

Fouille de collections de documents en vue d'une caractérisation thématique de connaissances textuelles

Abdenour Mokrane, Gérard Dray, Pascal Poncelet

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
{abdenour.mokrane, gerard.dray, pascal.poncelet}@ema.fr

Résumé. De nos jours, les entreprises, organismes ou individus se trouvent submergés par la quantité d'information et de documents disponibles. Les utilisateurs ne sont plus capables d'analyser ou d'appréhender ces informations dans leur globalité. Dans ce contexte, il devient indispensable de proposer de nouvelles méthodes pour extraire et caractériser de manière automatique les informations contenues dans les bases documentaires. Nous proposons dans cet article l'approche *IC-Doc* de caractérisation automatique et thématique du contenu de collections de documents textuels. *IC-Doc* 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. *IC-Doc* permet ainsi une extraction automatique de *KDMs* (*Knowledge Dynamic Maps*) sur les contenus des bases documentaires. Ces *KDMs* permettent de guider et d'aider les utilisateurs dans leurs tâches de consultations documentaires. Ce papier présente également une expérimentation de notre approche sur des collections de documents textuels.

Mots-Clefs. Caractérisation thématique, Similarité textuelle, Partage de contextes, Knowledge Dynamic Map.

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. Chaque jour, en particulier en raison de l'essor des communications électroniques, le nombre de documents disponibles croît de manière exponentielle et l'utilisateur (entreprise, organisme ou individu) se trouve submergé par la quantité d'informations disponibles. Ces utilisateurs ne sont donc plus capables d'analyser ou d'appréhender ces informations dans leur globalité.

De nombreux travaux de recherche, notamment issus du Web Mining et du Text Mining, s'intéressent aux traitements de bases de documents textuels (Baldi et Di meglia 2004,

Chung et al. 2003, Hongyuan et al. 2001, Mokrane et al. 2004b, Poibeau 2003, Ihadjadene 2004). Ces travaux ont donné naissance à des systèmes de catégorisation et de cartographie de documents comme *Kartoo* (Chung et al. 2003) ou *Mapstan* (Spinat 2002). Ces outils retrouvent des liens entre les différents documents ou sites Web et représentent ces liens sous forme de cartes de navigation. Cependant les modèles d'informations proposés sont peu représentatifs du contenu global ou de chacun des documents par rapport aux différentes thématiques des bases documentaires. Ces modèles s'inspirent des outils de recherches documentaires qui demandent à l'utilisateur de décrire l'information qu'il n'a pas (Ihadjadene 2004). En outre, ces systèmes sont peu adaptés à une caractérisation thématique en vue d'une navigation par le contenu dans une collection de documents. Il devient donc indispensable de proposer de nouvelles méthodes et systèmes pour extraire et caractériser de manière automatique les informations contenues dans les bases de documents textuels.

Dans cet article, nous proposons l'approche *IC-Doc* de caractérisation automatique et thématique du contenu de collections de documents textuels. *IC-Doc* 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. *IC-Doc* permet ainsi une extraction automatique de *KDMs* (*Knowledge Dynamic Maps*) à partir du contenu d'une base documentaire. Ces *KDMs* permettent de guider les utilisateurs dans leurs tâches de consultations documentaires.

L'article est organisé de la manière suivante. La section 2 présente les étapes générales de caractérisation de collections de documents, les différents pré-traitements linguistiques ainsi que l'analyse statistique pour l'extraction des termes représentatifs. La section 3 détaille la méthodologie d'extraction de connaissances textuelles en vue de la construction de *KDMs*. La section 4 expose les 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. Enfin, la section 6 résume notre approche et présente les perspectives de recherche associées.

2. Approche IC-Doc

L'approche *IC-Doc* proposée dans cet article fait suite à nos travaux sur la fouille de données textuelles (Mokrane et al. 2004a, Mokrane et al. 2004b). Cette approche permet une caractérisation automatique et thématique du contenu de collections de documents textuels. Les différentes étapes de cette approche sont les suivantes :

I. 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*).

II. 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. Dans le cadre de notre modèle un contexte correspond à une phrase. De manière générale, dans les différentes approches existantes, un contexte peut être une phrase, un paragraphe ou même l'ensemble du document. Dans le cadre de notre modèle, 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 termes représentatifs

Avant de préciser comment extraire les termes représentatifs, nous donnons quelques définitions et notations utilisées par la suite.

