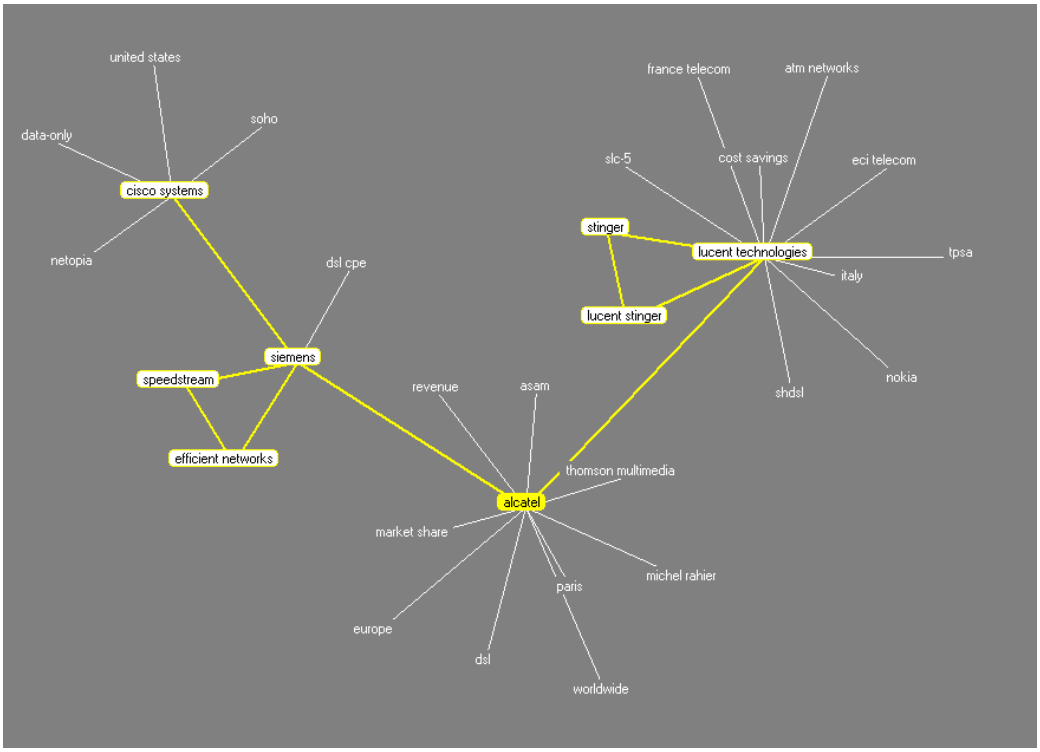


Fig 2. Cluster Alcatel

Détection de signaux faibles

Il existe, en très petite quantité, des clusters révélant des signaux faibles. Ces clusters ont été identifiés par des experts. Par exemple, le cluster Nokia présente son implantation géographique en Chine au cours de la période considérée.

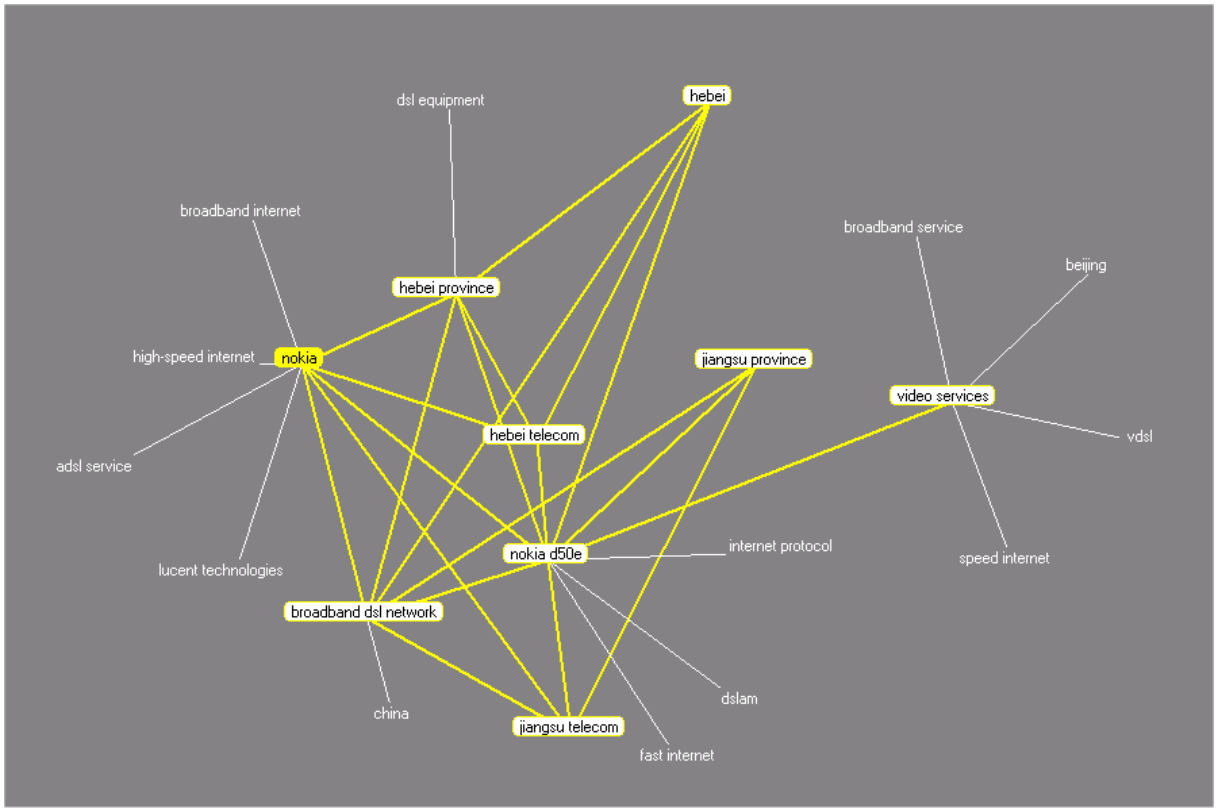


Fig 3. Cluster Nokia

Si Nokia est reliée à des termes relatifs à son activité (*broadband Internet, high-speed Internet, adsl service…*), cette société est également connectée à deux provinces chinoises et aux opérateurs locaux correspondants : *hebei province* and *hebei telecom*, *jiangsu province* and *jiangsu telecom*.
Cette activité récente de Nokia en Chine a été confirmée par une étude de marché *Idate.* Le cluster permet de mettre en évidence l'activité du constructeur durant cette période

Dans le deuxième exemple, on observe que le terme "*Federal Communication Commission"* est reliée aux termes *Telecom Act* et *Cable Providers* . Cette remarque est intéressante pour l'expert qui, en se référant aux dépêches d'origine, a pu confirmer un changement dans la réglementation concernant les câblo-opérateurs susceptible d'avoir un impact fort sur l'activité des opérateurs DSL.

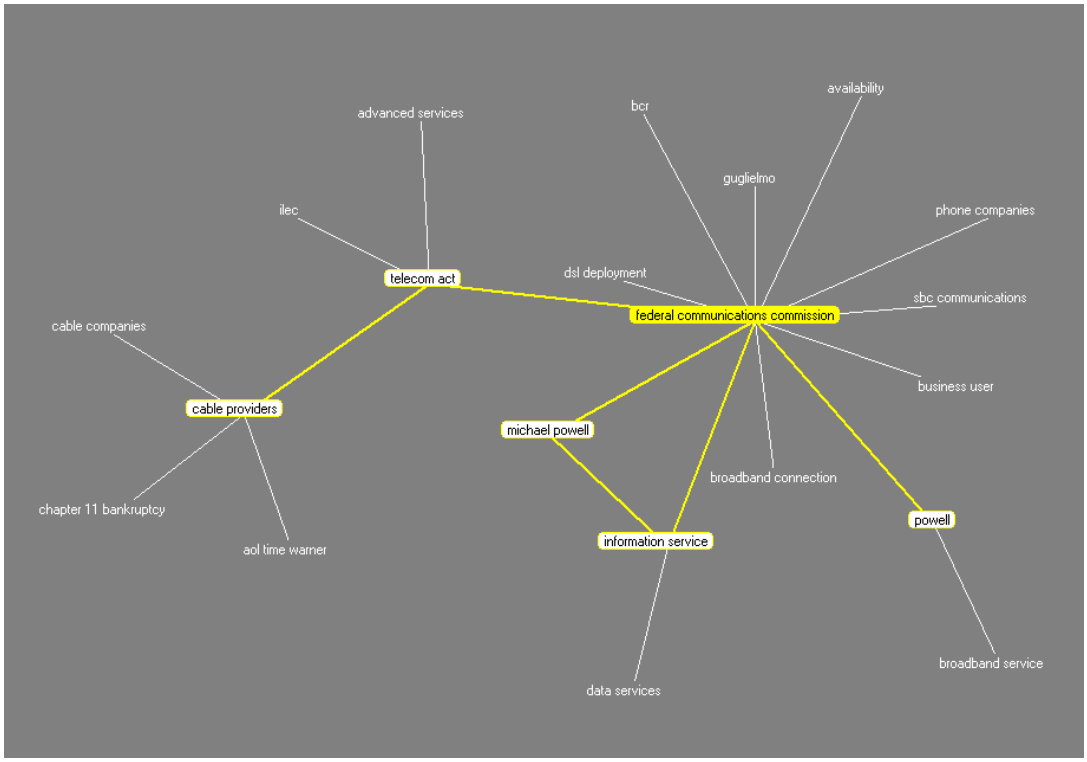


Fig 4. Cluster «Federal communication commission »

Synthèse sur l'analyse des clusters

Il existe deux types de cluster, chaque type ayant un intérêt pour une catégorie différente d'utilisateurs :
Les clusters qui permettent une bonne vue d'ensemble du domaine. Ces clusters permettent de découvrir les principaux acteurs, les technologies en jeux ainsi que les zones géographique où l'activité commerciale est la plus forte. Au sein de ces clusters, la cooccurrence est variable et l'indice d'équivalence souvent faible.
Les clusters recelant un signal faible. Ces clusters reflètent de faibles cooccurrences mais un indice d'équivalence fort. Ils contiennent une information à forte valeur ajoutée pour l'expert et mettent en scène les acteurs du domaine.

CONCLUSION

Cette expérience a été réalisée dans le but de valider ou d'invalider l'utilisation de la méthode des mots associés dans un contexte économique et financier. Elle nous a également permis de vérifier la pertinence de la méthode sur des dépêches de type Reuters.

Nous devons tout d'abord souligner que l'analyse est fortement dépendante de l'extraction terminologique, une extraction différente menant à un autre index et à la constitution de clusters différents. Nous avons essayé l'extracteur du logiciel *Leximine* sur le même corpus avec un résultat sensiblement différent (une meilleure extraction des multitermes mais des possibilités de modification manuelle de l'index obtenu moindre)

Peu de clusters sont directement interprétables (une quinzaine sur 72) De plus, les liens internes n'ont souvent aucun sens. L'interprétation, lorsqu'elle est possible, conduit à deux types de raisonnement :

- Le cluster est une interconnexion de termes du même ordre (pays, acteurs, technologies). Ces clusters sont utiles pour un profane souhaitant découvrir le domaine.

- Le cluster contient un signal faible ou un épiphénomène : Il est alors d'une grande aide pour l'expert.

La structure du corpus ne permet pas d'envisager l'évolution du domaine considéré. Les signaux faibles ne traduisent pas l'émergence de tendances fortes, mais traduisent une actualité ponctuelle.

En conclusion, la méthode des mots associés expérimentée au travers du logiciel Sampler fait sens, mais de façon différente. Si elle permet la détection de signaux faibles, elle n'est globalement pas aussi pertinente. De plus le temps de mise en œuvre est relativement important par rapport aux résultats obtenus.

D'autre travaux sur ce sujet sont en cours. Ils permettront une analyse plus fine des résultats obtenus et permettront à terme une généralisation des conclusions. Une étude « en dynamique » est actuellement en cours pour observer l'évolution des résultats présentés ici.

BIBLIOGRAPHIE

Eppstein R., (2001), Création d'un système d'information stratégique dans le domaine des technologies de l'information et de la communication, PhD Thesis – Université de Marne-la-Vallée.

Eppstein R., Datchary F. (1999), Complementarities between statistical and semantical analysis. in *17th Codata Conf Proceedings*.

Small, H. (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents, JASIS, vol 24 p 265-269.

Small, H., Griffith, B.C., (1974) The structure of the scientific literature: identifying and graphing specialities, Science Studies, vol 4, p17-40

Turenne N., Rousselot F., (1998) Evaluation of four clustering methods used in text mining, *ECML'98 text mining workshop*

Y. Ding, G. Chowdhury, S. Foo, (2000) Bibliography of information retrieval research using co-word analysis" in *Information Process Management* 517, p1-26, Elsevier Ed.

Courtial J.P., Callon M., Laville F. (1991). Co-words analysis as a tool for describing the networks of interaction between basic and technological researches : the case of polymer chemistry, *Scientometrics*, Vol 22, Nº 1,

Callon M., Courtial J.P., Turner W., Bauin S., (1983) From translation to problematic networks : an introduction to co-word analysis, *Social Science Information* n°22, 1983.

Callon M., Courtial J.P., Turner W., (1991) La méthode Leximappe, un outil pour l'analyse stratégique du développement scientifique et technique, *Gestion de la recherche : nouveaux problèmes nouveaux outils, ed. de Boeck.*

Grivel L.,François C., (1995) Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique, *Solaris n°2, Presses Universitaires de Rennes*.

Peyrichoux, I., (2000) Sampler, un logiciel d'analyse textuelle : de la conception aux usages, *Internal Report CNAM/INTD*.

Polanco X., (1993) Analyse stratégique de l'Information Scientifique et Technique : Construction de clusters de mots-clés, *Sciences de la société,* n° 29.

Jouve O., (1999) Les outils d'analyse et de filtrage d'information : l'exemple du projet Sampler, *IDT'99* - Paris.

Jouve O., (1996) Sampler Manual, Cisi - Paris.

Turenne N., Rousselot F. (1998) Application of clustering in a system of query reformulation, *KAW'98*.

Law, J., Whittaker, J., (1992) Mapping acidification research: a test of the coword method, Scientometrics, Vol 23, p 417-461.

Courtial, J.P (1994) A coword analysis of scientometrics, Scientometrics, VOL 31, p251-260.

ENSEIGNEMENT A DISTANCE – L'EXPERIENCE ACQUISE AU COURS DE LA REALISATION DE LA MAITRISE A DISTANCE NTIDE (NOUVELLES TECHNOLOGIES DE L'INFORMATION POUR LE DEVELOPPEMENT DES ENTREPRISES).

Henri Dou,
dou@crrm.u-3mrs.fr

Céline Riffaut,
celine@crrm.u-3mrs.fr

Hervé Rostaing,
rostaing@crrm.u-3mrs.fr

Adresse professionnelle
CRRM, Centre Scientifique de Saint Jérôme, Université Aix Marseille III
13397 Marseille cedex 20
http://crrm.u-3mrs.fr

Résumé : Dans cette présentation, nous récapitulons brièvement les points de blocage, les avantages, les freins et les leviers mis en évidence par l'expérience de la création d'une formation à distance délivrée principalement via l'Internet depuis 4 ans. Par la suite, les développements prévisibles de telles formations sont évoqués en précisant l'importance de l'enjeu pour le système éducatif français qui se voit soumis à une concurrence internationale pour la conquête de ces « nouveaux marchés » que représentent la formation et l'enseignement.

Abstract : In this presentation, we briefly recapitulated the breaking points, the advantages, the slowing downs and the spurs highlighted by the experiment of the creation of a distance learning training by Internet for 4 years. Thereafter, the foreseeable developments of such formations are exposed specifying the importance of the stakes for the French education system in the international competition for the conquest of these "new markets" that the training and teaching represent.

Mots-clés : Enseignement à distance, e-learning, formation non-présentielle, formation ouverte à distance, formation continue à distance

Keywords : Distance learning, e-learning, platform to create knowledge, life-long learning

Enseignement à distance – L'expérience acquise au cours de la réalisation de la Maîtrise à distance NTIDE (Nouvelles Technologies de l'Information pour le Développement des Entreprises).

INTRODUCTION

Depuis quatre ans, la maîtrise NTIDE (http://ntide.u-3mrs.fr), se déroule au CRRM. Cette maîtrise (Diplôme National), est exclusivement réalisée via l'Internet et l'envoi de certains CD-ROM (lorsque les fichiers sont volumineux). Nous récapitulons brièvement dans cette présentation les points de blocage, les avantages, les freins et les leviers que cette expérience nous a permis de mettre en évidence. Les objectifs poursuivis sont présentés dans la figure 1 :

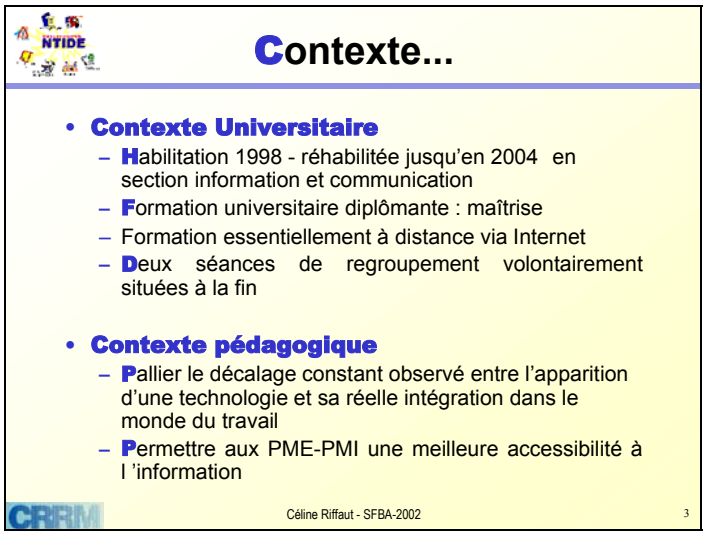


Figure 1 : les objectifs poursuivis

1 - ENTREE EN MATIERE :

Il n'est pas aisé de mettre en place dans le système éducatif (surtout français) des changements importants qui bouleversent les approches classiques de l'étudiant présentiel et de la toute puissance des enseignants par l'octroi de locaux, de salles et de discussions, entre autre, au niveau des emplois du temps des répartitions de salles, etc… Pourtant, il n'est plus de mise, dans un système où la formation tout au long de la vie est nécessaire, d'exiger d'un apprenant qu'il quitte son travail et vienne pendant un temps assez long (de l'ordre de six mois) à l'Université. Outre les problèmes d'emploi et familiaux, les contraintes de coût ne seraient plus supportables. C'est en prenant en compte ces différents paramètres que nous avons résolu de développer dans un cadre utile aux entreprises un diplôme national, qui outre sont impact sur la connaissance générale et le développement des entreprises, permettra aussi par la validation des acquis professionnels d'aider à la promotion des employés. Dans la figure 2 nous présentons le fonctionnement de la maîtrise :

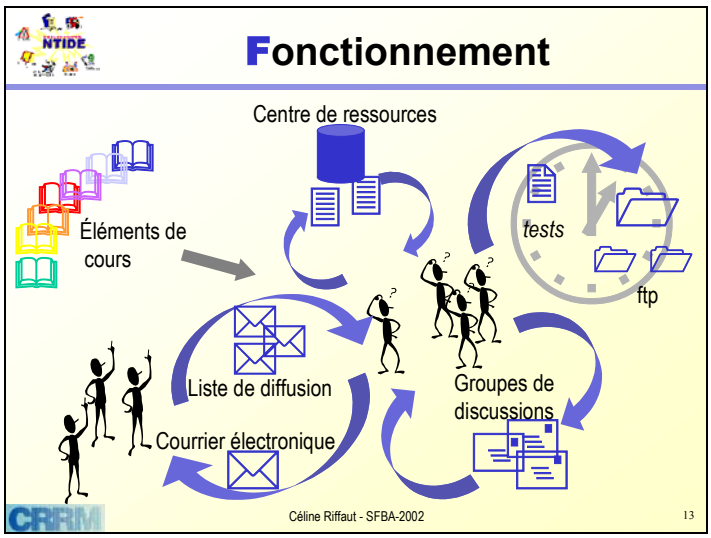


Figure 2 : le fonctionnement de la maîtrise

2 – LES CONTRAINTES

Nous allons procéder à une analyse par contrainte et par levier, afin de mettre en perspective le cadre général de ce nouvel enseignement.

- Tous les types d'enseignement (disciplines) ne se prêtent pas à l'enseignement à distance, certains domaines sont privilégiés par rapport à d'autres. Dans notre cas, l'enseignement à distance, par l'intermédiaire de l'Internet nourrit la maîtrise qui doit utiliser cette technologie pour développer les usages des NTI (Nouvelles Technologies de l'Information) dans les entreprises.

- Une des contraintes majeures lors de la diffusion des cours sur l'Internet reste la propriété (copyright) de ces derniers. Notre surprise a été assez forte lorsque nous avons constaté que pour certaines disciplines (les enseignants étaient venus nous consulter), les cours avaient principalement pour sources des livres et autres écrits dont les enseignants qui les utilisaient n'étaient pas les auteurs. Ainsi, il n'était pas possible dans ce cas de les utiliser. (sauf après avoir demandé l'avis ou l'accord des auteurs ou des éditeurs pour les livres ou les publications).

- Au niveau de la propriété des cours, lorsque l'Université ne reconnaît pas statutairement l'effort fait par les enseignants pour transformer leurs cours de manière interactive et électronique, il faut que les enseignants rétrocèdent leurs cours à une entreprise dont un des objets est l'édition. De ce fait l'Université ne sera pas propriétaire de ces derniers (comme cela est le cas lors de la réalisation d'un ouvrage). Les cours peuvent être utilisés gratuitement par clause spéciale, mais la propriété de l'auteur est sauvegardée.

- Sur le plan de l'Université, la non reconnaissance du temps passé dans l'encadrement des étudiants à distance (ce qui est plus long que dans le présentiel) n'est pas pris en compte. Non pas que l'Université ne le reconnaisse pas, mais pour ne pas créer de précédent. Cette difficulté à gérer les problèmes ne laisse évidemment pas bien augurer de la poursuite en France de telles expérimentations.

- Un autre frein réside dans le fait que certaines formations lorsqu'elles font l'objet d'un financement par des organismes publics doivent obéir à un certain format qui ne prend en compte que le présentiel. Cependant, nous avons pu inscrire la maîtrise dans le PRF[53] de la région Provence Alpes Côte d'Azur, ce qui à l'évidence montre que cet obstacle peut très bien être surmonté. Nous avons aussi développé des outils pour suivre le travail des étudiants, ceci est présenté dans la figure 3 :

[53] Plan Régional de Formation

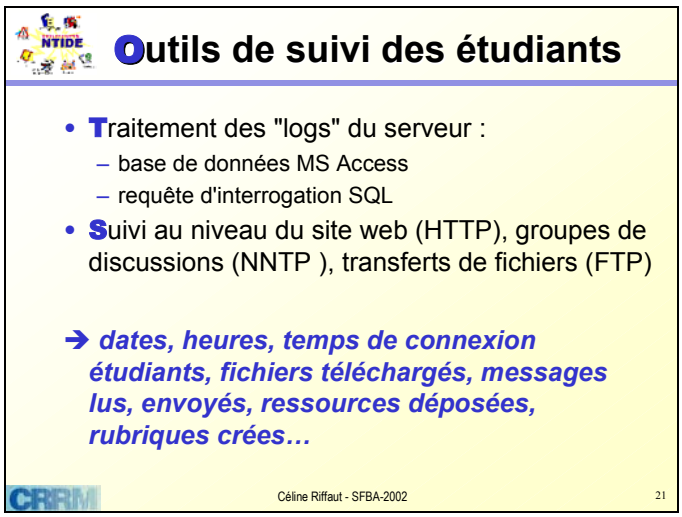


Figure3 : exemples d'outils de suivi des étudiants

L'accès via l'Internet au serveur de la maîtrise NTIDE, peut être un obstacle lorsqu'on utilise un simple modem avec une ligne téléphonique. En effet, dans ce cas la vitesse de transfert est assez lente et malgré le fait que les fichiers sont les plus compacts possibles, le coût d'accès peut être important. Nous conseillons donc, pour le confort de travail de prendre en compte les récentes propositions d'abonnement à l'Internet haut débit (ADSL) lorsque cela est possible. Malheureusement, le territoire français est mal desservi, ce qui impose des contraintes supplémentaires aux usagers[54].

Nous ne parlerons pas des droits du sport, où des formalités d'inscription qui doivent pour la cohérence du système ne pas être pris en compte ou être réalisé à distance. Il en va de même pour les accès aux bibliothèques universitaires, etc…

3 - LES LEVIERS

Ils sont très nombreux et de qualité (McGorry, 2002). En effet la possibilité pour les étudiants de bénéficier d'un apport de connaissance via les technologies modernes de l'Internet permet une économie de coût, un accès délocalisé à l'enseignement, une plus grande facilité pour gérer les acquis professionnels.

- Lorsque l'enseignement est complètement nouveau, ce qui est notre cas, les cours réalisés par les professeurs sont de leur propre création, ce qui permet de ne pas avoir de problème de copyright. Ceci n'est pas vrai dans toutes les disciplines, où les auteurs utilisent largement les publications, livres et cours d'autres personnes.

- Contrairement à ce que l'on croit, le travail est plus convivial dans un cours à distance que dans un cours présentiel. En effet, les groupes de travail mis en place (Forums) permettent un travail de groupe et la résolution de la majorité des problèmes par échanges entre les étudiants. Ce n'est que lorsqu'il y a un blocage, que les professeurs vont intervenir. Ce point est important, car il ne faut pas laisser de manière directe l'enseignant en frontal avec les étudiants. En effet la facilité d'envoyer un e-mail à propos de tout et de rien, placera l'enseignant devant un flux de demandes souvent inutiles, l'étudiant évitant de réfléchir devant la facilité de poser une question, de façon relativement impersonnelle, par rapport à la démarche présentielle.

- Sur le plan du matériel il y a plusieurs cas à considérer :
  - Le serveur de la maîtrise : il faut avoir un serveur indépendant de l'Université et surtout du centre informatique. Ceci permet de pouvoir intervenir comme l'on veut et quand bon vous semble, sans subir les contraintes de systèmes lourds et inadaptés.
  - La plate-forme de travail : elle doit permettre le développement de la création de connaissances. Nous recommandons de choisir la plate-forme en fonction des enseignements à dispenser. Il n'y a pas de plate-forme spécifique qui répondront à tous les problèmes, si non cela est au prix d'un alourdissement trop important. La figure 4 met en évidence sur l'écran de base du serveur (http://ntide.u-3mrs.fr) les outils mis à la disposition des étudiants. Seule la partie associée à une clé est privative. Les autres parties peuvent être consultées librement.

[54] Consulter à ce propos le rapport Jean Michel Yolin, http://www.yolin.net

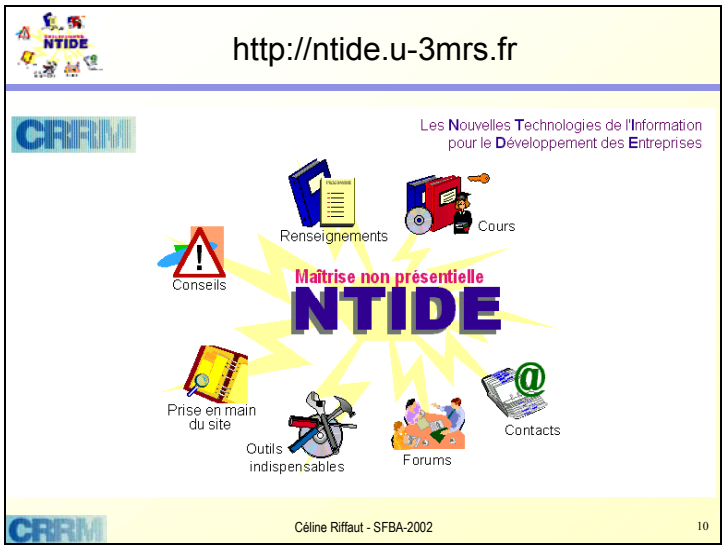


Figure 4 : les outils de la maîtrise

Les étudiants doivent avoir à disposition un ordinateur. Il est évident qu'en France le retard pris dans l'utilisation de l'Internet à grande vitesse (ADSL ou Câble), empêche le développement du WEB TV (serveur RealPlayer par exemple).

Le contrôle des connaissances doit être assuré de deux manières : par des exercices en cours d'années avec rendu via l'Internet, mais aussi par une série de questions (en présentiel en fin d'année), pour vérifier si l'étudiant a réellement travaillé de par lui-même. D'autre part, le mémoire réalisé en entreprise doit être rendu sous forme électronique, ce qui lors de sa présentation en présentiel permet aussi une autre série de contrôles.

4 – LE DEVELOPPEMENT PREVISIBLE

Un fait encourageant, pour ce type d'activité est que, sans publicité, nous assistons à une augmentation régulière des effectifs. Les étudiants sont répartis en France, dans les départements d'outre mer ainsi qu'à l'étranger (Canada, Portugal, Belgique). La figure 4 met en évidence la croissance des effectifs :

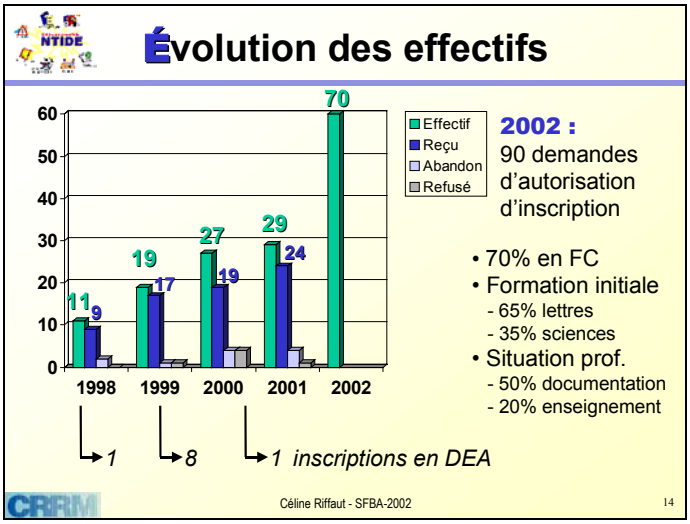


Figure 5 : évolution des effectifs de la maîtrise

Un enseignement à distance a des coûts qui ne sont pas pris en compte au plan du présentiel. Il faut entretenir un serveur, faire évoluer les cours (l'étudiant se rendant plus rapidement compte de l'obsolescence de certaines présentations), il faut gérer les groupes d'étudiants souvent de manière plus lourde qu'en présentiel. Cependant l'attitude de l'Université qui ne peut pas « organiquement » par modèle prendre en compte la totalité du problème en fonction des simples droits d'inscription, doit accepter le surcoût. C'est ce que nous faisons au CRRM et ce qu'a accepté l'Université Aix-Marseille III. De ce fait, une partie des ressources financières ainsi acquises permettent le paiement « d'assistants de bon niveau » pour l'encadrement des groupes. Mais alors, on quitte le système « classique » pour aller vers un système marchand (Whitty et Power, 2000), où l'étudiant, par le coût de son enseignement va devenir plus exigeant. Ceci introduit un esprit profondément différent de celui très conservateur de l'Université française. En outre avec l'évolution des enseignements vers les Masters, on ne sait pas si l'année équivalente à la maîtrise pourra toujours prendre en compte les surcoûts. Si cela n'était pas le cas, et si la contrepartie financière n'était pas octroyée par le Ministère, cela conduirait à la fermeture pure et simple de cet enseignement.

Cette fermeture ne voudrait pas dire que l'enseignement et l'expertise seraient perdus, en effet on aboutirait alors à un transfert de ces enseignements et de ces savoir-faire vers des universités sans doute étrangères qui prendraient ces diplômes en charge, dans la langue française, et avec les équivalents européens (cf la convention de Bologne[55]) et sans doute avec les mêmes enseignants travaillant à distance pour encadrer les étudiants, et ceci en dehors de leur service dans l'Université française, puisque tout se réalise à distance (Enguerand 2000). L'enjeu, sur le plan éducatif français est d'importance, on peut espérer que le défit sera relevé en temps utile.

Une autre alternative possible sera de faire passer ces enseignements dans le cadre d'Universités d'entreprises (Aizicovici 1997 et Posseme-Rageau 2000)[56], en les modifiant légèrement pour les rendre plus adaptés aux besoins de l'entreprise. Ce nouveau marché, souligné par de nombreux articles de presse

[55] "A Europe of Knowledge is now widely recognised as an irreplaceable factor for social and human growth and as an indispensable component to consolidate and enrich the European citizenship, capable of giving its citizens the necessary competences to face the challenges of the new millennium, together with an awareness of shared values and belonging to a common social and cultural space." 19 Juin 1999

[56] Lire à ce propos le chapitre du livre e-learning de Marc J. Rosenberg (2001), qui explique bien la différence entre le learning (apprendre) et le training (apprendre des instructions), et la balance qui doit constamment être ajustée entre ces deux aspects.

et par de nombreuses publications est sans doute un challenge à relever pour l'Université classique (Mills et Hrubetz, 2001).

CONCLUSION

Le travail à distance est une solution fiable et satisfaisante pour les étudiants qui par contraintes familiales ou professionnelles ou financières ne peuvent pas suivre les cours en présentiel. Par contre cet enseignement a un coût qui doit être pris en compte lors de l'inscription. Les achats de cours[57], la création d'applications, le secrétariat, les ordinateurs serveurs (duplication et sauvegardes), les ordinateurs portables des enseignants (qui peuvent ainsi intervenir de n'importe où), le maintien d'un accès à distance (RAS) via le RTC à plusieurs entrées sur le site du CRRM, la réalisation de CD-ROM, la constitution d'une base de ressources (http://crrm.u-3mrs.fr) de questions-réponses, des mémoires des étudiants…. sont des contraintes qui ne sont pas visibles dans les enseignements présentiels et qui induisent une logistique nouvelle à cet enseignement. Au niveau des enseignants, l'investissement est important est doit être pris en compte. Si non, on verra rapidement apparaître des enseignements de type privé qui feront largement appel aux spécialistes du public, sans que cela puisse faire l'objet d'un contrôle.

BIBLIOGRAPHIE

McGorry Susan Y (2002), « Online, but on target? Internet-based MBA courses. A case study », *Internet and Higher Education*, N°119, p. 1–9

Whitty Geoff, Power Sally (2000), « Marketization and privatization in mass education systems », *International Journal of Educational Development*, N°20, p. 93–107

Enguerand R. (2000), « Entreprises et nouvelle économie, les universités virtuelles bouleverseront les systèmes éducatifs », *Le Monde*, 3 juillet

Aizicovici F. (1997), « Les Universités d'entreprise favorisent l'intégration des cadres », *Le Monde Economie,* Enjeux et stratégies initiatives, 4 Février

Rosenberg Marc J. (2001), *E-learning*, McGraw-Hill

Posseme-Rageau G. (2000), « Réactivées, les universités d'entreprise permettront aux jeunes de se projeter dans l'avenir. Un projet de 43 millions de dollars », *Le Monde*, Campus, Economie, Stratégie d'entreprises, 16 mai, p. 25

Mills Andrew C., Hrubetz Joan (2001), « Strategic development of a master's program on the world wide web », *Journal of Professional Nursing* ,Vol. 17, N°4, p. 166-172

[57] Nous achetons par exemple le cours de bureautique au CRDP de Marseille, avec un exemplaire par étudiant inscrit. Le cours a été transformé en cours électronique pour les besoins de l'enseignement.

ANALYSE DU TRANSFERT DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE ENTRE LE SECTEUR PUBLIC ET LE SECTEUR PRIVE. ETUDE DES CO-PUBLICATIONS DANS LES REVUES SCIENTIFIQUES ESPAGNOLES.

Elea Giménez Toledo
E.Gimenez@cindoc.csic.es
Universidad de Navarra, Facultad de Comunicación, Pamplona 31080
Tel. (34)~(9)48-425617 - Ext. 2832
http://www.unav.es/fcom/

Adelaida Román Román
Adelaida@cindoc.csic.es
CINDOC-CSIC, C/ Pinar, 25. 3 stage, 28006 Madrid (Espagne)
Tel: (34)~(9)1 411 10 98 Ext. 249
http://www.cindoc.csic.es

Hervé Rostaing
Rostaing@crrm.u-3mrs.fr
CRRM, Université Aix-Marseille III, 13397 Marseille Cedex 20
Tel : (33)~4 91 28 87 46
http://crrm.u-3mrs.fr

Résumé : Ce travail expose comment les techniques bibliométriques peuvent répondre en partie à des démarches d'évaluation des échanges entre les acteurs du système Science-Technologie-Société. Cette étude bibliométrique s'est plus particulièrement focalisée sur l'évaluation des collaborations scientifiques par une analyse des co-publications entre les acteurs espagnols du secteur public (représentants la Science) et ceux du secteur privé (représentants la Technologie) pour la période 1995-1999. Le corpus étudié a été constitué à partir des bases de données ICYC et ISOC produites par le CINDOC. Ce corpus a été soumis plusieurs phases de codifications pour le préparer à l'analyse bibliométrique. Les références ont été codifiées en 8 secteurs scientifiques et techniques et les affiliations des auteurs ont été normalisées et agrégées selon 3 niveaux (organismes, catégories institutionnelles, catégories publiques ou privées). Quelques résultats bibliométriques représentatifs de chacun de ces 3 niveaux d'agrégation sont exposés. La méthode bibliométrique de l'analyse réseau a été employée pour la caractérisation de la régularité, de l'intensité et la structuration des échanges entre organismes. En conclusion, les résultats bibliométriques ont permis de quantifier les échanges entre les acteurs du secteur public et du secteur privé espagnols au niveau national ainsi que par secteur scientifique et technique ou par Région Autonome. Les techniques bibliométriques s'avèrent des outils très pertinents pour l'évaluation des échanges dans le système Science-Technologie-Société.

Abstract : This work shows how bibliometric techniques can partly answer the exchange evaluation among the actors of the Science-Technology-Society system. This bibliometric study was particularly focused on the evaluation of scientific collaborations by a co-authorship analysis of the Spanish actors from the public sector (representing Science) and from private sector (representing Technology) over the period 1995-1999. The studied dataset of references was collected from the databases ICYC and ISOC produced by the CINDOC. Coding processes to prepare the bibliometric analysis were applied to this dataset. The references were codified in 8 scientific and technical sectors and the affiliations of the authors were standardized and combined according to 3 levels (organizations, institutional categories, public or private categories). Some bibliometric results representing each of these 3 levels of aggregation are discussed. The bibliometric method of network analysis was used for the characterization of the regularity, the intensity and the structuring of the exchanges among organizations. In conclusion, the bibliometric results allow quantifying the exchanges among the actors of the public sector and the private sector in Spain at the national level, for scientific and technical sectors or for Autonomous Regions. The bibliometric techniques prove to be very relevant tools for the evaluation of the exchanges in the Science-Technology-Society system.

Mots-clés : Système Science-Technologie-Société / Echanges secteur public – secteur privé / Politique de la recherche / Analyse bibliométrique / Analyse des collaborations / Analyse des co-publications / Analyse réseaux / Espagne

Keywords : Science-Technology-Society System / Public–private sector exchanges / Research policy / Bibliometric analysis / Collaborations analysis / Co-authorship analysis / Networks analysis / Spain

PALABRAS CLAVE : SISTEMA CIENCIA-TECNOLOGIA-SOCIEDAD / INTERCAMBIOS SECTOR PÚBLICO–SECTOR PRIVADO / POLÍTICA LA INVESTIGACIÓN / ANÁLISIS BIBLIOMÉTRICO / ANÁLISIS DE LAS COLABORACIONES / ANÁLISIS DE COAUTORÍAS / ANÁLISIS REDES / ESPAÑA

Analyse du transfert de l'information scientifique et technique entre le secteur public et le secteur privé. Etude des co-publications dans les revues scientifiques espagnoles.

INTRODUCTION

Le processus d'innovation dans notre société s'expose bien souvent selon un système structuré en trois composantes : Science–Technologie–Société (Casillas Bueno et Roldán Salgueiro 1996, COTEC 1998, Quintanilla 1992). Le processus d'innovation est d'autant plus efficace que ces trois composantes sont parfaitement bien interfacées. Un élément important pour le succès de ce processus est que les acteurs impliqués dans ces trois composantes soient dynamiques. Mais surtout, c'est la collaboration active entre ces ensembles d'acteurs qui sera la clé de la réussite du processus d'innovation. Notre travail se focalise tout particulièrement sur l'élaboration d'outils d'évaluation des échanges entre les acteurs du système STS et plus particulièrement entre les acteurs impliqués dans la composante Science et dans la composante Technologie. Les méthodes d'évaluation des échanges entre acteurs permettent une meilleure compréhension des modèles d'organisation et des pratiques sous-jacents au processus d'innovation. Mieux comprendre et évaluer ces modèles est utile à la réflexion pour l'amélioration de l'articulation des échanges entre les différents acteurs et par-là même l'amélioration du système global. Toute cette démarche est fondée sur l'hypothèse initiale suivante qui suppose que les résultats produits par le système STS répondent en général à des besoins de progrès dans notre société. Ainsi rendre le système STS plus efficace et plus performant permettrait une accélération des progrès dans notre société.
Dans ce contexte, nous exposerons comment les techniques bibliométriques peuvent répondre en partie à des démarches d'évaluation du système STS. Pour étayer cette réflexion, une application des techniques bibliométriques pour l'évaluation des échanges entre les acteurs espagnols du système STS est présentée. Cette étude bibliométrique s'est plus particulièrement focalisée sur l'évaluation des collaborations scientifiques entre les acteurs espagnols du secteur public et ceux du secteur privé.

LA BIBLIOMETRIE ET L'EVALUATION DES ECHANGES DANS LE SYSTEME STS

Les acteurs impliqués dans le système STS du processus d'innovation peuvent être regroupés en 5 principales catégories (Figure 2).
Trois catégories sont tout particulièrement impliquées dans la création même des progrès de la société :

- les acteurs impliqués dans des actions de recherche
- les acteurs impliqués dans des actions de production
- les acteurs impliqués dans des actions de diffusion des résultats de la production au sein de la société

Alors que les deux dernières catégories peuvent être plutôt considérées comme des acteurs d'aide et de soutien au processus d'innovation

- les acteurs des organismes d'interface
- les acteurs des administrations publiques

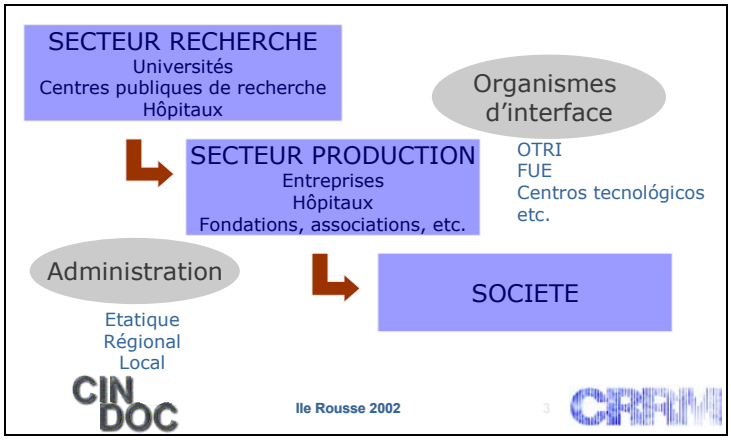


Figure 2 : les acteurs du système Science-Technologie-Société

Le bon fonctionnement d'un tel système dépend essentiellement de l'activité des échanges entretenus entre ces différents acteurs. Un des échanges clé pour la pérennité à long terme de l'ensemble des acteurs est le transfert des informations et des connaissances entre tous ces acteurs.
L'évaluation du transfert des informations et des connaissances entre ces acteurs fait l'objet d'un intérêt tout particulier parmi les différents organismes nationaux et internationaux (Commision of the European Communities 2001, COTEC 2001, OCDE 2001). Les méthodes usuellement employées ont pour objet l'élaboration de tableaux d'indicateurs d'évaluation des systèmes d'innovation. Ceux-ci évaluent les interactions entre les acteurs à partir d'indicateurs mesurant les « inputs », les « outputs » ou le système STS en lui-même. La bibliométrie est une méthode parmi d'autres permettant la mesure du transfert des informations entre ces acteurs à partir d'indicateurs de type « output ».
Différentes approches bibliométriques peuvent être envisagées de manière à analyser ces échanges selon des objectifs variés. Par exemple, nous pouvons citer les techniques bibliométriques suivantes : analyse des co-publications, analyse des citations entre travaux scientifiques, analyse des co-dépôts de brevets, analyse des citations entre brevets et analyse des citations entre brevets et travaux scientifiques.
Par rapport à d'autres approches d'évaluation, ces méthodes bibliométriques offrent des mesures quantitatives permettant de qualifier les échanges pour un très grand nombre d'acteurs.

OBJECTIF DE L'ETUDE BIBLIOMETRIQUE

L'étude bibliométrique exposée dans cette communication fait partie d'une étude plus complète abordant plusieurs méthodes d'évaluation des échanges entre les acteurs du système STS. Cette étude complète est décrite dans le mémoire de thèse de Giménez Toledo (2002). Nous n'exposerons dans cet article que l'analyse des co-publications entre un nombre restreints d'acteurs du système STS. Cette étude des co-publications a pour objectif l'évaluation des collaborations scientifiques entre les acteurs espagnols du secteur public et ceux du secteur privé pour la période 1995-1999.
Dans le système STS espagnol, les principaux acteurs de la composante *Recherche* appartiennent à des centres de recherche publics alors que les acteurs de la composante *Production* sont essentiellement dans des organismes privés. L'étude bibliométrique des co-publications scientifiques entre les acteurs du secteur public et ceux du secteur privé permet de donner une assez bonne estimation de l'activité des échanges des informations et des connaissances entre la composante *Recherche* et la composante *Production* du système STS pour l'Espagne.

Cette étude tente de répondre tout particulièrement aux interrogations suivantes :

- Les échanges entre les acteurs du secteur public et du secteur privé sont-ils bien réels en Espagne ?
- Peut-on le quantifier ?
- Ces échanges sont-ils semblables dans les différents secteurs scientifiques et techniques ?
- Peut-on caractériser la régularité de ces échanges ?

METHODOLOGIE

L'étude des co-publications entre les acteurs espagnols est obtenue à partir des références bibliographiques issues des bases de données ICYC[58] et ISOC[59] produites par le CINDOC[60]. Ces deux bases sont très certainement les meilleures sources pour obtenir une bonne représentation de l'activité scientifique espagnole.
L'étude ne s'est pas portée sur l'ensemble des références bibliographiques correspondant aux articles publiés entre 95 et 99. Seules les publications se rapportant à des secteurs scientifiques ou techniques privilégiés[61] ont fait l'objet de cette évaluation. Les 7 secteurs étudiés prioritairement sont les suivants : Chimie, Pharmacie, Alimentation, Mécanique, Électricité et Électronique, Bâtiment et travaux publics, et Télécommunications.
L'extraction des articles relevant de ces secteurs privilégiés est fondée sur l'identification des revues espagnoles les plus représentatives de chacun de ces secteurs. Les références des articles publiés par ces revues entre 95 et 99 constituent le corpus étudié. Le Tableau 2 de l'Annexe 1 présente le nombre de revues pris en compte pour chacun de ces 7 secteurs. Il est à noter qu'un huitième secteur a été ajouté aux 7 précédents rassemblant tous les

[58] ICYC = base de données spécialisée en Science et Technologie. Base de données équivalente à la base Pascal mais restreinte à la production espagnole rédigée en langue espagnole

[59] ISOC = , base de données spécialisée en Sciences Sociaux et Sciences Humaines. Base de données équivalente à la base Francis mais restreinte à la production espagnole rédigée en langue espagnole.

[60] CINDOC = Centre d'Information et Documentation Scientifique de l'Espagne. Centre équivalent à l'INIST en France.

[61] Ces secteurs ont été déterminés comme étant les plus porteurs économiquement pour la région de Madrid. La Région Autonome de Madrid étant commanditaire de cette étude, il fallait avant tout offrir une parfaite évaluation des domaines économiques prioritaires à cette région.

articles publiés dans des revues spécialisées en Economie et entreprise. Il s'est avéré nécessaire de rajouter cette nouvelle catégorie même si elle ne correspond pas un secteur économique réel car les revues en Économie&Entreprise publient très souvent des travaux appliqués qui ont plus de chances d'avoir été obtenus à la suite d'une collaboration entre le secteur public et le secteur privé que des travaux de la recherche fondamentale.
Le nombre d'articles publiés par les 72 revues de ces 8 secteurs s'élève à 10 832. Ces articles sont rédigés par 23 176 auteurs différents issus de 12 913 institutions différentes.
Ce corpus se réduit presque de moitié lorsqu'on ne considère que les articles mettant en jeu une collaboration entre au moins 2 individus. Seul 5 513 articles sont rédigés par au moins deux auteurs. Ceci indique que près de la moitié des publications sélectionnées initialement correspond à des recherches résultant d'un travail solitaire d'un acteur espagnol. Cette très forte proportion dévoile très nettement que les acteurs espagnols ne sont pas enclins à la collaboration scientifique.

L'objet de l'étude n'étant pas de se limiter à l'évaluation des collaborations entre les individus mais plus précisément entre les organismes et surtout entre les organismes publics et les organismes privés, un très gros travail de codification des affiliations des auteurs a été entrepris. Les affiliations des 5 513 références ont été normalisées. Lors de cette codification, les affiliations ont aussi été classées selon 12 grandes catégories institutionnelles (voir Tableau 3 en Annexe 2). Cette codification a permis de restreindre à nouveau le corpus analysé aux seules publications mettant en jeu des collaborations entre 2 affiliations distinctes. L'évaluation des échanges ne devant porter que sur des collaborations entre personnes relevant d'organismes différents, ce laborieux travail de codification était indispensable pour garantir la qualité des résultats.
Le corpus des travaux issus d'une collaboration entre au moins 2 organismes différents n'est constitué plus que de 1455 références de publications. Ce qui signifie que 74% des travaux rédigés par au moins deux auteurs sont le fait d'un travail au sein d'un même organisme. La Figure 3 montre bien la part très faible des travaux issus d'une collaboration entre organismes différents par rapport à l'ensemble des travaux espagnols. Non seulement les acteurs espagnols n'ont pas tendance à collaborer entre eux mais de plus lorsqu'ils le font, il est assez rare que les collaborations dépassent les murs de leurs propres organismes.

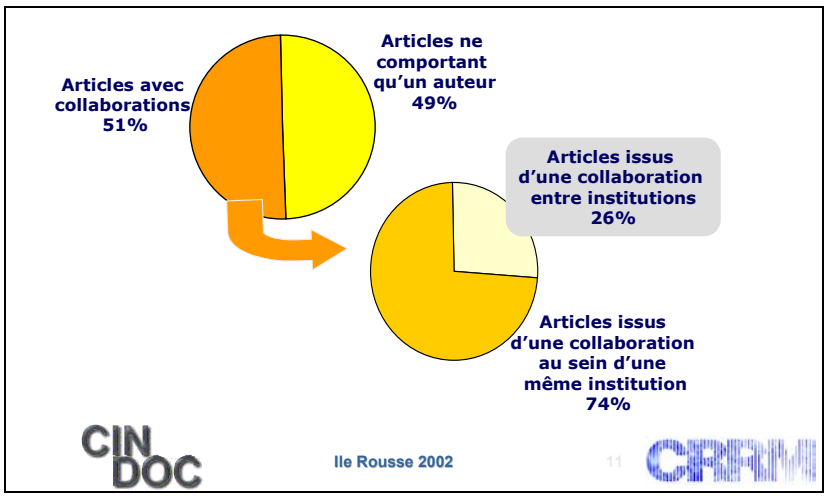


Figure 3 : Processus de filtrage des travaux issus d'une collaboration entre organismes

Les 1 455 références étudiées comportent néanmoins 3 341 organismes différents. Ce qui laisse pressentir un très grand saupoudrage des collaborations entre une très grande variété d'organismes espagnols. Les 3341 organismes étant classés en grandes catégories institutionnelles, un troisième niveau de codification des organismes permettant de préciser ceux qui relèvent du secteur public ou privé a été facilement réalisé.

La Figure 4 reprend le synopsis complet du protocole de sélection et de codification du corpus étudié en définitive.

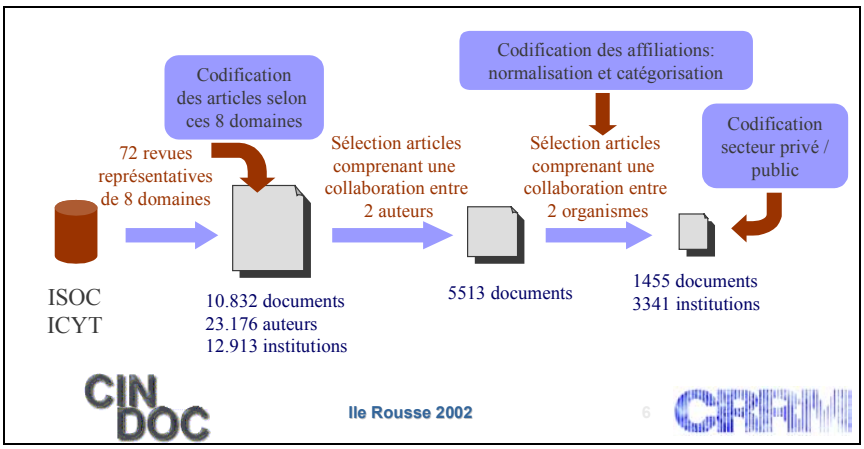


Figure 4 : Synopsis de la sélection et de la codification du corpus analysé

PRINCIPAUX RESULTATS

Les 3 niveaux de codification des affiliations ont permis l'élaboration d'indicateurs selon 3 niveaux d'agrégation des données. Nous exposerons dans ce paragraphe quelques résultats représentatifs de chacun de ces 3 niveaux d'agrégation. Nous commencerons par des résultats offrant une vision « macro » des collaborations entre le secteur public et le secteur privé, puis une vision « méso » sera exposée par l'analyse des collaborations entre catégories institutionnelles et pour terminer une analyse « micro » fournira des renseignements sur les organismes impliqués dans ces collaborations.

Evaluation macro : collaborations entre le secteur public et privé espagnol

Cette évaluation macro est basée sur l'analyse des collaborations selon 3 catégories :

- collaboration entre organismes du secteur public
- collaborations entre organismes du secteur privé
- collaborations entre au moins un organisme public et un organisme privé.

La Figure 5 décrit la part de ces 3 catégories de collaborations dans le corpus des 1455 articles issus des revues espagnoles. La majorité des collaborations (60%) correspond à des travaux effectués entre des organismes du secteur public. En seconde position viennent les collaborations entre secteur public et secteur privé avec une part représentant 24% des collaborations. Et finalement, les collaborations entre des organismes du secteur privé ne représentent plus que 9%. On peut noter que 7% des collaborations n'a pu être attribué à une de ces 3 catégories de collaborations par manque de renseignements sur les affiliations des auteurs.
A en juger par ces premiers résultats, la part des collaborations entre le secteur privé et le secteur public en Espagne est loin d'être négligeable : près d'un quart des collaborations entre organismes espagnols (pour les secteurs évalués par cette étude).

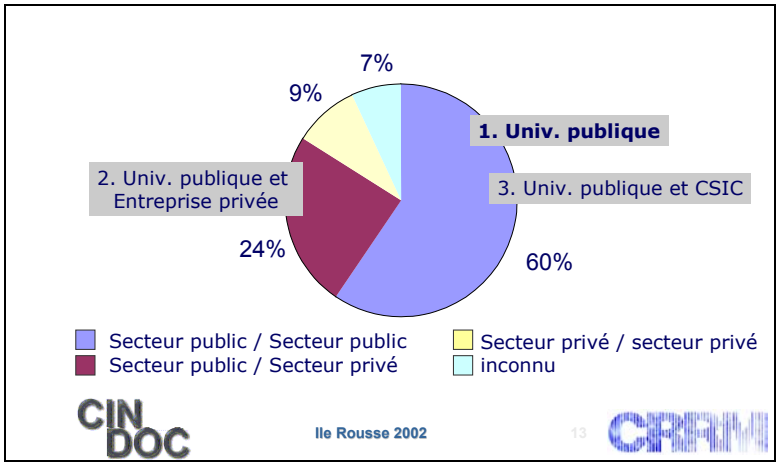


Figure 5 Part des collaborations entre le secteur public et le secteur privé

La Figure 6 présente la répartition des collaborations pour les 8 secteurs scientifiques et techniques étudiés selon ces 3 catégories de collaborations. Cette figure permet de caractériser la nature des collaborations pour chaque secteur. Elle met en évidence 4 grandes familles de comportements de collaborations :

- Pharmacie, Alimentation, Economie&Entreprise : dans ces secteurs les collaborations s'effectuent essentiellement entre des organismes publics (60 à 80% des collaborations) ; il y a peu de collaborations entre le secteur public et privé (17 à 23%) et pratiquement pas entre des organismes privés.
- BTP, Télécommunication : ces deux secteurs maintiennent une proportion équilibrée entre les collaborations internes au secteur public et les collaborations impliquant le secteur privé avec une part très faible des collaborations entre organismes privés (inférieur à 12%)
- Electricité&Electronique, Chimie -> les organismes du secteur privé sont majoritairement présents (supérieur à 50%) dans les collaborations soit dans des collaborations entre secteur public et privé soit exclusivement entre organismes privés. Ce dernier type de collaborations est assez important (de 18 à 20%).
- Mécanique -> ce secteur a un comportement très atypique puisque les collaborations entre organismes du privé prennent une part très importante (31%) presque égale aux collaborations entre organismes du public (45%) et bien supérieur aux collaborations entre secteur public et privé (17%).

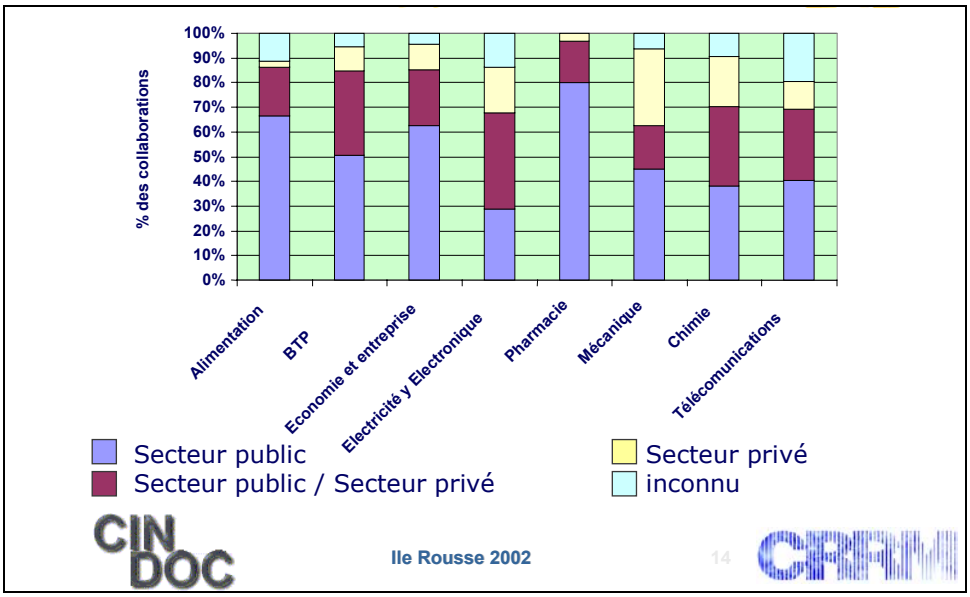


Figure 6 : Répartition des type de collaborations secteur public / secteur privé par secteur

Evaluation méso : collaborations entre les grandes catégories institutionnelles espagnoles

L'évaluation méso cherche à connaître les collaborations entre les grandes catégories institutionnelles espagnoles (voir Annexe 2).

Sur la Figure 5, les trois plus fréquentes collaborations entre catégories institutionnelles ont été rajoutées selon leur appartenance aux trois 3 catégories de collaboration secteur public / secteur privé. Ainsi, les trois plus fréquentes collaborations entre catégories institutionnelles sont :

- en première position les collaborations entre les Universités publiques (collaboration du type secteur public / secteur public)
- en seconde position les collaborations entre une Université publique et une Entreprise privée (collaboration du type secteur public et privé)
- et en troisième position les collaborations entre une Université publique et le CSIC[62] (collaboration du type secteur public / secteur public).

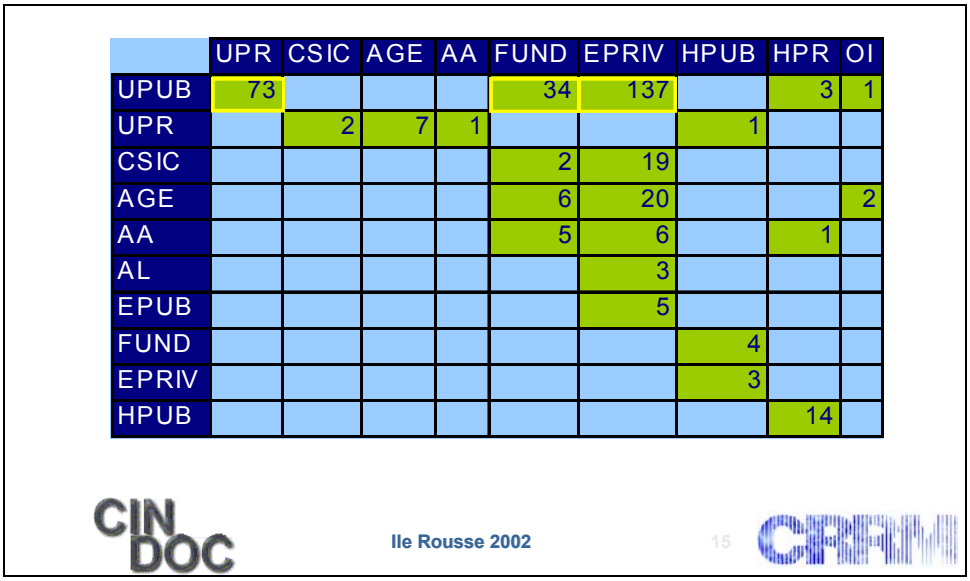
Comme l'étude s'intéresse plus particulièrement aux collaborations entre secteur public et privé, les documents contribuant à ce type de collaborations ont fait l'objet d'une analyse plus fine. Cette analyse cherche à connaître les principales catégories intitutionnelles qui participent aux collaborations entre le secteur public et privé.

[62] CSIC = Consejo Superior de Investigaciones Científicas, équivalent espagnol du CNRS français.

Le Tableau 1 répertorie l'ensemble de ces collaborations en les ventilant selon les différentes catégories institutionnelles qui y participent.

Ce tableau montre que de très loin les collaborations prédominantes sont celles rapprochant des Universités publics à des entreprises privés (39%). Viennent ensuite les collaborations entre des Universités publiques et les Universités privés (21%) et bien plus loin, avec seulement (10%), les collaborations entre Universités publiques et associations & fondations.

Ces résultats de l'analyse méso mettent en évidence le rôle central des Universités publiques espagnoles dans les collaborations entre le secteur public et le secteur privé. Il devient intéressant de savoir quelles Universités publiques espagnoles sont les plus actives dans ces collaborations. L'analyse micro permet d'identifier ces acteurs principaux.



	UPR	CSIC	AGE	AA	FUND	EPRIV	HPUB	HPR	OI
UPUB	73				34	137		3	1
UPR		2	7	1			1		
CSIC					2	19			
AGE					6	20			2
AA					5	6		1	
AL						3			
EPUB						5			
FUND							4		
EPRIV							3		
HPUB								14	

Tableau 1 : Collaborations entre secteur public et privé réparties par catégories institutionnelles (voir Annexe 2 pour connaître la signification des abréviations)

Evaluation micro : collaborations entre les organismes espagnoles

Dans une étude micro des collaborations, deux indicateurs sont souvent employés pour identifier les acteurs ayant un rôle important dans l'ensemble des collaborations. Le premier prend en compte l'ouverture et le second l'intensité des collaborations. L'indicateur d'ouverture informe sur le nombre d'affiliations avec lesquels un organisme collabore tandis que l'indicateur d'intensité renseigne sur le nombre de collaborations effectuées entre deux organismes. L'indicateur d'ouverture estime la capacité qu'un organisme a à collaborer avec une très grand nombre d'acteurs. L'indicateur d'intensité permet d'évaluer si ces collaborations sont récurrentes ou sans suite et donc permet d'estimer les collaborations qui contribuent à un processus continu et constructif d'échanges des collaborations peu fructueuses.

L'étude micro exposée ici se fonde sur l'analyse réseau obtenue par la chaîne de traitement Dataview[63]-Matrisme[64]. Le lecteur pourra se reporter à la thèse d'Eric Boutin (1999) pour une description plus détaillée de l'analyse réseau et du logiciel Matrisme, mais aussi à plusieurs travaux montrant l'application de cette méthode à divers domaines : l'analyse des réseaux d'auteurs (Boutin, 1996), l'analyse de syntagmes (Rostaing et al, 1998), l'analyse de liens entre sites web (Rostaing, 2001). Les analyses réseau obtenues par le logiciel Matrisme expriment parfaitement bien l'indicateur d'ouverture et l'indicateur d'intensité sous forme graphique. Dans le cadre de notre étude, les graphes de réseaux construits par ce logiciel livrent bien d'autres informations. Ils

[63] Dataview est un logiciel bibliométrique qui a été développé par le laboratoire CRRM : consulter http://crrm.u-3mrs.fr/commercial/software/dataview/dataview.html.

[64] Le logiciel Matrisme a été développé à l'issu d'une collaboration entre le laboratoire CRRM et le laboratoire LEPONT (http://lepont.univ-tvt.fr) : consulter http://crrm.u-3mrs.fr/commercial/software/matrisme/matrisme.html

fournissent une vue globale de l'ensemble des collaborations où se dégage, sous forme de groupes d'organismes plus ou moins réticulés, la structure de ces collaborations. Pour chaque organisme, ces graphes renseignent non seulement sur le degré d'ouverture d'un organisme et sur les intensités des collaborations mais aussi sur l'implication de cet organisme dans la construction du tissu des collaborations.

Seuls cinq graphes sont présentés en exemple. Les trois premiers sont issus d'une série de graphes réalisés pour l'analyse des collaborations entre les organismes espagnoles appartenant au même secteur scientifique et technique. Seuls les réseaux de collaborations du secteur Electricité&Electronique (Figure 7), du secteur Chimie (Figure 8) et du secteur BTP (Figure 9) sont reproduits. Les deux autres graphes sont issus d'une série de réseaux réalisés pour comparer les tissus des collaborations pour chaque Région Autonome espagnole. Seuls les réseaux de collaborations pour la Région de Catalogne (Figure 10) et pour la Région de Madrid (Figure 11) sont présentés.

Les analyses réseau exposées ci-dessous ne représentent pas tous les organismes espagnols impliqués dans des collaborations. Certains organismes ne sont pas représentés en tant que tels mais sont considérés sous la forme agrégée de leur catégorie institutionnelle. Ainsi, les organismes appartenant aux Universités publiques, Centres de recherche publics, Administrations nationales sont représentés individuellement alors que les organismes appartenant aux catégories Universités privées, Entreprises privées, Administrations régionales, Administrations locales, Associations&fondations, Hôpitaux privés, Hôpitaux publics et Instituts étrangers sont représentés par leur catégorie institutionnelle. Comme l'identification de l'origine géographique de ces organismes est importante, surtout s'ils sont regroupés sous leur catégorie institutionnelle, il a été décidé d'associer systématiquement le nom de l'organisme (ou de la catégorie institutionnelle) à la ville espagnole ou étrangère où l'organisme est implanté.

Etude comparative des réseaux de collaborations par secteur (Figure 7, Figure 8, Figure 9) :
Ces organismes sont représentés par les nœuds d'un réseau de collaborations et prennent la forme d'une « boite » dans les graphes. Le texte inscrit dans cette boite correspond au nom de l'organisme (ou à la catégorie institutionnelle) suivi de la ville espagnole ou de la ville étrangère. Le code couleur de la boite donne une indication sur le nombre total de collaborations auxquelles l'organisme a contribuées dans le corpus complet. Les organismes ayant participés à une ou deux collaborations sur la période 95-99 se trouvent représentés par des boites jaunes. Ceux qui ont contribués à 3 ou 4 collaborations sont en vert et ceux qui ont participés au moins à 5 collaborations sont en bleu. Le nombre exact de travaux effectués en collaboration est précisé dans la boite à la suite du nom de l'organisme. Pour l'analyse des collaborations par secteur, seuls les organismes ayant contribués au moins à trois collaborations ont été représentés dans les graphes. Les organismes n'ayant contribués qu'à une ou deux collaborations sur le période de cinq ans étudiés (95-99) ont été considérés comme étant des acteurs mineurs.

Les collaborations entre les organismes sont représentées par des arcs entre les nœuds du réseaux. Trois codes couleur sont utilisés pour indiquer l'intensité de la collaboration entre deux nœuds. Cette intensité est mesurée selon la fréquence de cette collaboration. Ainsi, les arcs de couleur rouge-foncé indiquent que les deux organismes ont au moins 4 travaux en communs, les arcs de couleur bleu-clair représentent des collaborations se reproduisant deux ou trois fois et les arcs de couleur orange des collaborations n'ayant eu lieu qu'une seul fois. Ces derniers arcs n'ont pas été dessinés dans les graphes pour améliorer la lisibilité de ces réseaux.

Les graphes obtenus pour les trois secteurs choisis ont des physionomies bien différentes ce qui laisse penser que ces 3 secteurs n'ont pas la même pratique des collaborations et des échanges entre les universités et le secteur privé.

Le secteur Electricité&Electronique (Figure 7) est très peu réticulé. Trois petits groupes indépendants composés seulement de deux organismes se dégagent. Un groupe un peu plus important composé de six organismes se présente sous la forme d'un chaînage de collaborations. On peut estimer que les organismes dans ce secteur ont un degré d'ouverture très faible et que très peu d'organismes entretiennent des collaborations récurrentes. Dans ce secteur, les universités clé pour la participation aux échanges avec le secteur privé sont : l'*Université Polytechnique de Catalogne*, l'*Université de Deusto* et l'*Université Polytechnique de Madrid*. Ce graphe montre aussi que l'*Université d'Oviedo* et l'*Université de Cordoue* ont des relations privilégiées et soutenues avec des institutions étrangères.

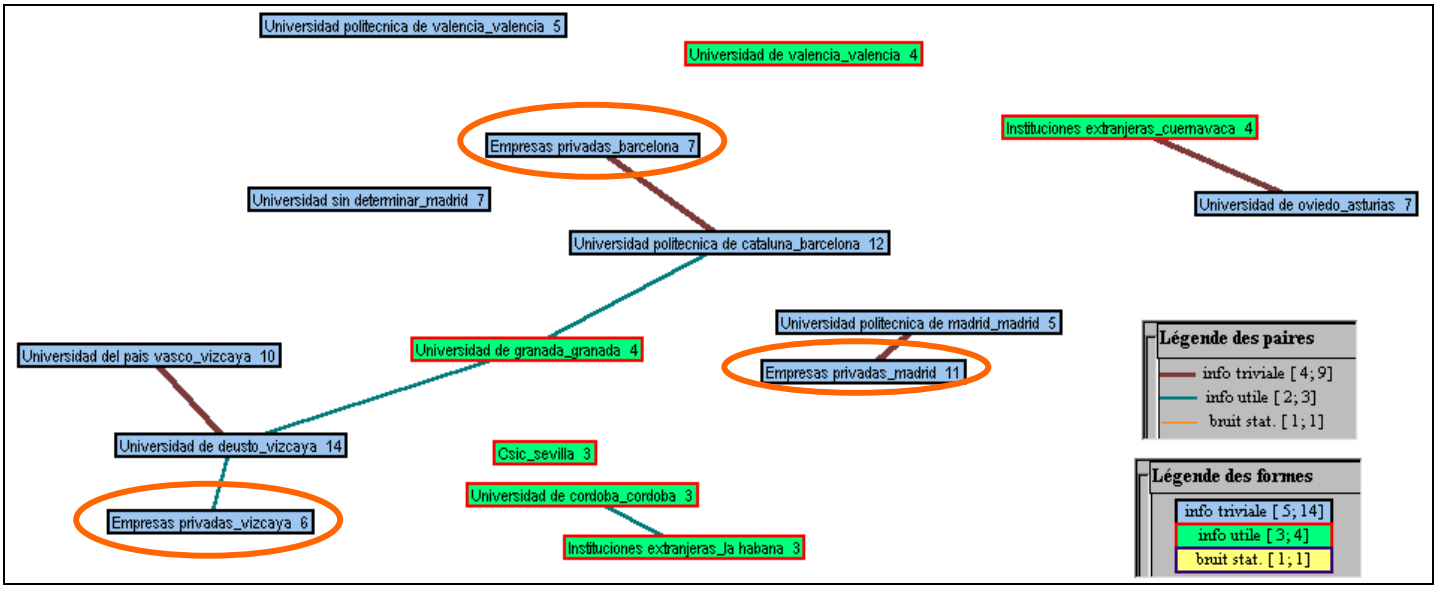


Figure 7 : Réseau des collaborations entre les organismes du domaine Électricité&Électronique

Le secteur Chimie (Figure 8) paraît avoir un comportement assez similaire à celui du secteur Electricité&Electronique dans la pratique des collaborations. Les organismes n'ont pas un degré d'ouverture important et peu d'organismes sont fidèles dans leurs collaborations. Il faut tout de même noter la particularité de la composition du principal groupe de collaborations. Il est structuré autour d'un noyau de collaborations entre entreprises de régions différentes. Sur ce noyau viennent se greffer l'*Université Polytechnique de Catalogne* et l'*Université du Pays Basque* ainsi que des institutions étrangères. Ces deux universités collaborent prioritairement avec les entreprises de leur ville. Par ailleurs, l'*Université A Coruña* participe de façon isolée aux transferts des informations avec les entreprises de sa ville.

