Introduction de la sémantique d'un document sous le modèle de langage

Arezki Hammache* — Mohand Boughanem — Rachid Ahmed-Ouamer***

* *Laboratoire LARI, Université Mouloud Mammeri*
15000 Tizi-Ouzou, Algérie
{arezki20002002, ahm_r}@yahoo.fr

** *Laboratoire IRIT, Université Paul Sabatier*
118route de Narbonne 31062 Toulouse Cedex 09, France
bougha@irit.fr

RÉSUMÉ. *La plupart des systèmes de recherche d'information classiques se basent sur une indexation par termes simples. Cependant, ces derniers délivrent beaucoup de résultats en réponse aux requêtes des utilisateurs. Ceci est du en partie au fait que le contenu sémantique d'un document (ou d'une requête) ne peut pas être capturé précisément par un simple ensemble de mots clés indépendants. Deux directions sont explorées pour incorporer la sémantique dans les modèles de langage. La première se base sur l'exploitation des liens entre termes tout en utilisant une même unité d'indexation. La seconde se base sur l'utilisation d'unités d'indexation plus complexes en plus de l'utilisation de termes simples. Dans ce papier est détaillée l'approche que nous proposons pour incorporer la dimension sémantique de document, et qui rentre dans le cadre de la seconde direction.*

ABSTRACT. *Most traditional information retrieval systems are based on simple terms indexing. However, they deliver massive results in response to users queries. This is partly due to the fact that semantic content of a document (or a request) can not be accurately captured by a simple set of independent keywords. Two directions are investigated to incorporate semantics in the language models. The first is based on the exploitation of terms dependency while using the same indexing unit. The second is based on the use of more complex indexing units. In this paper we detail our approach to incorporate the semantic dimension of document.*

MOTS-CLÉS : *Recherche d'information, modèles de langage, indexation sémantique.*

KEYWORDS: *Information retrieval, language model, semantic indexing.*

1. Introduction

La plupart des systèmes de recherche d'information (SRI) actuels privilégient la minimisation de temps de réponses par rapport à la qualité des documents retournés à l'utilisateur. En effet, ces derniers délivrent de grandes quantités de documents en réponse aux requêtes des utilisateurs, ce qui génère ainsi, une surcharge informationnelle dans laquelle il est difficile de distinguer l'information pertinente de l'information secondaire ou même du bruit. L'une des raisons qui a engendré ceci est la non prise en compte de toutes les caractéristiques d'un document dans le processus d'indexation et de recherche. En effet, les SRI implémentent les techniques traditionnelles de la RI, qui considèrent un document comme un ensemble de termes (sac de mots). Cependant, l'une des critiques formulées à l'utilisation des termes simples comme unité d'indexation est que le contenu d'un document ne peut pas être capturé précisément par un simple ensemble de mots clés indépendants.

Le modèle de langage offre un cadre probabilistique pour la description du processus de la RI. Parmi les propriétés qu'offre ce modèle est la combinaison des différentes représentations d'un document comme l'intégration des informations sémantiques d'un document. Deux alternatives ont été explorées pour intégrer le contenu sémantique d'un document dans ces modèles. La première se base sur l'exploitation des liens entre termes tout en utilisant une même unité d'indexation. La seconde se base sur le développement de modèles pour une représentation plus détaillée du contenu des documents et des requêtes, et cela par l'utilisation d'unités d'indexation plus complexes en plus de l'utilisation des termes simples. Notre approche s'inscrit dans cette deuxième orientation.

L'utilisation d'unités d'indexation complexes en RI nécessite la prise en compte de plusieurs paramètres à savoir : la technique d'extraction des termes composés utilisée, la prise en compte de l'adjacence et de la directionnalité des termes composants, la pondération des termes composés et l'intégration des termes composés dans le modèle de Ranking. L'apport de notre approche concerne les deux derniers paramètres.

Nous organisons ce papier comme suit : Dans la section 2 sont abordées la modélisation de langue en recherche d'information et l'intégration de la composante sémantique dans ces modèles. La section 3 est consacrée à la présentation de l'approche que nous proposons pour intégrer les informations sémantiques dans les modèles de langage. Un exemple d'illustration de l'approche proposée est donné en section 4. La dernière section fait la synthèse de cette étude.

2. Modélisation de langue et incorporation de la sémantique

2.1. *La modélisation de langue en recherche d'information*

L'approche de modélisation de langue part d'un principe différent de ceux des approches traditionnelles ; on ne tente pas de modéliser directement la notion de pertinence (à l'exception de (Lavrenko *et al.,* 2001)) ; mais on considère que la pertinence d'un document face à une requête est en rapport avec la probabilité que la requête puisse être générée par un modèle de langue d'un document. Ainsi un modèle de langue (Md) est construit pour chaque document, et le score d'un document est déterminé par la probabilité de génération de la requête sachant le modèle de ce document, P (Q | Md). Cette approche (ML) permet de combiner les deux composantes (indexation et Ranking) dans un seul modèle unifié.

La plupart des modèles de langue développés pour la RI utilisent le principe de génération de la requête par un modèle de document. Les approches de modélisation de langage pour la RI peuvent être classées en trois catégories :

– Génération de la requête par le modèle de document (Ponte *et al.,* 1998) (Hiemstra, 1998).

– Génération de document à partir du modèle de la requête : cette approche procède dans le sens inverse que la première. Ainsi, un modèle de langage de la requête est construit, ensuite les documents sont classés selon leurs probabilités que leur contenu soit généré par le modèle de la requête. Le travail de Lavrenko et Croft (Lavrenko *et al.,* 2001) s'in scrit dans cette catégorie d'approche .

– Similarité entre modèle de document et modèle de la requête : Dans cette approche, un modèle de langage est construit pour chaque document et un autre pour la requête. Un score de similarité est calculé entre ces deux modèles (Lafferty *et al.,* 2001).

Deux problèmes fondamentaux liés à l'utilisation des modèles de langage en RI sont la source de plusieurs études. Le premier problème concerne la clairsemence de données (Data Sparseness) : qui consiste à attribuer la probabilité zéro à tout document ne contenant pas tous les termes, même s'il est pertinent. Pour y remédier la technique de lissage est utilisée, et permet d'attribuer une probabilité non nulle pour les termes non observés dans le document. Cette technique (Smoothing) peut jouer divers rôles, par exemple la combinaison de multiples sources d'information sur un document (informations sémantiques). Le second problème est lié à l'utilisation de modèle de langage uni -gramme qui ne prend pas en compte la dépendance entre les termes (document, requête).

2.2. *Incorporation de la sémantique dans les modèles de langage*

Deux directions ont été explorées pour incorporer la sémantique dans les modèles de langage : la première se base sur l'exploitation des liens entre termes

(requête, document) tout en utilisant une même unité d'indexation (termes simples). La seconde se base sur le développement de modèles pour une représentation plus détaillée du contenu des documents et des requêtes, et cela par l'utilisation d'unités d'indexation plus complexes en plus de l'utilisation des termes simples.

2.2.1. *Utilisation des termes simples comme unité d'indexation*

Plusieurs travaux ont été réalisés pour incorporer les relations existantes entre unités d'indexation, selon deux approches. La première consiste à incorporer les relations entre les termes des documents ou de la requête (Gao *et al.,* 2004), les relations entre termes expriment les dépendances. Dans la seconde approche l'incorporation des relations entre les termes de la requête et les termes des documents est réalisée soit : en modifiant la représentation (modèle) de la requête en ajoutant les termes liés, cette opération est nommée expansion du modèle de la requête (Bai *et al.,* 2005) (Lafferty *et al.,* 2001) (Wei *et al.,* 2007) ; soit en modifiant la représentation de document, en donnant une probabilité plus grande aux termes liés aux termes de document que pour les autres termes. Cette opération est nommée expansion de modèle de document (Berger *et al.,* 1999) (Cao *et al.,* 2005) (Cao *et al.,* 2007) (Liu *et al.,* 2004) (Tao Tao *et al.,* 2006) (Wei *et al.,* 2006).

2.2.2. *Utilisation de termes composés et de termes simples comme unité d'indexation*

Pour incorporer la sémantique dans les modèles de langage, une autre piste peu explorée consiste à utiliser des unités d'indexation plus complexes en plus de termes simples ; ces unités d'indexation peuvent être vues comme des concepts, plusieurs vocables sont utilisés (phrase, N-gramme, collocation, termes composés) ; ce qui améliore la représentation des documents et des requêtes. L'intérêt d'utiliser des termes composés comme unité d'indexation est que les termes composés sont moins ambigus et plus précis que les termes simples. Par exemple le terme « Java » est ambigu, par contre les termes composés « Ile de java » et « Langage java » sont non ambigus. Et le terme « Voiture Electrique » est plus spécifique que les termes « Voiture » et « Electrique » pris isolement.

Les termes composés permettent de construire des unités d'indexation non ambiguës et plus précises et peuvent par conséquent améliorer la précision de la RI. Miller et al (Miller *et al.,* 1999) ont proposé d'intégrer les bi-grammes dans leur modèle initial. Song et Croft (Song *et al.,* 1999) ont proposé un modèle de langage qui combine le modèle bi-gramme et le modèle uni-gramme en utilisant l'interpolation linéaire. Srikanth et Srihari (Srikanth *et al.,* 2002) ont développé un modèle dit bi-terme (modèle bi-gramme dans l'ordre est ignoré). Jiang et al (Jiang *et al.,* 2004) ont proposé un modèle de langage pour incorporer des phrases (deux termes) en utilisant la méthode de lissage Backoff. Alvarez et al (Alvarez *et al.,* 2004) ont proposé l'incorporation des termes composés sans contraintes d'adjacence ou d'ordre.

Les techniques qui permettent l'identification des termes composées sont scindées en trois catégories : linguistiques, statistiques et mixtes. Les techniques statistiques auxquelles nous nous intéressons sont basées sur des informations tirées de corpus d'où leur flexibilité et leur portabilité (ie : elles ne dépendent ni de la langue du corpus ni du domaine traité par le corpus). Dans les techniques statistiques les termes composés sont extraits en se basant soit sur leurs fréquences observées dans le corpus soit en utilisant des mesures d'association qui déterminent le degré d'association entre les termes composants. Les mesures d'association permettent de calculer « un score d'association » pour chaque paire de termes candidat dans le corpus ; ce score indique le potentiel de ce candidat d'être reconnu comme un terme composé. Plusieurs mesures d'association ont été proposés dans la littérature telles que : l'Information Mutuelle, le coefficient de Dice, X2 score, etc. (Liu *et al.*, 2004).

3. Approche proposée

Notre objectif est de représenter au mieux le contenu sémantique des documents. Nous réalisons cela sous le cadre des modèles de langage. Comme nous l'avons vu auparavant, deux alternatives ont été explorées pour intégrer les informations sémantiques dans les modèles de langage.

La première consiste à exploiter les liens entre les termes des documents, requêtes (requête → document). Dans le cas de l'expansion de la requête l'exploitation des liens entre termes est réalisée au moment de la recherche, ce qui affecte négativement les temps de réponses, de plus l'efficacité d'un système de recherche d'information dépend fortement du nombre de termes de la requête. Dans le cas de l'expansion du modèle du document des inconvénients surgissent selon l'approche adoptée.

L'autre alternative qui consiste à intégrer la sémantique dans les modèles de langage est l'utilisation d'unités d'indexation plus complexes à côté d'unités d'indexation simples. Notre approche s'inscrit dans cette optique. Six paramètres sont généralement à considérer dans l'utilisation des termes composés comme unité d'indexation à coté de l'utilisation des termes simples à savoir : la technique d'extraction des termes composés utilisée, l'adjacence et la directionnalité des termes composants, la taille des termes composés, la pondération des termes composés et la manière d'intégrer ces termes dans le modèle de Ranking.

Nous présentons ci-dessous notre approche pour incorporer les informations sémantiques dans le modèle de langage en définissant les caractéristiques mises en jeu dans l'indexation par des termes composés. Plus particulièrement on décrit en détail la formule de pondération des termes composés proposée et l'intégration des termes composés dans le modèle de langage. Pour les autres caractéristiques : (identification des termes composés, directionnalité, taille des termes composés et distance entre termes composants) nous adoptons des solutions existantes.

– *Identification des termes composés.* Nous optons pour l'approche statistique pour l'extraction des termes composés du fait que nous voulons élaborer une approche plus générale qui ne dépend pas de la langue ou du domaine du corpus. Pour l'identification des termes composés nous utilisons la mesure d'Information Mutuelle (PMI) basée sur l'étude menée par Petrovic et al (Sasa *et al.*, 2006) qui ont montré que cette mesure donne de meilleurs résultats que ceux obtenus par les autres mesures telles que le coefficient de Dice ou la mesure Chi-square.

– *Directionnalité.* Srikanth et Srihari (Srikanth *et al.*, 2002) ont montré que la prise en compte de la directionnalité (cas de deux termes) est plus précise pour la RI. En se basant sur ce constat nous adoptons la directionnalité des termes composés dans notre approche.

– *Taille des termes composés.* En principe les termes composés peuvent être de n'importe qu'elle longueur (supérieure ou égale à 2). Dans notre cas nous limitons la taille des termes composés à deux qui est une pratique commune, et scalable pour de grandes collections hétérogènes.

– *Distance entre termes composants.* La cooccurrence des termes est une source d'information importante et efficace pour la désambiguïsation des termes (Agirre *et al.*, 2001). Nous traitons cette relation de cooccurrence en terme d'extraction des termes composés dont les termes composants sont adjacents, exemple : « Génie Logiciel », « Microsoft Word » seront utilisées comme unités d'indexation car les termes composés sont moins ambiguës et plus précis que les termes qui les composent, de ce fait le contenu sémantique des documents et des requêtes est plus précis. Un deuxième type de relation de cooccurrence consiste en l'extraction des termes de voisinage d'un terme donné. Exemple : docteur (hôpital, infirmière, aide soignant) peut être utilisé dans une étape ultérieure (expansion de la requête).

– *Pondération des termes composés.* La pondération des termes composés est un problème non résolu en RI. En effet il n'existe pas de schéma bien accepté pour la pondération des termes composés. Des alternatives ont été proposées ; parmi elles l'adaptation de schéma bien connu de pondération de termes simples TF-IDF. Cependant, les schémas de pondération proposés dans (Baziz *et al.*, 2005) et (Liu *et al.*, 2004) ne tiennent pas compte d'un facteur important qui est l'importance des termes composants dans le terme composé ; dans les schémas précédents cette importance est considérée identique. Or, dans la réalité un des termes composants peut être plus important que les autres termes. Exemple : le terme « ordinateur » est plus important que le terme « personnel » dans le terme composé « ordinateur personnel ». Cette dominance de terme est déterminée par la spécificité du terme, cette dominance est généralement supposée qu'elle est en corrélation avec l'IDF du terme. Ainsi, nous proposons d'exprimer l'importance d'un terme composant « t » dans un terme composé « tc » de la manière suivante :

$$imp(t / tc) = \frac{idf(t)}{\sum_{t_i \in t_c} idf(t_i)} \quad [1]$$

En supposant que l'auteur d'un document utilise les termes composants isolément pour exprimer le terme composé comme abréviation après un nombre d'occurrences de terme composé. Par exemple un document contenant le terme composé « énergie électrique », l'auteur utilise le terme « énergie » simplement pour désigner le terme composé « énergie électrique ». Néanmoins, un problème surgi lorsqu'un terme composant est partagé par deux voire plusieurs termes composés, dans ce cas il faut trouver le terme composé auquel renvoie le terme simple. Par exemple : si on a le terme « énergie fossile » dans le document précédent. Alors il faut choisir à quel terme composé le terme « énergie » renvoie. Nous proposons d'utiliser un facteur qui combine l'importance du terme composant dans les termes composés et la fréquence des termes composés dans le document pour désigner le terme composé adéquat. Le terme composé qui maximise ce facteur est choisi.

Termes / Facteurs	Energie électrique	Energie fossile
Importance (énergie)	0.6	0.1
Nombre d'occurrences	20	12
Produit	12	1.2

Table 1. *Exemple*

Dans cet exemple on voit bien que le terme composé « énergie électrique » maximise le produit des deux facteurs : importance du terme « énergie » dans le terme composé et la fréquence du terme composé dans le document, par conséquent le terme « énergie » utilisé isolement dans le document renvoie au terme composé « énergie électrique ». La fréquence d'un terme composé « tc » dans un document dépend du nombre d'occurrences de ce terme dans le document et du nombre d'occurrences des termes composants pour lesquels le terme composé maximise le facteur discuté auparavant. Formellement elle est exprimée ainsi :

$$F(tc) = nbr(tc) + \sum_{i=1}^{2} imp(t_i / t_c) \times nbr(t_i) \quad \text{si}$$

$$imp(t_i / t_c) \times nbr(tc) = \max_{t_i \in t_{c'}} ((imp(t_i / t_{c'}) \times nbr(tc')) \qquad [2]$$

Tel que « tc' » est l'ensemble des termes composés qui contient le terme « ti », *F(tc)* représente la fréquence du terme composé « tc », *nbr(tc)* est le nombre d'occurrences du terme composé « tc », $imp(t_i/t_c)$ est l'importance du terme « ti » dans le terme composé « tc » et *nbr(ti)* est le nombre d'occurrence du terme « ti ». Exemple : supposant que le nombre d'occurrences de terme « énergie » dans l'exemple précédent est 10. Alors la fréquence du terme composé « énergie électrique » est : F ('énergie électrique') = nbr('énergie électrique') + imp('énergie'/ 'énergie électrique') * nbr('énergie') = 20 + 0.6*10 = 26.

– *Intégration des termes composés dans le modèle de langage.* Dans notre approche nous considérons qu'un document « D » de la collection « C » est représenté de deux façons différentes : la première est la représentation par des termes simples notée « Dt », la seconde est la représentation par des termes composés notée « Dtc » ; ici la notion d'un terme composé se réfère à tous les termes de longueur « un » ou « deux », les termes de longueur « un » sont les termes qui ne participent pas a la composition des termes de taille « deux ». Le même raisonnement est appliqué pour la collection, ainsi la collection est représentée par deux représentations : une représentation avec des termes simples « Ct » qui est obtenue par la concaténation des représentations par des termes simples des documents formant la collection, et une représentation avec des termes composés « Ctc » qui est obtenue par la concaténation des représentations par des termes composés des documents formant la collection. A partir de ces deux représentations de document « Dt » et « Dtc » nous proposons deux modèles de document exprimés ainsi :

– Pour la représentation avec des termes simples le modèle obtenu est le modèle uni-gramme, en interpolant le modèle de document avec celui de la collection, exprimé ainsi :

$$P(t / M_{Dt}) = \lambda P_{ML}(t / M_{Dt}) + (1-\lambda) P_{ML}(t / C_t) \qquad [3]$$

Tel que : $P_{ML}(t / M_{Dt}) = \frac{tf(t,D)}{|D|}$ et $P_{ML}(t / C_t) = \frac{df(t)}{\sum_{t_i \in C_t} df(t_i)}$

Où ? est un facteur d'interpolation compris entre 0 et 1, *tf(t, D)* est la fréquence de terme « t » dans le document D, *df(t)* est le nombre de documents contenant le terme « t » et Ct est le vocabulaire d'indexes (termes simples).
Les documents de la collection vis-à-vis d'une requête « Q » sont alors ordonnés en utilisant la formule suivante :

$$P(Q / M_{Dt}) = \prod_{t_i \in Q} P(t_i / M_{Dt})$$

– Pour la représentation avec des termes composés nous exprimons le modèle ainsi :

$$P(t / M_{Dtc}) = \sum_{t_c} P(t / t_c) \times P_{ML}(t_c / M_{Dtc}) \qquad [4]$$

Comme nous l'avons noté auparavant un terme composant peut être partagé par deux voire plusieurs termes composés, dans ce cas il faut trouver le terme composé auquel renvoie le terme simple. Nous avons proposé un facteur qui permet de déterminer le terme composé concerné. Notons « t_{cmax} » ce terme composé. Ainsi la probabilité suivante devient alors :

$\sum_{t_c} P(t / t_c) = imp(t / t_{c\max})$ calculée avec la formule [1]

En remplaçant cette probabilité dans la formule [4] celle-ci devient :

$$P(t / M_{Dtc}) = imp(t / t_{c\max}) \times P(t_{c\max} / M_{Dtc}) \quad [5]$$

Et $$P(t_c / M_{Dtc}) = \lambda P_{ML}(t_c / M_{Dtc}) + (1-\lambda) P_{ML}(t_c / C_{tc}) \quad [6]$$

Est obtenue en interpolant le modèle de document avec celui de la collection, *PML (tc/MDc)* est estimée ainsi :

$$PML(tc / MDc) = \frac{\mathrm{tf}(\mathrm{t_c})}{\sum_{\mathrm{t_{ci}}} \mathrm{tf}(\mathrm{t_{ci}})}$$

Tel que $tf(t_c)$ est la fréquence du terme composé « tc » dans le document D, calculée selon la formule [2]. Et *PML (tc/Ctc)* est calculée ainsi :

$$PML(tc / Ctc) = \frac{df(t_c)}{\sum_{t_{ci} \in C_{tc}} df(t_{ci})}$$

Où $df(t_c)$ est le nombre de documents contenant le terme composé « tc » et Ctc est le vocabulaire d'indexes (termes composés). Les documents de la collection vis-à-vis d'une requête « Q » sont alors classés en utilisant la formule suivante :

$$P(Q / M_{Dtc}) = \prod_{t_i \in Q} \left(\sum_{t_i \in t_{c\max}; t_{c\max} \in D} imp(t_i / t_{c\max}) \times P_{ML}(t_{c\max} / M_{Dtc}) \right)$$

Les documents sont représentés par deux représentations différentes : avec des termes simples et avec des termes composés. Le modèle de Ranking doit combiner ces deux représentations. Le modèle obtenu est donné ainsi :

$$P(t / M_D) = \alpha P(t / M_{Dt}) + (1-\alpha) P(t / M_{Dtc})$$

$$P(t / M_D) = \alpha [\lambda P_{ML}(t / M_{Dt}) + (1-\lambda) P_{ML}(t / C_t)] + (1-\alpha)$$

$$[\lambda P_{ML}(t / M_{Dtc}) + (1-\lambda) P_{ML}(t / C_{tc})]$$

$$P(t / M_D) = \lambda [\alpha P_{ML}(t / M_{Dt}) + (1-\alpha) P_{ML}(t / M_{Dtc})] +$$

$$(1-\lambda)[\alpha P_{ML}(t / C_t) + (1-\alpha) P_{ML}(t / C_{tc})]$$

4. Exemple d'illustration

Dans l'exemple suivant « Ct » est le vocabulaire d'indexes (termes simples), « Ctc » est le vocabulaire d'indexes (termes composés), « C » est l'ensemble de documents de la collection et Q1, Q2, Q3, Q4 des requêtes.

Ct = {m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11, m12, m13, m14, m15}

Ctc = {m1, m2, m3, m6, m8, m9, m11, m13, m14, m15, m1m2, m1m10, m5m6, m12m4, m7m8, m12m2}
C = {d1, d2, d3, d4}; Q1= m1m2 ; Q2= m5m6 ; Q3= m1 ; Q4= m12m4 ;
Dt1={m1(5), m2(2), m5(6), m6(3), m8(2), m10(1), m15(1)};
Dtc1={m1m2(3.58), m1m10(1), m5m6(4.83) , m8(2), m15(1)}
Dt2={m2(1), m3(4), m4(2), m6(3), m8(1), m12(4)};
Dtc2={m2(1), m3(4), m6(3), m8(1), m12m4 (2.77)};
Dt3={m1(2), m4(3), m9(3), m12(5)}; Dtc3={m1(5), m2(3), m9(3), m12m4 (2.77)}
Dt4={m2(3), m5(6), m6(5), m7(6), m8(2)};
Dtc4={m2(3), m5(6), m6(5), m7m8(5.43)}
Dt5={m1(5), m11(8), m14(3), m15(2)}; Dtc5={m1(5), m11(8), m14(3), m15(2)}
Dt6={m2(3), m8(2), m9(1), m11(7) m12(2), m13(2)};
Dtc6={m8(2), m9(1), m11(7), m13(2) , m2m12(2.208)}

	d1			d2		
	scts	sctc	c(scts,sctc)	scts	sctc	c(scts,sctc)
Q1=m1m2	**0,022299**	**0,21330**	0,11801	0,00240	0,00206	0,00223
Q2=m5m6	**0,02992**	**0,28338**	0,15686	0,00293	0,00222	0,00258
Q3=m1	0,20147	0,16867	0,18506	0,02647	0,02222	0,02434
Q4=m12m4	0,00046	0	0,00023	0,02365	0,18713	0,10539
	d3			d4		
	scts	sctc	c(scts,sctc)	scts	sctc	c(scts,sctc)
Q1=m1m2	**0,03895**	**0,05134**	0,04514	0,00369	0,00314	0,00341
Q2=m5m6	0,00047	0,00025	0,00035	**0,03869**	**0,04598**	0,04234
Q3=m1	0,23235	0,27633	0,25434	0,02647	0,02222	0,02434
Q4=m12m4	0,01912	0,16319	0,09115	0,00046	0	0,00023
	d5			d6		
	scts	sctc	c(scts,sctc)	scts	sctc	c(scts,sctc)
Q1=m1m2	0,00974	0,00722	0,00848	0,00358	0,00074	0,00216
Q2=m5m6	0,00046	0	0,00023	0,00031	0	0,00015
Q3=m1	0,22091	0,21667	0,21879	0,02647	0,02222	0,02434
Q4=m12m4	0,00046	0	0,00023	0,00154	0	0,00077

Table 2. *Représentation des scores des documents.*

On peut noter les remarques suivantes à partir de la table 2 dans laquelle scts, sctc et c(scts,sctc) désignent respectivement le score d'un document avec des termes simples, le score d'un document avec des termes composés et la combinaison des deux scores précédents :

– Avec la requête Q1= « m1m2 » on note que pour le document « d1 » qui contient le terme composé « m1m2 » son score passe de **0,022299** avec des termes simples à **0,21330** avec des termes composés. Par contre pour le document « d3 » qui contient les termes m1 et m2 séparément son score passe de **0,03895** avec des

termes simples à **0,051334** avec des termes composés. Cela montre que le document « d1 » est plus approprié pour la requête Q1= « m1m2 » que le document « d3 » car le premier contient le terme composé, par contre le second contient les termes composants séparément.

– Avec la requête Q2= « m5m6 », on remarque que pour le document « d1 » qui contient le terme composé « m5m6 » son score passe de **0,029922** avec des termes simples à **0,28338** avec des termes composés. Par contre pour le document « d4 » qui contient les termes m5 et m6 séparément son score passe de **0,038669** avec des termes simples à **0,045598** avec des termes composés. Cela montre que le document « d1 » est plus approprié pour la requête Q2= « m5m6 » que le document « d4 » car le premier contient le terme composé, par contre le second contient les termes composants séparément. Cela montre que la représentation avec des termes composés influe considérablement sur les scores des documents en réponse aux requêtes contenant des termes composés.

– Avec la requête Q3= « m1 », on note que le changement des scores obtenus par des termes simples par rapport aux scores obtenus par des termes composés est presque identique pour tous les documents. Cela montre que la représentation avec des termes composés n'influe pas sur les scores des documents en réponse aux requêtes contenant uniquement des termes simples.

5. Conclusion

Nous avons exposé dans cet article une approche qui permet de représenter au mieux le contenu sémantique d'un document. Et cela par l'utilisation des termes composés comme unité d'indexation à coté des termes simples. Pour cela nous avons défini les six caractéristiques mises en jeu dans l'indexation par des termes composés, dans le cadre de modèle de langage. L'exemple d'illustration que nous avons présenté indique que l'approche répond bien à l'objectif fixé . La prochaine étape consiste à mettre en œuvre cette approche, la tester et la comparer aux autres approches (ex : bi-gramme).

6. Bibliographie

Agirre E. et Martinez D. *Knowledge sources for word sense disambiguation* 2001.

Alvarez C., Langlais P. et Nie J-Y. « Word pairs in language modeling for information retrieval », *Proc. of the conference on computer assisted information retrieval*, 2004

Bai, J. et al. « Query expansion using term relationships in language models for information retrieval », *CIKM*, 2005, p. 688-695.

Baziz M. et al. « Semantic cores for representing documents », *20th ACM symposium on applied computing, SAC'2005*, Santa Fe, New Mexico, USA, 13 - 17 mars 2005. ACM-SAC: Press, New York, NY, USA, 2005, p. 1011 - 1017.

Berger A. et Lafferty J. « Information retrieval as statistical translation », *Proc. of 1999 ACM SIGIR conference on research and development in IR*, 1999, p. 222-229.

Cao G., Gao J. F. et Nie J. Y. *Extending query translation to cross-language query expansion with Markov chain.* 2007.

Cao, G., Nie, J.Y. et Bai, J. « Integrating word relationships into language models », *Proc. of 17th ACM SIGIR conference*, 2005, p. 298–305.

Gao J. F.et al. « Dependence language model for information retrieval », *Proc. of 27th ACM SIGIR conference on research and development in IR*, 2004.

Hiemstra D. « A linguistically motivated probabilistic model of information retrieval », *Second european conference, ECDL'98*, Nicolaou C. et Stephanides C. (Eds.), Research and advanced technology for digital libraries, Springer Verlag, 1998.

Jiang M., Jensen E. et Beitzel S. « Effective use of phrases in language modeling to improve information retrieval », *Symposium on AI & math*, Special session on intelligent text processing, Florida, January 2004.

Lafferty J. et Zhai, C. « Document language models, query models, and risk minimization for information retrieval », *Proc. of 24th annual international ACM-SIGIR conference on research and development in information retrieval* , 2001, p.111-119.

Lavrenko V. et Croft W. B. « Relevance-based language models », *Proc. of 24th annual international ACM-SIGIR conference on research and development in IR*, Croft W.B. et al. (Eds.), New Orleans, Louisiana, 2001, p.120-127.

Liu, X. et Croft, W. B. « Cluster-based retrieval using language models », *Proc. of 27th ACM SIGIR*, 2004, p. 186-193.

Miller D. R. H., Leek T. et Schwartz R. M. « A hidden Markov model information retrieval system », Hearst *et al*. (Eds.), 1999, p. 214–221.

Ponte J.M. et Croft W. B. « A language modeling approach to information retrieval », Croft *et al*. (Eds.), 1998, p. 275–281.

Sasa P. *et al.* « Comparison of collocation extraction measures for document indexing », *Journal of Computing and Information Technology*, vol. 14, n°4, 2006, p. 321–327.

Song F. et Croft W. B. « A general language model for information retrieval » *Proc. of SIGIR'99*, 1999.

Srikanth M.et Srihari R. « Biterm language models for document retrieval », *Proc. of 25th annual international ACM SIGIR*, Finland, 2002, p. 425–426.

Tao Tao, *et al.* « Language model information retrieval with document expansion », *Proc. of the human language technology conference of the north American chapter of the ACL*, New York, 2006, p. 407–414.

Wei, X. et Croft W. B. « LDA-based document models for ad-hoc retrieval », *Proc. of 29th annual international ACM SIGIR conference on research and development on IR*, 2006.

Wei, X. et Croft, W. B. « Investigating retrieval performance with manually-built topic models » Proc. *of RIAO 2007 - 8th conference large scale semantic access to content (text, image, video and sound)*, paper number 12, 2007.