île Rousse 2005
Journée sur les systèmes d'information élaborée

MESURE DE L'APPORT INFORMATIONNEL DES CORPUS A L'ORGANISATION DE L'ACTIVITE COLLECTIVE

Mathilde de Saint Leger(*), William Turner(**)
stleger@limsi.fr, turner@limsi.fr

(*) : DIST-CNRS, BP 133, Bat. 508 Université de Paris Sud, 91403 Orsay

(**) : LIMSI-CNRS, BP 133, Bat. 508 Université de Paris Sud, 91403 Orsay Cedex

Mots clés : infométrie, organisation d'une activité scientifique, bibliométrie qualitative

Résumé
L'objectif de ce texte est de montrer comment établir une mesure de l'information fondée sur le pouvoir que cette information exerce sur la structuration d'un champ de recherche. Trois hypothèses sont à l'origine de ce travail :

- une information pertinente pour comprendre l'état d'une activité collective est celle qui permet de comprendre l'équilibre des forces travaillant à la consolidation ou, au contraire, à la problématisation d'un champ de recherche (Callon M., Courtial J.P., Turner W.A, 1991). Plus concrètement, un champ de recherche est caractérisé par un ensemble de thèmes qui sont reconnus comme étant centraux et ceux qui sont, au contraire, périphériques. Puisque la répartition des ressources (ressources humaines, argent, …) est généralement plus favorable aux premiers, l'émergence de nouveaux thèmes suppose la remise en cause de cette stratégie de consolidation du « mainstream ». Une information pertinente est une information qui aide à comprendre où est tracée la frontière entre le « mainstream » et le « périphérique » dans un champ de recherche.
- Un corpus documentaire permet au professionnel de l'information de faire apparaître la structure d'un champ de recherche puisque la frontière « mainstream/périphérique » est intrinsèque (de Saint Leger, M. 1997). Les calculs servant à l'identifier doivent tenir compte :
  - de la « typicalité » d'un document – à savoir, sa contribution spécifique à l'établissement de la frontière ;
  - de l'impact de sa publication dans le déplacement de la frontière entre la consolidation et la remise en cause des structures du champ.

Enfin, l'information pertinente est portée par l'usage des mots en contexte. C'est ce que nous allons montrer à partir de l'étude d'un corpus comportant plus de 1000 textes présentés dans le cadre du premier colloque de l'Association Française de Sociologie (l'AFS) en 2004 (de Saint Leger, M., van Meter, K, 2005). Ce corpus représentait fidèlement la recherche française en sociologie organisée autour d'une quarantaine de réseaux thématiques (RT) identifiés par les organisateurs du colloque. Chaque RT était sensé définir un axe spécifique d'activité et c'est précisément ce que nous nous proposons d'évaluer à l'aide de nos mesures.

1. Introduction

Dans le cadre du premier colloque de l'Association Française de Sociologie (l'AFS) en 2004, plus de 1000 contributions ont été présentées. Ce corpus ainsi rassemblé représentait fidèlement la recherche française en sociologie organisée autour d'une quarantaine de réseaux thématiques (RT). Partant de ce corpus, et en s'appuyant sur une mesure de l'information fondée sur son pouvoir de structuration d'un champ de recherche, nous allons montrer comment comparer, décrire et mettre en relation le contenu et les acteurs qui constituent trois des RT observés : "Sociologie du droit", "Migrations et relations interethniques" et "Travail, rapport sociaux, rapport de genre".

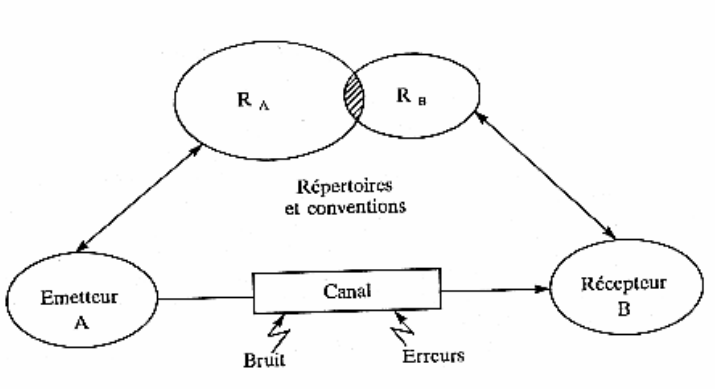
Les différentes mesures de l'information souvent proposées sont principalement liées aux problèmes de filtrage de l'information et localisation des ressources pertinentes. Nous avions proposés d'ajouter au panel de ces outils infométriques déjà existants, une mesure de pertinence qui ne serait pas d'abord fonction du bruit et du silence (ou encore de la précision et du rappel) introduits dans une collecte, mais plutot une mesure de la capacité d'une information à permettre l'ajustement dynamique d'une requête pour un filtrage pertinent (de Saint Leger, M. 1997). En partant des mêmes principes, nous proposons aujourd'hui une mesure de pertinence dont l'objectif est de mieux comprendre où est tracée la frontière entre le "mainstream" et le périphérique dans un champ de recherche. Nous faisons l'hypothèse que le flux d'information analysé est pertinent, en l'occurrence, ce corpus de 1046 contributions présentées au congrès de l'AFS en 2004, représente la recherche française en sociologie en 2004.

