Cartographie automatique du contenu d'un corpus de documents textuels

A. MOKRANE, R. AREZKI, G. DRAY, P. PONCELET

Groupe connaissance – LGI2P – EMA
Parc scientifique Georges Besse, Site EERIE
30035 Nîmes Cedex 1, France
Tél : +33 (0)4 66 38 70 33 Fax : +33 (0)4 66 38 70 74
Email : {abdenour.mokrane, rachid.arezki, gerard.dray, pascal.poncelet}@ema.fr

Abstract

Considering the quantity of documents available nowadays, the user (organization, company, individual, etc.) is overloaded by information. This one is unable to analysis or apprehend this information in their globality. With the Web, the not structured textual documents became prevalent. Useful information being hidden in the text, it becomes essential to propose new systems allowing the analysis, the organization and the representation of the various textual contents. In this paper we propose a new system, called IC (Information Cartography), for the automatic cartography of the contents of textual documents corpus. IC is based on an original approach for the selection of the representative terms of the content. It helps a community of users working on a thematic in its documentary consultations, by proposing an information chart on the total contents of the corpus or/and for each document.

Keywords : Text mining, Information cartography, Contextual co-occurrence, Representative term.

Résumé

La quantité d'informations et de documents disponibles de nos jours, entraîne une « surinformation » de l'utilisateur final (entreprise, organisme, individu, etc.) qui n'est donc plus capable d'analyser ou d'appréhender ces informations dans leur globalité. Avec le Web, les documents textuels non structurés sont devenus prédominants. L'information utile étant enfouie dans le texte, il devient indispensable de proposer de nouveaux systèmes permettant l'analyse, l'organisation et la représentation des différents contenus textuels. Dans cet article nous proposons un nouveau système, appelé IC (Information Cartography), pour la cartographie automatique du contenu d'un corpus de documents textuels. Le système IC est basée sur une approche originale pour le choix des termes représentatifs du contenu d'un corpus documentaire d'une thématique. Il permet d'aider une communauté d'utilisateurs travaillant sur une thématique donnée dans ses consultations documentaires en lui proposant une carte d'information sur le contenu global du corpus et/ou de chaque document. Nous illustrons IC sur un corpus d'articles de presse.

Mots-clés : Fouille de textes, Cartographie d'information, Co-occurrence contextuelle, Terme représentatif.

1. Introduction

Chaque jour, le nombre de documents disponibles croît d'une manière exponentielle, en particulier en raison de l'essor des communications électroniques. Nous sommes entrés dans l'ère de l'information où l'utilisateur est submergé par la quantité d'information disponible. Avec le Web, les documents textuels non structurés sont devenus prédominants. Les informations utiles étant enfouies dans les textes. Les utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Pour cela, il est nécessaire de mettre en œuvre des systèmes permettant d'analyser les contenus des documents, de les organiser et de les représenter automatiquement.

Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent à la fouille de corpus de documents textuels volumineux (Poibeau, 2003; Besançon 2001; Chen et al, 2001; He et al, 2001) Ces travaux ont donné naissance à des systèmes de catégorisation, voire de cartographie de documents tel que Kartoo (Chung et al, 2002) ou Mapstan (Spinat 2002). Ces outils retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous forme de carte de navigation. Cependant les informations du contenu du corpus documentaire ou de chacun des documents sont peu représentées.

L'objectif de ce papier est de présenter un système, IC (Information Cartography), de cartographie automatique du contenu d'un corpus de documents textuels, basé sur une approche originale pour la sélection des termes représentatifs du contenu. L'article est organisé de la manière suivante. La section 2 résume les travaux antérieurs liés à notre problématique. La section 3 décrit l'architecture fonctionnelle du système IC. Dans la section 4 nous détaillons les différentes étapes de cartographie de contenus textuels. La section 5 illustre la mise en œuvre de notre approche sur un corpus d'articles de presse. Enfin, la section 6 conclue ce papier et présente les perspectives de recherche associées.