2.2.1 Définitions et notations

Soit $\{T_1, T_2, \dots, T_n\}$ l'ensemble des termes de la base de données textuelles obtenus à l'étape des pré-traitements linguistiques.

Co-occurrence contextuelle (CO) : Deux termes T_i et T_j appartenant, en même temps au même contexte, forment une co-occurrence appelée CO et notée $\{CO : T_i - T_j\}$. 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 contenant 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 ($i=1..N, j=1..N$).

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

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 E_1 et E_2 avec $E_1 \subset E$ et $E_2 \subset E$, 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 $E2$. ($i= 1..M, j= 1..K$).

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

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 Extraction des termes représentatifs

Dans le cadre de notre modèle, nous sélectionnons l'ensemble des termes représentatifs à l'aide de l'Algorithme 1. Plus de détails sont disponibles dans (Mokrane et al. 2004a, Mokrane et al. 2004b).

Algorithme 1 : Termes Représentatifs TR

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

Vecteur $Vdist = \langle (T_1, FTD_1) \dots (T_i, FTD_i) \dots (T_n, FTD_n) \rangle$; $n= |E|$; $i=1..n$

Output: TR (ensemble des Termes Représentatifs)

Begin

1. $TR \leftarrow \emptyset$;

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 $\{CO : T_i-T_j\} > \beta$ **then** $TR = TR \cup \{T_i\} \cup \{T_j\}$;

End

Les paramètres α et β sont les seuils de sélection des termes (Mokrane et al. 2004b).

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. L'objectif visé est de classer les

termes représentatifs par thématiques et de construire des *KDMs*. Avant de détailler les différentes étapes de notre méthode, nous définissons la notion de *KDM*.

Une *KDM* (*Knowledge Dynamic Map*) est un graphe $G = \langle X, U \rangle$ où X est un ensemble de N sommets modélisant N termes représentatifs et U un ensemble d'arêtes représentant les relations textuelles. Les sommets du graphe sont des hyperliens (liens dynamiques) à deux fonctionnalités. La première fonctionnalité permet d'organiser d'une manière automatique le graphe autour d'un thème central. La seconde permet d'atteindre une nouvelle *KDM*. La dimension d'une *KDM* correspond au nombre de ses termes représentatifs.

3.1 Représentation des termes

Pour représenter l'ensemble des termes TR qui seront utilisés lors du calcul des mesures de similarités, nous définissons les deux relations suivantes : Soient T_i et T_j deux termes représentatifs, nous notons $(T_i \wedge T_j)$ l'ensemble des termes représentatifs appartenant à des contextes d'apparition de T_i et de T_j . Cet ensemble est défini comme suit :

$$(T_i \wedge T_j) = \{T_k \in E / \{CO : T_i - T_k\} \wedge \{CO : T_j - T_k\}\} \quad (3)$$

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

$$(T_i \wedge \neg T_j) = \{T_k \in E / \{CO : T_i - T_k\} \wedge \neg\{CO : T_j - T_k\}\} \quad (4)$$

où $\neg\{CO : T_j - T_k\}$ signifie que le couples de termes $\langle T_j, T_k \rangle$ 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 entre les termes représentatifs. Les deux matrices $MatR1$ et $MatR2$ sont calculées suivant l'*Algorithme 2*.

Algorithme 2 : Représentation de l'ensemble TR

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

Output: Matrices $MatR1$ et $MatR2$

Begin

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

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

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

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

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

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

End

3.2 Mesures de similarités, Clustering et KDMs

Après la représentation de l'ensemble des TR , nous calculons, à partir des deux matrices $MatR1$ et $MatR2$, les mesures de similarités entre les différents termes représentatifs TR de la base documentaire. Nous notons $KDMAT$, la matrice de mesures de similarités entre 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) :

$$Dist1(i, j) = \sqrt{\sum_{k=1}^m [MatR1(i, k) - MatR1(j, k)]^2} \quad (6)$$

$$Dist2(i, j) = \sqrt{\sum_{k=1}^m [MatR2(i, k) - MatR2(j, k)]^2} \quad (7)$$

Les expérimentations ont permis de fixer le paramètre α à 0.3. (i.e. 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.

