L'exploitation des relations d'association de termes pour l'enrichissement de l'indexation de documents textuels

Férihane Kboubi, Anja Habacha Chabi, Mohamed BenAhmed

Laboratoire RIADI, ENSI, Campus Universitaire de la Manouba 2010

Résumé

Notre travail se situe dans le cadre d'un projet d'annotation descriptive, conceptuelle et thématique de corpus textuel. Dans le présent article, nous focalisons notre attention sur l'annotation conceptuelle et plus précisément sur la tâche d'extraction des termes pertinents. Pour ceci, nous proposons trois méthodes d'extraction de termes. Ces méthodes diffèrent du point de vue sources d'information utilisées (texte intégral, résumé et/ou annotations descriptives) et de point de vue méthodes d'estimation de la pertinence des termes (calculs basés sur la fréquence des termes, calculs basés sur les relations d'association entre les termes). Afin d'évaluer ces trois méthodes nous avons eu recours à un ensemble de 19 classifieurs. Une discussion sur nos résultats met en évidence l'importance de l'exploitation des réseaux de cooccurrences dans un processus d'extraction de termes notamment dans l'objectif d'enrichir l'indexation conceptuelle de documents textuels.

Abstract

Our work is situated within the framework of a project of descriptive, conceptual and thematic annotation of textual corpus. In the present paper, we focus our attention on the conceptual annotation and more exactly on the task of relevant term extraction. We propose three methods of term extraction. These methods are different in too points: the sources of information (complete text, summarized text and/or descriptive annotations) and the way the terms are weighted (term frequency, cooccurrence relationship). To evaluate these methods we used a set of 19 classifiers. A discussion about our results show the importance of using the cooccurrence networks in a process of term extraction in particular to enrich the conceptual indexing of textual documents.

Keywords: term extraction, cooccurrence network, association relationship, document annotation

1. Introduction

Dans le cadre du Web sémantique, Euzenat (2002) formalise l'annotation comme étant une relation entre deux ensembles d'objets: l'ensemble *D* de documents et l'ensemble *C* des représentations formelles. Cette relation définit deux fonctions inverses. Premièrement, l'annotation, qui cherche à identifier les représentations formelles pertinentes pour chaque document. Deuxièmement, l'indexation, qui cherche à associer à chaque représentation formelle l'ensemble de documents qui lui sont relatifs.

Dans cet article, nous focalisons notre attention sur le processus d'extraction des termes pertinents en vue d'associer à chaque document une représentation formelle sous forme d'annotations conceptuelles. Dans la littérature, les méthodes d'extraction de termes se classent en deux grandes catégories à savoir: les méthodes statistiques (Bellot and El-Bèzel, 2001 ; Vercoustre et al., 2006 ; Kboubi et al., 2008) et les méthodes linguistiques (Forest and Meunier, 2004 ; Hernandez, 2005). Les méthodes statistiques se basent sur le calcul de la fréquence

des termes et sont indépendantes des langues. De l'autre côté, les méthodes syntaxiques sont basées sur l'analyse du rôle grammatical des mots. De ce fait, elles ont besoin de processus syntaxiques qui sont spécifiques à la langue traitée.

Dans notre travail, nous nous intéressons aux méthodes statistiques. Mais au lieu de nous baser uniquement sur des calculs reposant sur la fréquence des termes, nous proposons aussi d'exploiter les relations d'associations des termes et ceci par la construction de réseau de cooccurrences de termes. Cette double approche permet de bénéficier des avantages des deux démarches. En effet, les relations de cooccurrences permettent de mettre en évidence les termes qui peuvent ne pas être très fréquents mais qui ont des relations de cooccurrences fortes et qui sont susceptibles d'être pertinents.

Notre article est organisé comme suit. Dans les sections 2, 3 et 4 nous proposons respectivement trois méthodes d'extraction de termes. La première se base pour la pondération des termes sur des calculs de fréquence et consiste à combiner les résultats de TextTiling et de la méthode d'indexation sémantique latente (LSI). La deuxième méthode se fonde sur la construction d'un réseau local de cooccurrence. La troisième méthode repose sur la construction et l'exploitation d'un réseau global de cooccurrences pour l'enrichissement des termes générés par la première méthode. Dans la section 5, nous présentons et discutons les résultats de l'évaluation des trois méthodes proposées. Finalement, dans la section 6, nous discutons des perspectives du travail de recherche présenté dans cet article.

2. Extraction de termes par combinaison de TextTiling et LSI

Notre idée est ici d'exploiter l'outil TextTiling (Hearst, 1997), qui est un résumeur automatique de document textuel, pour l'extraction des termes candidats. Un terme candidat est un terme qui est susceptible d'être pertinent et peut être considéré comme un terme clé. Par la suite, nous proposons de sélectionner les termes clés par l'application de la technique d'indexation sémantique latente (LSI), puis d'appliquer un filtrage statistique pour éliminer les termes non pertinents.

