Experiments on two Query Expansion Approaches for a Proximity-based Information Retrieval Model

Bissan Audeh[1]

Institut Henri FAYOL
École Nationale Supérieure des Mines de Saint-Étienne
158 cours Fauriel
42000 Saint-Étienne
France

audeh@emse.fr

ABSTRACT. *Query expansion is a well-known technique used to overcome the word-mismatch drawback of keyword retrieval models. Fully automated query expansion comes with the risk of query drift. In our work we faced this phenomenon while trying to expand boolean queries for a Proximity-based information retrieval model. This model gets good precision in evaluation campaigns but gives a small number of results. Our experiments are focused on two different query expansion approaches: a global approach using WordNet synonyms and a local approach using pseudo relevance feedback based on LSA (Latent Semantic Analysis) to create a query-time thesaurus. The results we've got show an important query drift effect for both approaches. In this paper we present these experiences with an analysis of the results and the perspectives we are currently working on.*

RÉSUMÉ. *L'expansion de requêtes est une technique bien connue pour dépasser l'exigence de recherche par mot exact en recherche d'information. Cependant, l'expansion automatique vient avec le risque de dérive de la requête. Dans ce travail nous avons eu ce problème en essayant d'étendre des requêtes booléennes pour un modèle de recherche basé sur la proximité. Ce modèle obtient une bonne précision dans les campagnes d'évaluation mais il rend très peu de résultats. Dans ce travail, nous avons utilisé deux approches : une approche globale qui utilise les synonymes de WordNet, et une approche locale basée sur le retour de pertinence et l'utilisation de LSA pour créer un thésaurus local. Les résultats que nous avons obtenus montrent un effet important de dérive de la requête. Dans ce papier nous présentons ces expériences avec une analyse des résultats et les perspectives que nous envisageons.*

KEYWORDS: *Query Expansion, Fuzzy Proximity model, Query Drift, Latent Semantic Analysis, thesaurus, WordNet, Relevance Feedback.*

MOTS-CLÉS: *Expansion de requêtes, Modèle de proximité flou, Dérive de la requête, Retour de pertinence, thésaurus, LSA, WordNet*

---

[1]. Directeur de thèse : Olivier BOISSIER, Co-Encadrants: Philippe BEAUNE et Michel BEIGBEDER

1. Introduction

Any information need should be expressed by words in order to use an information retrieval system. In reality it is not uncommon for users to use terms that don't necessarily appear in all relevant documents. The severity of this problem tends to decrease for long disjunctive queries, as there is more chance of some important words co-occurring in the query and relevant documents (Xu & Croft, 1996); this is the reason why query expansion is a natural solution to this issue.

Our work uses query expansion in order to increase the recall of an information retrieval model based on fuzzy proximity matching. This model is extremely selective, which is the reason why it gets high precision rates. On the other hand, its strong selectivity minimizes the number of results returned by this model.

In this paper we present the application of local and global query expansion approaches for a proximity-based information retrieval model. Section 2 introduces the fuzzy proximity model. In section 3 we give an overview on query expansion in information retrieval. The two proposed query expansion approaches are explained in Section 4; the experiments achieved on these approaches are in section 5. Finally the last section concludes this work and presents our perspectives.

2. Fuzzy Proximity Matching Model

The fuzzy proximity model (Beigbeder & Mercier, 2005) is based on the idea that the distance between query terms in a document is important when deciding the score of the document: the smaller this distance is, the more relevant the document tends to be. To apply this idea, the model calculates a proximity function between a position in the text and one term. The query in this model is represented as a tree where the leaves are terms and the internal nodes are boolean operators. To score a document, the proximity function is applied recursively to query terms, while operations are interpreted as min (for *AND* operator) and max (for *OR* operator) combinations over that function.

This model was tested on TREC WT10g collection where it achieved good results when compared to OKAPI vector model; these good results were approved later in other evaluation campaigns (CLEF and INEX). In these evaluations it was noticed that the number of returned documents for a query is very small, this is due to the strong association exigency demanded by this model among query terms in a document. To overcome this problem query expansion could be one solution.

3. Query expansion in information retrieval

In the literature, query expansion techniques are categorized in different ways. In this paper we refer to methods that use a subset of documents collection for query

expansion as local approach, while methods that use a thesaurus based on the whole documents collection or an external resource are considered as global approach.