Dans le but de classer les termes représentatifs (TR) par thématiques et de construire des $KDMs$ cohérentes, 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 (Jain et al. 1999), qui nous a permis de mettre en œuvre notre approche.

Nous appliquons cette méthode de la manière suivante : soit NB le nombre de thématiques de la base documentaire et DK la dimension d'une KDM définie a priori de façon à ne pas surcharger l'utilisateur. Dans une première étape, 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). De la même manière, l'étape suivante du processus de clustering consiste en l'application du k-means(DK) aux sous ensembles des termes représentatifs obtenus à l'étape 1. A la fin de cette seconde étape, chaque cluster de termes représente un ensemble de sommets d'une KDM . Le processus de clustering est relancé sur chacun des ensembles de termes d'une KDM si la dimension de cette dernière dépasse DK . Le processus de clustering de cette seconde étape est itératif. L'objectif du processus de clustering itératif est de permettre d'éclater un cluster en plusieurs autres clusters, i.e. une KDM en plusieurs autres $KDMs$, permettant ainsi une visualisation des résultats du clustering et une navigation par le contenu dans une base documentaire.

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, cordial analyseur (Web 1 – Cf. Références) qui intègre un étiqueteur morphosyntaxique et un lemmatiseur fonctionnant pour les documents textuels en Français. Nous avons développé une collection d'outils permettant de mettre en oeuvre l'approche *IC-Doc* ainsi qu'un prototype de visualisation des résultats en *KDMs* (FIG 2 illustre l'interface de ce prototype).

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 dans le tableau 1.

| Documents analysés par étape | Economie <i>Nb_Doc</i> | Informatique <i>Nb_Doc</i> | Cinema <i>Nb_Doc</i> |
|---------------------------------|---------------------------|-------------------------------|-------------------------|
| 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 :

$$\text{Précision} = |S \cap V| / |S| \quad \text{Rappel} = |S \cap V| / |V|$$

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 se 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 illustre TAB 2, la précision ou le rappel ne peuvent chuter pour une thématique que dans le cas de thématiques pauvres dans une collection (par exemple 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 collection de documents

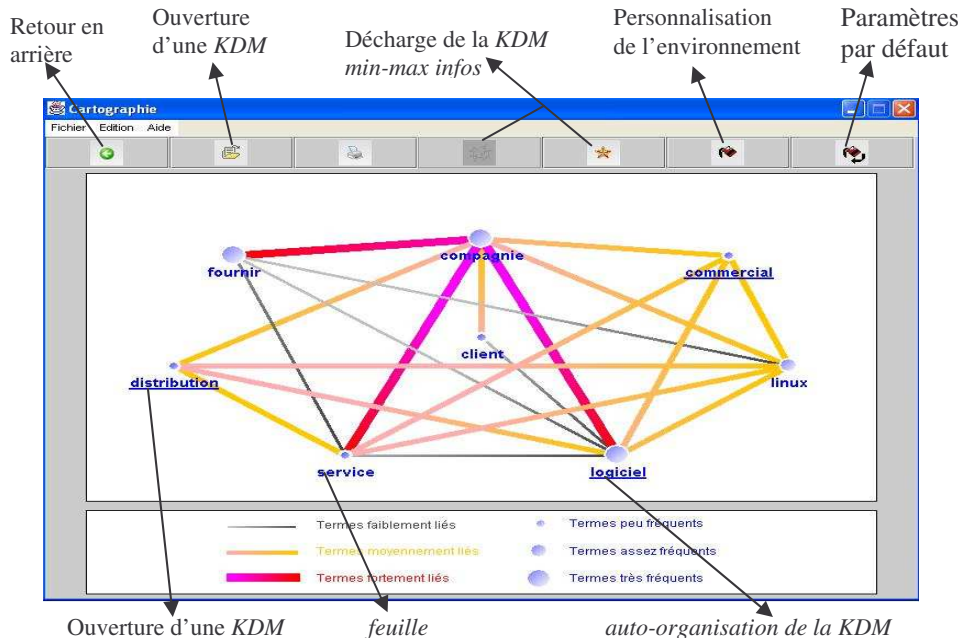


FIG 2 – Interface du prototype de visualisation des résultats en KDMs

5. Travaux connexes

Les outils et les méthodes de fouille de textes permettent l'acquisition, le classement, l'analyse, l'interprétation, l'exploitation et la visualisation 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 (Baldi et Di meglio 2004, Besançon R 2001, Chen et al. 2001, Hongyuan et al. 2001, Han et Kamber 2000, Turenne 2000, Ihadjadene 2004). 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, 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 IWASAKI (Tanaka et Iwasaki 1996) utilisent la matrice de co-occurrences pour la désambiguïsation des termes. Dans (Besançon R 2002) 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, à la fois, 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 caractérisation de bases documentaires est donc limitée dans la mesure où elle ne permet pas une extraction pertinente et représentative des informations sur les différents contenus textuels.