L'extraction des termes candidats se fait par la détermination de l'ensemble des termes distincts à partir du titre, du résumé des auteurs, du résumé de TextTiling, des mots-clés des auteurs et des mots-clés de TextTiling. Les résumés étant courts, par nature, nous proposons alors de ne pas utiliser aucune méthode de sélection de termes. En effet, les termes ont tous des fréquences faibles dans les résumés, l'utilisation d'indicateurs de pondération de termes tels que *tf.idf* n'apportera pas une information importante relativement à la pertinence des termes. Nous considérons alors comme candidats tous les termes distincts obtenus à partir des deux résumés après l'étape de prétraitement, auxquels nous ajoutons les termes du titre, les mots clés des auteurs et de TextTiling.

Pour sélectionner les termes représentatifs à partir de cet ensemble de termes, nous proposons d'utiliser la méthode LSI. Notre choix de LSI est justifié par le fait que cette technique est adaptée au traitement d'un grand nombre de données et qu'elle soit indépendante des langues car elle ne requiert pas de ressources externes dont la qualité est trop dépendante de la personne qui les construit. Le module LSI reçoit ainsi en entrée le document intégral et le vecteur des termes distincts (les termes candidats) puis pondère chaque terme candidat en fonctions de sa densité dans le document et ceci selon la Formule 1:

$$densité(t_i) = \frac{fr(t_i)}{N} \times k \times 100 \quad (1)$$

où $fr(t_i)$ représente la fréquence de t_i dans le document, N représente le nombre total de termes dans le document, et k représente le nombre de mots constituants le terme (si le terme est simple k=1; sinon, le terme est composé de plusieurs mots, alors k est égal à ce nombre de mots).

La pondération des termes par la densité permet de sélectionner les termes représentatifs qui ont une valeur de densité élevée et d'éliminer les termes ayant une densité faible: c'est l'étape de filtrage statistique. À la fin de cette étape nous obtenons un vecteur de termes clés où la pertinence de chaque terme est déterminée en fonction de sa fréquence dans le document (la densité étant calculée en fonction de la fréquence du terme). Les termes à faible fréquence se trouvent alors écartés même si en réalité certains d'entre eux peuvent être pertinents et représentatif du contenu du document. Pour remédier à cette limite, nous proposons une autre méthode d'extraction des termes dans laquelle la pertinence du terme est mesurée en fonction de ses relations d'association en plus de sa fréquence. Nous présentons le principe de cette deuxième méthode dans la section qui suit.

3. Extraction de termes par analyse des relations d'association locales

Avec cette deuxième méthode, nous partons de l'hypothèse qu'un terme ayant plusieurs relations d'associations fortes avec les autres termes des documents est susceptible d'être pertinent. Les relations d'associations sont déterminées en fonction des relations de cooccurrences des termes. La réalisation de cette deuxième méthode nécessite la construction du réseau de cooccurrence du document.

Un réseau de cooccurrence est un réseau de termes où chaque nœud représente un terme et un arc entre deux nœuds représente la relation de cooccurrence entre les deux termes concernés. Ce réseau permet d'identifier les termes qui apparaissent souvent ensemble au sein d'une même fenêtre mais pas nécessairement juxtaposés. Il s'agit de représenter chaque terme par le vecteur de ces cooccurrences avec les autres termes. Notre méthode de construction du réseau de cooccurrence repose sur la recherche de termes dans les textes. Elle suppose que les associations fréquentes de deux termes peuvent être révélatrices de relations sémantiques et constituer ainsi des éléments pouvant être intégrés dans le réseau.

La construction du réseau de cooccurrence d'un document se base en entrée sur le titre, le résumé des auteurs, les mots clés des auteurs et le corps du document. Les relations d'association entre les termes sont déterminées en glissant une fenêtre de taille L d'un mot en un mot sur le texte du document. À chaque position, nous enregistrons les cooccurrences entre le mot de tête et les autres mots de la fenêtre. Puis nous filtrons les cooccurrences non significatives. Une association entre deux termes t_1 et t_2 est notée comme suit : $t_1 \rightarrow t_2$. Elle signifie que si t_1 apparaît alors t_2 apparaît également dans une fenêtre de taille L. Deux indices sont couramment utilisés pour mesurer la pertinence des règles d'association. Ces indices sont le degré de support et le degré de confiance. Le support représente le nombre de documents qui sont décrits par les termes présents en partie gauche et droite de la Règle 2:

$$\sup(t_1 \rightarrow t_2) = \left|t_1 \wedge t_2\right| \tag{2}$$

avec $|t|$ le nombre de documents contenant le terme t. C'est la probabilité d'apparition de l'ensemble des documents correspondant à $t1 \wedge t2$.