For local approach, the most popular method is relevance feedback based on Rocchios' formula (Rocchio, 1971), which is adapted to the vector space model. Many variations of this basic method have been proposed, and other alternatives exist for probabilistic models. In general, relevance feedback showed improvement to retrieval effectiveness when choosing the right values for the different variables (Salton & Buckley, 1990). For simplicity, pseudo relevance feedback is often used, but it is not as effective as user-based feedback. Other local methods were proposed based on creating local thesaurus from top ranked documents (Attar & Fraenkel, 1977).

For global approach, many techniques tried to create a collection-based thesaurus, like term clustering (Jones, 1971), Latent Semantic Analysis (Deerwester, Furnas, Landauer, & Harshman, 1990) and similarity thesaurus (Frei, 1993). Most of these methods are known to be robust, but they require corpus-wide statistics, which is not easy for certain high consuming methods like LSA especially with considerable sizes of nowadays collections. In contrary, methods based on simple co-occurrence were not very successful (Peat & Willett, 1991). Another alternative is the use of an external resource independent from the collection. In recent works it is frequent to use WordNet[2], combined in some cases to a domain-based ontology. Global methods could suffer from term ambiguity, which decreases the precision of the expanded query. Croft (Xu & Croft, 1996) shows that combining global (collection-based) techniques with local approaches can give more effective and predictable results than using only local approaches, plus it is less expensive than using the exhaustive global approach.

Finally, query expansion is aimed to increase the recall. Nevertheless, automatic addition of new terms can cause query drift, which is an "alteration of the focus of a search topic" (Mitra, Singhal, & Buckley, 1998).

4. Query Expansion for the Fuzzy Proximity Model

Investigating the state of the art of query expansion shows that the success of this technique was not confirmed in all cases (He & Ounis, 2009). In addition, to our knowledge no experiments have been done on query expansion for the case of the proximity-based model. In order to have an idea of the effect of query expansion on this model we tried local and global query expansion approaches. In both cases the new expansion terms are added to the initial boolean query using *OR* operator between each term and its extensions, while the main structure of the query is kept intact.

---

[2]. WordNet is a lexical database of English. (link at year 2011: http://wordnet.princeton.edu/)

4.1. *Local Approach*

As mentioned in Section 3, Latent Semantic Analysis was already used for query expansion. The application of this method to a proximity-based information retrieval model has not been investigated. In this approach we want to use the benefits of LSA without being limited to its performance issues. We use pseudo relevance feedback to obtain a subset of top ranked documents. Supposing that these documents are relevant, we use them to create a local matrix A containing *Term×Document* tf values, to which we apply LSA. More specifically, we apply Singular Value Decomposition (*SVD*) to obtain the three matrices U, S and V. Finally; dimensionality reduction to level k is performed to get the matrices U_k, S_k and V_k forming a new "Concept Space" representation. These matrices can reproduce an approximation to the matrix A using the equation: $A \approx U_k S_k V_k^T$.

In our approach, we are interested in matrix U_k that contains the vectors of terms in the new "Concept Space". Using this matrix, we can find most related terms for each query term, by measuring cosines similarity between two terms vectors.

4.2. *Global Approach*

The second approach is based on WordNet. WordNet groups words and their synonyms into sets called Synsets; each of these Synsets presents one unique concept. The concepts are then connected to each other regarding their semantic and lexical relations. For experimentation purpose, we expand each query term by adding terms in all Synsets containing the term. In spite of its simplicity, this experience will help us understand query drift effect on proximity-based model.

5. Experiments and Discussion

5.1. *Experimental Setup*

We used the INEX2009 collection, which contains about 2600000 Wikipedia english articles in XML format. These articles are annotated using YAGO[3] semantic tags. Titles of INEX2009 topics were used as queries. Other fields could be used for disambiguation but it's not the focus of this work which studies query expansion effects in general. Porter method is used for stemming and phrases are not expanded.

For the local approach, no big difference was noticed when using 10, 30 or 50 top documents as feedback. The number of dimensions of LSA is chosen to be half the number of retained documents. For global approach all synonyms of each query term are taken, while we took 5 most similar terms in local approach.

---

[3] YAGO is a semantic database derived from wikipedia and wordnet. (link at year 2011: *http://www.mpi-inf.mpg.de/yago-naga/yago/index.html)*

5.2. *Results*

Figure 1 shows different measures for proximity-based model without expansion and the effect of applying local and global query expansion.