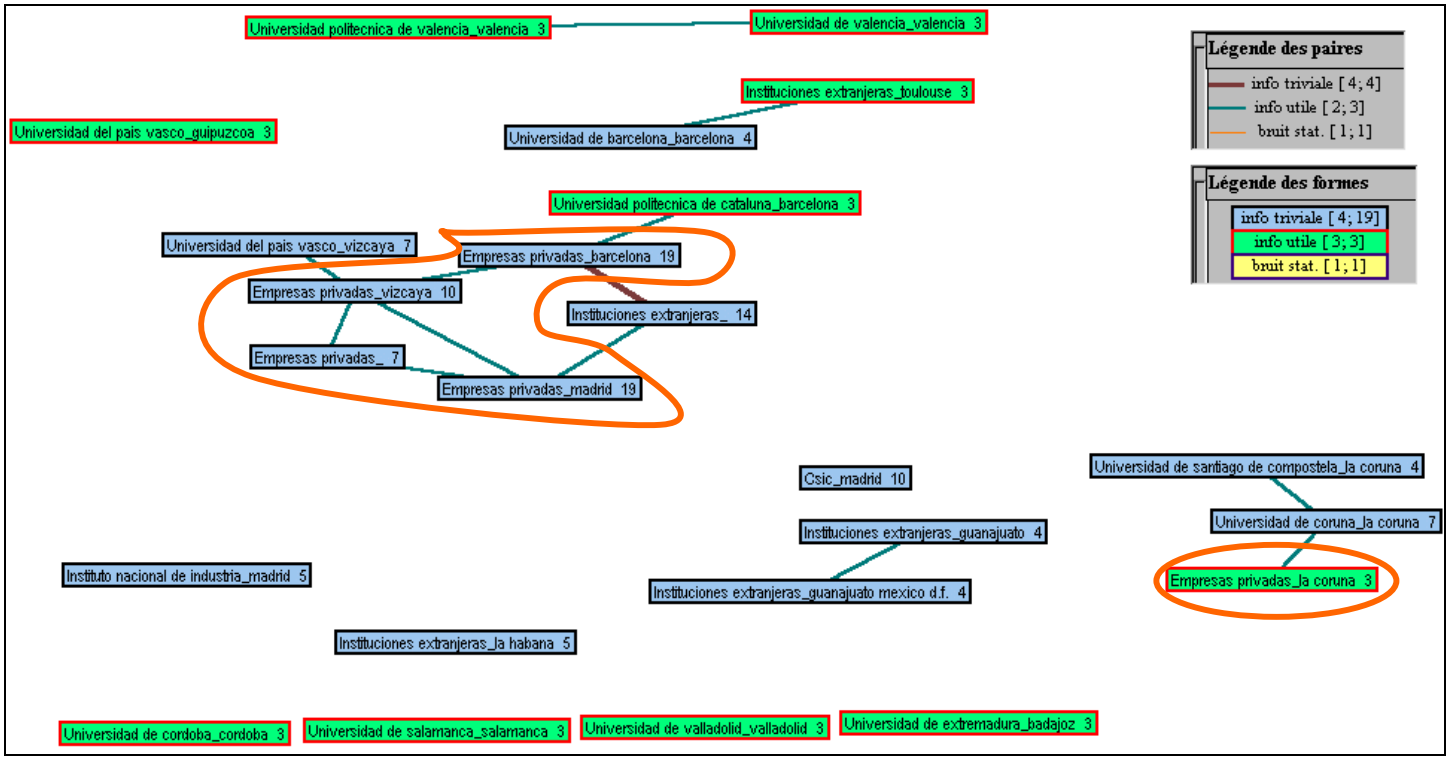


Figure 8 : Réseau des collaborations entre les organismes du domaine Chimie

Le réseau des collaborations du secteur BTP (Figure 9) a une structure bien différente des deux précédentes. Le graphe est très fortement réticulé. Le degré d'ouverture à la collaboration est bien plus important que pour les deux secteurs précédents ce qui n'empêche pas que des collaborations récurrentes s'établissent entre les acteurs.

Ce graphe dévoile une grande spécificité au secteur BTP dans la structuration des collaborations. Les institutions des Administrations Nationales sont très fortement impliquées dans les échanges effectués entre les universités et le secteur privé. Le *Ministère des Travaux Publics* est un centre névralgique dans la structuration des collaborations et plus particulièrement ses représentants dans les Régions de Madrid et d'Asturias. Le *Ministère de l'Environnement* et ses antennes à Valence et à Madrid ainsi que l'*Institution National de l'Industrie* (ancienne appellation) à Madrid contribuent eux aussi à l'animation des collaborations dans ce secteur. Les universités qui participent activement à des échanges avec des entreprises sont principalement l'*Université Polytechnique de Catalogne*, l'*Université Polytechnique de Madrid*, le *CSIC* et l'*Université de Grenade*.

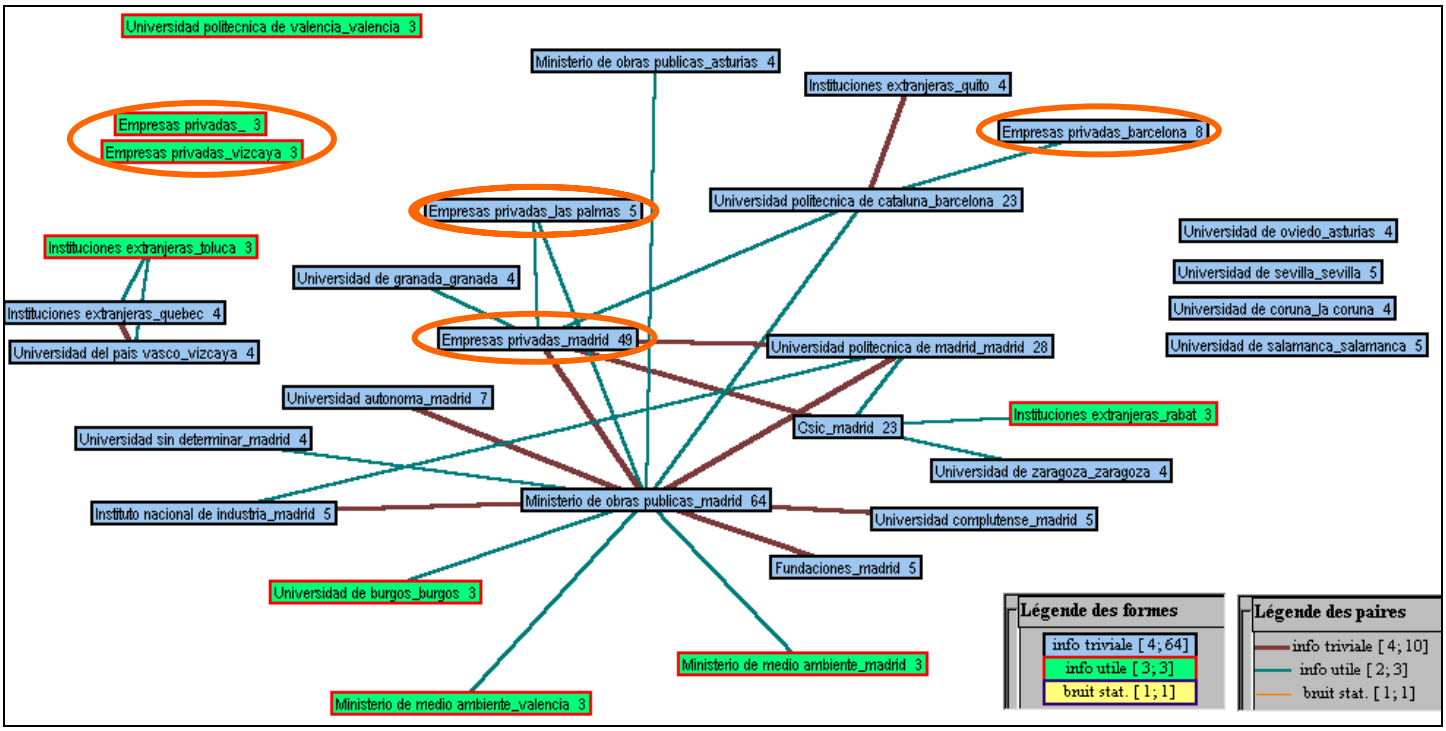


Figure 9 : Réseau des collaborations entre les organismes du domaine BTP

Etude comparative des réseaux de collaborations par Région Autonome espagnole :

Dans le souci d'évaluer le dynamise et l'organisation des collaborations par Région Autonome ainsi que pour identifier les acteurs clé dans ces collaborations, une analyse comparative des réseaux de collaborations par Région espagnole a été entreprise. Uniquement deux graphes sont présentés ici : les réseaux de collaborations pour les organismes de la Région Autonome de Catalogne (Figure 10) et pour les organismes de la Région Autonome de Madrid (Figure 11).

Pour ces deux représentations graphiques, tous les organismes de chaque région ont été pris en considération quel que soit le nombre de travaux auxquels ils ont contribués. Toutes les collaborations entre organismes ont été dessinées sur le graphe de la Région de Catalogne tandis que pour la Région de Madrid les collaborations ne s'étant produites qu'une seule fois ont été supprimées de façon à améliorer la lisibilité du réseau. Ainsi, ces graphes restituent tous les acteurs présents dans le corpus et pratiquement toutes les collaborations entre ces acteurs pour ces deux régions.

Ces deux graphes permettent d'identifier non seulement les universités fortement impliquées dans les transferts des connaissances avec le secteur privé mais aussi les organismes publics qui soutiennent et accompagnent ces transferts.

Dans le cas de la Région Autonome de Catalogne (Figure 10), les organismes qui structurent le tissu des collaborations dans sont à la fois des universités (*Université Polytechnique de Catalogne, Université de Barcelone, Université Pompeu Fabra*), mais aussi des hôpitaux (publics et privés), les *CSIC* et l'*Administration Régionale Autonome*. Deux principales sous-structures se dégagent, l'une composée des hôpitaux de la région (ensemble marron sur la figure), l'autre composée des universités de la région (ensemble bleu sur la figure). Il faut noter que les administrations et les institutions jouent le rôle d'interface entre ces deux sous-ensembles.

Pour la Région Autonome de Madrid (Figure 11), les organismes qui contribuent le plus à la structure des collaborations sont principalement des universités et des centres de recherche publics (ensemble bleu sur la

figure). Le *Ministère des Travaux Publics,* l'*Institut National de l'Industrie* et quelques fondations complètent l'animation de l'organisation des collaborations.

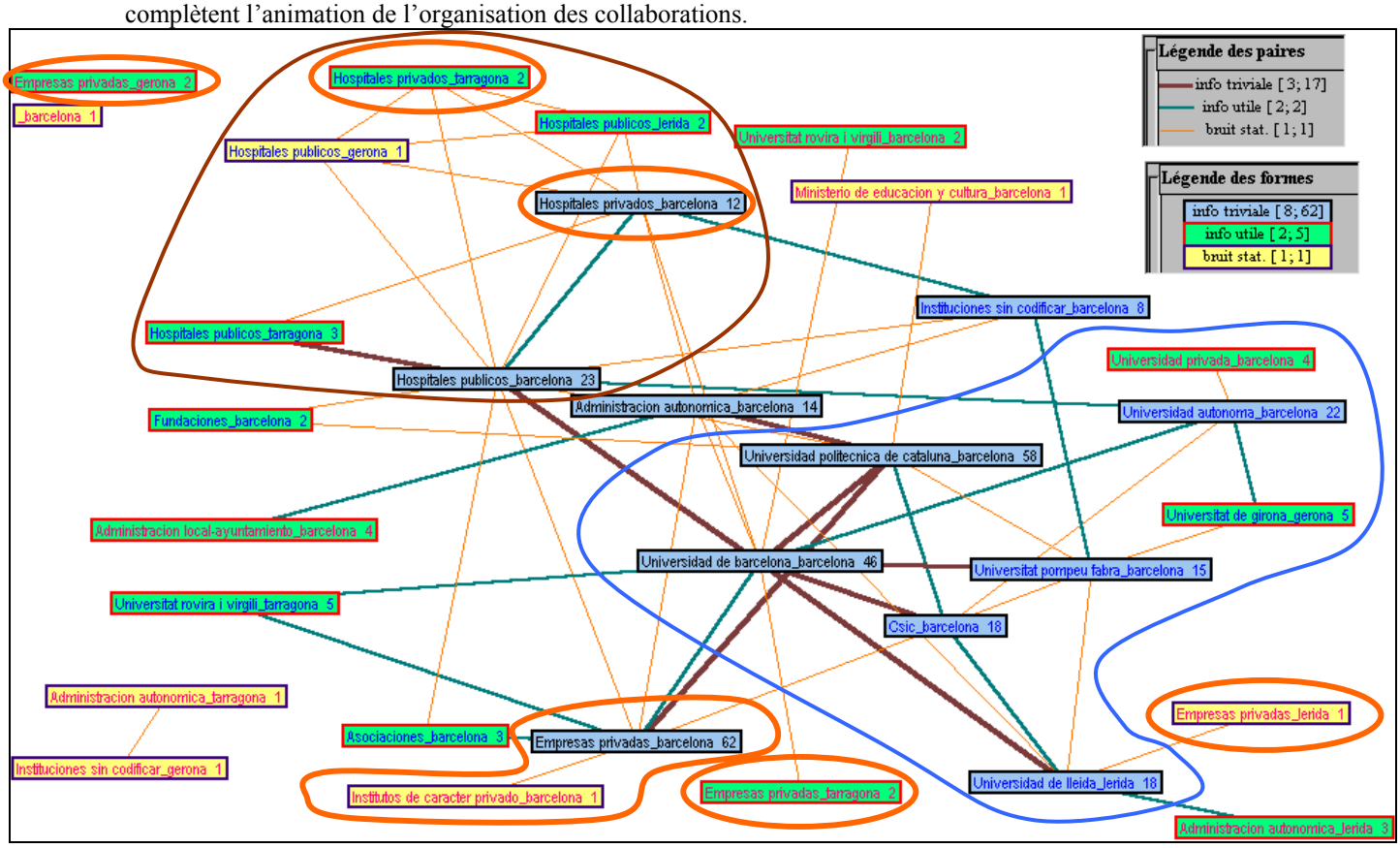


Figure 10 : Réseau des collaborations entre les organismes de la Région Autonome de Catalogne

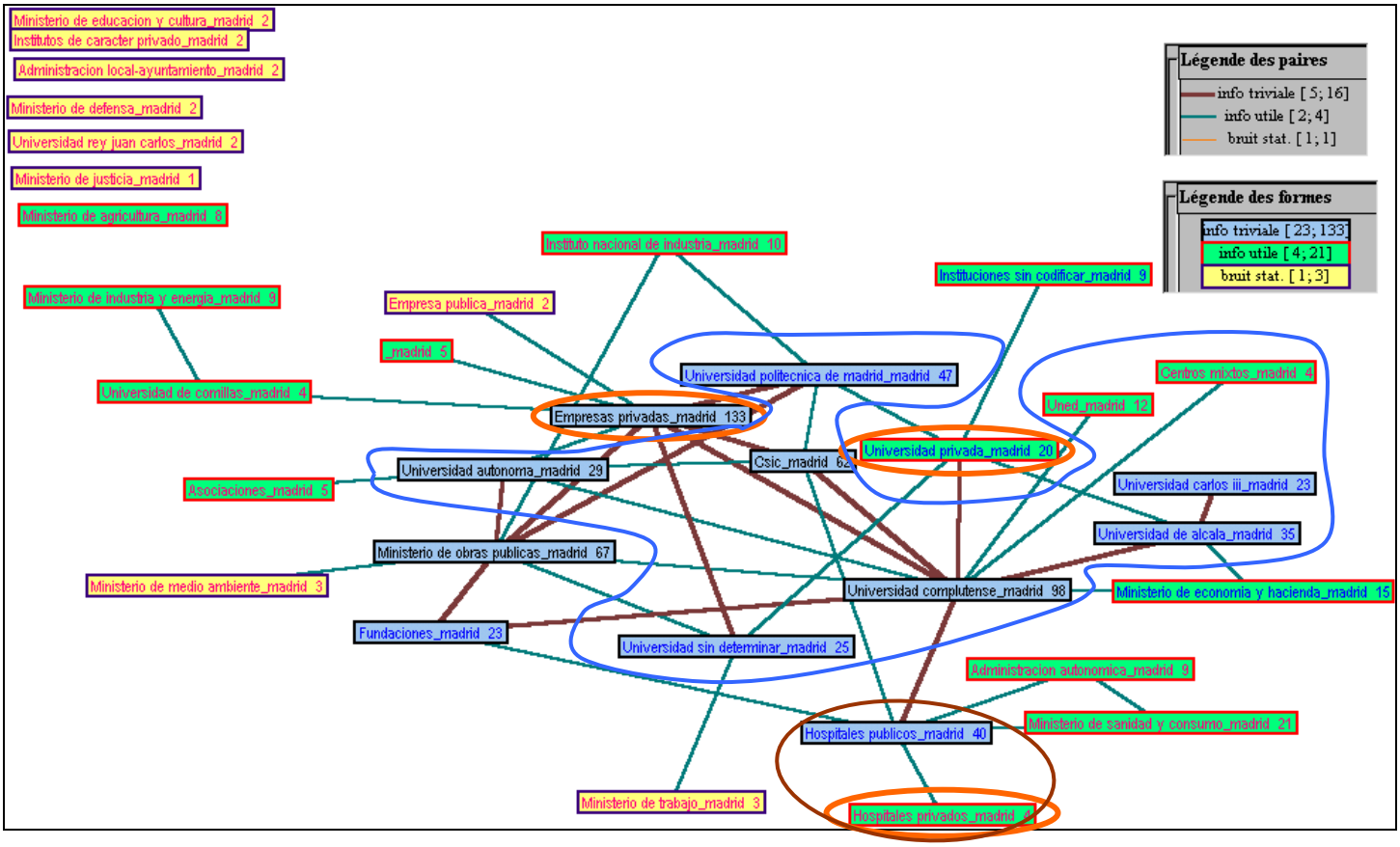


Figure 11 : Réseau des collaborations entre les organismes de la Région Autonome de Madrid

En ce qui concerne les organismes publics qui collaborent le plus avec le secteur privé, les principaux acteurs de la Région de Catalogne sont *l'Université Polytechnique de Catalogne*, l'*Université de Barcelone* et l'*Université Rovira i Virgli* pour des collaborations avec les entreprises privées de Barcelone appuyés par des associations et l'*Université de Lérida (*Lleida) pour des collaborations avec des entreprises privées de Lérida. Par ailleurs, les hôpitaux publics de Barcelone, Tarragone, Gérone et Lérida établissent des collaborations peu fréquentes avec les hôpitaux privés de Barcelone et Tarragone. Etrangement, les entreprises privées de Gérone ne collaborent pas avec des organismes de sa région mais avec des entreprises privées de la Région de Madrid.

Pour la Région de Madrid, les organismes publics impliqués dans des collaborations avec le secteur privé sont principalement l'*Université Autonome de Madrid*, l'*Université Polytechnique de Madrid*, l'*Université de Complutense* et l'*Université de Comillas* avec les entreprises privées de Madrid et l'*Université d'Alcalá* avec les universités privées de Madrid.

CONCLUSIONS

Deux grandes séries de commentaires sont à tirer d'une telle étude.

Les premiers commentaires concernent essentiellement les réponses que nous offrent les résultats bibliométriques à la question posée initialement : Quels échanges existe-il entre les acteurs secteur public et du secteur privé en Espagne ? Par le truchement de l'évaluation des collaborations scientifiques et techniques, les résultats nous permettent de relever que :

- Les échanges entre le secteur public et le secteur privé ne représentent que 24% des échanges.
- Le domaine où les échanges entre le secteur public et privé sont les plus nombreux est le secteur Electricité&Electronique avec une part de 40% des collaborations effectuées dans ce secteur.
- Les domaines qui arrivent juste derrières sont le BTP, la Chimie et les Télécommunications avec une part tournant autour des 30% de collaborations au sein de chaque secteur.
- Tous domaines confondus, ces échanges sont principalement obtenus à la suite d'une collaboration entre une université publique et une entreprise privée : 39% des échanges entre le secteur public et le secteur privé sont issus d'une telle collaboration.
- De façon générale, les collaborations entre les acteurs du secteur public et ceux du secteur privé sont très peu souvent récurrentes. Il y a un très fort taux de renouvellement des acteurs dans les collaborations entre le secteur public et privé.

La seconde série de commentaires expriment la pertinence, ressentie à l'issu de cette étude, des techniques bibliométriques dans l'évaluation du système Science-Technologie-Société :

- L'étude bibliométrique a permis l'élaboration d'indicateurs quantitatifs exprimant plusieurs niveaux d'agrégations des données caractérisant le système STS.
- L'étude bibliométrique a facilité l'identification des acteurs participant aux échanges des informations et des connaissances (dans le cadre de collaborations scientifiques et techniques) que ce soit dans les secteurs économiques porteurs et dans les principales régions espagnoles.
- L'étude bibliométrique a fourni des renseignements tangibles sur les habitudes et les pratiques de ces échanges comme par exemple : le taux de renouvellement des acteurs, la persistance des collaborations dans le temps, l'intensité des collaborations ou le degré d'ouverture des acteurs.

BIBLIOGRAPHIE

Boutin E (1999), Le traitement d'une information massive par l'analyse réseau : méthodes, outils et applications, Thèse soutenue le 14 janvier 1999 à l'Université d'Aix-Marseille III, France

Boutin E, Dumas P, Rostaing H, Quoniam L (1996), « Les réseaux comme outils d'analyse en bibliométrie. Un cas d'application : les réseaux d'auteurs. », Cahiers de la documentation, N°1, p. 3-13

Casillas Bueno JC, Roldán Salgueiro, JL (1996), « La articulación del sistema español de Ciencia y Tecnología », Cuadernos de Ciencias Económicas y Empresariales, 30, pp. 81-102.

Commission of the European Communities (2001), 2001 Innovation scoreboard. Commission staff working paper, SEC (2001) 1414. Brussels. http://trendchart.cordis.lu/Reports/Documents/SEC_2001_1414_EN.pdf

COTEC (1998), Libro blanco. El sistema español de innovación. Diagnósticos y recomendaciones, Madrid: COTEC. http://www.cotec.es/cas/publicaciones/libro.html

COTEC (2001), Indicadores de innovación. Situación en España, Madrid: COTEC. http://www.cotec.es/cas/publicaciones/pre_est_20.html

Giménez Toledo E (2001), Análisis de la trnsferencia de información entre el sector público y el sector privado a partir de las producciones y los consumos de información científica y técnica, Thèse soutenue le 10 septembre 2001 à l'Université Carlos III, Espagne

OECD (2001), Science, Technology and Industry scoreboard. Towards a konwledge-based economy. Paris: OECD

Quintanilla MA (1992), « El sistema español de Ciencia y Tecnología y la política de I+D », Arbor, 141, 554-555, pp. 9-29.

Rostaing H (2001), « Le Web et ses outils d'orientation. Comment mieux appréhender l'information disponible sur l'Internet par l'analyse des citations ? », Bulletin des biblitohèques de France, à paraître : Vol. 1, p. 68-77, [url : http://bbf.enssib.fr/bbf/html/2001_46_1/2001-1-p68-rostaing.xml.asp]

Rostaing H, Ziegelbaum H, Boutin E, Rogeaux M, Quoniam L (1998), « Analyse de commentaires libres par la technique des réseaux de segments. », Actes du colloque : Fourth International Conference on the Statistical Analysis of Textual Data, JADT'98, p. 697-704

ANNEXES

Annexe 1

SECTEUR ETUDIE	Nombre de revues
BTP	7 revues
Chimie	3 revues
Pharmacie	12 revues
Alimentation	6 revues
Electricité et électronique	6 revues
Mécanique	6 revues
Télécomunications	5 revues
Economie et Entreprise	27 revues
	72 revues

Tableau 2 : nombre de revues considérées comme représentatives pour chaque secteur étudié

Annexe 2

Code	Catégorie institutionnelle
UNIV	Universités
CSIC	CSIC et centres de recherche publique
AGE	Administrations nationales
AA	Administrations régionales
AL	Administrations locales
FUNDS	Associations & fondations

EPRIV	Entreprises privées
EPUBL	Entreprises publiques
HPRIV	Hôpitaux privés
HPUB	Hôpitaux publics
IE	Instituts étrangers
OI	Organismes internationaux

Tableau 3 : catégories institutionnelles utilisées pour caractériser les affiliations

PROPOSITION A L'INTEGRATION DES PROFILS DANS LE PROCESSUS DE RECHERCHE D'INFORMATION

Anis Benammar
Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 Toulouse Cedex, France
benammar@irit.fr

Gilles Hubert
Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 Toulouse Cedex, France
hubert@irit.fr

Josiane Mothe
Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne, 31062 Toulouse Cedex, France
mothe@irit.fr
Institut Universitaire de Formation des Maîtres Midi-Pyrénées
http://www.irit.fr/~Josiane.Mothe

Résumé : La formulation du besoin d'information est un des éléments clés pour obtenir des résultats pertinents dans un processus de recherche. Pour aider à cette formulation, des travaux proposent d'introduire la notion de profil d'un utilisateur. Les profils regroupent des informations qui permettent d'enrichir ou d'affiner l'expression des besoins d'un utilisateur mais également de filtrer des documents parmi les résultats de recherche. Les travaux présentés dans cet article s'inscrivent dans ce cadre. Nous proposons un système qui intègre des profils exploités au cours des différentes étapes d'un processus de recherche. Un profil correspond à l'expression d'un besoin en information. Il évolue en fonction des résultats de recherche obtenus en appliquant des mécanismes issus des méthodes de reformulation de requêtes. Les profils sont mémorisés et peuvent être partagés au sein d'un groupe d'utilisateurs. Ils peuvent ainsi être réutilisés et enrichis en tirant profit des recherches d'utilisateurs concernés par le même besoin d'information. Notre système de profils illustre une variante de la recherche collaborative.

Abstract : The formulation of the user's information need is a key step to get relevant results. In order to assist users to better express their information need, many systems introduce the user profile concept. Profiles include several information in order to improve the expression of the user's need and also to filter the retrieved results. The system presented in this paper joins this framework. Our system integrates the use of the profiles in all the steps of the information retrieval process. A profile corresponds to the user's information need. It evolves along the search sessions based on query reformulation mechanisms. Profiles are memorized

and shared by many users. Doing so, profile can be reused and improved by benefiting from researches of users concerned by the same information need. Our profile system illustrates a variant of the collaborative search.

Mots-clés : recherche d'information, profil d’interrogation, profile d’identification, reformulation automatique, recherche collaborative permettre

Keywords : information retrieval, querying profile, identification profile, automatic reformulation, collaborative search.

Proposition à l'intégration des profils dans le processus de recherche d'information

INTRODUCTION

Les systèmes de recherche d'information (SRI) ont pour objectif de restituer les documents correspondant au mieux aux besoins des utilisateurs. Différents éléments influent sur l'efficacité des mécanismes mis en œuvre. Parmi ces éléments, nous pouvons citer la représentation des documents (ou indexation), la représentation du besoin d'information de l'utilisateur et la mise en correspondance de ces deux éléments. Différents travaux s'intéressent à personnaliser la représentation du besoin d'information afin d'améliorer la satisfaction de l'utilisateur [Korfage97]. L'utilisation des profils utilisateurs entre dans ce cadre [Lieberman95], [Jeribi01]. Deux types d'utilisation de profil se distinguent. D'une part, un profil peut être utilisé dans une étape de pré-recherche pour aider l'utilisateur à formuler ou reformuler son besoin [Jeribi01]. Il peut s'agir par exemple d'affiner l'expression d'une requête proposée par l'utilisateur en fonction de son (ses) profil(s). D'autre part, un profil peut également être utilisé dans une étape post-recherche pour filtrer les résultats d'une recherche [Lainé99].
Différents formats de profils ont été étudiés [Korfhage97]. Un profil peut être simple [Baudisch97] ou étendu [Lieberman95]. Un profil simple se présente sous la forme d'un ensemble de mots-clés et éventuellement des poids associés. Un poids traduit l'importance de chaque terme dans le profil. Un profil étendu inclut, en plus des mots-clés et de leur poids, une série d'informations qui décrivent le contexte de la recherche (caractéristique de l'utilisateur, but de la recherche, thème de la recherche, etc.).
Nos travaux s'intéressent à la définition d'un système qui intègre les profils dans le but d'assister l'utilisateur tout au long d'un processus de recherche d'information. Selon notre approche, un profil représente l'expression du besoin en information d'un utilisateur. Ce profil évolue au fur et à mesure des interrogations du système en exploitant les résultats successifs au travers de mécanismes issus des techniques de reformulation de requêtes [Croft00]. Par ailleurs, les profils estimés efficaces peuvent être sauvegardés. Ces profils peuvent alors être partagés et réutilisés au sein d'un groupe permettant ainsi aux utilisateurs de tirer mutuellement profit de leurs recherches. Nous favorisons ainsi une forme de recherche d'information collaborative [Kurzke98]. La recherche d'information dans ce cas, n'est plus seulement orientée vers un unique utilisateur mais tient compte des besoins d'un groupe d'utilisateurs. De plus, notre système cherche à alléger la tâche de l'utilisateur au niveau de la gestion des profils. Son intervention est dans la plupart des cas optionnelle afin de minimiser son interaction avec le système. Elle reste cependant obligatoire dans certains cas afin que les profils puissent être utilisés de façon optimale.
Cet article détaille le système de profils que nous proposons. Dans la deuxième section nous examinons la problématique à laquelle nous nous intéressons à travers l'utilisation des profils. Dans la troisième section, nous spécifions les différentes composantes d'un profil dans notre système. Dans la dernière section, nous décrivons les fonctionnalités de notre système de profils.

1 – PROBLEMATIQUE ET OBJECTIFS

Un objectif des systèmes de recherche d'information et de permettre à l'utilisateur de retrouver l'information la plus pertinente pour lui. Parmi les d'études menées dans le cadre de la recherche d'information, différents travaux proposent d'introduire la gestion des profils des utilisateurs pour améliorer les recherches. Les profils sont utilisés pour filtrer un ensemble de documents afin de ne restituer à l'utilisateur que ceux qui correspondent le mieux à son (ses) profil(s). Alternativement, ils sont utilisés pour affiner l'expression des besoins en information de l'utilisateur. Deux types de gestion de profils ont été proposés à travers les différents systèmes existants : directe ou indirecte.
Dans la plupart des systèmes, la gestion des profils est indirecte [Lieberman95], [Jeribi01], c'est à dire qu'elle est transparente à l'utilisateur. L'avantage de ce type de gestion est qu'elle permet de minimiser l'interaction de l'utilisateur avec le système et ainsi de ne pas alourdir sa tâche de recherche. Cependant elle implique que le système soit capable de détecter les changements des intérêts de l'utilisateur au cours d'une même connexion. En effet, dans la gestion indirecte des profils, l'utilisateur n'ayant pas connaissance de l'existence des profils, il n'indiquera pas au système qu'il change de contexte de recherche. La difficulté est dans ce cas de déceler les changements d'intérêt de l'utilisateur pour modifier le profil de recherche.
A l'opposé, dans la gestion directe des profils, comme par exemple dans l'éditeur de profil de Baudish [Baudish97], l'utilisateur doit intervenir dans toutes les étapes du processus de recherche pour gérer ses profils. Cette stratégie permet à l'utilisateur d'avoir plus de contrôle sur ses profils puisqu'il les gère directement depuis leur création. La gestion directe implique cependant de lourdes charges pour l'utilisateur qui peuvent constituer un frein à l'utilisation des profils.

Pour limiter les inconvénients introduits par les deux types de gestion précédemment décrits, nous proposons une gestion de profils intermédiaire qui exploite les avantages de chaque solution. Dans certaines étapes du processus de recherche des mécanismes automatiques sont mis en œuvre par défaut, tout en laissant à l'utilisateur la possibilité d'intervenir directement. Dans d'autres étapes, l'intervention de l'utilisateur est obligatoire. Ces éléments sont détaillés dans la section 4.

Le but du système que nous proposons est d'aider l'utilisateur dans son processus de recherche d'information. L'aide est mise en œuvre via des fonctionnalités permettant à l'utilisateur :

- **d'enrichir et d'affiner l'expression de son besoin en exploitant des résultats de recherche successifs et ainsi concentrer progressivement les résultats autour du contexte de la recherche,**
- **de réutiliser les recherches passées propres à un utilisateur ou partagées par un groupe d'utilisateurs et validées comme fournissant un résultat satisfaisant. Les utilisateurs peuvent d'une part garder trace de leurs recherches et d'autre part échanger leurs expériences de recherche avec les autres.**

La réutilisation des recherches passées permet à l'utilisateur de développer un même besoin en information d'une manière incrémentale sur plusieurs sessions de recherche. Plusieurs utilisateurs peuvent aussi collaborer pour rechercher une information puisque les profils peuvent être utilisés successivement par différents utilisateurs. Les profils regroupent ainsi les expériences de recherche d'un ou de plusieurs utilisateurs pour un même besoin d'information. Dans ce cas, l'utilisation des profils est une forme de la recherche collaborative [Kurzke98].
Dans les sections suivantes, nous détaillons les différentes composantes d'un profil ainsi que l'utilisation des profils dans un processus de recherche d'information.

2 - MODELISATION DES PROFILS

Nous avons défini des formats de profils spécifiques permettant de mémoriser les recherches effectuées par différents utilisateurs. Un profil contient les informations suivantes :

- un sous-profil d'identification de l'utilisateur.
- un sous-profil d'interrogation : il correspond à la description du contexte de recherche d'un utilisateur pendant une seule session de recherche (court terme) ou plusieurs sessions de recherche (long terme).

Chaque utilisateur maintient un seul sous-profil d'identification et un ensemble de sous-profils d'interrogation qui traduisent son contexte de recherche par thème d'étude.

2.1 Le sous-profil d'identification

Cette première composante du profil sert à identifier un utilisateur à travers une série d'informations. Ce sous-profil spécifie des informations sur l'utilisateur ainsi que sur ses droits d'accès aux profils d'interrogation des autres utilisateurs dans le même groupe de travail. Ces informations, gérées par chaque utilisateur, permettent de réglementer l'échange des profils d'interrogation dans un groupe d'utilisateurs. Chaque utilisateur spécifie les droits d'accès qu'il attribue à ses propres profils pour les autres membres du groupe. Ainsi, le profil d'identification est défini à la première connexion au système de profils et est mis à jour par incrémentation à chaque création d'un profil d'interrogation.

2.2 Le sous-profil d'interrogation

Un sous-profil d'interrogation peut être assimilé à une requête. Il traduit le besoin en information de l'utilisateur. Dans les systèmes existants, une recherche est généralement caractérisée uniquement par un ensemble de mots-clés éventuellement pondérés. Elle est ainsi séparée de son contexte (le but de la recherche, son thème, etc.). Dans le but d'optimiser la réutilisation des profils et faciliter la compréhension des profils, nous optons pour l'association de la recherche à son contexte. Nous incluons ainsi dans la structure du sous-profil d'interrogation, en plus des mots-clés, des informations telles que la description courte et détaillée de la recherche, le thème de la recherche, le lien avec les autres profils et la notion de durée de vie d'un profil. Ainsi, nous distinguons les profils d'interrogation à court terme qui correspondent à une seule session de recherche (même utilisateur, même besoin) et qui ne sont pas réutilisables et les profils d'interrogation à long terme qui peuvent être issus de plusieurs sessions de recherche et qui peuvent être réutilisés. Ces deux types de profils sont liés puisqu'un profil à court terme peut être sauvegardé sous la forme d'un profil à long terme s'il correspond à un besoin d'information récurrent ou utile pour d'autres utilisateurs.
Le schéma suivant, illustre les liaisons entre les deux types du profil d'interrogation.

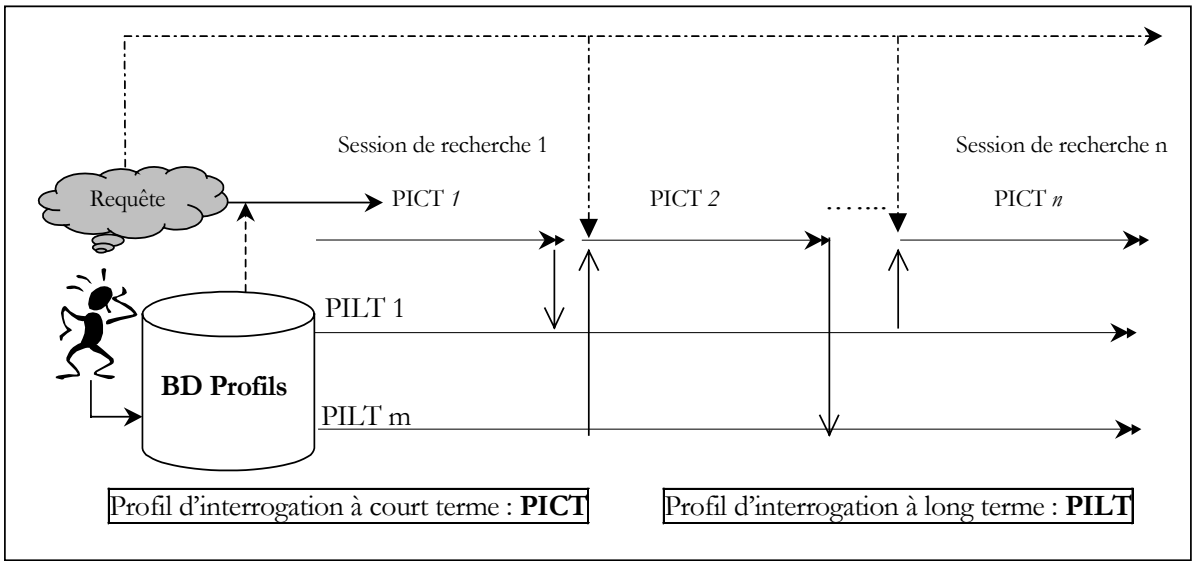


Figure 1 : Interaction entre les composantes du profil d'interrogation

2.2.a Le sous-profil d'interrogation à court terme

Initialement, un sous-profil d'interrogation à court terme correspond soit à une requête soumise au système de recherche par l'utilisateur soit à un profil d'interrogation à long terme choisi par l'utilisateur parmi ceux existant dans le système de profil.
Un profil à court terme est employé tout au long d'une session de recherche dans le processus de recherche. Son contenu évolue durant la session pour aboutir à une expression du besoin de l'utilisateur plus adaptée au contenu de la collection interrogée. Cette évolution est basée sur des mécanismes dérivés des processus de reformulation automatiques de requêtes par réinjection de pertinence [Rocchio71], [Croft00].

2.2.b Le sous-profil d'interrogation à long terme

Le sous profil d'interrogation à long terme correspond à la mémorisation d'un profil à court terme ou à une nuance ou évolution d'un profil à long terme existant. Dans ce cas, il peut s'agir de développer et de faire évoluer un même besoin en information dans le temps. Le lien d'antécédence ou de dérivation entre profils à long terme est mémorisé dans le système de profil.
La création ou mise à jour d'un profil d'interrogation à long terme intervient à la fin de chaque session de recherche. Ainsi, les profils d'interrogation à long terme peuvent être utilisés dans les sessions futures pour relancer de nouveaux processus de recherche.

3. - EXPLOITATION DES PROFILS AU COURS D'UN PROCESSUS DE RECHERCHE

La gestion des profils est assurée tout au long du processus de recherche :

- au début du processus de recherche : l'utilisateur peut se baser sur un profil d'interrogation (à long terme) existant. Cette aide à l'expression du besoin correspond à une réutilisation directe des profils qui permet à l'utilisateur de débuter une recherche à partir d'un profil qui correspond ou qui est proche de son besoin de recherche.
- au cours du processus de recherche : des procédures de reformulation automatiques sont proposées à l'utilisateur pour mettre à jour le profil d'interrogation qu'il utilise [Benammar02]. Elles sont basées sur les principes de réinjection de pertinence aveugle (totalement automatique) ou de réinjection des jugements de pertinence des documents retrouvés spécifiés par l'utilisateur.
- à la fin du processus de recherche : le profil utilisé lors de la recherche peut être sauvegardé pour d'éventuelles utilisations futures.

3.1 Au début du processus de recherche

Durant cette étape, notre but est d'assister l'utilisateur dans la formulation de son besoin en information en lui proposant de réutiliser des profils de recherche existants. Cette aide à l'utilisateur est fondamentale dans un système de recherche d'information compte tenu des difficultés que l'utilisateur peut rencontrer pour formuler d'une manière efficace son besoin en information.
La réutilisation des profils présente un grand intérêt pour l'utilisateur au moment où il formule son besoin en information. En effet, dans certains cas l'utilisateur cherche une information dans un domaine sans avoir d'idées

précises de ce qu'il cherche et sans savoir comment l'information pertinente est représentée dans la collection interrogée.
Pour un utilisateur, les formulations passées mémorisées dans ses propres profils ou dans ceux des autres utilisateurs peuvent faciliter l'expression de son besoin en information. L'utilisateur peut partir d'un profil existant pour le développer ou bien pour en créer un nouveau en s'inspirant des profils existants.

Figure 2 : Spécification du besoin en information de l'utilisateur

En revanche, il peut s'avérer fastidieux pour un utilisateur de consulter tous les profils existants. Des critères de sélection définis par l'utilisateur (figure 3) lui permettent de limiter le nombre de profils à consulter. Les critères de sélection que l'utilisateur peut appliquer sont les suivants :

- utilisateur : limitation aux profils d'un utilisateur donné,
- thème de recherche : limitation aux profils qui correspondent au thème de recherche de l'utilisateur,
- date de dernière modification : limitation aux profils modifiés (utilisés) depuis une certaine date.

Figure 3 : Recherche des profils existants

Le résultat de sélection des profils (figure 4) détaille à l'utilisateur les profils (description, thème associé, liste des mots-clés, etc.).
Le système offre également à l'utilisateur la possibilité de naviguer entre les profils. En effet, pour garder un historique de l'évolution des profils, ces derniers sont liés entre eux par une relation de type prédécesseur/successeur. Un profil peut posséder un prédécesseur et un ou plusieurs successeurs. L'utilisateur peut exploiter cette relation d'enchaînement entre les profils pour bien comprendre l'évolution du besoin en information et effectuer le bon choix du profil d'interrogation à réutiliser. La relation de prédécesseur/successeur peut dans certains cas être assimilée à une relation de généralisation/spécialisation. En effet, un profil d'interrogation est réutilisé pour affiner ou inversement pour développer un besoin en information. Dans certains cas, l'utilisateur peut trouver que le contexte de recherche rattaché à un profil est assez générique. En partant sur les successeurs de ce même profil, l'utilisateur pourra retrouver des spécialisations qui correspondent à son contexte de recherche précis. Inversement, les prédécesseurs d'un profil peuvent intéresser l'utilisateur parce qu'ils présentent le cadre général de son besoin.

Figure 4 : Résultat de recherche des profils

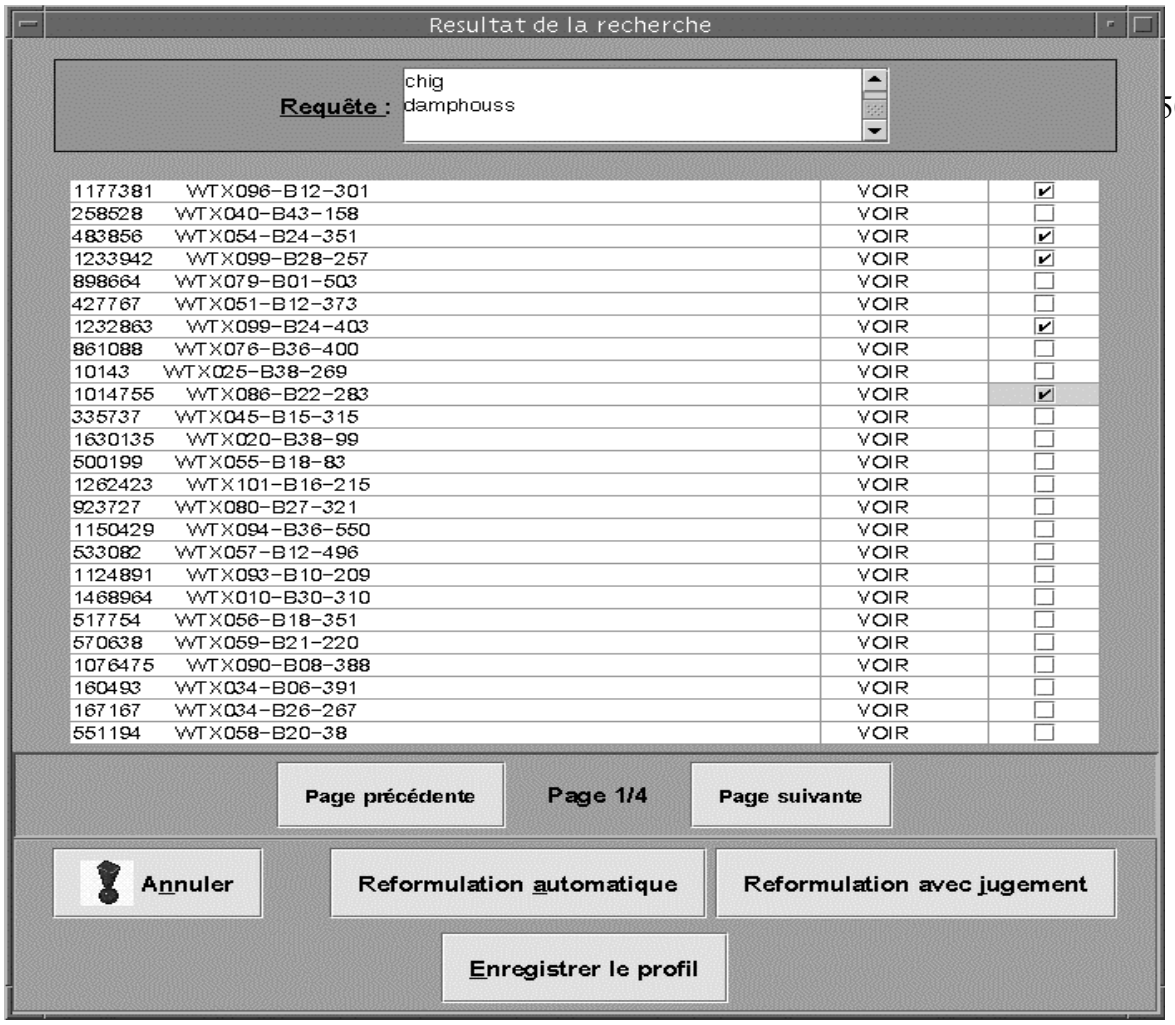
3.2 Au cours du processus de recherche

Durant cette étape, l'expression du profil d'interrogation est affinée afin d'aboutir à une meilleure expression du besoin de l'utilisateur. Nous avons adapté les mécanismes de réinjection de pertinence habituellement utilisés pour la reformulation automatique de requête à la modification automatique des profils d'interrogation [Benammar02].

Dans notre cadre d'étude, l'intervention de l'utilisateur au cours du processus de reformulation est optionnelle. Dans le cas où l'utilisateur précise des jugements par rapport au résultat de la recherche (documents restitués), la procédure de reformulation est exécutée sur la base des documents jugés pertinents par l'utilisateur. Dans le cas contraire, une procédure permet d'estimer les préférences de l'utilisateur en terme de documents pertinents. Cette procédure analyse les documents restitués et applique un filtrage pour ne garder que ceux qui sont susceptibles d'être les plus pertinents. La procédure de reformulation automatique du profil est alors exécutée sur la base de ces documents.

Dans les deux cas, la procédure de reformulation se base sur une analyse de cooccurrence des termes dans les documents pertinents et dans l'expression du sous-profil d'interrogation.

Figure 5 : affichage des résultats de recherche



3.3 A la fin du processus de recherche

Durant la phase finale d'une session de recherche, l'intervention de l'utilisateur est obligatoire afin de faire face au problème de changement des intérêts. En effet, durant une même connexion, l'utilisateur peut changer à plusieurs reprises son contexte de recherche. A chaque changement de contexte de recherche, l'utilisateur doit l'indiquer au système en annulant la recherche en cours, ou en sauvegardant le profil associé. Ainsi, le profil d'interrogation employé est modifié dès que l'utilisateur change de direction de recherche. L'intervention de l'utilisateur durant cette étape a pour but de valider le résultat de recherche qu'il vient d'obtenir. L'utilisateur doit préciser si le profil obtenu à la fin d'une recherche permet de retrouver l'information qui correspond à son besoin. Si ce n'est pas le cas, profil ne pourra pas être sauvegardé. En revanche si l'expression du besoin courante s'est avérée efficace, le profil d'interrogation correspondant pourra être sauvegardé afin d'être utilisé lors des prochaines interrogations. Dans le cas où il est issu d'un autre profil à long terme (choisi au début de la session), cette sauvegarde pourra soit correspondre à une mise à jour du profil initial, soit aboutir à la création d'un nouveau profil, lié à son "père". Nous traduisons ainsi, l'intérêt de l'utilisation des profils comme étant des expériences de recherche dont l'utilisateur pourrait tirer profit dans ses recherches futures.

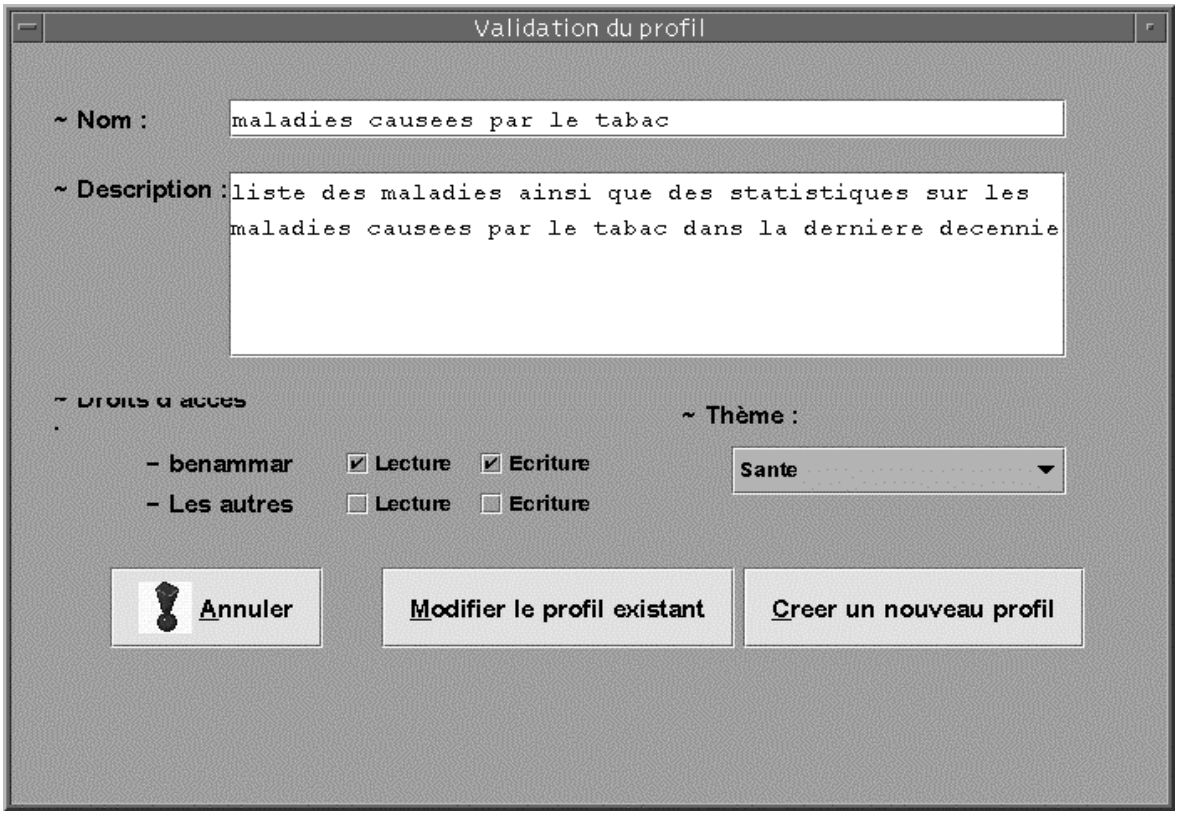


Figure 6 : Enregistrement des propriétés d'un profil

CONCLUSION

Dans cet article, nous avons proposé un système de gestion de profils pour la recherche d'information. L'objectif de ce système est d'aider l'utilisateur à retrouver les documents les plus pertinents tout en mutualisant les expériences de recherche.

Un profil représente un besoin en information. Ce profil peut évoluer au fur et à mesure des interrogations effectuées au cours d'une session (même contexte de recherche, même utilisateur) ou au cours de différentes sessions. Cette évolution est basée sur des techniques de reformulation de requêtes automatiques. Ainsi, le besoin en information exprimé initialement est progressivement enrichi ou affiné. D'autre part, le système permet à un utilisateur de sauvegarder les profils considérés comme fournissant des résultats pertinents. Ces profils peuvent être partagés entre différents utilisateurs ayant les mêmes besoins d'information. Chaque utilisateur peut donc profiter des expériences validées par d'autres utilisateurs afin d'améliorer les résultats de ses recherches.

Dans notre système, un profil est composé de deux éléments : le sous-profil d'identification et le sous-profil d'interrogation. Le sous-profil d'identification permet d'identifier un utilisateur à travers un ensemble d'information comme son nom et ses droits d'accès par rapport aux profils des autres utilisateurs. Le sous-profil d'interrogation décrit le besoin en information de l'utilisateur. Nous distinguons deux niveaux de profil d'interrogation. Le sous-profil d'interrogation à court terme correspond au besoin courant de l'utilisateur et sa durée de vie est celle de la session de recherche. Un sous-profil d'interrogation à long terme décrit un besoin d'information issu de plusieurs sessions de recherche.

Le système employé permet d'assister l'utilisateur durant toutes les étapes de son processus de recherche. L'intervention de l'utilisateur selon les étapes est soit facultative, soit obligatoire. Elle est facultative par exemple lors des reformulations de profil où le système utilise les jugements de l'utilisateur lorsqu'il les spécifie. Elle est obligatoire lorsque l'utilisateur juge l'efficacité du sous-profil d'interrogation à court terme par rapport au résultat de recherche retrouvé, afin de sauvegarder son évolution en tant que sous-profil d'interrogation à long terme.

Les travaux actuels concernent notamment les aspects de reformulation automatiques des profils d'interrogation en exploitant les liens hypertextes qui peuvent exister entre les documents. Un autre aspect concerne la stabilité des profils à long terme et la gestion des avis divergents des utilisateurs concernant la pertinence des documents par rapport à un profil.

RÉFÉRENCES BIBLIOGRAPHIQUES

{Croft00} Croft W.B, Xu J., *Improving Effectiveness of information retrieval with local context analysis.* ACM Transaction on Information systems Volume 18, Number 1, pages 79–112, 2000.

{Benammar02} Benammar A., Mothe J., Hubert G., *Automatic profile reformulation using a local document analysis*, European colloquium on IR research , Glasgow. Springer-Verlag , pages 124-134, 2002.

{Baudisch97} Baudisch P., *The Profile Editor: designing a direct manipulative tool for assembling profiles*, In Proceedings of Fifth DELOS Workshop on Filtering and Collaborative Filtering, pages 11-17, Budapest, November 1997.

{Jeribi01} Jeribi L., Rumpler B., Pinon J. M., *Système d'aide à la recherche et à l'interrogation de bases documentaires, fondé sur le réutilisation d'expériences*. XIXème Congrès INFORSID, pages 443-463, Genève, 2001.

{Korfhage97} Korfhage R., *Information storage and retrieval.* Wiley Computer Publishing 0-471-14-338 3, 1997.

{Kurzke98} Kurzke C., Galle M., Bathelt M., *WebAssist : a user profils specific information retrieval assistant*, 1998.

http://decweb.ethz.ch/WWW7/1903/com1903.html

{Lainé99} Lainé-Cruzel S., *ProfilDoc : filtrer une information exploitable,* Bulletin des Bibliothèques Françaises (BBF), T. 44, Numéro 55, pages 60-64, 1999.

{Lieberman95} Lieberman H., *Letizia: An agent that assists web browsing*, Proceedings of the fourteenth International Joint Conference on Artificial Intelligence, pages 924-929, 1995.

{Rocchio71} Rochio J. J., *Relevance Feedback in Information Retrieval*, in «The SMART retrieval system- Experiments in Automatic Document Processing » Prentice Hall Inc., pages 313-323, 1971.

DATA MINING & DEVELOPMENT POLICY:
WHICH HELP ON BUILDING TERRITORIAL INDICATORS ?

Yann BERTACCHINI
Maître de Conférences S.I.C
Expert près l'U.E
Laboratoire LePont, Université de Toulon & du Var
Dépt "services & Réseaux de Communication"
200, Avenue Victor Sergent, 83.700 Saint-Raphaël, France
mailto:bertacchini@univ-tln.fr

Sumário : Agora em, o ambiente de atores territoriais está composto de local, redes internacionais nacionais. Além do mais, por causa do novo sistema de telecomunicações e multimedia, a tecla a novas estacas tornou-se uma prioridade. Agradece a redes e para seuas aplicações, uma nova inteligência coletiva será feita possível para resolver problemas provados mais e mais complexo devido a informação massiva. Em hoje'mundo de s, quantia de informação sabido é essencial, e dá esses, territórios incluído, quem tem uma abundância de informação, uma borda acima outros. Todos atores locais, organizações públicas privadas, existem numa idade onde todas espécies diferentes de conhecimento são essenciais e importantes sobreviver e prosperar. Sem conhecimento, nós nao saberíamos que o que acontecia ao redor do próximo canto, nem ria't entende a jamais-mudança, mundo jamais crescendo. Um território local que competitivamente pode colecionar dados com êxito pode permanecer em topo de o que's ir em ao redor deles, e alcança níveis altos de entendimento que leva a êxito. É aceitado em parte toda, isso se um sabe, que um receberá, e isto segue tudo e todo o mundo. Se obter a informação muito necessária era tão fácil quanto pode soar, ter havido muito mais êxito, regular mais cedo que falamos, por sobrepujar qualquer barreiras que podem vir no meio. Os territórios em hoje'mundo de s também deve lidar com níveis altos de competição, e seu êxito anda de alto nos ombros dos atores locais eles servem. Por cavar fundo em seus fregueses' meio de vida, meio de pensar, territórios podem são capazes de usar esta informação como uma ferramenta positiva ajudar prediz e "lê" seu freguês'mentes de s, baseado em experiências de passado. O pensado que é proposto neste artigo aponta num processo de mediation, como nós também chamamos Sistema Territorial de Inteligência de Informação. Para encorajar o synergy entre atores locais. De fato, novos câmbios -parcialmente virtual- vai ir cohabit com outros lados de relações rompidas. Existir e renovar, esses espaça devem apresentar redes e meio de transferência de conhecimento por disposalsType educacional, econômico democrático ou texto de pasta aqui. Embarcamos para explicar como complexo o local e seu desenvolvimento é e, por quê dá o meio à definição e para o escrito de um procedimento de informational baseado em mineração de Dados para fingir a territoriality como para um território. Também explicamos por quê é a construção territorial uma questão para uma aproximação de systemic cujo propósito central, em nosso interesse, sobra a capacidade do território organizar o relational rede local de vistoria pericial e fazer seu meio a um projeto territorial de inteligência.

Summary : From now on, the environment of territorial actors is composed of local, national and international networks. Moreover, because of the new system of telecommunications and multimedia, the key to new stakes has become a priority. Thanks to networks and their applications, a new collective intelligence will be made possible in order to solve problems proved more and more complex due to massive information. In today's world, amount of information known is essential, and gives those, territories included, who have an abundance of information, an edge above others. All local actors, private and public organisations, exist in an age where all different kinds of knowledge are essential and important to survive and thrive. Without knowledge, we would not know what was happening around the next corner, or wouldn't understand the ever-changing, ever growing world. A local territory which can competitively collect data can successfully stay on top of what's going on around them, and achieve high levels of understanding which leads to success. It is accepted everywhere, that if one knows, than one will receive, and this follows everything and everyone. If obtaining the much needed information was as easy as it may sound, there would have been much more success, even earlier than we speak, by overcoming any barriers that may come into the way. Territories in today's world must also deal with high levels of competition, and their success rides high on the shoulders of the local actors they serve. By digging deep into their *customers*' way of life, way of thinking, territories might be able to use this information as a positive tool to help predict and "read" their *customer*'s minds, based on past experiences. The thought which is proposed in this article aims at a process of mediation, as we also called Territorial Information Intelligence System. In order to encourage the synergy between local actors. Indeed, new exchanges -partly virtual- will cohabit with other sides of disrupted relations. To exist and renew, those spaces should present networks and means of transfer of knowledge through democratic, economic and educational disposals. We set out to explain how complex the local and its development is and, why it gives the way to the definition and the writing of an informational procedure based on Data mining in order to pretend to territoriality as a territory. Also we explain why is the territorial construction a matter for a systemic approach whose central purpose, in our concern, remains the ability of the territory to organize the relational local expertise network and to make its way to a territorial intelligence project.

Palavras-chave: Os dados; Informação; Conhecimento, Local; Apólice; Processo.

Keywords : Data; Information; Knowledge, Local; Policy; Process.

Data mining & Development policy:

WHICH HELP ON BUILDING TERRITORIAL INDICATORS ?

INTRODUCTION

Data mining methods

ASSOCIATION

Investigates growth differences
Private School Location
The evaluation of socio-economic and cultural activities of cities possessing a university.

SEQUENCE-BASED ANALYSIS

Shifts in Regional Employment
Territorial Development and Networking

CLUSTERING

Knowledge, innovation and collective learning
Classification
Innovation and New Technologies

FUZZY LOGIC

The creation and development of incubators for business ideas in the ICT sector
Estimation
Measuring the level of quality of life of local governments

GENETIC ALGORITHM

Spatial Planning, Structural Change and Regional Development policies

NEURAL NETWORKS

Spatial Interaction Modeling of Interregional Commodity Flows
Data Mining of Advanced Database

INTELLIGENT AGENTS

The regional system of innovation and regional development
Multidimensional analysis tools
Immaterial resources and regional development : a territorial approach

CONCLUSION

BIBLIOGRAPHIC REFERENCES

INTRODUCTION

What is data mining? The process of automatically extracting valid, OR "You press the data until they confess".
In today's world, amount of information known is essential, and gives those who have an abundance of information, an edge above others. We all exist in an age where all different kinds of knowledge are essential and important to survive and thrive. Without knowledge, we would not know what was happening around the next corner, or wouldn't understand the ever-changing, ever growing world and the fast pace around us which increases everyday with new coming events and new ways of solving problems. Territories level as organisation need to see the need to acquire knowledge to keep up with the glocal (global-local) competition. A

territorial level which can competitively collect data can successfully stay on top of what's going on around them, and achieve high levels of understanding which leads to attractivity.
Data mining is compatible with territorial competitiveness. We will associate them in
our communication. The territory is carrying direction and establishing an activity of territorial watch before which would use amongst, other things average, data mining, can reinforce its attractivity.

Data mining methods

By digging deep into their inhabitants' way of life, way of thinking, their habits to do something, territories might be able to use this information as a positive tool to help predict and "read" their inhabitants' minds, based on past experiences. Much of the knowledge that can help out and assure success is hidden, not readily available to the right person at the right time. With the past limiting capabilities that have existed, emphasis has been placed on finding and sorting out this information to benefit the organization at hand[1]. The main problem here that needs to be overcome is the difficulty of extracting this hidden information about any particular system that has been studied from collected data. But the main and whole idea and act of discovering relationships and connecting variables is a database sum of the main idea and point of Data Mining.
There are various capabilities that are provided by data mining techniques. There are many techniques that Data Mining uses to do exactly that, and these techniques used depend on what type of information one is trying to develop and extract. In the present context of the emergence of the knowledge economy, the role of immaterial resources in regional or local development, in particular as a factor of attractivity and competitiveness, is widely acknowledged. However anyone who tries to consider both resource management and regional development gets confronted to a variety of content, to elements of different nature and origin, moreover to circular definitions[2].
The objective of this paper is to enter the black box of regional resources, more specifically of immaterial resources, and understand its territorial dimension in association with Data Mining methods. It proposes a conceptual framework allowing us to distinguish the various aspects of this immaterial 'magma' and to seize its dynamic through time and space.

ASSOCIATION

This technique approaches to address a problem that concerns a space-based analysis, treating the variables which are investigating on first step growth differences in the european urban system. The main goal is to locate trends and similar behaviors across a large number of indicators to be used later to understand. This information may be very useful to organisations to adjust aspects of their activities.

Investigates growth differences

There is evidence strongly suggesting that equilibrating flows between cities are highly constrained in the EU. Models in which growth of real GDP p.c. are the dependent variable perform well and make it possible to test significant hypotheses. Evidence is found which is supportive of a spatial adaptation of the endogenous growth model with the relative size of the university sector having a highly significant role in explaining growth differences[3].

Private School Location

A simple logic specification produces evidence of the importance of substitutes to private schools in deterring entry to a community. These are represented by public school spending per pupil and the extent of local library lending.

The evaluation of socio-economic and cultural activities of cities possessing a university.

The focus is to test socio-economic and cultural capacities of cities in which are located university to measure whether the university administration necessity of receives or not. On two successive steps is shown student attractive strength of this cities.

SEQUENCE-BASED ANALYSIS

This is another *space*-basket analysis dealing with the total factors in one professional *transaction*. This type of an analysis can identify a set of variables that could forecast the next anticipated behavior in labour market.

Shifts in Regional Employment

The U.E has undergone deep structural changes in the 1980s, 1990s with strong implications for the labor markets. The general move with overall decrease (80s) then overall increase (90s) in employment levels in U.E is evident from official statistics.

However, the sectoral and regional mapping of this shift has not been properly studied yet. Which appropriate model is able to measure the impact of output change, productivity gain and other non-labor factors on employment change in the period ?

Territorial Development and Networking

The purpose is to explore the main features and the significance of the SME sector development for approaching and to identify the signs and the perspectives of SME territorial networking phenomenon[4]. It also seeks to identify phenomena characteristic to territorial networking in which SMEs can bring a significant contribution.

CLUSTERING

This Data Mining method addresses problems with segmentation and grouping. Data with a big group of attributes are automatically assigned by clustering algorithms identifying unique characteristics into smaller sets of groups. This is often one of the first steps in a data mining analysis, identifying related record groups. These groups may then be used for further exploration and analysis of relationships. When combined with other data mining techniques, other types of characteristics may be determined.

Knowledge, innovation and collective learning

Innovative capacity of firms has traditionally been explained through intra-firm characteristics, being firms size the most important. A wave of empirical studies identifies small firms as the engines of technological change and innovative activity, at least in certain
industries. In the recent literature much emphasis has been put to determinate which are external to the firm; these external factors are called knowledge spillovers, and refer to the positive externalities that firms receive in terms of knowledge from the environment in which it operates. Both industrial and regional economists underline the importance of knowledge spillovers. The concept of relational capital is fundamental in this respect. Relation capital is in fact defined as the set of all relationships - market relationships, power relationships, co-operation - established between firms, institutions and people, which stem from a strong sense of belonging and a highly developed capacity of cooperation typical of culturally similar people and institutions. It is, indeed, reasonable to expect that relational capital will play a different role in different regional, sectoral and firm's contexts.

Classification

Classification is the most commonly applied data mining method, and is used to develop models the can classify large amounts of records. Pre-classified examples are used to determine a set of parameters needed to properly separate other data.

Innovation and New Technologies

Many European Regions with a high autonomous government degree have done an effort to promote research and technological development (R&D) to achieve welfare and economic growth. Several Regions have organized their efforts in science and technology[5], but, is really this budget effort been transferred to new technologies adopted by firms? Are science-technology flows increasing? Is there a regional gap between science and technology? Which are the factors determining these differences? What is the role of scientific knowledge to promote technological development in several Europeans Regions.
How is the basic science used by industrial firms? Which are the most dynamic sectors using scientific knowledge? Which are the most demanded scientific fields by private firms? What type of institutions do use the scientific knowledge? Which is the lag time to translate science to technology? The used methodology to explain the link between science and technology is based on scientific inputs in patents, or non-patent citations (NPC).

FUZZY LOGIC

Fuzzy logic is a superset of conventional (Boolean) logic that can handle a concept of partial truth whether values are "completely true" or "completely false", or anything in-between.

The creation and development of incubators for business ideas in the ICT sector

How to increase the number of business ideas inside an University, then to select the best ones and then to fill the incubators of companies ?

Estimation

Estimation deals with the classification technique changed around to involve assigned "scores" based on a pre-scored training set. This replaces assigning a binary classifier as used in classification.

Measuring the level of quality of life of local governments

How to construct the set of indicators to measure the level of quality of life[6] of local governments with emphasis on sustainability and to apply it to local governments. In addition, it attemps to identify whether it can be applied to other foreign countries or not. In measuring the level of quality of life, many statistical methods including facor analysis and AHP technique are employed.

GENETIC ALGORITHM

Basically this technique is an optimisation method using processes such as genetic combination, mutation and natural selection, based on the design of the concepts of natural evolution.

Spatial Planning, Structural Change and Regional Development policies

As numerous Europeans regions, they have been shaped through old industrial manner. These areas were ordinary built from steel and coal. These elements and the related industrial activities shaped the region[7], brought its growth and also, after almost a century its decline between the 1950s and the 1970s. From that moment on, it took a few years to the regional authorities to get a full perception of the scale of the structural crisis, as it can be seen with the first reports which have made the diagnosis of the economical evolution of the region, published at the beginning and middle of the 1960s.

NEURAL NETWORKS

Stated earlier were a collection of the most commonly used techniques and methods in Data Mining. One widely used technique, Neural Networks, is described in detail below. Neural Networking provides a new and different approach to using computers in the workplace today. By comparing it to the structure of neurons in a human brain, one can visualize the computer's architectural pattern, a huge web of electronic neurons, which send signals to each other through hundreds and thousands of connections. Neural Networks are used to determine patterns and relationships between various parameters and characteristics in information that is presented. This data, collected from various places such as market research, business decisions, is used to apply a neural network, stepping away from the traditional approaches to information processing.

Spatial Interaction Modeling of Interregional Commodity Flows

Understanding the determinants of interregional commodity flows is critical for both transportation infrastructure planning (highways, railroad tracks, river/port facilities) and regional development policies (location of activities, reducing regional disparities)[8]. Unfortunately, limited data availability has, in the past, hindered empirical research in this area. Based on input-output considerations and in order to differentiate intermediate from final commodity demands, the variables include more detailed descriptions of the economies of the origin and destination states, such as employment and value added for the commodity sector at the origin state, wholesale employment at both ends, manufacturing employment at the destination state, and population and per-capita income at both ends. In addition, the average establishment size for the commodity at the origin is intended to measure scale or diversification effects. The competitive or agglomerative effects of the economic spatial structure are captured with competing destination and intervening opportunities variables.

Data Mining of Advanced Database

Due to the amount of information companies store on their databases, the databases has gone through an advancement in technology that effect the size of the database, the speed need to operate the database, and a way to organize the data. Advancement on the database correlated with the advancement required on the data mining tools. Such tools are intelligent agents and multidimensional analysis tools. Because of the huge amount of data, the data has to be organized and any errors should be eliminated to increase process speed.