La confiance mesure le degré de validité d'une règle, c'est-à-dire lorsqu'il existe des contre-exemples de documents qui vérifient t_1 et pas nécessairement t_2. En terme probabilistes, la confiance mesure la probabilité conditionnelle de t_2 sachant t_1. Lorsque la confiance vaut 1, la règle est dite totale. Dans le cas de notre exemple la confiance (Formule 3) vaut:

$$conf(t_1 \rightarrow t_2) = \frac{\sup(t_1 \rightarrow t_2)}{|t_2|} = \frac{|t_1 \wedge t_2|}{|t_1|} \in [0,1] \quad (3)$$

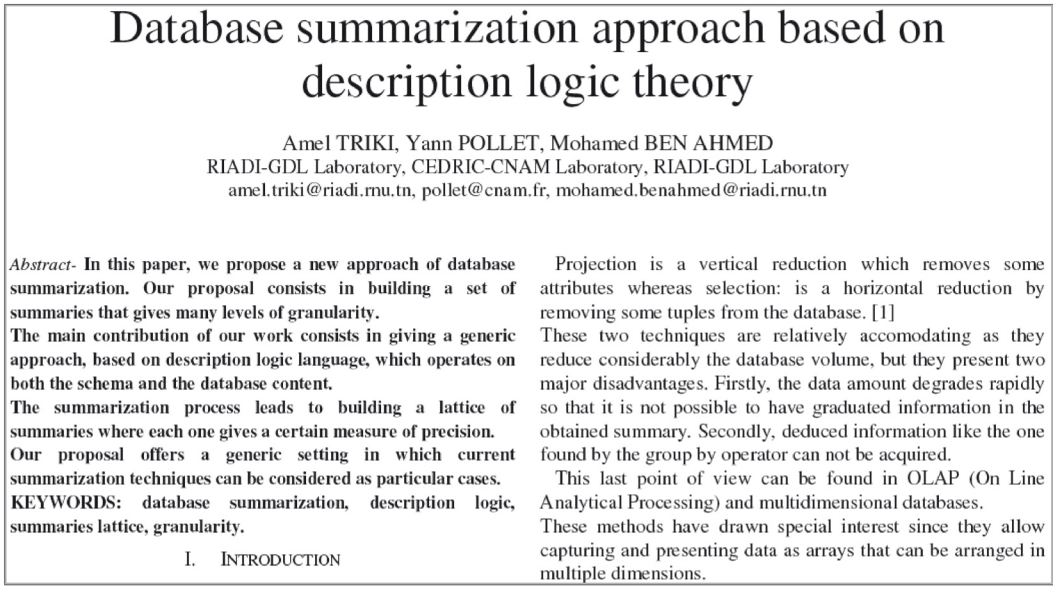
Pour la construction du réseau de cooccurrences, nous nous inspirons du travail de Ferret et al. (1997). Cependant, au lieu de pondérer chaque arc par la fréquence de cooccurrence des deux termes qu'il relie, nous proposons d'utiliser plutôt le degré de confiance. Les avantages de l'utilisation du degré de confiance par rapport à la fréquence absolue sont multiples. Premièrement, ne pas pénaliser les cooccurrences des termes qui n'ont pas une fréquence élevée mais qui peuvent être pertinents, comme dans l'exemple 1 :

Exemple 1 : $facial \rightarrow imag$
fréquence de facial = 7, fréquence de cooccurrence (*facial, imag*) =7,
$conf(facial \rightarrow imag) = 1$

Deuxièmement, ne pas surestimer l'importance des cooccurrences des termes très fréquents, comme dans l'exemple 2 :

Exemple 2 : $system \rightarrow provid$
fréquence de system= 180, fréquence de cooccurrence (*system, provid*) =17,
$conf(system \rightarrow provid) = 0.094$

Nous illustrons par les Fig. 1 et Fig. 2 la construction du réseau de cooccurrences d'un document de notre corpus. Les labels des arcs entre deux termes représentent les degrés de confiance de la relation d'association entre ces deux termes. Ce réseau est construit avec une fenêtre de taille 15 et un seuil de 0,4 pour le filtrage des relations de cooccurrences. Nous constatons sur le réseau la formation de trois agglomérations autour des termes suivants: *Database*, *Summarization* et *Logic*. Ces termes sont en fait, les principaux termes clés du document.



Database summarization approach based on description logic theory

Amel TRIKI, Yann POLLET, Mohamed BEN AHMED
RIADI-GDL Laboratory, CEDRIC-CNAM Laboratory, RIADI-GDL Laboratory
amel.triki@riadi.rnu.tn, pollet@cnam.fr, mohamed.benahmed@riadi.rnu.tn

Abstract- **In this paper, we propose a new approach of database summarization. Our proposal consists in building a set of summaries that gives many levels of granularity.**
The main contribution of our work consists in giving a generic approach, based on description logic language, which operates on both the schema and the database content.
The summarization process leads to building a lattice of summaries where each one gives a certain measure of precision.
Our proposal offers a generic setting in which current summarization techniques can be considered as particular cases.
KEYWORDS: database summarization, description logic, summaries lattice, granularity.

