Classification non supervisée Floue des termes basée sur la proximité pour les systèmes de recherche d'information

Ilyes Khennak[1]

USTHB, LRIA, Département d'Informatique,
BP 32 El-Alia, Alger, Algérie

RÉSUMÉ. *Le regroupement des termes basé sur la mesure de proximité est une stratégie menant efficacement à trouver les documents pertinents. Contrairement à ce qu'ont montré les études récentes qui ont utilisé la proximité des termes pour le classement des documents, le processus de recherche d'information est entièrement revu dans ce travail en ce qui concerne les étapes d'indexation et d'interrogation. Par conséquent, un Fichier Inverse Etendu est construit en exploitant le concept de proximité des termes et en utilisant les technologies de classification non supervisée. Trois approches d'interrogation sont alors proposées, la première utilise l'expansion de la requête, la seconde est basée sur le Fichier Inverse Etendu et la dernière hybride les méthodes de recherche. De nombreuses expérimentations sur OHSUMED ont été effectuées et les résultats obtenus sont très prometteurs.*

ABSTRACT. *Term clustering based on proximity measure is a strategy leading to efficiently yield documents relevance. Unlike the recent studies that investigated term proximity for documents ranking, the information retrieval process is thoroughly revised on both indexing and interrogation steps in this work. Consequently, an extended inverted file is built by exploiting the term proximity concept and using clustering technologies. Then three interrogation approaches are proposed, the first one uses query expansion, the second one is based on the extended inverted file and the last one hybridizes the retrieval methods. Extensive experiments on OHSUMED have been performed and the achieved results are very promising.*

MOTS-CLÉS : *recherche d'information, proximité des termes, association des mots, regroupement flou.*

KEYWORDS: *information retrieval, term proximity, word association, fuzzy clustering.*

1. Directeur de thèse: Habiba Drias

1. Introduction

Avec le développement de systèmes de communication électronique et l'accroissement de la quantité d'information disponible en ligne, il devient important d'aider les utilisateurs à accéder rapidement à l'information recherchée.

La recherche d'information classique a été largement explorée, elle a fait l'objet de nombreux ouvrages tels que : (Baeza-Yates *et al.,* 1999) (Manning *et al.,* 2008) (Chu, 2010) et (Kowalski, 2011). Elle offre des techniques et des outils permettant de retrouver une information intéressante qui satisfait un besoin en information à partir des bases de données. Elle permet de sélectionner parmi un volume d'information, les informations pertinentes à une requête utilisateur. Les systèmes de RI classiques considèrent les documents comme des ensembles de mots dépourvus de sens. De ce fait, ils permettent de retrouver seulement des documents qui sont décrits par les mêmes mots que la requête. En d'autres termes, ils traitent la RI en se basant seulement sur l'aspect morphologique du texte du document. Intuitivement, quand d'autres considérations telles que les caractéristiques sémantiques sont prises en compte, le système de recherche d'information devrait fonctionner plus rapidement et plus efficacement.

1.1. *Revue de la littérature*

Afin de remédier aux limites de la recherche d'information classique, de nouvelles approches basées sur les sens des mots ont été proposées (Cummins *et al.,* 2009) (He *et al.,* 2011) (Mingjie *et al.,* 2007) (Vechtomova *et al.,* 2006) et (Wei *et al.,* 2007). Elles utilisent des ressources sémantiques et des techniques statistiques pour améliorer la performance des systèmes de recherche d'information.

Dans les articles (Cummins *et al.,* 2009) (Mingjie *et al.,* 2007) et (Vechtomova *et al.,* 2006), les auteurs proposent des mesures de proximité des termes qu'ils combinent avec la fonction de pondération des termes classique afin de mieux traduire l'importance des termes dans un document.

Dans les papiers (He *et al.,* 2011) et (Wei *et al.,* 2007), les auteurs présentent une étude théorique sur la modélisation de la proximité des termes et montrent que l'utilisation de la proximité des termes permet d'améliorer considérablement l'efficacité du système.

1.2. *Objectifs assignés*

Habituellement on entend par la proximité des termes, le nombre minimum de mots qui séparent deux termes apparaissant dans le même document. Lorsque ce nombre est petit, la proximité est plus importante et lorsqu'il est égal à 1 cela signifie que les deux termes sont adjacents, tels que « recherche d'information ». Plus précisément, on parle d'association des termes. Dans cette étude, nous focalisons notre intérêt sur l'exploitation du concept de proximité, car il est plus général que le concept

d'association. En outre, un terme peut avoir une proximité conséquente avec plusieurs autres mots et dans plusieurs documents. Par exemple, le mot « information » peut être associé à « recherche », « science » ou « technologie ».