2. Travaux antérieurs

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'interprétation, l'exploitation et la visualisation systématiques d'informations contenues dans des documents textuels. (Poibeau, 2003). Actuellement, de nombreux travaux de recherche, notamment issus du Web Mining (Kosala et Blockeel, 2000) et du Text Mining, s'intéressent à la fouille de corpus documentaires volumineux (Andrieu, 2000; Besançon, 2001; Chen et al, 2001; Han et Kamber, 2000; He et al, 2001; Turenne, 2000). 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 (Poibeau, 2003). 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 (Pereira et al 1993) cette co-occurrence est utilisée pour la classification des termes selon la distribution de leurs contextes syntaxiques. (Tanaka et al 1996) utilise la matrice de co-occurrences pour la désambiguïsation des termes. (Besançon et al 2002) ont proposé un modèle de filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents et la recherche documentaire. Tous ces travaux de fouille de textes, ne prennent pas en considération les co-occurrences dans le choix des termes de l'espace vectoriel ou les termes représentatifs du contenu d'un corpus documentaire. Ce qui implique une pénalisation d'une partie importante des relations de co-occurrences. Cette

pénalisation est due au choix de termes basé principalement sur leurs fréquences d'occurrences (Salton et al 1975, Salton et Buckley 1988). L'application de ces approches pour la cartographie du contenu des documents est donc limitée dans la mesure où elle ne permet pas d'extraire des informations pertinentes sur le contenu des documents. En effet, ces méthodes ne contiennent pas de stratégie, prenant en considération les associations des termes et leurs occurrences, pour sélectionner les termes représentatifs d'un corpus documentaire d'une thématique donnée.

3. Architecture fonctionnelle du système IC

Le système IC (Information Cartography) se charge de la cartographie du contenu d'un corpus de documents textuels. Il est composé de trois sous-systèmes dédiés aux trois étapes de notre méthodologie de cartographie du contenu : analyse linguistique et prétraitements, analyse statistique des données textuelles, modélisation et réalisation des cartes d'information. La figure 1 illustre l'architecture fonctionnelle du système IC.