INTELLIGENT AGENTS

A huge database needs a tool that would be able to prioritize and/or filter the data that's being updated frequently. Intelligent agents do such as that and act as a secretary in filing and fetching data. There are several categories of intelligent agents available. Some agents are designed to launch automatically to perform specific queries or to search for patterns in data. Others automatically predefined intervals, performing a task or monitoring a condition in the background and returning an alert as required. IBM has proprietary, patent-pending techniques to analyse gigantic data sets for cross marketing or affinity marketing and look for patterns. These patterns may not just be the obvious, but rather ones that an analyst might miss. Other intelligent agents use fuzzy logic and neural network algorithms to help analyze and control real time technical processes. Yet another agent is able to look for correlation by forming, testing, and modifies its own hypotheses.

The regional system of innovation and regional development

Innovations and the capacity to innovate are crucial factors in the development of a firm and its ability to adapt to changes in the external environment. Growing attention has been paid to the mechanism facilitating innovation in firms, both in large, small and medium-sized enterprises. As a consequence, increasing attention has been on the role of innovation policy in regional development[8].

The purpose of this project is to analyse the linkages between the business advisory systems efforts to promote innovation and the innovative firm with special attention to small and medium sized enterprises. The analysis of this paper deals with entities and relations of the innovative environment. The focal point is the interaction between the analysed business entity and the external environment as a part of a broader network of innovative relations covering intra-firm as well as extra-firm relations and processes[9].

Multidimensional analysis tools

There are two types of multidimensional analysis tools and they are multidimensional analysis (MDA) and On-Line Analytical Processing (OLAP)[10]. The idea is to load a multidimensional server with data that is likely to be combined. MDA represents data as n-dimensional matrices called hypercubes. OLAP let the user calculate metrics over a subset of available data, by exploring combinations of one or more dimensions of data. Delta Airlines, for example, has use OLAP to gain insights into its frequent-flyer program. It is able to consolidate data from a 100-GB Teradata into six far more accessible multidimensional databases totalising 6 GB. OLAP servers are great for time-series analysis, recursive calculations, and data with up to about 15 dimensions.

Immaterial resources and regional development : a territorial approach

The analytical framework[11] is based on the idea that resources do not constitute a stock but rather a process. A process through which the objects (immaterial and material) of our surrounding environment are at a certain moment and in certain locus identified as being potentially useful for the production system[12,13]. The main questions here are the following: what are immaterial resources[14]? What are their characteristics? How do they evolve/move through time and space? How are they created? Destroyed? How are they identified? Managed? By whom? Where? How is learning (in it's different aspects) organised in time and space?

CONCLUSION

Yet one of the common problems is that databases storing company's information are infested by errors, duplicate data, and unnecessary information pertaining to the data-mining applications. For data mining to produce valid results, data has to be cleaned and structured for consistency[15]. Such consistent format could be as simple as assigning 0's and 1's, instead of 'female' and 'male'. The process of cleaning up may be a slow one, but this would pay off in the end. The diversity and integration requirements of data are still a problem and solutions are being developed. Many organizations are still grappling with how to make the whole process work and how to fit it in with other initiatives such as data warehousing and other decision support areas. We have in this paper tried to show the footbridge which existed between dated mining and local development through studies specific. Other applications are under studies.

BIBLIOGRAPHIC REFERENCES

[1]CHOO, C W & BERGERON, (2001), P "Issue editorial." [Special issue on Environmental scanning and competitive intelligence.] *Information Research*, **7**(1).

[2]BERTACCHINI, Y.,(2002), Territoire & Territorialités; Vers l'intelligence territoriale; *volet 1*, 190 pages, Collection *Les E.T.I.C,* Saint-Raphaël (Var-France).

[3]BERGERON, P., HILLER, C.,(2001), «Competitive intelligence in Cronin, B., » (ed). *Annual Review for Information Science and Technology* (ARIST). Vol 36.

[4]BERTACCHINI, Y.,(2001), DOU, H., «The Territorial competitive intelligence: a network concept», Colloque Proceedings VSST '2001 p 101, Barcelone 15-19.

[5]VON KROGH, G., ICHIJO, K., & NONAKA,(2000), I. Enabling knowledge creation. How to unlock the mysery of tacit knowledge and release the power of innovation.New York: *Oxford University Press*.

[6]BERTACCHINI, Y., LEBRETON, M., (2002),«Acteurs locaux, Réseaux et Territoire: représentation du potentiel d'action local», 70e Congrès de l'Acfas, Montréal, Canada.

[7]BARQUERO, A. (eds.),(1998), Organisation of production and territory: local models of development, Gianni Iuculano Editore, Pavia.

[8]LUNDVALL, B. A.,(1998), National systems of innovation, Pinter Publisher, London.

[9]NEMETI, F. AND PFISTER, M.,(1998), Aspects de la compétitivité de l'industrie microtechnique suisse, EDES, IRER, University of Neuchâtel, Neuchâtel.

[10]MORIN & AL., Base de données et statistiques,(2002), *Collection Sciences Sup*, 384 p, Dunod, Paris.

[11]BENOIT, J.,(1999), «Une communauté de pratique en réseau : Le forum de discussion et la base de connaissances des inspecteurs de la Commission de la santé et de la sécurité du travail» (CSST). (Collection infomètre). Québec : CEFRIO.

[12]BROWN, J. S. & DUGUID, P.,(2000), «The social life of information. », Boston, MASS : *Harvard Business School Press*.

[13]HILDRETH, P., KIMBLE, C. & WRIGHT, P.,(2000), Communities of practice in the distributed international environment. *Journal of Knowledge Management*. 4 (1), 27-38.

[14]BUMBO, N., COLEMAN, D.,(1998), Case study : Building a knowledge community at Viant : the case for using a holistic approach.

[15]HAN, J., KAMBER, M.,(2000), Data Mining: Concepts and Techniques, édition *Morgan Kaufmann Publishers*.

MAITRISE DE L'INFORMATION, AMELIORATION DES SYSTEMES DE SANTE ET AMENAGEMENT DU TERRITOIRE. L'EXEMPLE DE LA CATALOGNE (ESPAGNE) ET DE LA REGION MIDI-PYRENEES (FRANCE)

Christian BOURRET,
Université de Marne-la-Vallée, 77454 Marne-la-Vallée cedex 2, France
bourret@univ-mlv.fr

Jaume TORT i BARDOLET,
Generalitat de Catalunya, Servei Català de la Salut (CatSalut), Barcelone, Espagne
jtort@catsalut.net

Résumé : Voisines et frontalières de part et d'autre des Pyrénées, la Catalogne espagnole et la région Midi-Pyrénées ont des systèmes de santé différents. Depuis 1978, la Catalogne a affirmé son autonomie au sein de l'Espagne, notamment dans le domaine de la santé. Avec les « comarques », elle a défini ses structures territoriales de base. Plus récemment, la France a commencé à explorer la voie des « pays ».
Le système de santé catalan est basé sur les centres d'attention primaire. La France expérimente la voie des réseaux de santé. Le plan de développement stratégique du Service public Catalan de la Santé (*CatSalut*) mise beaucoup sur la création d'un système d'information performant pour améliorer son efficience et les services à ses usagers dans une perspective d'aménagement du territoire. Sans compétence institutionnelle en santé, la région Midi-Pyrénées a néanmoins favorisé le développement de la télémédecine.
Dans le cadre de l'affirmation de la société de l'information et du savoir, un nouvel État national est en train de se dessiner, en particulier en Espagne. Son rôle et les nouvelles répartitions des compétences avec les collectivités territoriales et avec le secteur privé reposent largement sur l'utilisation des technologies de l'information dans une perspective d'aménagement du territoire et d'amélioration des services à des citoyens, devenus des « clients » à fidéliser.

Abstract : Neighbour and located on both sides of the Pyrenees Mountains, Spanish Catalonia and French Midi-Pyrénées "région" have different health systems. Since 1978, Catalonia has asserted its self government within Spain, particularly in the health field. With the "comarques", Catalonia has defined its basic structures of territoriality. More recently, France has begun to experiment the way of "pays" (countries).
The Catalan health system is based on primary care centers. France is experimenting healthcare networks. The strategic development plan of the Catalan health public service *(CatSalut*) bets much on creating a high performance information system to improve efficiency in services towards users in a territory development perspective. Without

institutional competence in the field of health, Midi-Pyrénées "région" has nevertheless favoured the development of telemedicine.
Within the context of information and knowledge society development, a new national State is emerging, particularly in Spain. Its functions and new competences shared with local communities and with the private sector largely lay on the use of information technologies aiming at territory development and services improvement towards citizens. These citizens are becoming "customers" whose loyalty must be strengthened.

Mots clés : systèmes d'information, management des systèmes de santé, aménagement du territoire, réseaux, nouvelles technologies.

Key words : information systems, health care management, territory development, networks, new technologies.

Maîtrise de l'information, amélioration des systèmes de santé et aménagement du territoire. L'exemple de la Catalogne (Espagne) et de la région Midi-Pyrénées (France)

INTRODUCTION

Régions voisines, frontalières dans la zone des Pyrénées, la Catalogne espagnole (1) et la région Midi-Pyrénées relèvent de deux États nationaux différents : l'Espagne et la France. Les affinités et les relations ont été très importantes entre les deux régions qui, aux XII^{e}-XIII^{e} siècles, ont failli constituer un royaume « transpyrénéen » unique avec pour capitale Barcelone (2). Les deux régions ont eu ensuite des destins différents. L'évolution récente : création des Communautés Autonomes en Espagne à partir de 1978, régionalisation beaucoup plus lente en France, affirmation progressive de l'Union européenne, les rapproche à nouveau plus fortement.
Dans cet article, nous examinons tout d'abord les grandes lignes et les spécificités des deux systèmes de santé dont relèvent ces régions. Nous analysons ensuite les évolutions récentes et les choix stratégiques du système de santé catalan, à la fois en termes de systèmes d'information et d'aménagement du territoire, puis les évolutions du système de santé français et les particularités de la région Midi-Pyrénées. Nous étudions enfin comment, dans le cadre d'une Europe des régions, ces deux systèmes peuvent converger en s'appuyant sur les nouvelles technologies de l'information et de la communication pour favoriser une politique de santé de proximité, élément majeur d'aménagement du territoire et de citoyenneté.

1 - FRANCE ET CATALOGNE : DEUX SYSTEMES DE SANTÉ DIFFÉRENTS

Le système de santé français dont relève la région Midi-Pyrénées et le système de santé catalan ont deux histoires très spécifiques. Ils relèvent de deux choix politiques différents. Le système français, mis en place à partir de 1945 dans le cadre de l'affirmation de l'État providence (la Sécurité Sociale), est un système dit bismarckien (du nom du chancelier allemand du troisième quart du XIX^{e} siècle : Bismarck). Il est basé sur le paritarisme (cogestion par les organisations patronales et les syndicats) et est financé par des cotisations sociales. Il est donc basé sur le travail. Le système catalan, beaucoup plus récent, relève d'un autre modèle : celui d'un système national de santé, à l'image de celui mis en place à partir des recommandations de lord Beveridge en Grande-Bretagne (*National Health Service*). Il est financé par l'impôt et son critère d'affiliation est la résidence.
Les coûts des deux systèmes de santé sont différents : environ 8000 FF (1200 euros) par an pour un Catalan en 1998 contre environ 12 000 FF (1800 euros) pour un Français. Mais, en raison du vieillissement de la population et des besoins d'investissement (nouvelles structures, notamment hôpitaux, et équipements) ceux de la Catalogne vont augmenter très vite alors que ceux de la France ne sont pas maîtrisés.

Le système français comporte deux volets très différents et qui, trop souvent, s'opposent : la médecine de ville et l'hôpital. Y coexistent un secteur public majoritaire et un secteur privé (cliniques mais aussi service privé au sein de l'hôpital public). Pour la médecine de ville, le système est dit « libéral ». Le patient a le libre choix de son médecin qui peut s'installer sur tout le territoire. Le patient peut accéder directement (sans l'intermédiaire d'un médecin généraliste) à la fois aux médecins spécialistes et à l'hôpital, notamment par les urgences. Plus de 80 % des Français sont affiliés à des caisses complémentaires (mutuelles et compagnies d'assurances) qui remboursent les prestations au-delà du tarif de la Sécurité Sociale mais n'assurent pas de prestations médicales supplémentaires.

Bismarckien à l'origine et basé sur le travail, c'est-à-dire avec l'idée d'assurer des revenus de remplacement en cas de maladie et non de gérer la santé, le système français a évolué. Avec la mise en place d'un financement également par l'impôt (CSG : Contribution Sociale Généralisée), de la LFSS (Loi de Financement de la Sécurité Sociale), de la CMU (Couverture Maladie Universelle) et le départ du MEDEF (Mouvement des Entreprises de France) des organismes de gestion paritaire de l'Assurance Maladie, il est devenu un système quasi national. Par ailleurs, le système français est fragmenté en de très nombreux organismes et régimes différents : Mutualité Sociale Agricole, CANAM (caisse des professions indépendantes), Mines, SNCF … Le principal régime, qualifié assez abusivement de « général », relève de la CNAMTS (Caisse Nationale d'Assurance Maladie des Travailleurs Salariés) et de ses différentes structures locales, en premier lieu les CPAM (Caisses Primaires d'Assurance Maladie). La CNAMTS n'a qu'un rôle de remboursement (à la fois des médecins et des usagers) et très peu de gestion. C'est l'État (ministère de la Santé) qui gère directement l'hôpital qui représente 46 % des dépenses de santé.

Les ordonnances d'avril 1996 ont voulu faire évoluer le système français vers la régionalisation : affirmation du rôle des URML (Unions Régionales des Médecins Libéraux) et, surtout, création des URCAM (Unions Régionales des Caisses d'Assurance Maladie), qui ont un rôle de coordination des différents régimes mais pas de gestion directe, et des ARH (Agences Régionales de l'Hospitalisation), ayant en charge la supervision à la fois des secteurs public (hôpitaux) et privé (cliniques).

Ces ordonnances ont également voulu faciliter l'expérimentation des réseaux de santé. Très fortement individualistes, les médecins français sont en Europe les moins intégrés dans des structures de groupe. Les réseaux de santé, d'origines très diverses, ont été consacrés par la loi du 4 mars 2002 sur les *Droits des malades et la qualité du système de santé*. Ils constituent une voie privilégiée pour décloisonner le système de santé (généralistes, spécialistes, hôpitaux publics de statuts et de tailles très différents, cliniques …), assurer la continuité des soins et essayer de mettre (enfin !) le système de santé au service des patients et non des organisations et des professions. Mais il y a d'autres voies possibles. On parle aussi de « maisons médicales » notamment pour régler le problème des services des urgences dans les hôpitaux et cliniques ou de cabinets de groupes pluriprofessions sur le modèle des structures de proximité relevant des *Primary Care Trusts* du *N.H.S.* britannique ou des Centres d'Attention Primaire catalans, sur lesquels nous reviendrons. La réduction du temps de travail a accentué les tensions dans les hôpitaux et rend particulièrement urgente la recherche de solutions opérationnelles principalement basées sur le « virage ambulatoire » : transfert de certaines activités de l'hôpital vers la médecine de ville. Par ailleurs, l'extension de la régionalisation et en particulier la création d'ARS (Agences Régionales de Santé) est évoquée.

Le système catalan est beaucoup plus récent. Il est aussi très différent. La nouvelle constitution espagnole de 1978 (le général Franco étant mort en 1975) ou statut des autonomies a comme axe majeur la coopération de l'État national (Madrid) et de 17 Communautés Autonomes. La Généralité de Catalogne avait été rétablie en 1977 (3). Un système national de santé financé par l'impôt fut mis en place sur le modèle du *NHS* britannique, un système privé continuant à subsister. Un *Instituto Nacional de la Salud (Insalud)* fut créé pour essayer de faire converger systèmes publics et privés et assurer les relations avec les différentes Communautés Autonomes. D'autres entités continuèrent à exister, comme par exemple la *Muface* (Mutuelle de Fonctionnaires de l'Administration de l'État) ou, pour l'Armée et la Marine.

La constitution de 1978 précisait que la législation de base de la sécurité sociale relevait de l'État national mais que l'exécution des services pourrait être transférée aux Communautés Autonomes. En fait, il allait s'agir d'un véritable transfert de compétences. Ce transfert de compétences fut effectif pour la Catalogne dès 1981, pour l'Andalousie en 1984, le Pays Basque et Valence en 1988, la Galice et la Navarre en 1994, les Iles Canaries en 1994.
Première Communauté Autonome espagnole à avoir obtenu le transfert de la gestion de la santé publique, la Catalogne définit sa carte sanitaire en 1983 et créa en 1985 la *Xarxa Hospitalària d'Utilització Pública* (XHUP) ou réseau d'hospitalisation publique, avec des structures principalement héritées d'*Insalud*. Le système national de santé catalan fut créé en 1986. Il présente deux caractéristiques principales : séparation stricte du financement et des fournisseurs de soins, diversité et mixité des structures (système public, système privé sans profit (4), système privé à vocation de profit). Pour planifier, financer et évaluer les services de santé du secteur public fut créé en 1990 le *Servei Català de la Salut*, devenu depuis *CatSalut*. Système national public et organismes privés coexistent donc en Catalogne comme dans toute l'Espagne. On peut parler de système mixte diversifié. Mais la séparation public / privé, tant au niveau du financement que des coopérations, est beaucoup plus tranchée qu'en France. A l'exception des fonctionnaires (*Muface*) qui peuvent choisir, les Espagnols financent les systèmes de santé publics des Communautés Autonomes en tant que contribuables et paient une seconde fois s'ils souhaitent bénéficier de prestations privées. A ce jour, le réseau de santé public catalan propre à la Généralité assure 84 % des soins primaires, 30 % de l'hospitalisation et 5 % des centres de santé mentale.

Pour le reste, *CatSalut* passe contrat avec des organisations privées, le plus souvent du secteur « privé sans profit ». Les contrats avec le secteur « privé à profit » sont beaucoup plus limités : au transport (ambulances), aux soins à domicile, à la dialyse, à l'imagerie médicale … et aux petites équipes d'assistance primaire autogérées (EBA).

CatSalut, qui a vocation à coordonner tout le secteur d'utilisation publique en Catalogne a créé des consortiums pour favoriser la rénovation des anciens hôpitaux municipaux : Vilafranca del Penedés, Igualada, Vic, Blanes-Calella, Sabadell … Barcelone constitue un cas particulier. La municipalité y possède deux hôpitaux et quelques centres d'attention primaire qui sont gérés par des organismes municipaux : IMAS ou PAMEM (5) avec la participation financière de la Généralité.

La Catalogne a défini son premier Plan de Santé triennal (*Plan de Salud de Cataluña*) en 1993. Le 3e plan a couvert la période 1999-2001, le 4e couvre la période 2002-2004. Le principal défi est de répondre au vieillissement de la population. Le système de santé public est articulé autour des 360 Centres d'Attention Primaire (CAP) appelés *ambulatorios* dans le reste de l'Espagne. Ils regroupent des médecins généralistes (une spécialité en Espagne), des pédiatres et des infirmières, dont le rôle est beaucoup plus valorisé qu'en France. Sur des critères de résidence, tous les Catalans sont affiliés à un CAP. Pour le moment, ils ne peuvent pas en changer. Les médecins généralistes qui sont fonctionnaires (dans la partie réseau public de soins primaires relevant de la Généralité) doivent exercer 6 heures par jour dans le service public. Leur rémunération étant relativement modeste, ils peuvent ensuite exercer en privé. Dentistes et pharmaciens relèvent uniquement du secteur privé. Les CAP gèrent l'accès aux spécialistes et à l'hôpital. Les files d'attente pour certaines opérations à l'hôpital constituent un point particulièrement sensible qui peut faire basculer un pourcentage non négligeable de la population vers un secteur privé déjà largement majoritaire dans le domaine des soins secondaires (hôpitaux privés et cliniques).

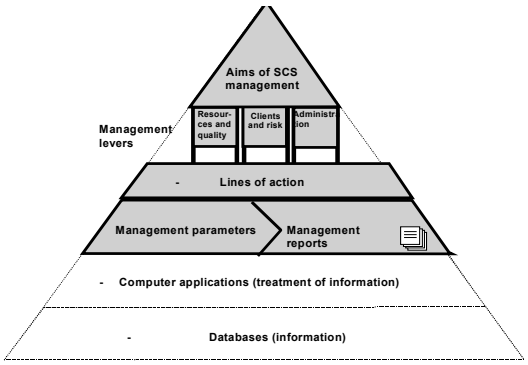
Depuis le 1er janvier 2002, l'intégralité de la gestion des services de santé est déléguée aux 17 Communautés Autonomes. Le rôle des Communautés Autonomes dans le domaine de la santé et dans son articulation avec l'aménagement du territoire ne peut que s'accentuer.

2 - ÉVOLUTIONS RÉCENTES ET CHOIX STRATÉGIQUES EN TERMES DE SYSTÈMES D'INFORMATION ET D'AMÉNAGEMENT DU TERRITOIRE

Le système de santé public catalan comprend près de 360 centres d'attention primaire, 66 hôpitaux (XHUP), 83 centres sociaux et 19 hôpitaux psychiatriques répartis en 8 régions sanitaires différentes : Lleida, Girona, Centre, Barcelonès Nord i Maresme, Barcelona Ciutat, Costa de Ponent, Tarragona, Tortosa. Les CAP constituent à la fois la structure de proximité et la base du système. Depuis 1987, la Catalogne a pu se doter (mais sans supprimer les 4 provinces imposées par Madrid en 1833) de son organisation territoriale de base particulière : les 41 « comarques ». Administrées par un président et un conseil comarcal élus, elles ont certaines attributions spécifiques, notamment pour la gestion du patrimoine et du tourisme, mais n'ont à ce jour aucune compétence en santé.

Le système de santé public dispose de deux instruments majeurs de pilotage : au niveau global, les différents plans de santé et, à un niveau plus opérationnel, le plan stratégique de *CatSalut*. Les journées de travail d'avril 2002 consacrées à la définition des lignes stratégiques de *CatSalut* (6) ont été axées sur la recherche du meilleur modèle de gestion pour la santé, avec deux pistes privilégiées. La première est le recours aux méthodes de gestion des entreprises privées (« reengineering » et gestion par processus, management par la qualité, gestion par projet, gestion des savoirs et des compétences …) mais dans un but d'efficience (efficacité au meilleur coût) dans le sens de l'intérêt général des usagers du système de santé et non de profit. La seconde est l'utilisation des nouvelles technologies de l'information et de la communication. L'ensemble se place dans une perspective d'aménagement du territoire : équité en termes de qualité et d'accès aux soins pour les différentes régions sanitaires et « comarques » (objectif : pas de Catalan à plus de 30 minutes d'un hôpital) et implication plus forte des usagers notamment dans les structures de proximité. En fait, *CatSalut* se positionne comme une *Health Maintenance Organization* (H.M.O.) américaine en appliquant les méthodes de *Managed Care* mais sans objectif de profit.

L'outil majeur de cette politique est la mise en place du système d'information de *CatSalut*. Au début, *CatSalut* a utilisé les systèmes d'information des fournisseurs de soins. Il a décidé de créer le sien propre et d'en faire un outil stratégique de planification, de coordination et d'évaluation du système de santé catalan. Commencé il y a deux ans et basé prioritairement sur les technologies internet, il devrait être opérationnel en 2003.
Le système peut être schématisé par une pyramide à trois niveaux. Le premier étage (la base du système) est celui des bases de données (supports d'information). Le niveau du milieu est celui du traitement de l'information. Le niveau supérieur est celui du management du système d'information.



Main elements of the SCS management and information

Le projet se décompose en deux phases distinctes. Une première étape vise à articuler étroitement le nouveau système d'information avec ceux des fournisseurs de soins. La seconde étape est beaucoup plus ambitieuse : elle doit assurer le lien direct entre les citoyens catalans et les fournisseurs de soins. Le projet entre dans le cadre plus vaste de la création d'une véritable *Administració Oberta de Catalunya* basée sur l'utilisation des technologies internet avec une seule carte d'entrée pour toutes les prestations de tous les services de la Généralité de Catalogne.

Un des éléments importants du projet est le centre d'appels *Sanitat Respon*. Conçu sur le modèle du *NHS Direct* britannique, il a commencé à fonctionner pour la ville de Barcelone en octobre 2001. L'objectif est d'en faire un véritable centre d'orientation 24 heures sur 24 : un centre de contact permanent en santé pour l'ensemble de la population catalane. En s'inspirant encore des réalisations du *NHS*, les NTIC vont également constituer un outil majeur pour l'amélioration du traitement des urgences dans les hôpitaux.

En comparaison, les réalisations de la région Midi-Pyrénées sont plus modestes. Entre les deux régions, les degrés d'autonomie, comme les moyens financiers, sont très différents. Consacrées comme collectivités territoriales par les lois de décentralisation de 1982 (avec les départements et les communes), les régions ont des compétences limitées à la formation professionnelle, au développement économique, aux transports, à la culture, à l'aménagement du territoire et à l'entretien des bâtiments (lycées). Elle n'ont pas de compétences en santé. La politique de régionalisation en santé a été jusqu'à présent menée dans un cadre de déconcentration de services de l'État, notamment avec la création des ARH (Agences Régionales de l'Hospitalisation) et non d'autonomie régionale (décentralisation).

Il convient néanmoins de signaler le développement de la télémédecine par la région Midi-Pyrénées. Le principe de la télémédecine est de ne plus faire déplacer le patient mais circuler l'information. Elle repose sur la transmission de données, notamment multimédia (imagerie médicale) pour l'aide au diagnostic, à la prescription, l'obtention d'un second avis médical, et la télésurveillance des malades à domicile. « La télémédecine favorise un accès égal aux soins et un droit de tout citoyen à la santé en tout point du territoire. Il en résulte que la santé doit être à la fois une politique sociale, une politique territoriale et l'expression d'une citoyenneté en acte » (7). En septembre 1998 a été créé le GCS RTR (Groupement de Coopération Sanitaire Réseau Télémédecine Régional) réunissant 52 établissements publics et privés de la région Midi-Pyrénées : CHU de Toulouse, Centre Régional de lutte contre le cancer, centres hospitaliers de Rodez, Cahors, Lourdes, Luchon, centre hospitalier intercommunal de Val d'Ariège … ensuite Auch, Saint-Gaudens, Figeac, Albi, Lannemezan, Castres … puis, à terme, tous les établissements publics et privés de Midi-Pyrénées.

Des politiques de proximité dans le domaine de la santé ont été mises en place grâce à des initiatives locales, en particulier réseaux de santé, émanant à la fois de médecins ou d'organismes de protection sociale comme des Caisses Primaires d'Assurance Maladie (CPAM), par exemple le réseau de soins palliatifs des Hautes-Pyrénées. Une Coordination Régionale de Réseaux de Santé concernant des réseaux moins institutionnels (souvent précarité ou toxicomanie) a vu le jour. Un Comité Régional des Réseaux Midi-Pyrénées regroupant des représentants des organismes de protection sociale (URCAM, CPAM, MSA) et des services déconcentrés de l'État (ARH ou DDASS) s'affirme à la fois comme structure de coordination de l'attribution des financements publics aux différents réseaux de santé de Midi-Pyrénées et comme structure de réflexion et d'incitation de convergence des expériences vers des organisations plus généralistes. La réalisation du réseau régional Diamip (diabète en Midi-Pyrénées) est soutenue conjointement par l'ARH et l'URCAM, préfigurant peut-être les futures ARS.

Au niveau local, les lois de 1995 et de 1999 ont favorisé la renaissance des « pays », en les liant à l'intercommunalité. Les « pays » avaient été balayés par la grande tourmente égalitariste de la Révolution française qui avait assimilé les particularismes locaux (structures d'autonomies, dialectes …) à l'Ancien Régime (8). En février 2001, la Conférence Régionale d'Aménagement et de Développement du Territoire (CRADT) de Midi-Pyrénées a reconnu comme espace de projet le périmètre de « pays Couserans », celui de « Foix-Haute Ariège » et celui de « Pays d'Olmes-Mirepoix ». A terme, le département de l'Ariège devrait compter quatre « pays ». En Haute-Garonne, la même CRADT a défini le périmètre d'un « pays Comminges ».

La renaissance des nouveaux « pays » risque de bouleverser l'ensemble des limites de juridiction des structures nées de la Révolution française (cantons, arrondissements et départements). Par exemple, le périmètre du « pays Couserans » ne correspond pas aux limites de l'arrondissement de Saint-Girons. Il regroupe 7 cantons : les 6 de l'arrondissement de Saint-Girons et celui de La-Bastide-de-Sérou (Séronnais), relevant de l'arrondissement de Foix.

Mais quel est l'avenir de ces « pays » ? Doivent-ils simplement constituer un cadre de référence pour les projets des services déconcentrés de l'État ? Ou, tout au contraire, doivent-ils constituer de nouvelles collectivités territoriales aux compétences à définir ? Avec alors le risque, typiquement français, de créer une nouvelle entité territoriale, sans en supprimer d'anciennes alors que la France compte déjà trop d'entités territoriales.

En France, dans le cadre de ce que l'on a appelé dans les années 1980-1990 au Canada le « virage ambulatoire » informatisé, le « pays » peut constituer le lieu privilégié du mariage des NTIC et de la santé. Le « virage ambulatoire » se propose d'essayer de maîtriser le dérapage des dépenses de santé et notamment de l'hôpital en recourant aux NTIC pour lutter contre les cloisonnements et les dysfonctionnements du système de santé en privilégiant la médecine de ville et en mettant en réseaux techniques les établissements entre eux, mais aussi la médecine de ville et les hôpitaux ou les cliniques (réseaux de santé) et en développant la télémédecine.

Les inégalités d'accès au système de santé se sont accentuées, non seulement en termes de revenus, mais aussi en termes géographiques et d'accès à l'information. Le rapport sur le *Paysage sanitaire français à l'horizon 2020* (9) parle de la nécessité d'une politique active pour faire évoluer le système vers le fonctionnement en réseau. En particulier dans les zones rurales, il appelle au « développement d'un fonctionnement en réseau, notamment des généralistes, prenant appui, lorsqu'il existe, sur un hôpital local renforcé ».

La santé est devenue un enjeu de l'aménagement du territoire mais aussi de l'emploi. Dans beaucoup de petites villes, l'hôpital est désormais le principal (et souvent le dernier !) employeur. C'est notamment le cas à Saint-Girons, à Saint-Gaudens, à Lannemezan, à Lavaur, à Auch, à Cahors ou à Rodez. Mais c'est aussi le cas pour beaucoup de grandes villes : dans 14 régions sur 22, l'hôpital est le premier employeur (10). A Toulouse, il emploie plus de personnes que l'industrie aéronautique.

Dans le cadre des nouveaux « pays », en s'affirmant comme un pôle de compétences (cœur d'un réseau de santé ou de toute autre structure de coopération forte ville-hôpital) et en liaison notamment avec le développement des soins à domicile face au vieillissement des populations (accentué par le retour de beaucoup à l'âge de la retraite) et au développement de pathologies liées au vieillissement, l'hôpital peut devenir en termes de service public un pôle majeur de la politique d'aménagement du territoire et de l'identité des « pays ». Un réseau de santé ville-hôpital Comminges fonctionne et un autre réseau est envisagé à partir de l'hôpital de Saint-Girons.

3 - VERS UNE CONVERGENCE DES POLITIQUES DE PROXIMITÉ EN SANTÉ GRÂCE AUX NOUVELLES TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION ?

L'État-nation, tel qu'il s'est imposé au XIX^{e} siècle et a étendu son emprise dans la seconde moitié du XX^{e} siècle, avec notamment le modèle français d'un État-providence dirigiste, est doublement remis en cause. Il est tout d'abord fortement concurrencé par la « mondialisation » et l'affirmation d'aires supranationales : dans le cas de la France et de l'Espagne, l'Union européenne. L'État-nation qui, au XIII^{e} siècle, avait commencé à s'affirmer dans le domaine monétaire en supplantant les monnaies féodales, vient de perdre cette prérogative essentielle : l'euro a remplacé le franc et la peseta au 1er janvier 2002.

L'État national est à la fois trop grand pour répondre aux problèmes quotidiens des citoyens et trop limité à l'échelle de la mondialisation. Concurrencé par l'Union européenne et par les multinationales de la communication, l'État-nation est également fortement ébranlé par la réémergence du local. C'est tout particulièrement vrai en Espagne où les Communautés Autonomes (notamment en Catalogne et au Pays Basque) ont affirmé leur importance et leurs spécificités. Cela commence à être aussi vrai en France avec les risques de cloisonnements que cela comporte.

En Grande-Bretagne également, l'Écosse a obtenu la reconnaissance de sa spécificité dans le domaine de la santé. Il existe désormais un *NHS Scotland*. Il a établi une stratégie particulière, fortement basée sur l'utilisation des technologies de l'information et de la communication, qui s'est traduite par de remarquables réalisations : réseau d'information (*SHOW : Scottish Health On the Web*), d'éducation et de prévention (*HEBS : Health Education Board for Scotland*).

Une nouvelle répartition des pouvoirs au niveau local et des compétences de l'État et des collectivités territoriales a eu lieu en Espagne. En France, elle commence à peine dans le cadre d'une Union européenne qui sera largement une Europe des régions. L'État national a fait la France, cela n'a pas été le cas en Espagne où les identités régionales (qui s'affirment comme « nationales ») ont toujours été très fortes, notamment dans ses régions périphériques. Son poids devrait demeurer plus fort en France.

C'est en fait tout un nouveau maillage territorial qui commence à se dessiner. Dans le domaine de la santé, en s'appuyant sur les réseaux et la télémédecine, la DATAR parle de « polycentrisme maillé » et réfléchit à une organisation idéale qui irait du niveau régional (et même interrégional européen) jusqu'au niveau de proximité des pays, centrés sur les hôpitaux locaux (10).

Se pose la question de la meilleure structure d' « intermédiation » et de services de santé de proximité pour des usagers devenus des partenaires, voire à terme des cogestionnaires du système de santé : centres d'attention primaire, réseaux de santé ou cabinets de groupe pluriprofessions. Dans le cadre des « comarques » ou des « pays », ces nouvelles structures de santé de proximité peuvent constituer le bon niveau d'implication des citoyens (« démocratie sanitaire »), voire de concurrence régulée. Le nouveau système d'information de *CatSalut* devrait à terme permettre aux Catalans de choisir leur CAP d'affectation et d'en changer s'ils n'en sont pas satisfaits. La Catalogne favorise aussi l'émergence de consortiums (organismes privés à but non lucratif) gérés par les médecins eux-mêmes et propose de les conventionner avec les CAP.

La Généralité de Catalogne a fait de gros efforts d'investissement pour marier santé et aménagement du territoire en termes de « pays », par exemple en construisant l'Espitau du Vall d'Aran (Vielha) et l'Hospital Comarcal del Pallars (Tremp). Jusqu'à la création de l'hôpital de Vielha, des Aranais, notamment en hiver, étaient transportés à l'hôpital de Saint-Gaudens. Désormais, ce seront peut-être des Commingeois de Fos ou Saint-Béat qui iront à Vielha.

La Cerdagne, divisée en deux depuis 1659 et si spécifique, constitue un cas particulièrement intéressant des relations entre le domaine de la santé et la politique de « pays ». Pourquoi des habitants de Cerdagne française (Font-Romeu, Saillagouse …) ne se feraient-ils pas soigner à Puigcerdà, dont l'hôpital a été considérablement amélioré par la Généralité ? Les problèmes de refacturation entre États nationaux ne sont guère différents de ceux entre Communautés Autonomes en Espagne. La « réunification » de la Cerdagne commencera-t-elle par les coopérations transfrontalières dans le domaine de la santé pour, à terme, déboucher sur un « pays » cerdan réunifié, avec pour chef-lieu Puigcerdà ? En Cerdagne, la mise en place des « pays » pourrait ainsi bousculer les frontières nationales.

En sens inverse, les files d'attente dans les hôpitaux catalans pour certaines opérations peuvent pousser certains patients à aller se faire opérer en France (Perpignan, Toulouse ou Montpellier). *CatSalut* multiplie les efforts pour les éviter.

Le développement de la télémédecine a déjà bousculé les vieilles frontières nationales au niveau des grands centres d'excellence de Toulouse et de Barcelone, en attendant sa prochaine diffusion au niveau des hôpitaux locaux et d'autres structures de proximité. Des coopérations institutionnelles commencent à être envisagées entre *CatSalut* et des organismes de protection sociale de la région Languedoc-Roussillon.

Mais, si elle pose le problème de l'articulation avec les structures territoriales de proximité (« comarques » ou « pays »), l'affirmation des autonomies régionales pose avant tout celui des compétences laissées aux États nationaux. Ils doivent demeurer les garants d'une certaine solidarité nationale, en veillant à établir des compensations financières entre régions riches et régions plus pauvres. Ils doivent notamment fixer le niveau de prestations minimales pour tous les citoyens. En Espagne, le ministère de la Santé de Madrid devrait définir un « panier de soins » national minimal pour toutes les Communautés Autonomes. De leur côté, les autonomies régionales doivent organiser la concurrence (définition de cahiers des charges) entre structures de soins publiques et structures privées et la réguler.

Se pose aussi la question de l'articulation de ces autonomies régionales et des structures de santé de proximité. Ces structures de proximité peuvent être le lieu d'adaptation de procédures et de diffusion d'informations locales : gardes, urgences, spécificités des populations … notamment à travers des portails de santé adaptés.

CONCLUSION

En Catalogne et dans une moindre mesure en France, la renaissance des collectivités territoriales constitue une sorte de revanche de l'Histoire et des traditions locales sur le centralisme égalisateur des États-nations. L'affirmation des communautés autonomes et des régions en a été la première étape. Celle des « comarques » et des « pays » peut en être la seconde. Si elle sait concilier identité, proximité et ouverture, cette revanche peut être riche d'avenir par sa rencontre avec les NTIC au sein de la nouvelle société de l'information et du savoir qui bouleverse chaque jour un peu plus les structures et les mentalités.

C'est une solution d'équilibre. Dans une société mondialisée et « virtualisée », les individus ont de plus en plus besoin de repères concrets. Le « pays » est le bon niveau pour la restauration de liens sociaux. En ce début de XXI^{e} siècle, il s'agit d'un enjeu vital. Pour le moment, leur spécificité est davantage reconnue en Catalogne qu'en Midi-Pyrénées. Les « comarques » n'ont néanmoins pas de compétences dans le domaine de la santé.

Les coopérations transfrontalières comme en Cerdagne ou en Val d'Aran – Haut Comminges vont se développer et effacer les vieilles frontières nationales, sans parler des réseaux de transmission d'informations du type télémédecine.

Un nouvel État national est en train de se dessiner. Il sera un garant et un régulateur, notamment d'actions menées par le secteur privé. Il reposera largement sur la coopération avec les collectivités territoriales avec des domaines de compétences spécifiques bien définis.

Les aspirations à l'émergence de structures de proximité sont très fortes en France comme en Espagne. La société dite de l'information et du savoir rejoint la volonté de démocratie locale, y compris sanitaire, et d'amélioration des services publics au bénéfice des usagers dans les cadres territoriaux de leur vie quotidienne. Avec l'aide des nouvelles technologies de l'information, la santé peut constituer un des premiers domaines privilégiés d'application de la nouvelle citoyenneté locale et un des volets essentiels de la politique d'aménagement du territoire.

NOTES

(1) Il existe aussi une Catalogne française. La Catalogne, soulevée une fois de plus contre la Castille en 1640, avait fait appel à la France. Le traité dit des Pyrénées de 1659-1660 la partagea. Les Catalans parlent de « mutilation ». La Catalogne du Nord (Roussillon et une partie de la Cerdagne) forma une nouvelle province française avec pour capitale Perpignan. En 1790, ces territoires formèrent l'essentiel du département des Pyrénées-Orientales. Les 5/6 de la Catalogne sont demeurés espagnols. Cette Catalogne avec pour capitale Barcelone rassemble 6 millions d'habitants. Les Pyrénées-Orientales en comptent 370000. Depuis 1977, la Catalogne a fortement réaffirmé sa spécificité au sein de l'Espagne des Autonomies, les Pyrénées-Orientales relevant de la région Languedoc-Roussillon.
C'est l'évolution du système de santé cette Catalogne du Sud ou Catalogne espagnole que nous étudions dans cet article.

(2) Dans le cadre de la croisade contre les Albigeois (ou Cathares), la bataille de Muret (sud de Toulouse) en septembre 1213 et la mort pendant le combat du comte de Barcelone et roi d'Aragon Pierre II, marqua la fin de la tentative de royaume « transpyrénéen » de la maison de Barcelone qui remontait à 1060. En 1271, le comté de Toulouse fut définitivement rattaché à la couronne de France.

Pour les liens privilégiés (historiques, culturels, linguistiques, commerciaux …) entre la région Midi-Pyrénées et la Catalogne, voir BOURRET (Christian), *Les Pyrénées centrales du IX^{e} au XIX^{e} siècles. La formation progressive d'une frontière*, préface de LE ROY LADURIE (Emmanuel), 1995, 463 p.

(3) La Généralité de Catalogne (*Generalitat de Catalunya*) constitue le symbole majeur de l'autonomie et de l'identité catalanes. Elle est apparue dès 1359. Les « Corts » catalanes ou « Cortes » en castillan, s'étaient progressivement affirmés au cours du XIII^{e} siècle. Regroupant des représentants du clergé, de la noblesse et de certaines villes, ils avaient comme fonction principale de consentir à l'impôt levé par le comte de Barcelone également roi d'Aragon et de traiter certaines affaires de gouvernement. Les convocations des « Corts » étaient irrégulières. En 1359, Pierre IV dut accepter la création de la Généralité, comme représentation permanente des « Corts », qui assurerait la continuité dans l'intervalle des sessions des « Corts » et les préparerait. La Généralité de Catalogne fut supprimée par le roi Bourbon Philippe V en 1714. Rétablie en 1931, elle fut supprimée par le général Franco en 1939. Elle a été rétablie en 1977.

(4) Le secteur dit « privé sans profit » traduit l'histoire spécifique de la Catalogne. Il correspond à des institutions caritatives ou mutuelles mises en place par l'Eglise, la Croix Rouge, les syndicats ouvriers, le patronat.

(5) IMAS : Institut Municipal d'Assistance Sanitaire.

PAMEM : Institut de Prestacions d'Assistència Mèdica al Personal Municipal.

(6) *Noves perspectives del model sanitari català. Les línes estratègiques del CatSalut,* Generalitat de Catalunya, Sitges, 15-16 avril 2002.

(7) Professeur Louis Lareng, *SANTETIC : Les Technologies de l'Information et de la Communication dans le domaine de la Santé au sein du grand Sud-Ouest*, 2000, p. 104

(8) BOURRET (Christian), *La renaissance des « pays » et la société de l'information : une revanche de l'Histoire tournée vers l'avenir ? L'exemple des Pyrénées centrales*, Salon du Livre de l'Ariège, Mirepoix, 2002, à paraître, 17 p.

(9) POLTON (Dominique) sous la dir. de, *Quel système de santé à l'horizon 2020 ?*, La Documentation française, octobre 2000, 358 p.

(10) GLATRON (Marion), JACOB (Jean-Yves), VIGNERON (Emmanuel), « Santé publique et aménagement du territoire », *11e festival international de Géographie de Saint-Dié*, juillet 2000, 14 p.

BIBLIOGRAPHIE

ANDREU SEGURA (Benedicto), « Salut i Sanitat a Catalunya », *Benestar i polítiques socials a Catalunya. La Societat Catalana*, 1997, pp. 655 – 667.

BOURRET (Christian), LAURENT (Daniel), SCARBONCHI (Elisabeth), « Une réponse en termes de systèmes d'information aux défis de la protection sociale : les réseaux de santé », *V.S.S.T.' 2001 : Veille Stratégique, Scientifique et Technologique*, Barcelone, octobre 2001, Actes, tome II, pp. 23 – 32.

BRENDER (Anton), *La France face à la mondialisation*, Repères, La Découverte, 1998, 128 p.

CARRÉ (Dominique), LACROIX (Jean-Guy) sous la direction de, *La santé et les autoroutes de l'information. La greffe informatique*, L'Harmattan, juillet 2001, 312 p.

DUBOIS (Jean-Pierre), « Une révolution territoriale silencieuse : vers une nouvelle répartition des pouvoirs », *Esprit*, janvier 2002, pp. 122-136.

GLATRON (Marion), JACOB (Jean-Yves), VIGNERON (Emmanuel), « Santé publique et aménagement du territoire », *11e festival international de Géographie de Saint-Dié*, juillet 2000, 14 p.

GROS (Josette), *Santé et nouvelles technologies de l'information*, rapport adopté par le Conseil Economique et Social, Paris, 10 avril 2002.

IBERN (Pere), *El papel de la competencia en el Sistema Nacional de Salud*, Universitat Pompeu Fabra, Barcelone, 2000, 11 p.

LAMBERT (Denis-Clair), *Les systèmes de santé. Analyse et évaluation comparée dans les grands pays industriels*, Economie humaine, Seuil, avril 2000, 529 p.

LUCAS GABRIELLI (Véronique), NABET (Norbert), TONNELIER (François), « L'offre de soins de proximité : le modèle catalan », dans *Les soins de proximité : une exception française ?*, CREDES, juillet 2001, pp. 59-65.

Noves perspectives del model sanitari català. Les línes estratègiques del CatSalut, Generalitat de Catalunya, Sitges, 15-16 avril 2002.

Plan de Salud de Cataluña 1999-2001, Generalitat de Catalunya, Departament de Sanitat i Seguretat Social, marzo 2000, 196 p.

POLTON (Dominique) sous la dir. de, *Quel système de santé à l'horizon 2020 ?*, DATAR-CREDES, octobre 2000, La Documentation française, 358 p.

SANTETIC : Les Technologies de l'Information et de la Communication dans le domaine de la Santé au sein du grand Sud-Ouest, Mission Interministérielle Interrégionale d'Aménagement du Territoire, Inter-Images, juin 2000, 139 p. + annexes.

Servei Català de la Salut, rapport d'activité 2000, Generalitat de Catalunya, octobre 2001, 59 p.

VIGNERON (Emmanuel), *Santé et territoires*, Paris, La Documentation française, septembre 2000.

NOUVEAUX METIERS DANS LE DOMAINE DE LA SANTE : MAITRISE DE L'INFORMATION, TRANSVERSALITE DES COMPETENCES ET AUTRES EXIGENCES

BOURRET Christian,
Université de Marne-la-Vallée, Cité Descartes, Champs-sur-Marne
bourret@univ-mlv.fr

SALZANO Gabriella,
Université de Marne-la-Vallée, Cité Descartes, Champs-sur-Marne
gabriella.salzano@univ-mlv.fr

CALISTE Jean-Pierre
77454 Marne-la-Vallée Cedex 2
** Université de Technologie de Compiègne, 60205 Compiègne Cedex
jean-pierre.caliste@utc.fr

Résumé : La crise des systèmes de protection sociale dans les pays développés, l'émergence de nouvelles technologies, notamment de l'information et de la communication, la nécessité de maîtriser les coûts ainsi que l'affirmation du rôle des patients favorisent l'apparition dans le domaine de la santé de nouvelles organisations transversales et de proximité : réseaux de santé ou cabinets de groupes pluriprofessions.
Au sein de ces organisations collégiales et innovantes, dont la vocation principale est de créer et de gérer les liens entre plusieurs partenaires, de nouveaux métiers émergent. En utilisant différentes perspectives, nous analysons et comparons plusieurs expériences, menées en France et à l'étranger, pour déterminer les principales exigences conduisant à la caractérisation de ces nouveaux profils : décloisonnement, coopération et transversalité, gestion de l'information et maîtrise des connaissances, innovation et professionnalisation, qualité.

Abstract : The health domain must cope with striking changes, in social, economical, scientific and technological dimensions : on the one hand, the crisis of the social protection system in the developed countries, with the costs containment requirement, the more active part the patients take in their relation with health professionals, and, on the other hand, the specialization of medical sciences and the explosive development of information and communication technologies.
In regard to the actual organizations, health networks and pluriprofessional practices constitute innovative organizations, whose main challenge is to build and develop transversal links between several partners. To accomplish these objectives in such a complex context,

new competences and skills are needed. In this paper we make use of various perspectives to analyze French and international experiences and to determine requirements for these new profiles, which are characterized by one or more expertises in transversal areas : information and knowledge management, social analysis and management as well as cooperative practices and quality management.

Mots clés : santé, métiers, information et communication, compétences, transversalité, management des connaissances, information partagée, qualité.

Key words : health, skills, information and communication, skills, transversality, knowledge management, information sharing, quality.

Nouveaux métiers dans le domaine de la santé : maîtrise de l'information, transversalité des compétences et autres exigences

INTRODUCTION

Avec l'amélioration du niveau de vie, l'évolution des techniques médicales et chirurgicales, le vieillissement des populations, les pays occidentaux sont confrontés à une dérive des dépenses de santé. En 2000, selon l'OCDE, elles représentaient 7,3 % du P.I.B. (Produit Intérieur Brut) au Royaume-Uni, 7,8 % au Japon, 8,1 % aux Pays-Bas, 9,5 % en France, 10,6 % en Allemagne et 13 % aux États-Unis.

En France, la réduction du temps de travail, la diminution annoncée du nombre de médecins et d'infirmières et la complexité croissante des nouvelles pathologies (notamment liées au vieillissement) imposent la recherche de nouvelles solutions efficientes reposant sur la coordination des activités.

Des pays anglo-saxons : Grande-Bretagne, Canada, Australie, … ont fait le choix de privilégier la médecine de ville en s'appuyant sur les nouvelles technologies de l'information et de la communication. En France, la loi du 4 mars 2002 a consacré les réseaux de santé comme une voie privilégiée pour améliorer le système de santé, en coordonnant médecine de ville et structures hospitalières.

La maîtrise de l'information, sa qualité et sa disponibilité sont au cœur de l'évolution des systèmes de santé et de l'amélioration de leurs performances. Par ailleurs, les patients affirment de plus en plus leur rôle et s'imposent comme des acteurs désormais incontournables du système de santé. Ces enjeux contribuent à l'émergence de nouveaux métiers de proximité, nécessairement transversaux.

Dans cette communication, en analysant les expériences menées en Grande-Bretagne, en Espagne (Catalogne), aux États-Unis, au Canada ou en Australie, nous mettons tout d'abord en évidence l'importance de la maîtrise de l'information pour l'évolution des systèmes de santé, avec en particulier le rôle majeur des technologies de l'information et de la communication et les enjeux du dossier médical patient. Nous présentons ensuite la voie originale des réseaux de santé en France. Nous montrons enfin, à partir de l'exemple des réseaux de santé et en élargissant la problématique à d'autres structures collégiales et transversales comparables en France et à l'étranger, comment la maîtrise et le partage de l'information conditionnent de nouvelles répartitions des compétences et imposent l'émergence de nouveaux métiers.

1 -LA MAÎTRISE DE L'INFORMATION AU CŒUR DE L'ÉVOLUTION DU SYSTÈME DE SANTÉ

Dès 1978, le rapport Nora-Minc (1) avait analysé les enjeux de l'informatisation de la société française et les profonds changements qu'elle impliquerait. Peter Drucker (2) a ensuite mis en évidence l'importance de l'information pour l'évolution des organisations : *« Les activités qui occupent la place centrale ne sont plus celles qui visent à produire et à distribuer des objets, mais celles qui produisent et distribuent du savoir et de l'information ».*

L'analyse de projets internationaux ambitieux illustre ces évolutions majeures dans le domaine de la santé :

- Des pays anglo-saxons ont clairement placé la maîtrise de l'information et sa qualité au cœur de l'évolution des systèmes de santé. Il y a déjà plus de vingt-cinq ans, aux États-Unis, le *Managed Care* et en particulier les *Health Maintenance Organizations* (H.M.O.), ont mesuré toute l'importance de la maîtrise de l'information pour la gestion des soins et mis en place des systèmes d'information performants. Malheureusement, la concurrence exacerbée entre les différents acheteurs de soins et les coûts du marketing ont très vite nui à la qualité des soins pour déboucher sur leur rationnement.
- En 1997-98, le *National Health Service* (N.H.S.) britannique a défini la maîtrise de l'information à la fois comme l'enjeu majeur et l'outil principal de son évolution. Le rapport *Information for Health : an Information Strategy for the Modern NHS* a concrétisé ses réflexions. Le Premier Ministre britannique Tony Blair a souligné l'importance de ce tournant stratégique mais qui doit impérativement être au service des patients et non des organisations de santé : « *The challenge for the NHS is to harness the information revolution and use it to benefit patients* » (3).
- Le Canada (« virage ambulatoire informatisé ») et l'Australie ont également misé sur les nouvelles technologies de l'information et de la communication pour faire évoluer leurs systèmes nationaux de santé en privilégiant, comme la Grande-Bretagne, la restructuration des soins primaires autour des infirmières et

des médecins généralistes chargés de la coordination avec les médecins spécialistes et les structures hospitalières. Pays immenses, ils ont accordé une attention particulière à l'utilisation de la télémédecine pour des populations dispersées.

- En Espagne, depuis le 1er janvier 2002, toutes les Communautés Autonomes nées en 1978, ont obtenu la compétence complète pour la gestion de la santé de leurs habitants. La Catalogne, où le transfert de compétences avait commencé dès 1983, a progressivement affirmé sa spécificité. Elle aussi a misé sur les technologies de l'information pour améliorer la coordination de ses structures de santé et l'efficience (efficacité au meilleur coût) de son système national. Pour une meilleure maîtrise des coûts et la transparence de sa gestion et de ses choix, le Service Catalan de la Santé (devenu *CatSalut)* veut gérer le secteur public du système de santé catalan comme une entreprise privée, c'est-à-dire finalement comme un H.M.O, mais, différence capitale avec le système américain, la concurrence des offreurs de soins doit être maîtrisée et ne pas déboucher sur des profits privés mais sur l'amélioration de l'ensemble du système au profit de ses usagers (4).

Les enjeux de la maîtrise de l'information dans le domaine de la santé sont illustrés par le schéma 1**.**

Le schéma proposé s'inspire des réflexions du *NHS* et de *CatSalut* pour une meilleure utilisation de l'information pour favoriser l'amélioration des performances au service des patients, en les resituant dans le modèle d'amélioration continue de la Qualité ou « roue de Deming ».

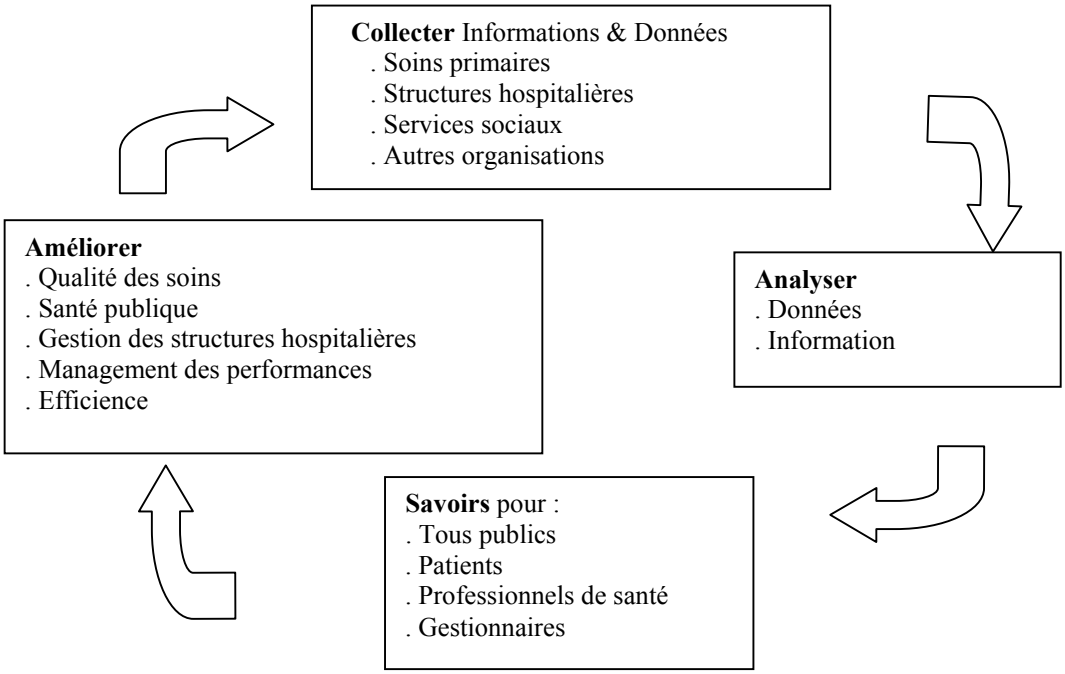


Schéma 1 : Collecter – Analyser – Savoirs – Améliorer = CASA
Un modèle intégré pour améliorer la qualité de la santé par le management de l'information

2 - LA MISE EN PLACE DE NOUVELLES ORGANISATIONS COLLÉGIALES ET TRANSVERSALES

Une meilleure utilisation de l'information doit permettre d'améliorer considérablement la qualité des services aux patients en assurant la continuité et la coordination des soins. Elle est indissociable de la définition de nouvelles structures organisationnelles. La question essentielle est celle de la mise en place de structures de proximité coopératives et transversales.

Les expériences internationales nous offrent des pistes de réflexion intéressantes :

- En Espagne, la Catalogne a fait de ses 360 Centres d'Attention Primaire (CAP) les cellules de proximité de son système « national » de santé en les ancrant dans les structures territoriales : « comarques » ou « pays ». Depuis 1987, la Catalogne est

divisée en 41 « comarques » traduisant l'histoire et les spécificités locales et qui disposent d'une certaine autonomie de gestion mais pas dans le domaine de la santé.

- La Grande-Bretagne a fait le choix de PCTs (*Primary Care Trusts)*, en misant sur la complémentarité de compétences d'infirmières (*nurses*) et de médecins généralistes *(General Practitioners*), formés à la fois à de nouvelles pratiques de soins (prévention, référentiels, systèmes d'aide à la décision et aux prescriptions …) et à des techniques de management : gestion d'équipe, management des savoirs, contrôle des coûts et qualité. Ces nouvelles fonctions s'exercent en articulation avec des centres d'appels accessibles en permanence, jouant à la fois un rôle d'information et d'orientation.
- Les dérives des H.M.O. américains (concurrence par les coûts, rationnement des soins …) ne doivent pas faire oublier la qualité de la coordination des soins souvent mise en place et les performances de leurs systèmes d'information.

En France, les réseaux de santé se sont progressivement développés depuis le début des années 1980. Ils ont correspondu à deux approches très différentes. Tout d'abord, notamment face au développement du Sida, celle de médecins de terrain qui ont cherché à trouver des réponses collégiales de proximité en améliorant la continuité des soins entre la médecine de ville et l'hôpital. La seconde approche est davantage inspirée du *Managed Care* et des H.M.O. américains. Assez souvent à l'initiative de CPAM (Caisses Primaires d'Assurance Maladie), d'URCAM (Unions Régionales des Caisses d'Assurance Maladie, nées en 1996) ou d'hôpitaux, mais aussi de groupes d'assurances ou de laboratoires pharmaceutiques, elle insiste davantage sur la coordination pour éviter les actes inutiles et maîtriser les coûts. De nombreux réseaux de santé (vision plus large) ou de soins (vision plus restrictive), se sont constitués. Beaucoup de ces réseaux et notamment ceux du premier type sont peu structurés et assez souvent informels : les estimations du nombre de réseaux de santé en France varient de 500 à plus de 2000. Tous ces réseaux sont rarement « globaux » mais plutôt dédiés à des groupes de patients ou à des pathologies spécifiques : Sida, précarité, toxicomanie, santé mentale, soins palliatifs, diabète, asthme, oncologie, cardiologie …

Il était essentiel de faire converger ces visions et d'institutionnaliser les réseaux de santé avec comme défi, pour reprendre la formule de Bernard Kouchner, de « formaliser sans stériliser » (5). La loi du 4 mars 2002 sur *Les droits des malades et la qualité du système de santé* (6) s'y attache, en voulant faire des réseaux de santé une troisième voie de gestion de la santé, articulant la médecine de ville (en France, très majoritairement libérale) et les structures hospitalières.

> *« Les réseaux de santé ont pour objet de favoriser l'accès aux soins, la coordination, la continuité ou l'interdisciplinarité des prises en charge sanitaires, notamment de celles qui sont spécifiques à certaines populations, pathologies ou activités sanitaires. Ils assurent une prise en charge adaptée aux besoins de la personne tant sur le plan de l'éducation à la santé, de la prévention, du diagnostic que des soins. Ils peuvent participer à des actions de santé publique. Ils procèdent à des actions d'évaluation afin de garantir la qualité de leurs services et prestations »* (Chapitre V, article 84).

Tout le monde n'est pas d'accord avec la préférence accordée aux réseaux de santé, jugés trop dédiés à des groupes de population et à des pathologies spécifiques. Certains sont beaucoup plus favorables à des solutions de proximité plus modestes : maisons médicales, maisons de santé ou cabinets de groupe pluriprofessions et pluripathologies (médecins généralistes, spécialistes, sages-femmes, dentistes, infirmières, kinésithérapeutes, travailleurs sociaux … également plus adaptées, selon eux, à la médecine libérale et au paiement à l'acte (les bases historiques du système français). Le nouveau ministre de la Santé, Jean-François Mattei, est favorable aux « maisons médicales », mais plus particulièrement pour résoudre le problème de la sursaturation des services d'urgences des hôpitaux et cliniques.

Ces visions ne sont pas inconciliables, elles peuvent même être complémentaires. Les cabinets de groupe n'ont souvent pour objet que la rentabilisation de ressources partagées (locaux, équipements, secrétariats, …), sans forcément un objectif de coordination entre leurs membres. En revanche, les réseaux de santé veulent constituer de nouvelles organisations innovantes, apprenantes et coopératives, reposant à la fois sur la complémentarité des compétences, le partage des savoirs, la collégialité (et non la hiérarchie) et la transversalité. Il y a en outre une différence d'échelle. Les réseaux de santé peuvent coordonner plusieurs cabinets de groupe pluriprofessions et des structures hospitalières sur un bassin de vie.

La question de l'articulation de ces nouvelles organisations de santé avec les collectivités territoriales dans une perspective d'autonomie, que ce soit sous forme de déconcentration ou de décentralisation, est essentielle pour

rapprocher le système de santé des citoyens. Elle est au cœur de l'évolution du *NHS*, de *CatSalut*, comme du système de santé français où l'on parle de la création d'ARS (Agences Régionales de Santé).

La mise en place de nouvelles organisations collégiales de proximité dans le domaine de la santé est aussi largement conditionnée par l'affirmation du rôle nouveau des patients au sein des systèmes de santé.

3 - INFORMATION ET AFFIRMATION DU RÔLE DU PATIENT

Le rôle des patients au sein des systèmes de santé a considérablement évolué. Il a été favorisé par le développement des nouvelles technologies de l'information et de la communication et en particulier d'Internet, qui bouleverse considérablement les conditions d'exercice de la médecine. Le médecin n'est plus désormais le seul à posséder et à monopoliser un savoir qu'il imposait à un patient soumis et respectueux. Le savoir, sous différentes formes et à différents niveaux, est désormais partagé entre le médecin et un patient de mieux en mieux informé (sites Internet), devenu acteur de sa santé et revendicatif.

Le sens de la médecine est en train de changer. Les pratiques des médecins commencent à être standardisées et encadrées par des référentiels et des guides de bonnes pratiques. Les craintes de contrôle liées à ces changements expliquent en partie le malaise de médecins français qui ont l'impression d'être instrumentalisés et transformés en techniciens supérieurs de la santé. On glisse aussi d'une obligation de moyens à une obligation de résultats.

Le *NHS* a, en 1998, fait référence au rôle majeur des patients au sein de la nouvelle organisation de santé qu'il voulait développer pour la Grande-Bretagne, en insistant sur les droits des patients mais aussi sur leur responsabilisation, notamment dans la prévention, l'information et l'accès à certains services comme les urgences. La France a préféré un texte privilégiant les droits des patients (loi du 4 mars 2002).

Les patients sont ainsi partout reconnus comme des acteurs majeurs du système de santé. Les nombreux documents sur la réorganisation du système de santé français, et notamment sur les réseaux de santé, parlent tous de mettre le patient et non les pathologies et les organisations au cœur du système. Il s'agit de louables déclarations d'intention. Le chemin est encore long à parcourir pour parvenir à une vraie démocratie sanitaire (surtout à l'hôpital !), mais il est désormais tracé. Les patients se regroupent et se structurent en associations qui, aux États-Unis, sont même parfois devenues gestionnaires de H.M.O. Leur rôle doit en tout cas être déterminant pour le contrôle des évaluations de la qualité des nouvelles organisations mises en place.

L'affirmation du rôle du patient dans le système de santé ne constitue qu'un cas particulier de l'affirmation de celui du consommateur – usager – client. Jeremy Rifkin (7) a parlé d'un « âge de l'accès » où l'usage de services et leur accès privilégié remplacent progressivement la propriété des biens. Cette révolution du système capitaliste façonne les mentalités et est à la base de ce que certains appellent la « nouvelle économie ».

Dans les pays développés, les services (secteur tertiaire) occupent plus des deux tiers de la population active. On est passé d'une logique de production à une logique de services. Ces services reposent en priorité sur une meilleure utilisation de l'information et sur le rôle central donné au client qu'il convient de fidéliser. La production standardisée de biens anonymes (système « taylorien-fordien ») est moins fondamentale. On insiste davantage sur la « customization » qui consiste à adapter de manière personnalisée différents composants produits en série.

Les organisations, désormais centrées sur des services destinés à des clients, investissent dans une double démarche, de réingénierie et de management de la qualité. Impulsé par le haut, le *Reengineering* anglo-saxon impose une reconfiguration radicale de l'organisation à partir de ses processus de services et non plus de ses fonctions traditionnelles. En revanche, le Management de la qualité, en partie d'origine anglo-saxonne, mais avec de forts apports japonais, mise en priorité sur l'amélioration continue des processus et des services (*kaizen*) grâce à des initiatives suggérées à tout niveau de l'organisation.

Mais *Reengineering* et Management par la Qualité s'accordent sur la priorité absolue à donner au client (interne ou externe), à la transversalité des compétences et à la gestion par processus : ces défis reposent principalement sur la maîtrise de l'information et donc sur la mise en place de systèmes d'information performants.

Yves Cannac a défini l'évolution profonde en cours : *« C'est la recomposition de l'entreprise par flux et processus et non par fonctions. On vivait sur une organisation hiérarchique selon une logique de propriétaire. La transversalisation, c'est la logique du service. Le vrai patron, c'est le client ou le responsable de projet et non de fonction » (8).*

Ces nouveaux paradigmes ont correspondu à une évolution de la vision des emplois et des fonctions au sein des organisations. On est passé progressivement de la notion étroite de qualification (née du modèle hiérarchique de division des tâches cher à F. Taylor) à la notion de compétences et de partage des savoirs entre des individus plus autonomes travaillant en équipes. Selon Guy Le Boterf : *« La compétence n'est pas un état ou une connaissance possédée. Elle ne se réduit ni à un savoir ni à un savoir-faire … Il n'y a de compétence que de compétence en acte … La compétence ne réside pas dans les ressources (connaissances, capacités …) à mobiliser mais dans la mobilisation même de ces ressources … La compétence fait ses preuves dans l'action »* (9).

Ces nouvelles organisations transversales et innovantes sont des organisations apprenantes. Pour Jean-Claude Tarondeau : *« L'organisation transversale est une organisation apprenante car c'est une organisation ouverte, composée d'individus ou de groupes autonomes combinant des savoirs dans l'action »* (10).

Dans le domaine de la santé, la formation, dont les perspectives ont été considérablement modifiées par les nouvelles technologies de l'information et de la communication, concerne désormais non seulement les praticiens pour acquérir de nouvelles compétences ou transmettre leurs savoirs mais aussi les patients (prévention, responsabilisation).

Le contexte évolue profondément. On passe progressivement d'une médecine où le médecin était quasiment au service individuel de ses patients à une médecine visant globalement l'amélioration de l'état de santé de l'ensemble d'une population dans un contexte de ressources limitées. Ainsi, par exemple, les médecins vont devoir arbitrer entre strict intérêt individuel de leurs patients et priorités globales définies par les institutions de santé ou les collectivités territoriales (11). Avec ces nouvelles contraintes, les médecins devront prendre les décisions de soins conjointement avec leurs patients et les leur expliquer. Les arbitrages, voire les conflits, entre intérêts individuels et politiques de santé, seront difficiles à gérer dans une société de plus en plus individualiste.

Toutes ces évolutions : affirmation des nouvelles technologies de l'information et de la communication, nécessité d'améliorer le système de santé en privilégiant des structures de proximité, réflexions sur l'évolution des organisations, gestion des compétences, rôle nouveau du patient-client, … ont favorisé l'émergence de nouveaux métiers transversaux et de nouvelles organisations dans le secteur de la santé. Elles sont de deux types : structures d' « intermédiation » entre le patient, la médecine de ville et l'hôpital (réseaux de santé, maisons médicales, associations d'hospitalisation à domicile …) ou entreprises spécialisées dans l'application des NTIC au secteur de la santé.

4 - NOUVEAUX MÉTIERS TRANSVERSAUX ET DE PROXIMITÉ

Ces nouveaux métiers transversaux exigent des compétences pluridisciplinaires concernant des domaines multiples que l'on peut schématiser de la manière suivante :

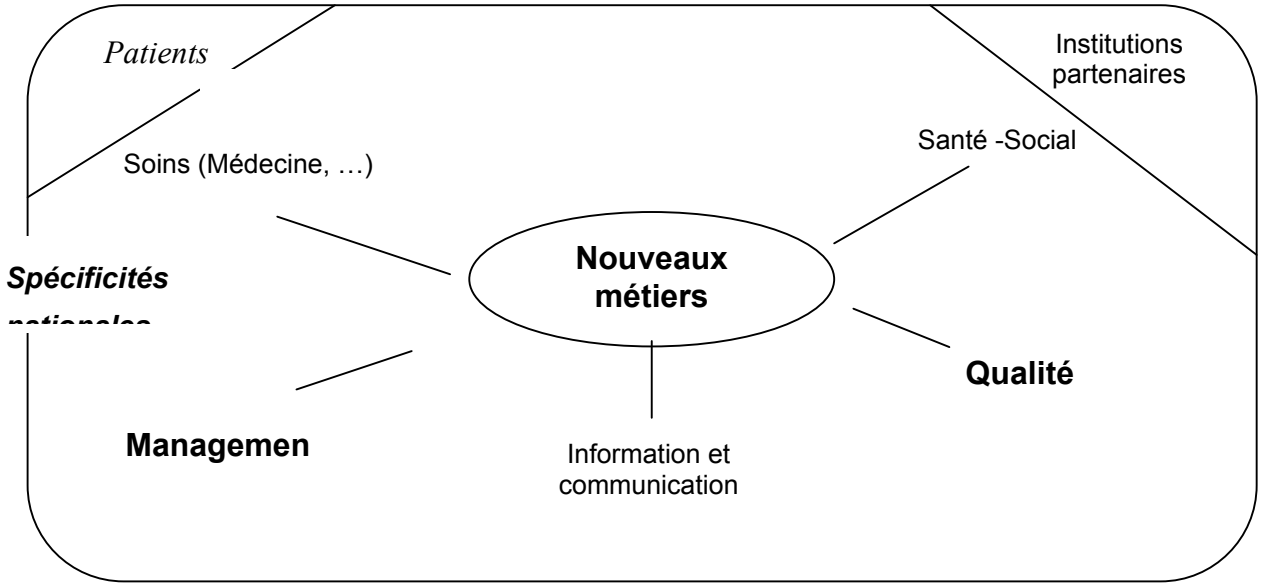


Schéma 2 : De nouveaux métiers transversaux exigeant des compétences pluridisciplinaires

Nous allons tout d'abord évoquer ces nouveaux métiers liés à la maîtrise de l'information et à sa meilleure utilisation dans le domaine de la santé en prenant l'exemple français des réseaux de santé.

Rappelons que parmi les principaux défis qui se présentent aux projets des réseaux de santé, émerge plus particulièrement l'appréhension précise du problème de santé publique et les spécificités des populations concernées : ainsi, toute nouvelle structure doit répondre à un besoin clairement identifié. Viennent ensuite la formalisation des objectifs communs (charte du réseau quand elle existe), la définition juridique de la structure de partenariat (association, coopérative, groupement d'intérêt économique (GIE) …) puis des outils facilitant les pratiques communes : définition de processus, formalisation de procédures, définition de trajectoires des patients, référentiels, guides de bonnes pratiques, système d'information, modalités d'évaluation …

Ces étapes conditionnent l'existence de la nouvelle organisation et l'accompagnent dans son développement. Organisations innovantes en constante évolution, les réseaux de santé font émerger de nouveaux métiers (coordonnateurs, évaluateurs, « hébergeurs » de données …) et enrichissent des métiers existants (concepteur, réalisateur, chef de projet, …). Ces nouveaux métiers se structurent à un moment charnière de l'histoire des réseaux de santé : leur passage de la phase pionnière à celle de la formalisation. Tout cela s'est développé en France, dans le contexte de crise du système hospitalier : services d'urgences sursaturés et application de la réduction du temps de travail.