2. Construction de cette mesure

2.1 Pourquoi une nouvelle mesure de l'information ?

La mesure de l'information dépend de l'approche systémique qui en est faite. De notre point de vue, elle est le résultat de l'interaction permanente entre les données et les usagers et de ce fait, elle n'existe pas en soi, « l'information ne peut être vue sans son contexte d'usage » (Turner 96). C'est donc un processus dynamique d'interaction entre les données et les usagers. Dans le cadre des publications scientifiques, l'information prend sa valeur effective dans l'échange et le partage, "L'article exprime donc de façon autorisée ce que font, veulent les acteurs humains ou non humains qu'il met en scène et dont il s'établit un porte-parole légitime ... Comme l'interlocuteur industriel, comme l'équipement, comme le chercheur recruté, le document rend présente dans le laboratoire une série d'autres actants, humains et non humains, dont il porte la parole et auxquels il donne accès" (Callon M., 1989). Pour étayer notre démarche nous proposons de l'opposer à la théorie de l'information de Shannon qui a pris corps il y a une quarantaine d'année. Elle est issue des recherches sur l'optimisation de la transmission des messages entre deux systèmes physiques A et B appelés respectivement émetteur et récepteur (voir figure suivante). L'émetteur peut se mettre dans un nombre d'états finis (x, y, et z par exemple) connus à l'avance par le récepteur. Un message émis à travers le canal de transmission est une combinaison de ces états (xxz par exemple). Or le bruit inhérent à toute transmission quel que soit le canal, masque le message. L'objectif de cette théorie est de mesurer la quantité d'information émise pour que le récepteur puisse interpréter correctement le message reçu. La quantité d'information que peut délivrer une source est d'autant plus élevée que le nombre de ses états possibles (donc ses messages) est grand : avec deux états on a au plus quatre combinaisons possibles, avec trois états on a au plus huit combinaisons etc. De plus la quantité d'information émise par une source est fonction de la probabilité d'apparition de chaque événement. En poursuivant ce raisonnement, Shannon montre que l'information au sens physique est une grandeur mesurable qui suppose connus tous les états possibles de l'émetteur (R_A) et leur probabilité d'apparition dans un message (Raisbeck G.,1964).

Un codage et un décodage adéquats des messages, respectivement à l'émission et à la réception permet de pallier le bruit apparu au cours de la transmission.



- Schéma général de la transmission d'information au sens de Shannon (Recoque A., 1991) -

Dans un système physique tel qu'il a été étudié par Shannon, les référentiels R_A et R_B sont identiques et le système est indépendant de son utilisateur. Comme par exemple les spams du courrier électronique qu'il est devenu fréquent de recevoir. Cette quantité d'information délivrée sous forme de signal électrique codé est transmise puis décodée et délivrée sous sa forme initiale. On peut construire une mesure de l'information avec ce système où l'utilisateur est exclu, cela permet de dire si des documents reçus sont identiques aux documents émis dans leur forme et leur contenu : l'information présentée à l'entrée du système peut au mieux se conserver jusqu'à sa sortie, il n'y a pas de création de sens et le bruit généré au cours de la transmission est gênant, il faut donc des algorithmes de vérification pour le réduire.
Dans notre approche systémique pour établir une mesure de l'information, nous mettons l'usager au cœur du système. L'objectif ici est de mieux comprendre l'état d'une activité collective et l'équilibre des forces travaillant à sa consolidation ou, au contraire, à sa problématisation (Callon M., Courtial J.P., Turner W.A, 1991). Plus concrètement, un champ de recherche est caractérisé par un ensemble de thèmes qui sont reconnus comme étant centraux ou, au contraire, périphériques. Une information pertinente est de notre point de vue, celle qui aide à comprendre où est tracée la frontière entre le « mainstream » et le « périphérique » dans un champ de recherche. Ainsi quand nous disposons d'un corpus documentaire representant un champ de recherche, le premier objectif est de le découper en sous-ensembles thématiquement homogènes et de pouvoir dire si tel sous-ensemble traite de thématiques centrales ou au contraire périphériques. Bien entendu le classement ne peut être binaire, mais l'idée est de mieux comprendre comment s'articule la frontière entre ces deux catégories. Nous proposons de le faire en modélisant le poids des mots du corpus suivant leur pouvoir de structuration du champ ou encore suivant leur capacité à agréger des termes pour former un réseau lexical qui indexe un sous-ensemble du corpus thématiquement homogène. Nous aurons alors 3 types de termes : central, périphérique et les autres. Le *bruit*[1] provoqué dans le système par les facteurs aléatoires de l'environnement ne serait plus un bruit au sens classique documentaire, à partir du moment où il serait utilisé par le système comme facteur d'organisation autrement dit, si l'usager se l'approprie pour l'integrer dans ces futures publicatoins. Le poids, appelé alors *pouvoir d'attraction* du mot, devient une mesure qui permet de rechercher des documents avec des termes centraux, et/ou périphériques et/ou bruités.

2.2 Découpage du corpus en classes thématiques homogènes et mesure de l'information

Pour le découpage du corpus en classes thématiquement homogènes, plusieurs méthodes sont possibles. Nous avons choisi une classification hiérarchique ascendante avec la méthode des mots associés fondée sur la réponse à la question suivante : "Comment faire apparaître dans les différents

[1] La notion de bruit, ne s'entend pas ici au sens classique documentaire, un mot est un bruit s'il n'est pas suffisamment coocurrent dans le corpus pour indexer un sous-ensemble documentaire, mais il est peut être en emergence.

documents qui circulent (articles scientifiques, brevets, rapports, modes d'emploi...) les mises en relation opérées par les acteurs ?" (Callon M., Courtial J.P., Turner W.A; 1991). Cette méthode de classification est fondée sur la coocurrence des termes dans les documents. La mesure de *ressemblance* ou *lien de similarité* est la cooccurrence *normalisée*[2] des paires de mots dans un même document. Ainsi des réseaux lexicaux se forment, les premiers étant les plus cooccurents dans les documents qui sont de ce fait, les plus proches thématiquement. Chaque cluster construit, est caractérisé alors par deux variables :

- sa *densité* calculée à partir de ses liens internes (ou liens intra cluster). Elle reflète la cohérence des pôles thématiques représentés par le cluster. Plus les liens internes sont forts, plus les mots liés sont systématiquement associés entre eux dans les documents.
- sa *centralité* calculée à partir des liens externes ou (liens extra clusters). Elle reflète le pouvoir de liaison des pôles thématiques représentés.

