île Rousse 2005
Journée sur les systèmes d'information élaborée

CARACTERISATION THEMATIQUE DE COLLECTIONS DE DOCUMENTS TEXTUELS

Abdenour Mokrane, Gérard Dray, Pascal Poncelet
abdenour.mokrane@ema.fr, gerard.dray@ema.fr, pascal.poncelet@ema.fr

Groupe Connaissance et Systèmes Complexes - LGI2P
Site EERIE – EMA - Parc scientifique Georges Besse, 30035 Nîmes cedex 1 - France
Tél. +33 (0)4 66 38 70 94 Fax. +33 (0)4 66 38 70 74

Mots-Clés : Caractérisation thématique, Représentation de termes, Connaissances textuelles, Similarité textuelle, Clustering.

Résumé

Nous proposons dans cet article l'approche *IC-Doc*, permettant une représentation automatique de collections de documents textuels pour une caractérisation thématique du contenu. *IC-Doc* est basée sur une méthode originale de représentation des termes prenant en considération à la fois les notions de co-occurrences contextuelles et de partage de contextes, en vue du calcul de mesures pertinentes de similarités textuelles. Ce papier présente également une expérimentation de l'approche *IC-Doc* sur des collections de documents textuels.

1. Introduction

La fouille de données textuelles vise essentiellement à résoudre les problèmes de surabondance d'informations et faciliter l'extraction des connaissances enfouies dans les documents disponibles sur les bases de données ou sur le *Web*. Les documents textuels sont devenus prédominants sur le *Web* et les informations utiles sont souvent enfouies. De nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent aux traitements de bases de documents textuels [1, 4, 8, 10, 11, 16]. Ces travaux ont donné naissance à des systèmes de catégorisation et de cartographie de documents tels que *Kartoo* [4] ou *Mapstan* [13]. Cependant les modèles d'informations proposés sont peu représentatifs du contenu global par rapport aux différentes thématiques des bases documentaires. Ces modèles s'inspirent des outils de recherches documentaires qui demandent à l'usager de décrire l'information qu'il n'a pas. Il est plus facile aux usagers de découvrir ou de repérer quelque chose qui peut les intéresser, que de produire des descriptions formelles [16]. En outre, ces systèmes sont peu adaptés à une représentation et une caractérisation automatiques du contenu.

Il devient donc indispensable de proposer de nouvelles méthodes et systèmes pour extraire, représenter et caractériser d'une manière automatique les informations contenues dans les bases de documents textuels. Dans ce contexte, nous proposons dans ce papier le système *IC-Doc* de représentation automatique de collections de documents textuels en vue d'une caractérisation thématique du contenu. Cette caractérisation thématique du contenu est basée sur une méthode originale de représentation des termes prenant en considération à la fois les notions de co-occurrences contextuelles et de partage de contextes.

L'article est organisé de la manière suivante. La section 2 présente les étapes générales de représentation de collections de documents et la caractérisation thématique du contenu ainsi que les différents pré-traitements linguistiques et l'analyse statistique des données. La section 3 détaille la méthodologie d'extraction des connaissances textuelles. La section 4 expose quelques résultats de nos expérimentations sur des collections de documents. La section 5 synthétise brièvement les travaux de fouille de données textuelles liés à notre problématique de représentation et caractérisation du contenu. Enfin, la section 6 conclue ce papier et présente les perspectives de recherche associées.

2. Approche IC-Doc

Les différentes étapes de représentation et de caractérisation thématique de collections de documents textuels se résument principalement en deux phases (PI et PII), à savoir :

PI. Pré-traitements linguistiques et analyse statistique des documents

(a) Lemmatisation et étiquetage morpho-syntaxique.

(b) Elimination des mots vides et détection des contextes.

(c) Analyse statistique en vue de l'extraction des Termes Représentatifs (*TR*).

PII. Extraction des connaissances textuelles

(a) Représentation des termes.

(b) Mesures de similarités.

(c) Clustering et caractérisation thématique.

Dans la suite de cette section, nous présentons succinctement la phase I concernant les pré-traitements linguistiques et l'analyse statistique des documents. La phase II de l'approche *IC-Doc*, objet de cet article, sera décrite dans la section 3.