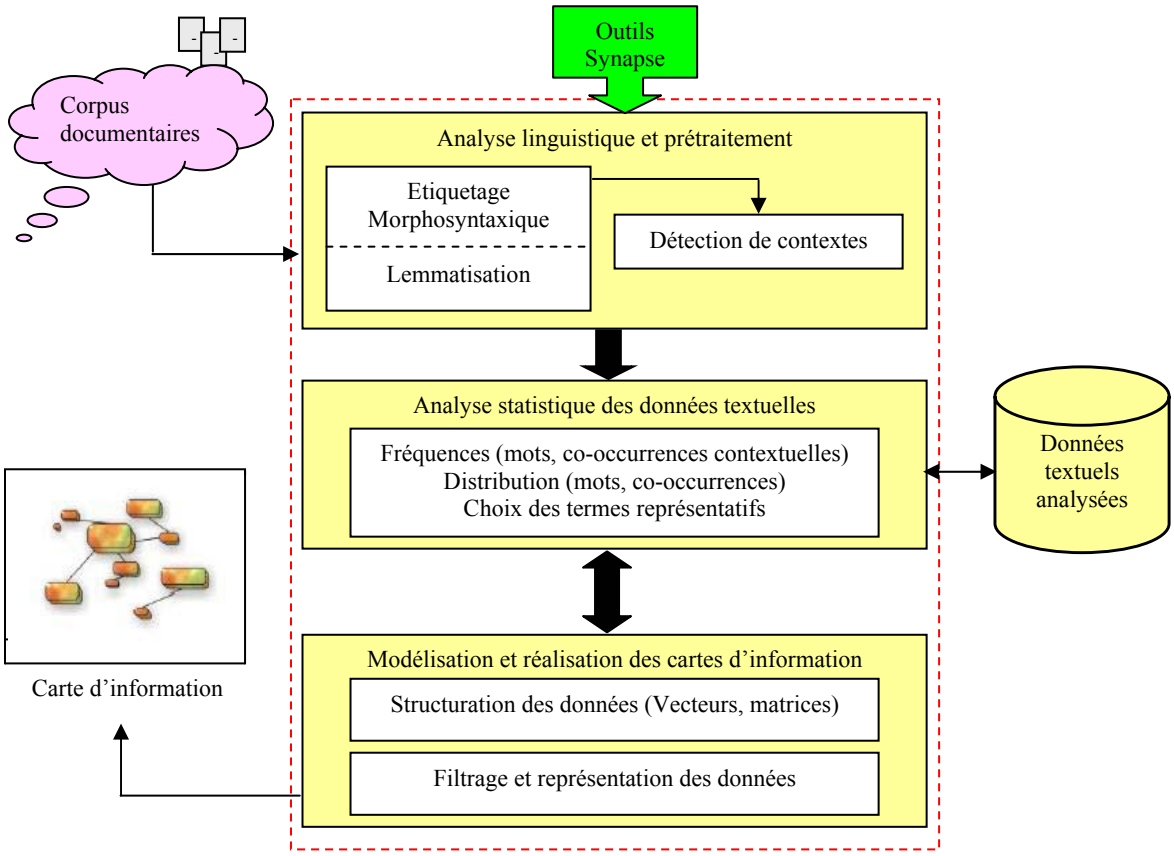

Figure 1. *Architecture fonctionnelle du système IC*

4. Fonctionnement du système IC

Etant donné que nous ne nous intéressons pas ici au traitement automatique du langage naturel (TALN), nous utilisons 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. Via ces deux outils, le corpus de documents textuels est transformé en documents étiquetés et lemmatisés. Le prétraitement concerne le nettoyage du corpus des mots vides (articles, pronoms, prépositions, etc.) et la détection des différents contextes. A l'aide des étiquettes, nous conservons principalement les noms, les verbes et les adjectifs.

4.1. Analyse statistique des données textuelles

Comme nous l'avons spécifié dans la section travaux antérieurs, la notion de co-occurrence, prend des dimensions différentes suivant la tâche à accomplir. Pour la recherche d'information, on définit généralement des co-occurrences entre des mots et des documents contenant ces mots. Pour la détection des contextes de termes, la co-occurrence est entendue entre le terme et une fenêtre de mots. Enfin une autre approche consiste en l'acquisition de co-occurrences entre les différentes unités d'un corpus de référence. Ce type d'approche a été adopté par exemple dans le cadre du modèle DSIR (Distributional Semantic for Information Retrieval) (Besançon 2002). Le modèle DSIR intègre la notion de co-occurrences entre termes pour une nouvelle représentation vectorielle de documents. Cependant, le modèle proposé, ne prenant pas en considération les co-occurrences dans le choix des termes de l'espace vectoriel, implique une pénalisation d'une partie importante des relations de co-occurrences. Cette pénalisation est due principalement au choix des mots clés basé sur leurs fréquences d'occurrences. Pour répondre à cette problématique, nous proposons dans la suite de cette section, une nouvelle stratégie pour établir le choix des termes représentatifs du contenu global ou de chacun des documents. Cependant, dans un premier temps, nous adaptons les définitions de co-occurrence à notre contexte.

4.1.1 Définitions

De manière générale, dans les différentes approches existantes, un contexte de co-occurrence peut être une phrase, un paragraphe ou même l'ensemble du document. Etant donné que nous considérons que les éléments pertinents sont généralement proches dans un document, nous considérons dans notre modèle qu'un contexte correspond à une phrase et ainsi la détection des contextes va correspondre à l'annotation des différentes phrases du corpus documentaire. Dans ce cadre, nous pouvons définir la notion de co-occurrence contextuelle de la manière suivante :

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}*.

Pour chaque terme, il est possible de connaître d'une part sa fréquence d'apparition dans le corpus mais également le nombre de documents qui contiennent ce terme.

Fréquences d'un terme (FTC et FTD) : La fréquence *FTC* d'un terme *T* dans un corpus de documents textuels correspond au nombre d'occurrences du terme *T* dans le corpus. La fréquence *FTD* d'un terme *T* dans un corpus de documents textuels correspond au nombre de documents contenants *T*.

La définition précédente est bien entendue spécifique à la notion de terme et les définitions suivantes sont adaptées à la notion de co-occurrence.

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

A partir des différentes fréquences obtenues, il est possible de créer les matrices suivantes :

Matrice de co-occurrences brute (MATCO) : Soit *N* le nombre de termes d'un corpus documentaire. La matrice de co-occurrence brute notée *MATCO* correspond à une matrice de *N* ligne et *N* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* du corpus et la colonne *j* de la matrice correspond à un terme *Tj* du corpus (*i= 1..N, j= 1..N)*.

Si (i ≠ j) MATCO (i,j)=FCC de *{CO : Ti—Tj}* Sinon *MATCO (i,j)=FTC de Ti*

La première analyse statistique consiste à calculer les *FTC, FTD, FCC et MATCO* (définies ci dessus) dans l'objectif de choisir les termes représentatifs. A partir de la matrice *MATCO* et des termes représentatifs du corpus documentaire, nous construisons la matrice de co-occurrences réduite définie de la manière suivante :

Matrice de co-occurrences réduite (RMATCO) : Soit *E* l'ensemble des termes d'un corpus documentaire et considérant l'ensemble $R \subset E$, contenant *M* termes.

La matrice de co-occurrences réduite notée *RMATCO* correspond à une matrice de *M* ligne et *M* colonnes. La ligne *i* de la matrice correspond à un terme *Ti* de l'ensemble *R* et la colonne *j* de la matrice correspond à un terme *Tj* de l'ensemble *R*. (*i= 1..M, j= 1..M).*

Si (i ≠ j) RMATCO (i,j)=FCC de {CO : Ti—Tj} sinon RMATCO (i,j)=FTC de Ti

Nous expliquons dans la section suivante comment nous sélectionnons les termes représentatifs.

4.1.2. Choix des termes représentatifs

Le choix des termes représentatifs du contenu du corpus prend en considération les co-occurrences contextuelles les plus fréquentes en plus des fréquences de termes, en tenant compte de la distribution *FTD* de chacun des termes. L'ensemble *E* des termes représentatifs du corpus est constitué par l'union des deux ensembles *E1* et *E2* : $E = E1 \cup E2$.

E1 est l'ensemble des termes *Ti* dont la fréquence d'occurrence dans le corpus *(FTCi)* pondérée par le nombre de documents contenant *Ti* (*FTD)* est supérieure à la moyenne. Ce qui se traduit par ceci :

$$\left(FTC_i / FTD \right) > MoyFTC$$

où *MoyFTC* est calculée de la manière suivante : soit *T* l'ensemble des termes du corpus documentaire et *M* le nombre de ces termes. Nous calculons la moyenne pondérée *MoyFTC* des *FTCi* de chacun de ces termes avec prise en considération des *FTD* (nombre de documents contenant chaque terme).

$$MoyFTC = \frac{\sum_{i=1}^{i=M} \left(FTC_i / FTD \right)}{M}$$

E2 est l'ensemble des couples de termes *Ti, Tj* du corpus qui forment une co-occurrence contextuelle *CO* dont la fréquence *(FCC)* est supérieure à la moyenne. Ce qui se traduit par ceci :

$$FCC \text{ de } \{CO : Ti—Tj\} > MoyFCO$$

Où *MoyFCO* est calculée par : soit *N* le nombre de co-occurrences *CO* du corpus documentaire. Nous calculons la moyenne pondérée *MoyFCO* des *FCCi* de chacune des co-occurrences du corpus.

$$MoyFCO = \frac{\sum_{i=1}^{i=N} FCC_i}{N}$$

Après avoir construit l'ensemble *E* des termes représentatifs (union des deux ensembles *E1 et E2*), nous pouvons procéder à la construction de la matrice *RMATCO* (Cf. définitions) sur l'ensemble *E*. Cette réduction de la matrice *MATCO* a pour objectif d'éviter à l'utilisateur le problème de surcharge d'informations.

4.2. Modélisation et réalisation des cartes d'information

La carte d'information représentant le contenu général du corpus documentaire n'est qu'une représentation sous forme d'un graphe de la matrice *RMATCO*. Les sommets du graphe modélisent les termes représentatifs et les liens représentent les co-occurrences contextuelles entre ces termes avec leurs degrés d'importance dans le corpus documentaire. Les sommets du graphe sont représentés avec des tailles différentes correspondant aux fréquences des termes dans le corpus. Les différents liens entre les termes sont représentés par des arrêtes de largeurs et de couleurs différentes. Les cartes associées à chaque document du corpus sont construites avec le même modèle appliqué au corpus documentaire.

5. Illustration

Dans le cadre du projet CARICOU (Capitalisation de Recherches d'Informations de Communautés d'Utilisateurs) mené par le LGI2P, nous avons développé l'outil *CO2MAT (COntextual CO-occurrence MATrix)* permettant de construire la matrice *RMATCO qui* stocke les informations reliées à la carte représentant le contenu global du corpus documentaire. Afin de montrer l'intérêt du système IC, nous avons appliqué notre approche via l'outil CO2MAT sur un corpus documentaire d'environ 1000 documents de presse. Le corpus documentaire porte sur la thématique « crise irakienne ». Ces documents sont issus de journaux internationaux avant le déclenchement de la deuxième guerre en Irak.

La première étape du choix des termes représentatifs consiste à générer l'ensemble des termes en ne prenant en considération que les fréquences *FTC* des termes dans le corpus général et le nombre de documents *FTD* contenant chaque terme. L'ensemble *SE1* des termes générés par la première étape et non détecté par la seconde ainsi que les termes les plus fréquents de cette étape est le suivant :

SE1 = {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, **France, ONU**}

La deuxième étape consiste à extraire des termes représentatifs en prenant en considération les fréquences de co-occurrences *FCC*. Les co-occurrences les plus fréquentes de l'étape 2 ainsi que celles qui ont participé à l'extraction de nouveaux termes non détectés à l'étape 1 sont décrites dans le *Tableau 1* et sont triées par ordre décroissant.

1	CO : Irak – Guerre	**6**	CO : Bush – Faire	**11**	CO : Président – Bush
2	CO : Bush – Guerre	**7**	CO : Faire – Guerre	**12**	CO : Président – Saddam
3	CO : Irak – Etats Unis	**8**	CO : Aller – Guerre	**13**	CO : Irak – Président
4	CO : Irak – Américain	**9**	CO : Bush – Irak	**14**	CO : Irak – Irakien
5	CO : Bush – Aller	**10**	CO : Faire – Irak	**15**	CO : Irak – Inspection

Tableau 1. *Les co-occurrences les plus fréquentes*

L'ensemble des termes représentatifs correspondant au tableau 1 est le suivant :

SE2 = {**Etats-Unis**, Irak, Irakien, Guerre, Américain, Bush, Aller, Faire, Président, Saddam, **Inspection**}

L'ensemble *SE* des termes représentatifs résultant de l'union de *SE1* et *SE2* est donc :

SE= {Guerre, Irak, Américain, Bush, Faire, Aller, Saddam, Président, Irakien, ***France, ONU, Etats-Unis, Inspection***}.

SE1—SE2 = { **France, ONU** }, les termes appartenant à *SE1* et non à *SE2*.
SE2—SE1 = { **Etats-Unis, Inspection** }, les termes appartenant à *SE2* et non à *SE1*.

Nous remarquons que la simple utilisation des fréquences des termes nous prive d'informations importantes sur les Etats-Unis et les Inspections (*SE2—SE1*). La prise en considération des co-occurrences seule mène au même problème, ce qui se traduit par la privation des informations sur l'ONU et la France (*SE1—SE2)*. Il est clair que les deux étapes de sélection des termes représentatifs sont complémentaires. La figure 2 illustre le choix et la carte d'information associée à ces termes représentatifs du contenu, via notre approche.

Dans notre modèle, malgré que le terme Etats-Unis ait une fréquence faible dans le corpus, il a été sélectionné comme terme représentatif. Cela grâce à la co-occurrence contextuelle importante entre Irak et Etats-Unis (CO : Irak — Etats-Unis). De même pour le terme Inspection qui a été sélectionné par rapport à la co-occurrence entre Irak et Inspection (CO : Irak — Inspection}. Ces nouveaux termes vont permettre de représenter des nouvelles co-occurrences très utiles pour l'interprétation et la représentation du contenu du corpus documentaire (ces co-occurrences ne figurent pas dans le tableau 1). Par exemple les co-occurrence entre la France et les Etats-Unis ou bien Inspection et ONU. Ces nouvelles co-occurrences sont faibles dans le corpus mais apportent des informations importantes sur le contenu du corpus documentaire. Notamment en ce qui concerne les inspections en Irak ou bien la relation entre la France et les Etats-Unis, etc. Ces nouvelles co-occurrences apportent de nouvelles informations.