➢ ***Coordonnateur***

Nouvelles organisations coopératives voulant améliorer la continuité des soins en luttant contre les cloisonnements du système de santé français, les réseaux de santé ont très vite vu émerger le métier spécifique de coordonnateur. Premier à apparaître, il est le plus spécifique au réseau et aussi probablement le plus fondamental. Le coordonnateur est le cœur, l'âme du réseau. Il en est le fédérateur. Il crée du lien entre toutes les composantes du réseau et tous ses participants (praticiens, institutionnels, patients …), gère les conflits. Il est aussi un « passeur » entre les cultures des différentes professions et institutions partenaires. Il doit aussi être capable de gérer les temps et les rythmes différents des professions et institutions partenaires.

Très pluridisciplinaire, ce nouveau métier de coordonnateur peut se décomposer en deux, voire en trois profils d'emplois différents pour les réseaux les plus importants, car il suppose à la fois des compétences techniques, mais aussi et surtout des compétences relationnelles, axées sur la communication et l'affirmation de valeurs partagées, l'animation, la gestion de conflits, sans oublier les compétences de management : gestion financière, gestion des savoirs ou du changement. Un des principaux problèmes rencontrés par les réseaux de santé est celui de la définition précise des responsabilités. Le métier de coordonnateur suppose une répartition des rôles claire et non conflictuelle avec le président souvent aussi fondateur du réseau, homme-orchestre providentiel qui peut avoir tendance à s'occuper de tout.

Ce métier essentiel de coordonnateur peut, selon le moment où il est recruté par le réseau, se décliner en deux volets très distincts, qui peuvent d'ailleurs correspondre à deux métiers différents. Lors de la conception du projet de réseau de santé puis de sa mise en place, les activités du coordonnateur relèvent nettement de l'assistance à maîtrise d'ouvrage. S'il a été impliqué dans le projet dès son origine, le rôle du coordonnateur devient ensuite un rôle d'accompagnement de projet.

Avec le développement des expériences de « maisons de réseaux de santé » pour regrouper en un même lieu plusieurs réseaux de santé, vont aussi s'affirmer des activités de coordination spécifiques, en quelque sorte de coordination de coordonnateurs.

➢ ***Évaluateur***

Il n'y a pas d'expérimentation sans évaluation. Le second métier favorisé par le développement des réseaux de santé est celui d'évaluateur. Il peut être à la fois interne et externe au réseau, le coordonnateur pouvant animer l'aspect interne de l'évaluation. Conçue dans une perspective de gestion de projet et comme une phase de la démarche qualité (amélioration), l'évaluation aide le réseau à s'adapter et à évoluer. Elle suppose à la fois l'évaluation de la structure réseau dans son ensemble, de ses différentes composantes, de ses processus et outils : référentiels, trajectoires de patients, système d'information … et, à plus long terme, des comparaisons avec d'autres structures similaires et de leurs effets sur les populations concernées.

➢ ***Hébergeur***

Le dernier métier à être apparu est celui d' « hébergeur » de données ou « infomédiaire », consacré par la loi du 4 mars 2002. Il suppose à la fois des compétences techniques et des connaissances éthiques et juridiques. L' « hébergeur », le plus souvent externe au réseau, assume la responsabilité de la sécurité, de la confidentialité, mais aussi de la propriété des données traitées et de leur transmission. Il gère les droits d'accès, les interfaces, les priorités, le partage et la sécurisation des données. Comme le met en évidence le rapport Gros (12), ce métier est appelé à avoir de plus en plus d'importance en raison du glissement (qu'elle regrette) des données médicales des mains des médecins vers ces intermédiaires. Ce nouveau métier devrait être soumis à des procédures d'agrément et être très strictement encadré, notamment par la CNIL (Commission Nationale de l'Informatique et des Libertés) pour éviter toute dérive commerciale dans l'utilisation des données ou toute atteinte à la vie privée.

➢ ***Médiateur Internet***

La plupart des systèmes d'information et les dossiers médicaux informatisés des patients reposent désormais principalement sur l'utilisation des technologies internet. Le site Internet du réseau va devenir à la fois la vitrine du réseau vers l'extérieur (portail de présentation) mais va aussi constituer un outil de communication interne au réseau et surtout d'orientation du patient adhérent au réseau en diffusant à son intention des informations locales : médecins de garde, procédures / urgences, … Le site Internet du réseau contribuera également à valider l'information médicale diffusée par des sites santé grand public ou assurera des liens avec d'autres sites spécifiques dédiés à certaines pathologies (labellisation de « favoris »). Le métier de « webmaster » appliqué aux réseaux de santé va ainsi probablement favoriser l'apparition d'un métier spécifique au réseau que l'on pourrait qualifier de « médiateur Internet ».

➢ ***Télésurveillance***

L'affirmation de la télémédecine, dont les activités sont indissociables de celles des réseaux de santé, va considérablement modifier les conditions d'exercice de la médecine, les relations entre les patients et les médecins et l'articulation entre médecine de ville et l'hôpital. Pour les réseaux de santé, cela conduit à l'affirmation des métiers relevant de la télésurveillance à distance, en incluant de nombreuses activités et en particulier le suivi de l'hospitalisation à domicile, de l'observance des prescriptions …

On peut également envisager le développement au sein des réseaux les plus structurés de métiers plus classiques de « *commerciaux* » ou de « *gestionnaires* ». Un métier strictement commercial : démarchage de nouveaux partenaires et surtout de nouveaux « clients » du réseau se justifierait avec la mise en concurrence systématique des organisations de santé. Aux États-Unis, elle impose des frais de marketing important aux H.M.O. : parfois de l'ordre de 20 à 30 % de leurs budgets globaux. Nous n'en sommes pas encore là en France. A ce jour, ces activités peuvent encore facilement s'intégrer aux nouveaux métiers que nous venons de définir : commercial et gestionnaire / coordonnateur, gestionnaire / évaluation interne.

Le développement d'organisations innovantes dans le domaine de la santé est fortement liée à la mise en place de systèmes d'information de qualité : ceci nécessite l'intégration de ***métiers liés à l'ingénierie des systèmes d'information***, comme des concepteurs, réalisateurs ou chefs de projet. Or, d'une part, de tels métiers, bien reconnus dans d'autres domaines, comme l'industrie ou la gestion, sont très innovants dans le domaine de la santé. D'autre part, les compétences requises par ces métiers ne peuvent pas être "importées" sans des adaptations profondes aux spécificités du domaine, en élaborant, une fois de plus, des démarches pluridisciplinaires et transversales.

Les « nouveaux » métiers, ainsi que les métiers « adaptés » que nous venons d'analyser dans le cadre des réseaux de santé peuvent également se décliner dans le cadre de cabinets de groupe pluriprofessions. Ils posent la question de la taille critique de ces organisations pour leur permettre de s'exercer correctement en s'appuyant sur des outils et des applications logicielles utilisables à différentes échelles.

Ces nouveaux métiers exigent souvent des compétences différentes de celles de la période d'émergence un peu anarchique des projets. Après l'ère des pionniers vient celle des administrateurs. Le défi est de savoir conserver un rôle à ces pionniers, de structurer tout en conservant esprit d'innovation, dynamisme et enthousiasme. Le problème est crucial pour le métier de coordonnateur, très souvent assumé par le promoteur du réseau.

5 - NOUVEAUX MÉTIERS EN SANTÉ ET SPÉCIFICITÉS NATIONALES

Une autre question essentielle est celle de l'articulation de ces nouveaux métiers pluridisciplinaires, collégiaux et transversaux par rapport au fonctionnement très individualiste de la médecine libérale en France (qui a trop tendance à confondre libéralisme et individualisme) et par rapport aux logiques professionnelles pour ne pas parler de corporatismes. Cette complexité peut être abordée en analysant le positionnement de ces nouveaux métiers pluridisciplinaires par rapport aux métiers traditionnels, par exemple médecins et infirmier(e)s mais aussi la coopération entre ces différents métiers traditionnels qui doit transcender les vieux clivages. En étudiant les expériences menées à l'étranger, nous remarquons que :

- La Grande-Bretagne a opté pour des cabinets de groupe relevant des nouveaux *Primary Care Trusts* articulés autour de la coopération entre infirmières et médecins généralistes en relation avec des centres d'appels. Ils jouent un rôle d'orientation vers les spécialistes des hôpitaux. Dans ces organisations, les responsabilités non médicales des médecins sont déléguées.
- D'autres documents insistent sur l'importance du rôle de managers de ces structures et proposent de le confier aux médecins généralistes, mais en leur enlevant certaines responsabilités de soins plus quotidiens attribuées à des professions intermédiaires comme infirmières, ou par exemple optométristes par rapport aux ophtalmologistes.
- Aux États-Unis, dans les H.M.O., comme en Catalogne dans les Centres d'Attention Primaire, les infirmières ont aussi un rôle privilégié.

Les spécificités de notre système de santé rendent quasi impossible d'envisager à ce jour une répartition similaire des rôles en France où les médecins ont bloqué le développement de professions intermédiaires (sauf les sages-femmes) et limité le rôle des infirmières. Les spécificités nationales, nées de cultures particulières, doivent donc être prises en compte pour la définition des nouveaux métiers.

Pour l'ensemble des métiers analysés, nous identifions quelques mots clés : professionnalisation, décloisonnement, transversalité, qualité, coopération, innovation. Ils reposent largement sur l'utilisation des nouvelles technologies de l'information et de la communication. Ils sont très liés à la mise en place de systèmes d'information et, en premier lieu, du dossier médical informatisé du patient ou plutôt des dossiers partagés du patient.

Dans ce domaine, malgré des cultures nationales très différentes, les problèmes identifiés en France et à l'étranger sont assez similaires en ce qui concerne les questions de l'identifiant national et le domaine de l'interopérabilité des données :

- la définition du dossier médical électronique du patient a été retenue comme une priorité par le plan de développement stratégique du *NHS.* La Grande-Bretagne est traditionnellement très vigilante sur la protection des libertés publiques. Il n'y a pas de carte d'identité ni d'identifiant national type INSEE-Sécurité Sociale
- en France, la CNIL est défavorable à l'utilisation de tout identifiant unique. Il sera dès lors assez difficile de faire en 2005 de la nouvelle version de la carte Vitale (qui pour le moment sert uniquement à la transmission électronique de feuilles de soins) une carte réellement médicale.

La spécificité des cultures nationales est plus importante dans d'autres domaines et conditionnent l'évolution des métiers. Les choix opérés en France dans les années 1945-1950 pèsent encore lourdement. Le système de santé français dit libéral repose sur la liberté de choix du praticien par le patient et sur le paiement à l'acte des prestations réalisées. Un tel système a imposé la mise en place de structures (CPAM en général au niveau des départements et bureaux locaux qui en relèvent) dont le rôle est prioritairement de traiter (« liquider » !) les feuilles de soins et d'assurer à la fois le paiement des praticiens et le remboursement des patients ou de leurs mutuelles. La carte Vitale n'a que partiellement amélioré la situation. En imposant le passage obligé par un généraliste pour l'accès aux spécialistes et à l'hôpital (inscription des patients sur une liste) et le paiement des médecins à la capitation, le système britannique a moins de contraintes administratives.

L'utilisation de la carte Vitale comme carte de paiement et la suppression totale des feuilles de soins papiers imposerait la définition de nouveaux métiers pour près de 200 000 salariés des différents organismes de Sécurité Sociale. Des réseaux de santé animés par des CPAM (comme celui de soins palliatifs des Hautes-Pyrénées) constituent des pistes intéressantes de définition de nouvelles activités pour l'assurance maladie. Dans le cadre de structures de proximité relevant d'Agences Régionales de Santé, on pourrait leur confier des responsabilités de gestion du risque santé (et plus seulement de payeur aveugle !) mais aussi des missions non seulement d'information des patients sur leurs droits mais aussi d'orientation des patients vers les structures de santé les plus adaptées (ce qui n'est pas pour le moment autorisé en France où seule l'information est permise). Cette mission est réalisée par le *NHS* avec le centre d'appel *NHS Direct*.

Le *NHS* envisage par ailleurs de favoriser les échanges de pratiques entre médecins à travers *NHSnet*. En s'inspirant de *NHS Direct*, la Catalogne met progressivement en place depuis octobre 2001, un centre d'appel *Sanitat Respon*. Dans ce domaine, la France est assez en retard. Les différents centres d'appels mis en place par les Caisses Primaires d'Assurance Maladie ou ceux des compagnies d'assurances (AXA…) doivent se contenter d'informer les usagers sur leurs différents droits.

Système national conscient de ses responsabilités à l'égard des Britanniques, le *NHS* est très sensible à la qualité de l'information, très inégale et trop souvent orientée, diffusée sur les très nombreux sites médicaux sur Internet. *NHS Scotland* a développé le programme SHOW : Scottish Health On the Web. Le Canada s'est aussi engagé dans cette voie. A travers le réseau Canadien de la Santé, *Santé Canada* veut fournir au public une information fiable en matière de santé (13). Les exigences de qualité concernant la diffusion, via l'Internet, de l'information médicale sont génératrices de besoins en compétences et métiers nouveaux, comme le témoignent les différents efforts de certification accomplis au niveau international.

6 - ÉVOLUTION DU RÔLE DE L'ÉTAT

En élargissant notre réflexion au delà des réseaux de santé et des cabinets de groupes pluriprofessions, nous remarquons que l'évolution même du rôle de l'Etat dans le domaine de la santé donne une forte impulsion à l'apparition d'autres métiers. Moins gestionnaire, l'Etat national devient garant de l'intérêt général et coordonnateur des activités des services publics et des prestations assurées par le secteur privé (rapports Bangemann et Gros). Cette évolution se traduit en France par l'apparition d'agences : ANAES (Agence Nationale d'Accréditation et d'Évaluation en Santé), Agence du médicament devenue AFSSAPS (Agence Française de Sécurité Sanitaire des Produits de Santé), ARH …

Ces agences ont un rôle majeur de régulation et de contrôle, voire de labellisation et d'accréditation. Le métier d'hébergeur, c'est-à-dire de gestionnaire de données personnelles particulièrement sensibles, correspond à une mission de service public qui suppose procédures d'accréditation et de contrôle de la qualité et de la déontologie (respect de chartes, de procédures …) par l'Etat. Les nouvelles missions de contrôle des activités des hébergeurs de données personnelles, de la qualité de l'information diffusée par les sites Internet en santé, des activités de télémédecine … et, bien sûr, à travers l'ANAES, d'accréditation des établissements hospitaliers, voire, à plus long terme, des médecins libéraux, peuvent conduire à l'apparition de nouveaux métiers.

Ces nouveaux métiers se dessinent dans le contexte de l'évolution globale de la notion de service public. Longtemps basé sur l'application de la seule égalité juridique des citoyens, le service public doit désormais tendre de plus en plus à rechercher l'équité dans le service rendu.

CONCLUSION

Les nouvelles technologies de l'information et de la communication favorisent ainsi l'apparition de nouveaux métiers dans le secteur de la protection sociale. Les NTIC commencent à gommer les spécificités nationales

comme les frontières entre les métiers existants pour faire émerger des métiers basés sur des compétences pluridisciplinaires : gestion de l'information, communication, management, coordination, évaluation …

Ces nouveaux métiers au service d'un patient devenu « client » et qui affirme son rôle reposent aussi sur l'éthique et la déontologie. Il n'y a pas de diffusion et de partage de l'information sans confiance et valeurs partagées, à la fois par les praticiens des structures de groupe et par les patients. Les nouveaux métiers, garants du respect de la vie privée et de la confidentialité des données personnelles, se définissent dans le cadre de structures locales de proximité dont il faut définir les liens avec l'État. L'État interventionniste, qualifié de « fordien-keynésien », qui s'est largement développé en particulier en France après 1945, est en crise. La déréglementation et la libre concurrence, dogmes de l'Union européenne, s'imposent dans de nombreux secteurs : télécommunications, transport, énergie … mais pas encore dans les domaines de la santé et de l'éducation nationale, où il y a tout de même des cliniques et des écoles privées. Mais la concurrence ne doit pas s'exercer sans règles. L'État a un rôle de garant à jouer. Même aux États-Unis, les associations de consommateurs et de médecins réclament son intervention, notamment dans le domaine de la santé face aux dérives des H.M.O.

Un nouvel État davantage garant, régulateur et incitateur se dessine. Ses domaines de compétences et leur articulation notamment avec les collectivités territoriales (en France : régions, départements, communes et bientôt nouveaux « pays »), mais aussi avec toutes les initiatives émanant du secteur privé, doivent être précisés.

C'est dans ce contexte que de nouvelles organisations transversales de proximité (réseaux de santé, cabinets de groupes pluriprofessions …) s'affirment. Leur développement est indissociable de celui de nouveaux métiers alliant compétences d'animation, de coordination et de management et reposant largement sur l'utilisation d'informations fiables et facilement accessibles. Dans une société de plus en plus individualiste, il s'agit de l'émergence de nouveaux métiers d'intermédiation au sein d'organisations intermédiaires. L'avenir des systèmes de santé dépend largement de l'affirmation de ces nouvelles organisations de proximité et de ces nouveaux métiers.

REMERCIEMENTS

Nous tenons à remercier le Dr Penny Bevan et Valérie Aston (*NHS* régional Londres) ainsi que Jaume Tort i Bardolet (*CatSalut*) pour le temps qu'ils nous ont consacré et les documents qu'ils ont bien voulu nous communiquer.

NOTES

(1) NORA (Simon) MINC (Alain), *L'informatisation de la société*, La Documentation française, 1978, Points, Seuil, 1980, 162 p.

(2) DRUCKER (Peter), *Au-delà du capitalisme. La métamorphose de cette fin de siècle*, Paris, Dunod, 1993.

(3) *Information for Health. An Information Strategy for the Modern NHS 1998 – 2005,* NHS Executive, 1998, p. 5

(4) *Noves perspectives del model sanitari català. Les línes estratègiques del CatSalut,* Generalitat de Catalunya, Sitges, 15-16 avril 2002

(5) KOUCHNER (Bernard), secrétaire d'Etat à la Santé, 3e Congrès de la Coordination Nationale des Réseaux de Santé, Paris, 23 et 24 juin 2001.

(6) Loi du 4 mars 2002 sur *Les droits des patients et la qualité du système de santé*, Journal Officiel du 5 mars.

(7) RIFKIN (Jeremy), *L'âge de l'accès. La révolution de la nouvelle économie*, La Découverte, 2000, 396 p.

(8) CANNAC (Yves) dans TARONDEAU (J.-C.), JOLIBERT (A.) et CHOFFRAY (J.-M.), « Le management à l'aube du XXIe siècle », *Revue française de gestion*, n° 100, 1994, p. 19.

(9) LE BOTERF (Guy), *De la compétence. Essai sur un attracteur étrange*, Paris, éd. d'Organisation, 1994, pp. 16-18

(10) TARONDEAU (Jean-Claude), *Le management des savoirs*, PUF, 1998, p. 110.

(11) MOORE (Gordon T.), *Managing to do better : general practice in the 21st century*, Office of Health Economics, London, 2000, pp. 16-20.

(12) GROS (Jeannette), *Santé et nouvelles technologies de l'information*, rapport adopté par le Conseil Économique et Social, 10 avril 2002, pp. I-27 et II-46.

(13) site http://www.canadian-health-network.ca

BIBLIOGRAPHIE

. BANGEMANN (Martin), *Europe and the Global Information Society*, Recommendations to the European Council, Bruxelles, 1994, 42 p.

. BEUSCART (Régis), *Les enjeux de la Société de l'Information dans le domaine de la Santé*, rapport au Premier ministre, mai 2000, 37 p.

. BOURDON (Antoine), LEBEL (Christine), MAGNIEN (Laurent), *Le système de protection sociale*, avril 2002, Ellipses, 64 p.

. BOURRET (Christian), LAURENT (Daniel), SCARBONCHI (Elisabeth), « Une réponse en termes de systèmes d'information aux défis de la protection sociale : les réseaux de santé », *V.S.S.T.' 2001 : Veille Stratégique, Scientifique et Technologique*, Barcelone, octobre 2001, Actes, tome II, pp. 23 – 32.

. *Canada Health Infoway : Paths to Better Health*, Final Report, Advisory Council on Health Infostructure, Health Canada Publications, 1999.

. CARRÉ (Dominique), LACROIX (Jean-Guy) sous la direction de, *La santé et les autoroutes de l'information. La greffe informatique*, L'Harmattan, juillet 2001, 312 p.

. *Noves perspectives del model sanitari català. Les línes estratègiques del CatSalut,* Generalitat de Catalunya, Sitges, 15-16 avril 2002

. DRUCKER (Peter), *Au-delà du capitalisme. La métamorphose de cette fin de siècle*, Paris, Dunod, 1993.

. HAMMER (M.), CHAMPY (J.), *Le Reengineering. Réinventer l'entreprise pour une amélioration spectaculaire de ses performances*, Dunod, 1993, 247 p.

. FAUROUX (Roger), SPITZ (Bernard) et al., *Notre Etat. Le livre vérité de la fonction publique*, Robert Laffont, janvier 2001, 805 p.

. GROS (Jeannette), *Santé et nouvelles technologies de l'information*, rapport adopté par le Conseil Economique et Social, 10 avril 2002.

. *Health Online : A Health Information Action Plan for Australia*, Commonwealth of Australia, Canberra, 1999.

. Actes du congrès *HIMSS 2002 : Healthcare Information and Management Systems Society*

. *Information for Health. An Information Strategy for the Modern NHS 1998 – 2005*, NHS Executive, 1998, 123 p.

. KERVASDOUÉ (Jean de), « Panser ou repenser le système de santé », in *Notre Etat ...*, op. cit., janvier 2001, pp. 359 – 388.

. KIMBERLY (J.R.), MINVIELLE (E.), *The Quality Imperative. Measurement and management of Quality in Health Care*, London, Imperial College Press, 1995, 214 p.

. LAURENT (Daniel), « Assurance maladie : diagnostic et remèdes », *Sociétal*, n° 30, 4e trim. 2000, pp. 16 – 21.

. LE BEUX (Pierre) et BOULLIER (Dominique) sous la dir. de, « L'information médicale numérique », *Les cahiers du numérique*, Hermes Science, novembre 2001, 207 p.

. LE BOTERF (Guy), *Ingénierie et évaluation des compétences*, Paris, Editions d'Organisation, 3e éd., 2001, 539 p.

. LEBRUN (Marcel), *Théorie et méthodes pédagogiques pour enseigner et apprendre. Quelle place pour les TIC dans l'éducation ?*, De Boeck, 2002.

. MARCINIAK (Rolande), ROWE (Frantz), *Systèmes d'Information, Dynamique et Organisation,* Economica, 1997, 111 p.

. MINTZBERG (Henry), *Le Management. Voyage au centre des organisations*, Editions d'Organisation, juillet 2001, 570 p.

. MOORE (Gordon T.), *Managing to do better : general practice in the 21st century*, Office of Health Economics, London, 2000, 62 p.

. *National Health Information Standards Plan*, Commonwealth of Australia, february 2001.

. *The New NHS : Modern . Dependable*, december 1997.

. NORA (Simon) MINC (Alain), *L'informatisation de la société*, La Documentation française, 1978, Points, Seuil, 1980, 162 p.

. PRAX (Jean-Yves), *Le Guide du Knowledge Management. Concepts et pratiques du management de la connaissance,* Dunod, mai 2000, 266 p.

. REIX (Robert), *Systèmes d'information et management des organisations*, Vuibert, 4e éd., 2002, 444 p.

. RIFKIN (Jeremy), *L'âge de l'accès. La révolution de la nouvelle économie*, La Découverte, 2000, 396 p.

. SHORTELL (S.M.) et al., *Remaking Health Care in America. Building Organized Delivery Systems*, 1996, 369 p.

. TARONDEAU (Jean-Luc), *Le management des savoirs*, P.U.F., décembre 1998, 127 p.

. VELTZ (Pierre), *Le nouveau monde industriel*, Le débat-Gallimard, mai 2000, 230 p.

LES LIMITES DU TOUT-TECHNOLOGIQUE DANS LA CAPITALISATION DE L'INFORMATION "MARCHE" AU SEIN DE GIAT INDUSTRIES

Patrick CANSELL, responsable Info Stratégique
Giat Industries, DCAI/DMAG
13, route de la Minière
78034 Versailles Cedex
01 30 97 35 62
p.cansell@giat-industries.fr

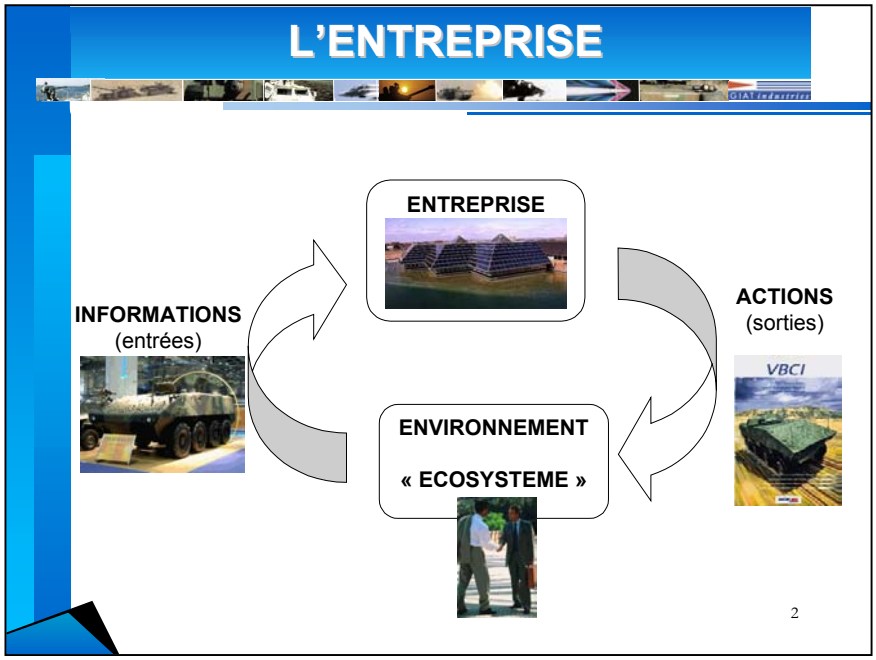
Résumé : Face aux développement d'outils ou de systèmes complets de capitalisation / diffusion d'informations, il importe de recentrer la problématique du knowledge management dans la stratégie de l'entreprise : que voulons nous obtenir ? quel est l'objectif du management des connaissances ?
La démarche adoptée au sein de Giat Industries repose sur les bases méthodologiques de l'intelligence économique et a permis, en exploitant certaines pratiques du knowledge management, de créer une mémoire d'entreprise dédiée au marché (connaissance des acteurs, clients, concurrents, produits, partenaires…) afin d'améliorer notre positionnement commercial et notre compétitivité.

Mots clés (français) : Intelligence économique, knowledge management, veille, compétitivité, marketing, mémoire corporate, système de partage d'information, warroom, culture d'entreprise…

Mots clés (anglais) : Competitive intelligence, knowledge management, marketing, corporate memory, information sharing system, warroom, culture …

Les limites du tout-technologique dans la capitalisation de l'information "marché" au sein de Giat Industries

L'entreprise, en tant que système complexe interagissant avec son environnement, se doit non seulement de développer une culture de l'observation de son milieu (on pourra parler de "culture du renseignement[65]"), mais également une politique de management de l'information et des connaissances, de façon à capitaliser, partager, exploiter au mieux ses ressources "informationnelles" tant internes qu'externes.



Le feed back généré par l'entreprise du fait des actions menées et des messages émis vient compléter les informations d'environnement qu'il est nécessaire de collecter en vue d'une meilleure compréhension de la problématique "marché". La démarche d'intelligence économique poursuit dans ce cadre trois axes d'effort :

- La mémoire collective à travers :
- la création de bases de données de référence
- la constitution de tableaux de bord concurrence, clients…
- la création d'un réseau d'experts

- La dynamisation des flux d'informations, à travers :
- l'instrumentalisation du « feed-back »
- la création de communautés d'intérêts
- une logique de warroom pour le traitement des affaires

- La valorisation et l 'exploitation des ressources, à travers :
- l'optimisation du ciblage des argumentaires
- l'amélioration de la relation clients
- la stratégie commerciale

En matière de capitalisation et de partage des informations, il est essentiel de prendre en compte les freins générés par le poids du relationnel dans la qualité de l'échange, la culture souvent orale des opérationnels comme des experts, la vision « métier » souvent transposée sur l'environnement extérieur et enfin la confiance souvent faible vis à vis des interlocuteurs d'autres métiers . Ces aspects culturels sont non seulement à prendre en

[65] Séminaire sur la Culture Française du Renseignement, 1995-1998, sous la présidence de l'Amiral Pierre Lacoste (CESD).

compte mais à traiter de façon à permettre l'indispensable fluidification des flux d'information en interne de l'entreprise.

Giat Industries développe désormais son management de l'information "marché" selon quatre axes :
- la création d'un système de partage d'informations, GINet (Giat Intelligence Network)
- le développement de la gestion « transverse » d'affaires commerciales (warroom)
- le développement de réseaux (d'experts, d'opérationnels, de stagiaires, d'assistantes... communautés d'intérêt, rencontres informelles sur des thématiques précises…)
- la captation des savoirs "marchés" à travers un processus du transfert de l'oral au numérique (le Club Intelligence Economique)

En définitive, il importe de prendre en compte le fait que culture d'entreprise et facteurs humains déterminent la démarche KM de l'entreprise. Les technologies (bases de données, portails, forum…) doivent s'imposer par la qualité des résultats obtenus et exploitables par chacun, car il ne peut y avoir de véritable adhésion à la démarche si les différents acteurs n'obtiennent pas un **retour sur investissement** immédiatement perceptible. L'intelligence économique permet dans cette perspective, en donnant du sens à la démarche et en soulignant les objectifs à atteindre, de valoriser les apports des individus dans une perspective collective immédiatement perceptible et acceptée : le BUSINESS !

Deux impératifs sont à prendre en compte pour obtenir une réelle participation des acteurs ciblés : valorisation et appropriation de la démarche. Cela signifie **valorisation des acteurs et des informations** fournies à la collectivité, **appropriation de la démarche** par ces acteurs à travers la sensibilisation, la formation, le maintien d'une relation de confiance. Giat Industries a obtenu à titre expérimental de réels succès dans cette démarche sur la période 1999-2001. Depuis près de 18 mois désormais, la démarche a été considérée comme étant opérationnelle au même titre que la plate-forme de partage d'informations développée, GINet, qui se développe désormais pour répondre tant à des impératifs d'alimentation que d'exploitation de la mémoire collective "marché" de l'entreprise.

LES INFRASONS ENTRE SCIENCE ET MYTHE : LA BIBLIOMETRIE PEUT-ELLE CONTRIBUER A CLARIFIER UNE VERITE SCIENTIFIQUE CONTROVERSEE ?

GOUJARD Bertrand
Ecole d'Ingénieurs en Génie des Systèmes Industriels (EIGSI)
26, rue de Vaux-de-Foletier
17041 La Rochelle Cedex 1- France
+ 33 (0)5 46 45 80 26
bertrand.goujard@eigsi.fr

Résumé : L'effet des infrasons (émissions sonores de fréquences inférieures à 20 Hz) sur l'organisme humain constitue une question scientifique à la fois marginale et incomplètement élucidée, y compris à travers les publications récentes ; ce sujet a généré de surcroît sur certaines pages de l'Internet un véritable mythe, alimenté sans doute par l'intérêt de ces ondes acoustiques pour la surveillance des essais nucléaires.
Lorsque cette rumeur débouche sur des questions posées par ses clients à un industriel, le veilleur scientifique et stratégique se doit de trouver des éléments qui permettent, dans un délai court, de prendre position face au problème. Si l'on dispose ensuite de plus de temps, il est possible d'approfondir l'analyse, en complétant les sources d'information consultées et en effectuant des traitements bibliométriques plus poussés.
L'objet de la présente communication est donc, à partir de ce cas concret, d'examiner d'abord comment une analyse bibliométrique à partir d'une base de données adéquate, jointe à celle des divers documents trouvés sur l'Internet, peut permettre d'évaluer rapidement l'acuité du problème posé. On montrera ensuite dans quelle mesure une méthodologie plus élaborée (examen des mots-clés, cartographie plus large des réseaux d'auteurs impliqués) est susceptible de contribuer à mieux cerner une vérité scientifique quelque peu indécise.

Abstract : The effect of infrasound noise (acoustic waves with frequencies below 20 Hz) on human body functions appears both marginal and poorly cleared up as a scientific issue, including in recent papers ; besides, the subject has reached on a number of Internet pages a mythical dimension, owing presumably to the relevance of these acoustic waves for nuclear test detection.
When questions from customers are arisen as a result of the rumour, the scientific and strategic intelligence specialist must quickly find the clues to take up a position on the issue. With more time available, it becomes possible to go deeper into the question, more information sources can be included and more sophisticated bibliometric treatment applied. With this concrete case in view, the present investigation strives to demonstrate how a bibliometic analysis from one relevant database in addition to the miscellaneous documents

found on the Internet allows to evaluate quickly how critical the issue is. It is then discussed how far a more worked out methodology (that includes keywords analysis and a broader cartography of the network of the scientists involved) is liable to clarify a somehow confused scientific truth.

Mots-clés : infrason – homme – effet biologique – nuisance acoustique – gêne (nuisance) – rumeur – information scientifique technique – analyse bibliométrique – cartographie – scientométrie

Keywords : infrasound – human – biological effect – noise pollution – annoyance – rumour – scientific technical information – bibliometric analysis – cartography – scientometrics

LES INFRASONS ENTRE SCIENCE ET MYTHE : LA BIBLIOMETRIE PEUT-ELLE CONTRIBUER A CLARIFIER UNE VERITE SCIENTIFIQUE CONTROVERSEE ?

INTRODUCTION

Les infrasons sont en général définis comme les ondes acoustiques de fréquence inférieure à 20 Hz, où l'on fixe la limite conventionnelle de la sensibilité en fréquence d'une oreille humaine. Vers le milieu des années 1960, certains chercheurs relatèrent des effets de forte ampleur des infrasons sur l'homme, et mentionnèrent en particulier des perturbations graves de la pression sanguine, du rythme cardiaque, de la respiration, du fonctionnement gastrique, ainsi que des altérations du système nerveux : perte d'équilibre, accroissement du temps de réaction, perte de vigilance, etc…

Ces premières constations, alarmantes compte tenu de leurs implications en médecine de travail et en médecine des transports routier et aérien, incitèrent d'autres équipes à se pencher sur le sujet et à mener de nouvelles expérimentations ; assez rapidement, dans la seconde moitié des années 1970, celles-ci suggérèrent que les premiers effets avaient été surestimés et qu'on pouvait généralement conclure à l'absence d'effets nocifs significatifs des infrasons sur l'état physique et mental des personnes, y compris à des niveaux d'exposition atteignant ou dépassant les 130 dB, au moins pour des expositions de courte durée.

Si les premières équipes avaient pu être victimes d'artefacts, cela est dû au caractère pluridisciplinaire du sujet. Dans le domaine des sciences physiques, l'expérimentation sur les ondes infrasonores est délicate, pour la génération comme pour la mesure ; et l'appréciation d'effets physiologiques suppose en même temps de bien maîtriser les méthodologies de la médecine expérimentale (analyses factorielles d'enquêtes sur des sujets humains).

Dans les années 1980, des travaux se poursuivirent, en particulier dans les pays nordiques (Danemark et Suède), en URSS et en Europe de l'Est, ainsi qu'au Japon, afin de mieux cerner les mécanismes d'interaction avec l'organisme et les effets de plus longue durée, et définir des niveaux d'exposition acceptables dans le cadre de préoccupations d'hygiène industrielle. Plus récemment, c'est la sensibilité croissante à la dégradation de l'environnement et des conditions de vie dans les grandes métropoles urbaines qui ont motivé les investigations des scientifiques.

Toutes ces études se caractérisent par une grande variété d'approches, dans les intentions comme dans les types d'expérimentations réalisées. Les résultats présentés sont parfois contradictoires et le plus souvent impossibles à comparer entre eux. Il semble admis par beaucoup que les infrasons sont en réalité audibles à partir d'une certaine intensité et que ce seuil d'audibilité est très proche du seuil de gêne (gêne ressentie souvent comme une pression dans l'oreille) – beaucoup plus proche que pour des fréquences plus élevées. Lorsque les infrasons sont perçus, un ralentissement des fonctions physiologiques (fatigue, perte de concentration) pourrait intervenir, mais ceci est parfois contesté. Selon d'autres auteurs, qui continuent à détecter expérimentalement des incidences sur l'équilibre et les rythmes respiratoire et cardiaque, certaines personnes seraient plus sensibles que d'autres, ou, hypothèse plusieurs fois formulée mais restée sans vérification, le deviendraient par suite d'une exposition prolongée. A côté d'expériences de laboratoire montrant des effets très légers, les infrasons continuent d'être désignés comme responsables des troubles les plus divers lors d'études menées suite à des plaintes liées aux conditions environnementales, alors même qu'il paraît difficile de les distinguer des vibrations, des bruits à fréquences plus élevées ou d'autres facteurs de pollution urbaine.

Les infrasons perturberaient-ils les capacités d'investigation et de raisonnement scientifiques ? Parmi les articles les plus récents, l'un mentionne sans le moindre début de preuve une sensibilisation de l'organisme soumis aux infrasons sur une longue durée. Un autre semble ignorer complètement les travaux menés dans les années 1980. Inversement, dans deux publications d'origines différentes, on retrouve des paragraphes entiers identiques au mot près. Dans un autre cas, on lit au détour d'une page des considérations surprenantes sur la « Nature » et l'évolution humaine pour expliquer l'absence d'organes perceptifs des ondes sismiques. Il arrive aussi, à la demande d'un couple âgé se plaignant d'insomnie, de sensation de fatigue et de picotements, qu'on traque dans un immeuble, à grands frais et pendant des semaines, un bruit d'appareil électroménager d'intensité *plus faible* que le seuil de sensibilité de l'oreille ; les mesures de bruit effectuées sont d'ailleurs perturbées par le trafic d'une rue fréquentée qui passe à dix mètres des fenêtres du logement, trafic dont ces personnes affirment n'être

pas affectées ! En somme, le processus cumulatif de la connaissance scientifique – expérimentations, échanges et confrontation rationnelle de résultats tangibles – ne semble pas aussi limpide qu'on pourrait le souhaiter.

De plus, les débats initiaux ont trouvé un écho dans le formidable médium de communication que constitue l'Internet, où la rumeur fait état, le plus sérieusement du monde, avec plus de trente ans de retard, d'effets plus ou moins dévastateurs produits par les infrasons.

Lorsqu'on est interrogé dans un cadre industriel sur l'effet sur les voyageurs des infrasons éventuellement produits par les différentes sources mécaniques du véhicule, la formulation d'une réponse claire au problème n'est évidemment pas facilitée par le foisonnement d'informations disparates voire parfois contradictoires.

Nous allons essayer de montrer sur ce cas particulier comment des techniques bibliométriques plus ou moins élaborées peuvent permettre de mieux dominer la question. Ceci en deux temps : d'abord comme aide à la décision, pour évaluer rapidement le niveau d'attention à accorder au sujet, ensuite pour mieux comprendre la confusion qui y règne et tenter de dégager des éléments d'appréciation sur le fond du problème posé.

1 - ANALYSE DES PAGES DE L'INTERNET :

Par la diversité des documents qu'on y trouve, la consultation de l'Internet permet de se faire rapidement une idée générale des différents aspects d'un sujet, y compris dans ses dimensions les plus subjectives. C'est d'ailleurs dans cette optique un outil irremplaçable.

Pour cette investigation nous avons utilisé le moteur de recherche Google qui présente l'avantage de présenter les pages dans l'ordre d'un indice de notoriété : ainsi les sites les plus référencés par d'autres sites à travers les liens hypertextes se trouvent-ils placés en tête de liste.

Les interrogations ont été les suivantes :

- Recherche dans les pages France avec le mots-clef « infrasons »
- Recherche dans les pages mondiales Web avec le mots-clef « infrasound » ;

Dans les deux cas les 100 premières pages ont été visitées et classées par catégories. Ces catégories sont les suivantes :

1) ***acoustique :*** il s'agit de pages qui présentent des informations sur les phénomènes acoustiques et qui évoquent les infrasons, sans mentionner d'effets particuliers sur la santé.
2) ***pertinents :*** il s'agit de pages qui présentent un intérêt particulier sur le sujet et dont nous détaillerons plus loin le contenu.
3) ***rumeur :*** il s'agit de pages qui font explicitement mention d'effets néfastes sur la santé et qui matérialisent donc les traces de cette rumeur.
4) ***médical :*** il s'agit de pages qui mentionnent l'utilisation des infrasons en kinésithérapie (électrophysiothérapie) ou dans le cadre d'autres usages thérapeutiques.
5) ***détection :*** il s'agit de pages qui mentionnent l'utilisation de systèmes de mesure par infrasons afin d'identifier des phénomènes atmosphériques ou sismiques d'origine naturelle (tornades, entrée de météorites dans l'atmosphère, éruptions volcaniques) ou artificielle (explosions, avions supersoniques, essais nucléaires). Une grande partie de ces sites sont liés aux organismes chargés de vérifier l'application du Traité de Non Prolifération des Armes Nucléaires ; beaucoup évoquent le réseau mondial de stations de surveillance déployé à cette occasion. En France, le CEA apparaît comme le principal acteur de cette activité.
6) ***armes :*** il s'agit de pages qui mentionnent explicitement des usages militaires des infrasons autres que la détection atmosphérique et sismique signalée ci-dessus, en particulier les armes à infrasons et certains types de sonars.
7) ***applications :*** il s'agit de pages (de fait, uniquement des pages France) qui mentionnent deux applications des infrasons : le ramonage et les alarmes anti-intrusion.
8) ***animaux :*** il s'agit de pages qui mentionnent le recours aux infrasons par certaines espèces d'animaux afin de communiquer ou de s'orienter, aussi bien en milieu marin qu'en milieu terrestre.
9) ***sans intérêt :*** on a regroupé sous ce titre les pages sans contenu informationnel pour ce qui nous concerne : celles qui n'ont en réalité rien à voir avec le sujet (groupes de musique, etc…), les simples mentions accidentelles, et les pages non accessibles.

20 août 2002	***France***	***Monde***
acoustique	32	14
pertinents	2	4
rumeur	9	7
médical	12	6
détection	14	40
armes	2	3
applications	8	0
animaux	10	11
sans intérêt	11	15
Total	**100** ***(sur 457 pages, soit 21,8%)***	**100** ***(sur 8740 pages, soit 1,1 %)***

Les pages remarquables, classées en « pertinents » sont les suivantes :

- *Pages France :* l'un des documents traite des affections liées au bruit, le second est relatif à une dépêche de l'AFP sur une pollution sonore aux causes non identifiées dans une ville du Mexique.
- *Pages Monde :* les 4 documents mentionnés abordent sur des bases médicales et scientifiques sérieuses l'effet des ultrasons sur la santé. L'un d'entre eux ([Haneke, 2001]) est une synthèse bibliographique toute récente (novembre 2001) et très exhaustive du sujet, réalisée à l'instigation des autorités sanitaires nord-américaines (National Institute of Environmental Health Sciences) suite à certains malaises inhabituels de résidents d'une ville de l'Indiana ; ce document met d'ailleurs bien en relief l'ambiguïté de la problématique sur les effets réels des infrasons décrite dans l'introduction de la présente communication.

Il est difficile de déterminer si les différences observées par comparaison entre les deux colonnes sont dues à une spécificité française sur le sujet ou tout simplement à une profondeur d'exploration des pages de l'Internet très différente.

Cette investigation permet de se faire une idée globale du sujet ; parmi les éléments notables :

1. Une pratique physiothérapique reconnue de traitement par vibrations à basse fréquence ;
2. Une utilisation récente et importante pour la détection de phénomènes sismiques et atmosphériques, dont l'impulsion a été donnée par le Traité de Non Prolifération nucléaire ;
3. Quelques indices d'essais d'utilisation, sous forme d'arme offensive bien que prétendue « non létale », qui semblent avoir été menés dans différents pays, sans grand succès semble-t-il pour des raisons liées à la nature même des ondes (voir [Altmann, 2001] pour une analyse approfondie du sujet).
4. Quelques mentions sur les effets nocifs (purement imaginaires) des infrasons. S'il est assez facile d'identifier le caractère suspect de ces informations (sites non officiels, propos alarmistes, effets littéraires, densité des coquilles et erreurs diverses, absence de citations de sources scientifiques sérieuses ou récentes, répétition des mêmes anecdotes sur différents sites presque dans les mêmes mots d'une langue à l'autre), des publications sérieuses sont toutefois contaminées (http://www.techniques-ingenieur.fr/affichage/DispIntro.asp?ngcmID=G2790, consulté le 27/08/2002).

On peut esquisser le mécanisme de la rumeur, édifiée sur la crainte (non fondée) que les infrasons puissent produire des effets nocifs sans être perçus, et alimentée par les items 2. et 3. relevés ci-dessus. Le doute est accentué par le manque de netteté scientifique qu'on trouve sur le sujet, qui prend du coup parfois une dimension mythique. Par ailleurs, il y a confusion entre les vibrations (voie solidienne) et le bruit (voie aérienne). Comme le prouve l'item 1., la voie solidienne ne semble pas présenter d'aspect nocif (évidemment sous des intensités modérées et pour des expositions de courte durée).

A partir de ces éléments issus de l'Internet, il est donc possible au veilleur de formuler une première réponse à la question posée : absence de danger grave et avéré, caractère fallacieux de certaines craintes exacerbées par une information distordue, incertitude scientifique sur des effets secondaires, nécessité d'approfondir l'enquête.

Le type de traitement effectué peut apparaître extrêmement « rustique », voire laborieux, et surtout très partiel, puisqu'on ne fait qu'effleurer la surface de la partie la plus visible de l'Internet. Il présente l'avantage d'être facilement réalisable. Sous réserve d'être effectué avec un peu de méthode, il n'est pas non plus particulièrement gourmand en temps ; il permet d'esquisser une quantification du domaine abordé et aussi d'obtenir, avec un peu de chance, quelques documents très précieux, comme nous l'avons mentionné.

Par ailleurs, le traitement plus automatisé des pages de l'Internet, qui sont balisées dans un format spécifique et rédigées en langage naturel, nécessitent d'utiliser des logiciels spécialisés relativement onéreux et dont on dispose rarement. L'analyse directe des pages de réponses fournies par le moteur de recherche selon un traitement bibliométrique identique à des notices bibliographiques (les champs « Titre », « Texte » et « Adresse » par exemple pourraient être repérés) est aussi assez difficile, parce que ces notices sont à la fois brèves et en langage naturel : elles nécessitent donc souvent d'être interprétées.

2 -ANALYSE BIBLIOMETRIQUE DE MEDLINE :

Nous avons recherché des informations pertinentes sur les effets physiologiques des infrasons en consultant une base de données bibliographique spécialisée dans le domaine.

La base de données Medline est reconnue comme une référence et un reflet fidèle (pour tout ce qui est publié) des activités de recherche scientifique par les chercheurs du monde médical : la diversité dans l'origine des documents identifiés dans la présente étude le montrera clairement.

Cette base, librement consultable en ligne par le biais de l'Internet, est donc susceptible de nous donner une idée correcte de l'état des travaux menés sur les effets des infrasons sur la santé humaine et des résultats qui en sont issus.

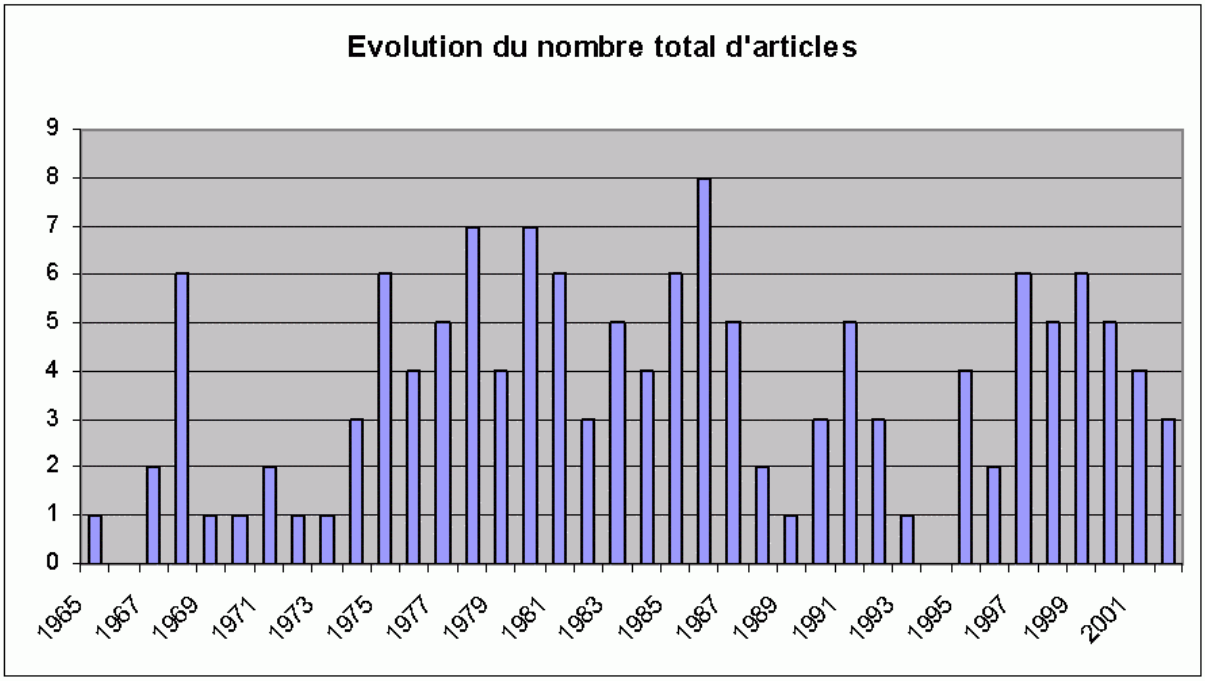
Nous avons donc interrogé Medline via l'interface PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) selon « infrasound OR infrasonic OR infrasonics » qui nous a fournit 138 références. Pour les articles pertinents, l'accès aux « Related Articles » a également été tenté, sans fournir de retours substantiels.

Afin de situer l'importance de la littérature relative aux infrasons dans l'ensemble de la problématique médicale telle qu'elle est reflétée par les publications, nous avons également interrogé avec des mots-clés relatifs à des affections majeures de nature diverse : la sclérose en plaques (« multiple sclerosis »), la lèpre (« leprosy »), le syndrome de Creutzfeldt-Jakob, et, ce qui relève de thèmes médicaux plus proches de nos préoccupations, les affections relatives à l'amiante (« asbestos ») et les effets de l'exposition au bruit dans un contexte professionnel (pour lesquels nous avons choisi une association de mots-clés parmi bien d'autres imaginables, qui a pour principal mérite de limiter … le bruit dans les réponses, mais qui ne constitue évidemment pas une exploration exhaustive de la base sur le sujet.)

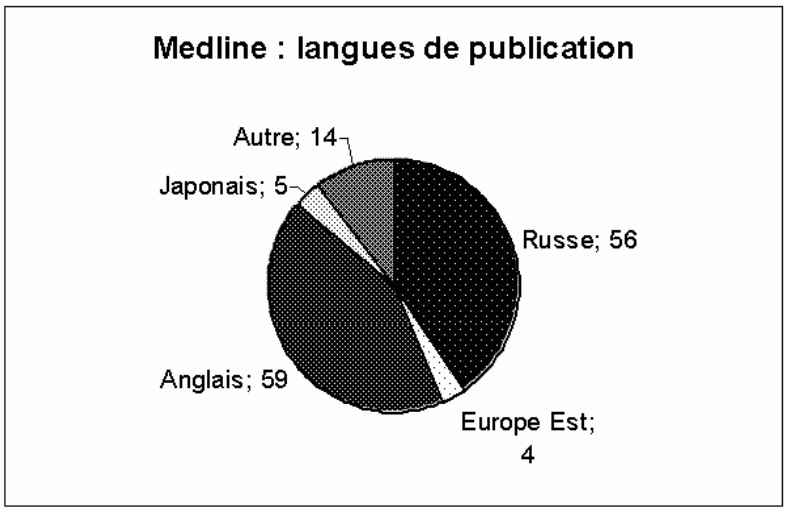
Mots-clés utilisés : le 31 août 2002	Nombre d'articles de la base Medline
« multiple sclerosis »	24 121
leprosy	13 846
creutzfeldt-jakob	3 866
asbestos	7 999
noise AND « occupational exposure »	367
infrasound OR infrasonic OR infrasonics	138

On s'aperçoit immédiatement à la lecture de ce tableau comparatif du **caractère extrêmement marginal**, dans la littérature publiée au moins, de l'étude des infrasons.

Il importe toutefois, au-delà de ces chiffres bruts, de saisir la dynamique des publications : on pourrait par exemple se trouver dans le cas d'un domaine en émergence. Il faut aussi examiner la nature des documents référencés pour s'expliquer les raisons d'une éventuelle absence de publication.

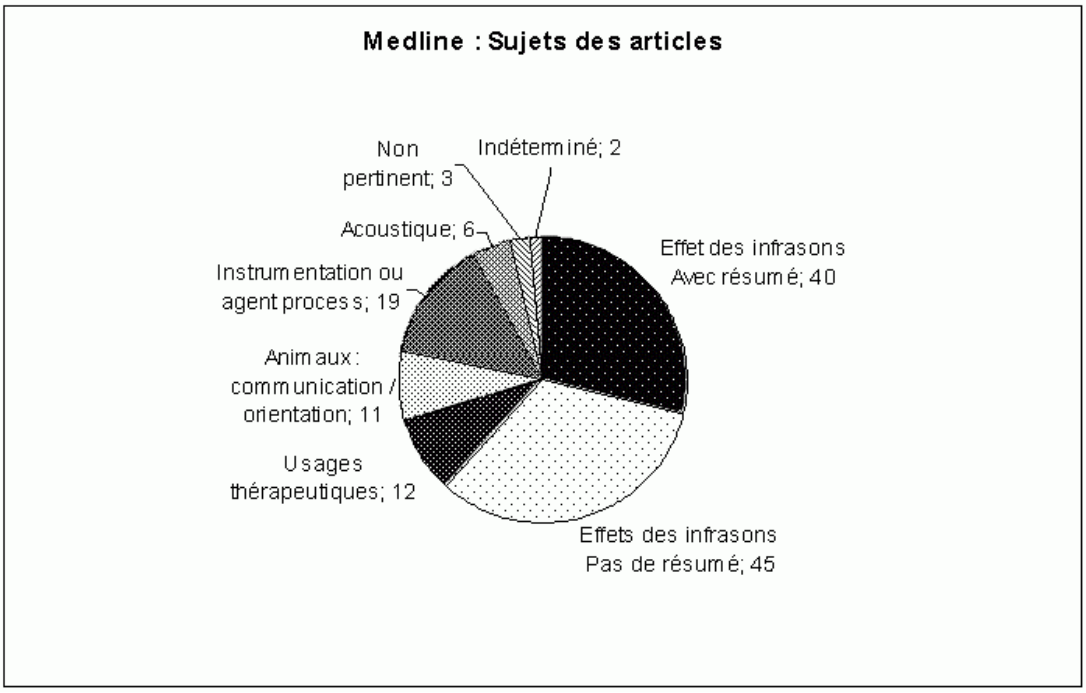


L'analyse de l'évolution du nombre d'articles du corpus total publiés au cours du temps fait apparaître au contraire un domaine relativement ancien, stagnant, avec une activité maximale entre 1975 et 1988 suivie d'une seconde vague à partir de 1995 (les données pour l'année 2002 étant incomplètes).



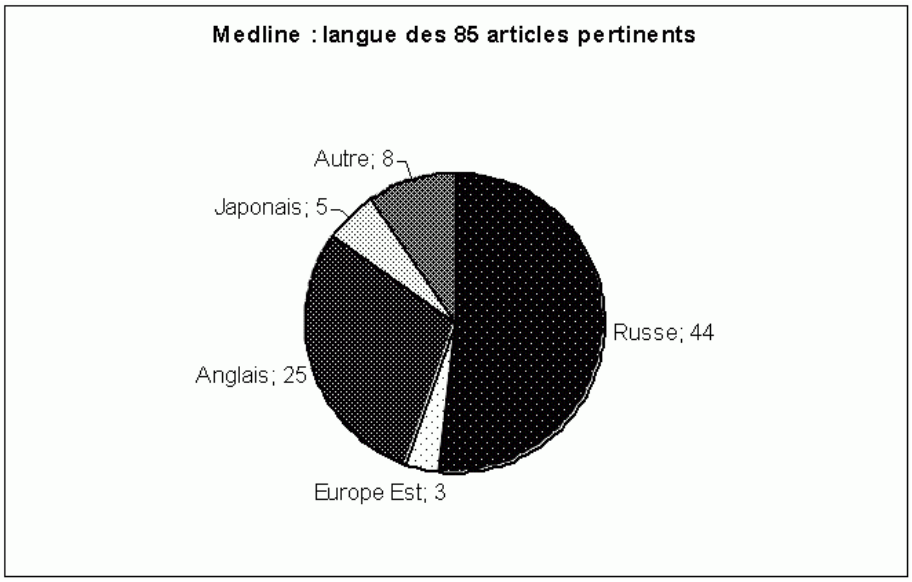
Pour ce qui concerne la langue des articles, qui donne une idée de l'espace de diffusion des publications, à défaut, en particulier quand il s'agit de l'anglais, de la nationalité précise d'origine des auteurs, une spécificité apparaît : une grande partie des articles ont été publiés en russe ou dans une langue d'Europe de l'Est.

Nous nous sommes intéressé plus en détail au contenu des documents : comme ils étaient assez peu nombreux, une analyse a été effectuée à partir de leur titre et, quand il était disponible sur Medline, de leur résumé, ce qui a permis de définir un classement en plusieurs catégories :



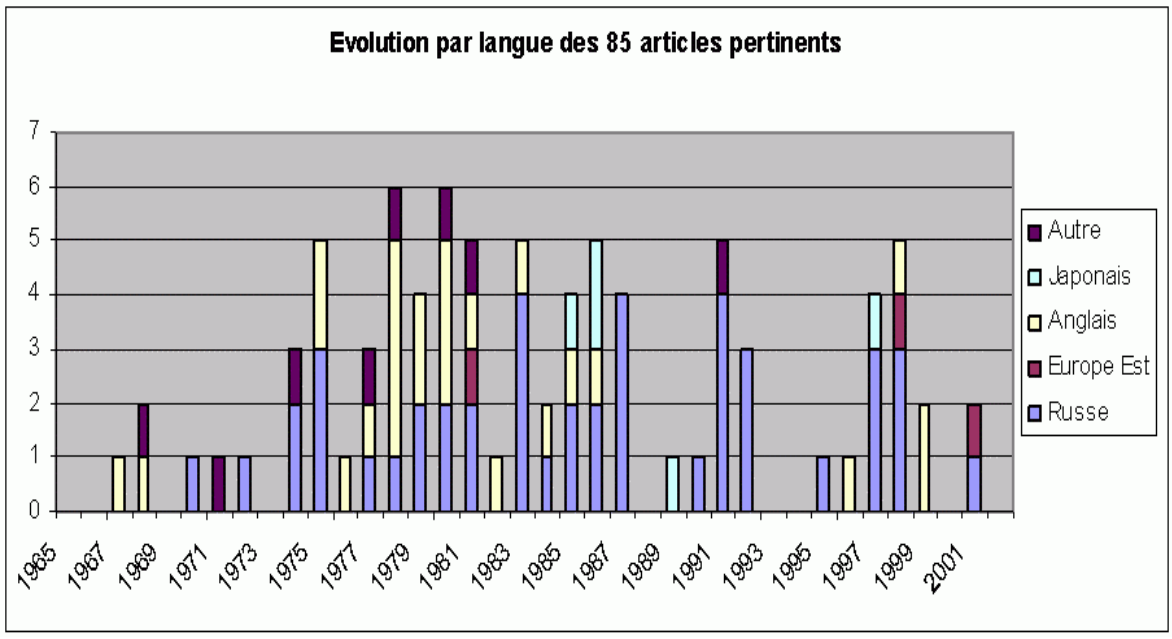
- ***Effets des infrasons :*** les documents qui traitent effectivement des effets des infrasons constatés sur des organismes vivants, homme ou animaux de laboratoire, ou sur des cellules in vitro. Certains de ces documents disposent dans Medline d'un résumé qui permet d'en mieux apprécier le contenu exact ; d'autres (principalement des publications en langues slaves) en sont dépourvus : l'incertitude sur le contenu est donc plus grand pour ce qui concerne ces derniers documents.
- ***Usages thérapeutiques :*** les documents qui décrivent sans ambiguïté des applications thérapeutiques des infrasons à des fins de traitement : en particulier, un certain nombre d'expérimentations ont été menées en Russie en ophtalmologie. Il est apparu souhaitable de distinguer ce type d'article des précédents.
- ***Instrumentation ou agent de process :*** les documents qui décrivent l'utilisation d'infrasons en instrumentation médicale ou dans certains process de préparation de solutions de laboratoire.
- ***Animaux : communication, orientation :*** les documents relatifs à l'utilisation des infrasons dans le règne animal à des fins de communication ou d'orientation dans l'espace.
- ***Acoustique :*** les documents qui traitent des aspects strictement acoustiques des infrasons sans référence à des aspects physiologiques.
- ***Non pertinent :*** les documents non pertinents (bruit dans la recherche).
- ***Indéterminé :*** quelques documents dont nous n'avons pas réussi à déterminer le contenu exact à partir des références de Medline.

Nous allons donc nous intéresser maintenant au 85 articles qui traitent effectivement d'éventuels effets nocifs des infrasons sur la santé. La répartition par langue de ces articles fait apparaître encore plus nettement la domination slave :



Il convient de noter que la plupart des articles en langue russe référencés dans Medline n'ont pas de résumé.

Si nous visualisons la dynamique de ces 85 publications par langue, une singularité géographique apparaît :



Le sujet, tel qu'il est reflété par Medline, semble avoir été abandonné après la poussée de 1985, sauf dans l'aire slave. Les trois articles en anglais de 1998-1999 ne doivent pas faire illusion : deux d'entre eux proviennent d'une équipe polonaise.

En résumé, d'après la vision qu'en donne Medline, le sujet de l'effet des infrasons sur la santé semble être une problématique très marginale en terme de santé publique. De plus, cette problématique n'est pas récente et apparaît au mieux en stagnation depuis le début des années 1980. Seuls à avoir conservé une certaine activité sur ce sujet, les Européens de l'ancien bloc Soviétique en ont pris la quasi-exclusivité : cette caractéristique, compte tenu des langues de publication, pourrait (on le vérifiera plus loin) ne pas faciliter l'accessibilité de leurs recherches par d'autres équipes. Mais ce désintérêt pour le sujet, qu'au-delà du contenu des articles nous révèle la bibliométrie, nous permet déjà, par cette analyse rapide, de préconiser à notre interlocuteur industriel de ne pas accorder trop de crédit à cette rumeur – dont l'examen de l'Internet nous donne par ailleurs les fondements.

Ces résultats ont été obtenus en un délai court et sans l'utilisation d'outils bibliométrique lourds ; il est clair que le faible volume du corpus a permis une telle opération, qui aurait toutefois été possible grâce à des outils bibliométriques classiques.

Nous allons maintenant examiner comment un examen plus approfondi du sujet va nous permettre de nuancer ces premiers éléments.

3 - ANALYSE BIBLIOMETRIQUE DE LA BASE DE DONNEES PASCAL :

Nous avons souhaité compléter les résultats donnés par Medline en interrogeant la base Pascal de l'Inist, qui présente l'avantage d'être pluridisciplinaire, ce qui convient bien pour couvrir les divers aspects de notre sujet.

a) Analyse préalable sous format B et sélection du corpus :

Nous avons interrogé la base en recherchant le mot « infrasound » dans les termes contrôlés (CT) et obtenu 138 réponses, que nous avons téléchargées pour une première analyse sous le format réduit B « Browse » . Ce format présente les informations suivantes :

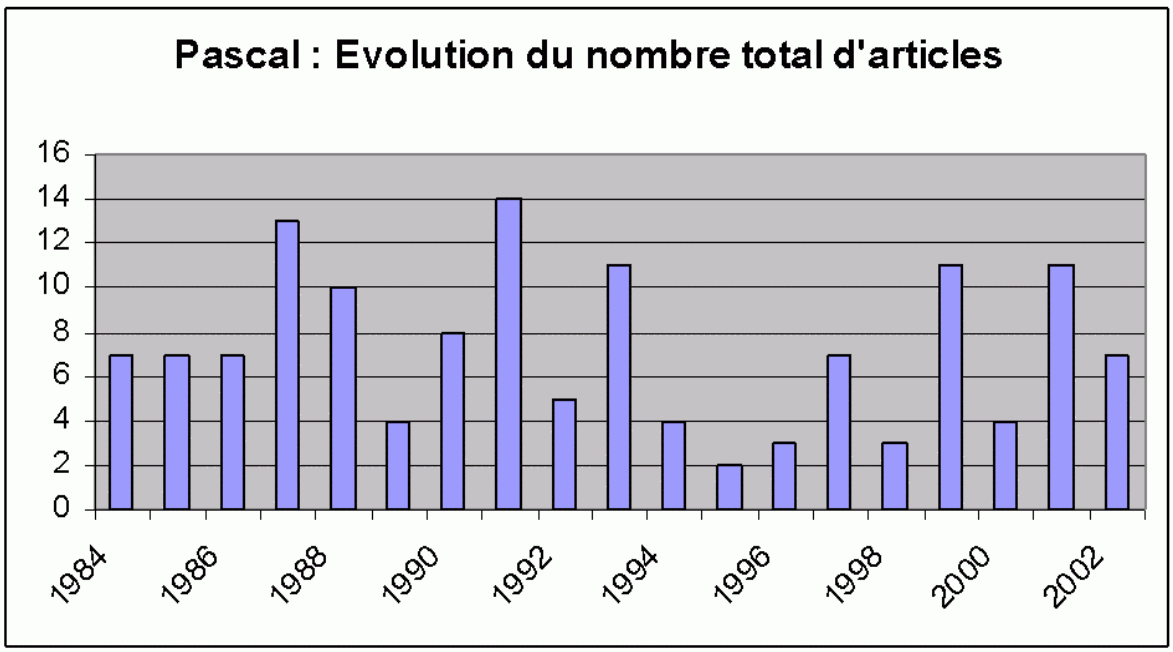
- le **numéro de référence** Pascal (Native Number) qui comporte l'année d'enregistrement du document dans Pascal, et donc, une information indirecte, mais néanmoins exploitable par comparaison, sur la date de l'article lui-même,
- le **titre** du document, en anglais et le cas échéant en français,
- le **code de classification** du document avec la liste des grands domaines dans lequel il peut être classé, en anglais, français et espagnol,
- les **termes contrôlés** (mots-clefs) en anglais et en français, parfois en espagnol,
- pour certains enregistrements, des **termes étendus** (broader terms) qui donnent des mots-clefs supplémentaires plus généraux.

Typing set: 1, record: 1
Native Number : 2002-0291368 PASCAL
Title in English : Danish guidelines on environmental low frequency noise, infrasound and vibration
Classification Code : 001B40C50; Physics; Acoustics; Pollution, Nuisances
Classification Code in French : 001B40C50; Physique; Acoustique; Pollution, Nuisances
Classification Code in Spanish : 001B40C50; Fisica; Acustica; Polucion, Contaminacion ambiental, Ruido ambiental
Physics and Astronomy Code : 4350
Controlled Terms : Noise pollution; 1/f noise; Infrasound; Ambient noise; Standards; Regulation
Controlled Terms in French : Nuisance acoustique; Bruit basse frequence; Infrason; Bruit ambiant; Norme; Reglementation; 4350
Controlled Terms in Spanish : Nocividad acustica; Ruido baja frecuencia; Infrasonido; Ruido ambiente; Norma; Reglamentacion

Ce format est donc susceptible de donner au bibliométricien des renseignements généraux sur le contenu domaine, y compris sur de grands corpus, sans engager de frais de téléchargement de notices prohibitifs, ce qui peut constituer une difficulté non négligeable dans ce genre d'étude. Il peut permettre alors de mieux cibler les documents dont on va télécharger les notices dans leur intégralité.

Nous avons donc traité le fichier texte des notices à l'aide d'un programme d'analyse bibliométrique expérimental de notre conception écrit en langage Perl, qui, à partir de la description du format des notices, extrait et classe les formes des différents champs, puis génère les paires de termes que l'on peut trouver dans les différentes notices.

Laissant de côté les titres en langage naturel dont l'analyse bibliométrique est plus hasardeuse, nous avons donc extrait les **dates**, **codes de classification** et **mots-clés en français** pour les différents documents.



L'évolution du nombre total de documents référencés par Pascal couvre un nombre d'années plus restreint que Medline, mais sa dynamique n'est pas différente malgré la dimention plus pluridisciplinaire de Pascal ; seule la période d'activité maximale est décalée vers le début des années 1990.

Si l'on s'intéresse maintenant aux disciplines auxquelles se rattachent les articles, les codes de classification les plus utilisés (notez qu'un article peut être référencé par plusieurs codes) avec leur fréquence d'apparition (nombre de références différentes où le code se trouve présent) montrent, dans la diversité des domaines couverts, une présence notable de documents relatifs à notre sujet (domaines en italique gras) :

Codes de classification	Fréquence
Physique	65
Acoustique	58
sciences de la vie	42
sciences biologiques	39
sciences appliquees	21
sciences de l'univers	18
pollution, nuisances	16
physiologie des vertebres, neurophysiologie des vertebres, systeme nerveux	14
geophysique externe	14
Meteorologie	10
psychologie	6
batiment, travaux publics, genie civil	5
metrologie	5

Il en est de même pour ce qui concerne les mots-clés les plus fréquemment cités :

Mots-clés	Fréquence
infrason	135
etude experimentale	21
homme	19
bruit	18
acoustique sous marine	16
basse frequence	15
nuisance acoustique	14

audition	13
bruit ambiant	11
acoustique atmospherique	10
stimulus acoustique	9
onde acoustique	9
propagation onde	8
mesure	8
frequence audible	7
effet biologique	7
vibration	6
perception	6
gene nuisance	6
appareil auditif	6
traitement signal	5
systeme nerveux central	5
seisme	5
oreille interne	5
explosion nucleaire	5
environnement	5
bruit industriel	5

Pour tenter de sélectionner les articles les plus pertinents, nous avons ensuite interrogé Pascal à l'aide des mots-clés les plus fréquemment cités qui ont paru répondre à la problématique, selon l'équation suivante :

INFRASOUND*/CT AND (***HUMAN***/CT OR* **NOISE***/CT OR «***NOISE POLLUTION***»/CT OR «***AMBIENT NOISE***»/CT OR «***BIOLOGICAL EFFECT***»/CT OR* **ANNOYANCE***/CT OR* **PERCEPTION***/CT OR «***CENTRAL NERVOUS SYSTEM***»/CT OR* **ENVIRONMENT***/CT OR «***INDUSTRIAL NOISE***»/CT*

Nous avons obtenu en retour 53 documents dont nous avons téléchargé les notices pour traitement. Nous avons ensuite extrait les termes des champs Auteurs et Mots-Clés pour les analyser.

En éditant les listes respectives des auteurs référencés dans Medline et Pascal, il est frappant de constater la forte complémentarité de ces deux bases : pour 121 auteurs cités dans le corpus de Pascal et 194 auteurs cités dans notre précédent corpus de Medline, on ne trouve que 12 auteurs cités à la fois dans les deux corpus :

1. Augustynska A.
2. Densert O.
3. Glinchikov V.
4. Kaczmarska- Kozlowska A
5. Kamedula M.
6. Klinke R.
7. Landström U.
8. Langner G.
9. Nekhoroshev A.S.
10. Pawlaczyk- Luszczynska M.
11. Scheich H.
12. Schermuly L.
13. Theurich M.