Un cluster indexe tous les documents qui ont participé à sa construction (ce sous-ensemble est appelé alors « méta documents »). S'il a une forte centralité et une forte densité, les documents qui y sont rattachés sont fédérateurs par leur pouvoir de liaison et structurant par leur cohérence interne.

Nous pouvons alors construire une mesure du poids du mot en fonction de sa participation à la construction des méta-documents en tenant compte de (de Saint Leger M., 1997):

- Sa *typicalité* à représenter spécifiquement un méta-document, un mot est d'autant plus typique, représentatif ou spécifique de ce dernier, que sa distribution fréquentielle y est concentrée.
- Sa *participation effective* à la construction d'un méta-document. Cette mesure est d'autant plus forte que la probabilité de présence d'un terme dans un méta-document comparée à son occurrence dans le corpus y est élevée.

Soit le mot m(i) du corpus, son *pouvoir d'attraction* est alors fonction de ces 2 paramètres :

$$P(i)= f(Ty,Pe)$$

En utilisant le modèle vectoriel classiquement employé dans la problématique de recherche d'information (Salton A. Wong C. S. Yang, 1985), le corpus est représenté par un vecteur dans l'espace des mots, ses coordonnées sont les poids des mots qui l'indexent. On peut alors definir la quantité d'information véhiculée par ce corpus comme étant sa norme dans l'espace des mots. Mais cela n'a de sens que si cette norme converge quelque soit la taille du corpus.

2.3 Convergence de la norme et mesure de proximité de 2 corpus

Dans l'espace vectoriel des mots, si toutes les coordonnées (ici pouvoir attracteur des mots) d'un vecteur, suivent une distribution[3] strictement majorée par une fonction de la forme : $C \cdot \frac{1}{i^a}$ où "C" est une constante dépendant du corpus, "i" est le rang du terme m(i) dans la distribution et "a" est une constante supérieure à ½, alors la norme converge. Nous allons montrer que c'est le cas ici.

Démonstration :

Pour commencer, nous rappelons l'expression des coordonnées du corpus:

$$p(i) = fontion(Ty.pe(i))$$

or Ty.pe(i,j) < C/i d'après sa définition.

Donc :

[2] La ressemblance ou le lien entre deux termes est la probabilité de trouver une paire de mots dans un même document.

[3]Par distribution, on entend le classement des coordonnées par ordre décroissant.

$$p(i) < C' \cdot \frac{1}{i} \qquad \textit{avec C'=constante qui dépend du corpus}$$

et le carré de la norme du vecteur corpus est alors majoré par :

$$C'^2 \cdot \sum_{i=1}^{n} \left(\frac{1}{i^2} \right)$$

$$\text{or : } \lim_{n \to +\infty} \sum_{i=1}^{n} \left(\frac{1}{i^2} \right) < \int_{x=1}^{+\infty} \frac{1}{x^2}$$

$$\text{et } \int_{x=1}^{+\infty} \frac{1}{x^2} = \left[-\frac{1}{x} \right]_1^{+\infty} = 1$$

Le carré de la norme du vecteur corpus est majoré quelque soit sa taille par C'², donc la norme converge et on peut appliquer le cosinus de Salton (Salton G. and Mac Gill M.J, 1984) comme mesure de proximité thématique entre deux corpus. Si p(i) désignent le pouvoir d'attraction du terme i d'un corpus et p'(i) son pouvoir d'attraction dans un deuxième corpus, alors :

$$s = \frac{\sqrt{\sum_i p(i).p'(i)}}{\sqrt{\sum_i p^2(i)} . \sqrt{\sum_i p'^2(i)}} \quad \text{avec } 0 \le s \le 1$$

s désigne alors la proximité thématique des deux corpus : si s=0, alors les 2 corpus n'ont aucun lien thématique, à l'inverse, si s=1 alors les deux corpus traitent exactement des mêmes thématiques.

Dans la suite de cet article, nous allons montrer concrètement l'intérêt de telles mesures dans l'étude de l'organisation de l'activité collective.

2.4 Applications de ces mesures

Partant d'un corpus de documents en texte intégral, on obtient une liste de mots valués et classés en 4 catégories :

- Catégorie de termes appartenant à des thématiques centrales et homogènes
- Catégorie de termes appartenant à des thématiques centrales et peu homogènes
- Catégorie de termes appartenant à des thématiques périphériques mais homogènes
- Catégorie de termes appartenant à des thématiques ni centrales ni homogènes

Ainsi, un document est fortement central s'il est indexé par de nombreux termes centraux, à l'inverse un document est périphérique s'il est indexé par de nombreux termes périphériques. Un troisième est un "bruit" s'il est indexé par des termes de la dernière catégorie.Pour mieux comprendre l'état d'une activité de recherche à partir d'un corpus la représentant fidèlement, on peut alors revenir sur les documents indexés par une ou plusieurs de ces catégories et construire une bibliométrie qualitative avec les autres champs du corpus (auteurs, affiliations, etc.).

D'autre part, l'analyse de 2 corpus permet de mesurer leur proximité thématique avec le cosinus de Salton, on peut récupérer les documents qui constituent cette proximité ou au contraire ceux qui sont spécifiques à l'un ou à l'autre. Nous allons illustrer dans le paragraphe suivant la pratique de telles mesures.