2.1 Pré-traitements linguistiques

La première étape des pré-traitements linguistiques consiste en la lemmatisation et l'étiquetage morphosyntaxique des documents. L'étape suivante concerne l'élimination des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. A l'aide des étiquettes, nous conservons les noms, les verbes et les adjectifs. De manière générale, dans les différentes approches existantes, un contexte peut être une phrase, un paragraphe ou même l'ensemble du document. Etant

donné que les éléments textuels pertinents, dans le cadre de notre modèle, sont généralement proches dans un document, nous considérons qu'un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases de la base documentaire.

2.2 Analyse statistique en vue de l'extraction des TR

2.2.1 Définitions et notations

Co-occurrence contextuelle (*CO*) : Deux termes *A* et *B* appartenant, en même temps au même contexte, forment une co-occurrence appelée *CO* et notée *{CO : A—B}*. L'ensemble des co-occurrences contextuelles d'une base documentaire *BDoc* est notée *COD* de *BDoc*.

Fréquences d'un terme (*FTC* et *FTD*) : La fréquence *FTC* d'un terme *T* dans une base de documents textuels correspond au nombre d'occurrences du terme *T* dans la base. La fréquence *FTD* d'un terme *T* dans une base de documents textuels correspond au nombre de documents contenants *T*. Les fréquences *FTC* et *FTD* d'un terme T_i sont notées respectivement FTC_i et FTD_i.

Fréquences d'une co-occurrence (*FCC* et *FCD*) : La fréquence *FCC* d'une co-occurrence *CO* dans une base de documents textuels correspond au nombre d'occurrences de *CO* dans la base. La fréquence *FCD* d'une co-occurrence *CO* dans un document *D* correspond au nombre d'occurrences de *CO* dans *D*.

Matrice de co-occurrences brute (*MATCO*) : Soit *N* le nombre de termes d'un corpus documentaire et *E* l'ensemble de ces termes. La matrice de co-occurrence brute d'une base documentaire notée *MATCO* de *E* correspond à une matrice de *N* lignes et *N* colonnes. La ligne *i* de la matrice correspond à un terme T_i de la base et la colonne *j* de la matrice correspond à un terme T_j de la base (*i= 1..N, j= 1..N).*

$$Si\ (i \neq j)\quad MATCO\ (i,j) = FCC \text{ de } \{CO : T_i—T_j\}\ \ sinon\quad MATCO\ (i,j)=FTC_i \qquad (1)$$

Matrice de co-occurrences réduite (*RMATCO*) : A partir de la matrice de co-occurrences brute d'une base documentaire, nous pouvons construire une matrice de co-occurrences réduite définie comme suit : soit *E* l'ensemble des termes d'un corpus documentaire et considérant les deux ensembles *E1* et *E2*, $E1 \subset E$, $E2 \subset E$, contenant respectivement *M* et *K* termes. La matrice de co-occurrences réduite notée *RMATCO* de *E1* sur *E2* correspond à une matrice de *M* lignes et *K* colonnes. La ligne *i* de la matrice correspond à un terme T_i de l'ensemble *E1* et la colonne *j* de la matrice correspond à un terme T_j de l'ensemble *E2*. (*i= 1..M, j= 1..K).*

$$Si\ (T_i \neq T_j)\quad RMATCO\ (i,j)=FCC\ de\ \{CO : T_i—T_j\}\quad sinon\quad RMATCO\ (i,j)=FTC_i \qquad (2)$$

L'analyse statistique de la base documentaire consiste à calculer tout d'abord les *FTC*, *FTD*, et *FCC* de l'ensemble des termes *E* de la base documentaire *BDoc*. Elle consiste ensuite à construire la matrice de co-occurrences brute *MATCO* de *E* pour l'extraction de l'ensemble des termes représentatifs.

2.2.2 Termes représentatifs

Il existe plusieurs méthodes pour le choix des termes représentatifs d'une base de documents textuels. Nous sélectionnons l'ensemble de ces termes suivant l'*Algorithme 1*. Plus de détails sur les méthodes existantes et les paramètres de l'*Algorithme 1* sont disponibles dans [10].