On vérifie une fois de plus sur cet exemple la nécessité de consulter plusieurs bases pour bien se documenter sur un sujet. En l'espèce, ce recouvrement paraît tout de même très limité ; il pourrait illustrer, outre la marginalité du sujet, la diversité des travaux menés en parallèle et un certain manque de cohésion induit par exemple par l'absence d' « autorités » susceptibles de générer des synthèses. D'où la difficulté de discerner l'état de l'art scientifique, y compris pour des chercheurs qui souhaiteraient aborder le domaine.

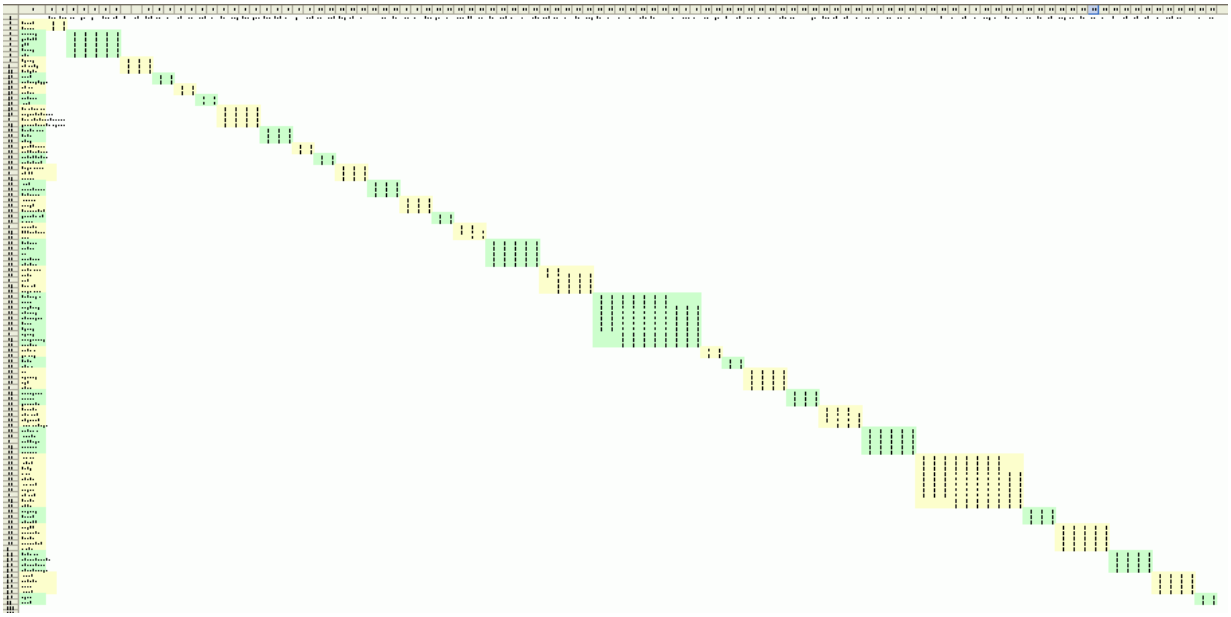
De plus, les auteurs présents dans ce recouvrement ne sont pas toujours les plus actifs dans le domaine. Surtout, on constate l'absence de part et d'autre d'auteurs indubitablement importants, comme par exemple H. Möller dans Medline et V.N. Alekseev dans Pascal.

Pour mieux apprécier le contenu des différentes problématiques traitées dans ces articles, il est classique de mettre en évidence les associations de mots-clés entre eux, à partir des paires de termes extraites des notices dans le champ des termes contrôlés. Nous avons pour cela effectué une sériation par blocs, procédé qui offre l'avantage, sans définir de catégories a priori, de regrouper les notices à travers des termes communs mis en évidence, et permet donc de structurer sémantiquement le corpus (voir sur cette technique bibliométrique [Baldit, 1994] et [Rostaing, 1996]). L'algorithme que nous utilisons est perfectible mais il offre l'avantage d'écarter les lignes qui correspondent à des termes triviaux, présents dans un nombre jugé excessif de références : cette précaution permet de constituer les blocs avec plus de netteté.

Il est clair que, comme tout traitement massif de données, l'apport de cette technique est particulièrement appréciable dans le cas de grands corpus. Pour ce qui nous concerne, au contraire, le faible nombre de notices présentes fragilise les conclusions qu'on pourra en tirer.

b) Analyse des relations entre les auteurs :

La sériation fait apparaître 30 groupes d'auteurs d'importance numérique variable (2 à 10) mais parfaitement distincts entre eux, ce qui confirme l'éparpillement du traitement de la problématique.



c) Analyse sur les mots-clés entre eux :

Seules ont été retenues pour la sériation les paires de mots-clés qui apparaissent deux fois ou d'avantage dans le corpus ; ceux-ci sont considérés comme seuls significatifs. La sériation révèle alors 6 groupes de mots-clés (1 à 6) nettement différenciés, plus 2 groupes qui s'interpénètrent partiellement l'un dans l'autre (7 et 8) :

MOTS-CLES	GROUPE
mesure	1
hydrophone	1
action vent	1

bruit oceanique	1
eau profonde	1
appareil auditif	2
canal semicirculaire	2
oreille interne	2
bruit ambiant	Commun à 1 et 3
acoustique sous marine	Commun à 1et 3
niveau bruit	3
interaction onde	3
microseisme	3
onde surface	3
non linearite	3
bruit industriel	7
exposition professionnelle	7
	7
acoustique structurale	7
psychoacoustique	7
fenetre	7
etude experimentale	7
japon	7
nuisance acoustique	7
trafic routier	8
ventilation	8
basse frequence	8
vibration	8
gene(nuisance)	8
encephale	8
effet biologique	8
systeme nerveux central	8
electrophysiologie	8
facteur milieu	8
niveau acoustique	8
appareil respiratoire	4
fonction vestibulaire	4
pression acoustique	4
son pur	4
hormone adenohypophysaire	5
hormone steroide	5
hormone peptide	5
acth	5
systeme hypophysocorticosurrenalien	5
hormone surrenalienne	5
environnement	Commun à 5 et 6
stimulus acoustique	6
audition	6
perception	6
frequence stimulus	6
stimulus infraliminaire	6
frequence audible	6
temps exposition	6

Les groupes 1 et 3 qui traitent d'acoustique sous-marine ne nous intéressent pas ici (nous n'aurions pas dû utiliser « bruit ambiant » dans l'interrogation, nous l'aurions évité en faisant cette analyse préalablement à partir

du format B). La pertinence du groupe 5 est hypothétique. Les autres groupes correspondent bien par contre à notre sujet.

Remarquons qu'il y a nettement moins de groupes de mots-clés que de groupes d'auteurs. Il est maintenant intéressant d'examiner la relation des auteurs avec les mots-clés.

d) Analyse sur les auteurs et les mots-clés :

Pour cette analyse, n'ont été retenues comme précédemment que les paires (auteurs x mots-clés) présentes deux fois dans le corpus ; ceci suppose que les auteurs considérés aient publié deux fois parmi les 54 articles, et avec les mêmes mots-clés … ce qui sera peu fréquent pour un corpus si restreint, sauf s'il y avait une problématique centrale dominante.

Si l'on élimine ce qui se rapporte à l'acoustique sous-marine, nous obtenons simplement le tableau suivant :

	landstrom u.	moller h.	xiang zhang	zhou fei	jia ke-yong	li zhi-gang	chen jing-zao	andresen j.	inukai y.	takigawa h.	sakamoto h.	nishimura k.
perception	2											
audition	2	2										
son pur		2										
encephale			2	2	2	2	2					
systeme nerveux central			2	2	2	2	2					
effet biologique			2	2	2	2	2					
homme	2	4	2	2	2	2	2	2		2	2	
infrason	2	4	2	2	2	2	2	2	2	2	2	3
fonction vestibulaire										2	2	
hormone steroide												2
Systeme hypophysocorticosurrenalien												2
hormone surrenalienne												2
hormone peptide												2
acth												2
environnement												2
Hormone adenohypophysaire												2

Vérification faite, les chercheurs de patronymes chinois sont co-auteurs des mêmes articles (expérimentations sur des animaux de laboratoire). Par contre, il est intéressant de trouver côte à côte Ulf Andresen et Henrik Möller, qui sont les chefs de file de deux équipes scandinaves qui ont beaucoup travaillé sur le sujet, sans avoir publié ensemble. On conçoit donc que sur des corpus plus importants, ce tableau confronté à celui des auteurs puisse être riche d'enseignement ; ici, il ne fournit pas d'information très utile, si ce n'est de confirmer l'absence de thèmes centraux.

En résumé, cette analyse rapide des notices de Pascal n'invalide en rien ce qui avait été constaté sous Medline : une problématique marginale, ancienne et peu dynamique, des équipes assez nombreuses mais dispersées sans collaborations communes, sur des thèmes eux aussi très divers. L'originalité et la diversité des démarches de recherche scientifique est évidemment positive ; dans les conditions présentes, elle pourrait toutefois favoriser des points de vues divergents voire contradictoires sur la portée réelle des nuisances infrasonores – et par conséquent laisser une incertitude propice à la rumeur. Une certaine cohérence semble donc émerger de ces observations.

4 - L'ENTRELACEMENT DES CO-CITATIONS ENTRE LES EQUIPES DE CHERCHEURS :

Nous avons voulu pousser plus loin l'analyse en essayant de dresser une cartographie de l'activité de recherche relative à l'effet des infrasons sur la santé humaine à partir des documents qui nous étaient accessibles. Pour ceci, nous avons examiné les notices de Pascal issues de notre interrogation et nous nous sommes procuré 29 articles qui nous paraissaient répondre à la problématique ; 26 ont été finalement retenus (les expérimentations sur les animaux, dont la transposition sur l'homme est des plus délicates, ont été laissées de côté). Pour chaque article, les références bibliographiques des auteurs cités ont été examinées et chaque document a été enregistré sous une forme simplifiée qui comporte uniquement **le nom** du ou des auteurs, dans l'ordre, et **l'année de publication**.

(On pourrait évidemment utiliser la base SCISEARCH de l'ISI à cette fin. Toutefois, le champ CR – ou CA – des notices ne donne qu'un seul auteur, ce qui ne permet pas de connaître directement les autres).

Nous avons ainsi été en mesure de tracer le réseau de citations des publications entre elles et leur entrecroisement au cours du temps. Nous espérions ainsi mieux comprendre l'élaboration au cours du temps de la création du savoir scientifique de chercheur en chercheur et de « collège » en « collège », et mieux comprendre comment les connaissances actuelles sur les infrasons avaient pu se constituer au cours des années.

L'approximation commise dans cette analyse consiste à confondre éventuellement deux publications qui auraient été faites la même année par le ou les mêmes auteurs (listés dans le même ordre). Ce cas existe évidemment et s'est d'ailleurs présenté dans notre corpus ; toutefois, ce qui nous intéresse ici, ce n'est pas la publication en tant que telle, c'est l'état de l'expérience et de la réflexion de son ou de ses auteurs à un moment donné, telles que les publications ont pu le révéler à ceux qui les ont lues et citées. Dans le cas de plusieurs publications du même auteur la même année, on peut penser qu'il n'y a pas eu d'évolution significative en un laps de temps si court. Par contre, nous avons considéré les publications comme distinctes lorsque l'ordre de citation des auteurs était différent, car dans ce cas nous avons jugé qu'il pouvait y avoir des nuances ou des apports significatifs. Ainsi un seul document « Möller H. 1984 » a été retenu, mais on a distingué « Möller H., Andresen J. 1984 » et « Andresen J. , Möller H. 1984 » comme deux publications distinctes.

Ont été aussi écartées toutes les citations à des normes qui se répétaient dans un certain nombre d'articles. Ce documents sont importants par leur contenu, mais d'une nature différente de celle des publications scientifiques.

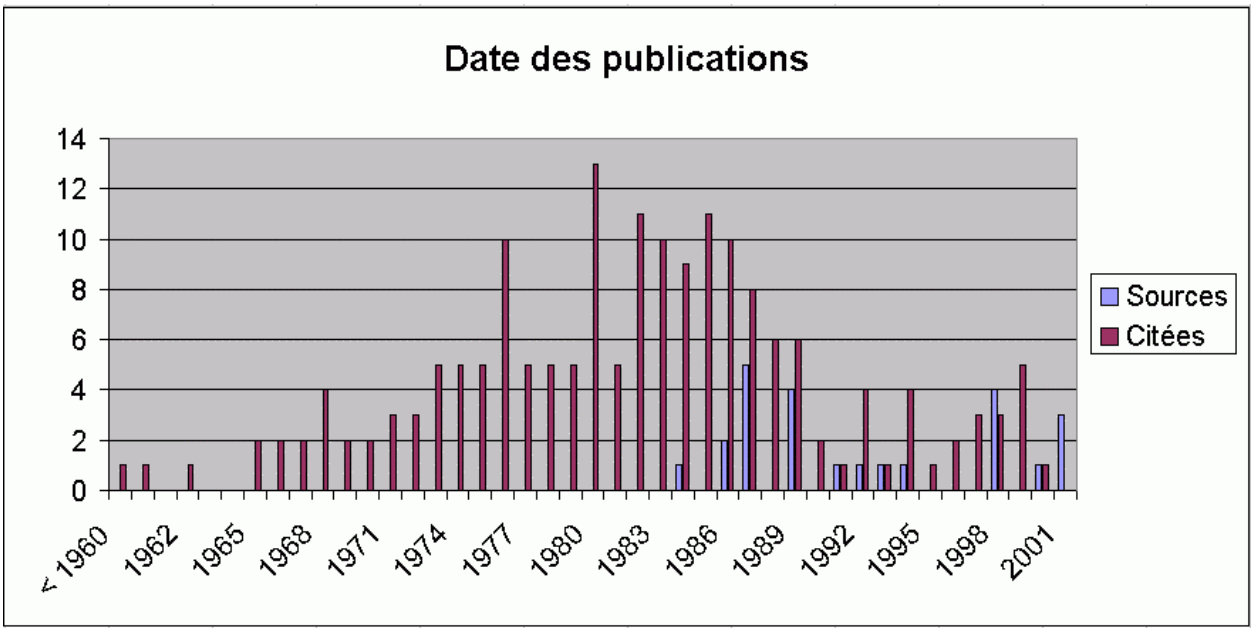
Comme nous n'avons travaillé qu'à partir de références issues de la base Pascal, cette démarche peut paraître très limitée : cette vision à travers une petite lorgnette est-elle réellement susceptible de nous faire découvrir la réalité, c'est-à-dire l'ensemble des travaux menés sur l'effet des infrasons sur l'organisme humain ? Le faible recouvrement des bases Pascal et Medline, que nous avons mentionné, semble appuyer une telle objection.

On peut y apporter deux réponses :

1) S'il semble clair que les publications référencées dans une base particulière ne couvrent qu'une partie des travaux menés pour le sujet, d'une part, et s'il en est de même pour les références citées dans chaque publication particulière, d'autre part, le champ qu'on découvre par contre en agglomérant l'ensemble devient plus vaste avec le nombre de publications consultées, et les risques d'erreurs (passer à côté de travaux importants) décroît aussi. Nous avons déjà signalé que le moteur de recherche Google définissait son indice de pertinence à partir d'un processus similaire.
2) La seconde réponse est plus fondamentale. La réalité perçue par l'homme est une construction de l'esprit ([Watzlawick, 1988]) : il est illusoire de croire qu'elle préexiste à sa prise de conscience. En l'espèce, quel que soit ce qui a été publié sur les infrasons, comment en prendre connaissance aujourd'hui ? Il n'y a pas beaucoup d'autres possibilités que de consulter quelques ouvrages fondamentaux et certaines bases de données (en pratique toujours en petit nombre), puis de se procurer les articles les plus accessibles (malheureusement pas ceux qu'on ne sait trouver ou qu'on ne saura déchiffrer du fait de leur origine géographique), de lire leurs références et d'essayer encore d'obtenir, si l'on en a les moyens et le loisir, ces autres documents ou les revues signalées. Ce processus comporte une part inévitable de risque et de contingence, même si l'on peut toujours compléter et aller plus loin. Il correspond à la démarche que nous avons suivie.

a) Chronologie des articles cités :

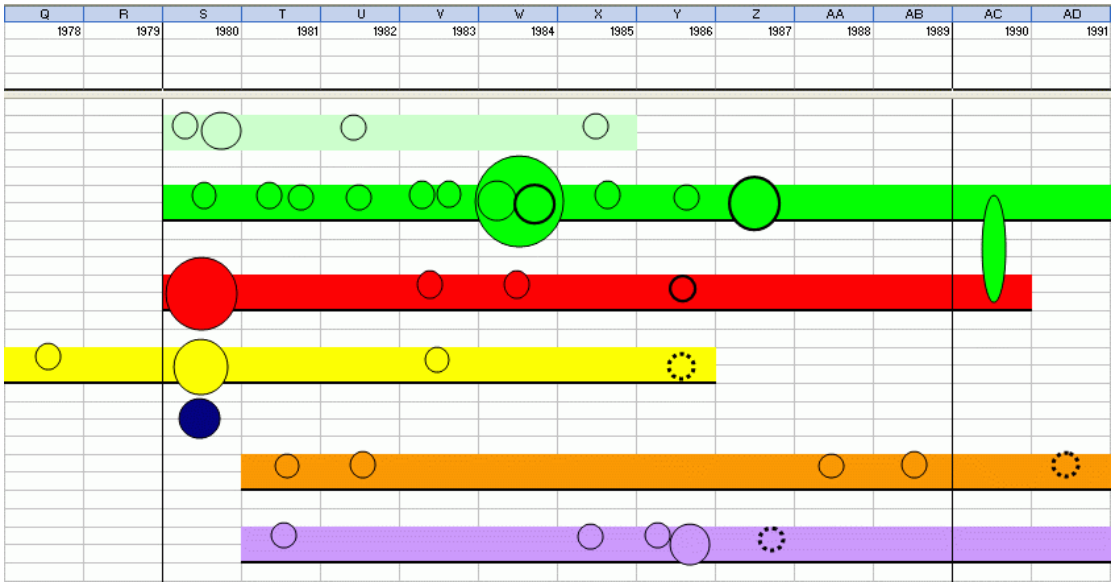
Pour commencer, nous avons représenté l'évolution des recherches sur les effets des infrasons à travers la date des documents cités en références de ceux dont nous disposions. Ceci élargit la perspective, comme on l'a dit, en mettant en évidence le socle à partir duquel s'élèvent les travaux qui correspondent aux publications consultées : nous obtenons un historique du sujet dressé par les chercheurs eux-mêmes. On vérifie que le sujet est actif à partir de 1975 avec un maximum entre 1980 et 1990 ; à partir de cette date, il apparaît nettement délaissé : bien sûr, les publications susceptibles d'être citées sont mécaniquement moins nombreuses à mesure qu'on s'approche du présent, mais la rupture de tendance est néanmoins très nette. Sous réserve d'une renaissance toujours possible, le cœur de l'activité sur le sujet apparaît aujourd'hui être du domaine d'un passé qui ne cesse de s'éloigner. En somme, la question semble dépassée.



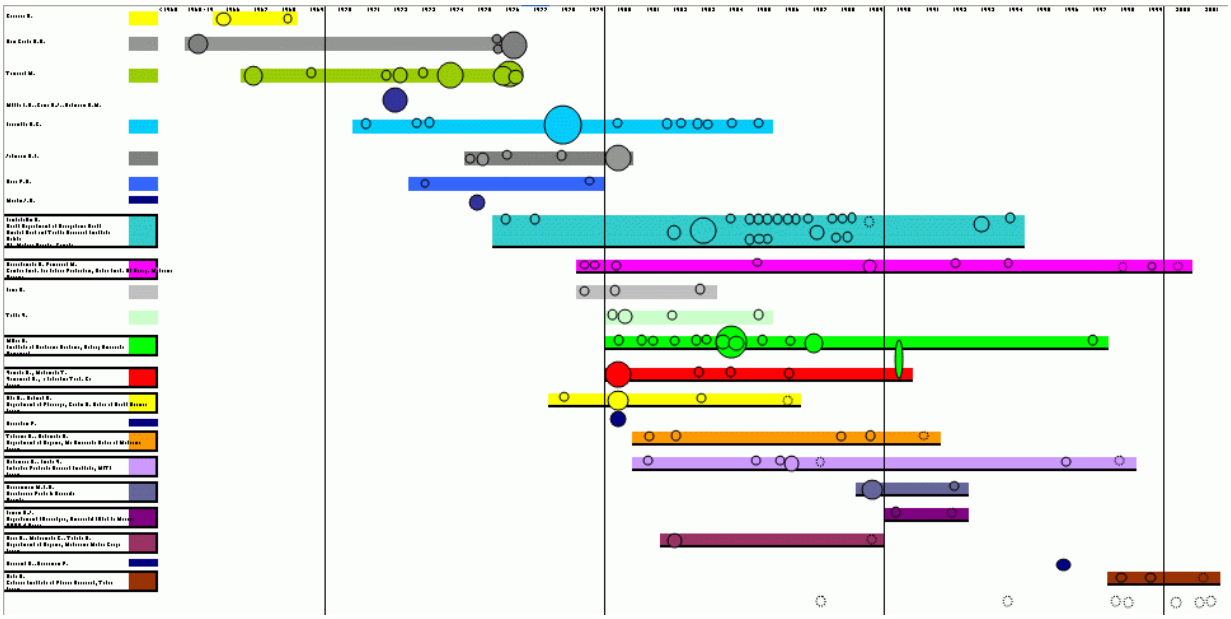
Ce qui veut dire aussi, pour ce qui nous concerne, que l'essentiel de la connaissance pertinente sur le sujet est à chercher dans les publications du milieu et de la fin des années 1980 : leurs conclusions pourront être interprétées rétrospectivement à la lumière de l'extinction prochaine de l'activité sur ce thème de recherche. Cette indication bibliométrique pure pourrait donc nous permettre de valider le contenu des articles concernés.

b) Cartographie des collèges :

Nous avons ensuite tenté de créer, à l'aide du tableur Excel, une cartographie susceptible de mettre en évidence la dynamique temporelle des travaux des chercheurs telle que les publications obtenues ou citées nous la révèle. Sur un axe horizontal, nous avons placé le temps. Les différentes équipes de recherche identifiées (les « collèges », éventuellement « invisibles »), repérées par une couleur particulière sur leur période d'activité apparente, ont été étagées de haut en bas selon les dates. Les publications ont été repérées sur ces barres par des disques de diamètre proportionnel au nombre de citations reçues. Celles que nous avons eues en main, sources de références, sont dessinées en gras ; lorsqu'elles n'étaient pas elles-mêmes citées par d'autres, le disque est en pointillé. Celles qui ne font pas partie d'un des « collèges », n'ont pas été représentées, sauf les sources de références regroupées en bas de la figure.



Détail de la cartographie : publications des collèges dans le temps, références citées une ou plusieurs fois, publications sources non cités.



Cartographie générale des publications.
Les lignes verticales correspondent aux années 1970, 1980, 1990 et 2000.

On met ainsi en évidence une vingtaine d'équipes actives sur le sujet entre le début des années 1960 et l'année 2001. D'après les publications émises, beaucoup d'équipes ont travaillé sur le sujet une dizaine d'années, voire le double, parfois de façon intensive pendant plusieurs années, parfois de façon plus sporadique. Par ailleurs, on observe que les articles les plus cités se localisent entre le début des années 1970 et le milieu des années 1980. Il est évident que les articles les plus récents ne peuvent recevoir beaucoup de citations ; mais, comme il y a naturellement un certain nombre d'autocitations dans les équipes présentes à partir de 1984 et dont nous avons eu les articles, cette répartition apparaît quelque peu déséquilibrée au détriment des publications récentes, pour laquelle la « reconnaissance par les pairs » semble plus limitée. Enfin, semble exister un seul cas de collaboration croisée, tardif pour les deux collèges considérés : T. Watanabe et H. Möller en 1990 (avec toutefois un risque d'homonymie sur ce patronyme japonais très répandu).

On peut considérer que, compte tenu du faible volume de publications sur le sujet, une telle configuration révèle que les différentes équipes de recherche ont abordé les infrasons comme un thème périphérique à leurs thématiques principales et de façon isolée les unes des autres (absence de publications communes), comme on l'avait déjà remarqué.

c) Analyse quantitative des citations :

Nous avons tenté d'avoir une idée globale de l'intensité et de l'actualité de la communication entre ces collèges à travers l'analyse des citations.

En effet, les échanges entre les différentes équipes de chercheurs paraissent très importants dans la construction des connaissances scientifiques, et les citations constituent un indicateur de ces échanges. Comme le cas étudié le montre clairement, une équipe de recherche fera bien entendu référence à ses propres travaux antérieurs pour étayer son argumentation ; elle pourra citer aussi les importants travaux « historiques » qui l'ont précédée. Mais l'on peut s'attendre aussi à ce qu'elle s'intéresse aux publications récentes des équipes « concurrentes », et qu'elle les cite, ne serait-ce que pour insister sur l'originalité de l'apport de son propre travail.

Pour chaque article, nous avons compté le nombre total de citations données, leur âge moyen et l'écart type correspondant, le nombre d'articles de moins de 10 ans et de moins de 5 ans. Nous avons ensuite écarté les autocitations (au sens des collèges précédemment définis) en comptant de même le nombre total de citations données aux chercheurs d'autres équipes, et dans celles-ci le nombre d'articles de moins de 10 ans et de moins de 5 ans.

L'âge moyen affiché et l'écart type correspondant ont été corrigés ; en effet certains auteurs citent des publications d'avant les années 1960 voire fort anciennes (1936), ce qui aurait indûment faussé les statistiques. C'est pourquoi les dates qui précèdent l'année 1960 ont été alignées cette date pour effectuer les calculs.

Article	Date	Nbre total cités	Age moyen corrigé	Ecart type	Nbre >= 10 ans	Nbre >= 5 ans	Nbre cités autres	Nbre cités autres >= 10 ans	Nbre cités autres >= 5 ans
1	1984	44	8,9	5,7	27	15	36	9	3
2	1986	2	4,0	2,8	2	1	0	0	0
3	1986	3	9,7	4,7	2	0	3	2	0
4	1987	11	7,7	6,0	8	4	8	5	1
5	1987	5	11,8	10,0	2	2	3	0	0
6	1987	13	2,5	1,8	13	12	9	9	8
7	1987	8	14,6	4,2	1	0	8	1	0
8	1987	0							
9	1989	21	8,6	8,6	15	10	11	5	1
10	1989	5	6,8	2,6	5	2	2	0	0
11	1989	8	9,8	5,0	5	1	8	5	1
12	1989	14	11,0	4,3	6	1	13	5	1
13	1991	14	13,4	9,1	7	4	10	3	2
14	1992	5	10,6	10,7	3	3	4	2	2
15	1993	12	11,9	4,9	6	0	7	2	0
16	1994	0							
17	1998	10	11,2	6,9	5	3	8	3	2
18	1998	3	6,7	5,0	2	1	1	1	0
19	1998	12	16,8	9,1	3	1	11	2	0
20	1998	4	9,8	7,2	2	2	4	2	2
21	2000	18	14,6	8,8	5	4	14	2	2
22	2000	3	3,3	2,5	3	2	2	2	1
23	2000	8	12,1	8,3	3	2	7	2	1
24	2001	4	11,5	10,4	2	2	2	0	0
25	2001	7	18,1	13,4	2	1	7	2	1
26	2001	6	11,3	4,5	2	1	6	2	1

Articles correspondant aux numéros indiqués dans le tableau :

1) Möller H.-1984
2) Yamada S., Watanabe T., Kosaka T., Negishi H., Watanabe H.-1986
3) Okai O.-1986
4) Landstrom U.-1987
5) Möller H.-1987
6) Inukai Y.,Taya H.,Nagamura N.,Kuriyama H.-1987
7) Tsunekawa S., Kajikawa Y., Nohara S., Azizumi M., Okada A.-1987
8) Densert D., Densert O.-1987
9) Moren B., Landstrom U., Nillson L., Sandberg U., Tornros J.-1989
10) Augustynska D.-1989
11) Vercammen M.L.S.-1989
12) Nagai N., Matsumoto M., Yamasumi Y., Shiraishi T., Nishimura K., Matsumoto K., Miyashita K., Takeda S., - 1989
13) Takigawa H., Sakamoto H., Murata M.-1991
14) Friman B.J., Ivannikov A.N., Zhukov A.N. -1992
15) Landstrom U., Pelmear P.L.-1993
16) Motylewski J., Zmierczak T., Nadolski W., Wasala T.-1994
17) Pawlaczyk M.-1998
18) Nakamura N., Inukai Y.-1998
19) Burt T.S.-1998
20) Lundin A.,Ahman M.-1998

21) Pawlaczyk M., Kaczmarska A., Augustynska D., Kamedula M.-2000
22) Rybak S.A., Rudenko O.V., Sobissevitch A.L., Sobissevitch L.Y., 2000
23) Sisto R., Lenzuni P., Pieroni A.-2000
24) Iwahashi K., Ochiai H.-2001
25) Crépon F.-2001
26) Jakobsen J.-2001

La lecture du tableau précédent permet de constater pour ces publications :

- la faiblesse en moyenne du nombre total de citations :
  - 2 ne citent aucun document
  - 17 sur 26 citent 5 documents ou plus
  - 10 sur 26 citent 10 documents ou plus
- la faiblesse (à une exception près : voir 6) du nombre de citations de documents récents d'autres équipes (le critère des 5 ans est un peu sévère à cet égard, mais avec 10 ans le résultat n'est pas très différent). Or, la présente étude montre justement qu'il y a eu, à toutes les époques, une certaine activité sur le sujet.

Il nous semble que les thématiques des travaux des différentes équipes de chercheurs étaient suffisamment similaires, sans bien sûr être identiques, pour permettre un taux de citations supérieur.

Le fait de ne pas citer le travail d'une autre équipe ne signifie évidemment pas qu'on n'en ait pas eu connaissance, ni qu'on ne s'y soit pas intéressé de près. Le doute existe néanmoins dès lors qu'on ne fait pas référence de façon explicite à d'autres publications, en particulier parce que nous avons vu que certaines des publications étaient peu accessibles. D'ailleurs, peu d'auteurs (voir les articles 1 et 15) ont tenté un effort de synthèse.

Il nous semblerait donc, au vu de la présente analyse, que les échanges entre les différents chercheurs ont été assez réduits ; cette étanchéité ne peut pas favoriser l'élucidation des expériences contradictoires générées au fil du temps.

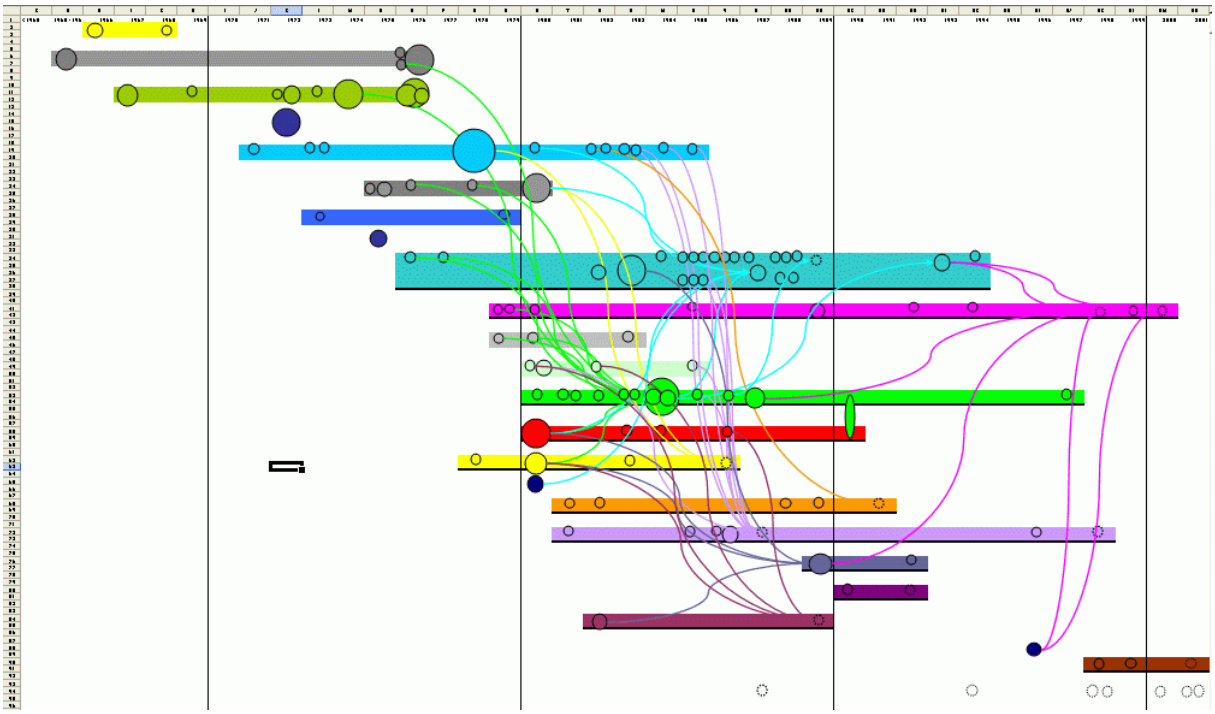
Pour ce qui concerne l'aide que la bibliométrie peut nous apporter pour tenter d'élucider la question scientifique sur le fond, on peut considérer que le crédit à accorder aux publications d'une équipe n'est pas seulement une fonction croissante du nombre d'articles qui les citent (ce qui est bien connu). Il dépend aussi du nombre d'articles récents des différentes autres équipes engagées sur le même sujet que les auteurs ont pris la peine de citer, ce qui est la preuve (probable faute d'être absolue, mais nous n'en avons pas d'autre) qu'ils ont été recherchés et examinés. En dehors bien entendu de la valeur intrinsèque des travaux réalisés, on peut supposer en effet que ces chercheurs ont été alors mieux en mesure de prendre en compte les différentes facettes d'un sujet controversé.

d) Analyse qualitative des citations :

Pour évaluer de façon plus détaillée les relations entre les différents « collèges », nous avons tenté de compléter la cartographie précédente en matérialisant les liaisons entre chaque publication et ses références. L'enchevêtrement des liens qui en a résulté a rendu la lecture de la carte difficile.

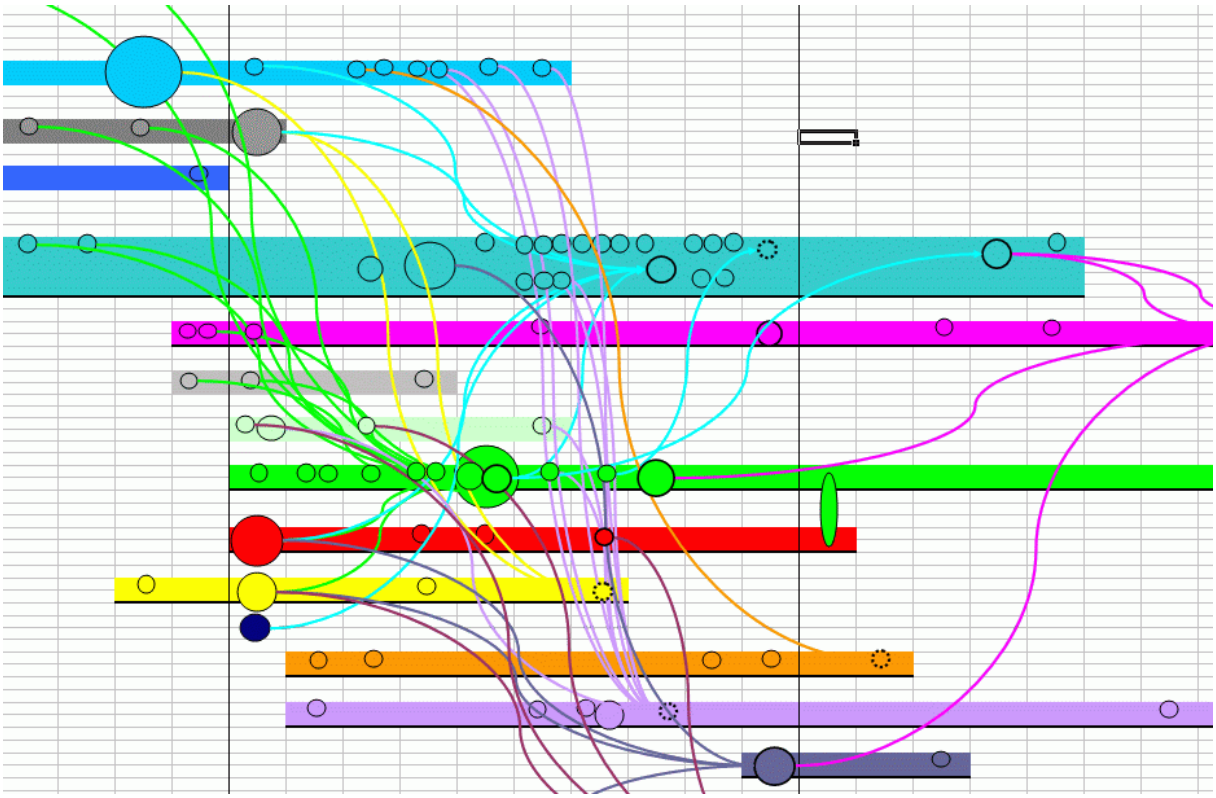
Après plusieurs tentatives, nous nous sommes limité aux liens jugés les plus cruciaux pour avoir une idée de l'activité des échanges entre collèges ; seules les citations de publications d'autres équipes datant de 10 ans ou moins ont été représentées. Nous avons considéré que les articles plus anciens appartenaient au « fonds culturel » de la Science et que le fait de les citer n'apportait plus guère d'information pertinente.

Les collèges dont nous avons eu à disposition les publications sont les seuls dont on puisse évidemment connaître les liaisons avec d'autres : ils sont repérés à gauche par un cadre noir, et la barre correspondante est également soulignée d'un trait noir.



Cartographie générale des publications avec les citations récentes.
Les lignes verticales correspondent aux années 1970, 1980, 1990 et 2000.

Sur ce schéma, les silences dans les citations de certaines écoles, les dissymétries de citations entre deux écoles sont des indices de singularités qu'on peut ainsi repérer et qu'on pourra éventuellement expliquer en se référant aux articles eux-mêmes – la carte, comme on sait, n'étant pas le paysage.



Détail de la cartographie pour les années 1980 à 1995

On peut ainsi effectuer une typologie des collèges matérialisés sur la carte :

- ceux dont les citations récentes d'autres équipes sont rares ou absentes,
- ceux dont les citations récentes concernent uniquement une ou deux autres équipes,
- ceux qui ont tenté un effort de synthèse en citant différents autres collèges.

Ce sont bien ces derniers qui nous permettent par la même occasion de vérifier la proximité thématique des différentes équipes de recherche du tableau : il s'agit bien de la même question scientifique. Partant, on peut reconnaître comme significatif le silence d'autres publications.

La similarité des citations de beaucoup de collèges laisse aussi bien supposer la proximité de leur thème d'étude ; de ce fait, quand la chronologie le permet, on pourrait s'attendre à une certain niveau de citations croisées entre eux : or ce n'est pas toujours le cas. Dans une trame de citations qui devrait être assez dense et homogène, on remarque de nombreux accrocs.

L'impression générale laissée par cette cartographie est qu'un certain niveau de synthèse a été atteint aux alentours de l'année 1975 et que, depuis, les problématiques n'ont cessé de diverger.

Même si beaucoup d'équipes ont pu être repérées, il faut insister à nouveau sur la caractère incomplet de cette cartographie, tout particulièrement l'absence presque totale des chercheurs russes.

5 - ANALYSE DES ITINERAIRES DES CHERCHEURS :

La cartographie précédente nous a permis de mettre en évidence la dynamique de la recherche dans le domaine des infrasons. Mais les investigations menées au cours de leur carrière par les différentes personnes identifiées dans cette étude ne se sont pas limitées à la thématique des infrasons. Pour compléter, il nous a paru intéressant d'essayer de répondre à cette question : qu'ont traité ces chercheurs en dehors des infrasons ?

Ce travail a été effectué sur Medline. Les organismes n'étant pas précisés, nous avons suivi le fil des auteurs, représentés dans la base sous la forme nom + initiale du prénom, en examinant les listes de références obtenues et certains résumés. Ce travail aurait pu être également effectué par un programme bibliométrique en croisant les dates, les auteurs et les mots du titre (à défaut des mots-clés) comme nous l'avons fait pour Pascal.

a) L'école russe :

Nous nous sommes intéressé en particulier aux auteurs russes, particulièrement référencés dans la base Medline comme on l'a vu précédemment. Beaucoup d'entre eux, parmi bien d'autres activités, ont abordé la thématique des infrasons à un moment de leur carrière. Beaucoup de ces chercheurs, selon Medline, ont publié au moins une fois ensemble. Nous allons faire ressortir les tendances principales.

Evgeniia Tsezarevna Andreeva-Galanina a défini en 1956 une classification des pathologies de la main et du bras produites par les vibrations qui a servi de référence mondiale en la matière. Elle a par la suite continué de travailler sur les effets des vibrations et du bruit sur l'organisme, en particulier au niveau du système nerveux. Cette personne semble avoir été une figure particulièrement charismatique compte tenu de la présence de plusieurs articles écrits à son hommage (sa dernière publication présente dans Medline date de 1973 : elle avait 85 ans…) ; elle est très probablement à l'origine de l'intérêt des chercheurs russes sur le sujet.

Certains de ses collaborateurs ont poursuivi ses travaux : citons S.V. Alekseev et G.A. Suvorov qui sont des auteurs centraux (et très actifs) dans le réseaux des chercheurs qui ont abordé le thème des infrasons. Le premier est resté sur cette problématique ; le second, après avoir travaillé sur la pathologie des bruits et vibrations, s'est consacré à des approches plus générales et variées sur les facteurs physiques qui interviennent dans les conditions de travail.

Proche de S.V. Alekseev, V.I. Svidovyi, qui s'est intéressé à d'autres facteurs concernant l'hygiène du travail, a été également très actif sur les infrasons : en particulier, il a souvent formé équipe avec V.V. Glinchikov, et A.S. Nekhoroshev pour étudier systématiquement les effets des infrasons, des bruits de basse fréquence et des vibrations sur le cœur, le foie, le système nerveux, les muscles.

V.N. Alekseev a beaucoup travaillé sur la fonction vestibulaire en particulier dans le contexte des vols spatiaux jusqu'au début des années 1990. (Il semble avoir un homonyme exact, ophtalmologue, qui a publié en 2001 un article sur les effets du bruit et des infrasons sur les organes de vision).

L'itinéraire de V.G. Ovakimov, 9 fois co-auteur avec G.A. Suvorov, est très intéressant. Il s'est spécialisé dans la résistance aux radiations (sans aucun doute dans un contexte spatial) entre 1965 et 1977 avant de se tourner vers les problèmes de bruit et vibrations en usine ou dans les hélicoptères, puis à partir de 1990, sans abandonner ce sujet, vers l'hygiène du travail et les maladies professionnelles d'une façon plus générale.

En résumé, le volume de littérature sur les infrasons et sa persistance s'explique par une spécialisation de longue date des équipes soviétiques sur l'hygiène industrielle (encouragée pour des raisons idéologiques probablement), en particulier les pathologies relatives aux phénomènes vibratoires propres à l'industrie métallurgique lourde et semi-lourde ; semble avoir amplifié cette tendance la reconversion, sur des problèmes plus terre à terre si l'on ose dire, de chercheurs engagés dans des expérimentations dans le cadre des ambitieux programmes spatiaux soviétiques, pour lesquels, comme aux Etats-Unis, ont été menées des investigations sur la résistance de l'organisme à des conditions particulières.

D'une façon générale, cette littérature semble riche et variée, avec un degré de connexité élevé du réseau des chercheurs ; l'on peut regretter sa méconnaissance en Occident que nous a révélée notre cartographie à partir de Pascal.

b) Les écoles scandinaves :

Nous avons également essayé de reconstituer l'itinéraire thématique d'Ulf Landström. On trouve, entre 1975 et 1979, 9 publications sur l'embryologie : il est possible qu'il s'agisse d'un homonyme. Puis à partir de 1982, apparaît la problématique de l'effet des vibrations sur la peau. A partir de 1991, le chercheur s'oriente vers les conditions de maintien de l'attention : influence du bruit et des vibrations, de la température (1999) puis des prises de nourriture (2000 à 2002).

Ce que semble suggérer cet itinéraire, c'est qu'après avoir travaillé autour des risques de somnolence suscités par les infrasons (en particulier lors de la conduite routière), ce chercheur a poursuivi cette problématique en intégrant d'autres facteurs. Ce glissement pourrait être significatif de la place des infrasons comme un facteur parmi d'autres – nullement prépondérant – de ce phénomène. Telle serait peut être la vérité que nous recherchions.

Nous avons tenté le même traitement dans Medline avec « Moller H » (dont on a vu l'absence dans le corpus relatif aux infrasons). Malheureusement (compte tenu de l'ambiguïté liée à l'« o » barré danois), on trouve beaucoup d'homonymes partiels (« Moller HB, Moller HJ, Moller HE ») et même au moins quatre homonymes exacts (un Allemand spécialiste de psychiatrie, un Danois du Centre de recherche de statistiques socio-médicales de Copenhage, un Américain de l'Ecole de médecine de Los Angeles, un Suédois odontologue de l'Université de Malmö). Comme Henrik Möller appartient à l'université d'Aalborg, « moller h AND aalborg » ne nous a donné dans Medline qu'un seul article de 1996 :

Hammershoi D, Moller H. , Sound transmission to and within the human ear canal.
J Acoust Soc Am. 1996 Jul;100(1):408-27.

Dans Pascal, la même interrogation (AU = « MOLLER H » AND AALBORG/CS) nous a donné 15 articles (avec un homonyme épidémiologiste !) ; à partir de 1990, l'auteur se consacre aux mécanismes physiques de l'audition humaine (perception de la stéréophonie, fonction de transfert acoustique de la tête humaine, enregistrement par tête artificielle, etc…).

Nous vérifions ainsi pour ces deux chercheurs, parmi les plus actifs dans la recherche sur notre thème, l'abandon des investigations purement centrées sur les infrasons au début des années 1990 que l'analyse des références citées nous avait déjà permis de constater.

La démarche qui consiste, à travers les mots-clés voire le contenu des titres des articles, à retracer l'évolution thématique des chercheurs peut donc être riche d'enseignement : elle pourrait constituer une vérification et une validation des résultats de leurs travaux antérieurs. Cette approche peut être compliquée par des problèmes d'homonymie, malgré les initiales des prénoms, comme nous l'avons constaté sur les noms scandinaves, russes et japonais – il en serait de même pour les patronymes chinois. Pour lever certaines ambiguïtés, il faudrait,

lorsqu'on a de telles informations à disposition (car toutes les bases ne les indiquent pas systématiquement et clairement pour chacun des auteurs de la publication), prendre en compte les organismes émetteurs : mais il faut en utiliser le libellé exact dans la base. Comme on l'a vu, le nom de la localité n'est pas très discriminant.

Ainsi, nous avons tenté de trouver dans Medline des références de deux chercheurs japonais d'un « collège » fréquemment cité par les publications issues de Pascal. Peu de chercheurs sont référencés plus d'une cinquantaine de fois dans Medline. Or, pour S. Yamada, on trouve (le 18 septembre 2002) 2344 articles, 5540 pour T. Watanabe, 82 pour « watanabe t. AND fukushima » ; il s'agit d'un homonyme probable qui travaille à la Faculté de Médecine de Fukushima, et non au Collège Technique. (Rajouter « college » n'est d'ailleurs pas discriminant et rajouter « technical » ne donne rien.)

On conçoit donc la nécessité dans de tels cas d'une méthodologie rigoureuse dans l'identification et l'utilisation du nom de l'organisme pour ne pas faire d'erreurs … en espérant que l'auteur soit assez sédentaire pour pouvoir embrasser son activité pendant quelques années !

CONCLUSION

L'investigation menée sur la question des ultrasons nous a permis de mettre en évidence quelques aléas de la construction de la vérité scientifique, et la difficulté de prendre du recul sur des sujets polémiques comme celui-ci.

La quantification du volume de publications sur le thème par rapport à d'autres similaires nous a permis d'abord d'en apprécier l'importance. L'évolution du volume des publications dans le temps et sa répartition en termes géographiques ont permis de caractériser la création de connaissance liée aux différents travaux menés. L'identification des « collèges », la visualisation de leur activité dans le temps, la mise en évidence de leurs publications-clés et la matérialisation du réseau des citations vers les autres équipes a permis de constituer une cartographie significative de l'ensemble de la thématique. Enfin, l'analyse des itinéraires thématiques des chercheurs est une voie particulièrement intéressante (mais qui comporte un risque du fait des homonymies) car elle traduit les avancées du front de la recherche et pourrait être hautement significative des découvertes réalisées.

Soulignons que la dynamique globale des publications que nous avons constatée à propos des infrasons est très particulière à ce domaine : il serait intéressant de mener des analyses identiques dans d'autres domaines de recherche.

S'il fallait prendre le risque de conclure sur le fond du problème, nous serions tenté de dire que les infrasons en eux-mêmes peuvent être certes une source d'inconfort à des niveaux élevés, mais que pour le reste des situations, ils ne constituent au mieux pour l'organisme, parmi ceux susceptibles de l'affecter, qu'un facteur secondaire, et difficilement dissociable de beaucoup d'autres (bruit dans les fréquences audibles, vibrations, conditions de température, qualité de l'air …). On admet d'ailleurs depuis quelques années que la notion de confort d'un espace habité est multifactoriel – et subjectif. Quand à supposer une sensibilisation progressive aux infrasons par suite d'une exposition prolongée, l'hypothèse souvent évoquée en est purement conjecturale : il ne semble pas exister à ce jour d'étude susceptible de la fonder scientifiquement.

Reste à comprendre pourquoi les conclusions, quelles qu'elles soient, apparaissent avec aussi peu de netteté, malgré une vingtaine d'années de travaux qui représentent un volume non négligeable.

Nous émettons l'hypothèse que souvent, comme nous l'avons constaté, la fixation des chercheurs sur les publications les plus anciennes (celles des années 1970) au détriment du travail plus récent des équipes contemporaines, jointe à un contexte mal défini, a pu contribuer à propager certaines erreurs initiales, dont, si l'on peut dire, la Science s'est difficilement remise.

Il nous semble aussi que la marginalité du sujet n'a pas permis à l'activité scientifique d'atteindre une visibilité et une taille critique qui auraient permis, à travers des colloques spécialisés en particulier, une synergie des travaux des chercheurs et une approche scientifique globale plus efficace.

La bibliométrie ne nous a pas donné la clef de la vérité scientifique sur les infrasons : pour cela, rien ne peut remplacer la lecture systématique, soigneuse et experte des articles eux-mêmes, heureusement ! Mais elle a l'avantage de remettre en perspective la contribution de chaque auteur et de chaque équipe ; elle permet aussi

d'interpréter les reconversions thématiques et les silences (car l'on ne publie plus sur un sujet quand on n'y trouve plus rien digne d'intérêt, et le retrait des chercheurs se fait sans commentaires), et, en prenant certaines précautions, de valider rétrospectivement les discours scientifiques.

REMERCIEMENTS

L'auteur remercie la société ALSTOM TRANSPORT (établissement d'Aytré) grâce à laquelle cette étude a pu être réalisée.

BIBLIOGRAPHIE :

Sur les infrasons, articles de Pascal utilisés pour la cartographie :

Möller Henrik, *Physiological and Psychological Effects of Infrasound on Humans*, Journal of Low Frequency Noise and Vibration, Vol. 3 No. 1, 1984, pp 1-17.

Yamada Shinji, **Watanabe T**oshio, **Kosaka T**oshifumi, **Negishi H**iromichi, **Watanabe I**deo, *Physiological Effects of Low Frequency Noise*, Journal of Low Frequency Noise and Vibration, Vol. 5 No. 1, 1986, pp 14-24.

Okai Osamu, *Effects of Infrasound on Respiratory Function of Man*, Journal of Low Frequency Noise and Vibration, Vol. 5 No. 3, 1986, pp 94-99.

Möller Henrik, *Annoyance of Audible Infrasound*, Journal of Low Frequency Noise and Vibration, Vol. 6 No. 1, 1987, pp 1-17.

Landström Ulf, *Laboratory and Field Studies on Infrasound and its Effects on Humans*, Journal of Low Frequency Noise and Vibration, Vol. 6 No. 1, 1987, pp 29-33.

Densert Barbara, **Densert O**ve, *Infrasonic Energy Transmission to the Inner Ear*, Journal of Low Frequency Noise and Vibration, Vol. 6 No. 2, 1987, pp 74-75.

Inukai Yukio, **Taya H**ideto, **Nagamura N**eiichi, **Kuriyama H**iroshi, *An evaluation Method of Combined Effects of Infrasound and Audible Noise*, Journal of Low Frequency Noise and Vibration, Vol. 6 No. 3, 1987, pp 119-125.

Tsunekawa S., **Kajikawa Y**., **Nohara S**., **Ariizumi M**. **Okada A**., *Study on the perceptible level for infrasound*, Journal of Sound and Vibration, 112(1), 1987, pp 15-22.

Vercammen M.L.S., *Setting Limits for Low Frequency Noise*, Journal of Low Frequency Noise and Vibration, Vol. 8 No. 4, 1989, pp 105-109.

Morén Bertil, **Landström U**lf, **Nilsson L**ena, **Sandberg U**lf, **Törnros J**an, *The influence of noise, infrasound and temperature on driver performance and wakefulness – A driving simulator study* (titre original en Suédois), VTIrapport 340, 1989, ISSN 03047-6030.

Augustynska Danuta, *Infrasonic Noise Emitted by Flow Machines, Its Sources and Reduction Methods*, Journal of Low Frequency Noise and Vibration, Vol. 8 No. 1, 1989, pp 9-15.

Nagai Naoko, **Matsumoto M**asanobu, **Yamasumi Y**asukiyo, **Shiraishi T**atsue, **Nishimura K**oh, **Matsumoto K**enji, **Miyashita K**azuhisa, **Takeda S**intaro, *Process and Emergence on the Effets on Infrasonic and Low Frequency Noise on Inhabitants*, Journal of Low Frequency Noise and Vibration, Vol. 8 No. 3, 1989, pp 87-99.

Takigawa H., **Sakamoto H**., **Murata M**., *Effects of infrasound on vestibular function*, Journal of Sound and Vibration, 151(3), 1991, pp 455-460.

Friman B.J., **Ivannikov A.N.**, **Zhukov A.N.**, *On the Influence of Infranoise Fields on Humans*, Journal of Low Frequency Noise and Vibration, Vol. 11 No. 4, 1992, pp 105-108.

Landström U., **Pelmear P.L.**, *Infrasound – A Short Review*, Journal of Low Frequency Noise and Vibration, Vol. 12 No. 3, 1993, pp 72-74.

Motylewski Jerzy, **Zmierczak T**omasz, **Nadolski W**ladyslaw, **Wasala T**adeusz, *Infrasound in Residential Area – A Case Study*, Journal of Low Frequency Noise and Vibration, Vol. 13 No. 2, 1994, pp 65-70.

Burt T.S., *Building acoustics and sick building syndrome*, EPIC 1998, 2ème conférence européenne, Lyon, 19-21 novembre 1998, pp 856-861.

Lundin Anders, **Ahman M**ats, *Case report : Is Low-Frequency Noise from Refrigerators in a Multi-Family House a Cause of Diffuse Disorders ?*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 17 No. 2, 1998, pp 65 -70.

Pawlaczyk-Luszczynska Malgorzata, *Occupational Exposure to Infrasonic Noise in Poland*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 17 No. 2, 1998, pp 71-83.

Nakamura Norio, **Inukai Y**ukio, *Proposal of Models which Indicate Unpleasantness of Low Frequency Noise using Exploratory Factor Analysis and Structural Covariance Analysis*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 17 No. 3, 1998, pp 127-131.

Pawlaczyk-Luszczynska Malgorzata, **Kaczmarska-Kozlowska A**nna, **Augustynska D**anuta, **Kamedula M**aria, *Proposal of New Limit Value for Occupational Exposure to Infrasonic Noise in Poland*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 19 No. 4, 2000, pp 183-193.

Sisto Renate, **Lenzuni P**aolo, **Pieroni A**ldo, *High Amplitude Infrasound in Railway Tunnels*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 19 No. 2, 2000, pp 83-92.

Rybak Samuil **A.**, **Rudenko O**leg **V.**, **Sobissevitch A**lexei **L.**, **Sobissevitch L**eonid **Ye.**, *Geo-ecological infrasound monitoring of highways and surrounding areas*, Acoustics Letters, Vol. 23, No. 10, 2000, pp 197-200.

Iwahashi Kiyokatsu, **Ochiai H**iroaki, *Infrasound Pressure Meter and Examples of Measuring Data*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 20 No. 1, 2000, pp 15-19.

Jakobsen Jorgen, *Danish guidelines on environmental low frequency noise, infrasound and vibration*, Journal of Low Frequency Noise, Vibration and Active Control, Vol. 20 No. 3, 2001, pp 141-148.

Crépon Francis, *Actualité en électrophysiothérapie - Infrasons, désencombrement bronchique et rééducation fonctionnelle*, Kinésithérapie scientifique, n° 410, Avril 2001, pp 55-56.

Sur les infrasons, tirés de l'Internet :

Haneke Karen, **C**arson **B**onnie, **Gregorio C**laudine, **Maull E**lizabeth, *Infrasound - Brief Review of Toxicological Literature*, Novembre 2001, http://ntp-server.niehs.nih.gov/htdocs/Chem_Background/ ExSumPdf/Infrasound.pdf, consulté le 30 août 2002.

Altmann Jürgen, *Acoustic Weapons – A Prospective Assessment*, Science & Global Security, 2001, Volume 9 pp 165-234, http://www.princeton.edu/~globsec/pdf/9_3altmann.pdf, consulté le 8 mars 2002.

Autres références :

[Watzlawick, 1988] **Watzlawick P**aul (dir.), L'Invention de la réalité – Comment savons-nous ce que nous croyons savoir ? Contributions au constructivisme, Editions du Seuil, 1988, 378 p.

[Baldit , 1994] **Baldit P**atrick, La sériation des similarités spécifiques : un outil pour la recherche de l'information stratégique, 1994, Thèse de doctorat, Faculté des Sciences et Techniques de Saint-Jérôme, Université d'Aix-Marseille III.

[Rostaing, 1996] **Rostaing H**ervé, La bibliométrie et ses techniques, Co-édition Sciences de la Société – CRRM, 131 p, 1996.

Dkaki Taoufiq, Une Méthode pour la détection et l'analyse des réseaux de collaborations dans le domaine de la recherche scientifique, V.S.S.T.'95, Toulouse, Octobre 1995, pp 143-152.

Dkaki Dkaki Taoufiq, **Dousset B**ernard, **Mothe J**osiane, Analyse d'informations issues du Web avec Tétralogie, V.S.S.T.'98, Toulouse, Octobre 1998, pp 159-170.

Marcon Christian, **Moinet N**icolas, La Stratégie-réseau, Editions 00h00.com, 2000, 235 p.

OUTILS ET MODELES DE TRAVAIL COLLABORATIF

Eric GIRAUD
giraud@unimeca.univ-mrs.fr

Jean-Francis RANUCCI

Adresse professionnelle
IUFM Site Château-Gombert
60, rue Joliot Curie
13453 Marseille CEDEX 13

Résumé : Formation à distance, knowledge management, institution apprenante, gestion de l'information, sont autant de concepts qui sollicitent de manière croissante l'utilisation de techniques de communication informatisées. L'instrumentalisation de l'outil informatique ne peut, en aucun cas, être envisagée comme une fin en soi, et, ne peut se substituer à une réelle démarche de fond dans la création d'outils opérationnels. Si la mise en place de systèmes de travail collaboratif semble constituer une approche pertinente dans le cadre de la création de groupes humains autour d'un projet commun, il convient au préalable de conduire une étude approfondie des acteurs, des objectifs ainsi que des usages, afin d'obtenir un cahier des charges répondant aux besoins des utilisateurs. Le précédent constat pourrait laisser penser qu'il n'y a pas de système d'information universel et que chaque problématique doit s'assortir d'une solution individualisée. Si d'emblée, l'idée d'une «solution universelle» semble être illusoire, nous pouvons toutefois proposer une réflexion globale tentant de s'orienter vers une famille de solutions techniques invariantes, véritables conditions nécessaires, mais non suffisantes dans le cadre du déploiement d'un système de travail collaboratif. La présente communication décrit la démarche d'analyse et de conception d'une plate-forme de travail collaboratif à distance, les auteurs présentent une synthèse des fonctions de base, nécessaires à la création de communautés virtuelles productives. Ces réflexions s'articuleront autour de l'étude d'un système collaboratif expérimental mis en place à l'IUFM d'Aix Marseille dans le cadre d'un programme de recherche universitaire et utilisé pour l'organisation d'actions de formation.

Abstract : Distance learning, knowledge management, smart institution, information management, are so many concepts which seek in an increasing way the use of computerized techniques of communication. Instrumentalisation of computerized tool can not be envisaged as an end in itself, and, can not substitute itself for a real thorough step in the creation of operational tools. If the implementation of collaborative working systems seems to establish a relevant approach within the framework of the creation of human groups around a common

project, it agrees beforehand to lead a detailed study of the actors, the objectives as well as the manners, to obtain a conditions of contract meeting the needs of users. The previous report could let think that there is no universal information system and that every problem should match of an individualized solution. If at once, the idea of an "universal solution" seems not to be realistic, we can however propose a global reflection trying to turn to a family of invariant technical solutions, real necessary, but not sufficient conditions within the framework of the display of a collaborative working system. The present communication describes the method of analysis and conception of a distance collaborative working platform. Authors present a synthesis of the basic functions, necessary for the creation of productive virtual communities. These reflections will articulate around the study of a collaborative experimental system set up in AIX-MARSEILLE'S IUFM within the framework of a research program university and used for the organization of actions of forming people.

Mots Clés : Travail collaboratif, systèmes d’information, communautés virtuelles, services web, gestion des connaissanes, knowledge management.

Outils et modèles de travail collaboratif

1. PREAMBULE

Les structures dans lesquelles nous évoluons mettent en place de plus en plus de réunions, de groupe de travail, de groupe de recherche et de développement auxquels il convient de participer si l'on désire maintenir son niveau d'information et garantir une marge décisionnelle. Ces activités viennent s'ajouter au travail quotidien, de sorte que nos plannings sont remplis et qu'il devient impossible de répondre aux demandes urgentes de dernières minutes sans devoir annuler un rendez-vous ou une réunion.

Travailler autrement, à distance, devrait permettre de regrouper à tous moments, les connaissances professionnelles nécessaires pour réagir. Le travail collaboratif «asynchrone» devrait faciliter, avec le support d'une technologie appropriée, la diffusion numérique immédiate des informations à un groupe professionnel donné, éloignés ou non, en raccourcissant les délais de recherche de l'information.

Nous faisons l'hypothèse que les T.I.C et particulièrement les systèmes autorisant le travail collaboratif devraient permettre des échanges d'informations, des résolutions de problèmes et des prises de décision, sans pour cela monopoliser tout notre temps. Dans ce cadre, nous présentons ici les résultats d'un travail de recherche portant sur l'élaboration d'une plateforme de travail collaboratif à distance dans le cadre de la formation continue des enseignants (Ranucci, 2001).

Formuler l'hypothèse qu'un outil peut se révéler utile dans des actions de travail collaboratif revient à adhérer à une des théories de G. Simondon (Simondon, 1989) selon laquelle un objet technique n'existe que s'il modifie des organisations sociales et si les pratiques sociales associées à l'existence de l'outil viennent influer sur les critères d'élaboration de le l'objet technique lui même. L'approche systémique, culturelle et réseau de D. Desjeux (Desjeux, 1994), qui analyse l'appropriation d'un objet technique ou d'une innovation par des groupes sociaux au travers de différents filtres (sociaux, culturels, ….), conforte notre hypothèse.

2. CONTEXTE

Nous présenterons ici les résultats d'une expérimentation conduite dans le cadre de l'aide à l'ingénierie pédagogique lors d'actions de formation continue.

La multidisciplinarité, l'éloignement du site avec les différentes structures gestionnaires, le recours à des personnels très diversifiés, en termes de spécialités, de cultures et de localisations, sont autant d'éléments qui contribuent à faire de notre structure un site idéal pour l'expérimentation en grandeur réelle de la Gestion des Connaissances et du Savoir (GCS), La plateforme expérimentale de travail collaboratif présentée ici assurera la fonction de médiation du concept de Gestion des connaissances (ou Knowledge Management).

Selon Jordan, l'objectif initial du Knowledge Management (KM) réside dans l'intégration, voire l'anticipation aux brusques changements technologiques en optimisant l'organisation et en modifiant les rapports humains dans l'entreprise. La gestion des connaissances, définie quelquefois comme une méta-compétence, est sensée créer et entretenir l'avantage compétitif des industries en redéfinissant les rapports de pouvoir, les statuts et les protocoles de communication au sein de l'entreprise. (Jordan, 1997)

Dans ce sens, il paraît clair que la mise en place d'un système de gestion des connaissances dans un établissement à vocation de recherche et d'enseignement peut apporter de grands services.

3. LES FLUX D'INFORMATION

Une autre hypothèse de travail que nous formulons se fonde sur les travaux de Shannon et Viterbi (Shannon, 1949) (Viterbi, 1979). Bien que fortement ancrée sur des concepts mathématiques, les théories de Shannon et de Viterbi sur la communication exposent des notions facilement vulgarisables :

- pour être perçue, une information doit s'élever au dessus du niveau moyen d'information de l'environnement (notion de rapport signal/bruit)

- et, corrélativement, il n'y a pas d'information lorsque le niveau moyen d'information est stationnaire (pas d'événements)

Un autre argument en faveur de notre hypothèse est apporté par une affirmation de G. Bachelard qui suppose qu'il n'y a réellement de création de connaissance (ou d'intelligence, ou de savoir …) que lors de ruptures épistémologiques. (Bachelard, 1968)

La création de valeur ajoutée sera ainsi indissociable du concept d'événement que nous introduisons dans la structure de la plateforme de travail collaboratif.

Nous analysons ici les flux d'information dont la manipulation est inévitable dans le contexte de l'étude.
Ces données ont fait l'objet d'une classification préalable selon un modèle développée récemment. (Giraud, 1997)

Les auteurs de ce modèle proposent une classification de l'information tenant compte de 5 paramètres :

La source
La destination
Le flux
La rigidité
Le support

Classes, auxquelles on adjoint le concept de ***système***, le système étant ici constitué des personnels impliqués dans l'action de formation (enseignants, stagiaires et personnel administratif), la matière d'œuvre reste, bien entendu, l'information.

Selon cette classification, nous recensons cinq principaux types de flux qu'il faudra manipuler :

- L'Information Purement Entrante (IPE)
- L'Information Purement Sortante (IPS)
- L'Information Interne (II)
- L'Information Structurelle Rigide (ISR)
- L'Information Structurelle Modifiable (ISM)

3.1. Information Purement Entrante (IPE)

L'information purement entrante trouve sa source à l'extérieur de la structure, le site l'absorbe en se l'appropriant. On trouvera ici des informations diverses telles que des données sur les fournisseurs, des catalogues, des contenus de prospectus, des appels à communication, des annonces de manifestations, des colloques, des appels d'offre de recherche ….

Dans le cadre d'activités de Gestion des Connaissances et du Savoir, chaque acteur doit pouvoir intervenir sur les flux d'IPE.

3.2. Information Purement Sortante (IPS)

L'information purement sortante est élaborée par la structure, elle est publique et concerne les opérations de communication externe, telles que les informations sur les différentes structures d'enseignement et de recherche, les programmes de cours, les emplois du temps, ainsi que les activités scientifiques courantes.

L'IPS a une vocation de médiatisation et de publicité, on conçoit tout à fait qu'une structure telle qu'UNIMECA éprouve le besoin de communiquer aussi bien à destination du grand public qu'aux personnels délocalisés.

3.3. Information Interne (II)

Ceci représente un ensemble d'informations générées par la structure, elles sont destinées à une diffusion interne unique.

L'II est la seule qui peut véhiculer des données confidentielles telles que les tâches de gestion de personnel, la diffusion des orientations stratégiques, des idées de développement de produits ou de services peuvent également en faire partie, ainsi que les différentes données sur les fournisseurs, les contrats de recherche, les notes de services.

Le terme **interne** est ici relatif à des groupes d'individus, organisés par missions ou par responsabilités et non par localisation géographique sur le site.
Ceci signifie que l'information qualifiée d'interne ne concerne pas forcément les membres du site géographique, mais aussi que l'on doit prévoir un moyen d'externaliser ces flux, et ce, de manière sécurisée.

3.4. Information Structurelle Rigide (ISR)

L'information structurelle rigide peut être élaborée en externe ou en interne, l'ISR d'origine externe peut être par exemple des textes de loi, des décrets, des règlements administratifs.

L'ISR interne est constituée par les différents règlements internes et décisions d'organisation, d'administration ou de gestion.

On conçoit très bien que ce genre d'information ne peut pas subir de modification, de par l'aspect réglementaire des contenus.

Aucune modification, dans le fond ou dans la forme, aucune analyse ni interprétation ne peuvent être opérées sur de l'ISR.

3.5. Information Structurelle Modifiable (ISM)

L'Information structurelle modifiable est fondamentalement complexe, dans le sens ou elle se fonde sur une valorisation des flux de type II et/ou IPE.

Une mise en relation de plusieurs expériences vis à vis d'un fournisseur, l'analyse et la synthèse d'appels d'offre de recherche, le partage d'expériences et de compétences, peuvent être classifiés sous le terme d'ISM.

On notera que la plus value ainsi réalisée est très souvent destinée à des fins de décision ou d'orientation stratégique. Elle constitue une part très importante du système de Gestion des Connaissances et du Savoir.

En général les flux d'ISM sont très véloces, très denses et possèdent une durée de vie très courte. Cette classe d'information est, sans aucun doute la plus délicate à manipuler, car elle doit solliciter un maximum d'acteurs destinés à élaborer la plus value, sans pour autant être « trop » diffusée pour des raisons de confidentialité ou, encore pour éviter d'être noyé sous le « bruit » masquant la valorisation apportée par certains intervenants.

4. LES METHODES DE GESTION DE FLUX

4.1. Critères d'organisation

Dans le cas de flux d'informations élaborés en interne, ce qui concerne les classes II, ISR et ISM, précédemment exposées, l'accent doit être mis sur des solutions techniques facilitant le travail collaboratif. On se contentera ici de rappeler les travaux de Mintzberg (Mintzberg, 1983) et de Mosvick (Mosvick, 1986) qui proposent des activités et des méthodes de travail de groupe à hauteur de 70 % du temps d'utilisation total du système d'information.

Ces mêmes auteurs ont ainsi défini le concept de « groupware » comme un ensemble compact de logiciels, de matériels, de techniques de communication et de méthodes d'organisation destinés à développer le travail en commun sur les flux d'information.

Nous préciserons ici que les outils et méthodes mis en œuvre dans le cadre de processus de travail collaboratif doivent être adaptés à des tâches (ou a des objectifs) bien précis(es). En effet prévoir un espace de travail collaboratif sans que les méthodes d'organisation, les objectifs à atteindre, les natures de flux manipulés n'aient été préalablement bien définis laisse trop de « flou » et crée une forte démotivation chez les utilisateurs.

Il est en effet prouvé que les systèmes mettant en œuvre des utilisateurs très différents et de l'information fortement déstructurée conduisent à de telles incertitudes quant à l'utilisation de l'outil et à la fiabilité des informations, à tel point que l'investissement des acteurs devient totalement improductif (Grudin, 1988).

L'idéal serait ainsi de permettre à chaque utilisateur de personnaliser aussi bien son interface que l'organisation (et /ou l'indexation) des données circulant sur le système, afin que chacun puisse utiliser ses propres mécanismes intellectuels de réflexion et de résolution de problèmes. Toutefois, des études ont montré que cette flexibilité dans l'utilisation du système de gestion de connaissances tend à séparer les individus en fonction de leurs

compétences en informatique. On voit ainsi apparaître sur les services d'information plusieurs classes d'individus, réparties selon le niveau de compétences techniques des utilisateurs, ce qui crée des divergences de préoccupations et se révèle ainsi totalement incompatible avec la notion de travail collaboratif (Chen, 1990).

Le système de gestion de connaissance mis en place se devait, bien entendu, de minimiser cet effet induit lié à la dispersion des compétences en informatique parmi le public concerné.

Nous devions donc concevoir un système assez souple pour que chacun puisse y trouver une marge de manœuvre, assez simple d'accès pour pouvoir être utilisable par n'importe quel acteur de la structure, mais dont l'utilisation et les possibilités soient compatibles avec des méthodes de traitement commun, aussi bien au niveau des documents eux-mêmes que sur le plan des règles et protocoles d'utilisation du système.