3. Description du corpus et méthodologie

Corpus analysé	Nb de documents	Thématique principale
Tout	1046	La recherche française en sociologie, en 2004
RT-2	55	sociologie de l'immigration/intégration
RT-13	43	sociologie du droit/justice
RT-24	57	Travail (productif et reproductif), rapports sociaux, rapport de genre

Le corpus analysé regroupe 1046 contributions, l'analyse des thématiques se fait sur le titre et le résumé. D'autres champs comme l'affiliation régionale des auteurs, leur genre ou encore leur nom apportent d'autres informations sur la structuration de ce champ de recherche. Nous avons choisi d'analyser 3 réseaux thématiques dont le nombre de documents atteint un seuil statistiquement pertinent et équivalent. Chaque Réseau Thématique (RT) définit un axe spécifique de la recherche en sociologie en France. L'objectif de cette analyse est de décrire la structuration de ces axes en partant de la mesure de l'information définie précédemment. Ce travail d'analyse est fait à l'aide du logiciel Calliope[4] qui intègre la méthode des mots associés et le calcul du pouvoir d'attraction des mots. Une première analyse a déjà été faite sur le corpus entier (de Saint Leger, M.; van Meter, K ; 2005), nous proposons ici de comparer des RT pour repérer les thèmes partagés ou au contraire spécifiques à un seul RT, et partant de là, d'établir des mesures d'indicateurs sur les objets de recherche qui organisent ces activités de recherche, comme par exemple les auteurs ou encore leur genre.

Les documents des corpus sont en texte intégral, Calliope Extract, le premier module de la suite logicielle Calliope, permet l'extraction des termes pertinents contenus dans ces documents en plusieurs étapes :

- La transformation des fichiers sources de documents en un seul fichier structuré, respectant un format xml minimal, pour être traité de façon automatique par les autres modules (Calliope Process et Calliope Viewer)
- L'extraction des termes contenus dans les documents est automatique et fondée sur des traitements statistiques et linguistiques (les segments répétés pour l'extraction, l'usage de dictionnaires de la langue pour l'élimination des mots vides et la lemmatisation). Au terme de cette étape, l'utilisateur dispose d'un lexique des termes extraits. Certains termes sont des synonymes, d'autres sont trop génériques ou vide de sens dans le domaine analysé. Il faut valider ces termes, créer un lexique de synonymie et un antilexique.

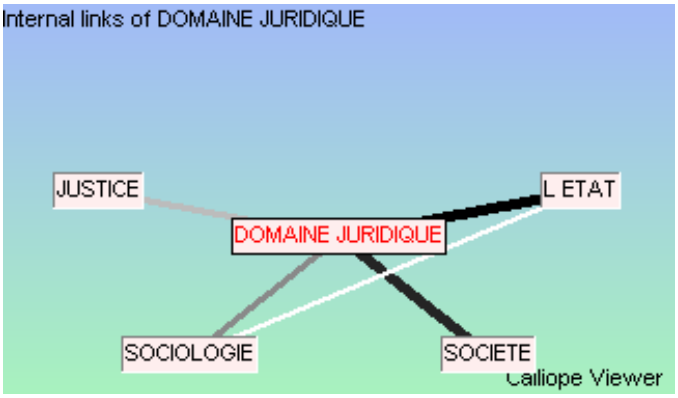
Ainsi, chaque document en texte intégral pointe sur une liste de termes validés représentant son contenu. Calliope Process, 2ème module de la suite logicielle, automatise le découpage thématique du corpus avec la méthode des mots associés, le calcul des *pouvoirs d'attraction* des termes et le lien thématique si plusieurs corpus sont analysés.

4. Résultats et exploitations

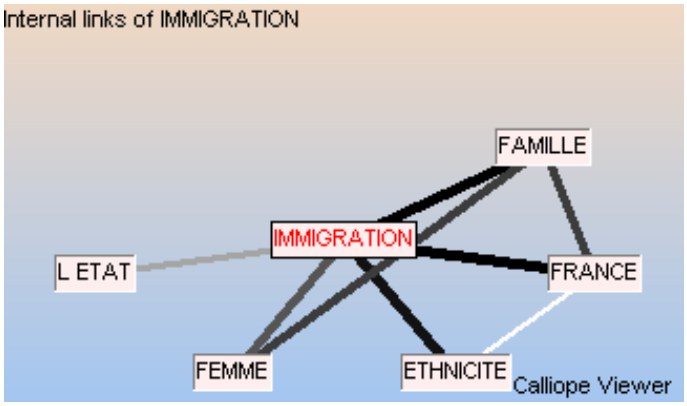
4.1 Thématique principale de chaque RT

[4] Calliope s'appuie sur le prototype DynaTools ayant servi à la "modélisation de la dynamique des flux d'information par le bruit", (de Saint Leger Mathilde, 1997)

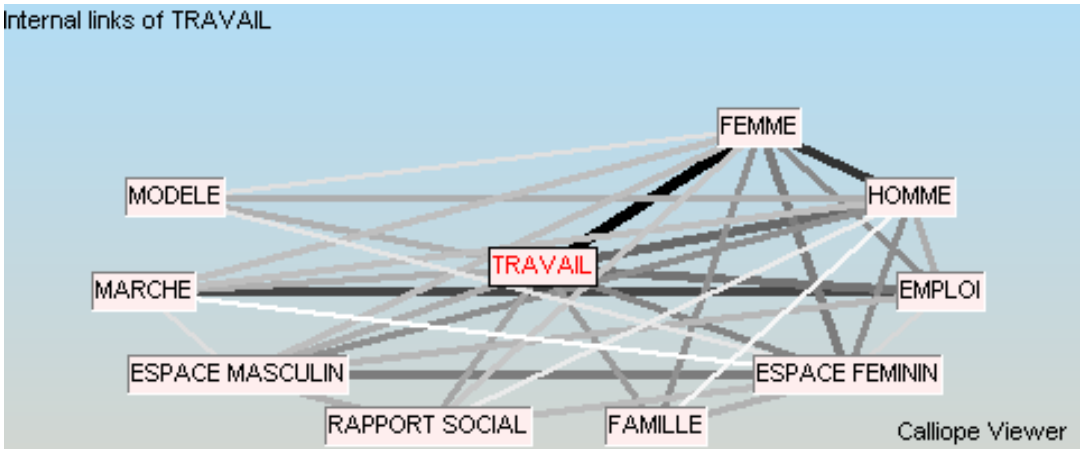
L'objectif ici, est de caractériser chaque RT par sa thématique principale ou *mainstream* en appliquant la méthode des mots associés. Les paramètres de clusterisation sont choisis automatiquement de manière à obtenir un seul cluster : c'est le *mainstream* (voir les réseaux ci-dessous), les mots de ce cluster sont les plus coocurrents dans le corpus. On obtient ainsi une partition du corpus initial : les documents *mainstream* qui ont permis la construction de la thématique principale et les documents *péripheriques* qui n'ont pas participé à cette construction.



-RT 13 : sociologie du droit/justice-



-RT 2 : sociologie de l'immigration/intégration-



-RT 24 : Travail (productif et reproductif), rapports sociaux, rapport de genre-

	nb de doc mainstream	nb de doc périph.
RT-2	50	5
RT-24	55	2
RT-13	33	10

-distribution des documents dans les 3 corpus-

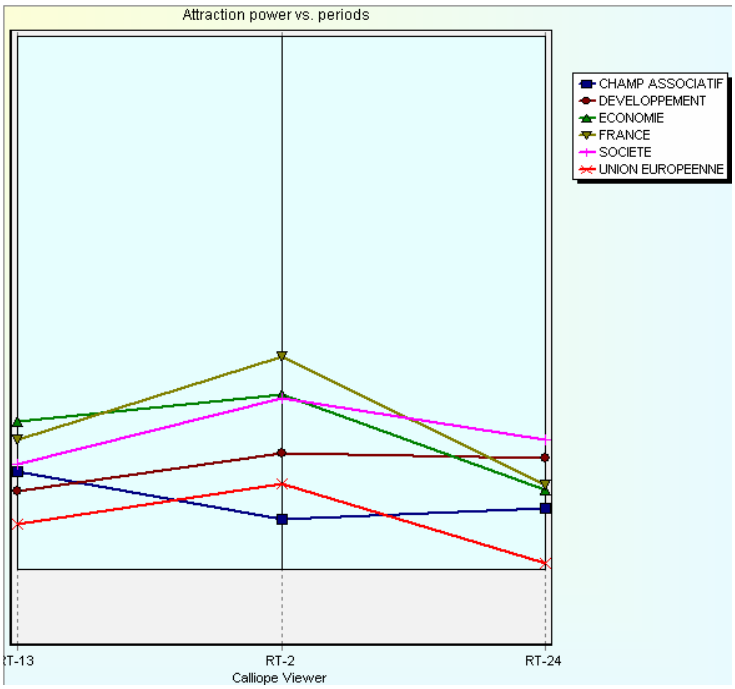
4.2 Informations pertinentes et usages des mots dans leur contexte

Calliope Process calcule le pouvoir d'attraction des termes pour chaque RT suivant les mesures présentées au paragraphe 2 de cet article. Cette partition faite, il est interessant aussi de comparer le contenu thématique de corpus représentant des axes différents d'un même domaine de recherche, typiquement ici, les 3 RT. Une mesure du cosinus de Salton quantifie alors cette mesure.

	Lien
RT13 RT2	0,464
RT13 RT24	0,414
RT2 RT24	0,724

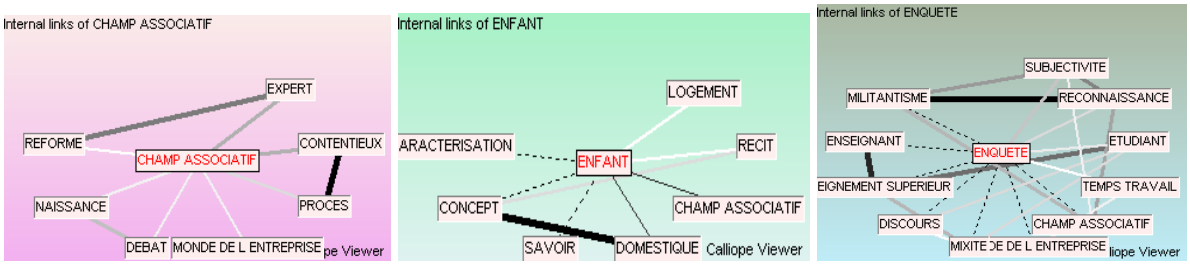
-Mesure de la proximité thématique entre les RT-

Les termes constituant cette proximité thématique sont ceux qui ont un pouvoir d'attraction équivalent dans les 3 RT. A l'inverse, ceux dont le pouvoir d'attraction est fort dans un RT uniquement, est alors spécifique à cet axe de recherche. Nous revenons sur quelques exemples dans ce qui suit.