Algorithme 1 : *Termes Représentatifs TR*

Input : *E* ensemble des termes d'une base documentaire *BDoc* ; *MATCO de E* ;

Vecteur $Vdist = <(T_1 , FTD_1) \ldots (T_i , FTD_i) \ldots (T_n , FTD_n)>$; $n = |E|$; $i=1..n$

Output : *TR* (ensemble des Termes Représentatifs)

Begin

1. $TR \leftarrow \varnothing$
2. **foreach** $T_i \in E$ **do**

   if $\frac{FTC_i}{FTD_i} > \alpha$ **then** $TR = TR \cup \{T_i\}$
3. **foreach** $\{CO : T_i — T_j\} \in COD$ de *BDoc* **do**

   if *FCC* de $\{CO : Ti—Tj\} > \beta$ **then** $TR = TR \cup \{T_i\} \cup \{T_j\}$

End

α et β seuils de sélection des termes, dans notre modèle α représente la moyenne des FTC des termes pondérées par les FTD et *β représente la moyennes des FCC de l'ensemble COD* [10].

A partir de l'ensemble des termes représentatifs (*TR*) et de la matrice de co-occurrences brute (*MATCO*) de la base documentaire, nous construisons la matrice de co-occurrences réduite *RMATCO* de *TR* sur *TR*, cette matrice est utilisée dans la phase II de l'approche *IC-Doc*, décrite dans la section suivante.

3. Extraction des connaissances textuelles

Dans cette section, nous présentons la méthodologie de représentation de l'ensemble *TR* en se basant sur les relations textuelles entre les termes, dans l'objectif de classer les termes représentatifs par thématiques.

3.1 Représentation des termes

Afin de représenter l'ensemble des termes *TR* pour le calcul des mesures de similarités entre ces différents termes, nous définissons les deux relations suivantes : Soient *A* et *B* deux termes représentatifs, nous notons $(A \wedge B)$ l'ensemble des termes représentatifs appartenant à des contextes d'apparition de *A* et de *B*. Cet ensemble est défini comme suit :

$$(A \wedge B) = \{T \in E / \{CO : A—T\} \wedge \{CO : B—T\}\} \qquad (3)$$

Nous notons $(A \wedge \neg B)$ l'ensemble des termes représentatifs appartenant aux contextes de *A* et non pas aux contextes de *B*. Cet ensemble est défini comme suit :

$$(A \wedge \neg B) = \{T \in E / \{CO : A—T\} \wedge \neg\{CO : B—T\}\} \qquad (4)$$

où $\neg\{CO : B—T\}$ signifie que le couples de termes $<B,T>$ ne forme pas une co-occurrence contextuelle.

Nous représentons les termes de l'ensemble *TR* par deux matrices notées respectivement *MatR1* et *MatR2*. La première matrice (*MatR1*) prend en considération la relation de co-occurrences contextuelles et la deuxième matrice (*MatR2*) prend en considération la notion de partage de contextes entres les termes représentatifs. Les deux matrices *MatR1 et MatR2* sont calculées suivant l'*Algorithme 2* ci-dessous.

Algorithme 2 *: Représentation de l'ensemble TR*

Input : $TR = \{T_1,...,T_m\}$; *RMATCO de TR sur TR* ; $m = |\ TR\ |$;

Output : Matrices *MatR1* et *MatR2*

Begin

for (i =1; $i <= m$; i++) **do**

for (j =1; $j <= m$; j++) **do**

$MatR1(i,j) = (FCC$ de $\{CO : T_i —T_j\}) / FTC_i$;

$A = |\ (T_i \wedge T_j)\ |$;

$B = |\ (T_i \wedge \neg\ T_j)\ |$;

$$MatR2(i,j) = \frac{A}{A+B}\ ;$$

End

3.2 Mesures de similarités et Clustering

Après la représentation de l'ensemble des *TR*, nous calculons, à partir des deux matrices *MatR1 et MatR2*, les mesures de similarités entres les différents termes représentatifs *TR* de la base documentaire. Nous notons *KDMAT*, la matrice de mesures de similarités entres les *TR*, cette matrice est calculée de la manière suivante :

Soit $T_i \in TR$, $T_j \in TR$ *et* $m = |\ TR\ |$; la similarité textuelle entre T_i *et* T_j est donnée par la formule (5).

$$KDMAT(i,j) = \alpha * Dist1(i,j) + (1-\alpha) Dist2(i,j) \quad (5)$$

où *Dist1(i,j)* et *Dist2(i,j)* sont des distances euclidiennes calculés à partir des matrices *MatR1* et *MatR2* suivant les formules (6) et (7) :

$$Dist1(i,j) = \sqrt{\sum_{k=1}^{m} [MatR\ 1(i,k) - MatR\ 1(j,k)]^2} \quad (6)$$

$$Dist2(i,j) = \sqrt{\sum_{k=1}^{m} [MatR\ 2(i,k) - MatR\ 2(j,k)]^2} \quad (7)$$

Les expérimentations ont permis de fixer le paramètre α à 0.3. (Le critère de co-occurrences contextuelles contribue à 30 % à la pertinence des résultats tandis que le critère de partage de contextes contribue à 70% à la pertinence des résultats présentés à la section 4).