Ces réflexions excluaient ainsi, de fait :

Un système de gestion de bases de données pur (SGBD type Access ou Oracle) car trop rigide d'accès et trop complexe pour les novices.

Un système de type tableau noir (Black Board), très simple d'utilisation mais trop pauvre en matière de gestion de l'information (pas d'index, aucune structuration des données)

Nous avons ainsi opté pour un système mixte, c'est à dire offrant une souplesse d'utilisation comparable à celle d'un site web tout en offrant, de manière partiellement cachée des possibilités de recherche, de classement et de tri d'informations. Un espace de travail collaboratif étant, bien entendu ménagé au cœur du système de gestion de connaissances. Les choix techniques seront plus amplement décrits au cours du paragraphe 4 du présent document.

4.2. Modèles de gestion

Parallèlement à ces considérations qui concernent essentiellement l'information qui est créée ou modifiée par les utilisateurs, il est indispensable de conduire une réflexion sur le mode d'échange et de diffusion de l'information, et, ce quelle que soit sa nature : interne ou externe, rigide ou modifiable.

Cette problématique conduit inévitablement au choix d'une méthode par « Push » ou par « Pull ». Nous rappelons ici que la méthode de Push consiste à envoyer l'information au destinataire (id. e-mail) tandis que le modèle Pull impose à l'utilisateur d'aller chercher l'information lui-même.

Etant donné que ces méthodes sont fondamentalement opposées, chacune trouve sa raison d'être dans une somme d'avantages et d'inconvénients. Nous tenterons ici de faire le point sur les caractéristiques de chacun de ces deux modèles de diffusion afin d'opérer un choix technologique pour le système mis en œuvre.

4.2.1. La méthode Push

D. Stenmark des laboratoires Volvo a réalisé une étude très intéressante sur les « effets de bords » liés aux technologies push (Stenmark, 1998).

Le modèle de diffusion push entraîne un certain confort de l'utilisateur, dans la mesure où les flux d'information sont directement dirigés vers son espace de travail, mais si cette méthode permet à tout novice d'avoir accès à l'information, plusieurs écueils peuvent prendre naissance :

L'utilisateur peut être inondé d'information et peut ainsi se « noyer » sous une masse de données pas toujours pertinentes (Davis, 1985).

Le débit du canal d'information peut être soit trop faible, soit trop élevé, ce qui conduit dans un cas comme dans l'autre à des contraintes dans le rythme de travail et dans l'organisation des individus, on perd ici le concept de souplesse et de personnalisation que nous avons évoqué plus haut.

Délivrer de l'information à des individus impose d'effectuer deux types de choix : la détermination de l'individu à atteindre et la sélection de l'information à diffuser vers le récepteur.

Des trois inconvénients cités précédemment, le dernier est sans aucun doute le majeur. En effet, les choix effectués posent les questions suivantes : (Hall, 1998)

Qui effectue ces choix ?

En fonction de quels critères ?

Formuler ces questions remet totalement en cause le concept de Gestion de Connaissance et de Savoir, car nous entrons ici dans une hypothèse de total déterminisme où chaque acteur de la structure est identifié comme un expert dans tel ou tel domaine, et où les compétences de chacun ne peuvent résider que dans le domaine de

spécialité auquel il appartient, un tel système entérinerait un état de fait selon lequel chaque individu aurait une place fixe et un rôle déterminé dans la structure, empêchant ainsi toute interaction avec les autres.

4.2.2. La méthode Pull

La collecte d'information selon la méthode « pull » contraint l'utilisateur à aller lui-même chercher les données. Selon un éclairage purement matérialiste, le modèle pull permet de ne pas dupliquer l'information vers chaque usager et ainsi réaliser des économies de place mémoire sur les unités de stockage centrales ou distantes.

Mais au delà de cette caractéristique, la diffusion d'information selon ce processus conduit à des situations fortement favorables à la création de connaissance.

Quelques auteurs ont souligné l'importance d'une implication active des utilisateurs, en mentionnant, notamment le caractère culturel et volontariste de la démarche, caractéristiques indissociables du concept de « partage des connaissances » (Holtz, 1996) .

A ce propos, Hackathorn s'est intéressé à la modélisation des échanges de flux au sein d'une logique « d'information pull » (Hackathorn, 1997).

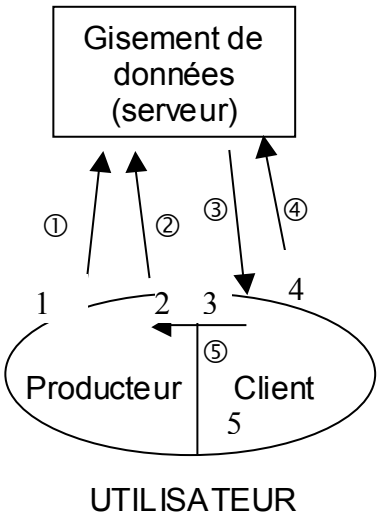


Figure 1 : Les cinq processus de l'information pull (selon Hackthorn)

Ce modèle fait apparaître la double personnalité de l'utilisateur qui peut être tantôt producteur, tantôt consommateur d'information associée aux cinq phases d'échanges de flux d'informations que nous détaillons ci-dessous : (Rq. La numérotation de phases n'introduit pas de synchronisme ni de référence temporelle)

Phase 1 : **Publication**
Lors de la phase de publication, un document est crée par le producteur, il est en outre rendu disponible au niveau du serveur d'information lié à l'intranet.

Phase 2 : **Rafraîchissement**
La phase de rafraîchissement est constituée par la mise à jour ou la modification des données primaires précédemment créées. Ce processus peut être déclenché sur la simple initiative du producteur en fonction du contexte ou bien à l'issue d'une interaction Client-Producteur (naissance d'une valeur ajoutée et/ou expertise, éléments nouveaux) cf. Phase 5

Phase 3 : **Consultation**
Représente la consultation des documents repérés par l'utilisateur et disponibles sur le serveur.

Phase 4 : **Recherche**
La phase de recherche constitue la part active du client, elle sous-tend l'existence d'outils de recherche sur le serveur, tels qu'index, moteurs de recherche, agents ….

A l'issue du processus de recherche, l'utilisateur doit pouvoir consulter des documents et avoir une vue assez exhaustive du contenu de la base de connaissance de la structure.

Phase 5 : **Interaction**
La phase d'interaction constitue en quelque sorte le « feed-back » du système. C'est à ce niveau que la confrontation entre plusieurs documents peut donner naissance à une nouvelle information à haute valeur ajoutée. Les compétences du client (expert) sont ici mise à contribution pour « créer de l'intelligence ».

Cette étude montre, de manière indéniable l'intérêt de « l'information pull » comme support de la gestion des connaissances et du savoir, mais requiert de la part de l'utilisateur deux agilités que l'on ne peut pas forcément exiger de tout utilisateur :

La connaissance des outils et méthodes de recherche et d'interrogation
La localisation des nœuds de stockage des informations

Pour cette raison, nous avons décidé d'opter pour un système mixte Push-Pull, au sein duquel, les clients potentiels sont avertis des changements, des nouveautés, des nouvelles sources d'information par messagerie (technique push), ce qui leur permet d'aller consulter et d'approfondir des domaines en se connectant au serveur et en allant chercher (ou déposer) l'information sur les nœuds désignés par le message (technique pull). De manière plus précise, pour s'affranchir du recours à un « animateur » c'est à dire à une personne chargée des opérations de messagerie, nous avons fait de chaque utilisateur un fournisseur d'information push en associant de manière transparente pour l'utilisateur une opération de messagerie, conjointe à un dépôt ou une modification d'information sur l'intranet.

Le paragraphe du présent document illustrera cette propriété.

5. LES SOLUTIONS TECHNIQUES ENVISAGEES

5.1. Cahier des charges

L'analyse des flux résumée au fil du paragraphe 3 du présent document conduit à l'élaboration du cahier des charges suivant.

A titre de précision, nous soulignerons la non exhaustivité dans le recensement des données à prendre en compte. En effet, l'expérience montre qu'un tel système s'auto-entretient, dans le sens où les flux d'information, n'étant pas fondamentalement indépendants, créent des intentions et donnent naissance à des besoins ou à des données dont on n'avait pas pressenti l'existence.

L'art de la conception consistera à concevoir un système assez rigide pour être facilement manipulable par des utilisateurs très différents, mais en laissant toutefois une place à l'évolutivité et à l'adaptabilité.
A titre de complément, nous recenserons ici quelques éléments et fonctions du cahier des charges.

Fonction principale :
La système doit assurer la fonction de Gestion des Connaissances et des Savoirs internes à la structure. Notamment en remplissant les fonctions :

- ***de recueil ;***
- ***de diffusion ;***
- ***d'analyse ;***
- ***de valorisation ;***
- ***de mémorisation des données***.

Fonctions contraintes :
Le système doit s'adapter sur la configuration système et réseau existants à savoir : un réseau TCP/IP en étoile d'environ 50 Micro-ordinateurs de type PC répartis sur un domaine Windows 2000 ®.
Les solutions techniques doivent permettre une réutilisation des données déjà manipulées.
Les habitudes de travail des utilisateurs doivent subir le moins de changement possible.
La manipulation technique des outils mis en place doit être possible pour tous les non-informaticiens.
La mise en place, l'évolutivité et la maintenance du système de gestion des connaissances doit pouvoir être assurée par les ressources humaines locales, à savoir deux ou trois personnes, non professionnels de l'informatique et réalisant cette tâche de manière cumulée avec leur fonctions principales.

5.2. Analyse des solutions

Il existe des solutions « clés-en-main » proposées par les représentants des grandes marques de produits informatiques : Microsoft, IBM, Bull, …, faisant appel à une étude détaillée des besoins et fournissant une réponse adaptée et personnalisée.

Cette solution constitue sans doute un optimum de performance et de fiabilité, mais présente toutefois certains inconvénients que nous n'avons pu ignorer :

- Le coût de la prestation ;
- La facilité d'utilisation ;
- Les besoins de maintenance.

En effet, opter pour une solution personnalisée entraîne des coûts d'étude et de mise en place très importants d'une part, mais aussi, étant donné le caractère professionnel d'un tel système, il faut prévoir une formation « lourde » des utilisateurs et s'attendre à une modification radicale de leurs habitudes de travail, chose que nous ne désirions pas.

D'autre part, les compétences en termes d'administration, d'évolutivité et de maintenance se trouvent délocalisées chez le prestataire de service, ce qui n'obéit pas, non plus, au cahier des charges.

Nous avons donc opté pour une solution hybride, à base de produits informatiques professionnels, mis en place par la cellule intranet du site UNIMECA.

Ceci nous permettait à la fois de ne pas avoir à développer de logiciels, étant donnée l'étendue de l'offre de produits professionnels, tout en maîtrisant les clés du système, permettant des adaptations éventuelles ainsi que la maintenance.

5.3. Mise en place

Le système mis en place est basé sur l'utilisation de Microsoft Internet Information Server Version 5 dans un environnement Windows 2000®. Les pages dynamiques seront réalisées en technologie ASP®.

5.3.1. Le cœur du produit

Afin de satisfaire aux théories mathématiques et Shannon et de Viterbi, ainsi qu'au modèle plus philosophique de Bachelard (voir paragraphe 3), nous proposons un système entièrement architecturé autour d'une base d'événements.

Cette base d'événements matérialise les canaux bidirectionnels d'information entre les différents éléments du milieu extérieur et opère le transcodage nécessaire à la mise en regard des différentes modalités d'information.
Un événement sera ainsi réalisé par :

Une communication au groupe
Le dépôt d'une ressource documentaire
La création et le suivi d'annotations sur un document

5.3.2. Mode opératoire et organisation

Outre les opérations traditionnelles de travail en commun sur des documents et de partage de données de type texte, permises par tous les réseaux, l'utilisation de produits tels que Microsoft Exchange Server ® permet la création de services de messagerie interne, il est ainsi possible :

- d'échanger en interne des notes, courriers ;
- de partager un carnet d'adresses ;
- de consulter et modifier le planning des tâches individuelles et communes ;
- de consulter et modifier les agendas.
- d'échanger des points de vue par l'intermédiaire de forums de discussion internes.

De plus le couple Outlook-Exchange offre aussi l'accès au courrier électronique traditionnel ainsi qu'au newsgroups publics.

Cette configuration permet le travail collaboratif aussi bien dans le cadre de l'intranet que par l'accès à des sources de données externes (web, newsgroups, mailing lists, email).

Chacun peut intervenir sur les données précédemment décrites et participer aussi bien à l'enrichissement de la mémoire du système que d'obéir à une démarche de gestion de projet efficace.

L'expérience montre que ce mode de gestion des connaissances est très efficace, notamment au niveau de la mise à jour et au renseignement de certains champs du carnet d'adresses.

Les sources d'information n'étant pas exclusivement électroniques, il a été nécessaire de prévoir un système de recueil des données, pour cela nous distinguons deux canaux d'information :

- L'information formelle écrite
- L'information informelle (orale ou intellectuelle)

Concernant l'information écrite, des postes munis d'un scanner permettent la numérisation de documents, tels que courriers, plaquettes, prospectus, articles de presse ou table des matières de périodiques.

Les images des textes sont transformées en documents PDF (à l'aide de Adobe Acrobat ®) et complétés d'une reconnaissance automatique des caractères. Cette étape est rapide et très simple (numérisation+ocr+création PDF en une seule opération), elle permet de garder une mémoire des documents manipulés ainsi qu'une recherche aisée dans la mesure où les documents dont les caractères ont été reconnus sont indexés en texte intégral. (Dou, 1997)

La recherche dans l'index et l'accès aux documents intégraux de type PDF est possible par le biais d'un serveur WEB implanté sur le système GCS.

L'information informelle concerne :

- Les réflexions et remarques personnelles ;
- La formalisation et/ou l'analyse de réunions, conversations et discours
- Les expériences concernant la résolution d'un problème particulier
- Les rapports d'étonnement,
- … etc.

Dans l'optique d'une collecte de ce type d'information, nous avons prévu la constitution d'une base de données munie d'une interface de type web (formulaire), le nombre de champs à renseigner est réduit à l'essentiel afin de ne pas démotiver les clients/acteurs. La faiblesse dans la structuration de tels documents (peu de champs descripteurs) sera amplement compensée par les propriétés d'indexation en texte intégral offertes par le serveur WEB.

Dans ce cas aussi, la recherche et l'accès aux documents se font à travers une interface WEB.

Selon ce type d'organisation, la répartition des pôles de collecte pour tous types d'informations permet à chaque individu d'être un acteur à part entière de l'opération de GCS.

6. PERSPECTIVES ET CONCLUSION

La plateforme présentée à l'occasion de cet article a fait l'objet d'une étude préalable très détaillée, il n'en demeure pas moins que, malgré un fonctionnement correct depuis plus de deux années, sur un volume d'usagers avoisinant les 150 individus, on peut la considérer comme étant expérimentale, dans la mesure où la cellule humaine chargée de son déploiement et de sa maintenance est très souvent sollicitée dans l'optique d'amélioration de fonctions existantes ou pour la création de services supplémentaires. Nous sommes ici au cœur de l'hypothèse de Simondon (cf. paragraphe 1) et nous voyons se former la boucle de rétroaction, par laquelle l'organisation sociale va réellement s'approprier l'objet.

La problématique réside dans les choix effectués à la base, que l'on se doit d'opérer de manière subjective et, si possible, les plus universels possibles, dans le sens où des modifications régulières doivent être possibles.

D'autre part nous insistons sur l'absolue nécessité de mettre en place des systèmes très simples d'accès, très conviviaux et très fiable, même au détriment des (théoriques) performances. (Maholtra, 1998)

Nous avons ainsi trop souvent assisté à des confrontations mettant en scène les Centres de Ressources Informatiques et les utilisateurs qui se révélaient complètement stériles du fait des conceptions totalement différentes de la notion de système d'information.

Notre expérience dans la conception de produits d'information auprès de plusieurs organismes nous a appris qu'une *« usine à gaz »* qui *« plante tout le temps »* décourage la majorité des utilisateurs et remet totalement en cause la stratégie d'information de la structure.

Un des grands enjeux des mois à venir résidera dans la mise en lumière d'indicateurs pertinents pour mesurer la réelle efficacité de l'outil proposé. Car il est évident que la seule mesure du taux de satisfaction des utilisateurs ne saurait suffire pour produire des résultats scientifiques fiables.

7. BIBLIOGRAPHIE

(Bachelard, 1968) - Bachelard G. – *Le nouvel esprit scientifique* - Paris : Les Presses universitaires de France, 1934 rééd. 1968.

(Chen, 1990) - Chen H., Dhar V. - *User misconceptions of online information retrieval systems* - International journal of man-machine studies, N°32, Vol.6, 1990.

(Davis, 1985) - Davis G., Olson M. - *Management information systems. Conceptual foundations, structure and development* - 2nd ed, Mc Graw Hill, 1985.

(Desjeux, 1994) – Desjeux D. – *Le sens de l'autre* – L'Harmattan, Paris, 1994.

(Dou, 1997) - C. Dou, E. Giraud - L'intégration du format PDF dans l'élaboration des systèmes d'information , Colloque International d'Information Elaborée - 12-16 mai 1997 - Ile Rousse

(Giraud, 1997) - E. Giraud, H. Dou - *Information flood management and multimedia integration in information systems* - International Sciences for Decision Making , N°1, 1997.

(Grudin, 1988) - Grudjn J. - *Why CSCW applications fail, problems in the design and evaluation of organizational interfaces* - Proc. Conference on Computer-Supported Cooperative Work, CSCW'88, Portland, 1988.

(Hackathorn, 1997) - Hackathorn R. - *Publish or perish* - Byte, N° 52, 1997.

(Hall, 1998) - Hall R. - *How to Avoid Unwanted Email*, Communications of the ACM - vol. 41, no. 3, 1998.

(Holtz, 1996) - Holtz S. - *The intranet advantage* - Macmillan Computer Publishing, 1996.

(Jordan, 1997) – Jordan J. – *Competing trough Knowledge : an introduction* – Technology analysis and strategic management , Vol. 9 , N°4 , 1997 , pp. 379 390

(Maholtra, 1998) - Maholtra Y., *Knowledge management, Knowledge Organisations & Knowledge workers* - Maeil Business, Feb. 1998

(Mintzberg, 1983) - Mintzberg H., *The nature of managerial work* - Harper and Row, New-York Press, 1983.

(Mosvick, 1986) - Mosvick R. & Nelson R. - *We've got to start meeting like this : a guide to successful business meeting management* - Scott Foresman & co., 1986.

(Ranucci, 2001) - Ranucci JF. - *Elaboration de contenus de formation au travers d'une plate forme de travail collaboratif* – Mémoire de DEA en sciences de l'information et de la communication, veille et intelligence compétitive, CRRM, Université Aix– Marseille III, 2001

(Simondon, 1989) : Simondon G. – *Du mode d'existence des objets techniques, Aubier*, Paris, 1989.

(Shannon, 1949) - Shannon C. - *Communication in the presence of noise*.- Actes d'IRE, pp.10-21, janvier 1949.

(Stenmark, 1998) - Stenmark D. - *Identifying Problems with Email-based Information Sharing* - Proceedings of IRIS21, Department of Computer Science, Aalborg University, Denmark, 1998.

(Viterbi, 1979) - Viterbi A. & Omura J. - *Principles of Digital Communication and Coding.* - McGraw-Hill, 1979.

INTEGRATION DE COMPOSANTS DE TEXT MINING POUR LE DEVELOPPEMENT D'UN SYSTEME DE RECHERCHE ET D'ANALYSE D'INFORMATION

Grivel Luc
TEMIS Text Mining Solutions
59, rue de Ponthieu 75 008 Paris
http://www.temis-group.com
luc.grivel@temis-group.com

Résumé : L'objectif de cet article est de montrer l'intérêt de l'emploi combiné de techniques d'analyse du texte (segmentation, lexicale, syntaxique, sémantique) et de diverses techniques d'accès à l'information (index, classification, catégorisation, cartographie) pour le développement d'un système de recherche et d'analyse d'information **qui soit adapté à des non-spécialistes des langages documentaires et qui s'intègre dans un processus de veille.** L'article montre comment ces techniques interviennent dans les fonctions d'un système d'analyse de l'information. L'originalité se situe dans l'approche (intégration de composants de text mining) qui est détaillée : reformatage XML des documents, visualisation des résultats, en passant par l'extraction des caractéristiques des documents et la classification.

Mots clefs : Fouille de données textuelles, extraction information, traitement du langage naturel, classification, hypertexte, cartographie

Abstract : The goal of this paper is to show the interest of combining various text analysis techniques (shallow parsing, semantic analysis, etc.) and some information access techniques (indexing, classification, clustering, mapping)) to developp an information analysis system to be used and customized by non-specialists of documentary languages. The paper shows how these techniques can be integrated to for a process chain including : XML reformating, information extraction, clustering, mapping.

Keywords : text mining, information extraction, natural language processing, classification, clustering, mapping, hypertext

Intégration de Composants de Text Mining pour le développement d'un système de recherche et d'analyse d'information

INTRODUCTION

L'objectif de cet article est de montrer l'intérêt de la combinaison de techniques d'analyse du texte (segmentation, lexicale, syntaxique, sémantique) et d'accès à l'information (index, classification, catégorisation, cartographie), rassemblées sous le nom de text mining, en prenant pour cadre le développement d'un système de recherche et d'analyse d'information **qui soit adapté à des non-spécialistes des langages documentaires et qui s'intègre dans un processus de veille**.

Dans ce processus itératif qu'est la veille, on peut distinguer quatre fonctions essentielles d'un système d'analyse de l'information :

- Automatiser la constitution et la mise à jour régulière d'une base documentaire, avec la meilleure couverture possible pour les axes de surveillance recensés.
- Annoter les documents par extraction d'information en vue de leur accès et leur organisation ultérieure
- Les stocker dans une base documentaire
- Fournir une interface conviviale permettant d'exploiter cette base selon différents modes de recherche et scénarios d'analyse, en combinant recherche, statistiques, catégorisation et classification du résultat de la recherche.

Comment les techniques citées plus haut interviennent elles dans ces fonctions?

Chacune des techniques est vue comme un composant aux fonctionnalités précises et délimitées. Chaque composant constitue un 'objet-serveur' ou objet distant et dispose d'une API (Application Programmatic Interface) Java publique lui permettant de s'intégrer facilement dans une application existante (Figure 1).

L'administrateur de sources 'crawle' différentes sources d'information, actionne des filtres de conversion XML des documents (en fonction du type de document et de la date de mise à jour) et les stocke dans un répertoire local.

Un serveur d'extraction pour dégager les concepts clé contenus dans les documents (noms de compagnies, dates, valeurs monétaires, fonctions, lieux, ou tout autre concept relatif à un domaine…) et génère les metadonnées décrivant chaque document.

Un serveur de recherche documentaire ou un SGBD stocke et indexe les documents et leurs metadonnées.

Un serveur incorporant un moteur de classification et un moteur de catégorisation classifie les documents (constitue des groupes), ou les catégorise (les place dans des groupes définis à priori).

DESIGN DU SYSTEME D'INFORMATION

Nous détaillons ici les composants et montrons comment ils interagissent dans cette architecture selon que l'objectif recherché est de naviguer dans une collection entière de documents ou de naviguer dans des résultats de recherche.

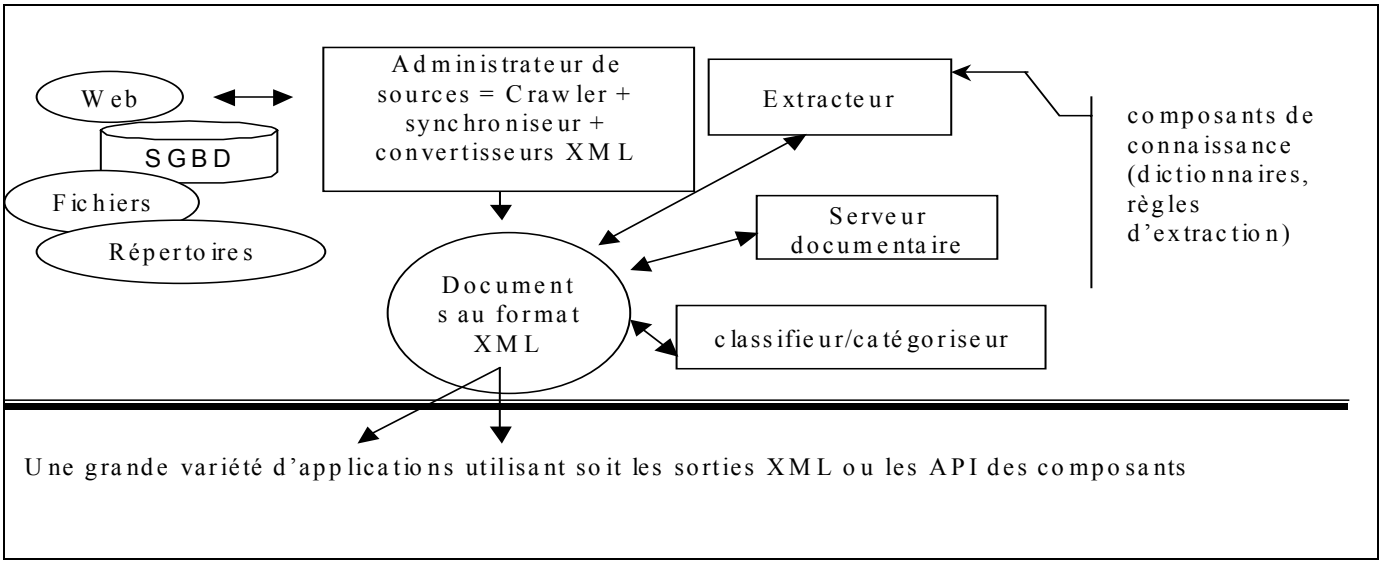


Figure 1 : Architecture

L'administrateur de sources

Il met à jour une base de données à partir de documents provenant de différentes sources pré-définies. Ce composant inclut les fonctions permettant de se connecter à différentes sources d'information, (Web, mail, news, data banks) dans différent formats (ascii, HTML, Word, PowerPoint, Excel, PDF) et de définir un profil de recherche sur ces sources. Tous les documents correspondant à ce profil sont périodiquement et automatiquement recherchés, copiés et stockés dans un répertoire local. Ils sont ensuite convertis dans un format homogène XML. Cette approche est aujourd'hui largement utilisée par les communautés des bases documentaires et des SGBD (orienté objet ou relationnel) lorsqu'il s'agit d'intégrer des documents hétérogènes. [MIC 98] [ABIT 97]

Dans l'exemple ci-dessous, une proposition d 'emploi comporte différents champs : Job Title, Description, Skills, Education, Location. Le reformatage XML a permis de conserver cette notion de zone de texte qui pourra être exploitée pour l'extraction et l'organisation de l'information.

