Exploring the Use of Term Proximity in Collocate-Ranking for Query Expansion

Ying Wang *(yingwang@engmail.uwaterloo.ca)*
University of Waterloo
Olga Vechtomova
(ovechtom@engmail.uwaterloo.ca)
University of Waterloo

The exponential increase in the amount of humanities information available in digital libraries and archives calls for better search techniques that can help information users to retrieve full-text documents matching their information need with high accuracy. *Information Retrieval* (*IR*) research can lead to improved search techniques that facilitate access to large collections of humanities literature.

IR researchers focus on various topics to improve the retrieval performance, such as the representation of documents, the formulation of queries, and document matching and ranking techniques. Many established retrieval models do not take into account relations between words in text. While they work well with short and semantically homogeneous documents, arguably they are less appropriate for long multi-topic and more semantically complex texts. Many documents in the humanities archives fall under the latter category. In this paper we report a study that was conducted to develop a new query expansion technique which uses term proximity information and statistical term association measures in selecting query expansion terms.

Query expansion is a technique commonly used in IR (e.g., Rocchio; Beaulieu) to improve the retrieval performance by reformulating the original query - either adding new terms or reweighing the original terms. Query expansion terms can be automatically extracted from the documents or taken from knowledge resources, such as thesauri. The main advantage of the former techniques is that they are collection-independent and cheaper to construct. Typically either top-ranked documents in the initially retrieved document set (blind or pseudo-relevance feedback) or documents judged relevant by the user in the retrieved set (relevance feedback) are used to extract query expansion terms. For short and incomplete queries, a substantial improvement can be achieved by using expanded queries (Sparck-Jones et al.). Terms added to the original query usually have the following common characteristics: a) they are semantically related to the original query terms; b) they are good at discriminating between relevant and non-relevant documents. Computational linguists use various statistical association measures to extract significant word associations (or co-occurrences), as these measures can judge the degree of closeness between words. Association measures have been also used in query expansion to select words that are closely related to query terms (Ishikawa et al.; Vechtomova et al.).

The query expansion method proposed in this paper is used to select words which co-occur with the original query terms in a certain proximity (such as the same sentence, paragraph or a fixed-size window) in the documents judged as relevant by the user. We refer to such terms as collocates of the original query terms. We experimented with a number of parameters for selecting such terms, such as their distance from the original query term(s) in text and their degree of association with the query terms.

Previous related studies investigated the effect of using the distance information between query terms for document ranking, whereas we investigated the effect of using term proximity in collocate-ranking for query expansion. Under the similar assumption "if the distance between two words is closer, the pair is considered as more associated with each other" , we proposed to use a distance factor in collocate-ranking formula to measure the association between collocates and query terms. The main contribution of our experimentation is that we combined the distance-weighting factor with the traditional word association measure of *Mutual Information* (*MI*).

The following three hypotheses were explored in this study:

Hypothesis 1: The use of term proximity in collocate-ranking for query expansion results in a significant performance improvement over no query expansion.

Hypothesis 2: The use of term proximity in collocate-ranking formula for query expansion can lead to significant performance improvements over the current best-performing term selection values. We used *Offer Weight* (*OW*) of the Robertson/ Sparck Jones IR model as the baseline (Sparck Jones et al.).

Hypothesis 3: The collocate-ranking formula using distance information results in a significant performance improvement over the formula without the distance factor.

The experiments were conducted using the *Okapi* IR system (Sparck Jones et al.), and *TREC* (*Text REtrieval Conference*) evaluation framework (Voorhees).

The collocate-ranking method is comprised of several formulae – *Cohesion score*, *Similarity score*, *MI score* and *Distance factor* formulae. *Cohesion score* and *Similarity score* were formulated in a similar way to those proposed by Gao et al.. The cohesion score is the final score to select query expansion terms; the similarity score of a pair (x, y) is the multiplication of the *MI*

score and the *distance factor*. *MI score* was formulated similarly as the local *MI score* proposed by Vechtomova et al.. As the goal of this study was to explore the use of distance in collocate-ranking, different distance factors that might improve the retrieval performance were investigated. The best distance factor proved to be Formula 4.

The cohesion between a collocate y and query topic T is defined in Formula 1.

$$Cohesion(y,T) = \log(\sum_{x \in T} SIM(x,y)) \quad (1)$$

The similarity score of a pair (x, y) is the multiplication of the MI score and the distance factor, shown in Formula 2.

$$SIM(x,y) = MI(x,y) * df(x,y) \quad (2)$$

Where *MI(x, y)* – Mutual Information score of pair (x, y);

df(x, y) - the distance factor of pair (x, y).

MI(x, y) is calculated using frequencies from the set of relevant documents, shown in Formula 3.

$$MI(x,y) = \log_2 \frac{\frac{f_r(x,y)}{R * V_x(D)}}{\frac{f_r(x)}{R} * \frac{f_c(y)}{N}} \quad (3)$$

Where $f_r(x, y)$ - the joint frequency of pair (x, y) in the set of relevant documents;

$f_c(y)$ – frequency of y in the corpus;

$f_r(x)$ – frequency of x in the relevant documents;

$V_x(D)$ – average document length in the relevant document set;

N – corpus size;

R – size of the relevant document set (in tokens).