I. INTRODUCTION

Projection is a vertical reduction which removes some attributes whereas selection: is a horizontal reduction by removing some tuples from the database. [1]
These two techniques are relatively accomodating as they reduce considerably the database volume, but they present two major disadvantages. Firstly, the data amount degrades rapidly so that it is not possible to have graduated information in the obtained summary. Secondly, deduced information like the one found by the group by operator can not be acquired.
This last point of view can be found in OLAP (On Line Analytical Processing) and multidimensional databases.
These methods have drawn special interest since they allow capturing and presenting data as arrays that can be arranged in multiple dimensions.

Figure 1 : Aperçu d'un extrait du document

Par ailleurs, l'utilisation des degrés de confiance permet aussi de mettre en évidence les termes qui apparaissent toujours ensemble (dans une même fenêtre) et qui très souvent forment en réalité des termes composés (des termes constitués de plus qu'un seul mot). La Fig. 3 illustre un exemple de cette situation. Les deux termes « *SMART* » et « *CARD* » ont une relation de

cooccurrence de degré de confiance égale à 1 : $conf(SMART \rightarrow CARD) = 1$. Ces deux termes sont alors regroupés pour former le terme composé « *SMART CARD* ». De même pour les termes « *VIRTUAL* » et « *MACHIN* » qui ont une relation de degré de confiance égal à 1, et qui sont regroupés pour former le terme « *VIRTUAL MACHIN* ».

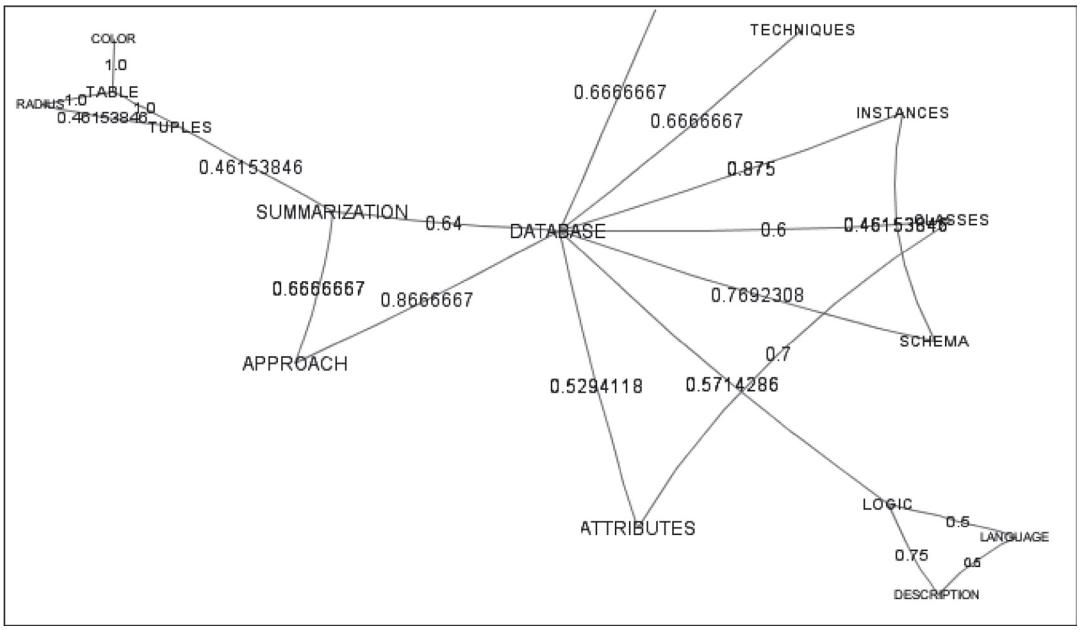


Figure 2: Réseau de cooccurrence du document

La sélection des termes pertinents ne se base pas uniquement sur le calcul du nombre de relations du terme candidat avec les autres termes, mais aussi sur le degré de confiance de ces relations. Un terme ayant un grand nombre de relations toutes avec un degré de confiance faible n'est pas susceptible d'être pertinent. Par exemple, le terme *approach* nous le retrouvons dans tous les domaines et dans la majorité des documents. Ce terme a un grand nombre de relations mais toutes de faible degré de confiance.

