Semantic query expansion for fuzzy proximity information retrieval model

Author **Bissan AUDEH**

Supervisors Philippe BEAUNE, Michel BEIGBEDER, Olivier BOISSIER

Studies/Stage Second year PhD student

Affiliation École Nationale Supérieure des Mines de Saint-Etienne

E-Mail audeh@emse.fr

Aims and Objectives of the Research

Our research aim is to ameliorate the recall of the Fuzzy Proximity Information Retrieval Model (*FPIRM*) of Beigbeder & Mercier (2005), their approach is very “precise” when evaluating internationally agreed upon collections of documents used for benchmarking. The precision in this case for each query is the ratio of relevant documents retrieved over documents returned. However, the recall is weak. By recall we mean the ratio of relevant documents retrieved over all relevant documents in the collection (Rijsbergen, 1979). *FPIRM* approach evaluates the relevance between a document and a query by using a fuzzy function that takes into account the distance between the occurrences of query terms in a document. Our research studies the use of semantic query expansion to increase the recall of their model.

Justification for the Research Topic

Query expansion is a well-known technique in information retrieval used to increase recall. It involves adding new terms to the query in order to obtain relevant documents that don’t contain the terms of the initial query. Thus, query expansion can increase recall, because “there is more chance of some important words co-occurring in the query and relevant documents” (Xu & Croft, 1996). With its high precision and weak recall, *FPIRM* seems a good candidate for query expansion.

Research Questions

Different approaches are used to expand a query (Manning, Raghavan, & Schütze, 2008). Our main question in this research is to know if using ontology-based query expansion could enhance the performance of the fuzzy proximity model. From this principal question, the following spring questions are generated:

- How to choose appropriate expansion words for each query term?
- How to choose the maximum number of expansion words?
- Would employing query expansion adversely affect precision of *FPIRM* and to what extend?
- What is the cost of using ontology for query expansion?
- Is there any advantage of using generic ontology (*YAGO*[1]) instead of a lexical thesaurus (*Wordnet*[2])?

[1] YAGO (Yet Another Great Ontology) is a semantic database derived from wikipedia and wordnet in 2012. (http://www.mpi-inf.mpg.de/yago-naga/yago/index.html*).*

[2] WordNet is a lexical database of English (downloaded in 2012 from http://wordnet.princeton.edu/).

Research Methodology

Our testbed is the collection *INEX*[3] *2009*, which contains all English articles of *Wikipedia* 2008 (2’600’000 documents). These documents are annotated using *YAGO*. We are considering the use of this ontology to find new terms that are semantically related to query terms.

Research Results to Date

During our study, we’ve experimented with the use of the lexical thesaurus *WordNet*, where we looked up all synonyms of each term in the query. This trivial use of WordNet produced the query drift. This is a common issue related to query expansion and caused by the “alteration of the focus of a search topic” (Mitra, Singhal, & Buckley, 1998). Figure 1 shows how query expansion using only *WordNet* synonyms affected the performance of the fuzzy proximity model.

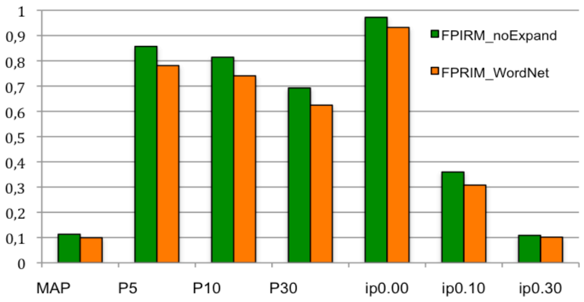

Figure 1. Proximity model without query expansion vs. WordNet based query expansion : *Mean Avarage Precision (MAP), Precision at {5,10,30}, Interpolated Recall-Precision at {0,10,30}*

References

Beigbeder, M., & Mercier, A. (2005). An information retrieval model using the fuzzy proximity degree of term occurences. *Proceedings of the 2005 ACM symposium on Applied computing - SAC ’05*, 1018. New York, USA: ACM Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval. Introduction to Information Retrieval* (pp. 177-194). Cambridge University Press.

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’98*, 206-214. New York, USA: ACM Press.

Rijsbergen, V. (1979). *INFORMATION RETRIEVAL*. Butterworth-Heinemann 1979.

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’96*, 4-11. New York, USA: ACM Press.

[3] INEX (INitiative for the Evaluation of XML retrieval) an information retrieval evaluation forum