Dans le but de classer les termes représentatifs (*TR*) et de regrouper les *TR* les plus proches sémantiquement par thématiques, nous appliquons un algorithme de clustering aux données de la matrice *KDMAT* adapté aux données de cette matrice. Nous avons choisi d'utiliser l'algorithme k-means, simple et robuste [5], qui nous a permis de mettre en œuvre notre approche. Nous appliquons le k-means de la manière suivante : soit *NB* le nombre de thématiques de la base documentaire. Nous appliquons k-means (*NB*) aux données de *KDMAT*. A l'issue de cette étape nous obtenons *NB* clusters, chaque cluster correspond aux termes représentatifs d'une thématique (sous ensemble des *TR*). Nous évaluons les résultats obtenus par les mesures de précisions et de rappels pour chacune des collections de documents. La section qui suit explique la démarche de mise en oeuvre de notre approche et sa pertinence via des expérimentations sur des collections de documents textuels.

4. Expérimentation

Etant donné que nous ne nous intéressons pas dans cet article au traitement automatique du langage naturel (*TALN*), nous avons utilisé pour l'analyse linguistique des documents, l'analyseur de la société

Synapse (http://www.synapse-fr.com) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Par ailleurs, nous avons développé une collection d'outils permettant de mettre en oeuvre l'approche *IC-Doc*.

De manière à valider notre proposition sur des collections de documents, différentes expérimentations ont été réalisées, dans l'objectif de montrer la pertinence et la capacité de notre approche pour une caractérisation thématique indépendamment des poids donnés aux thématiques dans les collections de documents.

4.1 Données

Nous expérimentons notre approche sur des collections de documents composées de trois thématiques qui sont : économie, informatique et cinéma. Les compositions des différentes collections de documents sont illustrées sur le tableau 1.

Documents analysés par étape	**Economie** ***Nb_Doc***	**Informatique** ***Nb_Doc***	**Cinema** ***Nb_Doc***
C1	10	10	10
C2	40	40	40
C3	100	100	100
C4	10	40	100
C5	40	100	10
C6	100	10	40
C7	40	10	100

Tab. 1 – Composition des collections de documents

4.2 Méthode

Après l'extraction des différents termes représentatifs *TR* à partir de chacune des collections de documents, nous appliquons le clustering suivant l'approche *IC-Doc* sur les *TR,* nous évaluons les résultats obtenus par les mesures de *Précision* et de *Rappel* sur les termes représentatifs extraits pour chacune des thématiques dans chaque collection de documents. La précision et le rappel dans le cadre de notre expérimentation sont définis comme suit : soit *S* l'ensemble des *TR* d'une thématique extraits par le système dans une collection de documents ; soit *V* l'ensemble des *TR* de la thématique dans la collection de documents, la précision et le rappel sont calculés comme suit :

$$\textbf{Précision} = |\boldsymbol{S} \cap \boldsymbol{V}| / |\boldsymbol{S}| \qquad \textbf{Rappel} = |\boldsymbol{S} \cap \boldsymbol{V}| / |\boldsymbol{V}|$$

La précision détermine la quantité d'informations extraite appartenant à chacune des thématiques ; le rappel détermine la quantité d'information extraite par rapport aux thématiques.

4.3 Résultats

Les résultats obtenus sont illustrés sur le tableau 2. L'objectif de l'expérimentation sur la collection C1 (10 documents pour chacune des thématiques) est de montrer que les résultats pour une thématique ne sont pas significatifs dans le cas d'une quantité très réduite de documents, en raison de données pauvres sur la thématique dans la collection de documents, ce qui ce traduit par des chutes de précisions ou de rappels.