2. Indexation avec la proximité des termes

Dans la phase d'indexation, les termes sont regroupés dans des classes selon leur proximité avec d'autres termes dans les documents. Cette phase est conçue en utilisant les techniques de classification non supervisée. Le regroupement des termes est implémenté afin de construire un Fichier Inverse Etendu. Ce fichier contient en plus des relations [Terme, Document], des relations [(Terme$_1$, Terme$_2$), Document] telles que : Terme$_1$ et Terme$_2$ appartiennent à la même classe.

2.1. *Regroupement des termes*

Pour le regroupement des termes, nous procédons à la modélisation de l'espace des termes par rapport aux concepts de Clustering comme suit : Un terme du dictionnaire étant l'*objet*, le poids du terme dans un document étant l'*attribut* et l'ensemble de termes co-occurrents souvent dans les mêmes documents étant une *classe* (cluster). La mesure *cosinus* est utilisée pour le calcul de la distance entre deux termes. La formule [1] permet de calculer ce score où x_i et y_i sont respectivement les poids du terme x et du terme y dans le document i.

$$distance(x,y) = \frac{\sum_{i=1}^{n} x_i * y_i}{\sqrt{\sum_{i=1}^{n}(x_i)^2 * \sum_{i=1}^{n}(y_i)^2}} \quad [1]$$

L'algorithme de regroupement Fuzzy k-means est plus adapté pour le regroupement des termes. Dans cet algorithme, un objet peut appartenir à plusieurs classes. Pour faire face à la question de la complexité de l'algorithme Fuzzy k-means, nous commençons par fixer le nombre de classes par rapport au nombre de termes du dictionnaire. Ensuite, nous affectons chacun de ces termes à une classe, chaque terme du dictionnaire correspond à un centroïde. Une fois les classes définies, nous assignons pour chacune d'entre elles les termes co-occurrents avec le centroïde de la classe. Cette assignation est faite à l'aide d'un calcul de *probabilité conditionnelle* entre le centroïde de la classe et chacun des termes du dictionnaire. Nous avons remplacé la distance par la probabilité conditionnelle de proximité, $proximite(t_1|t_2)$ (Wei *et al.,* 2007), qui est donnée par la formule [2]. $Count(t_1 \vee t_2)$ compte le nombre de documents où les

termes t_1 et t_2 apparaissent ensemble et $Count(t_2)$ compte le nombre de documents où le terme t_2 apparaît.

$$proximite(t_1|t_2) = \frac{P(t_1 \vee t_2)}{P(t_2)} = \frac{Count(t_1 \vee t_2)}{Count(t_2)} \quad [2]$$

Le résultat est un ensemble de classes où chacune est décrite par un terme du dictionnaire (*centroïde*) et un vecteur de proximités. Le vecteur des termes est trié selon les proximités. Ce résultat est ensuite stocké dans un fichier appelé *Clusters*.

2.2. *Création du Fichier Inverse Etendu*

Dans cette étape, nous récupérons pour chaque terme (différent de centroïde) d'une classe les documents dans lesquels il apparaît avec le centroïde de celle-ci. Nous sauvegardons ensuite l'index obtenu dans le fichier *Fichier Inverse Etendu*.

3. Interrogation avec la proximité des termes

Comme dans le SRI classique, les documents et la requête sont représentés par des vecteurs de poids des termes. Ainsi, la même fonction d'appariement est utilisée pour calculer la similarité entre un document et une requête. Dans cette section, nous présentons trois méthodes de recherche réalisées.

Expansion de la requête : Dans cette méthode de recherche, nous utilisons le fichier *Clusters* afin de récupérer les termes co-occurrents avec une certaine proximité avec ceux de la requête. Nous utilisons ensuite ces termes (en plus des termes de la requête) pour sélectionner les documents qui contiennent au moins un de ceux-ci.

Utilisation du Fichier Inverse Etendu : Dans cette méthode, nous sélectionnons tous les documents contenant au moins un couple de termes de la requête existant dans le *Fichier Inverse Etendu*.

Hybridation entre les méthodes de recherche : Pour cette méthode de recherche, nous exploitons le *Fichier Inverse* et le *Fichier Inverse Etendu* ensemble.

4. Résultats expérimentaux

4.1. *La collection OHSUMED*

De nombreuses expérimentations ont été effectuées sur la collection OHSUMED (corpus utilisé pour la tâche de filtrage de TREC-9). Cette collection est composée de 350 000 d'extraits de journaux médicaux datés de 1987 à 1991. En général, ces textes comportent un titre et un résumé. Les évaluations ont été effectuées en utilisant uniquement les titres. Afin d'évaluer nos approches, et plus particulièrement les méthodes

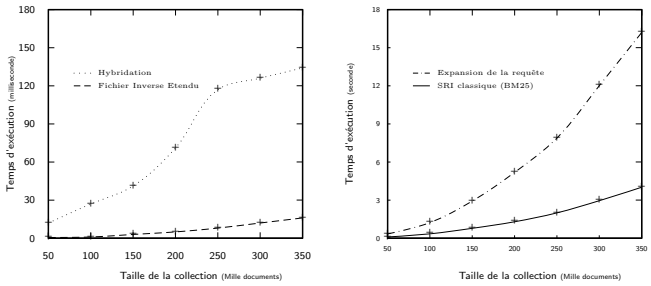

Figure 1 – Comparaison entre les approches de recherche par rapport au temps d'exécution

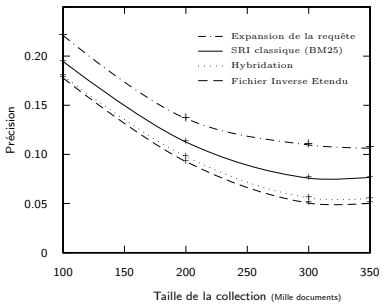

Figure 2 – Comparaison entre les méthodes de recherche par rapport à la pertinence des documents retournés

de recherche, nous avons effectué un partitionnement de la collection de documents en plusieurs sous-collections.

4.2. *Comparaison de notre SRI avec le SRI classique*

Dans cette section, nous comparons les résultats de nos méthodes de recherche avec ceux du SRI classique. Ce SRI est basé sur le BM25 qui est une extension du modèle probabiliste (Robertson *et al.,* 2009). Les expérimentations ont été effectuées sur la base de 106 requêtes. La figure 1. présente les courbes du SRI classique et des trois algorithmes de recherche proposés par rapport au temps d'exécution. Nous remarquons non seulement que le temps d'exécution de l'approche du Fichier Inverse Etendu est le plus court, mais il est également pratiquement constant et proche de 0. Du point de vue de la performance, figure 2. Illustre la supériorité de la stratégie d'expansion de la requête par rapport aux autres méthodes de recherche. Il est à noter que

d'une façon générale, les résultats obtenus avec l'utilisation du Fichier Inverse Etendu ont montré une baisse significative du temps d'exécution et du nombre de documents sélectionnés, associé à de très bonnes performances par rapport aux résultats obtenus par le SRI classique.

5. Conclusion

Ce travail nous a permis d'étudier et de réaliser un système de recherche d'information basé sur l'interprétation de la notion de proximité des termes. Dans le cadre de ce travail, nous avons proposé une méthode d'indexation basée sur le regroupement des termes à l'aide de méthodes statistiques et de méthodes de Clustering. La réalisation de cette proposition se résume par la création du *fichier Clusters* et du *Fichier Inverse Etendu*. Ces fichiers sont utilisés dans la phase d'interrogation par plusieurs méthodes de recherche afin de retrouver rapidement les documents pertinents. Ces méthodes sont très performantes et donnent des résultats très satisfaisants en manière de robustesse et de temps de calcul.

6. Bibliographie

Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, Addison-Wesley, New York, 1999.

Chu H., *Information representation and retrieval in the digital age*, Information Today, New Jersey, 2010.

Cummins R., O'Riordan C., « Learning in a pairwise term-term proximity framework for information retrieval », *SIGIR'09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, Boston, p. 251-258, 19-23 juillet, 2009.

He B., Huang J., Zhou X., « Modeling Term Proximity for Probabilistic Information Retrieval Models », *Information Sciences*, vol. 181, n° 14, p. 3031-3017, 2011.

Kowalski G., *Information Retrieval Architecture and Algorithms*, Springer, New York, 2011.

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

Mingjie Z., Shuming S., Mingjing L., Ji-Rong W., « Effective top-k computation in retrieving structured documents with term-proximity support », *CIKM'07 Proceedings of the 16th ACM conference on information and knowledge management*, ACM, New York, Lisboa, p. 771-780, 6-9 novembre, 2007.

Robertson S., Zaragoza H., *The probabilistic relevance framework : BM25 and beyond*, Now Publishers Inc, Hanover, 2009.

Vechtomova O., Wang Y., « A study of the effect of term proximity on query expansion », *Information Sciences*, vol. 32, n° 4, p. 324-333, 2006.

Wei X., Croft W., « Modeling Term Associations for Ad-Hoc Retrieval Performance Within Language Modeling Framework », *ECIR'7 Proceedings of the 29th European conference on IR research*, Springer-Verlag, Heidelberg, Rome, p. 52-63, 2-5 avril, 2007.