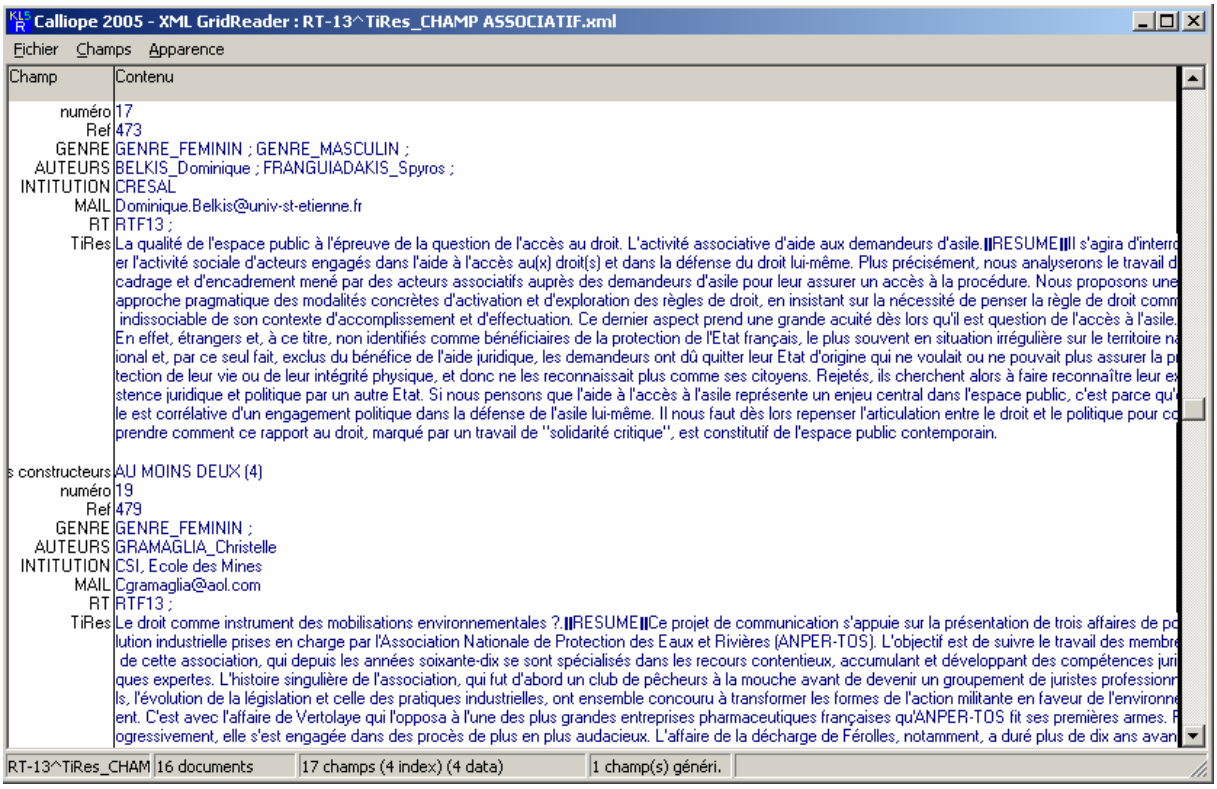


-termes d'importance équivalente dans les 3 RT-

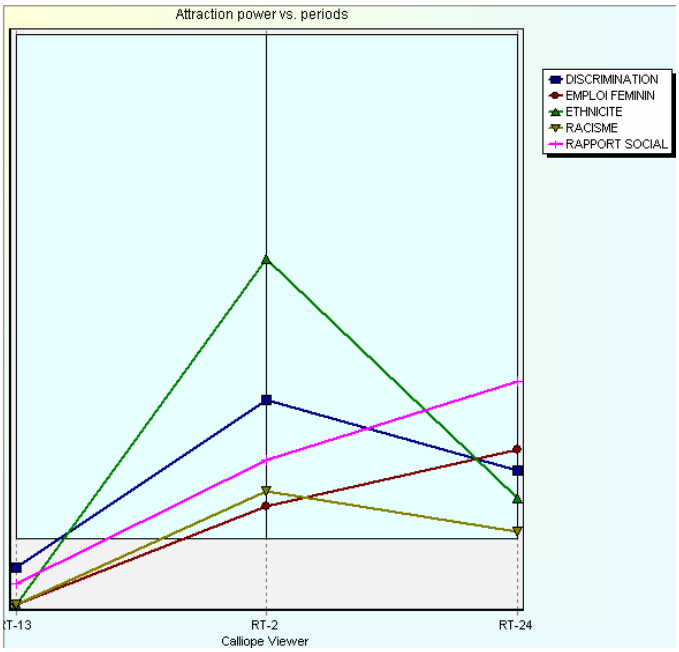
Un retour vers le contexte de ces mots donne un premier éclairage sémantique, comme pour le terme CHAMP ASSOCIATIF pour resp. RT-13, RT-2 et RT-24 ci-dessous.



Au-delà de ce contexte lexical, il est important de revenir vers les documents qui ont permis sa construction pour approfondir notre compréhension du domaine grâce notamment à d'autres champs présents comme indiqués ci-dessous :