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<doc_list xmlns:dc="http://purl.org/dc/elements/1.1/">
<DOC>
<dc:Identifier>bjo_1</dc:Identifier>
<dc:Title> OFFICE ADMINISTRATOR </dc:Title>
<text zone="location">
SC Missile Defense &amp;amp;Space Control - Anaheim
</text>
<text zone="description">
Site CM Office Administrator will be responsible for supporting one or more site resident CM work groups. Site CM OA performs various
tasks contributing to office organization and efficiency. Duties included preparation and distribution of memos, reports, forms and business
 graphics; directs phone calls and electronic mail; schedules and coordinates meetings and business travel. Uses advanced computer SW
features. Edits documents for correct grammar, punctuation, spelling, context and format.
   Anticipates what information and data will be needed for appointments, meetings and business travel. Follows up on assigned action items.
Anticipates the need for office supplies. Establishes records, files and logs. This  position will include duties and assignments in support of
the NMD program CM  organization as required. This position is located in Arlington, VA.
</text>
<text zone="skills">
Team player with good oral and written communication skills. Nominally skilled in the use of the MS Office Suite of
    tools. Ability to work in a fast paced, multi-tasking environment.
</text>
<text zone="education">
Prefer vocational school training or equivalent work     experience; proficiency in operation of office equipment and business
software; and strong organizational, communication, and interpersonal skills.
</text>
</DOC>
…
```

L'extracteur et les composants de connaissance (Skill Cartridge™)

Outre la récupération des meta-données existantes (par exemple, les champs qui sont présents dans le cas de documents provenant d'une base bibliographique), il est indispensable de générer des meta-données décrivant le contenu du document si on désire implanter des mode de recherche qui ne soient pas uniquement fondés sur une simple indexation de tous les mots du texte. C'est ici qu'intervient la phase d'extraction d'information.

Actuellement les systèmes d'extraction industriels les plus évolués s'appuient sur des analyseurs morpho-syntaxiques basés sur la technologie des transducteurs[66]. Ils prennent éventuellement en compte des règles d'extraction ou des ressources terminologiques. Les analyseurs morpho-syntaxiques sont de plus en plus performants (20 Mo/h pour Xelda de Xerox) et capables de traiter de plus en plus de langues différentes. Citons également les systèmes basés sur INTEX [Silb 99] et FASTR [Jacq 94] dans le monde universitaire.

Le serveur d'extraction développé par TEMIS est capable de traiter 7 langues (anglais, français, allemand, espagnol, hollandais, portugais, italien). Il est paramétré par un ou plusieurs composants de connaissances (Skill Cartridges™) qui définissent les éléments à extraire. Le système recherche les patterns décrits dans les règles et affecte aux éléments trouvés une étiquette syntaxique ou sémantique tels que des noms de compagnies, des relations (fusion de X avec Y, achat de W par Z), des noms de lieux, des dates, des prix, ….

Le résultat de l'extraction est une annotation du document (ici une offre d'emploi) par les éléments extraits. Le fait de connaître la sémantique associée à une zone de texte permet éventuellement de 'contrôler' les meta-données générées pendant la phase d'extraction d'information en développant des règles d'extraction adaptées à cette zone de texte. Par exemple, les fonctions dans un secteur d'activité donné, les diplômes requis, le nombre d'années d'expérience.

La méthodologie de construction de règles d'extraction développée par TEMIS est basée sur la notion d'organisation de règles d'extraction en niveaux hiérarchiques pour contrôler leur exécution sur les corpus[Bus 02]. Les problèmes relatifs à l'acquisition de la terminologie [Fau 00], [Bou 99], la création d'ontologies [Roc 00], la création de règles d'extraction [Gris97] sont complexes. Un programme de recherche [Gri 01a] est en cours pour aider les linguistes dans leur tâche. L'objectif est de développer un environnement logiciel pour aider à la customisation des Skill Cartridges™. Pour ce projet, TEMIS s'est vu décerner récemment le label technologie-clé par l'**ANVAR**.[Aub 02]

Organisation des données a priori versus organisation des résultats de recherche

A l'issue de l'étape d'extraction, chaque document est donc annoté par des concepts ou par des éléments syntaxiques (noms, verbes, adjectifs).

Les techniques usuelles pour faciliter l'accès à l'information sont :

- Recherche par mots-clés ou sur des concepts et classement des documents selon leur pertinence par rapport à la requête (Salton)
- Classification et cartographie: la classification (clustering) permet de regrouper les documents similaires par thèmes sans a priori sur la structure thématique. La Cartographie est un moyen de présenter un résumé de la classification,
- Catégorisation de documents : sur la base d'un vecteur de caractéristiques comprenant, entre autres, les résultats de l'extraction, la catégorisation permet d'affecter des documents à des rubriques ou catégories prédéfinies.
- Analyse statistique (distribution de concepts, distribution d'auteurs, distribution par dates de publication, …)

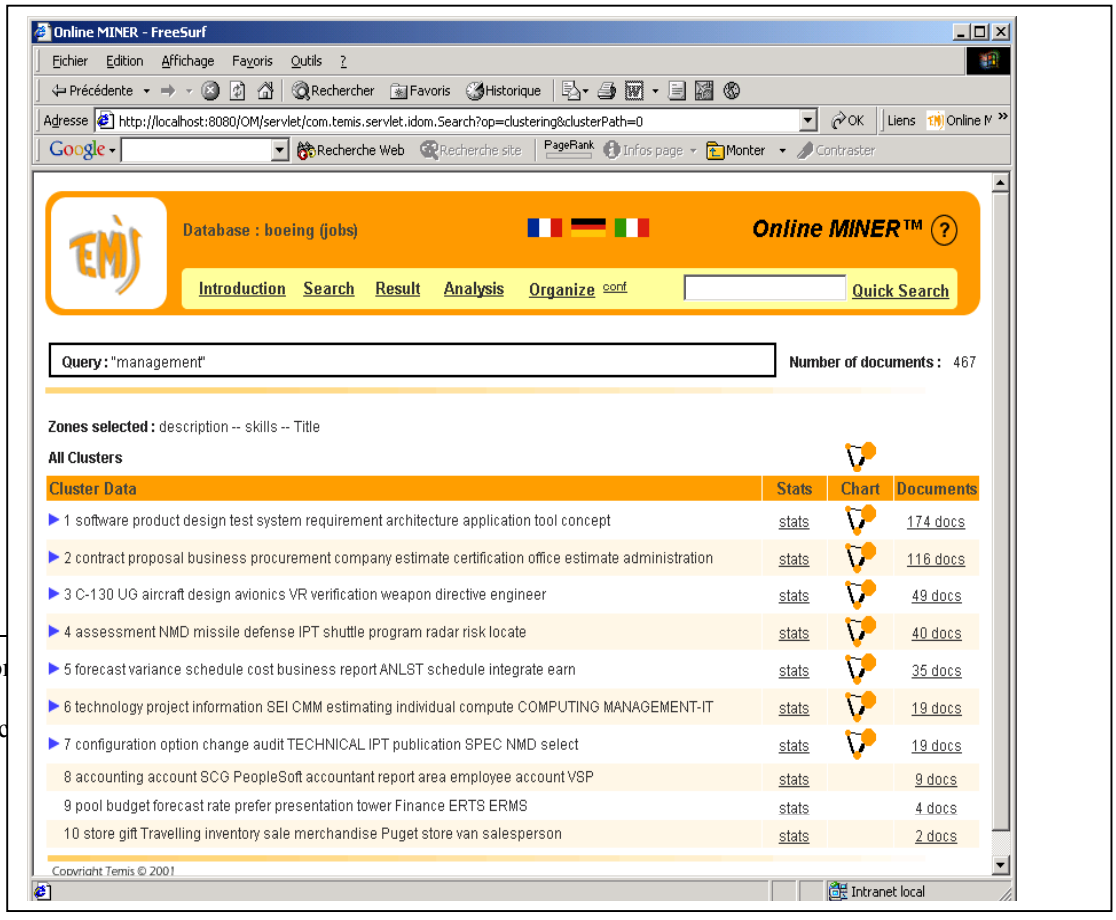
Ces techniques peuvent être combinées de manières différentes selon l'objectif recherché : naviguer dans une collection entière de documents ou naviguer dans des résultats de recherche.

[66] Le langage de description des règles d'extraction est de la famille des langages réguliers. Les transducteurs (automates d'états finis) sont un moyen efficace d'opérer sur ces langages. Cette efficacité garantit de pouvoir traiter de gros volumes de données.

Dans le cas où l'objectif est de permettre une navigation dans une collection de documents, il est nécessaire de présenter à l'utilisateur une structure fixe. Si on dispose pour chaque catégorie d'un ensemble de documents représentatifs, l'emploi d'un moteur de catégorisation[67] basé sur un modèle d'apprentissage s'impose-. Si l'on ne dispose pas d'exemples de documents pour chaque catégorie, le moteur de classification peut être utilisé. C'est l'approche qui était développée dans HENOCH (INIST) [GRIVEL 2000, 2001] et dans TEWAT (IBM) [COUPET 1995, 1998]. Après nettoyage et validation des résultats de classification sur la collection de documents, on peut utiliser ces mêmes résultats pour initialiser un moteur de catégorisation.

Dans le cas où l'objectif est d'aider à la navigation dans des résultats de recherche**,** la classification doit donc être effectuée à la volée, ce qui a un impact sur le choix de l'algorithme utilisé. Des exemples d'algorithmes adéquats sont donnés dans [Dou92]. Le serveur de classification de TEMIS est de type non hiérarchique (partition en K classes, K étant un paramètre fixé à priori), déterministe (on obtient le même résultat sur un même groupe de documents ordonnés). Il prend en entrée une description vectorielle des documents sous forme de hiérarchie de concepts et tient compte de cette hiérarchie pour la mesure de la similarité entre deux documents. Il est aussi possible de tenir compte des types de meta-données qui caractérisent les documents à classer, en excluant par exemple le contenu certains champs que l'on ne veut pas voir intervenir dans le calcul de similarité inter-documents.Une carte globale ou un tableau permettent de résumer les caractéristiques des clusters (les termes les plus pertinents du cluster, les relations inter-clusters, les statistiques sur différents champs des documents de chaque cluster (titres, auteurs, dates, …). Sur la base de ce résumé, l'utilisateur peut sélectionner un cluster qui lui semble contenir les documents les plus pertinents et classifier à nouveau ces documents s'il désire un niveau de détail plus fin sur la structure de ce sous-ensemble. Cette approche est particulièrement avantageuse lorsque l'utilisateur ne veut pas ou ne peut pas exprimer une requête bien formalisée, nécessitant de connaître toutes les finesses du langage de requête, la structure de la base documentaire, les plans de classement éventuellement utilisés (cas des brevets ou des offres d 'emploi). Il peut exprimer des requêtes plus simples, comportant éventuellement des termes ambigus et s'appuyer sur la classification pour sélectionner les documents les plus intéressants. Ceux-ci sont, le plus souvent, rassemblés au sein d'un même cluster [Hea96]. Les 'clusters-poubelle' sont rapidement identifiés (peu de documents, peu cohérents). Ce qui a pour conséquence de réduire le temps passé à l'identification des documents intéressants lorsque la réponse à une requête comporte trop de documents pour être analysée séquentiellement ou dans un temps limité.

L'exemple ci-dessous (Figure 2) montre comment sur un ensemble d'offres d'emploi aspirées sur un site Web, on peut à partir du résultat d'une recherche sur un terme général tel que 'management' distinguer les ensembles de document ayant trait au 'business management', au 'sofware management', 'risk management', 'project management', 'option and configuration management', …



[67] La Catégor
d'utilisateur.
lui-même la c

Figure 2 :Résultat d'une classification sur les champs 'description', 'skill' et 'titles'

CONCLUSION

Plus simples à développer, plus robustes car testés dans des contextes différents, ces composants peuvent être assemblés pour créer des applications dans des domaines variés et dans différentes langues. Citons par exemple :

- Grunher+Jahr en Allemagne, premier éditeur de presse en Europe, filiale de Bertelsmann : un système d'indexation et de catégorisation d'articles de presse,
- un consortium italien (TELCAL) : un système de veille dans le domaine de l'artisanat, l'agriculture et le tourisme,
- et des applications de gestion des relations humaines, de gestion de la relation clientèle ou de veille concurrentielle ou technologique en Suisse, Allemagne, USA, et France.

BIBLIOGRAPHIE

[ABI 97] ABITEBOUL S., CLUET S., CHRISTOPHIDES V., MILO T., MOERKOTTE G., SIMEON J. - Querying Documents in Object Databases -, International Journal on Digital Libraries, 1(1), 5-19, 1997.

[AUB 02] Aubry, C., Grivel, L., Guillemin-Lanne, S., Lautier, C., « Une méthodologie et un environnement d'aide à la construction de composants de connaissance pour l'Extraction d'Information » *CIFT'02, **Colloque International sur la Fouille de Texte 20-23 octobre 2002***, Hammamet-Tunisie, 2002.

[BOU 99]Bourigault D. et Jacquemin, C. (1999) : Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology. *In Proceedings, 8th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, Bergen, pages 15-22.

[COU 95]Coupet P., Grandjean N., Huot C.et Chellali T.: Application du logiciel Technology Watch à l'analyse du développement pharmaceutique pour le domaine des maladies neurodégénératives, Les systèmes d'information élaborée, *Congrès S.F.B.A.,* juin 1995.

[COU 98]Coupet P., Hehenberger Michael.: Text Mining applied to patent analysis, Les systèmes d'information élaborée, *Annual Meeting of American Intellectual Property Law Association (AIPLA) Airlington.,* octobre 1998.

[DOU 92] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey, *Scatter/Gather: ACluster-based Approach to Browsing Large Document Collections*, SIGIR '92, Pages 318 – 329, 1992.

[FAU 00] Faure D. Poilbeau T. Extraction d'information utilisant INTEX et des connaissances sémantiques apprises par ASIUM, premières expérimentations, RFIA 2000, F. Schmitt et I Bloch Editeurs, Paris, France.

[Gei 00.] Geißler S. "The DocCat system - Automatic Indexing in a pratical Application" in: "Medien Informations Management Praxis, Projekte Präsentationen" Verlag für Berlin-Brandenburg, Potsdam Hrgs. Ralph Schmidt Potsdam, 2000,Seiten 48-55

[GRI 97]Grivel L., Polanco X., Kaplan A. : 'A computer System for Big Scientometrics at the Age of the World Wide Web', *Scientometrics*, vol.40, n°3, 493-506, 1997.

[GRI 95] Grivel L., Mutschke P., Polanco X.: 'Thematic mapping on bibliographic databases by cluster analysis : a description of SDOC environment with SOLIS', *Journal of Knowledge Organization*, Vol. 22, n°2, 70-77, 1995.

[GRI 95] Grivel L., Francois C. : 'Une station de travail pour classer, cartographier et analyser l'information bibliographique dans une perspective de veille scientifique et technique' - *Solaris* n°2 "Les sciences de l'Information : Bibliométrie, Scientométrie, Infométrie", Presses universitaires de Rennes, p.81-113, 1995.

[GRI 00] Grivel L. : L'hypertexte comme mode d'exploitation des résultats d'outils et méthodes d'analyse de l'information scientifique et technique, thèse de doctorat en Sciences de l'information et de la communication, Université Aix-Marseille III,. 10 janvier 2000.

[GRI 01] Grivel Luc, Guillemin-Lanne Sylvie, Lautier Christian, Mari Alda La construction de composants de connaissance pour l'extraction et le filtrage de l'information sur les réseaux Filtrage et résumé automatique de l'information sur les réseaux, 3ème congrès du Chapitre français de l'ISKO International Society for knowledge Organization, 5-6 juillet 2001

[GRISH 97] Grishman, R. (1997). Information Extraction: Techniques and Challenges. In Pazienza, M. T., editor, Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Frascati, Italie. LNAI Tutorial, Springer. 1418.

[JAC 94] Jacquemin, C.. FASTR : A unification grammar and a parser for terminology extraction from large corpora. In *Proceedings, Journées IA'94*, pages 155-164, Paris. Paris : EC2. (1994e)

[HEA 96] Hearst M.and Pedersen J., Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results, *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996

[MIC 98] MICHARD A. 'XML Langage et application' Editions Eyrolles, 361 p, 1998

[RIL 99] Riloff E. Jones R. Learning Dictionnaries for Information Extraction by multi level Bootstrapping Proceedings of Sixteenth National Conference on Artificial Intelligence, AAAI 1999, Orlando Floride.

[ROC 00] Roche C. « Corporate Ontologies and Concurrent Engineering », in: *Journal of Materials Processing Technology* volume 107, pages 187-193, Elsevier Science, 2000.

[SOD 99] Soderland S. Learning Information Extraction Rules for Semi-Structured and Free Text, Machine Learning 34, 233-272, 1999.

[SIL 99] Silberztein, 1999. *INTEX: a Finite State Transducer toolbox*, in Theoretical Computer Science #231:1, Elsevier Science

[TOU 98] Toussaint, Y., Namer, F., Daille, B., Jacquemin, C., Royauté, J. et Hathout, N. (1998). *Une approche linguistique et statistique pour l'analyse de l'information en corpus*. In P. Zweigenbaum, editor, *Proceedings, TALN'98*, pages 182-191.

[ZWE 00] Zweigenbaum, P. and Grabar, N. (2000). Expériences d'acquisition automatique de connaissances morphologiques par amorçage à partir d'un thesaurus. In Schmitt, F. and Bloch, I., editors, (RFIA'2000), volume II, pages 101--110, Paris, France

PLATE-FORME D'ENSEIGNEMENT A DISTANCE ET ENSEIGNEMENT EN ALTERNANCE :

EXEMPLE DE LA LICENCE PROFESSIONNELLE TOURISME ET NOUVELLES TECHNOLOGIES DE L'INFORMATION DE L'UNIVERSITE DE MARNE-LA-VALLEE.

LACOUR Marie Christine
Université de Marne la Vallée, Institut Francilien d'Ingénierie et des Services (IFIS)
2, allée du Promontoire 93160 NOISY LE GRAND
Tél. : +33 (0) 1 49 32 91 13
lacour@univ-mlv.fr

DOU Jean Marie
IMCS (Information Management Consulting & Solution)
8, rue Crillon 13005 Marseille France
Tél. : +33 (0) 491 333 801

BARON François
Université de Marne la Vallée, Institut Francilien d'Ingénierie et des Services (IFIS)
2, allée du Promontoire 93160 NOISY LE GRAND
Tél. : +33 (0) 1 49 32 91 13
f.baron@voila.fr

Résumé : Le développement des formations professionnalisées en alternance (apprentissage, formation continue, stages …) s'appuie sur la promotion de partenariats privilégiés avec les entreprises : c'est le cas de la licence professionnelle " Tourisme et NTIC " de l'Université de Marne-la-Vallée.
Cette formation en alternance nécessite un suivi attentif des étudiants en période d'entreprise. Nous proposons l'expérimentation d'une plate-forme pour améliorer ces relations. La plate-forme, IMCS, permet un travail coopératif et à terme la création puis la gestion et la valorisation de connaissances.
Nous passerons rapidement en revue les modes opératoires et les difficultés principales rencontrées par les enseignants pour le suivi des étudiants hors université.

Abstract : The development of professionnal training is based on preferential partnership with enterprises, this is the case of the professionnal Bachelor " Tourism and New Technologies " of Université de Marne la Vallée.
This training, both in University and enterprise requires a very careful management of the student during the enterprise period. We put forward the experimentation of one platform in

order to improve these relations. The I.M.C.S (Information Management Consulting & Solution™) platform, allows a collaborative work, and in the short term the creation, the management and the valuation of knowledge.
We review the operating processes and the main difficulties encontered by the teachers for the management of the students during the enterprise period.

Mot-clés : plate-forme de travail coopératif, gestion des connaissances, relations entreprises, suivi de projet.

Key words : collaborative work platform, knowledge management, enterprise relations, project management

Plate-forme d'enseignement à distance et enseignement en alternance : exemple de la licence professionnelle Tourisme et nouvelles technologies de l'information de l'Université de Marne-la-Vallée.

CARACTERISTIQUE DE LA PLATE-FORME

La plate-forme IMCS (Information Management Consulting & Solution) est un extranet pour la formation en ligne à distance. C'est un dispositif évolutif qui convient à l'expérimentation que nous avons tentée en licence professionnelle Tourisme option Nouvelles Technologies.

Ses caractéristiques sont celles d'un site WEB privé et sécurisé, accessible en permanence, avec une gestion individualisée des utilisateurs (nom utilisateurs+mot de passe), une gestion de groupes utilisateurs, et surtout une différenciation animateurs et utilisateurs classiques.

PREMIERE EXPERIMENTATION

Dans cette première période d'utilisation du site nous avons modestement utilisé la messagerie interne, le forum de discussion interactive, l'agenda de groupe et la base de connaissances (KBase). D'autres fonctionnalités existent et il est également possible d'en ajouter.

Notre approche fut empirique : les étudiants et la plupart des enseignants ont utilisé d'un commun accord le forum et les e-mail. Les autres approches n'ont pas été immédiates.

Les seuls enseignants à mettre spontanément leurs cours en ligne ont été ceux qui ont une solide culture informatique, les autres n'ont pas adhéré au projet de mise en ligne de cours avec des objections différentes, notamment relatives à l'exploitation ultérieure non contrôlée des données mises à dispositions. La majorité des enseignants n'ont utilisé que le forum et les e-mails parfois accompagnés de documents de travail.

Côté étudiants la visite du forum a été constante pendant les périodes en entreprise, et moins assidue pendant les périodes à l'université.

CONCLUSION ET AFFINEMENTS DES MODALITES D'USAGE

En nous basant sur l'expérience de cette année, nous tirons en conclusion que pour améliorer la fréquentation de la plate forme et encourager les enseignants à mettre leurs cours en ligne et les étudiants à les consulter il faudra prévoir une courte période d'initiation au logiciel. Par ailleurs nous souhaitons aussi mettre en œuvre des devoirs en ligne.

Nous avons constaté que l'usage de cette plate-forme a renforcé la cohérence du groupe et créer des liens de nature différente des liens habituels entre le duo enseignants - étudiants et celui étudiants - étudiants. De nouveaux aspects collaboratifs et une entraide efficace se sont instaurés entre les étudiants du groupe.

PERSPECTIVES

Au-delà des difficultés culturelles et techniques rencontrées par les acteurs de cette expérience, on peut s'interroger sur le futur du dispositif surtout sur l'aspect d'une création de communauté d'utilisateurs, notamment sur le rôle que pourront prendre ou avoir les " anciens " dans ces dispositifs. En effet, une fois habitué à des échanges électroniques, il semble intéressant de garder cohérent cette notion de communauté et de la développer. Cela peut permettre non seulement des facilités de stages, emplois,... mais aussi un retour d'expérience de terrain sur le contenu, voire la pédagogie, de la formation.

Ce qui suppose, à terme, la création d'un espace d'échanges "multi-dimensionnels" Prof – Elèves Anciens – Profs Anciens – Elèves - Tuteurs Entreprise. Celui-ci permettra l'approfondissement et la concrétisation des idées échangées, la capitalisation des savoir-faire pour tendre vers la pérennisation des relations étudiants-enseignant-entreprises autour de projets à moyen et long terme s'étalant sur plusieurs années.

FILTRAGE AUTO-ADAPTATIF BASE SUR L'ANALYSE DE LA VARIANCE.

KAROUACH Saïd,
karouach@irit.fr

DOUSSET Bernard,
dousset@irit.fr

BOUTILLAT Nicolas
boutillat@irit.fr

Adresse professionnelle
IRIT-SIG, Université Paul Sabatier, 118, route de Narbonne
31062 Toulouse cedex 04
Tél : 05.61.55.67.81

Résumé : Nous abordons, ici, le problème délicat du filtrage de l'information et de ses différentes applications en amont du processus de veille scientifique et technique. Toute recherche d'information nécessite une phase de ciblage et de validation qui permet un recentrage sur le sujet choisi. Notre laboratoire s'est depuis longtemps penché sur ce problème et a proposé des techniques de filtrage basées essentiellement sur les réseaux de neuronaux **[BOUG01] [TMAR01]** et la notion de profil. Nous reprenons cette approche afin de proposer une alternative basée sur l'analyse de la variance. Son principe est le suivant : en partant d'une liste de documents corrélés positivement ou négativement avec un thème donné (profil), nous proposons de filtrer de nouveaux documents, pris à la volée ou dans un corpus existant. Un jugement sur la pertinence des réponses permet de recaler le modèle proposé par analyse de la variance et donc d'affiner ou d'adapter le filtrage soit en continu (dépêches d'agence, push) soit en boucle (validation d'un corpus sur des extraits). Nous proposons deux approches : modèle tout ou rien et arbitrage par pondérations. La pertinence de cette approche sera illustrée par des exemples et des comparaisons à des techniques déjà existantes.

Abstract : In this paper, we discuss the information filtering problem and its various applications connected with science and technology watch. Any information retrieval requires a phase of targetting and validation which allows a centring on the selected subject. Our laboratory was for a long time concentrated on this problem and proposed techniques of filtering based primarily on the neuronal networks **[BOUG01] [TMAR01]** and the concept of profile. We take again this approach in order to propose an alternative based on the analysis of the variance. Its principle is as follows: on the basis of a list of documents correlated positively or negatively with a given topic (profile), we propose to filter new documents,

taken with stolen or in an existing corpus. A judgement on the relevance of the answers makes it possible to readjust the model suggested by analysis of the variance and thus to refine or adapt filtering either uninterrupted (dispatches of agency, push) or buckles some (validation of a corpus on extracts). We propose two approaches: all or nothing model and arbitration by weightings. The relevance of this approach will be illustrated for examples and comparisons with already existing techniques

Mots-clés : Filtrage - Data mining - Veille stratégique scientifique et technique - Analyse de données.

Keywords : Filtering - Data mining - Science and technology watch - Data analysis.

Filtrage auto-adaptatif basé sur l'analyse de la variance.

1 - LE PROBLEME POSE

Initialement, nous partons de deux collections de documents :

- Les documents de référence (le profil)
- Les documents à filtrer (flot de documents, corpus, …)

La première collection est évaluée par l'utilisateur

- En mode tout ou rien : un document est pertinent ou ne l'est pas
- En mode pondéré : un document est plus ou moins pertinent sur une échelle ayant au moins trois niveaux fixée a priori.

Le profil est analysé pour en extraire la terminologie significative. Chaque terme possède donc une fréquence dans le profil ainsi que dans chaque document du profil. Nous allons désigner par F_j la fréquence relative du terme j dans le profil et par f_{ij} la fréquence relative de ce même terme dans le document i. Nous définissons ensuite une fonction de validité v_j qui est linéaire par rapport aux fréquences relatives et dont la valeur doit nous permettre de nous prononcer sur la pertinence de chaque document. Pour les documents du profil, cette fonction vaut théoriquement :

- -1 pour les documents non pertinents et +1 pour les documents pertinents (en mode tout ou rien)
- le niveau de pertinence attribué par l'utilisateur (en mode pondéré)

Il nous reste à trouver le modèle linéaire ou affine qui vérifie au mieux l'ensemble des équations ainsi produites :

$$v_i = \beta_0 + \sum_{j=1}^{n} \beta_j f_{ij} + e_i$$

où m est le nombre d'équations (i=1,m et m>n) et où e_i représente l'erreur sur l'équation i, β_j les coefficients du modèle, β_0 le terme constant dans le cas d'un modèle affine.

2 -LES DEUX APPROCHES POSSIBLES

2.1 - Méthode des moindres carrés

La méthode des moindres carrés consiste à minimiser la somme des carrés des écarts constatés entre les résultats de mesure et les valeurs obtenues à l'aide du modèle linéaire ou affine.

$$S = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} \left(v_i - \beta_0 - \sum_{j=1}^{n} \beta_j f_{ij}\right)^2$$

On peut aussi minimiser la racine carrée de S, qui n'est autre que la norme euclidienne du vecteur des erreurs dans R^m. Il s'agit en fait de trouver la pseudo solution du système linéaire à m lignes et n colonnes admettant comme second membre le vecteur des mesures de validité des documents du profil.

$$[f_{ij}][\beta_j] \cong [v_i] \quad (1)$$

La meilleure solution au sens des moindres carrés de ce système est obtenue en résolvant le système de Cramer associé :

$$[f_{ji}][f_{ij}][\beta_j] = [f_{ji}][v_i] \quad (2)$$

Mais dans notre cas, l'ensemble des fréquences relatives est lié par une équation et ce pour chacun des documents du profil :

$$\sum_{j=1}^{n} f_{ij} = 1$$

Cette particularité rend instable la résolution du système linéaire associé (2), nous avons donc opté pour la suppression de la dernière colonne ($j=n$) car la valeur f_{in} peut être déduite des précédentes par complément à 1.

2.2 - Méthode du maximum de vraisemblance

Comme précédemment, nous réalisons n mesures de fréquences relatives sur chacun des m documents du profil et nous devons obtenir les m résultats escomptés suivants:

$$v_i \quad i=1,m$$

Nous représenterons cette série d'évaluations par un point V de l'espace R^m.

$$V=(v_1,v_2,\cdots,v_m)$$

Nous avons donc dans R^m une fonction de répartition:

$$d=f(V)=f(v_1,v_2,\cdots,v_m)$$

De plus, cette fonction de répartition dépend des $n+1$ paramètres λ_j du modèle affine optimal que nous essayons de déterminer (un pour chaque fréquence relative calculée, un autre pour le terme constant) :

$$d=d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)$$

La probabilité de réalisation de V est:

$$d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)dv=d(v_1,v_2,\cdots,v_m;\lambda_0,\lambda_1,\cdots,\lambda_n)dv_1\,dv_2\cdots dv_m$$

Nous cherchons la probabilité de réalisation des m expériences en sachant qu'elles sont indépendantes et identiquement distribuées:

$$Prob(v_1\cap v_2\cap\cdots\cap v_m)=\prod_{i=1}^{m} Prob(v_i)$$

$$Prob(v_i)=f(v_i,\lambda_0,\lambda_1,\cdots,\lambda_n)dv_i$$

$$\prod_{i=1}^{m} Prob\ (v_i)=\prod_{i=1}^{m} f(v_i,\lambda_0,\lambda_1,\cdots,\lambda_n)\prod_{i=1}^{m} dv_i$$

La méthode du maximum de vraisemblance consiste à maximiser la probabilité de réalisation de la série d'expériences V. Pour cela nous pouvons maximiser le logarithme de cette probabilité afin de transformer les produits en sommes.

<u>Remarque</u> : cadre des hypothèses de l'analyse de la variance:

Nous allons supposer que les erreurs commises sur les expériences par le modèle affine vérifient les propriétés suivantes:

1. $Esp(e_i)=0$,
2. Les e_i sont des variables aléatoires indépendantes,
3. les e_i ont même variance σ^2.

Nous pouvons donc en déduire que les e_i ont un comportement de loi normale centrée de variance σ^2: $N(0,\sigma^2)$.

Si l'approximation par le modèle linéaire s'écrit:

$$v_i = \lambda_0 + \sum_{j=1}^{n} \lambda_j f_{ij} + e_i \quad \forall i = 1, m$$

On en déduit que le v_i suivent aussi une loi normale:

$$v_i \approx N(\lambda_0 + \sum_{j=1}^{n} \lambda_j f_{ij}, \sigma_2)$$

La fonction de répartition des v_i peut alors s'écrire sous la forme:

$$f(v_i, \lambda_1, \lambda_2, \cdots, \lambda_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(v_i - \lambda_0 - \sum_{j=1}^{n} \lambda_j f_{ij})^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{e_i^2}{2\sigma^2}}$$

$$L = \frac{1}{(\sigma\sqrt{2\pi})^m} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{m} e_i^2}$$

$$S = Log(L) = -m Log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma_2}\sum_{i=1}^{m} e_i^2$$

Le problème consiste alors à trouver le maximum de *S*, ce qui revient à trouver le minimum :

$$Min(\sum_{i=1}^{m} e_i^2)$$

car:

$$-m\, Log(\sigma\sqrt{2\pi}) \quad \text{est une constante,} \quad \sum_{i=1}^{m} e_i^2 \geq 0 \quad \text{et} \quad \frac{1}{2\sigma^2} \geq 0$$

Nous sommes donc conduits à résoudre le même problème qu'avec la méthode des moindres carrés. Les estimateurs obtenus par ces deux méthodes sont donc rigoureusement identiques.

3 - PRINCIPE GENERAL DE L'OUTIL DE FILTRAGE

3.1 - Constitution du profil initial

Le principal problème d'un procédure de filtrage se situe lors de la phase de démarrage. En effet, il est difficile de définir à priori quel seront les représentants les plus pertinents (documents ou mots-clés) qui seront capables de bien cibler le sujet et surtout quels seront les éléments non pertinents qui permettront d'éliminer une part importante du bruit (faux amis, sujets voisins, domaines d'application indésirables, …). Une collection hétérogène initiale reste donc à trouver pour permettre le démarrage du processus, mais les premiers filtrages sont souvent catastrophiques, car toutes les sources d'erreurs n'ont pas été prévues. Le mieux est d'analyser une collection assez vaste et de sélectionner des documents qui ne sont pas voisins, afin de recouvrir aux mieux l'espace des possibilités. Cette phase peut parfaitement être réalisée depuis notre plate-forme Tétralogie. Mais, comme dans le cas d'un estimateur, le nombre de documents doit être supérieur au nombre de termes utilisés dans le modèle, nous n'hésitons pas à prendre une collection assez vaste (une bonne centaine de documents semble un point de départ honorable).

3.2 - Choix des termes discriminants

Dans notre cas, nous devons trouver un équilibre entre les terminologies positives (corrélées aux document acceptés) et négatives (corrélées aux documents rejetés). Une zone d'incertitude doit aussi être aménagée afin de conserver certains termes présents dans les deux groupes : on évite ainsi une source d'instabilité de la méthode. Pour effectuer ce choix nous pouvons utiliser notre plate-forme Tétralogie. Une première étape consiste à extraire toute la terminologie du profil en évitant les mots vides. Ensuite chaque document est analysé pour déterminer la fréquence absolue de tous les termes retenus. Grâce au tableur on peut extraire la matrice des fréquences relatives et proposer une liste de termes caractéristiques de chaque groupe (acceptés ou rejetés) et de termes caractéristiques du profil sans discernement. La concaténation des trois listes est dédoublonnée et tronquée afin que le nombre de termes retenus soit inférieur au nombre de documents du profil. La meilleure solution est de prendre les termes les plus fréquents en évitant les termes équirépartis sur l'ensemble des documents.

3.3 - Filtrage des nouveaux documents

Une fois l'estimateur calculé (fonction affine des fréquences relatives), il suffit de calculer ces fréquences pour chaque document à filtrer et d'appliquer la fonction au vecteur associé. La valeur de la fonction est réelle mais elle s'approche soit du critère d'acceptation soit du critère de rejet.

Dans le cas du tout ou rien, les valeurs cibles sont +1 et –1, l'incertitude est donc au zéro. Suivant le nombre de documents acceptés on peut décaler la frontière en fonction de la capacité de lecture de l'utilisateur. Pour une utilisation électronique des documents filtrés, le problème ne se pose pas, il convient donc de tout garder au dessus de zéro.

Dans le cas d'une échelle de niveaux de pertinence, la restitution peut être plus nuancée. Les documents peuvent être triés ou classés en fonction de leur pertinence et toujours des capacités ultérieures de traitement. Il convient, dans ce cas, de conserver certains documents peu pertinents afin d'alimenter le profil pour bien tenir compte de la répartition en niveaux.

3.4 - Réactualisation du profil

Toute procédure de filtrage nécessite une phase d'apprentissage qui, si elle est bien menée, doit rapidement faire remonter l'efficacité du processus. Donc, à terme, le profil sera plus efficace qu'au départ, car les documents indésirables sont systématiquement ajoutés au profil ce qui permet de ne pas renouveler certaines erreurs. De même pour les documents les plus pertinents et dans le cas de niveaux de pertinence pour certains documents intermédiaires. Cette réactualisation du profil permet une auto-adaptation du processus de filtrage et donc un suivi de l'évolution du domaine ou des préoccupations de l'utilisateur. La procédure n'est donc pas figée dans le temps mais, au contraire, s'adapte aux fluctuations et aux ruptures dues souvent à la source elle même, au sujet, à l'utilisateur ou tout simplement à l'environnement (innovation, nouveaux acteurs, nouvelles applications, interfaçage avec d'autres domaines, évolution de la terminologie, maturité du sujet, …). Le principale difficulté est la participation active de l'utilisateur, qui doit réinjecter périodiquement dans le profil les documents les plus représentatifs soit de la variabilité du sujet, soit de nouvelles erreurs d'identification. De nouveaux termes peuvent aussi être élus et pris en compte dans la nouvelle définition du profil. Ce n'est pas une difficulté puisque le nombre de documents augment, le nombre de termes peut suivre. Bien sûr le système linéaire à résoudre est de plus en plus grand, mais ce n'est pas non plus un problème vu la puissance actuelle des processeurs. Le filtrage lui même ne nécessite que le calcul d'un produit scalaire par document, le plus difficile reste donc l'extraction des fréquences à la volée.

3.5 - Nouveau calcul de l'estimateur

Si la liste des termes retenus pour le modèle est reconduite, il suffit de déterminer la solution d'un nouveau système linéaire dont l'expression tient compte des nouveaux documents introduits dans le profil. Par contre, si la liste des termes est remise en cause (nouveaux mots-clés, nouvelles analyse du contenu sémantique), le nombre de paramètres va changer, l'ordre du système aussi. Dans les deux cas, il faut alors veiller à nettoyer le profil de documents devenus inutiles et il est recommandé d'effectuer une réévaluation de la pertinence du filtre (sur d'anciens document jugés) et éventuellement de réévaluer les documents qui avaient été préalablement rejetés et qui maintenant peuvent être acceptés par le nouveau profil. On peut ainsi récupérer des documents intéressants qui avaient été mal jugés précédemment (signaux faibles, émergences terminologiques, front de recherche) car le profil ne bénéficiait pas encore de la nuance sémantique du dernier apport.

4 - VALIDATION DE LA METHODE

4.1 - Base de tests utilisée

Nous avons utilisé une base de test très connue dans le monde du filtrage d'information : la base TREC. Nous avons généré plusieurs fichiers constitués d'environ 1800 documents déjà classés par les experts du domaine en documents pertinents et non pertinents. Pour représenter chaque document, nous ne gardons, dans un premier temps, que les 250 termes significatifs dont les fréquences sont les plus fortes Nous avons simulé, sur l'ensemble des documents de chaque série, l'évolution progressive d'un profil de filtrage. Au départ, le profil est initialisé par les 300 premiers documents du fichier, dont une faible partie est constituée de documents pertinents. Un premier estimateur est calculé, il est ensuite appliqué soit à la totalité des documents (y compris ce qui ont servi à calculer l'estimateur), soit appliqué aux documents restants. L'efficacité du filtrage est alors évaluée par les indicateurs que nous allons décrire ci-dessous. Nous supposons que l'estimateur est recalculé dès qu'une série de 100 documents supplémentaire a été jugée. Pour les besoins du test, nous avons placé la même quantité de documents pertinents dans chaque tranche complémentaire de 100 documents. Nous pouvons donc connaître l'évolution des performances du filtrage sans que cette mesure dépende du nombre de nouveaux documents pertinents introduit à chaque fois dans le calcul de l'estimateur.

4.2 - Indicateurs utilisés

Dans les problèmes de filtrage, deux objectifs antagonistes sont à prendre en compte :

- Trouver un maximum de documents pertinents ou la part maximum
- Dans les documents récupérés la part de documents pertinents doit être maximum

Dans le premier cas, on cherche à obtenir le maximum d'information même s'il y a du bruit, dans le second cas, on ne veut pas de bruit dans les documents proposés quitte à en rater un nombre plus important que précédemment.

Pour atteindre simultanément ces deux objectifs, il faut donc réaliser un compromis et trouver un indicateur fiable permettant de le mesurer.

Nous allons utiliser les notations suivantes :

- Les documents pertinents (+)
- Les documents non pertinents (-)
- Les documents bien jugés (*V*)
- Les documents mal jugés (*F*)

Nous allons définir quatre mesures :

- Le nombre de vrais positifs, documents pertinents (+) bien jugés (*V*) : V_+
- Le nombre de vrais négatifs, documents non pertinents (-) bien jugés (*V*) : *V*-
- Le nombre de faux positifs, documents pertinents (+) mal jugés (*F*) : F_+
- Le nombre de faux négatifs, documents non pertinents (-) mal jugés (*F*) : *F*-

Le premier objectif est satisfait si :

$$\text{Le taux de } \textbf{rappel } R_+ = V_+ /(V_+ + F_+) \text{ est maximum}$$

Le second si :

$$\text{Le taux de } \textbf{précision } P_+ = V_+ /(V_+ + V\text{-}) \text{ est maximum}$$

Une autre mesure utilisée est le taux de documents bien jugés soit :

$$\tau = (V_+ + V\text{-}) /(V_+ + V\text{-} + F_+ + F\text{-})$$

Dans notre cas qui est celui de la veille, l'objectif n'est pas de lire les documents filtrés mais de les analyser, il faut donc privilégier le taux de rappel. Par contre, si l'utilisateur final est lecteur et a donc une capacité de lecture limitée, il vaut mieux privilégier le taux de précision.
Pour valider notre démarche, nous avons défini un élément qui peut être paradoxal : le taux de rappel négatif. En l'associant au taux de rappel positif, nous obtenons une mesure du compromis cherché : une sorte de taux moyen.

$$R=\sqrt{R_+ R_-}=\sqrt{\frac{V_+ V_-}{(V_+ + F_+)(V_- + F_-)}}$$

En effet dans ce cas, les documents pertinents sont biens trouvés et les documents non pertinents bien écartés. Il faut bien entendu que ce taux s'approche de 1. Pour un filtrage aléatoire des documents (un sur deux) les valeur des différent indicateurs sont indiqués dans le tableau ci-dessous (le taux ***R*** est alors de 50%).

4.3 - Quelques résultats

Dans le cas de la base TREC, le nombre de documents pertinents est très inférieur au nombre de documents non pertinents. Nous avons donc constaté une disproportion flagrante dans la prise en compte par l'estimateur de ces deux population. Afin de rétablir l'équilibre, nous pondérons ces deux populations. La première idée est de simuler deux populations égales. Nous pouvons donc artificiellement gonfler la population des documents pertinents pour l'amener au niveau de celle des documents non pertinents, il suffit pour cela de multiplier chaque équation valant +1 par le coefficient suivant :

$$P = (F_+ + F\text{-}) / (V_+ + V\text{-})$$

Les erreurs commises sur ces équations étant plus fortes, l'estimateur tient mieux compte des documents positifs, leur reconnaissance est ainsi améliorée, par contre le jugement sur les documents négatifs peut diminuer en qualité.

Comme cette pondération est un peu empirique, nous avons aussi essayé une sous pondération par la racine de P et une sur pondération par P à la puissance 3/2. Donc une série de quatre expériences :

- P_0 sans pondération
- $P_{1/2}$ sous pondération
- P_1 pondération normale
- $P_{3/2}$ sur pondération

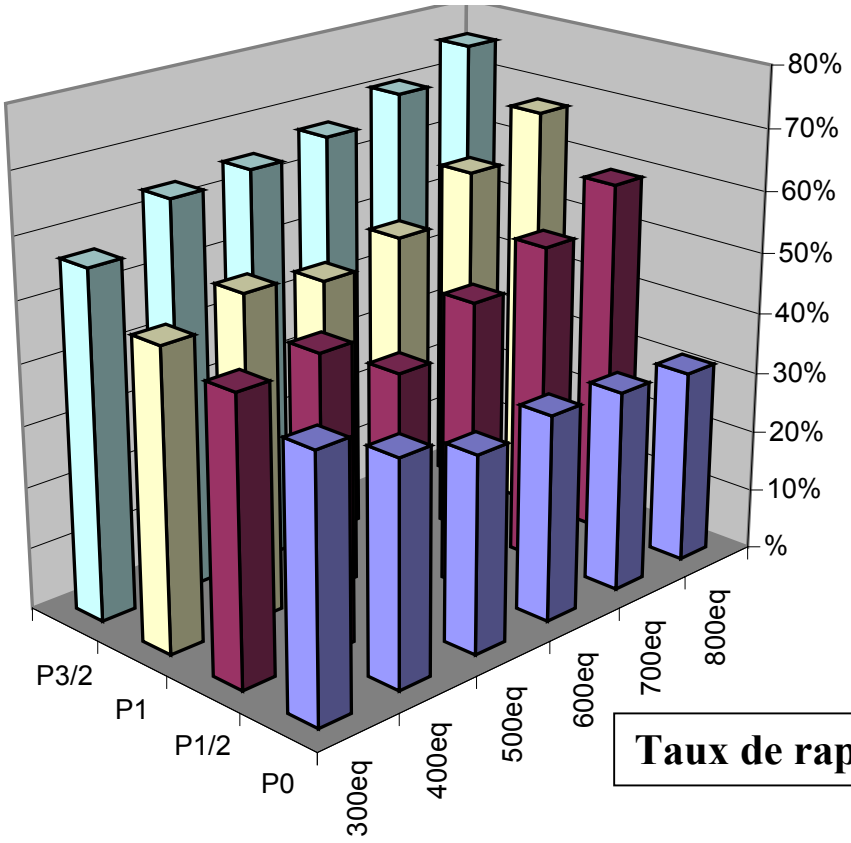
Les différents indices sont consignés dans les tableaux suivants :

256		**P0**			**P 1/2**			**P 1**			**P 3/2**	
1805	21%		J-	24%	J+	J-		J+	J-	28%	J+	J-
300	M+	108	1311	M+	118	1253	M+	125	1194	M+	145	1151
	M-	238	148	M-	296	138	M-	355	131	M-		111
	31%	0,6		29%	0,61	46%	26%	0,61	49%	27%	0,65	57%
	15%	J+	J-	18%	J+	J-	22%	J+	J-	25%	J+	J-
400	M+	93	1445	M+	122	1351	M+	135		M+	163	1196
	M-	104	163	M-	198		M-	273		M-	353	93
		0,58	36%		0,64	48%		0,66	53%	32%	0,7	64%
		J+	J-	15%	J+	J-	18%	J+	J-	25%		J-
500	M+	82			103	1430	M+	130	1351	M+	166	1186
	M-	48	174	M-	119	153	M-		126		363	90
		0,56	32%	46%	0,61	40%	40%	0,67		31%	0,7	65%
	12%	J+		15%		J-	19%	J+	J-	27%	J+	J-
600	M+		1507	M+		1417	M+	138	1319	M+		1140
		42	170	M-	132	135	M-	230	118	M-	409	85
	67%	0,57		48%	0,66		38%	0,68		29%	0,7	
	11%	J+	J-	13%	J+	J-	19%	J+	J-	28%	J+	J-
700	M+	84	1521	M+	134	1428	M+	156	1315	M+	181	1126
	M-	28	172	M-	121	122	M-	234	100	M-	423	75
	75%	0,57	33%	53%	0,69	52%	40%	0,72	61%	30%	0,72	71%
	11%	J+	J-	13%		J-		J+	J-	27%	J+	J-
800	M+	81	1524	M+	151	1415	M+	173	1284	M+	194	1122

	M-	25	175	M-		105	M-	265	83	M-	427	62
	76%	0,56	32%	53%	0,73	59%	39%	0,75	68%	31%	0,74	76%
τ	50%	J+	J-	τ	**J+**	**J-**					**R**	
Aléa	M+	128	775	**M+**	**V+**	**V-**	**Avec intégration de l'estimateur**					
	M-	775	128	**M-**	**F-**	**F+**						
P+		0,5	50%	**R+**								

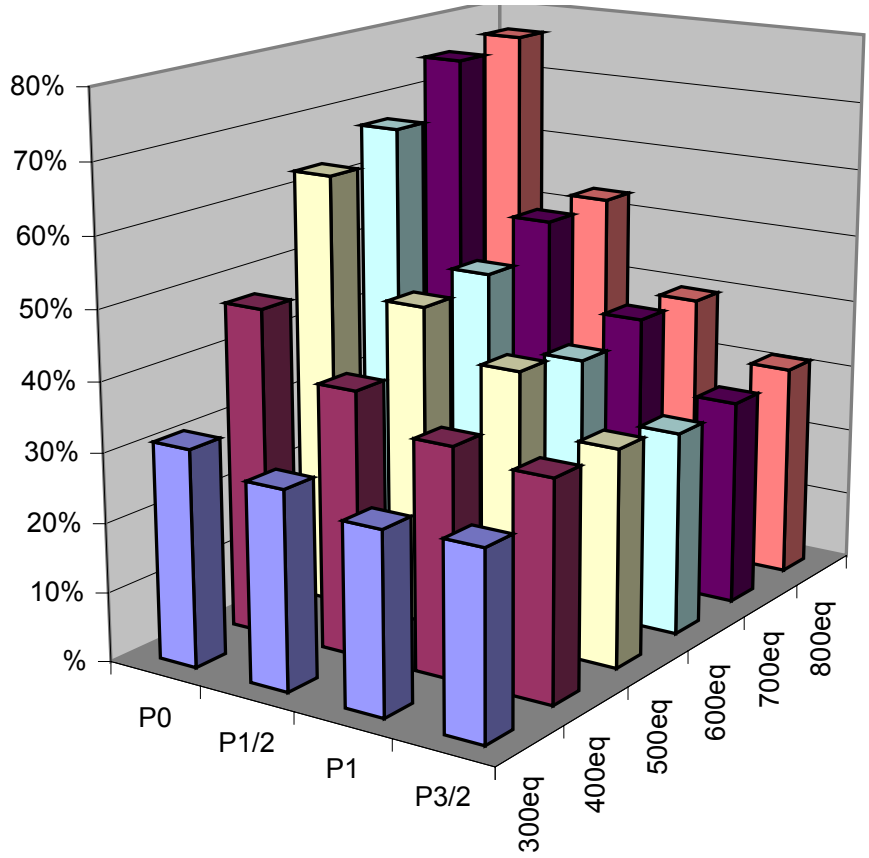
Nous pouvons remarquer que les bons scores de précision sont obtenus pour un faible taux de rappel et inversement. Par contre le taux moyen maximum est obtenu pour des scores intermédiaires de ces deux indicateurs. L'erreur globale quant à elle suit le taux de précision car le nombre de documents pertinents est très faible.

Ci-dessous, les évolutions du taux de rappel et du taux de précision en fonction de la taille de l'échantillon (profil) et de la pondération corrigeant le déséquilibre entre documents pertinents et non pertinents pris en compte dans le calcul de l'estimateur.

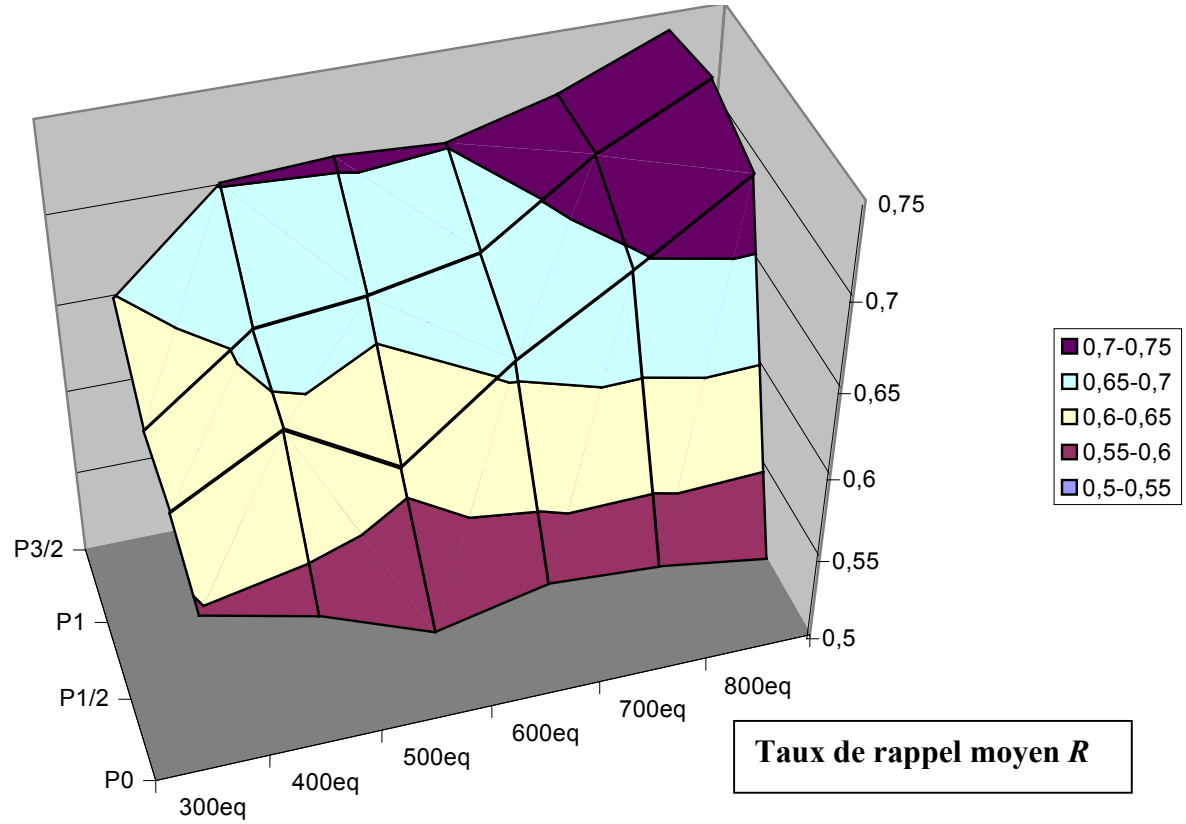


Taux de rappel R_+

Taux de précision P_+



Nous pouvons aussi constater que le taux moyen de réussite du filtrage augment avec le nombre de documents pris en compte du moment que ceux-ci sont pondérés pour égaliser exactement le poids des deux populations. Dans notre cas il n'y en en tout que 256 documents pertinents sur 1805, et dans le calcul des estimateurs leur nombre varie de 40 pour 300 documents à 90 pour 800 soit 10 de plus pour 100 nouveaux documents intégrant le profil.



4.4 - Intégration de l'outil

Pour l'instant, cet outil de filtrage fonctionne sur station SUN sous Solaris. Nous comptons prochainement l'introduire dans notre plate-forme Tétralogie (Unix et Linux) afin de nettoyer certains corpus du bruit dû à des équations de recherche trop larges (Internet, sujets difficiles à cerner, corpus généralistes, résultats de push en veille automatique). Une seconde utilisation peut être la recherche de sous corpus liés à un résultat d'analyse (production d'une équipe de recherche non isolée, ciblage d'un nouveau concept, d'une rupture, activité d'une communauté, environnement d'un sous domaine). En effet, ces entités sont un peu floues et ne peuvent pas être parfaitement identifiées par une simple liste de termes liés par une équation booléenne. L'idée est donc de partir de certains documents identifiés et d'en tirer toute la classe des documents voisins si possible triés par pertinence.

CONCLUSION

Parmi les différentes techniques de filtrage développées dans notre laboratoire (Vigie, Mercure), l'outils que nous proposons, basé sur l'algèbre linéaire, est plus conforme à la philosophie générale de notre plate-forme Tétralogie très orientée analyse de données à la française et donc très proche elle aussi de l'algèbre linéaire. Comme les premiers test sont particulièrement encourageants, nous comptons évaluer cet outil selon les mêmes conditions que les autres et offrir ainsi un alternative aux techniques statistiques et neuronales. Maintenant que le principe d'un estimateurs semble acquit, il nous faut améliorer et automatiser en partie les méthodes de sélection des termes les plus pertinents dans le cadre de cette approche. Faire évoluer le profil sera alors une tâche plus aisée, puisque le seul travail à la charge de l'utilisateur consistera à sélectionner de nouveaux documents pertinents en vue d'alimenter le profil. Les documents non pertinents quant-à eux pourront être sélectionnés pour entrer dans le profil en fonction des valeurs de leur fonction d'évaluation (les plus mauvais scores sont à privilégier afin de réaliser un recadrage efficace). Enfin, pour une utilisation sur la plate-forme, nous comptons introduire une alimentation mixte du profil : dictionnaires de termes isolés (listes d'acteurs ou de mots-clés) et collection de documents repérés (externes ou issus du corpus à filtrer).

BIBLIOGRAPHIE

[DOUS99] B. Dousset, M. Salles

La Veille Scientifique par l'Analyse des Informations Ouvertes. 26th International Conference, Information Technologies in Science, Education and Business, (Yalta-Gurzuf, Crimée, Ukraine), may 17-30 1999.

[KARO99] S. Karouach, T. Dkaki, B. Dousset

Visualisation interactive de classifications d'informations. 8ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, p. 45-61, 27 septembre-1° octobre 1999.

[ROUX99] C. Roux , B. Dousset

Une méthode de détection des signaux faibles: application à l'émergence des Dendrimères. Veille stratégique, scientifique et technologique : VSST'98, pp 349-357, (Toulouse, France), octobre 1998.

[DKAK00] T. Dkaki, B. Dousset, D. Egret, J. Mothe

Information discovery from semi-structured sources. Application to astronomical literature. Computer Physics Communications, Eds: Elsevier Science, V. 127 N° 2-3 , pp 198-206, 2000.

[DOUS00] B. Dousset, T. Dkaki, J. Mothe

Information mining in order to graphically summarize semi-structured document. 17th international CODATA Conference, (Baveno Italie), 15-19 octobre 2000.

[HUBE00] G. Hubert, J. Mothe, A. Benammar, T. Dkaki, B. Dousset, S. Karouach

Textual document Mining using graphical interface. International Human Computer Interaction, HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey, pp 918-922 (volume 1), 05-10 août 2001.

[SALL00] M. Salles, Ph. Clermont, B. Dousset

MEDESIIE : une méthode de conception de systèmes d'intelligence écono-mique. IDMME'2000, (Montréal Canada), 16-19 mai 2000.

[BOUG01] M. Boughanem, B. Dousset

Relation entre le push adaptatif et l'optimisation des abonnements dans les centres de documentation. Veille stratégique, scientifique et technologique : VSST'01, pp 239-252, Vol 1, (Barcelone, Espagne), octobre 2001.

[KARO01] S. Karouach, B. Dousset

Visualisation interactive pour la découverte de connaissances. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 291-300, (Barcelone, Espagne), octobre 2001.

[MOTH01] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret

Information mining: use of the document dimensions to analyse interactively a document set. 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

[MULT01] J.-L. Multon, G. Lacombe, B. Dousset

Analyse bibliométrique des collaborations internationales de l'INRA. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

[SOSS01] D. Sosson, M. Vassard, B. Dousset

Portail pour la navigation en ligne dans les analyses stratégiques. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

[TMAR01] M. Tmar

Apprentissage incrémental dans un système de filtrage adaptatif. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 313-320, (Barcelone, Espagne), octobre 2001.

[DOUS02] B. Dousset, S. Karouach

Collaboration interactive entre classifications et cartes thématiques ou géographiques. 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.

VISUALISATION DE RELATIONS PAR DES GRAPHES INTERACTIFS DE GRANDE TAILLE

KAROUACH Saïd,
karouach@irit.fr

DOUSSET Bernard
dousset@irit.fr
IRIT-SIG, Université Paul Sabatier, 118, route de Narbonne
31062 Toulouse cedex 04
Tél : 05.61.55.67.81

Résumé : Notre plate-forme Tétralogie dédiée à la découverte de connaissances implicites dans de grand corpus hétérogènes d'information électronique est essentiellement basée sur l'analyse relationnelle : entre les acteurs du domaine, les éléments terminologiques et le temps. Afin de mieux communiquer les résultats globaux ou partiels des ces analyses, nous développons actuellement des méthodes de visualisation interactive de graphes basées sur la notion d'attraction et de répulsion. Elle permettent à l'utilisateur de naviguer dans les données, en proposant des vues partielles ou simplifiées qui permettent de comprendre assez vite certaines organisations locales remarquables. Le couplage avec des méthodes de tri et de classification déjà présentes sur la plate-forme permet un pré-placement optimal des sommets ce qui conduit immédiatement à une meilleure lisibilité et à une découverte intuitive des relations les plus significatives.

Abstract : Our Tétralogie platform is dedicated to knowledge discovery in great heterogeneous electronic information corpus. It is primarily based on relational analysis: between the actors' field, terminological elements and time. In order to show total or partial results of these analysis, we develop some interactive graphs visualization techniques, based on attraction and repulsion concept. The user can navigate in data, by proposing partial or simplified views which help him to understand rather quickly some local organizations. The coupling with sort and/or classification methods already present on the platform, allows vertexes optimal preplacement which immediately leads to better legibility and intuitive discovery of the most significant relations.

Mots Clés : Veille scientifique et technique, Analyse relationnelle, Graphes, Visualisations interactives, Java, Internet, Intranet.

Keywords : Science and technology watch, Relational analysis, Graphs, Interactive vizualization, Internet, Intranet

VISUALISATION DE RELATIONS PAR DES GRAPHES INTERACTIFS DE GRANDE TAILLE

INTRODUCTION

Le graphe est largement utilisé comme moyen de représentation et de visualisation de données. Cette méthode de restitution est très appréciée par les utilisateurs, car elle ne nécessite pas de connaissances mathématiques particulières. Le tout est de trouver un graphe à la fois fidèle à la réalité et suffisamment lisible pour être exploité. Nous partons, soit d'une matrice relationnelle calculée au niveau de la plate-forme, soit d'un extrait de matrice récupéré, via un portail, depuis une table de base de données relationnelle (compilation d'une matrice issue de l'analyse initiale).

La matrice analysée appartient à un des types suivants :

- Présence absence (existence d'un lien dans au moins un document)
- Contingence (entre deux variables à item unique : journal, date, 1° auteur, langue)
- Co-occurrence (une des variables peut contenir plusieurs items : auteurs, mots-clés)

Cette matrice peut avoir subi un ou plusieurs traitements préalables :

- Tri par classes de simple connexité
- Tri par blocs diagonaux en mode absolu
- Tri par blocs diagonaux en mode relatif
- Normalisations de tous ordres
- Extraction de sous-graphes
- Seuillages
- Fusion d'éléments
- Elimination d'éléments ne générant pas de connexion

Le but étant toujours de trouver un modèle qui reflète le mieux possible la réalité.

1 - VISUALISATION DYNAMIQUE DE GRAPHES

1.1 - Principe général

A l'origine, Eades **[EADE 84]** comparait les liaisons dans les graphes à des ressorts en analogie avec la loi physique de Hook. Il associait les sommets à des masses et les arêtes à des ressorts reliant celles-ci. Un tel système engendre des forces entre les sommets ce qui entraîne des déplacements respectifs. Après une phase de transition, le système finira par se stabiliser. Eades supposait que le placement final des sommets pourrait correspondre à une configuration satisfaisante du graphe. L'algorithme d'Eades ajoute la notion de force de répulsion entre les sommets et une attraction matérialisée par les arêtes. La condition d'arrêt est simplement un nombre maximum d'itérations. L'analogie à la mécanique serait complète si la condition d'arrêt était fonction de l'énergie du système. L'équilibre est atteint pour un optimum énergétique du graphe. Cette démarche a fait l'objet de plusieurs développements **[KAMA 89], [FRUC 91], [FRIC 94]** conduisant à différents modèles dynamiques de type FDP (Force Directed Placement) dont voici l'algorithme général.

1.2 - Algorithme

Initialisation : Positionnement aléatoire des sommets
Tant que $k < Max\_iter$ **faire**
    Pour tout sommet *u* **faire**
        Pour tout sommet $v \mathrel{!}= u$ **faire**
        Si $distance(u ; v) < seuil$ **alors**
            Calcul de la force de répulsion entre u et v
    Pour toute arête *(u, v)* **faire**
        Calcul de la force d'attraction entre u et v

Pour tout sommet *u* **faire**
Cumul des forces
Déplacement de u en fonction de la température globale
Diminuer la température globale

Fin

1.3 - Forces d'attraction et de répulsion

La notion de la température globale du système a été introduite pour limiter le déplacement des sommets. Cette température diminue au cours du déroulement de l'algorithme ce qui implique que plus on approchera de l'équilibre , moins un sommet pourra se déplacer. En effet, le déplacement de tous les sommets est inversement proportionnel à la température globale.

Dans notre prototype, nous avons adopté la même démarche que celle décrite ci-dessus. Notre modèle diffère de celui-ci dans la définition des forces d'attraction et de répulsion. La force d'attraction entre deux sommets v_i et v_j est définie par :

$$f_a(v_i,v_j)=\beta_{ij}\, d_{ij}^{\alpha_a}/k$$

β_{ij} est le poids de l'arête (v_i, v_j), k est calculé en fonction de la surface de la fenêtre et du nombre de sommets du graphe, d_{ij} est la distance entre v_i et v_j sur le dessin. Si les sommets v_i et v_j ne sont pas reliées par une arête, alors $f_a = 0$.
La force de répulsion est donnée par :

$$f_r(v_i,v_j)=-k^2/d_{ij}^{\alpha_r}$$

la variable α_a (respectivement α_r) est une constante qui sert à définir le degré d'attraction (respectivement de répulsion) entre deux sommets.

Ce type d'algorithme de dessin de graphe donne de bons résultats pour des graphes relativement petits (100 sommets). Son utilisation devient très lourde pour des graphes de grande taille. Une solution consiste à transformer le graphe initial en une structure équivalente de taille moyenne. L'idée est alors de décomposer le graphe en sous-graphes (groupes), et d'appliquer ensuite l'algorithme de dessin sur le graphe des groupes. Ceci nécessite une technique de partitionnement de graphe efficace tenant compte de la taille du graphe initial.

1.4 - Partitionnement de graphe

La limitation du nombre de sommets à afficher améliore la clarté et augmente simultanément la rapidité d'exécution des tâches chargées du placement de ces sommets. Stijn van Dongen **[DONG 00a]** a introduit une technique de partitionnement de graphes de grande taille. Son algorithme MCL (Markov Cluster algorithm) est basé sur la simulation de l'écoulement stochastique dans un graphe. L'idée est de simuler plusieurs écoulements aléatoires dans le graphe, puis de renforcer l'écoulement là où il est déjà fort, et de l'affaiblir là où il est faible. Mathématiquement, l'écoulement est simulé par des opérations algébriques (**expansion** et **inflation**).sur la matrice stochastique (de Markov) associée au graphe. L'expansion de matrice correspond au calcul des probabilités des chemins aléatoires de plus grandes longueurs. Plus spécifiquement, l'expansion augmente l'écoulement par le calcul des puissances de la matrice stochastique, elle permet à l'écoulement de relier différentes parties du graphes, mais elle ne montrera pas la structure fondamentale des clusters. L'inflation favorise et rétrograde les probabilités des chemins dans le graphe. Autrement dit , cet opérateur sert à renforcer ou affaiblir l'écoulement là où c'est nécessaire.

Les étapes d'expansion et d'inflation s'effectuent alternativement sur la matrice de Markov jusqu'à ce qu'aucun changement ne puisse être détecté. La matrice de Markov est alors interprétée en tant que matrice résultat d'un regroupement. Pour plus de détails, nous conseillons de consulter les autres travaux de Stijn van Dongen **[DONG 00b, DONG 00c]**.

2 - RECHERCHE DE CLUSTERS ET DE CONNECTEURS

2.1 - Visualisation du graphe sans simplification

Avant d'avoir recours à l'arsenal des simplifications proposées en mode interactif, le mieux est de visualiser le graphe dans son intégralité. Très souvent, il est suffisamment lisible pour être interprété , mais un écran haute

résolution est bien entendu recommandé. Certaines manipulations élémentaires permettent ensuite de découvrir les caractéristiques principales de la structure relationnelle :

- Rechercher les éléments isolés
- Tirer sur un sommet pour évaluer ses liaisons
- Déplacer vers les bords de l'écran les structures homogènes
- Laisser agir le système d'attraction répulsion pour placer les sommets

Dans l'exemple ci-dessous, nous proposons une vue globale qui manque de lisibilité, mais qui avec un peu d'habitude peut déjà être exploitée en l'état. Les leaders du domaine sont plus colorés (matrices quantitatives), la densité des faisceaux dévoile les réseaux importants, la force des liens entre réseaux peut être évaluée par la lecture des valeurs des arêtes.

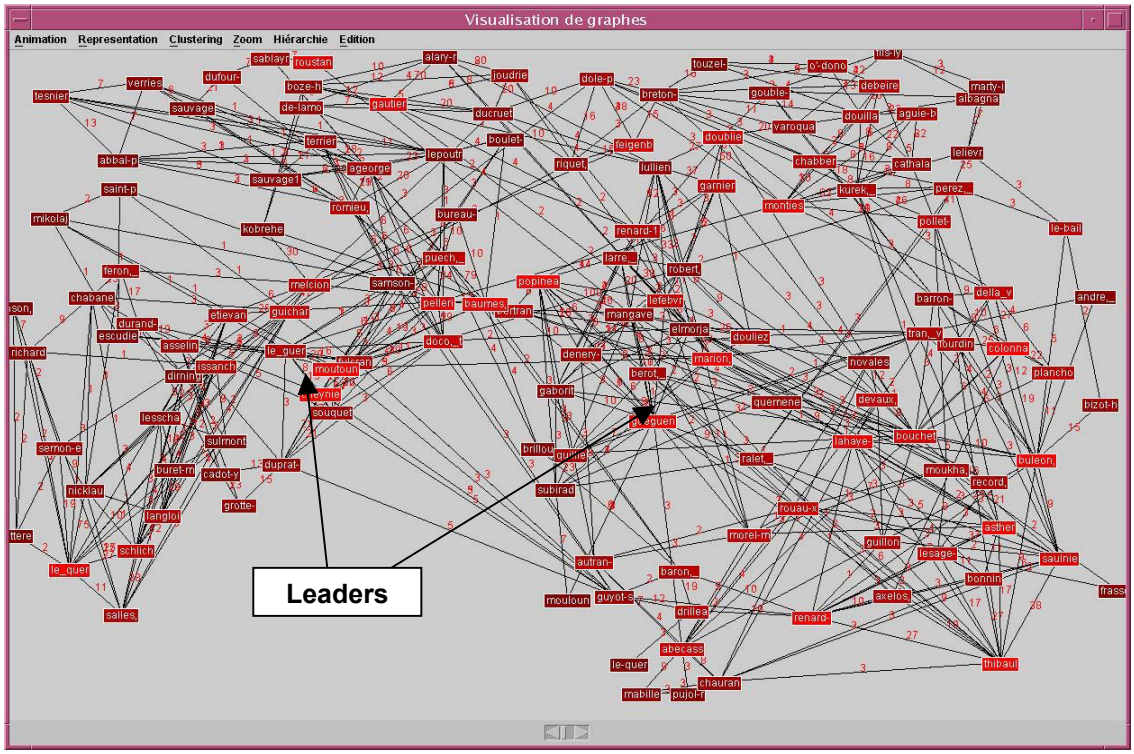


Figure 1. : Graphe de liens sans simplification : manque de lisibilité.

2.2 - Aide à l'analyse par simplifications interactives

Ci-dessous, une simplification par seuillage du graphe précédent qui permet de détecter les groupes isolés, ceux qui sont connectés et les individus qui servent d'interfaces. La complexité du graphe étant réglable, le processus de découverte peut se baser sur cette fonction pour faire apparaître progressivement les détails par abaissement du seuil une fois que les signaux fort ont été repérés et correctement répartis sur l'écran.

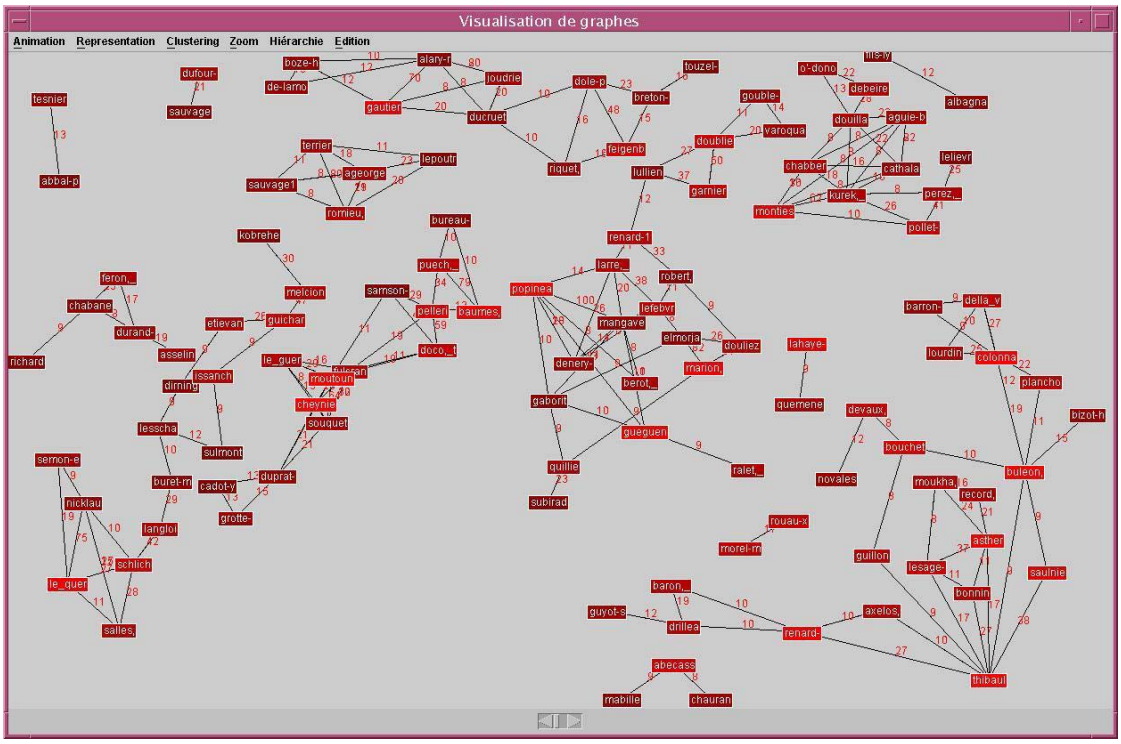


Figure 2. : Graphe obtenu après simplification : les éléments clés sont identifiés.

3 - CLASSES CONNEXES ET SOUS GRAPHES

3.1 - Tri par classes de connexité

Une matrice qu'elle soit carrée ou rectangulaire peut être triée en fonction de ses classes de simple connexité. Un zoom arrière en 2D fait alors clairement apparaître sa structure comme ci-dessous. Chaque classe peut alors être étudiée séparément afin d'en expliquer la structure et d'en découvrir les éléments clés : ici les connecteurs entre équipes de recherche.

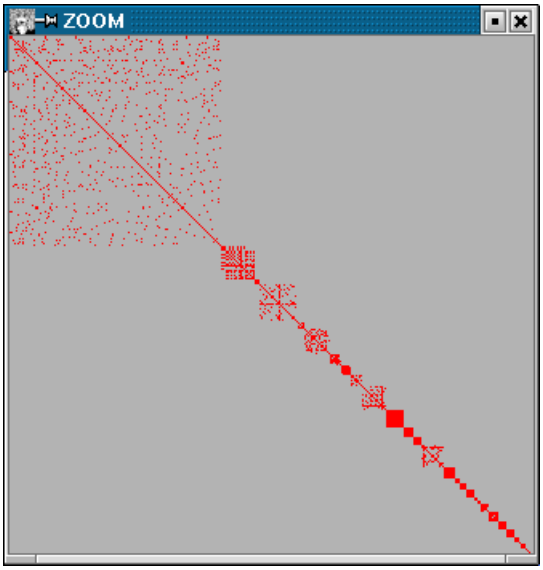


Figure 3. : zoom arrière 2D d'une matrice carrée Auteurs² triée par connexité.

Chaque classe est en fait un sous graphe totalement indépendant qui peut être visualisé et analysé à part. Ci-dessous les structures de deux équipes de la partie basse du zoom.

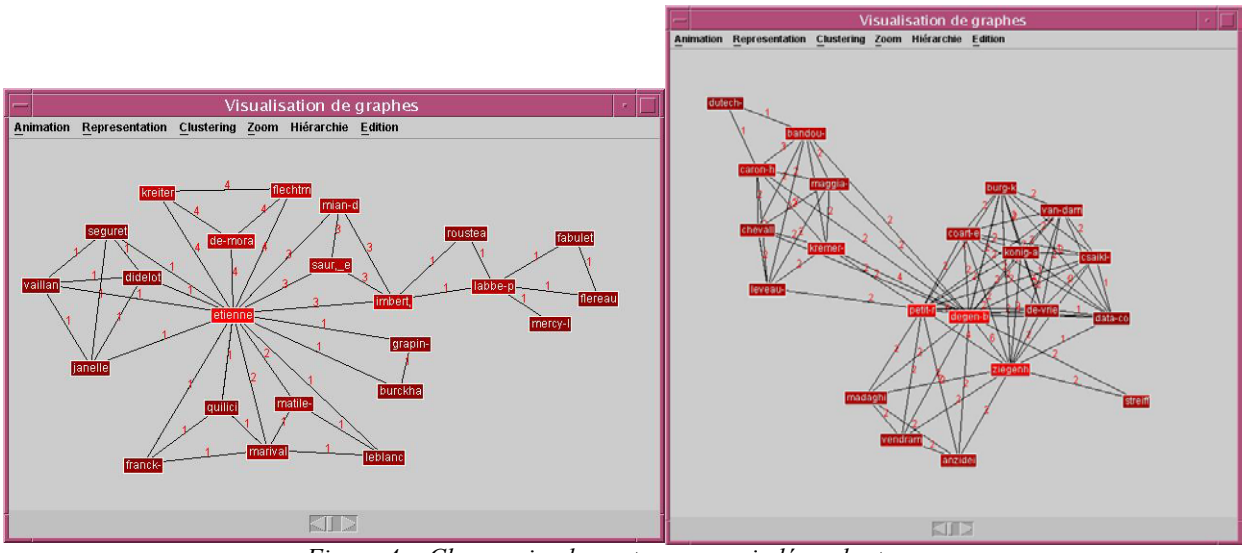


Figure 4. : Classes simplement connexes indépendantes.

3.2 - Extraction de sous graphes

La partie haute du zoom précédent représente un collaboratoire beaucoup plus important, qui est formé de plusieurs équipes cohérentes. Les collaborations entre ces équipes s'effectuent au travers de connecteurs qui sont parfaitement identifiables dans les graphe qui suit. Mais l'étude du fonctionnement interne de chaque équipe n'est pas facilitée par la densité de ce graphe. Le mieux est de couper les liens qui unissent une équipe à ses voisins sans perdre les connecteurs, donc extraire un sous graphe et l'étudier à part.

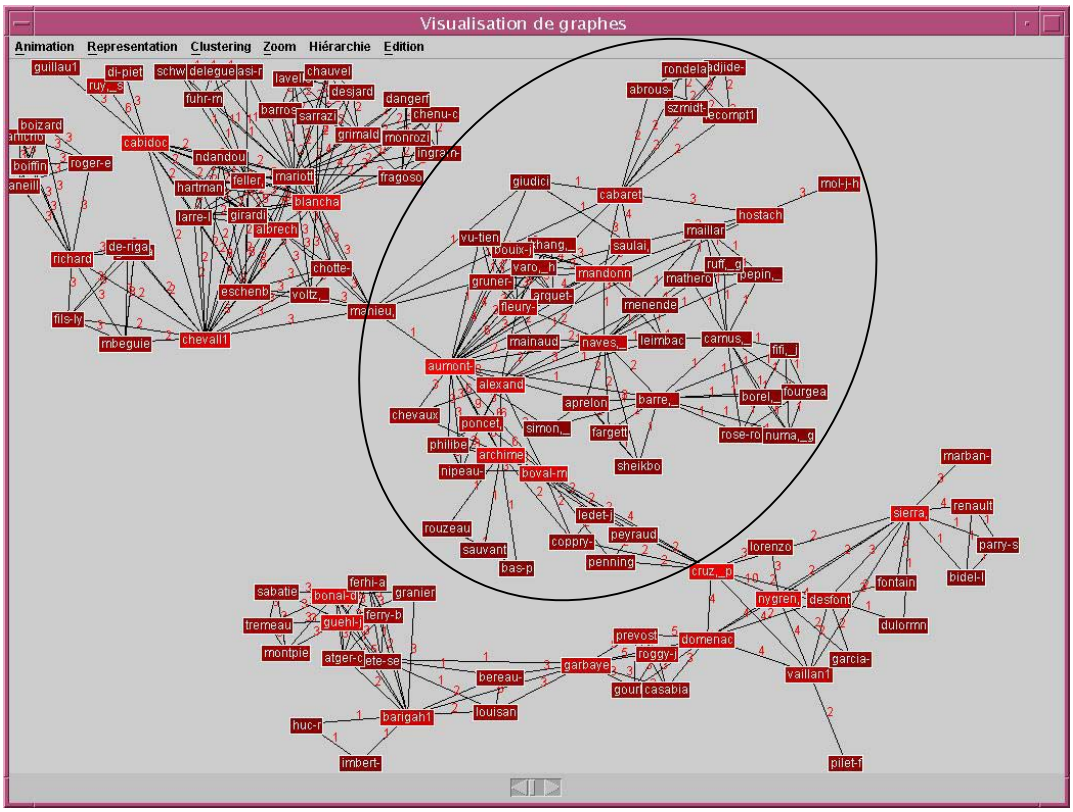


Figure 5. : Sous-graphes ou classes connexes indépendantes.

Pour un graphe biparti issu d'une matrice non symétrique, le problème est similaire : on peut étudier à part chaque composante simplement connexe, ou extraire des sous graphes en supprimant les liens qui les unissent au reste du graphe.

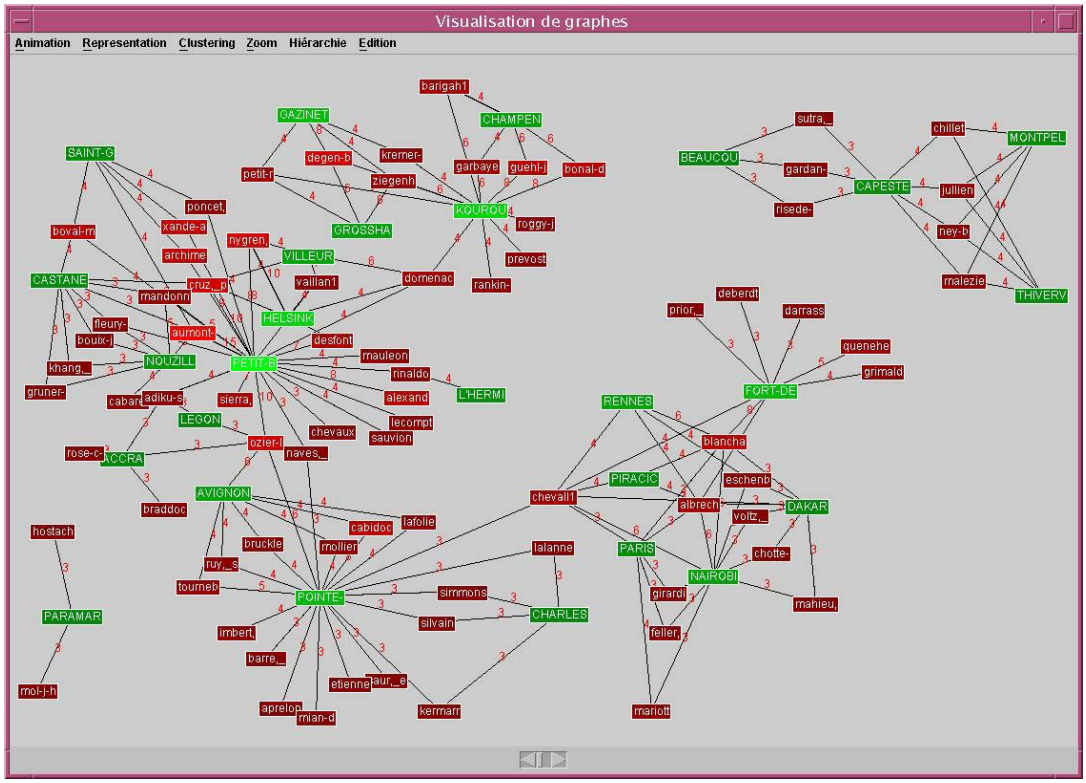
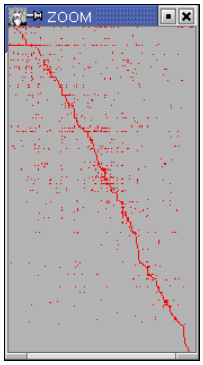


Figure 6. : Sous-graphes ou classes connexes indépendantes : graphe biparti.

Ci-dessus, un extrait d'une matrice Auteurs – Villes où des liens entre sous graphes sont alternativement dus à certains auteurs ou à certaines villes. Il est assez facile d'isoler quatre groupes en ne coupant qu'un ou deux liens. Les éléments isolés peuvent alors être listés afin de servir de filtre pour une analyse stratégique du groupe ainsi formé (par croisement avec toutes les autres informations disponibles : acteurs et thèmes)

4 - PLACEMENT DIRIGE DES SOMMETS



4.1 - Tri par blocs diagonaux