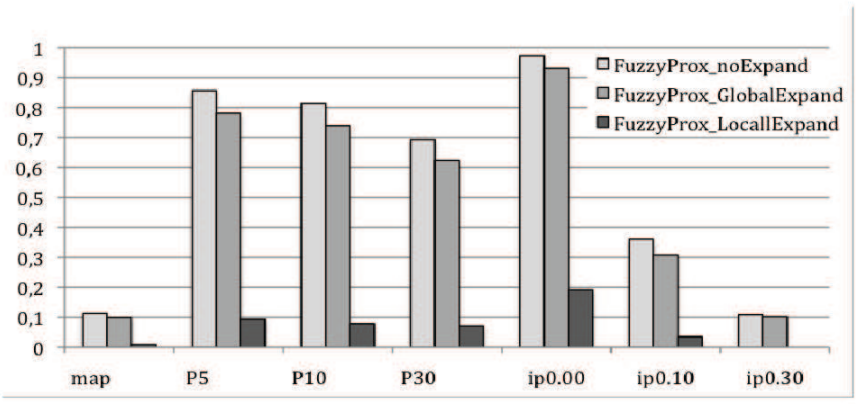

Figure 1. *Evaluation measures for the proximity model with local and global expansion: map, precision at {5,10,30}, interpolated recall-precision at {0,10,30}*

We notice a degradation of precision at different recall levels when using both query expansion approaches, though it is more important with local approach.

5.3. *Discussion*

The experiments presented in this paper are an explorative study. The high selectivity of the proximity model was very clear in our experiments, for some queries we got less than 5 results. This fact explains the low performance of the local approach that in order to be useful needs enough number of documents. On the other hand, we noticed that for many queries the new expansion terms were the same for all query terms. This could be explained by the nature of the fuzzy proximity model: it is designed to privilege documents that contain query terms as close as possible to each other; as a result, the top ranked documents contained almost the same frequency for each term in the query. This high association between query terms in top retrieved documents is the reason why these terms got practically the same vectors in the terms vectors matrix. In consequence the top similar terms vectors were the same for these terms. Thus the conjunctive aspect of the original query is lost, and the expanded query is pushed away from its desired direction.

For global approach, this phenomenon was produced differently; adding all the synonyms of all concepts represented by a term added different senses to the query. The severity of this fact though didn't prevent this approach of performing better than the local method. The reason of this is that using a resource that is independent

from the collection was more suitable than depending completely on the limited number of query results for the fuzzy proximity model.

6. Conclusion and perspectives

This work revealed the challenges and limitations when applying query expansion to the fuzzy proximity model. The experiments showed that a local query expansion approach could suffer from the limited number of results of this model. On the other hand, using an external resource for query expansion without choosing the right sense of each term didn't ameliorate the base-model results.

Currently we are considering more detailed study of an external resource usage. Depending on the state of the art of that case we will investigate more sophisticated use of WordNet, which involves selecting the "good" expansion terms of a word instead of taking all its synonyms. One idea to achieve this for the case of fuzzy proximity model is to combine global approach to a local approach to profit from the high precision of this model. We may also consider the use of YAGO ontology, which is already used to annotate our test collection.

7. Bibliography

Attar, R., & Fraenkel, A. S. (1977). Local Feedback in Full-Text Retrieval Systems. Journal of the Association for Computing Machinery. J.ACM, 24(3), 397-417.

Beigbeder, M., & Mercier, A. (2005). An information retrieval model using the fuzzy proximity degree of term occurences. Proceedings of SAC '05. New York, USA: ACM Press.

Deerwester, S., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis, 41(6).

Frei, Y. Q. and H. P. (1993). Concept Based Query Expansion. SIGIR '93 (Vol. 11, p. 212). NY: ACM.

He, B., & Ounis, I. (2009). Studying Query Expansion Effectiveness. Proceedings of ECIR '09 European Conference in Information Retrieval.

Jones, K. S. (1971). Automatic keyword classification for information retrieval. Butterworth's (London). Archon Books (1971).

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. Proceedings of SIGIR'98, 206-214. New York, USA: ACM Press.

Peat, H. J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. American Society forInformation Science, 42(5).

Rocchio, J. (1971). Relevance Feedback in Information Retrieval. in Salton: The SMART Retrieval System: Experiments in Automatic Document Processing.

Salton, G., & Buckley, C. (1990). Improving Retrieval Performance by Relevance Feedback. Society, 41(4).

Xu, J., & Croft, W. B. (1996). Query expansion using local and global documentanalysis. Proceedings of SIGIR '96. New York, New York, USA: ACM Press.