Dans tous les autres cas la précision dépasse les 75% et le rappel dépasse les 50% pour chacune des thématiques. Comme illustré sur la figure 1 et sur le tableau 1 la précision ou le rappel ne peuvent chuter pour une thématique que dans le cas de thématiques pauvres dans une collection (10 documents) comme dans la collection C6 et C7 pour informatique, C5 pour cinéma ou C4 pour économie.

Résultats	*Nombre des TR*	*Economie*		*Informatique*		*Cinéma*	
		Précision	*Rappel*	*Précision*	*Rappel*	*Précision*	*Rappel*
C1	533	0.996	0.852	1.000	0.223	0.490	0.387
C2	1526	0.915	0.671	0.985	0.683	0.821	0.533
C3	2383	0.943	0.675	0.997	0.616	0.902	0.557
C4	1631	0.660	0.559	0.994	0.664	0.958	0.565
C5	1510	0.868	0.649	0.996	0.611	0.316	0.348
C6	1391	0.992	0.905	0.920	0.080	0.751	0.513
C7	1394	0.982	0.721	0.575	0.381	0.982	0.622

Tab. 2 – Résultats par collections de documents

Nous avons fait varier les poids des thématiques dans les collections de documents pour montrer la pertinence des résultats indépendamment des poids des thématiques dans les collections, les résultats illustrés sur le tableau 1 sont aussi bien pertinents dans le cas de thématiques équilibrés (collection C2 ou C3) ou dans les autres cas (C4, C5, C6 ou C7).

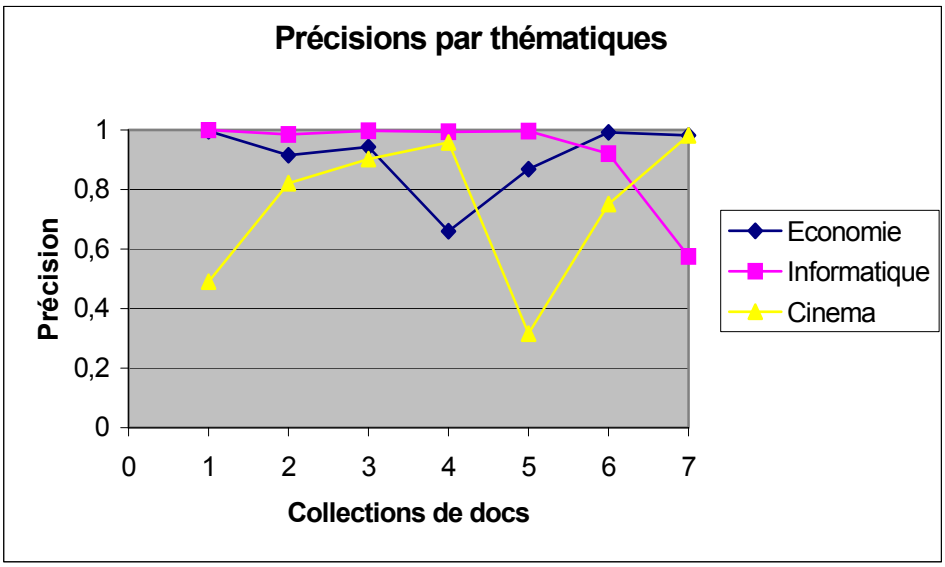

Fig. 1 - Précisions par thématiques dans les colletions de documents

5. Travaux connexes

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'exploitation et l'interprétation systématiques d'informations contenues dans des documents textuels [11]. Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining [7] et du Text Mining, s'intéressent à la fouille de bases de documents textuels [1, 3, 6, 8, 9, 15, 16]. L'objectif de ces travaux est généralement d'analyser le contenu des documents pour en extraire des termes significatifs ainsi que les liaisons qui peuvent exister entre ces différents termes. Dans ce cadre, les modèles de similarités textuelles et la notion de co-occurrences sont les plus utilisées pour l'analyse du contenu [11]. Dans un contexte proche, celui de la recherche documentaire, la recherche de co-occurrences a également été largement étudiée ces dernières années, elle consiste à rechercher les associations de termes les plus fréquentes dans les documents afin de retrouver rapidement les documents pertinents qui peuvent répondre aux requêtes de l'utilisateur. Dans [12] cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. TANAKA et IWASAKI [14] utilisent la matrice de co-occurrences pour la désambiguïsation des termes. Dans [2] un modèle de filtrages syntaxiques de co-occurrences est proposé pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux, ne prennent pas en considération la notion de co-occurrences avec la notion de partage de contextes pour l'extraction des

connaissances textuelles ou le choix des termes représentatifs d'une base de documents textuels. Ce qui implique une pénalisation d'une partie importante des relations textuelles. L'application de ces approches pour la représentation et la caractérisation thématique de collections de documents textuels est donc limitée dans la mesure où elle ne permet pas une extraction pertinente et représentative d'informations sur les différents contenus textuels.