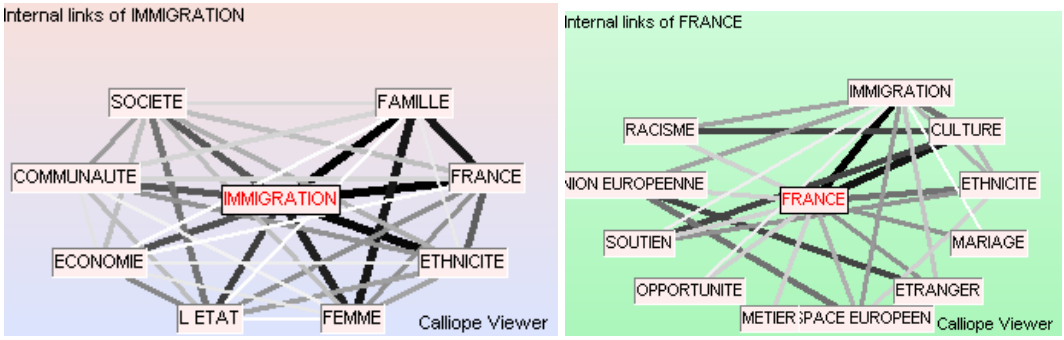


-documents du RT-13 indexés par le contexte lexical CHAMP ASSOCIATIF-

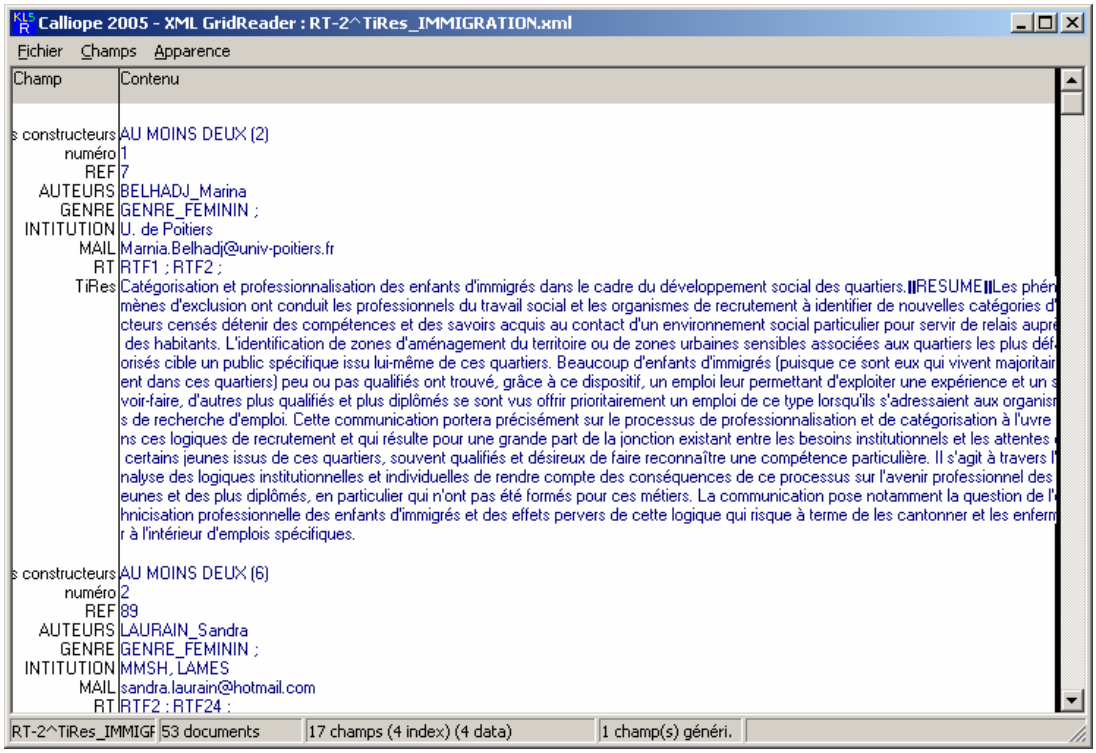


-termes spécifiques à RT-2 et RT-24-

Un retour vers le contexte de ces mots permet de mieux comprendre leur rôle.



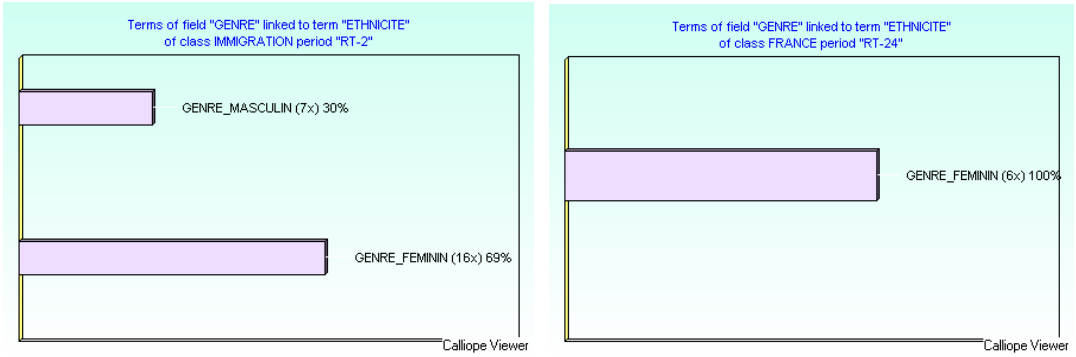
-ETHNICITE et ses différents contextes thématiques-



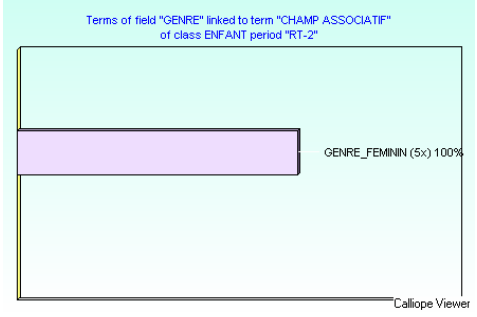
-documents du RT-2 indexés par le contexte lexical autour d'ETHNICITE-

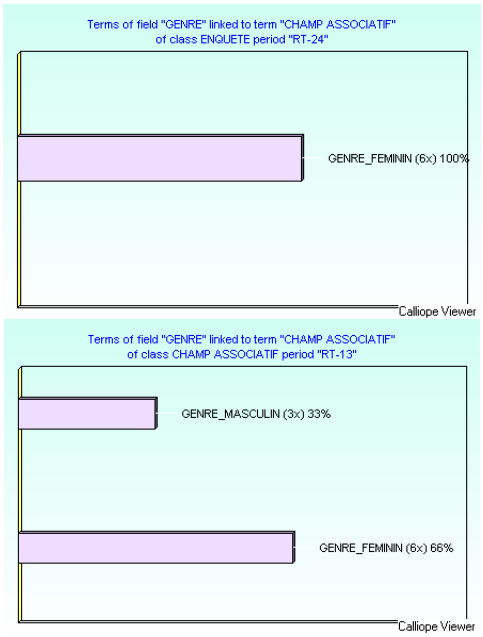
4.3 Structuration de ces axes de recherche

Au-delà des thématiques contenues dans un document, il est important de revenir vers d'autres champs pour comprendre qui fait quoi, où, etc. Nous proposons ici, de revenir vers le genre des auteurs avec la question suivante : certaines thématiques sont-elles plus étudiées par des auteurs femmes ou hommes ? Cette question prend son sens du fait que, dans le corpus regroupant toutes les communications à ce congrès de sociologie, il y a autant d'auteurs femmes qu'hommes.