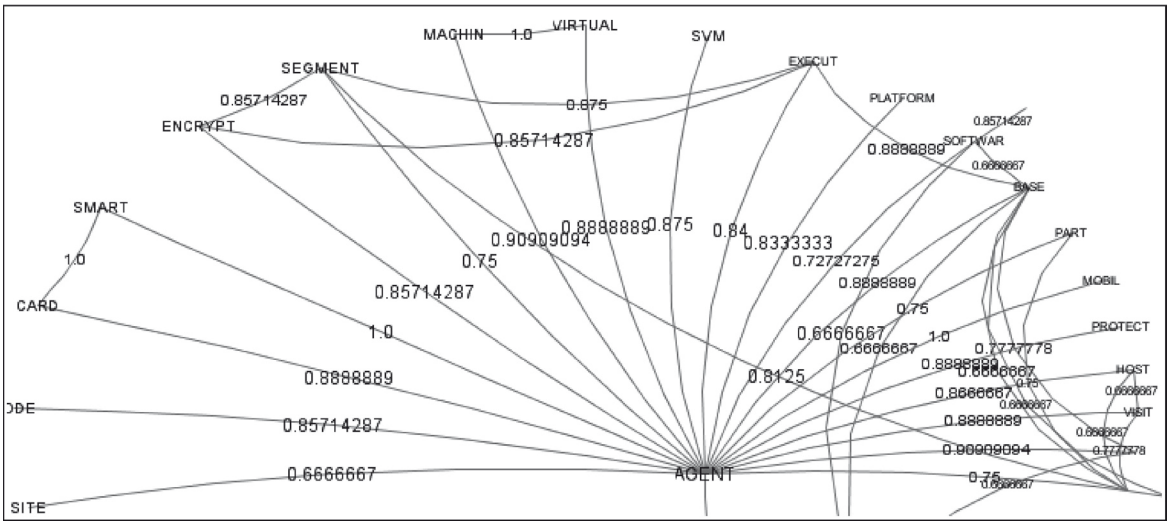


Figure 3: Exemple de détection de termes composés par l'identification des relations d'association

C'est pour cette raison que nous avons établit un seuil pour le degré de confiance en dessous duquel la relation de cooccurrence entre deux concepts n'est pas considérée comme pertinente.

Ce seuil est utilisé pour le filtrage du réseau de cooccurrence. Les termes restant dans le réseau après l'étape de filtrage sont considérés comme pertinents. La construction du réseau de cooccurrence d'un document permet de mettre en évidence les relations d'association entre les termes du document. Elle permet aussi d'estimer la pertinence des termes en fonction du nombre et de l'importance de leurs relations.

4. L'enrichissement des vecteurs de documents par l'exploitation de réseau global de cooccurrence

L'idée ici est d'étendre chaque vecteur de document formé par la combinaison de TextTiling et de LSI par les termes qui sont fortement associés aux termes du vecteur. L'intérêt de l'extension des vecteurs des documents par des informations déduites d'un réseau de cooccurrences réside dans le fait que les relations d'association entre les termes peuvent représenter implicitement des relations sémantiques, telles que : la synonymie (dans un paragraphe ont peut désigner un concept par plusieurs termes différents pour éviter la redondance), tout ou partie (*Database*, *Table*), dépendance forte (*Mobil agent*, *Network*). Le fait de décrire chaque terme par l'ensemble des termes qui lui sont fortement associés permettrait de prendre en compte de ces relations sémantiques pour la représentation des documents. Ces relations sémantiques sont matérialisées par des relations d'associations déduites à partir de la fréquence de cooccurrence des termes dans le même contexte (fenêtre de taille fixe de mots). Nous considérons une cooccurrence de deux termes t_1 et t_2 comme l'apparition de ces deux termes dans une même fenêtre. La cooccurrence de t_1 avec t_2 ne signifie pas obligatoirement qu'ils sont juxtaposés. Ils peuvent, en effet, être séparés par d'autres termes.

Notre démarche pour l'enrichissement d'un vecteur de document consiste à établir pour chaque terme ti du vecteur un ensemble de termes descriptifs. Pour construire cet ensemble il s'agit d'extraire à partir du réseau de cooccurrences global les termes avec lesquels ti a une relation de cooccurrence forte (degré de confiance > à un seuil). Par la suite, nous vérifions si les termes descriptifs obtenus appartiennent ou non au vecteur initial. Si un terme descriptif appartient au vecteur alors nous ajustons son poids en fonction de son degré d'association au terme ti. Sinon, nous l'ajoutons au vecteur. Le principe de cette étape est décrit dans l'algorithm :

- Pour chaque terme t_1 d'un vecteur de document et t_2 un terme associé à t_1 par une relation de cooccurrence de degré de confiance w
- Si t_2 est présent dans le vecteur
  Ajouter wp_2 au poids de t_1 et wp_1 au poids de t_2 *{p_1 et p_2 sont les poids initiaux de t_1 et de t_2}*
- Sinon *{t2 n'est pas présent dans le vecteur}*
  Ajouter t_2 *au vecteur du document avec un poids égale à* wp_1

La Fig. 4 illustre le principe de notre méthode d'enrichissement des vecteurs de termes. Nous identifions les relations d'association entre les termes à l'aide d'un réseau de cooccurrences préalablement construit. Ce réseau de cooccurrences n'est pas le même que celui construit dans l'étape précédente. Le réseau que nous utilisons ici est un réseau global. Il est construit sur la base de tous les documents du corpus et non pas sur un seul document comme dans l'étape précédente. Il s'agit ici d'avoir une idée sur les relations d'association globales dans le corpus et non pas locales à un document.