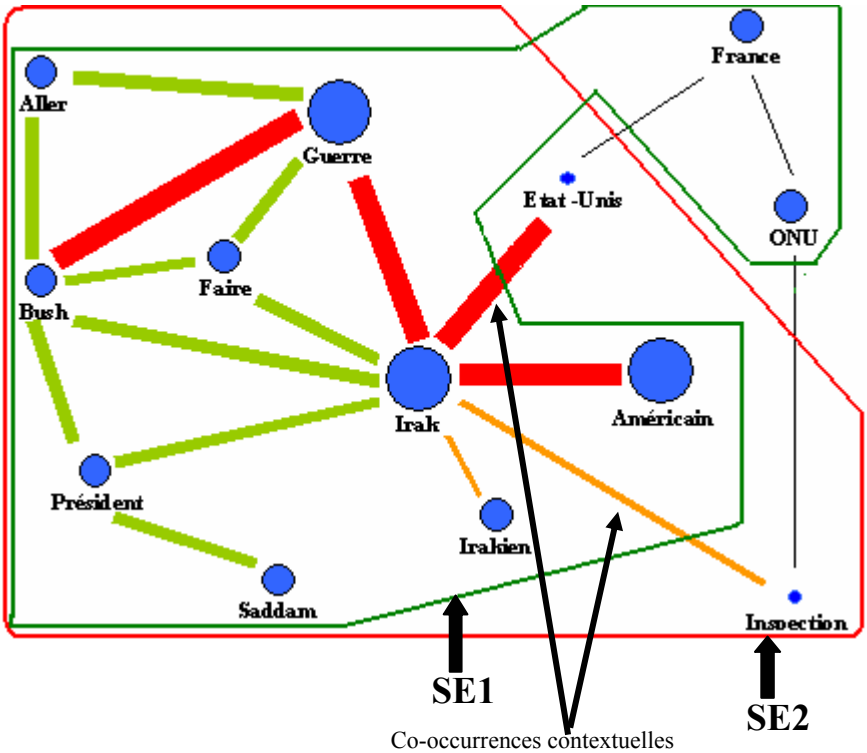

Figure 2. *Choix et représentation des termes de l'ensemble SE*

Une première consultation de cette carte, donne une idée général sur le contenu du corpus. On peut déduire rapidement que le corpus documentaire est concentré sur la guerre en Irak. On peut déduire aussi que la presse écrite considère le président Bush comme acteur principal de cette guerre. Il existe des informations sur la relation Etats-Unis / France dans ce corpus documentaire. On en parle beaucoup moins de l'Organisation des Nations Unies (ONU). Il y a aussi des informations sur les inspections en Irak, etc.

6. Conclusion

Dans cet article, nous avons proposé une nouvelle approche pour la cartographie automatique de contenus textuels. Nous avons concrétisé cette approche par le système IC, basé sur notre approche originale de sélection de termes représentatifs. Cette approche permet d'aider une communauté d'utilisateurs, travaillant sur une thématique donnée, dans ses consultations documentaires. Elle peut également être utilisée, dans le cadre d'un système de recherche documentaire, pour le choix des mots clés et la réduction de l'espace vectoriel. Nos travaux en cours portent sur l'exploitation efficace de cette approche par le développement d'un système de navigation par le contenu, à l'aide des cartes d'information, dans des corpus documentaires ainsi que l'application d'algorithmes de clustering flou pour la visualisation de l'ensemble des termes représentatifs de corpus documentaires volumineux et divers.

Références

Andrieu O. (2000). Créer du trafic sur son site Web. Edition Eyrolles.

Besançon R. (2002). Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents. *Conférence TALN*, Nancy.

Besançon R. (2001). Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes. PhD thesis, Ecole polytechnique Fédérale de Lausanne.

Chung W., Chen H. and Nunamaker J. (2002). 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.

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.

Kosala R. and Blockeel H. (2000). Web Mining research : A survey. *SIGKDD Explorations*, 2(1), pages 1-15.

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.

Han. J. and Kamber. M.(2000). Data Mining : Concepts and Techniques », Morgan Kaufmann Publishers, 550 pages.

Poibeau T. (2003). Extraction automatique d'information, du text mining au Web sémantique. Edition Lavoisier.

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.

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.

Salton G. and Buckley C. (1988). Term weighting approaches in automatic text retrieval. Information Processing and Management, 24(5) pages 513-523.

Salton G., Yang C.S. and YU C.T. (1975). A theory of term importance in automatic text analysis, *Journal of the American Society of Information Science*.

Tanaka K.and Iwasaki H. (1996). Extraction of lexical translations from non-aligned corpora. In *Proceedings of the 16th International Conference on Computational Linguistics*.

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.