A Framework for Selective Query Expansion *

Steve Cronen-Townsend Yun Zhou W. Bruce Croft

{crotown, yzhou, croft}@cs.umass.edu
Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts, Amherst, MA 01003

ABSTRACT

Query expansion is a well-known technique that has been shown to improve *average* retrieval performance. This technique has not been used in many operational systems because of the fact that it can greatly degrade the performance of some individual queries. We show how comparison between language models of the unexpanded and expanded retrieval results can be used to predict when the expanded retrieval has strayed from the original sense of the query. In these cases, the unexpanded results are used while the expanded results are used in the remaining cases (where such straying is not detected). We evaluate this method on a wide variety of TREC collections.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Query Formulation

General Terms: Experimentation

Keywords: language modeling, clarity, query expansion

1. INTRODUCTION

In this paper we develop a method for discriminating between queries and deciding when not to use the results of an expansion technique that is likely to hurt the retrieval performance *for that particular query*. We explore our method in a language modeling framework where ordinary retrieval is done by the query likelihood method[4] and expanded retrieval is done using relevance models[1]. The retrieval parameter settings are given in [3].

2. MODEL COMPARISON METHOD

We seek to predict queries that have highly negative changes in average precision on expansion, with a score that does not depend on relevance information. To do this, we compare a language model of the unexpanded retrieval ranked list (model A) with a language model of the ranked list produced with the expanded query (model B). With this comparison, our goal is to determine when the expanded retrieval has strayed from the sense of the original query. Our model comparison scores focus on important terms in the unexpanded query and are high when the documents in the expanded results use the terms much less frequently than do the documents in the unexpanded results. This often indicates a poor expansion outcome (highly negative change in average precision). In this case the system would show the user the unexpanded retrieval results instead of the expanded retrieval results. We call this strategy *selective query expansion*. We now define each of the component of this method.

For the first component, we estimate a ranked list language model (a distribution over terms) as

$$P(w|Q) = \sum_{D \in R} P(w|D)P(\text{rank of } D|Q), \qquad (1)$$

where w is any term, D is a document, Q is the query, and R is the set of all documents, or, in practice, the retrieved set. We approximate $P(\text{rank of } D|Q)$, the probability that a document at a certain rank under Q is relevant, as query independent. For this study we used equal probabilities of relevance for the top 100 documents, and zero for all others.

Now that we have shown how to construct ranked list language models for the two ranked lists (model A and model B) the second component is the comparison. For this, we use the weighted relative entropy[5]

$$D(A||B;U) = \frac{1}{E(A;U)} \sum_{events,i} u_i a_i \log_2 \frac{a_i}{b_i}, \qquad (2)$$

where A and B represent probability distributions and U represents a vector of weights over events. The normalization factor $E(A;U) = \sum_j a_j u_j$, where a_i and b_i represents the probability of event i according to the A and B distributions, respectively. The weighted relative entropy is the expectation value of the quantity $Log_2 \frac{A}{B}$ using a weighted version of the A distribution instead of the plain A distribution as in standard relative entropy(KL). In our case, A and B are the language models for the two ranked lists, $P_A(w|Q)$ and $P_B(w|Q)$ and the events are occurrences of terms from the vocabulary of the collection.

Differences in the usage of all terms are not equally important. To reflect this we pick the top T terms in contribution to the clarity score[2] of the unexpanded model,

$$\text{contrib}(w) = P_A(w|Q) * Log_2[P_A(w|Q)/P(w)], \qquad (3)$$

where $P_A(w|Q)$ is the probability of a term w in the model and $P(w)$ is the probability of the term in the entire collection. Since these are the terms in the unexpanded model that are most unusual relative to the overall collection statistics, this forms a suitable measure of importance in the

*A full version of this paper[3] is available as `http://ciir.cs.umass.edu/pubfiles/ir-338.pdf`

CIKM'04, November 8–13, 2004, Washington, DC, USA.
ACM 1-58113-874-1/04/0011.

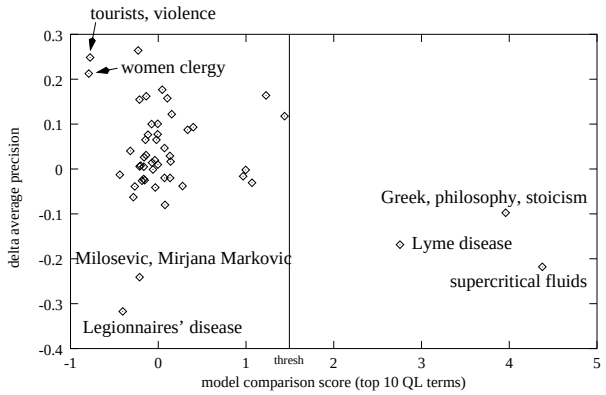

Figure 1: A scatter plot of Δ average precision and model comparison scores for TREC 8.

model. These top T terms are all given weight 1 and all other terms are given weight 0. Tests show that the method is not very sensitive to T as long as it is in the range 5 to 50. By $T = 100$ the value of the score as a predictor of expansion failures suffers noticeably.