Pour la construction du réseau de cooccurrences global nous avons suivi le même principe que pour la construction du réseau local, sauf qu'au lieu d'utiliser le texte intégral de chaque document nous utilisons plutôt le résumé automatique et les termes-clés de TextTiling. En effet, se baser sur toute la partie *body* de tous les documents du corpus est très couteux en

terme de temps d'exécution et d'espace mémoire nécessaire. L'utilisation des résumeurs automatiques est une alternative très intéressante pour remédier à ces limites. Ainsi, le réseau de cooccurrences global est construit à partir de tous les documents du corpus. Chaque document est représenté par son titre, le résumé et les termes-clés de l'auteur, et le résumé et les termes-clés de TextTiling.

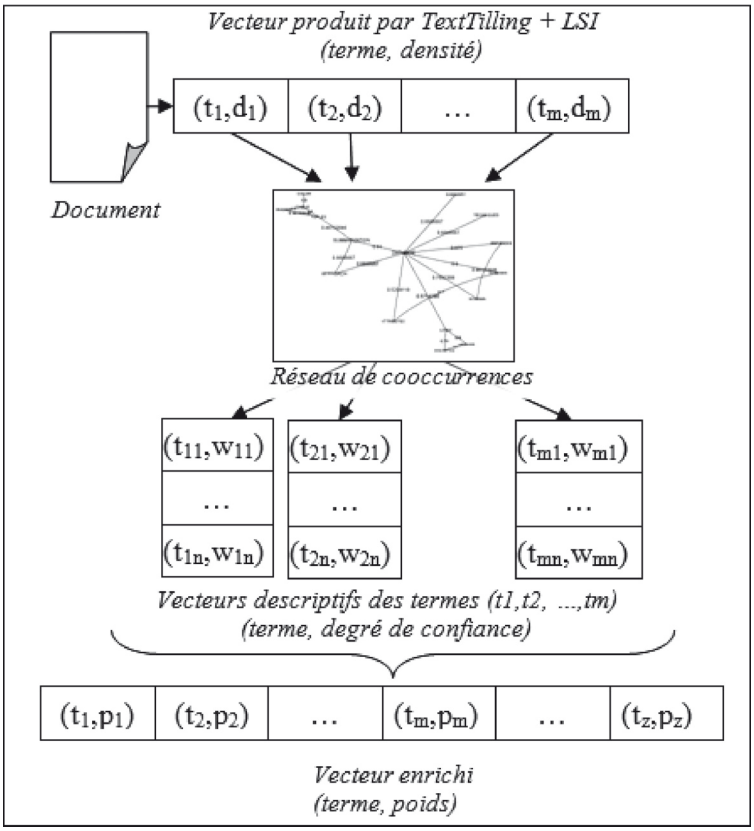


Figure 4 : Enrichissement des vecteurs de Termes

5. Expérimentations

5.1. Méthode d'évaluation

L'évaluation des méthodes d'extraction de termes est une tâche très problématique du fait que la pertinence des termes peut être estimée par plusieurs méthodes, chacune pouvant attribuer un poids différents à un même terme et pouvant sélectionner un ensemble de termes différents que les autres méthodes. Pour évaluer nos trois méthodes d'extraction de termes nous avons pensé à suivre une démarche qui se base sur l'utilisation d'algorithmes de classification supervisée. Pour ceci nous avons choisi un ensemble de 19 classifieurs utilisant des méthodes et des techniques différentes à savoir: six classifieurs bayésiens (*complementNaiveBayes, DMNBtext, NaiveBayes, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, NaiveBayesUpdateable*), un classifieur utilisant SVM (*SMO*), un classifieur *Réseau RBF*, les classifieurs *logistic*, *simpleLogistic* et neuf classifieurs de type arbre de décision (BFTree, FT, j48, j48graft, LADTree, LMT, NBTree, RandomForest, RandomTree) [1].

[1] Nous avons utilisé les implémentations des classifieurs de WEKA (http://wekaclassalgos.sourceforge.net/)

L'idée est de comparer les résultats de ces classifieurs en utilisant les trois types de vecteurs de termes générés par les trois méthodes d'extraction de termes. Se baser sur un ensemble de classifieurs au lieu d'un seul procure une fiabilité plus grande aux résultats d'évaluation par l'élimination de l'hypothèse qui affirme que la performance de certaines méthodes de sélection d'attributs (termes) varie selon le classifieur sur lequel elles sont testées (Brank et al., 2002). Les évaluations sont réalisées sur une base de test constituée par un ensemble de documents représentants les publications de notre laboratoire durant les trois années 2006, 2007 et 2008. Ces documents sont répartis en trois thèmes majeurs à savoir l'intelligence artificielle, la sécurité et les systèmes d'information

Tous les documents de la base ont passé par une étape de prétraitement. Lors de laquelle nous commençons tout d'abord par l'élimination des mots vides. Se sont les termes d'usage qui ont une fréquence souvent élevée et qui n'apporte aucune information pour la description du document. Ensuite nous effectuons un filtrage statistique, c'est-à-dire nous éliminons les termes ayant une fréquence inférieure à un seuil. En effet, les termes peu fréquents sont les plus nombreux, en conséquence le nombre des termes d'indexation peut augmenter de manière très importante avec la taille du corpus. L'élimination de ces termes est donc justifiée surtout si on connait que la plupart d'eux sont peu représentatifs du contenu d'un document et ne sont donc pas discriminant pour sa description. Finalement, nous procédons à la lemmatisation des termes restants à l'aide de l'algorithme de Porter (1980). La lemmatisation permet de regrouper en un seul attribut les multiples formes morphologiques de mots qui ont une sémantique commune.

5.2. Résultats et discussions

Dans ce qui suit, nous notons « *VTTLSI »* le vecteur produit par la combinaison de TextTiling et de LSI, *« Vcooc »* le vecteur des termes extraits du réseau local de cooccurrences, et « *VTTLSIEnrich »* le vecteur produit par l'enrichissement de « *VTTLSI »* par le réseau global de cooccurrences [2].

Le Tableau 1 présente les résultats des algorithmes de classification en utilisant ces trois de types de vecteurs. Les valeurs numériques reportées dans ce tableau représentent les taux de bonne classification calculés par WEKA suite à une validation croisée. Il est à indiquer que nous avons utilisé les algorithmes de classification de WEKA sans modifier leurs paramètres. Les valeurs grisées représentent les meilleurs taux de bonne classification obtenus par chaque classifieur. A partir de ce tableau, nous constatons que les meilleurs taux de classification sont engendrés par l'utilisation des vecteurs basés sur les réseaux de cooccurrences *VTTLSIEnrich* et *Vcooc*. Ces vecteurs ont généré les meilleurs résultats avec 13 classifieurs sur les 19 testés, et ont obtenu des résultats égaux à ceux de *VTTLSI* avec 4 classifieurs. En effet, les résultats engendrés par *VTTLSI* ont dépassé uniquement à deux reprises (avec *SimpleLogistic* et *LMT*) les résultats obtenus avec les deux autres vecteurs. Par contre, *VTTLSIEnrich* a donné les meilleurs taux de classification avec 5 classifieurs (*NaiveBayes*, *NaiveBayesUpdateable*, *FT*, *LADTree* et *RandomTree*). De son côté, *Vcooc* généré les taux les plus élevés avec 7 classifieurs (*complementNaiveBayes*, *NaiveBayesMultinomial*, *Logistic*, *BFTree*, *j48*, *j48graft*, *NBTree* et *RandomForest*).

[2] Il est à noter que le réseau local de cooccurrences est différent du réseau global de cooccurrences. Le premier est construit à partir d'un seul document. Il représente les relations d'association des termes au sein d'un même document. Le second est construit à partir de tous les documents du corpus. Il représente les relations d'association globales des termes du corpus. Nous l'utilisons pour l'enrichissement des vecteurs VTTLSI.

Classifieurs WEKA	*VTTLSI*	*VTTLSI-Enrich*	*Vcooc*
complementNaiveBayes	62,85	74,28	88,57
DMNBtext	88,57	88,57	85,71
NaiveBayes	80	85,71	80
NaiveBayesMultinomial	62,85	65,71	82,85
NaiveBayesMultinomialUpdateable	77,14	82,85	82,85
NaiveBayesUpdateable	80	85,71	80
Logistic	77,14	77,14	82,85
RBFNetwork	80	80	80
SimpleLogistic	85,71	82,85	74,28
SMO	82,85	82,85	82,85
BFTree	60	68,57	71,42
FT	85	85,71	77,14
j48	77,14	71	77,14
j48graft	74,28	71,42	77,14
LADTree	71,42	80	71,42
LMT	85,71	82,85	74,28
NBTree	71,42	65,71	77,14
RandomForest	80	77,14	82,85
RandomTree	71,42	80	60

Tableau 1 : Performances de classification (en %) avec les trois méthodes d'extraction de termes

Les résultats obtenus ont montré que, pour les mêmes documents et les mêmes classifieurs, les performances de classification varient en fonction des vecteurs de termes extraits. Ceci souligne l'importance de l'étape d'extraction des attributs des objets dans un processus de classification. Nous avons aussi remarqué que les méthodes d'extraction de termes basées sur l'analyse des relations d'association entre les termes donnent des résultats soit meilleurs soit équivalent aux résultats obtenus par la combinaison TextTiling-LSI qui est basée uniquement sur la fréquence des termes. Ce qui nous permet de dire que les approches basées sur les relations d'association et l'enrichissement par réseau de cooccurrences sont capables de générer des vecteurs de termes plus descriptifs que ceux générés par des méthodes basées sur des calculs de fréquence. Par ailleurs, nous avons aussi remarqué que les performances de classification produites par l'utilisation de Vcooc sont comparables à celles obtenues par l'utilisation de *VTTLSIEnrich*. En effet, à 8 reprises Vcooc dépasse *VTTLSIEnrich*, à 8 reprises *VTTLSIEnrich* dépasse Vcooc et à 3 reprises les résultats de Vcooc et *VTTLSIEnrich* sont égaux.

Sachant que la difficulté d'implémentation de Vcooc est nettement inférieure à celle de *VTTLSIEnrich*, l'utilisation de *Vcooc* devient une alternative attirante pour l'extraction des termes pertinents d'un document. En effet, la construction d'un réseau de cooccurrences d'un seul document nécessite moins de temps de calcul que dans le cas d'un corpus de documents. D'autre part, la construction d'un réseau de cooccurrences global est une tâche qui dépend beaucoup des thèmes inclus dans le corpus.

6. Conclusion et perspectives

L'extraction des termes pertinents d'un document est une tâche essentielle pour plusieurs processus tel que la classification, l'analyse thématique, l'annotation et l'indexation de documents. En effet, les performances de ces processus dépendent fortement des résultats de cette tâche. Dans le cadre de notre projet, nous nous intéressons à la tâche d'extraction des termes dans un objectif d'annotation conceptuelle et thématique de corpus textuels. Dans cet article nous avons proposé et évalué trois nouvelles méthodes d'extraction de termes. Les tests réalisés ont montré que les méthodes exploitant les relations d'association entre les termes peuvent améliorer considérablement les performances des méthodes statistiques d'extraction de termes. Comme perspective immédiate à notre travail, nous envisageons de tester la combinaison des réseaux locaux de cooccurrence avec la méthode basée sur TextTiling et LSI. Par ailleurs, il est nécessaire de poursuivre ce travail par l'intégration des différentes méthodes décrites dans cet article dans un processus d'annotation conceptuelle et thématique de documents textuels.

Références

Bellot P. and El-Bèze M. (2001). Classification et segmentation de texte par arbre de decision. *Technique et science informatiques*, 20(1) : 107-134.

Brank J., Grobelnik M., Milic-Frayling N. and D. Mladenic (2002). Interaction of Feature Selection Methods and Linear Classification Models. In *Proceedings of ICML-02, 19th Conference on Machine Learning, Workshop on Text Learning*, Sydney, Australia.

Bouckaert R.R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. and Scuse D. (2008). *WEKA Manual for Version 3-6-0*, December 18, 2008.

Euzenat J. (2002). Eight questions about Semantic Web annotations, Intelligent Systems. *IEEE*, 17(2): 55-62.

Ferret O., Grau B. and Masson N. (1997). Utilisation d'un réseau de cooccurrences lexicales pour améliorer une analyse thématique fondée sur la distribution des mots. In *1ères Journées du Chapitre Français de l'ISKO : Organisation des connaissances en vue de leur intégration dans les systèmes de représentation et de recherche d'information*. Presses Universitaires de Lille.

Forest D. and Meunier J. (2004). Classification et catégorisation automatiques: application à l'analyse thématique des données textuelles. In *JADT : 7ème Journées internationals d'Analyse statistique de données Textuelles*.

Hearst M. (1997). TextTiling : Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics*, 23(1) : 33-64.

Hernandez N. (2005). *Ontologies de domaine pour la modélisation du contexte en recherche d'information*. Thèse de doctorat, Institut de recherche en informatique de Toulouse, Université Paul Sabatier.

Kboubi F., Habacha A. and BenAhmed M. (2008). Méthodes d'extraction de termes basées sur une combinaison d'indicateurs. In *Colloque international sur le document électronique (CIDE11)*.

Porter M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3): 130-137.

Vercoustre A., Fegas M., Lechevallier Y. and Despeyroux T. (2006). Classification de documents XML à partir d'une représentation linéaire des arbres de ces documents. *EGC*: 433-444.