Afin de rendre plus lisible la visualisation de graphes complexes et connexes, nous avons tenu compte, dans le placement des sommets, de la sérialisation des items obtenue à l'issue d'un tri par blocs diagonaux. Les sommets sont alors distribués de façon circulaire, leurs principales connexions s'effectuent avec leurs voisins et seules les liaisons inter-clusters attirent maintenant l'attention. Ce principe est applicable aussi bien pour les matrices symétriques qu'asymétriques. Ci-contre, le zoom arrière 2D d'une matrice asymétrique ainsi réorganisée par permutation des lignes et des colonnes.

4.2 - Classification simultanée des lignes et des colonnes

Une autre méthode consiste à trier simultanément les lignes et les colonnes de la matrice. C'est par exemple possible après avoir réalisé une analyse factorielle des correspondances qui normalise les lignes et les colonnes et les plonge dans le même espace. Il suffit alors de placer les sommets du graphe dans l'ordre obtenu sur l'arbre planaire de classification hiérarchique en mettent les lignes à l'extérieur (plus nombreuses) et les colonnes à l'intérieur.

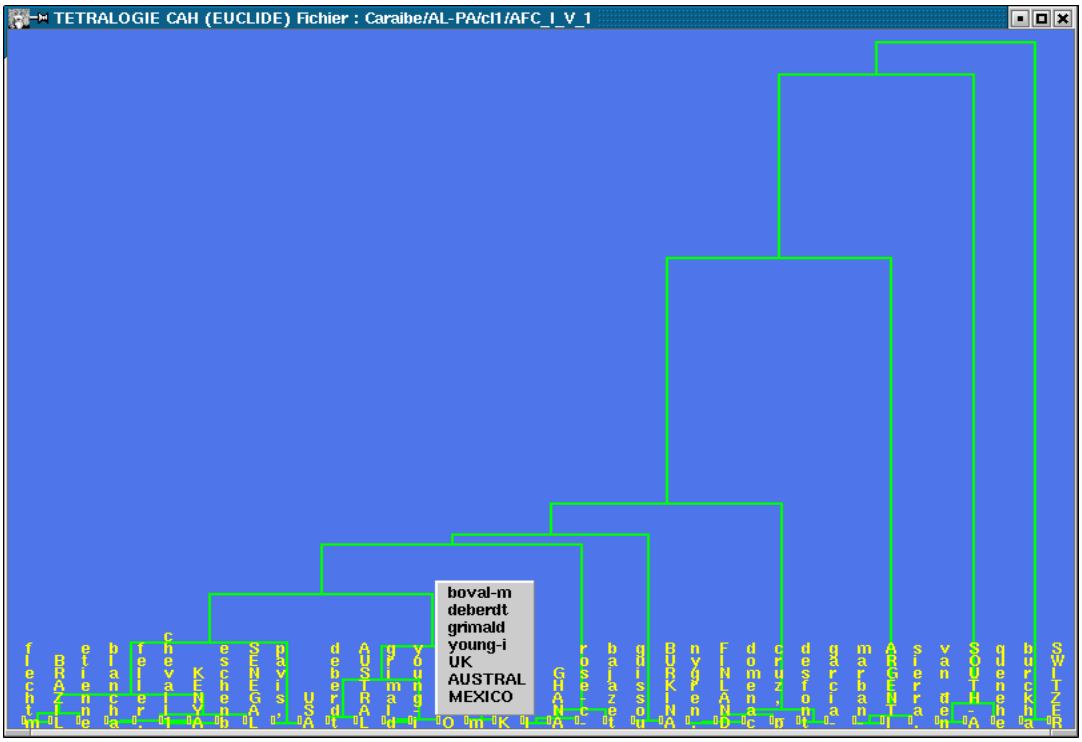


Figure 7. : Classification ascendante hiérarchique après AFC sur Auteurs Pays.

4.3 - Visualisation d'une matrice triée

Dans l'exemple suivant, où nous avons croisé des Auteurs et des Pays, les Auteurs (lignes de la matrice) sont disposés à la périphérie, les pays moins nombreux à l'intérieur. Nous pouvons immédiatement remarquer :

- Les pays totalement isolés,
- Les auteurs isolés dans un pays isolé,
- Les groupes d'auteurs qui collaborent à l'international,
- Les acteurs de ces collaborations à l'intérieur des groupes,
- Les pays liés au travers des auteurs,
- La force de ces liaisons,
- Les éléments clés (connecteurs uniques, plaques tournantes),
- Le leader de chaque groupe (par la coloration)
- La taille de chaque cluster qu'il soit ou non isolé.

En fait, les éléments remarquables (liens importants) sont ceux qui traversent la figure, ils représentent, ici, des collaborations internationales.

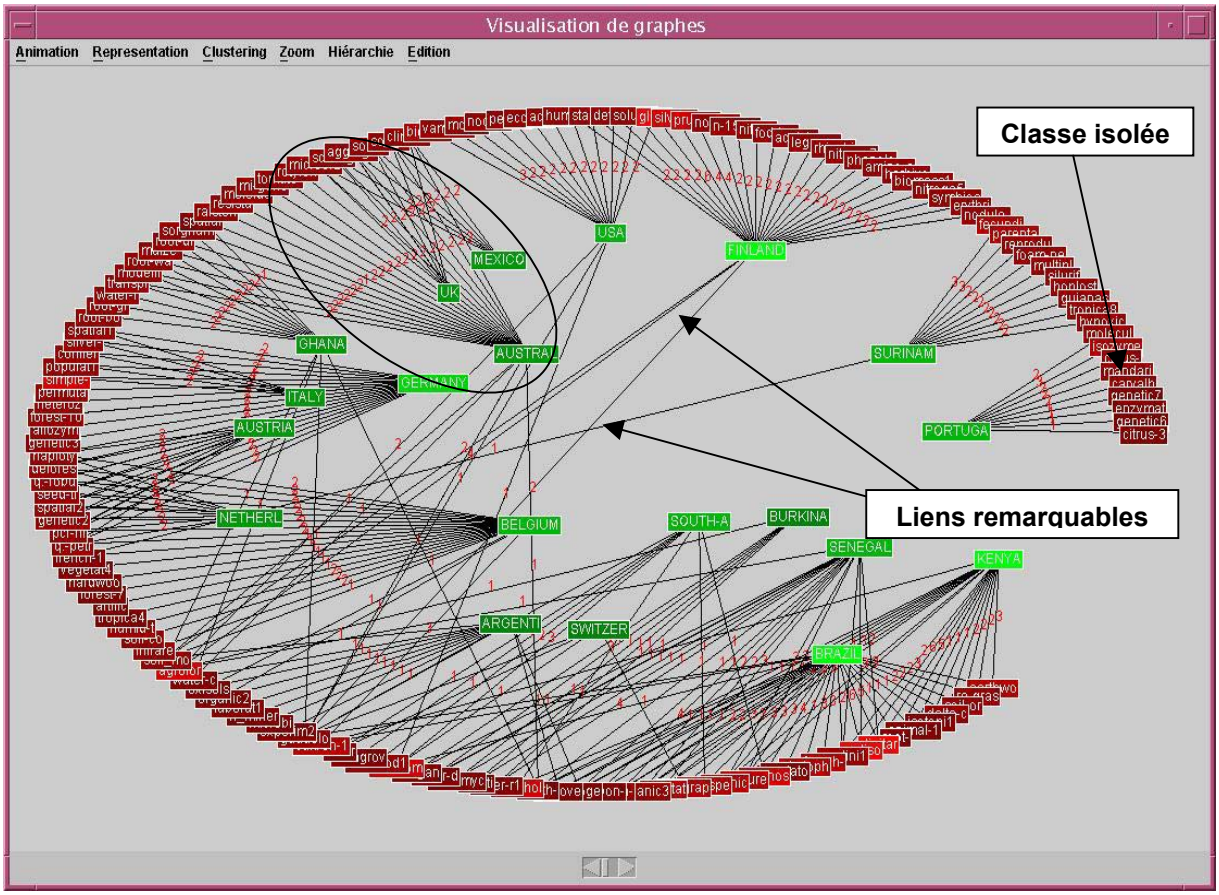


Figure 8. : Placement des items en fonction d'un tri par blocs diagonaux.

4.4 - Autre possibilité de partitionnement

D'autres méthodes de placement des sommets peuvent être issues des classifications proposées par Tétralogie, en particulier la classification par la méthodes des centre mobiles. Cette méthode propose en effet une partition en *n* classes effectuée à partir d'un représentant de chaque classe. Il est possible de choisir ces représentants directement sur le graphe comme on le fait déjà sur une carte factorielle en 3D. Une fois la partition réalisée les sommets sont placé de façon judicieuse en fonction de la taille de la fenêtre. Une autre utilité de ce partitionnement est l'extraction automatique de sous graphes plus faciles à analyser, cette possibilité prend toute son importance dès que le nombre de sommets dépasse plusieurs centaines.

5 - CAS PARTICULIER DES ARBRES PARTIELS

Une méthode radicale de simplification d'un graphe connexe est de ne conserver qu'un de ses arbres partiels extremum (ici maximum). Pour cela, il suffit, pour chaque sommet, de garder sa connexion la plus forte (première arête de chaque cocycle). Le graphe obtenu permet de savoir quels sont les éléments clés du domaine (connecteurs privilégiés). De plus, le placement des sommets obtenu à partir d'un de ses arbres planaires peut servir de base pour la reconstruction du graphe complet ou de l'une de ses simplifications moins poussées. En voici une illustration qui permet de juger de la lisibilité de cette méthode dans le cas d'une matrice asymétrique (graphe biparti).

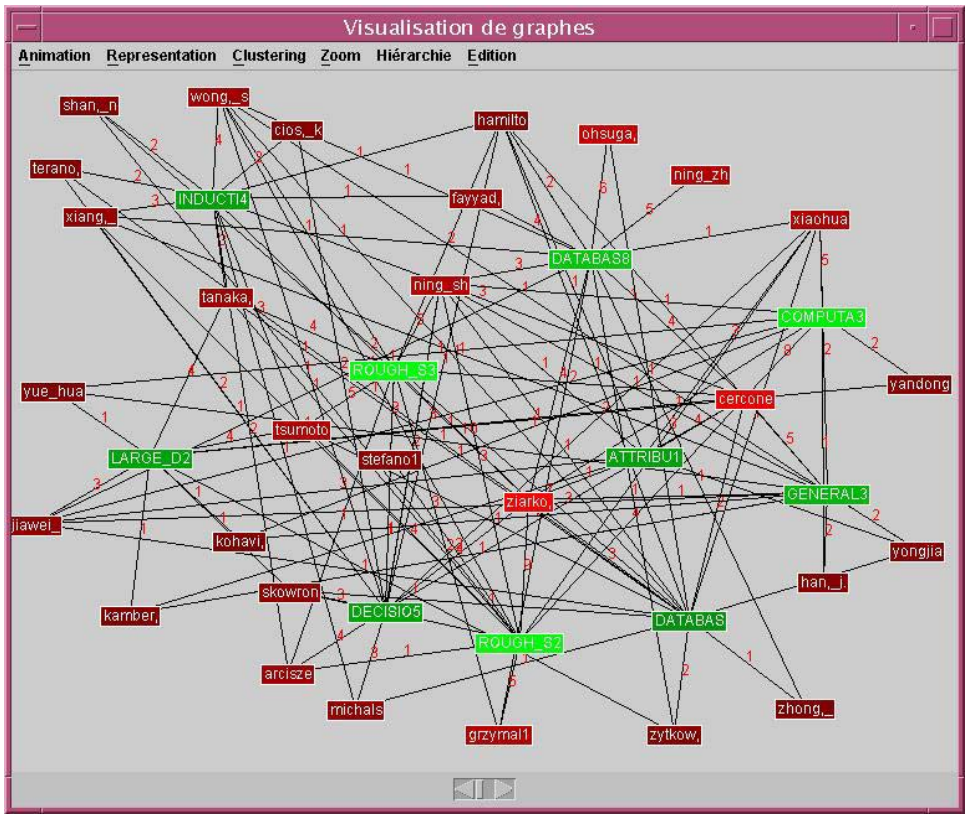


Figure 9. : Graphe simplement connexe de départ.

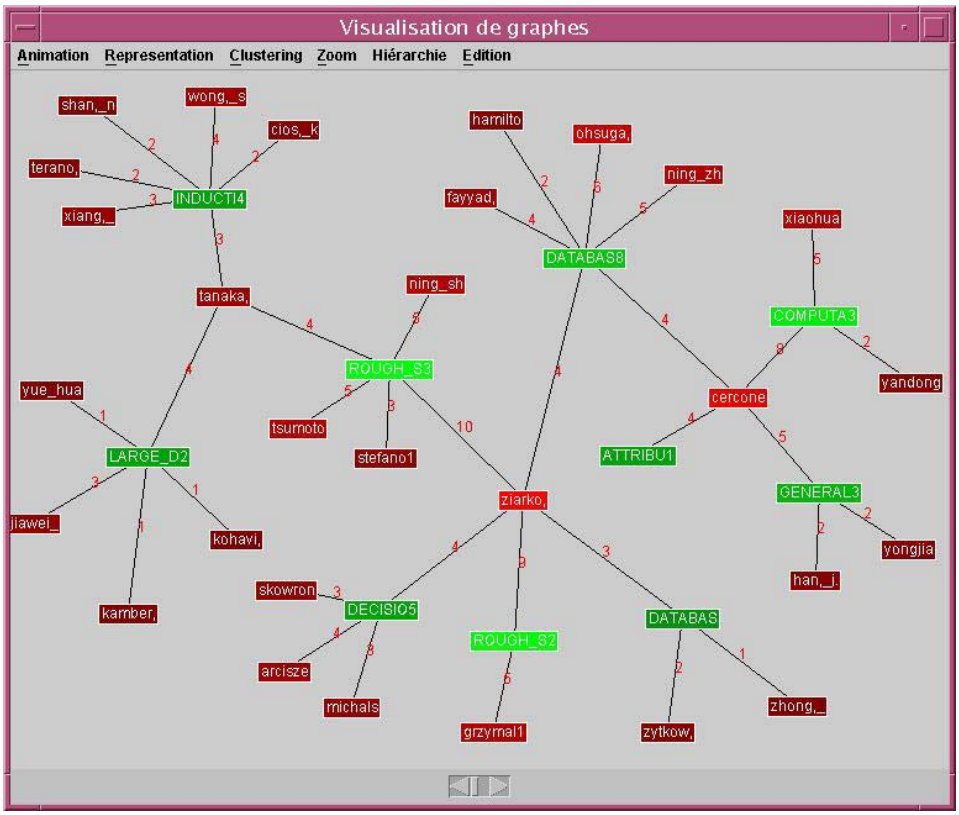


Figure 10. : Arbre partiel maximum extrait du graphe précédent.

CONCLUSION

La visualisation interactive et dynamique de graphes de grande taille apporte à notre plate-forme de veille un outil de restitution accessible à tous. En effet, la lecture d'un graphe ne nécessite pas de connaissances particulières comme celles exigées pour interpréter une carte factorielle, un arbre de classification ou un zoom interactif sur une matrice réorganisée. De plus, cet outil est disponible, grâce à Java, sur la plate-forme d'analyse elle-même, sur la restitution de l'analyse sous forme de site Web implantable sur CD/Rom ainsi que sur un portail ouvert sur l'ensemble des dictionnaires et matrices utilisés dans la macro-analyse et compilés sous forme de base de données relationnelle implantée **[SOSS 01]** en MySql et interfacée en Php. A tous les niveaux de la démarche de veille, ces graphes sont donc accessibles pour aider à la compréhension des phénomènes relationnels inhérents à toute activité humaine. L'étude des réseaux sert à découvrir les stratégies les plus fines qui ne sont pas révélées explicitement dans les écrits analysés mais qui s'y trouvent bel et bien et qu'il suffit de mettre en lumière par de telles méthodes. Les démarches basées sur l'analyse de données à la française nous permettaient déjà d'accéder à ce type d'informations stratégiques, mais leur mode de représentation graphique était assez mal adapté pour une restitution grand public. Le vide est maintenant comblé et ces nouveaux outils viennent compléter les modes de communication graphique et intuitive que nous privilégions actuellement comme les cartes géographiques interactives pour la géostratégie **[HUBE 00], [KARO 01], [DOUS 02].**

BIBLIOGRAPHIE

[EADE 84] P. Eades

A heuristic for graph drawing. Congressus Numeranium, Vol 42, pp. 149-160, 1984.

[KAMA 84] T. Kamada, S. Kawai

An algorithm for drawing general undirected graphs. Information Processing Letters, vol. 31, pp. 7-15, 1989.

[FRUC 91] T. Fruchterman, E. Reingold

Graph Drawing by Force-Directed Placement. Software Pratice an Experience, 1991.

[FRIC 94] A. Frick, A. Ludwig, H. Lehldau

A fast adaptative layout algorithm for undirected graphs. In Proceeding of Graph Drawing'94, vol. 894, pp. 388-403, 1994.

[DONG 00a] S. van Dongen

A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam. May 2000.

[DONG 00b] S. van Dongen

Graph Clustering by Flow Simulation. Thèse , Université d'Ultrecht, Allemagne, May 2000.

[DONG 00c] S. van Dongen

Performance criteria for graph clustering and Markov experiments. Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam. May 2000.

[HUBE 00] G. Hubert, J. Mothe, A. Benammar, T. Dkaki, B. Dousset, S. Karouach

Textual document Mining using graphical interface. International Human Computer Interaction, HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey, pp 918-922 (volume 1), 05-10 août 2001.

[KARO 01] S. Karouach, B. Dousset

Visualisation interactive pour la découverte de connaissances : GeoECD. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 291-300, (Barcelone, Espagne), octobre 2001.

[SOSS 01] D. Sosson, M. Vassard, B. Dousset

Portail pour la navigation en ligne dans les analyses stratégiques. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

[DOUS 02] B. Dousset, S. Karouach

Collaboration interactive entre classifications et cartes thématiques ou géographiques. 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.

EVALUATION DE TROIS MESURES DE SIMILARITE UTILISEES EN SCIENCES DE L'INFORMATION

Lelu Alain
INRA /unité Mathématiques, Informatique et Génomique
domaine de Vilvert, 78352 Jouy en Josas

Université de Franche-Comté / LASELDI
UFR SLHS, 25030 Besançon cedex
alain.lelu@univ-fcomte.fr

Résumé : sous-jacents à la diversité des méthodes et outils des sciences de l'information et de l'analyse des données textuelles, on trouve un petit nombre d'indicateurs de proximité entre textes ou entre descripteurs de ceux-ci . Nous comparons les deux plus répandus, le cosinus de Salton et la distance du khi-deux, à celui que nous avons choisi pour nos traitements : le cosinus dans l'espace distributionnel. Après un exposé et une comparaison théorique, nous procédons à une évaluation empirique par la méthode "rappel / précision" sur le corpus réel multilingue ELDA, fourni dans le cadre de la campagne d'évaluation Amaryllis 99.

Abstract : A small number of similarity indicators are underlying the rich diversity of methods and tools in the overlapping information sciences and text analysis fields. We test the major two of them, Salton's TF-IDF cosine and khi-square distance, against the one we have chosen in our text analysis tools : the cosine measure in the "distributional" vector space. After a theoretical presentation and comparison, we apply the "recall / precision" methodology to their empirical evaluation based on a real text database, provided by the ELDA multilingual corpus (Amaryllis'99 evaluation campaign).

Mots-clés : analyse de données textuelles, sciences de l'information, recherche d'information, évaluation, mesure de similarité, cosinus de Salton, distance du khi-deux, distance distributionnelle, gain d'information de Renyi.

Keywords : textual data analysis, information science, information retrieval, evaluation, similarity indicator, Salton's cosine, khi-square distance, distributional distance, Renyi's information measure.

Evaluation de trois mesures de similarité utilisées en sciences de l'information

INTRODUCTION ET GRILLE D'ANALYSE

Confrontés aujourd'hui à des quantités croissantes de textes bruts (pages Web, dépêches d'agences, articles en ligne, …) les méthodes et outils de la documentation automatique et des sciences de l'information commencent à chevaucher les frontières de ceux des autres analyses de corpus : analyses de discours, d'œuvres littéraires, d'interviews, études psychologiques ou psycholinguistiques, … Jusqu'à présent, chaque méthode utilisée dans ces domaines a souvent été présentée comme un tout en soi, le plus souvent " à prendre ou à laisser ". De fait chacune est caractérisée par une série de choix situés sur des registres très différents :

A - Découpage des unités textuelles : Ces unités peuvent être "logiques" (articles ou chapitres, sections, paragraphes, phrases, …), ou "naturelles" (pages, résumés bibliographiques), ou encore être issues d'un découpage à fenêtre glissante autour de chaque mot, de chaque phrase, etc.

B - Choix de codage des textes : A chaque texte on fait correspondre un ensemble de descripteurs codés – se pose donc la question du choix de ces descripteurs. Faut-il opter pour un codage par mots, ou bien par N-grammes (Lelu et al., 1998) ? Au sein du codage par mots : faut-il opter pour une indexation manuelle, automatique, assistée ? Au sein des indexations automatiques et assistées : pour une indexation par simples chaînes de caractères (" full text "), ou par lemmes (mots normalisés) issus d'un traitement morpho-syntaxique, avec ou sans filtrage de certaines catégories grammaticales, et/ou filtrage statistique ? Tous ces choix de codage conduisent *in fine* à la représentation de chaque unité textuelle par un vecteur – logique (valeurs binaires) ou de fréquences - des descripteurs codés.

C - Choix des opérations offertes aux usagers : Au-delà des classiques requêtes booléennes, les systèmes documentaires évolués permettent à l'utilisateur, depuis les travaux pionniers de Gerald Salton et Karen Sparck-Jones dans les années 1960, de trouver les documents les plus proches d'un document désigné comme pertinent (*relevance feed-back :* rétroaction de pertinence) ou proches d'un document "idéal" défini par une série de mots, voire une requête en langage naturel (requête de similarité vectorielle) ; ou encore trouver les mots les plus proches d'un mot donné (*query expansion*), proches au sens de leur co-occurrence dans les unités textuelles. Beaucoup de logiciels d'analyse de données textuelles proposent des fonction voisines, par exemple : extraire les termes environnant un un terme donné, avec des indicateurs statistiques pour valider la significativité de ces liaisons.

Plus récemment, certains systèmes documentaires en ligne ont mis à la disposition de l'utilisateur des algorithmes de classification automatique sur les documents (Zamir et Otzioni, 1999) ou sur les mots (Bourdoncle, 1997), ou même une représentation cartographique d'ensemble des classes ainsi créées (Kohonen et al., 1995) (Lelu et al., 1997). Si les logiciels d'analyse de données textuelles comportent souvent de telles classifications (par ex. classification descendante hiérarchique dans le cas du logiciel Alceste [ALCESTE ; Reinert 1993]), ils proposent également la possibilité d'effectuer une analyse factorielle des correspondances (AFC : Benzecri et al., 1981) sur un sous-ensemble d'unités textuelles et de mots. Dans le même ordre d'idées et sur des domaines d'applications psycho-pédagogiques et documentaires, la méthode Latent Semantic Analysis (LSA ; Dumais 1994) réalise une opération de réduction de dimensions et filtrage des données par décomposition aux valeurs singulières de la matrice complète des données - les composantes issues de cette décomposition ne sont pas interprétées, mais servent à réaliser des opérations de rétroaction de pertinence ou d'expansion de requête dans l'espace des 200 à 300 premières composantes.

Mais toutes ces opérations et algorithmes portent rarement sur les vecteurs bruts issus des choix précédents : ils peuvent utiliser ces vecteurs sous leur forme normalisée en ligne ou en colonne (calculs de cosinus), ou réaliser des transformations plus complexes sur ces éléments.

D - Transformations des vecteurs-données : De façon générale, chaque opération d'analyse des données peut être définie par une primitive standard - comme un calcul de cosinus, une décomposition aux valeurs singulières, une classification à centres mobiles, … - dans un espace de données transformé : chaque vecteur est déduit du vecteur brut par une certaine opération (normalisation, intersection avec l'hyperplan simplexe, …) ; il est doté

d'un poids (unitaire, ou déduit des marges de la matrice des données), et ce dans une métrique particulière, en donnant au mot métrique son sens mathématique de matrice **M** carrée généralement symétrique, définie positive, intervenant dans la définition générale du produit scalaire : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{M}} = \mathbf{x}' \mathbf{M} \mathbf{y}$

Cette métrique peut être euclidienne standard (**M** est alors une matrice identité), ou du khi-deux, etc.

Les combinaisons de ces divers registres de choix réalisés par chaque méthode comportent une grande part d'arbitraire, et peu d'auteurs se préoccupent de leur justification. Notre exposé se propose de comparer quelques transformations de données sur un plan théorique, puis nous présenterons une tentative d'évaluation pratique à partir d'une méthode reconnue dans le domaine documentaire : l'utilisation des courbes "rappel/précision" relatives à un ensemble de questions pour lesquelles des experts ont déterminé exhaustivement les réponses correctes. Dans ce but, seule est concernée l'opération d'expansion d'un document ou d'une requête vers les documents proches : à ce jour l'évaluation de la qualité d'une classification ou d'une cartographie textuelle reste un problème ouvert, sur lequel peu de propositions existent, et encore moins d'accord.

Notre travail se limitera donc à une comparaison/évaluation de 3 indicateurs de similarité ; il se rapproche de l'exposé (Lebart et Rajman 2000) qui compare de façon théorique et en détail une variante différente du cosinus de Salton à la distance du khi-deux et à la dissimilarité informationnelle de Kullback-Liebler, au sein d'un exposé très large sur l'ensemble des méthodes statistiques utilisées pour traiter la " matière textuelle " dans les multiples domaines qui la concernent (découpage d'unités textuelles, mesures de proximité, enrichies ou non par des connaissances *a priori*, fonction d'indexation, synthèses de corpus, phrases ou documents caractéristiques, classement de textes…).

1. TROIS MESURES DE SIMILARITE DANS L'ESPACE DES UNITES TEXTUELLES ET DES MOTS

1.1. Le cosinus de Salton

Notations utilisées :
x_{it} : fréquence du mot N° i dans le texte N° t ;
Sommes en ligne, en colonne, totale : $x_{i.} = \Sigma_t x_{it}$; $x_{.t} = \Sigma_i x_{it}$; $x_{..} = \Sigma_i \Sigma_t x_{it}$
Vecteurs (et matrices) : en minuscules (resp. majuscules) grasses. Ex. :
$\mathbf{x}_i$: vecteur-mot i ; $\mathbf{x}_t$: vecteur-texte t ; **Diag** [a] : matrice diagonale d'éléments a
{ } : ensemble des éléments d'un vecteur ou d'une matrice. Ex. :
$\{ x_{i.} \}$: vecteur " sommes en ligne " ; $\{ \mathbf{x}_i \}$: matrice **X** des données
Produit scalaire des vecteurs-colonnes **x** et **y** dans la métrique **D** : $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{D}} = \mathbf{x}' \mathbf{D} \mathbf{y}$
(la norme de **x** dans la métrique **D** est définie par : $\|\mathbf{x}\|^2_{\mathbf{D}} = \langle \mathbf{x}, \mathbf{x} \rangle_{\mathbf{D}}$)

Parmi les nombreux indices de similarité entre documents proposés et validés par G. Salton et son équipe (Salton, 1983), le plus répandu et apprécié est nommé par ses auteurs *best fully weighted system* (ou encore TF-IDF : term frequency, inverse document frequency) et défini par le produit scalaire usuel (métrique identité) entre vecteurs-documents dont les composantes sont définies (" pondérées ", selon la terminologie de Salton) par le terme $x_{it} \log(N/n_i)$, puis normalisés :

$$\mathbf{x}_t : \{ x_{it} \} \rightarrow \{ x_{it} \log(N/ n_i) / \sqrt{(\Sigma_i (x_{it} \log(N/ n_i))^2} \}$$

Où N est le nombre total de documents, et n_i le nombre de documents contenant le mot i, quelle que soit sa fréquence x_{it} dans chaque document.
Il revient à calculer un cosinus, c'est à dire un produit scalaire entre vecteurs normalisés de fréquences de mots dans un espace des documents pourvu de la métrique :

$$\mathbf{Diag} [(\log(N/n_i))^2]$$

Pour les besoins de notre comparaison, on peut remarquer que $\log(N/n_i)$ est égal à $\log(x_{..}/x_{i.})$ si les mots sont codés en présence/absence, ou en diffère peu, à des constantes additives près, pour des corpus volumineux et des documents longs : dans ce cas fréquent la métrique de l'espace des données peut donc être assimilée à :

$$\mathbf{Ds} = \mathbf{Diag} [(\log(x_{..}/x_{i.}))^2]$$

En résumé, Cos Salton $(t_1,t_2) \cong \langle \mathbf{x}_{t1}, \mathbf{x}_{t2} \rangle_{\mathbf{Ds}} / \| \mathbf{x}_{t1} \|_{\mathbf{Ds}} \times \| \mathbf{x}_{t2} \|_{\mathbf{Ds}}$

1.2. Distance du khi-deux

Il est bien connu (Lebart et al., 1977) que l'AFC est équivalente à une analyse en composantes principales (ACP) dans l'espace des données transformé comme suit :

- coordonnées de $\mathbf{x}_t$: $\{ x_{it} \} \rightarrow \{x_{it} / x_{.t} \}$
-

Les points sont alors sur le simplexe, hyperplan lieu des points **z** tels que $\Sigma_i z_i = 1$ On compare donc des profils relatifs, cas bien adapté aux données textuelles où le nombre absolu de mots dans un texte est indifférent, et seule leur répartition relative compte.

- métrique : **Diag** $[x_{..} / x_{i.}]$
- masses des points t : $\{ x_{.t} / x_{..} \}$

La propriété la plus intéressante de cet espace est qu'il est doté de la propriété d'*équivalence distributionnelle* : si on fusionne deux descripteurs de mêmes profils relatifs, les distances entre les unités textuelles sont inchangées. En d'autres termes, cette propriété assure la stabilité du système des distances au regard de l'éclatement ou du regroupement de catégories de descripteurs, tant que les unités textuelles s'y répartissent de façon un tant soit peu identique.

L'ordre des distances croissantes nous donne celui des similarités décroissantes par rapport à un document réel ou "idéal" posé en requête.

Résumons les similarités et différences par rapport à l'espace du cosinus de Salton :

. Dans l'espace de Salton, les points sont représentés sur une hypersphère unité : on compare donc des profils normalisés, et non des profils relatifs, comme en AFC. Toutefois, dans les deux cas la métrique est de la famille *inverse document frequency* (IDF) :

Diag $[x_{..}/x_{i.}]$ pour l'AFC, **Diag**$[(\log(N/n_i)^2)]$, souvent équivalent à **Diag**$[(\log(x_{..}/x_{i.}))^2]$, pour le cosinus TF-IDF de Salton

. L'AFC fait un usage intensif de la symétrie formelle entre lignes et colonnes de la matrice des données, alors que peu d'auteurs suggèrent d'utiliser le cosinus de Salton pour l'expansion d'un mot vers les autres mots, bien que rien ne s'y oppose.

1.3. Le cosinus dans l'espace distributionnel

Une lignée ancienne de travaux (Matusita 1955) (Escofier 1978) (Domengès et Volle 1979) (Fichet et Gbegan 1985) se sont intéressés à ce que ces derniers auteurs appellent *distance distributionnelle* : si l'on transforme les données brutes comme suit :

. Coordonnées : - des vecteurs-colonnes $\mathbf{x}_t : \{ x_{it} \} \rightarrow \mathbf{y}_t : \{\sqrt{ x_{it} } \}$
- des vecteurs-lignes $\mathbf{x}_i : \{ x_{it} \} \rightarrow \mathbf{y}_i : \{ \sqrt{ x_{it} } \}$

. Poids de ces vecteurs : unité
. Métrique : euclidienne standard

les cosinus entre vecteurs-colonnes $\mathbf{y}_t$ (resp. entre vecteurs-ligne $\mathbf{y}_i$) possèdent dans cet espace des propriétés intéressantes. Ils sont liés en effet à la notion de distance distributionnelle Dd par la relation :

$Dd(t_1,t_2)^2 = 2(1-\cos(t_1,t_2))$
(resp. $Dd(i_1,i_2)^2 = 2(1-\cos(i_1,i_2))$)

La distance distributionnelle est la distance entre les intersections de 2 vecteurs $\mathbf{y}_{t1}$ et $\mathbf{y}_{t2}$ avec l'hypersphère unité, c'est à dire la longueur de la corde correspondant à l'angle ($\mathbf{y}_{t1}$, $\mathbf{y}_{t2}$) - égale au plus à 2 quand ces 2 vecteurs sont opposés, égale à $\sqrt{2}$ quand ils sont orthogonaux. Cette distance semble triviale et arbitraire en apparence (pourquoi partir du tableau des racines carrées plutôt que celui des fréquences brutes ?), mais elle jouit de propriétés intéressantes :

- Escofier et Volle ont montré qu'elle satisfaisait à la propriété d'équivalence distributionnelle décrite plus haut, au même titre que la distance du khi-deux utilisée en AFC.
- Contrairement à celle-ci, elle peut prendre en compte des vecteurs ayant des composantes négatives, propriété utile pour certains types de codage " symétriques " (comme *Oui, Non, Ne sait pas*), ou pour des tableaux de flux divers, économiques ou autres.
- Elle est liée à la mesure du gain d'information de Renyi d'ordre ½ (Renyi 1955) apporté par une distribution $\mathbf{x}_q$ quand on connaît la distribution $\mathbf{x}_p$:

$$I^{(1/2)}(\mathbf{x}_q / \mathbf{x}_p) = -2 \log_2 (\cos(\mathbf{y}_p, \mathbf{y}_q)) = -2 \log_2 (1 - Dd^2/2)$$

- Elle est rapide à calculer dans le cas des données textuelles, où les vecteurs $\mathbf{y}_t$ sont très " creux ".

Si on calcule les directions propres $\mathbf{U} = \{ \mathbf{u}^{(k)} \}$ (resp $\mathbf{W} = \{ \mathbf{w}^{(k)} \}$) du nuage des points-colonnes $\mathbf{y}_t$ (resp., des points-lignes $\mathbf{y}_i$) défini plus haut (k est l'ordre de ces directions propres), on démontre 1) que les valeurs propres $\lambda^{(k)}$ sont communes aux deux nuages analysés, 2) que les cosinus ci-après se déduisent des directions propres :

$$\text{Cos}(\mathbf{y}_t, \mathbf{u}^{(k)}) = w_t^{(k)} \sqrt{(\lambda^{(k)} / x_{.t})} \quad (1)$$
$$\text{Cos}(\mathbf{y}_i, \mathbf{w}^{(k)}) = u_i^{(k)} \sqrt{(\lambda^{(k)} / x_{i.})} \quad (2)$$

Ces cosinus peuvent être considérés comme les facteurs d'un cas particulier et simple d'analyse factorielle sphérique, pour reprendre la terminologie de M. Volle, dite centrée sur le " tableau nul ". Nous les nommerons désormais respectivement $F_t^{(k)}$ et $G_i^{(k)}$. Ils sont liés entre eux et avec les composantes $w_t^{(k)}$ et $u_i^{(k)}$ par des formules de transition, en particulier :

$$F_t^{(k)} = \Sigma_i u_i^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$
$$G_i^{(k)} = \Sigma_t w_t^{(k)} \sqrt{(x_{it} / \lambda^{(k)})}$$

Ce qui permet de les déduire à partir de l'extraction des éléments propres du plus petit des deux nuages de points.

Nos algorithmes K-means axiales (Lelu, 1994) et Analyse en composantes locales (Lelu et Ferhan, 1998) réalisent la partition en classes des unités textuelles dans l'espace distributionnel, et extraient pour chaque classe les facteurs mots et documents définis ci-dessus. Chaque facteur est alors un indicateur de *centralité* (ou " typicité ") du document ou du mot dans sa classe.

Contrairement à la distance du khi-deux et au cosinus de Salton, les calculs de similarité dans cet espace ne font pas intervenir une métrique de type *inverse document frequency*. Ceci dit, il est à noter que le passage d'un ensemble de documents à l'ensemble des mots qui en sont les plus caractéristiques se fait en deux étapes :

1) Calcul du 1er vecteur propre (facteurs-documents, s'il y a moins de documents que de mots) et de sa valeur propre associée issus du tableau des racines carrées des fréquences brutes,
2) Calcul des facteurs-mots via la formule (2), qui fait intervenir une correction de type IDF.

On peut donc dire que la pondération de type IDF intervient aussi dans l'espace distributionnel, mais de façon moins évidente que dans le cosinus de Salton ou la distance du khi-deux. Nous comptons approfondir plus tard ce problème d'attribution de mots caractéristiques à un bloc de texte, dit de " fonction d'indexation ", en comparant la solution ci-dessus à d'autres, en particulier celle basée sur un modèle hypergéométrique présentée dans (Lebart et Rajman 2000).

Nos algorithmes ont fait la preuve depuis un bon nombre d'années de résultats " sensés et interprétables " sur de nombreux corpus de toutes tailles, de toutes provenances, et de tous types de codage.

Pour convaincre de leur bien-fondé d'autres personnes que les utilisateurs directs de nos méthodes, la nécessité d'une validation moins subjective se fait sentir, qui permette en particulier une comparaison avec les mesures utilisées de façon courante par ailleurs.

2. COMPARAISONS EMPIRIQUES

2.1. Le corpus utilisé

Parmi les corpus mis à disposition par la campagne d'évaluation de systèmes d'information Amaryllis 99 (ref. : Amaryllis 1999) figure celui de l'ELDA (ref. : Elda), constitué de 3511 questions écrites des députés européens à la Commission des Communautés européennes, avec les réponses de la Commission ; ce qui représente en français 8 Mo de textes bruts, sans mise en page, soit en moyenne autour de 2300 signes par texte Ces textes comportent également des versions dans les principales langues européennes, rassemblées pour alimenter la tâche multilingue de cette évaluation.

2.2. Les traitements effectués

Nous sommes partis pour la présente expérience de l'ensemble des mots lemmatisés et des candidats termes composés proposés par le logiciel de traitement morpho-syntaxique Nomino (Plante et al., 1997) : 82 572 termes hors les mots grammaticaux, les verbes opérateur comme *avoir* et *être* ; nous avons éliminé par seuillage les hapax et les mots de fréquence supérieure à 1500 : restent 20 034 formes lemmatisées différentes. Nous n'avons pas éliminé les verbes et les adjectifs, contrairement à notre pratique habituelle sur les corpus documentaires, afin de rester les plus fidèles possible au contenu d'origine. Afin d'enlever un maximum de bruit dû aux polysémies, nous avons retiré de l'indexation de chaque texte les mots simples dont la composition formait les mots composés - nous verrons plus loin que cette option a des répercussions importantes en matière de stratégie de recherche d'information.

2.3. Les requêtes de test et leurs expansions

Les organisateurs d'Amaryllis 99 ont fourni au total 30 requêtes, associées chacune avec l'ensemble de documents pertinents validé par des experts. Par manque de temps, nous n'avons exploité que les 15 premières. Nous avons traité comme un seul texte l'ensemble des champs de chaque requête - titre, sous-titre, explications, mots-clés – par le logiciel Nomino. Ainsi par exemple :

Q13. Pollution atmosphérique.
Pollution de l'air et lutte antipollution.
Tous les documents sur la pollution atmosphériques et intérieure par dioxyde et monoxyde de carbone, oxydes de soufre et d'azote, ozone, plomb, particules ... (réglementation, lutte antipollution, aspect économique).
Pollution air.
Lutte antipollution.
Pollution intérieure.
Emission particule.
Carbone dioxyde.
Carbone monoxyde.
Effet serre.
Azote oxyde.
Soufre oxyde.

Après importation des termes trouvés (tous n'étaient pas présents dans le corpus…) nous avons utilisé notre procédure d'expansion lexicale Proxilex (Lelu et al., 1998) sur chacun de ses composants, puis réalisé une expansion " sémantique " par similarité {mots}→ mots) dans notre environnement de contrôle de vocabulaire NeuroNav (NeuroNav) où nous avons sélectionné manuellement les termes pertinents parmi les 250 termes les plus associés aux termes de requête.

Ainsi avons-nous obtenu pour la requête N°13 une liste de 81 mots reformulant cette requête initiale, auxquels nous n'avions le plus souvent pas pensé à l'origine, et dont voici un extrait :

énergie
teneur
protection
qualité
fixer
installation
environnement
air
pollution
concentration
SO2
NO2
polluer
effet
seuil
valeur
limiter
réduire
substance
réglementation
risque
énergies_renouvelables
taxe
lutte
eau
charbon
gaz_à_effet_de_serre
émissions_de_dioxyde_de_carb
one
émissions_de_CO2
dioxyde
carbone
émission
CO2
gaz
combustion
serre
climatique
moteur
rejet
prévention
station_de_mesure
pollution_marine
pollution_atmosphérique
sol
soufre
azote
particule

Nous avons vérifié sur un exemple que le choix de termes effectué parmi une longue liste des 250 premiers éléments produit les mêmes termes pertinents, mais avec un ordre et un rang différent, quelle que soit la mesure de similarité utilisée pour l'expansion.

2.4. Etablissement des courbes Rappel / précision

La méthodologie d'évaluation des systèmes documentaires classique depuis (Salton 1968) définit les notions de taux de rappel et taux de précision :

Rappel = nombre de documents pertinents retrouvés / nombre de documents retrouvés
Précision = nombre de documents pertinents retrouvés / nombre de documents pertinents

Nos trois mesures de similarité (cosinus de Salton, distance du khi-deux changée de signe, et cosinus dans l'espace distributionnel) donnent pour notre liste de mots en requête trois listes de titres les plus proches. Pour chaque liste, et pour chaque titre pertinent trouvé par ordre de similarité décroissante, nous calculons les indices de rappel et précision. Pour un exemple de requête nous avons reporté ces valeurs sur la figure 1, où nous avons dessiné également, après lissage, les trois courbes Rappel / Précision respectives. Au total, la figure 2 montre les trois courbes Rappel / Précision pour l'ensemble des requêtes.

2.5. Résultat et discussion

Plus une courbe est haute et plus la concentration de documents pertinents est élevée en début de liste de documents restituée, ce qui est le but recherché : le cosinus distributionnel sort vainqueur de cette comparaison, sauf pour les valeurs moyennes du taux de rappel (0.4 à 0.6), où le cosinus de Salton fait jeu égal avec lui. La distance du khi-deux est systématiquement moins bonne, ce que confirme une expérience antérieure menée sur un corpus-jouet de 8 documents et 7 mots : après expansions exhaustives sur chaque document, la moyenne des corrélations de rang de Spearman entre cos. distributionnel et khi-deux est négligeable (pas de relation) alors qu'elle est de 0,25 entre cos. de Salton et khi-deux, de 0,45 entre cos. de Salton et cos. distributionnel.

Il est possible que le type d'indexation "pointue" adopté, principalement à base de mots composés lemmatisés et de mots simples non redondants avec ceux-ci, donc avec peu de termes génériques, défavorise les similarités avec pondération des mots de type IDF, comme le cosinus de Salton, et plus encore la distance du khi-deux. Des comparaisons dans le cadre de stratégies d'indexation différentes seraient à réaliser pour approfondir cette question, que nous n'avons pas trouvé mentionnée à notre connaissance dans la littérature concernant la recherche d'information.

Mais il faut aussi noter que ce type d'indexation donne d'excellents taux de rappel, 100% en général, en tête des listes de réponses, quelle que soit la mesure de similarité utilisée, du fait de l'absence d'ambiguïté et de la précision conceptuelle des termes composés.

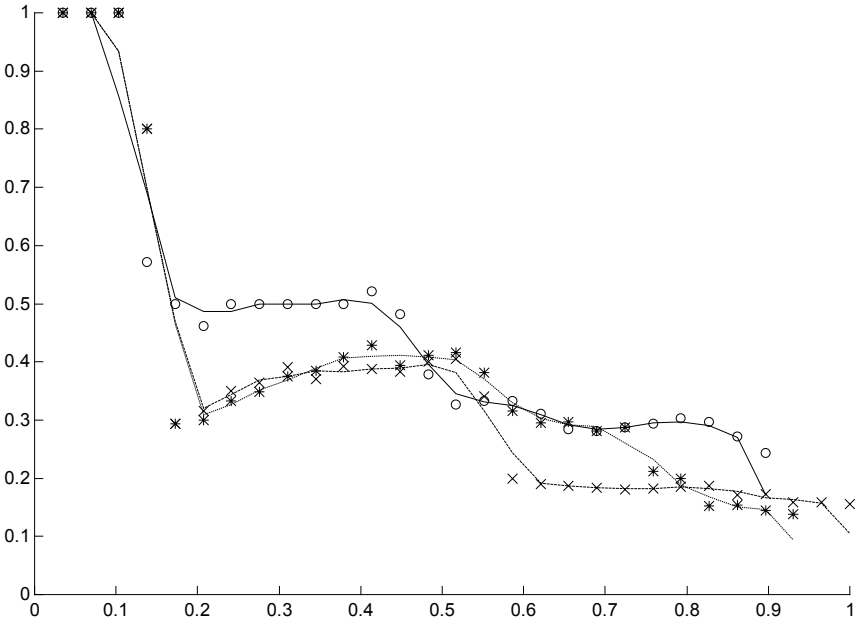


Figure 1 - Courbes Rappel / précision de la question 1 pour les 3 mesures de similarité.

En abscisse : taux de précision ; en ordonnée : taux de rappel

o : *cosinus dans l'espace distributionnel*

∗ : *cosinus de Salton*

× : *distance du khi-deux*

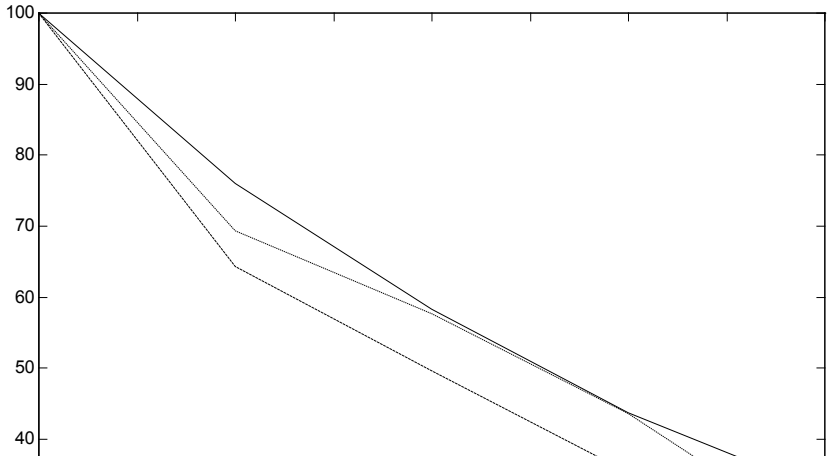


Figure 2 - Courbes Rappel / précision globales pour les 3 mesures de similarité.

En abscisse : taux de précision ; en ordonnée : taux de rappel

o : *cosinus dans l'espace distributionnel*

∗ : *cosinus de Salton*

× : *distance du khi-deux*

CONCLUSIONS ET PERSPECTIVES

Sur le strict plan de l'étude présentée ici, il serait intéressant de la prolonger par des comparaisons 1) par rapport à d'autres indicateurs de liens, comme ceux issus des théories de l'information, 2) par rapport à d'autres choix d'indexation – filtrer certaines catégories grammaticales ? Ne pas supprimer les composants des mots composés dans une même unité textuelle ? Voire " indexer " par des N-grammes, au lieu de mots (Lelu et al. 1998) ?

Un autre axe intéressant à explorer serait la comparaison des résultats de similarités, sur un même corpus, et à transformation égale de l'espace des données, entre l'indexation sémantique latente (LSI : cf. LSA) et l'indexation sémantique *patente* que réalisent nos algorithmes K-means axiales et Analyse en composantes locales – ils extraient des axes obliques et interprétables, au lieu d'axes orthogonaux et en majeure partie ininterprétables, sur lesquels les mots et les documents n'ont que des composantes positives ou nulles ; composantes qui traduisent une décomposition *additive* des données, plausible sur le plan cognitif (un texte, ou un mot, se rattache, avec plus ou moins de force, à un ou plusieurs thèmes sémantiques parmi de nombreux répertoriés dans le corpus). Cette comparaison devrait être éclairante quant au nombre de dimensions pertinentes sous-jacentes à un corpus textuel typique de quelques méga-octets : moins d'une dizaine de dimensions, d'après la pratique recommandée pour les logiciels SPAD/T (Lebart et Salem, 1994) et Alceste ? De 200 à 300, d'après l'équipe LSA ?

Nous espérons avoir contribué à faire avancer la question de l'évaluation comparée des méthodes en vigueur dans le domaine de l'exploitation des corpus textuels, au sens large, en montrant que pour au moins une stratégie d'indexation automatique le cosinus distributionnel est plus pertinent pour les calculs de similarité que le cosinus " TF-IDF " de Salton, lui-même meilleur que la distance du khi-deux. C'est seulement sur ce socle constitué par la transformation de l'espace des données (espérons que nos résultats seront complétés, nuancés ou, pourquoi pas, contestés par des tests ultérieurs !) que les diverses représentations synthétiques des données pourraient être comparées toutes choses égales par ailleurs – que ce soient les nombreuses variantes des analyses factorielles orthogonales ou obliques (Analyse en Composantes Principales, Latent Semantic Analysis, Independent Component Analysis, Analyse en Composantes Locales, Non-negative Matrix Factorization…), et celles des classifications supervisées ou non (pour les non-supervisées : K-means, K-means axiales, classifications ascendantes ou descendantes hiérarchiques, réseaux de Kohonen, …) ; en ce qui concerne les méthodes supervisées (analyses discriminantes, perceptrons multi-couches, réseaux RBF, Support Vector Machines, …) cette distinction permettrait de faire la part de l'algorithme proprement dit dans la qualité des résultats obtenus, par rapport à celle du pré-traitement des données.

REMERCIEMENTS :

Aux organisateurs de la campagne d'évaluation Amaryllis 1999.

A l'agence européenne ELDA (Evaluations and Language resources Distribution Agency) pour avoir fourni le corpus de test utilisé et les couples questions-réponses permettant les tests.

REFERENCES

ALCESTE : cf. www.image.cict.fr/

Amaryllis 1999 : cf. http://amaryllis.inist.fr/

Benzécri J.P. et coll. (1981). Pratique de l'Analyse des Données : Linguistique et Lexicologie. Dunod, Paris.

Bourdoncle F. (1997). LiveTopics : recherche visuelle d'information sur l'Internet. In C. Jacquemin editor, Proc. of RIAO'97 (Recherche d'Information Assistée par Ordinateur), CID, Paris.

Domengès D., Volle M. (1979). Analyse factorielle sphérique : une exploration. Annales de l'INSEE, 35-1979 :3-84, Paris.

Dumais S.T. (1994). Latent Semantic Indexing (LSI) and TREC-2. NIST special publication, N°500-215, pages 105-115, NIST.

ELDA : cf. www.elda.fr/

Escofier B. (1978). Analyses factorielles et distances répondant au principe d'équivalence distributionnelle. Revue de Stat. Appliquée, 26(4):29-37, Paris

Fichet B. et Gbegan A. (1985). Analyse factorielle des correspondances sur signes de présence-absence. In Diday et al. editors, 4e Journées Analyse des données et Informatique. INRIA, Rocquencourt.

Matusita, Kameo (1955). Decision rules, based on the distance for problems of fit, two examples, and estimation. Annals of Statistical Mathematics, pages 631-640, Tokyo.

Kohonen T., Kaski S., Lagus K., Honkela T. (1995). Very large two-level SOM for the browsing of newsgroups. Proc. of WWW'95 (5th International World Wide Web Conference), Paris.

http://websom.hut.fi/websom

Lebart L., Morineau A., Tabard N. (1977). Techniques de la description statistique. Dunod, Paris.

Lebart L., Rajman M. (2000). Computing Similarities. In Dale R., Moisl H., Somers H. editors : Handbook of Natural Language Processing, Marcel Dekker, pages 477-505, New York.

Lebart L., Salem A. (1994). Statistique textuelle. Dunod, Paris.

Lelu A., Tisseau-Pirot A.G. (1993). Emergence de catégories sémantiques à partir d'une base de résumés d'articles. In Anastex S.J. editor, Proc. of JADT'98 (2emes Journées Internationales d'Analyse Statistique des données Textuelles), pages 227-242, ENST, Paris.

Lelu A. (1994). Clusters and factors: neural algorithms for a novel representation of huge and highly multidimensional data sets. In E. Diday, Y. Lechevallier & al. editors. New Approaches in Classification and Data Analysis, pages 241-248, Springer-Verlag, Berlin,

Lelu A., Tisseau-Pirot A.G., Adnani A. (1997). Cartographie de corpus textuels évolutifs : un outil pour l'analyse et la navigation. Hypertextes et Hypermédias, 1(1):23-55, Hermès, Paris,

Lelu A., Ferhan S. (1998). Clustering a textual dataflow by incremental density-modes seeking. In Rizzi A. et al. editors, Proc. of IFCS'98 (6th Conference of the International Federation of Classification Societies), pages 206-209, Universita La Sapienza, Roma.

Lelu A., Hallab M., Delprat B. (1998). Recherche d'information et cartographie dans des corpus textuels à partir des fréquences de N-grammes. In Mellet S. editor, Proc. of JADT'98 (4emes Journées Internationales d'Analyse Statistique des données Textuelles), pages 391-400, UPRESA " Bases, Corpus et Langage ", Université de Nice.

LSA : cf. http://lsa.colorado.edu/

NeuroNav : cf. www.diatopie.com

Plante P., Dumas L. et Plante A. (1997). Atelier FX. ATO, Département de Linguistique, Université du Québec à Montréal.

http://www.ling.uqam.ca/Ato/FX

Reinert M. (1993). "Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars", Langage et Société, 66, 5-39.

Renyi A. (1966). Calcul des probabilités. Dunod, Paris.

Rhissassi H. and Lelu A. (1998). Indexation assistée et cartographie sémantique pour la génération automatique d'hypertexte. In Mojahid M. editors, Proc. of CIDE'98, pages 131-139, Europia Productions, INPT, Rabat, Maroc.

Salton G. (1968). Automatic Information Organization and Retrieval. Mac Graw Hill, NY.

Salton G. and Mac Gill M.J (1983). Introduction to Modern Information Retrieval. International Student Edition.

Zamir O., Etzioni O. (1999).Grouper : a dynamic Clustering Interface to Web Search Results. Proc. of WWW'99 (8th International World Wide Web Conference).

ANALYSE BIBLIOMETRIQUE DES COLLABORATIONS INTERNATIONALES DE L'INRA.

MULTON Jean-Louis,
INRA, Mission Relations Internationales, 147 rue de l'Université, 75338 Paris cedex 07
multon@paris.inra.fr

BRANCA-LACOMBE Geneviève,
INRA, DISI, BP 2078, 06606 Antibes cedex
lacombe@antibes.inra.fr

DOUSSET Bernard,
IRIT, SIG, Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse cedex 04
dousset@irit.fr

Résumé : Ce travail est la réactualisation d'une étude réalisée, il y a deux ans, dans des conditions similaires et qui ne portait que sur deux périodes : 1998 et 1999 avec le début de l'an 2000. Nous avons complété le corpus qui avait été constitué à l'époque à partir des bases bibliographiques sur CD/Rom : Current-Contents, Pascal et Sci. La version CD/Rom de Sci qui ne disposait ni du résumé, ni de mots-clés a été remplacée cette fois-ci par la version en ligne du Web of science qui est beaucoup plus complète mais assez difficile à décharger. Nous disposons maintenant de quatre années de production scientifique (janvier 1998 à Avril 2002) ce qui nous permet une détection beaucoup plus complète des acteurs et de leurs relations ainsi que des thématiques abordées. Après avoir analysé les collaborations internationales de l'INRA, nous avons pu réaliser un zoom sur deux entités de cet institut : le département TPV (Traitement des Produits Végétaux) et le positionnement de l'INRA aux Caraïbes (centres INRA des Caraïbes et pays de la région qui collaborent avec l'INRA local ou de métropole). Ce complément d'étude montre que le corpus initialement constitué pour les collaborations internationales renferme bien d'autres informations stratégiques. Chaque département, chaque laboratoire, chaque région peut ainsi prendre connaissance de son environnement, de ses forces et de ses faiblesses, de ses relations avec les autres organismes de recherche. Une politique de développement peut en être déduite en prenant en compte de façon rigoureuse des éléments objectifs mesurant la lisibilité de l'activité de recherche et de son évolution. Nous présentons les principaux résultats de cette analyse et les perspectives à moyen et long terme de ce type d'étude.

Abstract : This work is the actualization of a study which has been achieved in similar conditions two years ago; it was about only two periods 1998, 1999 and the beginning of 2000. We completed the corpus which was constituted with cd-rom bibliographic databases: Current-Contents, Pascal and Sci. The Sci cd-rom version which had neither abstract nor keywords, has been replaced here by the Web of Science online version, which is more

complete but not easy to load. We have now four years of scientific production (january 1998 to april 2002) which allow to detect actors, subjects and their relations. With the international collaboration analysis of INRA, we achieved a zoom on the two entities' institute: the TPV department (Treatment of Vegetal Products) and the INRA position in the Caribean.
This completed study shows that the corpus, initially constituted for the international collaborations, contains more strategic informations. By this way, each department, laboratory or area can read its environment, its forces and weaknesses, its relations with the others research organisms. Then a development policy can be deduced by the measurement of research activity and its evolution. We present the main results of the analysis and the middle and long-term views.

Mots-clés : Veille scientifique et technique - Bibliométrie - Collaborations internationales.

Keywords : Science and technology watch - Bibliometry - International collaboration

Analyse bibliométrique des collaborations internationales de l'INRA.

INTRODUCTION

A la suite de l'étude présentée au Colloque VSST'2001 (Veille Stratégique Scientifique et Technologique, 15-19 octobre 2001 – Barcelone) sur la période 1998-1999, une nouvelle étude bibliométrique d'évaluation quantitative des collaborations internationales de la Direction des Relations Internationales de l'INRA a été effectuée sur la période 1998 – 2001, complétée par deux zooms, l'un sur un Département de recherche et l'autre sur une région particulière : la zone Caraïbe.
Comme précédemment :

- Le logiciel retenu pour effectuer cette étude est « Tétralogie » (IRIT Toulouse) identifié comme étant celui qui répond le mieux au cahier des charges fixé. Cependant, la présente étude a bénéficié des nouvelles avancées de « Tétralogie » en matière de cartographie et d'analyse relationnelle.
- Le choix des bases bibliographiques (Current-Contents, Pascal et Science Citation Index,) à partir desquelles a été établi le corpus INRA, a été strictement imposé par la double nécessité d'avoir les noms et adresses complètes de tous auteurs et de recencer toutes les activités de l'INRA.
- Après extraction des données brutes utiles, un travail considérable de normalisation (notamment de traitement des synonymies) a du être réalisé au niveau des adresses, des noms de pays, de villes et d'organismes, travail pour lequel « Tétralogie » est particulièrement bien adapté.

1 - LA NATURE DES PROBLEMES POSES

1.1 - Etudier l'évolution des collaborations internationales de l'INRA

La démarche que nous avions utilisée il y a deux ans pour la première étude et que nous avons reconduite pour cette actualisation se décompose en cinq points :

- Rechercher des bases consignant l'ensemble des adresses de co-signataires
- Retrouver la totalité des publications dans lesquelles est intervenu l'INRA depuis 1998
- Standardiser l'information: pays, villes, organismes, adresses, auteurs, journaux, …
- Découper le temps en périodes significatives par rapport aux données disponibles
- Croiser les informations nécessaires à l'étude.

Les évolutions sur quatre périodes homogènes sont alors déduites de ces croisements : nombre de citations dans les bases retenues, nombre de publications nationales, nombre de publications internationales, nombre de co-signatures avec des pays étrangers, stratégies de coopérations, réorientations thématiques, stratégies de publication.

1.2 - Analyser l'activité du département TPV

Ce département n'ayant pas de localisation propre, nous sommes partis de la liste de ses chercheurs pour établir un filtre permettant de récupérer, dans le corpus INRA, l'ensemble de sa production scientifique. Toutes les variations orthographiques des auteurs (plus de 550) ont été détectées et elles ont été synonymées à TPV. Un « retour aux notices » sur ce terme en mode multi-bases nous a permis d'extraire en une seule fois tout le corpus TPV. Celui-ci a ensuite été analysé en détail : détection de tous les acteurs, de leurs relations et des éléments sémantiques associés. Une analyse de l'évolution nous a ensuite permis de dégager, pour ce département, les composantes stratégiques et les signaux faibles.

1.3 - Analyser l'influence de l'INRA aux Caraïbes

Pour extraire l'activité des laboratoires INRA des Caraïbes (Guadeloupe, Martinique, Guyane), nous avons travaillé dans un premier temps sur les adresses : recherche des ces trois noms et de tous leurs synonymes dans le corpus INRA. Comme certains organismes non INRA de ces départements travaillent avec l'INRA de métropole, nous avons ensuite éliminé manuellement une quinzaine de publications où l'INRA local n'intervenait pas explicitement. Le corpus Caraïbe ainsi constitué a ensuite été analysé suivant les mêmes principes que celui de TPV. Mais l'influence de l'INRA s'exerce aux Caraïbes essentiellement depuis la métropole, nous avons donc fait un zoom sur les collaborations internationales de l'INRA avec les Pays de la région. Nous avons détecté les laboratoires concernés des deux cotés et les sujets de recherche abordés. Cet inventaire devrait permettre de promouvoir une meilleure coopération locale (entre INRA Caraïbe et pays de la région) puisque la métropole peut servir de pivot pour l'initialisation de nouveaux contacts.

2 - DESCRIPTION DES BASES RETENUES

2.1 - Base Current-Contents

Principales caractéristiques de cette base :

- Deux formats avant et après janvier 98.
- Plusieurs adresses depuis janvier 98.
- En avance sur l'indexation (articles sous presse)
- Propose un résumé et des mots-clés.

Format de l'adresse :

RP: de Lajudie, P; IRD; INRA,AGRO M,CIRAD; Campus Baillarguet,BP 5035; F-34032 Montpellier; France

IN: IRD, INRA,AGRO M,CIRAD, Lab Symbioses Trop & Mediterraneennes, F-34032 Montpellier, France; State Univ Ghent, Microbiol Lab, Ghent, Belgium; IRD, Lab Microbiol Sols, Dakar, Senegal

Périodes télé-déchargées et retenues : CD/Rom de 1998 à début 2002.
Certaines publications parues en 96 et 97 sont indexées sur ces CD/Rom.

2.2 - Base Pascal

Principales caractéristiques de cette base :

- Deux formats avant et après janvier 96.
- Sciences exactes et appliquées.
- Mots-clés en français, anglais et espagnol.
- Présence d'un résumé.

Format de l'adresse :

CS- Departement Etudes et recherches du CETIOM, 174 avenue Victor Hugo, 75116 Paris, France^GIE Cartisol, 20 rue Bachaumont, 75001 Paris, France^Unite associee a l'INRA Organisation et variabilite des genomes vegetaux, Universite Blaise Pascal, Campus des Cezeaux, 63177 Aubiere, France^INRA, Station d'amelioration des plantes et de pathologie vegetale, Domaine de Crouelle, 234 avenue du Brezet, 63039 Clermont-Ferrand, France^INRA-ENSAM, Station de genetique et d'amelioration des plantes, 2 place Viala, 34060 Montpellier, France|

Périodes télé-déchargées et retenues : CD de 1998 à avril 2002.
Certaines publications parues en 96 et 97 sont indexée sur ces CD.

2.3 - Base SCI sur le Web of Science

Principales caractéristiques de cette base :

- Plus de 3500 journaux
- Sciences exactes et appliquées
- Pas de descripteur ni de résumé sur le CD/Rom (étude précédente)
- Descripteurs et résumé disponibles sur le Web of science (réactualisation)
- Les citations n'ont pas été récupérées dans le corpus INRA (problèmes techniques à répétition lors du déchargement en ligne).

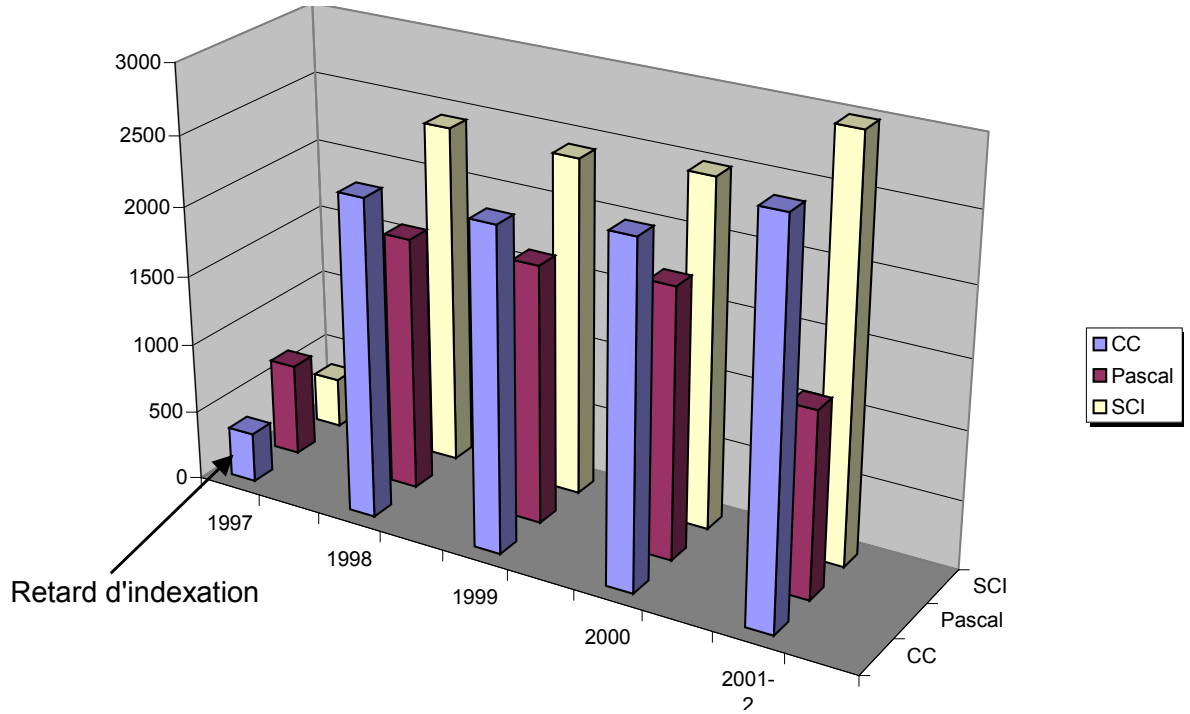
Format de l'adresse :

C1 Univ Munster, Dept Dermatol, Von Esmarchstr 56, D-48149
Munster, Germany
Univ Munster, Dept Dermatol, D-48149 Munster, Germany
Univ Hosp Rudolf Virchow, Inst Human Genet, Berlin, Germany
INRA, CEA, Lab Radiobiol Appl, Jouy En Josas, France
NIH, Skin Biol Lab, Bethesda, MD 20892 USA

Périodes télé-déchargées: 1998 à avril 2002.

2.4 - Répartition par bases et par années

Années/Bases	CC	Pascal	SCI	Sigma
1997	357	665	359	1381
1998	2290	1826	2445	6561
1999	2314	1858	2427	6599
2000	2446	1938	2501	6885
2001-2	2808	1337	3039	7184



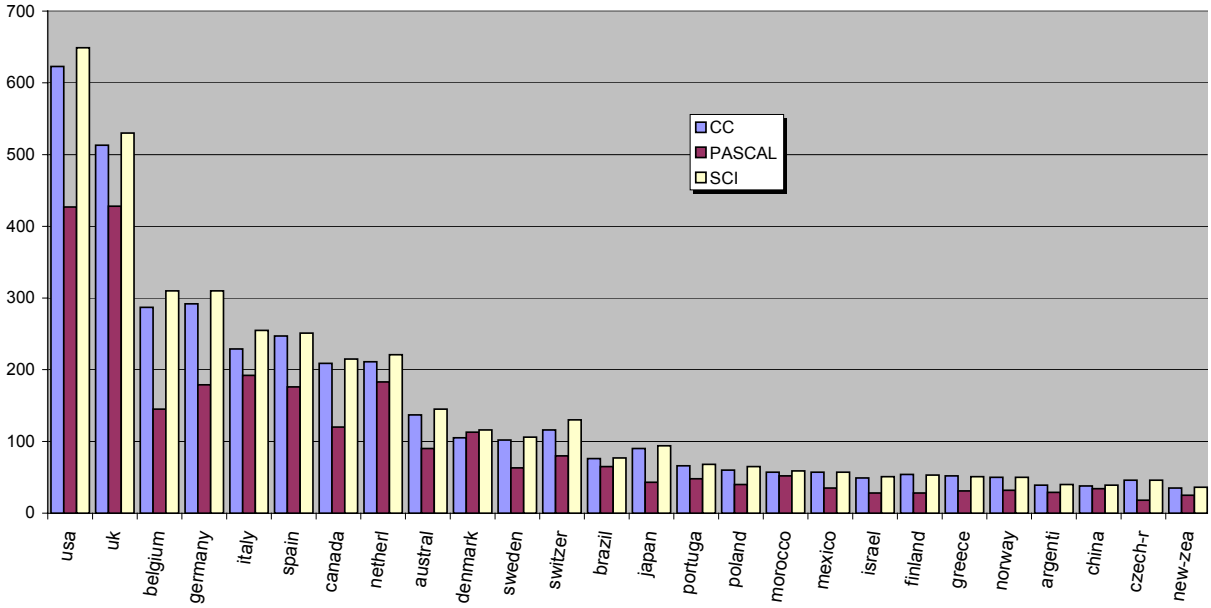
3 - L'ANALYSE DES COLLABORATIONS INTERNATIONALES

3.1 - Avantages et inconvénients de l'analyse multi-base

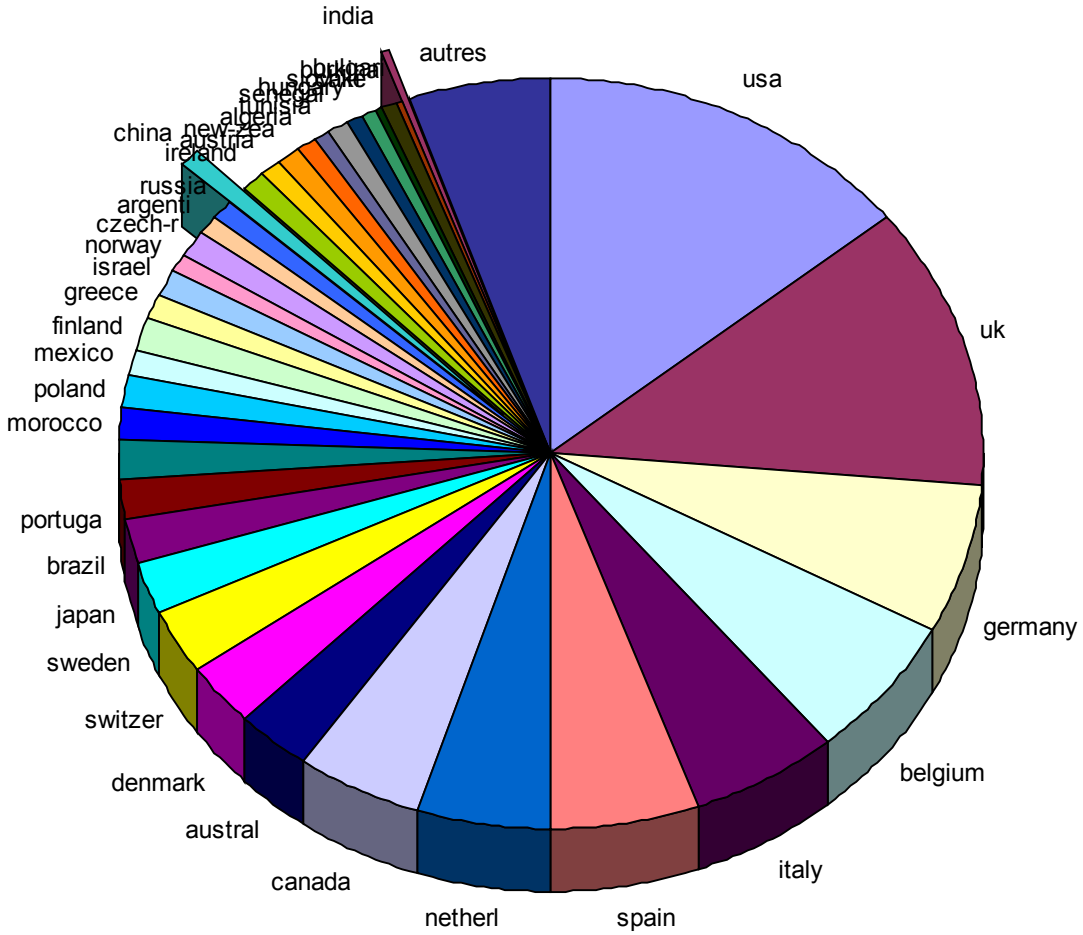
Cette analyse conduite simultanément sur les trois bases présente les caractéristiques suivantes :

- Réponse simultanée des trois bases.
- Etude de l'évolution la plus récente (avant les formats diffèrent et pas de CC).
- Assez bonne homogénéité des quatre sous périodes : 1998, 1999, 2000 et 2001-02.
- Léger décalage de CC et SCI avec PASCAL (retard d'indexation pour PASCAL).
- La dernière sous période (01-**02**) a donc un déficit sur PASCAL.
- La dernière sous période (01-**02**) est sous représentée. (Indexation encore incomplète).
- Les valeurs brutes ne sont pas vraiment significatives.
- Il faut travailler si possible en évolution relative

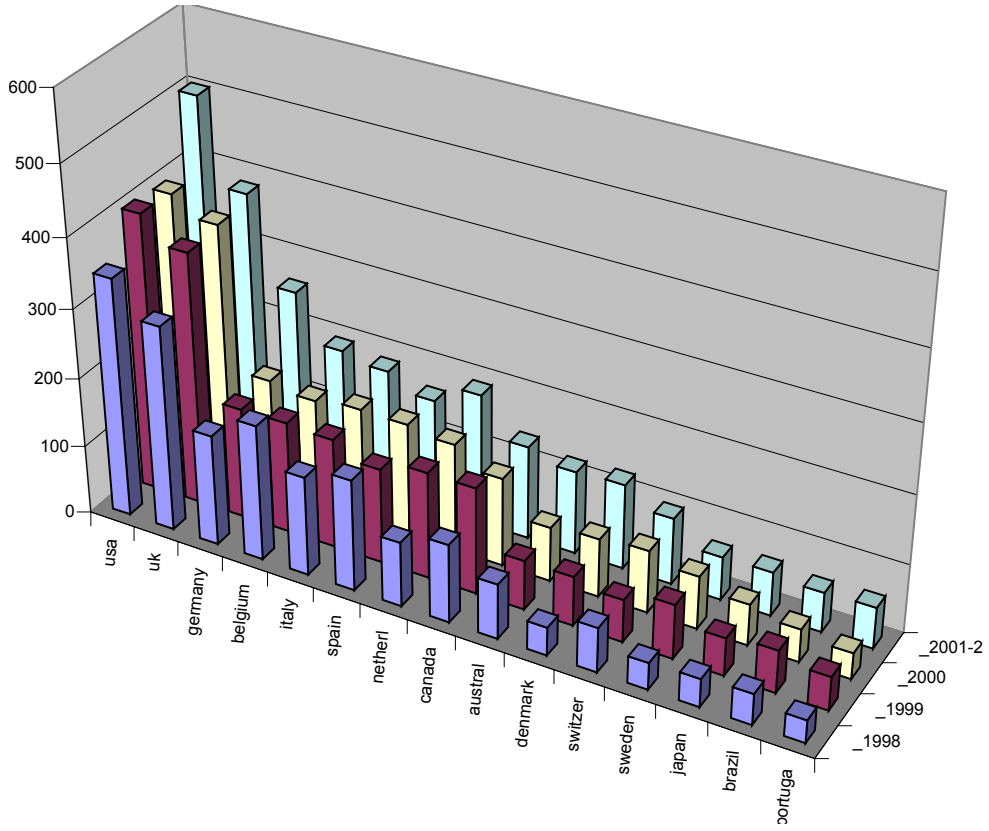
3.2 - Citations dans les trois bases par pays



3.3 - Répartition des pays pour l'ensemble du corpus



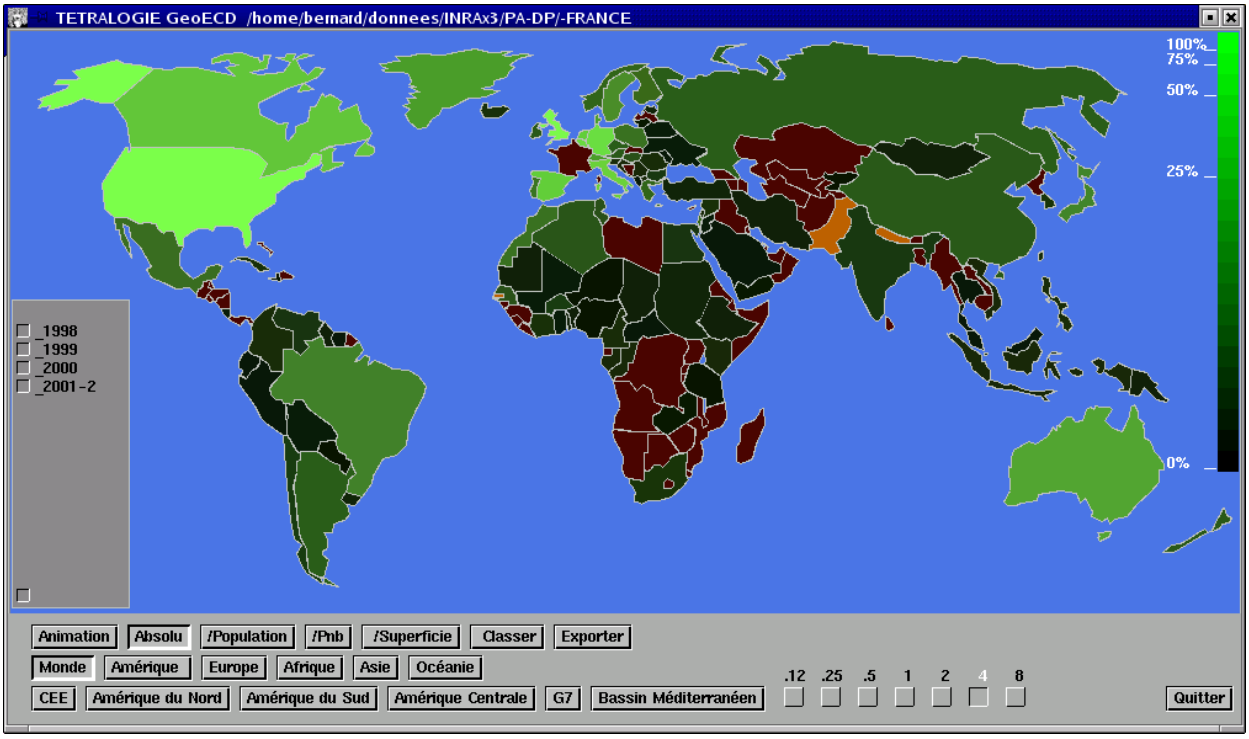
3.4 - Présence absolue dans les 3 bases sur les 4 périodes



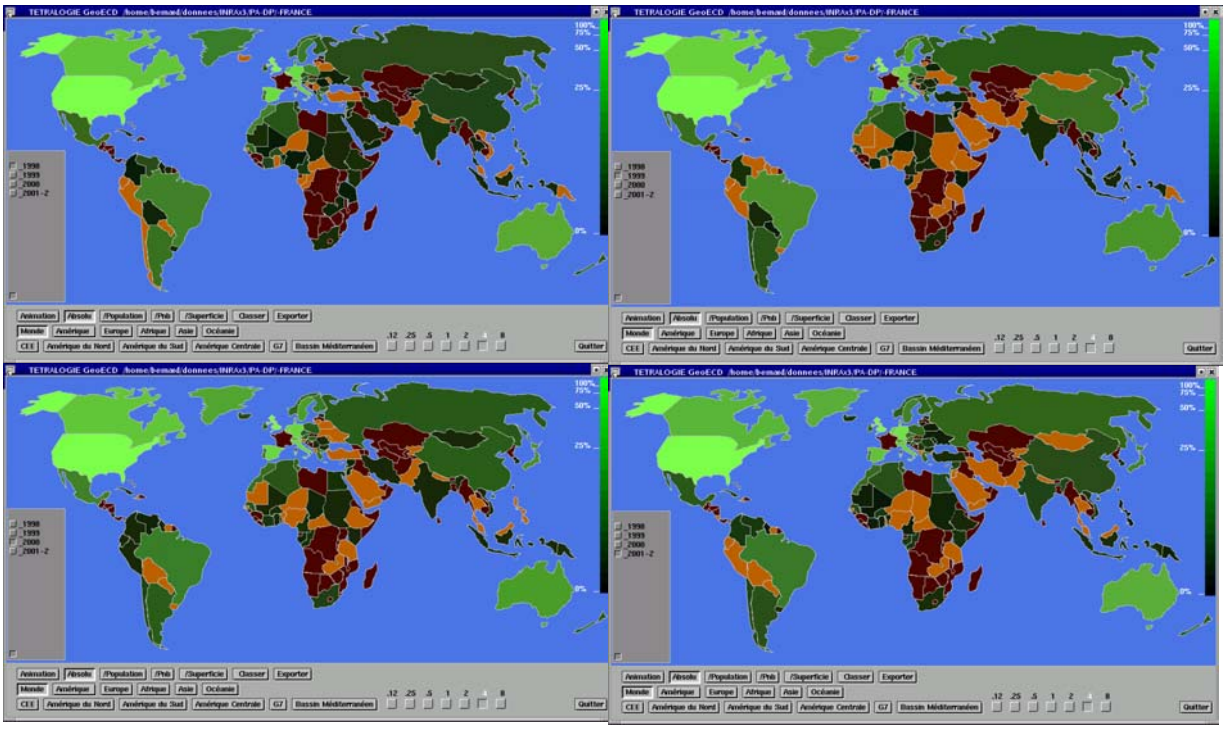
3.5 - Cartes d'implantation des collaborations INRA

3.5.1 - Carte couvrant l'ensemble des périodes de 1998 à début 2002

La France a été enlevée de cette représentation, où figurent en grenat les pays absents, en orange les pays présents avant 1998 et perdus depuis, en nuances de vert (du noir au vert vif) le nombre d'articles co-signés, suivant une échelle privilégiant les signaux faibles.



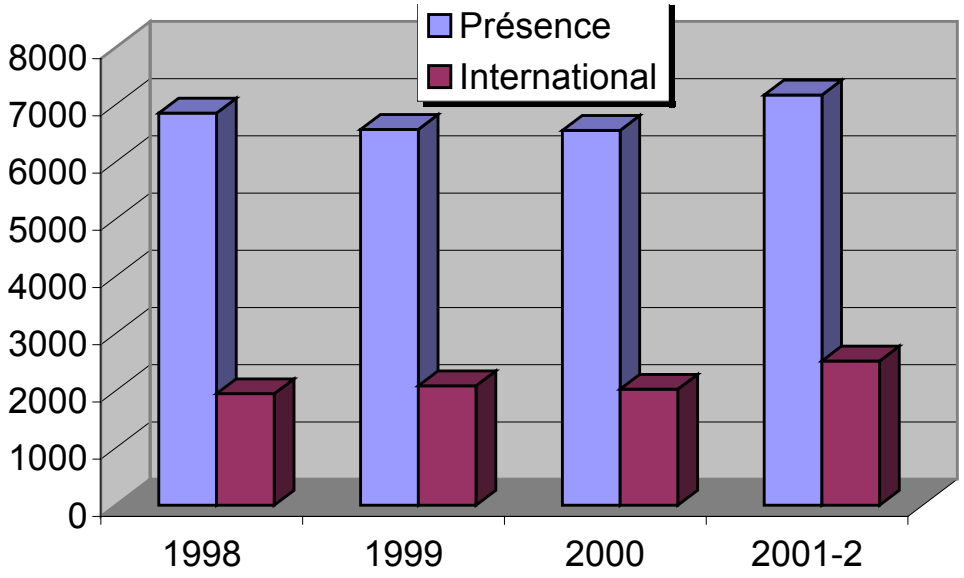
3.5.2 - Cartes comparatives pour les quatre périodes retenues



3.6 - Evolution des collaborations internationales sur la périodes 1998 à 2001-2

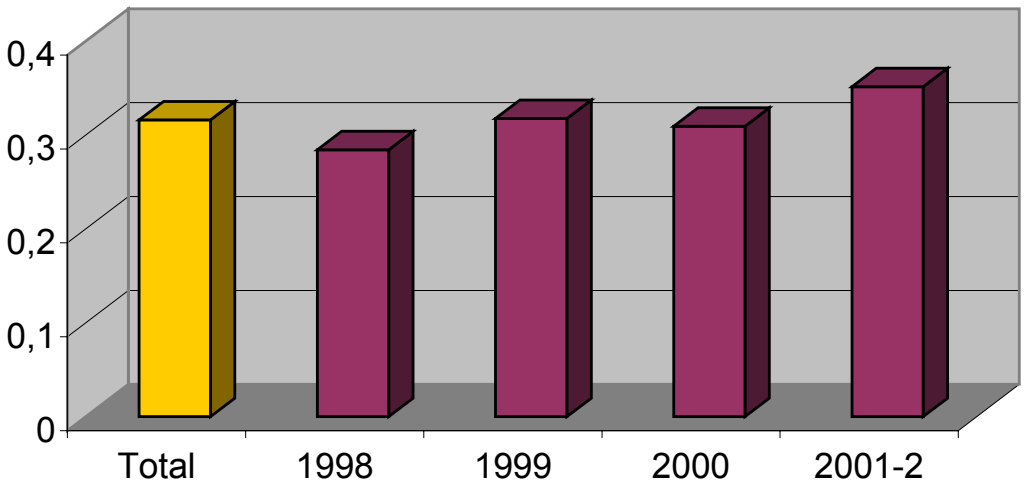
3.6.1 - Répartition des références dans le temps

En valeur absolue le nombre des publications internationales augmente. Mais cette mesure n'est pas fiable car les chiffres de chaque période ne sont pas obtenus dans les mêmes conditions. Les années 1998 et 1999 sont pratiquement indexées à 100%, pour 2000 il subsiste un léger déficit d'indexation surtout pour Pascal, enfin la période 2001-02 couvre 16 mois avec un fort déficit d'indexation pour 2001 et une indexation très faible des 4 premiers mois de 2002.



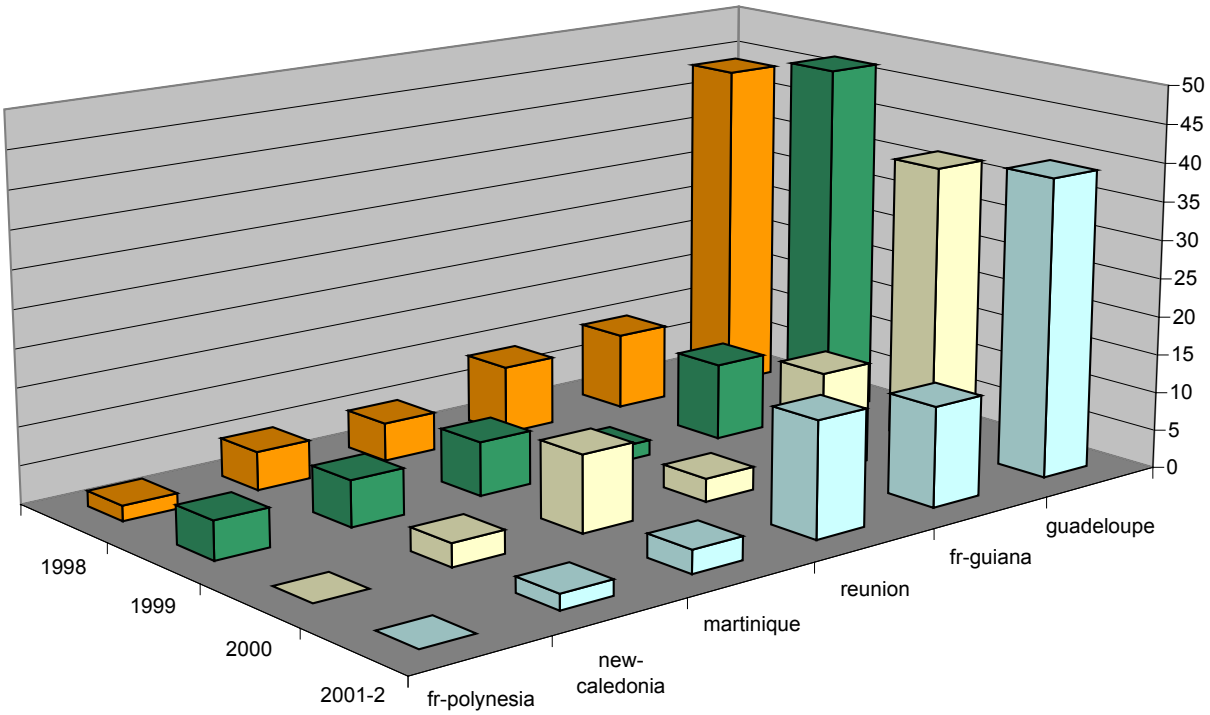
3.6.2 - Part des publications internationales

Si nous travaillons en proportion de publications internationales, nous remarquons une progression de 5 points par rapport à 1998. Les publications visibles de l'INRA sont donc de meilleure qualité, d'autant plus que la couverture des bases augmente et que des revues de second rang sont maintenant indexées.



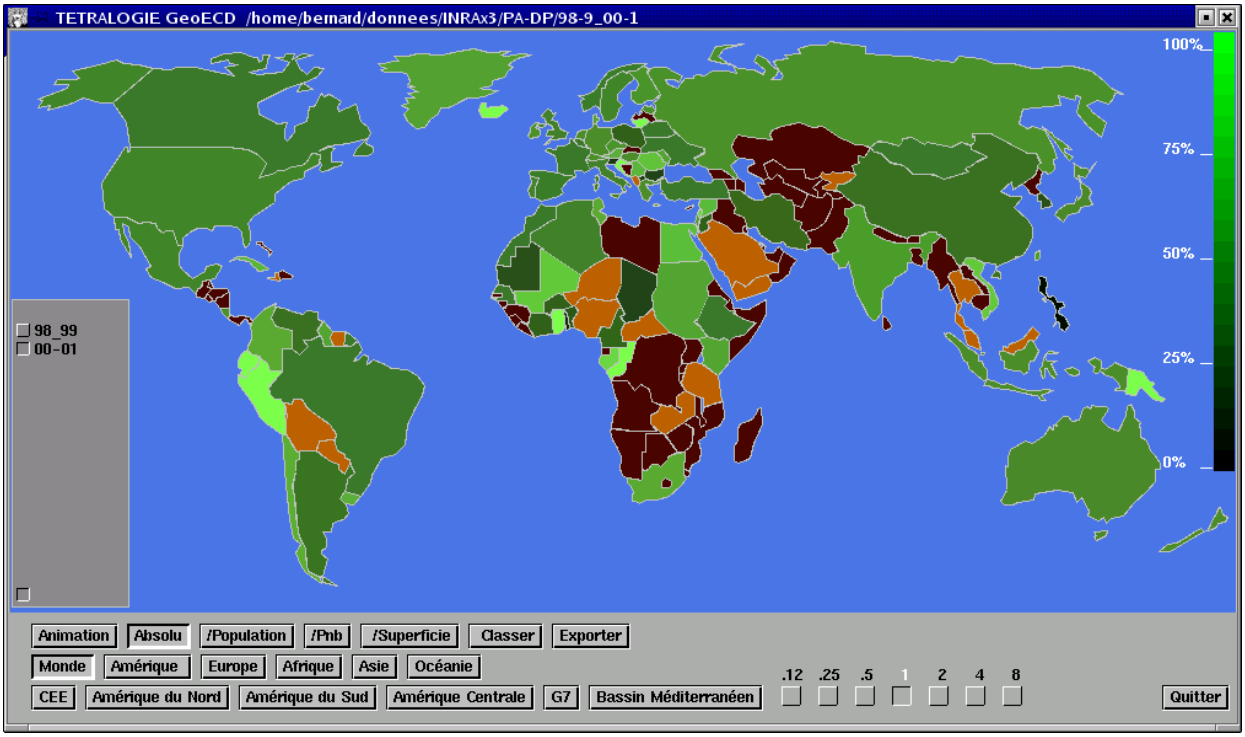
3.6.3 - Présence et évolution dans les Dom/Tom

Nous pouvons, par le biais des adresses, retrouver l'origine régionale des publications référencées. Nous avons ainsi pu isoler la production des Dom/Tom (INRA ou collaboration directe avec l'INRA). La Guadeloupe est sans conteste en première position, mais nous constatons un léger déclin des Dom/Tom sauf pour la Réunion.



3.6.4 - Evolution entre 1998-99 et 2000-01 dans les trois bases

Codes des couleurs utilisées dans les cartes géographiques :
En **grenat** : pays absents sur l'ensemble des périodes.
En **orange** : pays perdus après 2000 mais présents avant.
En **vert** part de la période 2000-01 sur la totalité.
Vert sombre diminution des collaborations
Vert moyen (50%) maintient
Vert clair augmentation des collaborations.
Dans l'exemple ci-dessous : l'Islande augmente, la Pologne diminue.



3.7 - Conclusions globales de cette nouvelle étude

Le nouveau système de cartes géographiques interactives propre à « Tétralogie » permet ainsi de formuler les principales conclusions suivantes :
Les publications internationales de l'INRA

- Se concentrent sur moins de pays
- Leur nombre augmente
- Leur proportion augmente (de 27 à 34%) par rapport aux publications franco-françaises.

On peut en conclure que la qualité et le niveau des publications augmentent (publications acceptées dans les grandes revues référencées dans les bases).
Par comparaison avec l'étude précédente (analyse 1998-99) , la période 2000-01 montre dans la dynamique des collaborations internationales de l'INRA :

- L'arrêt du déclin africain
- La confirmation du déclin en Asie mineur et centrale, sauf Méditerranée
- La retombée de la Chine
- L'augmentation de l'Inde et du Vietnam
- L'émergence de certains sujets de recherche à l'international.

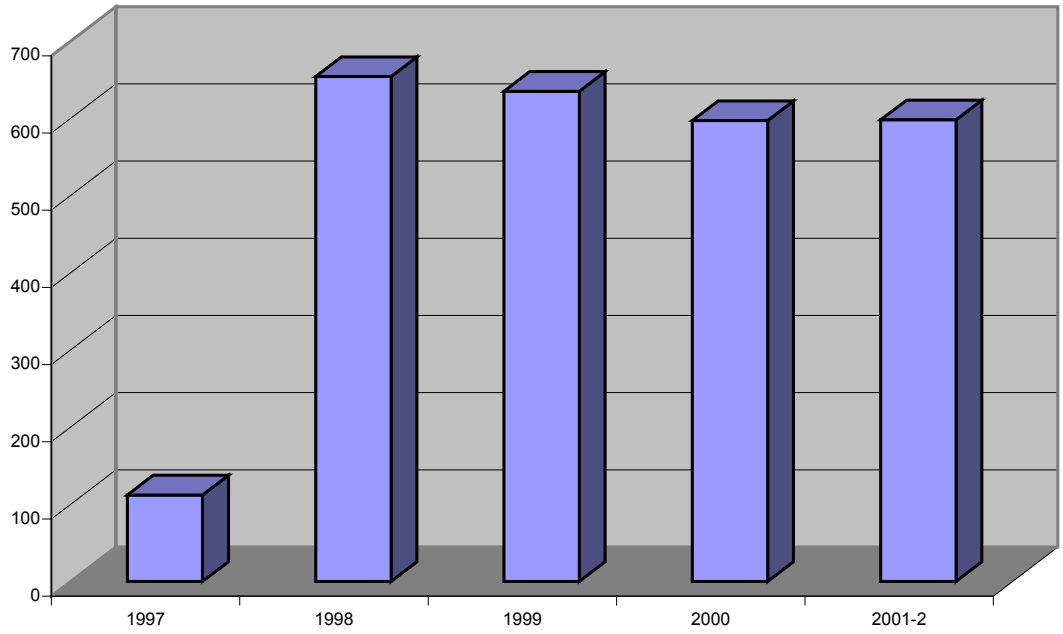
4 - ANALYSE DU DEPARTEMENT TPV

4.1 - Identification du département

La liste de tous les auteurs de TPV et des variations orthographiques recensées a été synonymée au mot TPV. Cette méthode permet soit de filtrer la production de TPV dans le corpus INRA, soit d'extraire un sous corpus TPV en recherchant automatiquement les notices bibliographiques correspondantes dans les trois bases étudiées.

ABBAL-P	TPV
ABBAL, P	TPV
ABECASSIS-JOEL	TPV
ABECASSIS-J	TPV
ABECASSIS, J	TPV
AGEORGES-AGNES	TPV
AGEORGES-A	TPV
AGEORGES, A	TPV
AGUIE-BEGHIN-T-V	TPV
AGUIE-BEGHIN-V	TPV
AGUIE-BEGHIN, V	TPV
AGUIEBEGHIN-V	TPV
AGUIE-BEGLUN, V	TPV

Ci-dessous la répartition dans le temps des articles retenus.



4.2 - Principaux résultats

Dans cette étude, nous avons détecté la totalité des acteurs connexes à TPV : auteurs, journaux, pays, villes, organismes, universités, autres centres INRA, les thèmes de recherche et les signaux faibles. Nous avons croisé toutes ces entités entre elles afin de savoir qui fait quoi avec qui et depuis combien de temps. Nous avons aussi étudié l'évolution des équipes de recherche et leurs collaborations nationales et internationales. Par contre, nous n'avons pas encore utilisé nos méthodes d'indexation du texte libre (Titres et Résumés) afin de rendre homogène les trois bases sur le plan sémantique. En effet, une intervention des experts du domaine est nécessaire pour valider les dictionnaires de mots composés qui sont extraits du corpus par des méthodes de détection statistique. Une fois cette étape réalisée, il sera possible de connaître, avec beaucoup plus de précision qu'en utilisant les mots-clés, l'étendue du front de recherche actuel et les émergences de nouveaux sujets.

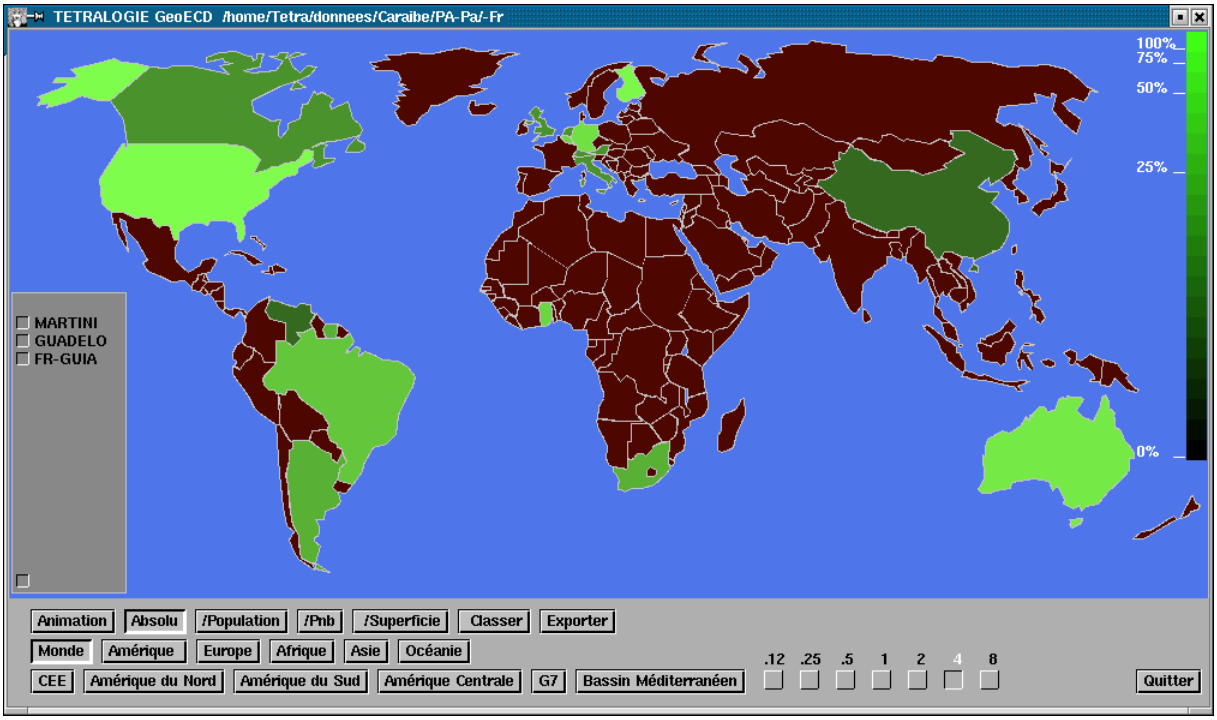
5 - ANALYSE DE L'INFLUENCE DE L'INRA AUX CARAÏBES

5.1 - Les collaborations avec les Caraïbes

	CC	PASCAL	SCI
GUADELOUPE	**58**	**53**	**63**
MEXICO	**57**	**35**	**57**
FR-GUIANA	**16**	**14**	**15**
CUBA	**9**	**7**	**9**
VENEZUELA	**10**	**5**	**13**
MARTINIQUE	**8**	**8**	**9**
COLOMBIA	**9**	**2**	**8**
GUYANA	**4**	**0**	**2**
COSTA-RICA	**1**	**4**	**1**
SURINAME	**1**	**1**	**1**
JAMAICA	**1**	**0**	**1**
HAITI	**0**	**1**	**0**

5.2 - Les laboratoires INRA des Caraïbes

Si on se restreint aux laboratoires INRA des Caraïbes explicitement identifiés par leur adresse, nous pouvons dresser la carte des collaborations internationales. Nous remarquons, par exemple, l'absence du Mexique, de la Chine. En fait, ces pays collaborent avec l'INRA de métropole et un laboratoire en agronomie de la région Caraïbe qui n'est pas explicitement l'INRA (Cirad, Orstom, Ifremer, …).



CONCLUSION

Cette seconde étude complétée par deux zooms spécifique a permis de montrer l'étendue des possibilités de la bibliométrie pour extraire d'un corpus conséquent (plus de 30 000 références) des informations stratégiques pour la conduite et le développement d'un grand institut de recherche. L'identification précise des acteurs et la détection de toutes leurs relations permet de mieux comprendre le fonctionnement d'une telle entité. Le suivi de ses relations internationales dévoile les forces et les faiblesses en terme de collaboration et de coopération ainsi que des stratégies peu visibles autrement. Chaque département peut être analysé séparément notamment sur le plan sémantique (domaine plus ciblé que l'agronomie dans son ensemble), ses caractéristiques structurelles et son fonctionnement sont alors mieux compris, ses orientations mieux évaluées, sa lisibilité mesurée, son influence délimitée. Pour une région comme celle des Caraïbes, le but d'une telle étude est de détecter tous les partenaires locaux potentiels, les contacts qu'ils ont déjà avec l'INRA et les domaines dans lesquels ils collaborent. Une extension naturelle de cette étude est la mise en ligne de l'analyse du corpus INRA sur un portail (Internet ou Intranet). Les dictionnaires et les matrices sont alors compilés dans une base de données afin de permettre à un utilisateur de faire sa propre analyse (micro analyse) en scrutant son environnement (acteurs et thèmes de recherche). Des statistiques en ligne doivent l'aider à appréhender les éléments stratégiques qui le concernent, à se situer dans le contexte national et international et à détecter à temps toute innovation pouvant l'intéresser. La macro analyse, que nous venons de produire, lui servant de référentiel tout au long sa navigation et de ses investigations.

BIBLIOGRAPHIE

[KARO99] S. Karouach, T. Dkaki, B. Dousset

Visualisation interactive de classifications d'informations. 8ièmes journées d'études sur les systèmes d'information élaborée: Bibliométrie - Informatique stratégique - Veille technologique, (Ile Rousse Corse France), CD-ROM, p. 45-61, 27 septembre-1° octobre 1999.

[ROUX99] C. Roux , B. Dousset

Une méthode de détection des signaux faibles: application à l'émergence des Dendrimères. Veille stratégique, scientifique et technologique : VSST'98, pp 349-357, (Toulouse, France), octobre 1998.

[DOUS00] B. Dousset, T. Dkaki, J. Mothe

Information mining in order to graphically summarize semi-structured document. 17th international CODATA Conference, (Baveno Italie), 15-19 octobre 2000.

[HUBE00] G. Hubert, J. Mothe, A. Benammar, T. Dkaki, B. Dousset, S. Karouach

Textual document Mining using graphical interface. International Human Computer Interaction, HCI International 2001 , New Orleans (USA). Lawence Erlbaum Associates - Publishers , Mahwah - New Jersey, pp 918-922 (volume 1), 05-10 août 2001.

[KARO01] S. Karouach, B. Dousset

Visualisation interactive pour la découverte de connaissances. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 291-300, (Barcelone, Espagne), octobre 2001.

[MOTH01] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, D. Egret

Information mining: use of the document dimensions to analyse interactively a document set. 23rd BCS European Colloquium on IR Research: ECIR, Darmstadt. BCS IRSG, pp 66-77, 4-6 avril 2001.

[MULT01] J.-L. Multon, G. Lacombe, B. Dousset

Analyse bibliométrique des collaborations internationales de l'INRA. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 261-270, (Barcelone, Espagne), octobre 2001.

[SOSS01] D. Sosson, M. Vassard, B. Dousset

Portail pour la navigation en ligne dans les analyses stratégiques. Veille stratégique, scientifique et technologique : VSST'01, Vol 1, pp 347-358, (Barcelone, Espagne), octobre 2001.

[DOUS02] B. Dousset, S. Karouach

Collaboration interactive entre classifications et cartes thématiques ou géographiques. 9ièmes rencontres de la société francophone de classification, (Toulouse France), 16-18 septembre 2002.