3. RESULTS

Figure 1 shows the model comparison scores for the TREC 8 queries. The queries with significantly higher scores are all bad choices for expansion. This separation between the scores of bad-to-expand queries' and others makes automatic thresholding of these scores possible[3]. Also, the scores are easily interpreted and reflect the average usage difference of the most important terms in the unexpanded model.

Table 1 shows the mean average precision for selective query expansion ("Model Comp") compared to using consistent relevance model retrieval ("Rel Model") and correctly choosing every time the method that performs better ("Perfect Choice"). We use a 95% threshold [3] to decide not to expand a given query. The mean average precision marked "(*)" are higher than using relevance model retrieval for every query, indicating our method helps more than it hurts, on average, for these test sets. "ident" indicates score distributions too similar to allow automatic threshold setting.

Table 2 shows the performance of queries higher than the threshold in tests of our method. The threshold is set to a model comparison score that exceeds 95% of one-term queries. Above-threshold queries are divided into three performance classes: "good" where Δ, the change in average precision, is greater than 0.05, "neut" where $-0.05 < \Delta < 0.05$, and "bad" where $\Delta < -0.05$. The method is successful in collections above the line and unsuccessful in collections below the line. We do not expect to see large *mean* average precision improvements since the method is tuned to detect a small percentage of queries that perform very poorly on expansion. For certain collections (e.g. TREC 6 and TREC 8), the effect of the method with this high automatic thresholding is quite good: all the queries above threshold are indeed bad to expand or neutral, hence some poor expansions are avoided, and more consistency is obtained.

4. CONCLUSIONS

We present a method for improving the consistency of retrieval results through automatically chosing among competing retrieval techniques. The method can compare models of the *ranked list* of documents from any two retrieval techniques, whether they are based on language modeling or not. The method measures when a new ranked list of documents (e.g. from expanded retrieval or feedback) has strayed significantly from the usage of important terms in an original ranked list of documents (e.g. from unexpanded retrieval). In these cases, *not* using the expansion results in will usually avoid one type of expansion failure by sensing that something has gone wrong. Our work provides the first steps toward solving a difficult, but very important, problem. We suggest one meaningful criterion that systems may use to help avoid showing users the results of techniques that may hurt performance for a particular query.

TREC	Queries	Rel Model	Model Comp	Perfect Choice
1+2+3	51-150 title	0.2490	ident	0.2589
5	251-300 title	0.1609	0.1644(∗)	0.1837
6	301-350 title	0.2013	0.2115(∗)	0.2468
7	351-400 title	0.2524	0.2212	0.2711
8	401-450 title	0.2715	0.2756(∗)	0.3011
Agg QT	51-100: 1804	0.2219	0.2188	0.2387

Table 1: Selective mean average precision with estimated Bayes optimum thresholds.

Collection	Rel Model	Model Comp	Above Threshold		
			good	neut	bad
TREC 5	0.1609	0.1621	0	2	1
TREC 6	0.2013	0.2197	0	3	3
TREC 8	0.2715	0.2812	0	0	3
TREC 1+2+3	0.2490	0.2451	1	0	0
TREC 7	0.2524	0.2394	2	1	1
QT agg	0.2219	0.2217	13	32	11

Table 2: Breakdown of above-threshold query performance for selective query expansion.

5. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by SPAWARSYSCEN-SD grant number N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

6. REFERENCES

[1] W. B. Croft and J. Lafferty, editors. *Language Modeling for Information Retrieval*, pages 11–56. Kluwer, 2003.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306, August 2002.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A language modeling framework for selective query expansion. Technical Report IR-338, Center for Intelligent Information Retrieval, University of Massachusetts, 2004.

[4] F. Song and W. B. Croft. A general language model for informatioon retrieval. In *Proceedings of the 22nd ACM SIGIR conference*, pages 279–280, 1999.

[5] H. C. Taneja and R. K. Tuteja. Characterization of a quantitative-qualitative measure of relative information. *Information Sciences*, 33:217–222, 1984.