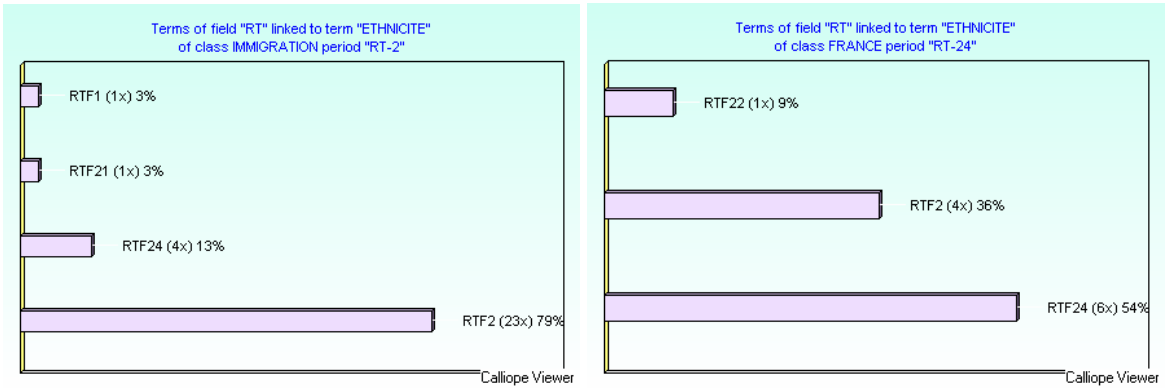
- Répartition du genre des auteurs pour les thématiques autour d'ETHNICITE-





-Répartition du genre des auteurs pour les thématiques autour de CHAMP ASSOCIATIF -

De même, on peut retrouver parmi les documents indexés par la thématique ETNICITE, ceux qui sont communs à d'autres RT :



-Répartition des documents par RT autour du terme ETHNICITE-

Ainsi dans le cadre du RT-2 (migrations et relations interethniques), 29 documents traitent de l'ethnicité dont 4 en commun avec le RT-24 (travail, rapports sociaux, rapports de genre). Alors que seuls 6 documents sont propres au RT-24. De même, RT-2 partage un document commun avec respectivement RT-21 (mouvements sociaux) et RT-1 (savoirs travail, professions). Alors que RT-24 partage un document avec le RT-22 (parcours de vie et dynamique sociale).

5. Conclusion

Nous avons proposé dans ce texte, une mesure de l'information fondée sur le pouvoir qu'elle représente dans la structuration d'un champ de recherche. Nous avons pu ainsi, montrer dans le cadre des publications du congrès de l'Association Française de Sociologie de 2004, comment une telle mesure permet de décrire l'organisation de l'activité collective :

- quelles sont les thématiques principales, avec l'usage des mots dans leur contexte,
- quels sont les différents axes d'activité, et leurs thématiques, partagées ou au contraire spécifiques, en analysant le *pouvoir d'attraction* des termes de corpus les décrivant.

- qui sont les acteurs.

On peut bien entendu établir ces mêmes mesures pour mieux comprendre l'évolution d'une activité de recherche décrite à travers un flux d'information pertinent dans le temps. En effet, dans ce processus d'acquisition de la connaissance, la valeur de l'information n'est pas mesurée seulement sur son pouvoir discriminatoire dans la réponse à une requête, comme dans le cas des moteurs de recherche sur Internet par exemple, mais surtout en fonction du renseignement et de la nouveauté qu'elle apporte quant à l'évolution du domaine étudié. Au-delà de sa valeur sémantique, un terme prend ou perd de l'importance si son rôle, son *pouvoir d'attraction*, dans la structuration d'un champ de recherche augmente ou diminue dans le temps.

6. Bibliographie

Callon M., Courtial J.P., Turner W.A ; La méthode LEXIMAPPE : une méthode pour l'analyse stratégique du développement scientifique et technique ; dans D. Vinck (ed.), Gestion de la Recherche, De Boeck-Wesmael, Bruxelles, 1991.

Callon M. ; La science et ses réseaux. Génèse et circulations des faits scientifiques ; Ed. La Découverte coll. "Anthropologie des sciences et des techniques", Paris, 1989.

De Saint Leger, M. ; Modélisation des flux d'information scientifique et technique pour un suivi de l'évolution des domaines de connaissances ; Thèse de Doctorat d'Université, CERESI/CNRS-CNAM ; 1997.

De Saint Leger, M., van Meter, K ; Cartographie du Premier Congrès de l'AFS avec la Méthode des Mots-Associés ; Paris : Bulletin de Méthodologie Sociologique ; 2005.

Latour B.; La science en action; Ed. La Découvertes/Textes à l'appui; Paris, 1989; ch. 3

Lecoadic Y. « *La science de l'information* » Presses Universitaires de France, Paris, 1994

Lyotard J. F., La condition postmoderne, Les éditions de Minuit, 1979

Pierrat M.-J., de Saint Leger M, Turner W ; *Une Plateforme pour le Travail Lexical Partagé; Congrès SFBA, Ile Rousse 2005*

Raisbeck G.; Théorie de l'information; Ed. Masson; Paris 1964

Recoque A., "Qu'est-ce que l'Intelligence Artificielle" in "Intelligence Artificielle et bon sens", Collection F. R. Bull; Ed. Masson, Paris 1991

Salton A. Wong C. S. Yang ; A Vector Space Model For Automatic Indexing ; Communications Of Acm Vol. 18 N° 11 pp. 613 620 ; Nov. 1985

Salton G. and Mac Gill M.J : Introduction to modern Information retrieval, New-York McGraw Hill Edition ; 1984

Zipf H. P.; Human Behavior And The Principle Of Least Effort; Addison-Wesley Cambridge Massachussets; 1949