6. Conclusion

Dans ce papier, nous avons présenté l'approche *IC-Doc* de caractérisation thématique de collections de documents textuels. Cette caractérisation est basée sur une méthode originale d'extraction et de classification de connaissances textuelles prenant en considération les co-occurrences contextuelles et le partage de contextes entre les différents termes représentatifs du contenu. Les résultats de nos expérimentations ont montré la pertinence du modèle proposé indépendamment des poids donnés aux thématiques dans les collections de documents. Etant donné que les thématiques se chevauchent les résultats du clustering peuvent être améliorés par des techniques de clustering flou. Ces techniques font l'objet de nos travaux en cours sur la représentation et la caractérisation automatiques et dynamiques de connaissances à partir de bases de données textuelles.

7. Bibliographie

[1] BALDI S. AND DI MEGLIO E. (2004) : « *A Text Mining Strategy based on local contexts of words* ». Proceedings of the JADT'04, Le poids des mots, vol 2, pages 79-87, Louvain-la-Neuve : Presses Universitaires de Louvain.

[2] BESANÇON R. (2002) : « *Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents* ». Conférence TALN, Nancy.

[3] BESANÇON R. (2001) : « *Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes* ». PhD thesis, Ecole polytechnique Fédérale de Lausanne.

[4] CHUNG W., CHEN H. AND NUNAMAKER J. (2003) : « *Business intelligence explorer : A knowledge map framework for discovering business intelligence on the Web* ». Proceedings of the 36 Hawaii International Conference on System Sciences (HICSS'03), Hawaii.

[5] JAIN A.K., MURTY M.N. AND FLYNN P.J. (1999). « *Data Clustering : A Review* ». ACM Computing Surveys, vol 31, Issue 3, pages 264-323.

[6] CHEN H., FAN H., CHAU M. AND ZENG D. (2001) : « *MetaSpider : Meta-searching and categorization on the Web* ». Journal of the American Society for Information Science and Technology, vol. 52, pages 1134 –1147.

[7] KOSALA R. AND BLOCKEEL H. (2000) : « *Web Mining research : A survey* ». *SIGKDD Explorations*, 2(1), pages 1-15.

[8] HE X., DING C., ZHA H. AND SIMON H. (2001) : « Automatic topic identification using Webpage clustering ». *Proceedings of 2001 IEEE International Conference on Data Mining*, Los Alamitos, CA.

[9] HAN. J. AND KAMBER. M. (2000) : «*Data Mining : Concepts and Techniques* », Morgan Kaufmann Publishers, 550 pages.

[10] MOKRANE. A, AREZKI. R, DRAY. G ET PONCELET. P. (2004) : « *Cartographie automatique du contenu d'un corpus de documents textuels* ». JADT'04, Le poids des mots, vol 2, pages 816-823, Louvain-la-Neuve : Presses Universitaires de Louvain.

[11] POIBEAU T. (2003) : « Extraction automatique d'information, du text mining au Web sémantique ». Hermès sciences publications, Paris. 240 pages.

[12] PEREIRA, F., TISHBY, N. AND LEE, L. (1993) : « Distributional clustering of English words ».In *Proceedings of the 31th Meeting of the Association for Computational Linguistics*, pages 183-190

[13] SPINAT E. (2002) : « Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? » *Colloque Cartographie de l'information : De la visualisation à la prise de décision dans la veille et le management de la connaissance*, Paris.

[14] TANAKA K. AND IWASAKI H. (1996) : « Extraction of lexical translations from non-aligned corpora ». In *Proceedings of the 16th International Conference on Computational Linguistics*.

[15] TURENNE N. (2000) : « Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles ». Thèse de doctorat, Université Louis Pasteur, Strasbourg.

[16] OUVRAGE COLLECTIF (2004) : « Méthodes avancées pour les systèmes de recherche d'informations ». Sous la direction de M. IHADJADENE. Hermès sciences publications, Paris. 247 pages.