The best distance factor proved to be Formula 4.

$$df(x,y) = fr(x,y) / D(x,y) \quad (4)$$

Where *fr(x, y)* - the joint frequency in the relevant document set;

D(x, y) - the average distance of the pair (x, y) within documents in the relevant set.

Formulae

We performed statistical analysis of the search results produced using (a) terms selected by our experimental collocate-ranking formula (b) original query terms only and (c) query expansion terms selected by using OW. Figure 1 shows the precision values at 11-Recall levels using the above three methods. The analysis results indicate that the experimental search run using our derived collocate-ranking formula significantly improved the retrieval performance compared with the no-expansion run, but did not outperform the OW run. The method using term proximity in collocate-ranking was proved to be effective. Hypothesis 1 was supported by the analysis, while Hypotheses 2 and 3 were not supported. The top 10 query expansion terms selected by MI, the best distance formula and OW for the query topic #432 are presented in Table 1.

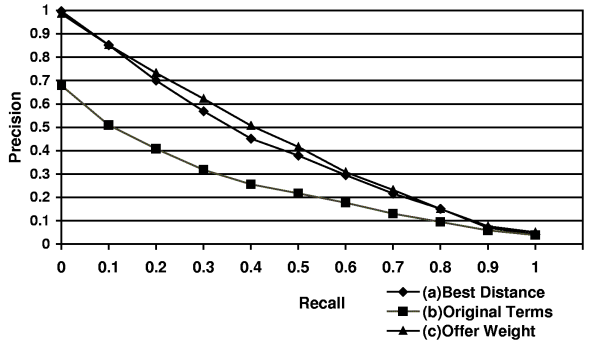

Figure 1: Precision Values at 11-Recall Levels

```
<num> Number: 432
<title> profiling, motorists, police
<desc> Description:
Do police departments use "profiling" to
stop motorists?
<narr> Narrative:
A relevant document will report or
discuss police department criteria for
identifying motorists considered likely
to be carrying contraband. Documents
discussing the detention of individuals
by foreign security forces are not
relevant.
```

Best Distance Formula	MI	OW
Jackson	non-law	sherriff
officer	Hannon	checkpoint
Hawthorne	Mirabella	drunk
checkpoint	avocation	neighborhood
black	woodyard	non-law
Dickey	Dickey	enforcement
I	remade	search
complaint	D-hayward	Hannon
search	feeler	Mirabella
department	Hardeman	15300

Table 1: The top 10 query expansion terms for sample topic #432

Some of the findings and recommendations from this study are: the distance factor has to be compatible with the collocate-extraction process and the MI score itself is an effective collocate-ranking formula compared with no query expansion. Further studies on the use of term proximity for query expansion need to be carried on through integrating other promising techniques, such as part-of-speech tagging, into the query expansion process.

This study contributes to advancing high-accuracy retrieval of documents from large resources and archives in humanities through the investigation of the role of distance and association between words in text for selecting useful terms that can be added to the search formulations and help searchers find more relevant documents. Retrieval techniques which capture relations between words in text are particularly promising for the high-precision retrieval of long multi-topic texts. Large proportion of the humanities literature consists of such texts. Query expansion techniques that assume term independence in text are less appropriate for such collections. The techniques presented in this paper can also be used in interactive query expansion. Searchers often have difficulty in formulating their information need. Previous studies showed that searchers prefer to formulate short queries and then browse through the document space and reformulate queries manually. Finding related terms that co-occur with the query terms and suggesting them to the searcher can facilitate this process.

Bibliography

Beaulieu, M. "Experiments with interfaces to support query expansion." *Journal of Documentation*, 53.1 (1997): 8-19.

Gao, J., J. Nie, H. He, W. Chen, and M Zhou. "Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations." *Proceedings of the 25th annual international ACM SIGIR conference on Research and Development in Information Retrieval.* Tampere, Finland, 2002. 183-190.

Ishikawa, K., K. Satoh, and A. Okumura. "Query term expansion based on paragraphs of the relevant document." *Proceedings of Sixth Text Retrieval Conference (TREC-6).* Gaithersburg, MD, USA, 1997. 577-584.

Rocchio, J.J. "Relevance feedback in information retrieval." *The SMART Retrieval System – experiments in automatic document processing.* Ed. G. Salton. Englewood Cliffs, New Jersey: Prentic-Hall, 1971. 312-323.

Sparck Jones, K., S. Walker, and S.E. Robertson. "A probabilistic model of information retrieval: development and comparative experiments." *Information Processing and Management* 36.6 (2000): 779-808 (Part 1); 809-840 (Part 2).

Vechtomova, O., S. Robertson, and S. Jones. "Query expansion with long-span collocates." *Information Retrieval* 6.2 (2003): 251-273.

"Overview of TREC 2003." *Proceedings of the twelfth Text Retrieval Conference (TREC 2003).* Gaithersburg, MD, USA, 2004. 1-13.