6. Conclusion

Dans le cadre de cet article, nous avons présenté l'approche *IC-Doc* de caractérisation thématique de connaissances textuelles à partir de collections de documents. *IC-Doc* 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 montre la pertinence de notre approche et sa capacité pour une caractérisation thématique des collections de documents indépendamment des poids donnés aux thématiques dans les collections de documents. Lorsque les thématiques se chevauchent, les résultats du clustering pourraient être améliorés par des techniques de clustering flou. Ces techniques font l'objet de nos travaux en cours sur l'extraction et la caractérisation automatique de connaissances textuelles à partir de diverses collections de documents, facilitant les consultations documentaires et favorisant les échanges d'expériences entre utilisateurs.

7. Références

- Baldi S. et Di meglio E. (2004), A text mining strategy based on local contexts of words, Proceedings of JADT'04, Le poids des mots, Presses Universitaires de Louvain, Vol. 2, pp 79-87.
- Besaçon R. (2002), Filtrages syntaxiques de co-occurrences pour la représentation vectorielle de documents, Actes de TALN'02, pp 135-144.
- Besaçon R. (2001), Intégration de connaissances syntaxiques et sémantiques dans les représentations vectorielles des textes, PhD thesis, Ecole polytechnique Fédérale de Lausanne, 2001.
- Chung W., Chen H. et Nunamaker J. (2003), Business intelligence explorer: A knowledge map framework for discovering business intelligence on the Web, Proceedings of HICSS'03, 10 p.
- Jain A.K., Murty M.N. et Flynn P.J. (1999). Data clustering: A review. ACM Computing Surveys, Vol. 31, Issue 3, pp 264-323.
- Chen H., Fan H., Chau M. et Zeng D. (2001), MetaSpider: Meta-searching and categorization on the Web, Journal of the American Society for Information Science and Technology, Vol. 52, pp 1134 –1147.
- Kosala R. et Blockeel H. (2000), Web Mining research: A survey. SIGKDD Explorations, 2(1), pp 1-15.
- Hongyuan Z., Xiaofeng H., Chris D., Ming G., et Horst S. (2001), Automatic topic identification using webpage clustering, Proceedings of ICDM'01, pp 25-31.
- Han. J. et Kamber. M. (2000), Data mining: concepts and techniques, Morgan Kaufmann Publishers, ISBN 1-55860-489-8, 2000.
- Mokrane A, Poncelet. P et Dray. G. (2004), Visualisation automatique du contenu d'une base de documents textuels via les hyper-cartes d'information, Actes des VSST'04, pp 239-250.
- Mokrane. A, Arezki. R, Dray. G et Poncelet. P. (2004): Cartographie automatique du contenu d'un corpus de documents textuels. Actes des JADT'04, Le poids des mots, Presses Universitaires de Louvain, Vol. 2, pp 816-823.
- Poibeau T. (2003), Extraction automatique d'information, du text mining au Web sémantique. Hermès sciences publications, ISBN : 2-7462-0610-2, 2003.
- Pereira, F., Tishby, N. et Lee, L. (1993), Distributional clustering of English words. Proceedings of the 31th Meeting of the Association for Computational Linguistics, pp183-190.
- 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.
- Tanaka K. et Iwasaki H. (1996), Extraction of lexical translations from non-aligned corpora, Proceedings of the 16th International Conference on Computational Linguistics, pp 580-585.
- 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 de Strasbourg, 2000.
- Ihadjadene M. (2004), Méthodes avancées pour les systèmes de recherche d'informations, Ouvrage collectif sous la direction de M. Ihadjadene, Hermès sciences publications, ISBN : 2-7462-0846-6, 2004.
- Web 1. Cordial analyseur, Site Web,
http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm