Pseudo-réinjection de pertinence basée sur un modèle de langue mixte combinant les termes simples et composés

Arezki Hammache[(1)], Mohand Boughanem[(2)], Rachid Ahmed-Ouamer[(1)]

[(1)]*Laboratoire LARI, Département d'informatique, Université Mouloud Mammeri 15000 Tizi-Ouzou, Algérie*

{arezki20002002,ahm_r}@yahoo.fr

[(2)] *Laboratoire IRIT, Université Paul Sabatier*
118 route de Narbonne 31062 Toulouse Cedex 09, France.

bougha@irit.fr

RÉSUMÉ. *Dans cet article nous présentons une nouvelle technique de reformulation de requête. Cette technique considère la requête comme un ensemble de termes composés et un ensemble de termes simples. Pour déterminer les termes d'expansion on additionne les poids des relations d'un terme candidat avec chacun des termes de la requête (simple, composé). Un terme candidat est choisi s'il est fortement en relation avec la plupart des termes de la requête. Cette technique est modélisée dans le cadre de modèle de langue. Les tests effectués sur deux collections TREC ont montré des améliorations par rapport à deux modèles: le modèle uni-gramme et le modèle d'expansion de requêtes basé sur la mesure Kullback-Leibler Divergence (KLD).*

ABSTRACT. *In this paper we present a new technique for query expansion. This technique considers the query as a set of compound terms and a set of simple terms. To determine the expansion terms we add the weight of a term relationships with each of the candidate query terms (simple, compound). A candidate term is selected if it is strongly related with most query terms. This technique is modeled in the context of the language model. Tests on two TREC collections showed improvements compared to the uni-gram model and KLD expansion model.*

MOTS-CLÉS : *Expansion de requête, Termes composés, Modèle de langue, cooccurrence.*

KEYWORDS: *Query expansion, Compound terms, Language model, Co-occurrence.*

1. Introduction

L'expansion de requêtes est la méthode la plus utilisée pour pallier au problème de disparité des termes (term mismatch) en Recherche d'Information. Les techniques d'expansion de requêtes peuvent être classées en tenant en compte de plusieurs paramètres (Carpineto *et al.*, 2012) : La sélection des termes d'expansion en considérant chaque terme de la requête individuellement, ou la requête dans son ensemble ; La représentation de la requête (document) comme un ensemble de terme simple (sac de mots) ou une représentation prenant en compte les relations entre termes ; le type des termes d'expansion sélectionnés (simples, composés); La méthode de sélection des termes d'expansion utilisée (la relation de cooccurrence, les mesures d'information, les techniques de classification) ; et les sources de données utilisées pour l'expansion de la requête.

Différentes sources de données sont utilisées pour sélectionner les termes d'expansion. Elles peuvent être (Manning *et al.*, 2008) : (1) Des sources externes telles que les ontologies, les thésaurus et la relation de cooccurrence entre termes. Les méthodes basées sur ces sources sont dites méthodes globales. (2) Les documents retournés par la première recherche. Les méthodes basées sur ces sources sont dites méthodes locales. Ces méthodes sont également connues sous le nom de réinjection de pertinence. La réinjection de pertinence a montré son efficacité avec différents modèles de la RI (Rocchio, 1971) (Salton *et al.*, 1990) (Robertson *et al.*, 1994) (Lavrenko *et al.*, 2001) (Lv *et al.*, 2009a) et a démontré de meilleurs résultats que les méthodes globales.

Afin de sélectionner les termes d'expansion dans les sources de données plusieurs méthodes ont été utilisées, Parmi elles la relation de cooccurrence. Cependant, les améliorations obtenues en utilisant cette relation restent minimes. Qiu et Frei ont indiqué dans (Qiu *et al.*, 1993) que l'une des raisons expliquant ces faibles améliorations dans les études antérieures est la suivante: «l'utilisation de la relation de cooccurrence d'une manière individuelle (indépendante) par les termes de la requête originale peut introduire beaucoup de bruit (inclusion de termes qui ne sont pas du contexte de la requête) ». Pour remédier à ce problème, ils ont proposé de déterminer les termes d'expansion en additionnant les poids des relations d'un terme candidat avec chacun des termes de la requête. Un terme candidat est alors choisi s'il est fortement en cooccurrence avec la plupart des termes de la requête.

Un des facteurs important de succès de l'expansion de requêtes basée sur la réinjection de pertinence est la qualité des documents obtenus lors de la première recherche. Plusieurs travaux ont montré que les modèles de recherche qui vont au-delà de l'hypothèse d'indépendance entre termes peuvent estimer la pertinence d'un document avec plus de précision (Song *et al.*, 1999) (Srikanth *et al.*, 2002) (Metzler *et al.*, 2005) (Lv *et al.*, 2009b) (Hammache *et al.*, 2011). Cependant, il existe relativement peu de travaux où l'expansion de la requête est réalisée en se basant sur ce type de modèles (Metzler *et al.*, 2007) (Lv *et al.*, 2010) (Miao *et al.*, 2012). Ces travaux ont rapporté des améliorations significatives. Dans l'optique de ces travaux,

nous proposons dans cet article une nouvelle approche pour l'expansion de la requête dans le contexte de modèle de langue; elle est caractérisée par les points suivants:

Premièrement, notre approche d'expansion de requêtes est basée sur un modèle de langue mixte, qui va au-delà du la représentation en sac de mots, tels que les modèles présentés dans (Song *et al.*, 1999) (Srikanth *et al.*, 2002) (Metzler *et al.*, 2005) (Lv *et al.*, 2009b) . Dans ces modèles toutes les dépendances adjacentes entre termes (bi-grammes) sont prises en considération et combinées avec le modèle uni-gramme. Cependant, seules quelques dépendances sont utiles. Dans notre modèle, nous ne considérons que les bi-grammes pertinents, appelés «termes composés» définis comme «une expression composée de deux ou plusieurs termes qui correspondent à une certaine manière conventionnelle d'exprimer des choses» (Manning *et al.*, 2003). « *Moteur de recherche* » et « *jeux olympiques* » sont par exemple de termes composés. De plus, nous introduisons une nouvelle méthode de pondération des termes d'expansion.

Deuxièmement, notre approche utilise la relation de cooccurrence entre les termes pour extraire les termes d'expansion comme cela est réalisé dans (Qiu *et al.*, 1993). Cependant, notre approche d'expansion de la requête est basée sur la réinjection de pertinence. De plus, elle est formalisée dans le cadre de modèle de langue.

Enfin, dans les méthodes antérieures d'expansion de requête, le type du terme d'expansion utilisé est généralement le terme simple. Cependant, l'utilisation d'unité composée à la phase de recherche initiale a montré son utilité pour la RI. Ainsi, nous explorons dans ce travail, l'utilisation d'unités composées lors de la phase d'expansion de la requête.

Nous avons évalué notre méthode sur deux collections TREC et nous l'avons comparé aux approches traditionnelles. Les résultats expérimentaux montrent que les performances de la RI ont été améliorées de façon significative.

Le reste de ce papier est organisé comme suit: la section 2 présente les travaux connexes. La section 3 présente notre méthode d'expansion de la requête basée sur le modèle de langue mixte combinant les termes simples et composés et nous détaillons la façon dont les termes d'expansion sont extraits et pondérés. Nous rapportons les résultats expérimentaux dans la section 4. Enfin, dans la section 5, nous concluons notre travail et énumérons quelques perspectives.

2. Etat de l'art

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

Le modèle de langue est un nouveau cadre probabiliste pour la description du processus de la RI (Ponte *et al.*, 1998). Les résultats obtenus avec ce modèle ont

montré des performances équivalentes voire supérieures à celles des modèles classiques (vectoriel, probabiliste) (Zhai *et al.*, 2004). La formule proposée pour calculer le score d'un document D est basée sur sa probabilité de générer la requête Q. Elle est exprimée comme suit:

$$P(Q|D) = \prod_{w \in Q} P(w|D)$$

Cette formule est généralement utilisée pour l'estimation de modèle du document $P(w|D)$. Afin d'intégrer d'autres informations pour l'estimation de modèle de la requête Lafferty et al (Lafferty *et al.*, 2001) ont proposé une autre implémentation de la fonction de calcul de score. Elle est basée sur la mesure de divergence de Kullback-Leibler (KLD : Kullback-Leibler Divergence) entre le modèle de la requête et celui de document. La formule d'ordonnancement est exprimée comme suit :

$$Score(Q, D) = \sum_{w \in V} P(w|Q) log P(w|D)$$

Où V est l'ensemble de vocabulaire de la collection.

Généralement, l'Estimation de Maximum de vraisemblance (ML : Maximum Likelihood) est utilisée pour estimer les deux modèles (document «$P_{ML}(w|D)$» et requête «$P_{ML}(w|Q)$ »). On obtient alors la formule suivante :

$$Score(Q, D) = \sum_{w \in V} P_{ML}(w|Q) log P_{ML}(w|D)$$

L'Estimation ML est basée sur l'hypothèse d'indépendance entre termes (document, requête) qui est une simplification facilitant grandement le calcul. Néanmoins, elle ne reflète pas la réalité, car les termes dans les documents et/ou la requête sont liés. Pour estimer d'une manière précise le modèle de langue de la requête et/ou de document, on incorpore les relations entre termes dans les modèles de langage (Gao *et al.*, 2004) (Maisonnasse *et al.*, 2008).

2.2. *Incorporation des relations entre termes en RI*

Les modèles uni-grammes sont basés sur l'hypothèse de l'indépendance entre termes. Compte tenu de la connaissance commune sur la langue, une telle hypothèse peut sembler irréaliste. Plusieurs modèles ont été proposés pour étendre le modèle uni-gramme. En particulier, dans le contexte de l'approche de modélisation de langage en RI, deux principales directions ont été investies.

La première se base sur l'utilisation de la proximité entre termes. Les fonctions de proximité capturent la mesure dans laquelle les termes de recherche apparaissent proches les uns des autres dans un document. Ces fonctions sont alors utilisées comme un facteur supplémentaire pour le classement des documents (Lv *et al.*, 2009b).

La seconde se base sur l'utilisation d'unités plus complexes, tels que des syntagmes ou des termes composés. Particulièrement, elle suppose que la requête (document) est composée de plusieurs unités de termes (n-grammes, termes

composés) et utilise les occurrences des unités dans le document pour l'appariement (Song *et al.,* 1999) (Srikanth *et al.,* 2002) (Metzler *et al.,* 2005). Cependant, dans ces modèles toutes les dépendances entre termes adjacents (bi-grammes) sont considérées et combinées avec le modèle uni-gramme. Néanmoins, seules quelques dépendances sont utiles, en d'autres termes la prise en compte de tous les bi-grammes peut introduire du bruit. Dans (Hammache *et al.,* 2011) un modèle de langue mixte a été proposé, qui ne considère que les bi-grammes pertinents, ce modèle a montré des améliorations par rapport au modèle développé dans (Metzler et al., 2005).

En résumé, tous les modèles, intégrant les relations entre termes, ont montré que la représentation (modèle) de document peut être estimée avec plus de précision. Cependant, il y a eu relativement peu de travaux réalisés dans le cadre de la réinjection de pertinence basée sur de tels modèles.

2.3. *La pseudo-réinjection de pertinence*

Dans cette section, nous discutons des approches couramment utilisées pour l'expansion de la requête dans le cadre du modèle de langue:

De nombreuses techniques et algorithmes de réinjection de pertinence ont été développés ces dernières décennies, principalement inspirés du travail de Rocchio (Rocchio, 1971). Carpineto et al (Carpineto *et al.,* 2001) ont proposé une méthode basée sur la mesure KLD entre les distributions de probabilité des termes dans les documents les mieux classés et dans la collection, pour pondérer les termes d'expansion.

Dans le cadre de modèle de langue, la réinjection de pertinence a été principalement appliquée pour (ré-) estimer le modèle de la requête. Zhai et al (Zhai *et al.,* 2001) ont proposé un modèle nommé model-based feedback, où le nouveau modèle de la requête est obtenu par l'interpolation de modèle original de la requête avec un modèle de requête obtenu en utilisant les documents jugés pertinents (retour de pertinence). Dans la même optique, Laveranko et al (Laveranko *et al.,* 2001) ont proposé un modèle de langue de pertinence. Cependant, ces modèles ont tendance à ignorer une évidence importante qui est la dépendance entre termes.

D'autres travaux ont exploité la dépendance (proximité) entre termes pour construire une bonne représentation (modèle) de la requête. Lv et al (Lv *et al.,* 2009b) ont proposé le modèle Positional Language Model (PLM), afin de prendre en compte la notion de proximité des termes de la requête dans les documents. Dans ce modèle, un modèle de langue pour chaque position (PLM) est défini, ensuite le score de document est calculé sur la base de l'ensemble des scores de ses PLMs. Ce modèle est utilisé dans un travail récent (Lv *et al.,* 2010) pour étendre le modèle de pertinence (Laveranko *et al.,* 2001). Et cela en surpondérant les termes candidats qui sont à proximité des termes de la requête.

Metzler et al (Metzler *et al.*, 2007) ont développé un modèle d'expansion de la requête nommé LCE (Latent Concept Expansion), ce dernier est basé sur le modèle MRF (Metzler *et al.*, 2005). Le modèle MRF permet de modéliser les relations entre termes de la requête, différentes variantes de relations ont été utilisées. Le modèle LCE permet de retrouver les concepts non exprimés par la requête utilisateur (les concepts latents), et cela en se basant sur leurs cooccurrences avec les concepts explicitement exprimés dans la requête originale, dans les documents pertinents ou pseudo-pertinents.

Contrairement à (Metzler *et al.*, 2007), nous proposons dans cet article une méthode d'expansion de requêtes basée sur un modèle de langue mixte, qui ne considère que les bi-grammes pertinents.

Pour confirmer l'efficacité de notre méthode, nous comparons ses performances avec celles du modèle de réinjection KLD (Carpineto *et al.*, 2001) dans la section 4. Les résultats expérimentaux montrent que notre modèle surpasse le modèle KLD.

3. Approche proposée

Dans cette section, nous présentons notre approche d'expansion de requête basée sur un modèle de langue mixte combinant les termes simples et composés. Premièrement, nous introduisons brièvement le modèle de langue mixte. Ensuite, nous décrivons en détail l'estimation de modèle de la requête.

3.1 *Le modèle de langue mixte*

Dans ce modèle, Nous considérons une requête Q représentée dans le vocabulaire $V = \{T_1, \dots, T_m\} \cup \{t_1, \dots, t_n\}$ contenant des termes simples t et des termes composés T. Le terme composé T est un n-gramme qui apparait fréquemment dans la collection, il est formé par deux ou plusieurs termes simples adjacents non vides. Nous décrivons en section 4.1 la procédure d'extraction des termes composés.

Pour le calcul de score d'un document D vis-à-vis d'une requête Q on utilise la mesure de divergence de Kullback-Leibler (KLD). Nous supposons que le modèle de document $P(w|D)$ peut être estimé à l'aide de deux modèles : un modèle de termes simples (M_{D_t}) et un modèle de termes composés (M_{D_T}). Sachant qu'une requête Q est exprimée par des termes simples et des termes composés, les deux modèles de document sont estimés comme suit:

$$P(t|D) = P\left(t|M_{D_t}\right) \quad [1]$$

$$P(T|D) = \lambda P\left(T|M_{D_T}\right) + (1-\lambda) \prod_{t_k \in T} P\left(t_k|M_{D_t}\right) \quad [2]$$

Où λ ϵ [0, 1] est un paramètre de lissage, $P\left(T|M_{D_T}\right)$ et $P\left(t|M_{D_t}\right)$ peuvent être évalués en utilisant n'importe quel modèle de langue uni-gramme. Dans ce travail le lissage Dirichlet est utilisé. Ces deux modèles sont exprimés comme suit:

$$P_{Dir}(t|M_{D_t}) = \frac{F(t, D_t) + \mu P(t|C_t)}{|D_t| + \mu}$$

Où $F(t, D_t)$ est la fréquence de terme t dans le document D, $P(t|C_t)$ est le modèle de langue de la collection, $|D_t|$ est la longueur du document exprimé avec des termes simples et μ est un paramètre de lissage. De la même manière on a:

$$P_{Dir}(T|M_{D_T}) = \frac{F(T, D_T) + \mu' P(T|C_T)}{|D_T| + \mu'}$$

Où $P(T|C_T)$ est le modèle de langue de la collection, $F(T, D_T)$ est la fréquence de terme composé T dans le document D et $|D_T|$ est la longueur du document représenté par les termes composés, et μ' est un paramètre de lissage.

3.2. *Modèle de langue de la requête*

Après avoir estimé le modèle de document dans la section précédente, nous présentons maintenant l'estimation de modèle de la requête.

La plus simple manière d'estimer le modèle de la requête $P(w|Q)$ dans la formule (1) est d'utiliser l'estimation de maximum de vraisemblance (ML) $P_{ML}(w|Q)$. Cependant, la mesure de la distance KLD entre une requête courte (quelques termes en général) et le modèle de document ne peut pas être précise. Pour obtenir une évaluation plus précise de cette distance il est impératif d'estimer plus précisément le modèle de la requête. Afin, d'y parvenir, il faut attribuer une probabilité non nulle non seulement aux termes exprimés dans la requête initiale, mais également aux termes qui sont en relation avec les termes de la requête. Pour considérer cela dans notre approche, nous proposons une méthode similaire à celle développée dans (Bai *et al.*, 2005). Dans laquelle, nous lissons le modèle de la requête initiale noté $P_{org}(w|Q)$ avec un autre modèle considérant les relations entre termes noté $P_R(w|Q)$. Ainsi, le nouveau modèle de la requête est exprimé comme suit:

$$P(w|Q) = \varphi \times P_{org}(w|Q) + (1 - \varphi) \times P_R(w|Q)$$

Où φ est un paramètre de lissage.

En remplaçant ce modèle dans la formule d'appariement, elle devient ainsi:

$$score(Q, D) = \sum_{w \in V} \left(\varphi \times P_{org}(w|Q) + (1 - \varphi) \times P_R(w|Q)\right) logP(w|D)$$

Cette formule est récrite ainsi :

$$= \varphi \times \sum_{w \in Q} P_{org}(w|Q) logP(w|D) + (1 - \varphi) \times \sum_{w \in Q \cup DP} P_R(w|Q) logP(w|D)$$

Où DP est l'ensemble de documents de réinjection.

On note que le premier terme de la formule est une sommation à travers les termes de la requête (non pas à travers tous les termes du vocabulaire), car $P_{org}(w|Q) = 0$ pour tout terme n'appartenant pas à la requête. Nous assumons que les termes reliés à la requête originale sont ceux qui apparaissent seulement dans les

documents de réinjection DP. Nous ne considérons qu'un sous-ensemble de $Q \cup DP$, noté G ; où $|G| = N$; où N est le nombre de termes à ajouter à la requête originale (i.e. les termes ayant une plus grande probabilité). Par conséquent, la dernière formule devient ainsi :

$$score(Q,D) = \varphi \times \sum_{w \in Q} P_{org}(w|Q) logP(w|D) + (1-\varphi) \times \sum_{w \in G} P_R(w|Q) logP(w|D) \quad [3]$$

3.2.1. *Le modèle de la requête considérant les relations entre termes* $P_R(w|Q)$

Maintenant, nous présentons l'estimation de modèle de la requête utilisant les relations entre termes défini dans la formule (3).

Sachant que la requête Q est considérée comme un ensemble de termes composés T et de termes simples t, on peut estimer la probabilité $P_R(w|Q)$ comme suit :

$$P_R(w|Q) = \sum_{t \in Q} P_R(w|t) \times P_{org}(t|Q) + \sum_{T \in Q} P_R(w|T) \times P_{org}(T|Q)$$

Le principe de cette formule est le même que celui de modèle de traduction (Berger *et al.*, 1999). Cependant, dans (Berger *et al.*, 1999), il est utilisé pour l'expansion du modèle de document, dans notre cas il est utilisé dans le contexte d'expansion de la requête.

3.2.2 *Estimation du modèle de requête initiale de terme simple et de terme composé* $P_{org}(t|Q)$ *et* $P_{org}(T|Q)$*:*

Pour estimer ces deux probabilités nous supposons que la contribution d'un terme dans la requête initiale est liée à deux facteurs: (1) la fréquence du terme dans la requête, (2) le type du terme (simple ou composé). Afin calculer ce dernier facteur nous considérons qu'un terme composé ajoute plus de précision à la requête qu'un terme simple, nous lions cette apport à la taille du terme. Nous exprimons alors ces deux probabilités comme suit:

$$P_{org}(t|Q) = \frac{F(t,Q)}{|Q|} \quad \text{Et} \quad P_{org}(T|Q) = \frac{F(T,Q) \times |T|}{|Q|} \quad [4]$$

Où $|T|$ est la longueur de terme composé et $|Q| = \sum_{T \in Q} F(T,Q) \times |T| + \sum_{t \in Q} F(t,Q)$ est la longueur de la requête.

3.2.3. *Estimation de la probabilité* $P_R(w|T)$*:*

Le calcul de cette probabilité est basé sur l'hypothèse suivante: « Nous supposons que l'auteur d'un document utilise les termes composants (simples) isolément pour exprimer le terme composé comme abréviation après un nombre d'occurrences de terme composé ». Par exemple dans un document qui contient le terme composé «*compression de données* » l'auteur peut utiliser le terme composant «*compression*» pour exprimer le terme composé. De plus, si un terme est lié à un terme composant, il peut être aussi lié à ses termes composés. Par exemple,

le terme «*entropie*» est lié au terme «*compression*», il est également lié au terme composé «*compression de données*».

Afin de prendre en compte cette hypothèse, nous proposons d'estimer la probabilité $P_R(w|T)$ comme une combinaison entre la probabilité de relation de terme w avec le terme composé T, notée $P_{R\_org}(w|T)$ et la probabilité de relation entre le terme w avec l'ensemble des termes qui composent le terme T. La probabilité $P_R(w|T)$ est alors exprimée comme suit :

$$P_R(w|T) = \alpha P_{R\_org}(w|T) + (1-\alpha)P_R\big(w|c(T)\big) \quad [5]$$

Où $c(T)$ est l'ensemble de termes simples composants le terme T. α est un facteur d'interpolation pour contrôler l'apport d'un terme composé par rapport à ses termes composants.

Afin d'estimer la seconde partie de la formule, nous proposons une méthode similaire à celle développée dans le modèle de traduction (Berger *et al.*, 1999). Par conséquent, nous exprimons cette probabilité comme suit:

$$P_R(w|T) = \alpha P_{R\_org}(w|T) + (1-\alpha) \times \sum_{t \in T} P_R(w|t)P(t|T)$$

Le calcul des deux probabilités $P_{R\_org}(w|T)$ et $P_R(w|t)$ et fait en utilisant la formule (10). Pour le calcul de la probabilité $P(t|T)$ nous posons l'hypothèse suivante : « les termes composants d'un terme composé ne sont pas de même importance ». L'un des termes peut être plus important que d'autres. Exemple : le terme «*ordinateur*» est plus important que le terme «*personnel*» dans le terme composé «*ordinateur personnel*».

Nous considérons intuitivement que la dominance d'un terme est déterminée par sa spécificité, nous proposons de l'estimer de la manière suivante:

$$imp(t) = {}^{M}/_{df(t)} \quad [6]$$

Où $df(t)$est le nombre de documents où le terme t apparaît, et M est le nombre de documents dans la collection C.

Nous calculons la probabilité de dominance d'un terme simple dans un terme composé $P(t|T)$ comme suit :

$$P(t|T) = \frac{imp(t)}{\sum_{t,t_i \in T} imp(t_i)} \quad [7]$$

3.2.4 *Estimation de la probabilité $P_R(w|t)$:*

Nous nous sommes basés sur l'hypothèse suivante pour calculer la probabilité $P_R(w|t)$: « Nous supposons que l'utilisateur lorsqu'il utilise un terme simple dans sa requête, généralement, il fait référence à un ou plusieurs termes composés ». Par exemple, un utilisateur qui utilise le terme «*énergie*» dans sa

requête peut faire référence au terme composé «*énergie solaire*». Ainsi, nous proposons d'étendre la requête originale non seulement avec les termes liés au terme simple, mais aussi avec des termes liés à des termes composés qui contiennent ce terme simple, et cela relativement à la dominance de ce terme simple dans le terme composé et la fréquence de ce dernier dans la collection.

Afin de prendre en compte cette hypothèse, nous proposons de lisser la probabilité de relation de terme w avec le terme simple t, notée $P_{R\_org}(w|t)$ avec la probabilité de relation de terme w dans l'ensemble des termes composés auxquels le terme t appartient. La probabilité $P_R(w|t)$ est alors exprimée comme suit :

$$P_R(w|t) = \beta P_{R\_org}(w|t) + (1-\beta) P_R\big(w|C(t)\big) \quad [8]$$

Où $C(t)$ est l'ensemble des termes composés auxquels le terme t appartient et β est un paramètre de lissage qui contrôle la contribution de terme simple relativement à ses termes composés. De même que pour la formule (5), nous utilisons une méthode de traduction (Berger *et al.*, 1999). Nous obtenons alors la formulation suivante :

$$P_R(w|t) = \beta P_{R\_org}(w|t) + (1-\beta) \times \sum_{t \in T} P_R(w|T) P(T|t)$$

Le calcul des deux probabilités $\mathrm{P_{R\_org}(w|t)}$ et $\mathrm{P_R(w|T)}$ et fait en utilisant la formule (10). Pour estimer la probabilité $P(T|t)$ nous appliquons le théorème de Bayes, et on obtient:

$$P(T|t) = \frac{P(t|T)P(T)}{P(t)}$$

Où $P(t|T)$ est calculée en utilisant la formule (7), et $P(T)$ est estimée comme suit:

$$P(T) = \frac{df(T)}{\sum_{T_m \in C(t)} (df(T_m))} \quad [9]$$

Où $df(T)$ est le nombre de documents contenant le terme composé T. $C(t)$ est l'ensemble des termes composés auxquels le terme t appartient.

3.2.5. *Estimation de la relation entre termes* $P_R(w|w_j)$:

L'estimation de la relation entre termes consiste à calculer la probabilité $P_R(w|w_j)$. Comme dans de nombreuses études antérieures, nous exploitons dans ce travail la relation de cooccurrence, cette probabilité est estimée ainsi:

$$\mathrm{P_R(w|w_j)} = \frac{\mathrm{Count(w,w_j)}}{\sum_{\mathrm{w_i}} \mathrm{Count(w,w_i)}} \quad [10]$$

Tel que : $w, w_i, w_j \in V$ et $Count(w, w_j)$ est la fréquence de cooccurrence de couple (w, w_j) dans une fenêtre de texte de taille F.

4. Expérimentations et résultats

4.1. Collections de test et configuration expérimentale

Nous avons évalué notre modèle noté (QE-MM), décrit dans les sections précédentes, en utilisant deux collections de test TREC : WSJ90-92 (Wall Street Journal, 1990-92) et AP88 (Associated Press, 1988). La Table 1 ci-dessous montre quelques statistiques sur les collections et les requêtes utilisées. Seule la partie titre des requêtes est prise en compte.

Collection	# Documents	Requêtes d'apprentissage	Requêtes de test
WSJ90-92	74520	101-150	51-100
AP88	79919	101-150	51-100

Table 1. Statistiques sur les collections et les requêtes utilisées

Dans nos expérimentations, nous avons utilisé la plate-forme Terrier (http://terrier.org/). Les termes vides sont éliminés et l'algorithme de Porter est utilisé.

Pour l'extraction des termes composés nous avons utilisé l'outil Text-NSP (Banerjee *et al.*, 2003) où nous avons pris en compte les facteurs suivants : (1) la directionnalité entre termes simples (la contrainte d'ordre est respectée), par exemple le terme «*recherche information* » et le terme « *information recherche*» sont considérés comme deux termes composés différents, (2) l'adjacence entre les termes qui composent le terme composé (un terme composé ne peut être formé que par des termes simples adjacents) et (3) la taille d'un terme composé est fixée à deux, qui est une pratique commune et scalable pour de grandes collections. Nous avons sauvegardé dans la liste de termes composés uniquement les bi-grammes ayant une fréquence dans la collection supérieure à un seuil donné, fixé à 20 dans nos expérimentations. Afin de détecter un terme composé dans un document lors de processus d'indexation, nous avons utilisé une technique ad-hoc qui repose sur la concaténation de deux termes simples (non vides) adjacents, puis de vérifier si le terme existe dans la liste des termes composés. Si c'est le cas le terme composé est alors gardé comme index.

4.2. Valeurs des paramètres:

Il ya plusieurs paramètres de contrôle à affiner dans notre approche. Afin de trouver les valeurs optimales des paramètres et une comparaison équitable entre notre modèle et les modèles Baseline, nous avons utilisé des requêtes d'apprentissage (101-150). Nous avons estimé les différents paramètres des modèles d'une manière empirique de façon à optimiser la valeur de MAP. La table 2 ci-dessous illustre les valeurs de ces paramètres pour les trois modèles.

QE-MM			KLD		MLU	
Collection	AP88	WSJ90-92	AP88	WSJ90-92	AP88	WSJ90-92
μ (paramètre de lissage Dirichlet)	2000	500	500	300	1000	500
μ' (paramètre de lissage Dirichlet)	2500	300				
φ (formule(3))	0,3	0,3				
α (formule(5))	0,5	0,3				
β (formule(8))	0,5	0,6				
Nombre de documents de réinjection	3	3	10	14		
Nombre de termes d'expansion	50	30	50	50		
F (Taille de la fenêtre de texte)	20	50				

Table 2. Valeurs des paramètres

4.3 Évaluation

Dans nos expérimentations, les modèles suivants ont été comparés:

QE-MM: Notre modèle d'expansion de requêtes basée sur le Modèle de langue Mixte.
QE-MM-α = 1: Est le modèle QE-MM avec la non prise en compte de lissage dans la formule (5).
QE-MM-β = 1: Est le modèle QE-MM avec la non prise en compte de lissage dans la formule (8).
QE-MM-α = β = 1: Est le modèle QE-MM avec la non prise en compte de lissage dans les formules (5) et (8).
MLU: Modèle de Langue Uni-gramme.
MM: Modèle de langue Mixte sans expansion de requêtes.
KLD : Est une méthode populaire de pondération des termes d'expansion. La méthode KLD a obtenu une meilleure Précision (MAP) sur un ensemble de méthodes standards dans TREC 2009 (Ye *et al.*, 2009).

Afin d'évaluer notre modèle et de le comparer aux autres modèles nous avons utilisé la mesure MAP (Mean Average Precision), qui est une mesure largement acceptée pour l'évaluation de la performance des systèmes de recherche d'information. En outre, nous avons utilisé la précision à 10 et 20 documents (P@ 10, P@20) et le rappel (le nombre de documents pertinents retrouvés) comme mesures supplémentaires.

Les tables 3 et 4 montrent la comparaison entre les différents modèles de recherche. Afin de vérifier la significativité des résultats obtenus, nous avons effectué le test de Student et nous avons joint $^{+}$ et $^{++}$ pour l'indice d'amélioration dans les différents tables des résultats lorsque le test passe respectivement 95% et 99%.

Modèles	P@10	P@20	Rappel	MAP	Amélioration
MLU	0,3286	0,3041	0,6275	0,2471	
MM	0, 3367	0, 3041	0,6578	0,2684	8,62 % $^{++}$
KLD	0,3612	0,3500	0,6968	0,3085	24,85% $^{++}$
QE-MM-β=1	0,3980	0,3714	0,7072	0,3246	31,36% $^{++}$
QE-MM-α=1	**0.4143**	0.3786	0,7115	0,3250	31,53% $^{++}$

QE-MM-α=β=1	0,4041	0,3684	0,7058	0,3223	30,43% $^{++}$
QE-MM	0,4082	**0,3867**	**0,7197**	**0,3289**	**33,10%** $^{++}$

Table 3. Résultats des différents modèles sur la collection AP88.

Modèles	P@10	P@20	Rappel	MAP	Amélioration
MLU	0,283 3	0,2635	0,6727	0,1971	
MM	0,2917	0,2708	0,6754	0,2063	4,67%
KLD	0,3083	0,2760	**0,7155**	0,2310	17,20% $^{++}$
QE-MM-β=1	0.3167	0.2833	0,7058	0,2338	18,62% $^{++}$
QE-MM-α=1	0,3229	0,2802	0,7090	0,2424	22,98 % $^{++}$
QE-MM-α=β=1	0,3104	0,2813	0,7063	0,2327	18,06% $^{+}$
QE-MM	**0,3292**	**0,2844**	0,7090	**0,2449**	**24,25 %** $^{++}$

Table 4. Résultats des différents modèles sur la collection WSJ90-92.

D'après les tables ci-dessus, nous pouvons tirer les remarques et conclusions suivantes:

Le Modèle de langue Mixte (MM) améliore le modèle Uni-gramme (MLU) en termes de précision et de rappel. Cela montre que l'utilisation des termes composés combinés avec les termes simples à l'étape de recherche initiale peut être utile pour la RI.

Les modèles KLD et QE-MM améliorent significativement le modèle Uni-gramme (MLU), cela reconfirme que l'expansion de requête à un apport effectif pour la Recherche d'Information. De plus, notre modèle surpasse le modèle KLD. Nous avons obtenu une amélioration de l'ordre de +24,25% et +33,10% par rapport au modèle MLU sur les collections AP88 et WSJ90-92 respectivement. Cela montre que l'expansion de requête basée sur un modèle de langue mixte peut sélectionner et pondérer les termes d'expansion mieux que le modèle d'expansion KLD.

Enfin, le modèle QE-MM donne de meilleurs résultats que ses trois autres versions (QE-MM-α = 1, QE-MM-β = 1, QE-MM-α = β = 1) où le lissage est ignoré respectivement dans les formules (5), (8) et (5) (8). Ceci montre qu'il est intéressant d'ajouter non seulement les termes qui sont liés au terme simple de la requête, mais aussi les termes qui sont liés aux termes composés qui contiennent ce terme simple. Et inversement ajouter non seulement les termes qui sont liés à un terme composé de la requête, mais aussi les termes qui sont liés à ses termes composants est aussi intéressant.

Pour illustrer les améliorations de notre modèle par rapport au modèle KLD, nous avons examiné manuellement quelques requêtes. Nous présentons ci-dessous un exemple de requête numéro 88: « *Topic: Crude Oil Price Trends* », où nous montrons les premiers termes d'expansion de la requête sélectionnés par les deux modèles, sur la collection WSJ90-92.

Modèle QE-MM	Modèle KLD
opec	*opec*
joint meet	*barrel*
opec member	*produc*
price committe	*meet*
opec produc	*cartel*
iran iraq	*cent*
barrel	*output*
world oil	***petroleum***
sourc	*lukman,*
ink	*suppli,*
group	*committe*
newspap	***energi***
oil produc	***gallon***
…..….	…..….

Table 5. Termes d'expansion (lemmatisés) générés par les modèles KLD et QE-MM

Comme on peut le voir dans la table 5., notre modèle d'expansion permet de sélectionner de nouveaux termes pertinents (**en gras**) que ceux sélectionnés par le modèle KLD. Par exemple, le terme composé «***opec production***» est plus précis que les termes «*opec*» et «*production*», sélectionnés séparément par le modèle KLD. Par conséquent, en utilisant cette requête notre modèle obtient une meilleure précision que le modèle KLD. Précisément, notre modèle obtient une précision moyenne égale à 0.1143, et les modèles KLD et MLU obtiennent respectivement : 0.0577 et 0.0501 de précision moyenne.

5. Conclusion

Dans cet article, nous avons décrit une nouvelle approche pour l'expansion de la requête basée sur un modèle de langue mixte combinant les termes simples et composés. Les expérimentations effectuées sur deux collections TREC ont montré que notre modèle améliore significativement le modèle uni-gramme et affiche de meilleurs résultats que le modèle utilisant la méthode d'expansion KLD.

Dans le futur, nous prévoyons d'explorer différents points. En premier lieu, l'utilisation d'autres types de relations pour sélectionner les termes d'expansion telle que l'information mutuelle. Deuxièmement, nous examinerons l'efficacité de notre méthode en utilisant la collection comme source de données pour l'expansion de la requête (méthode globale).

6. Bibliographie

Amati, G. Probabilistic models for information retrieval based on divergence from randomness, Ph.D. Thesis, Department of Computing Science, University of Glasgow, UK, 2003.

Bai, J., Song, D., Bruza, P., Nie, J. Y. and Cao, G., « Query expansion using term relationships in language models for information retrieval», *In ACM International Conference on Information and Knowledge Management*, 2005, p. 688-695.

Banerjee, S., and Pedersen, T., « The Design, Implementation, and Use of the Ngram Statistic Package», *In international conference on Computational linguistics and intelligent text processing*, 2003, P. 370-381.

Berger, A. and Lafferty, J., «Information retrieval as statistical translation», *In ACM International Conference on Research and Development in Information Retrieval*, 1999, p. 222–229.

Carpineto, C. DeMori, R., Romano, G., and Bigi, B. «An information theoretic approach to automatic query expansion», *In ACM Transaction Information System*, 19, 2001, p. 1–27.

Carpineto, C. and Romano, G., «A Survey of Automatic Query Expansion in Information Retrieval», *In ACM Computing Surveys*, Vol. 44, No. 1. 2012.

Gao, J.F., Nie, J.Y., Wu, G., and Cao, G. «Dependence Language Model for Information Retrieval». *In ACM International Conference on Research and Development in Information Retrieval*, 2004, p. 170-177.

Hammache, A., Boughanem, M., Ahmed-Ouamer, R., «A new language model combining single and compound terms», *In IEEE ACM Web Intelligence Conference*, 2011, p. 67-70.

Lafferty, J. and Zhai, C., «Document language models, query models, and risk minimization for information retrieval», *In ACM International Conference on Research and Development in Information Retrieval*, 2001, p. 111–119.

Lavrenko, V. and Croft, W. B., «Relevance based language models», *In ACM International Conference on Research and Development in Information Retrieval*, 2001, p. 120–127.

Lv, Y., Zhai. C., «A comparative study of methods for estimating query language models with pseudo feedback», *In ACM International Conference on Information and Knowledge Management*, 2009a, p. 1895-1898.

Lv, Y., Zhai. C. , «Positional language models for information retrieval», *In ACM International Conference on Research and Development in Information Retrieval*, 2009b, p. 299-306.

Lv, Y., Zhai. C., «Positional Relevance Model for Pseudo-Relevance Feedback», *In ACM International Conference on Research and Development in Information Retrieval*, 2010, p. 579-586.

Maisonnasse, L., Gaussier, E., Chevallet J.P., «Modélisation de relations dans l'approche modèle de langue en recherché d'information», *In Conférence en Recherche d'Information et Applications, 2008, p.305-319.*

Manning, C. D., Raghavan, P., Schutze, H., *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Manning, C., and Schutze, H. *Fondation of statistical Natural Language Processing* (6th ed.), MIT Press, Cambridge, MA, 2003.

Metzler, D., and Croft, W.B., «A Markov Random Field model for term dependencies», *In ACM International Conference on Research and Development in Information Retrieval*, 2005, p. 472–479.

Metzler, D. and Croft, W. B., «Latent concept expansion using markov random fields», *In ACM International Conference on Research and Development in Information Retrieval*, 2007, p. 311–318.

Miao, J., Huang, J.X.,Ye., Z., «Proximity-based Rocchio's Model for Pseudo Relevance Feedback», *In ACM International Conference on Research and Development in Information Retrieval*, 2012, p. 535-544.

Ponte, J. and Croft, W.B., «A language modeling approach to information retrieval», *In ACM International Conference on Research and Development in Information Retrieval*, 1998, p. 275-281.

Qiu, Y. and Frei, H.P., «Concept based query expansion», *In ACM International Conference on Research and Development in Information Retrieval*, 1993, p. 160-169.

Robertson, S.E., Walker. S., Jones, S., Hancock-Beaulieu, M. and Gatford, M., «Okapi at trec-3», *In TREC '94*, p. 109-126.

Rocchio, J. J., «Relevance feedback in information retrieval», *In the SMART Retrieval System: Experiments in Automatic Document Processing*, 1971, p. 313-323.

Salton., G, Buckley., C., «Improving retrieval performance by relevance feedback», *Journal of the American Society for Information Science*, 41, 1990, p. 288-297.

Song, F., and Croft, W. B., «A general language model for information retrieval», *In ACM International Conference on Research and Development in Information Retrieval*, 1999, p. 279-280.

Srikanth, M., and Srihari, R., «Biterm language models for document retrieval», *In International Conference on Research and Development in Information Retrieval*, 2002, p. 425–426.

Ye, Z., Huang, X., He, B., Lin, H., «York University at TREC 2009: relevance feedback track», *Proceedings of the 18th Text Retrieval Conference Gaithersburg*, 2009, p. 1–6.

Zhai, C. and Lafferty, J., «Model-based feedback in the language modeling approach to information retrieval», *In ACM International Conference on Information and Knowledge Management*, 2001, p. 403–410.

Zhai, C., Lafferty, J., «A study of smoothing methods for language models applied to information retrieval», *ACM Transactions on Information Systems (TOIS)* Volume 22, Issue 2, 2004, p. 179 – 214.