Enterprise Search: Identifying Relevant Sentences and using them for Query Expansion

Maheedhar Kolla
University of Waterloo
Canada. N2L 3G1
mkolla@uwaterloo.ca

Olga Vechtomova
University of Waterloo
Canada. N2L 3G1
ovechtom@uwaterloo.ca

ABSTRACT

In this paper, we discuss the experiments conducted in context of Document Search task of 2007 Enterprise Search track. Our method is based on selecting sentences from the given relevant documents and using those selected sentences for query expansion. We observed that our method of query expansion improves system's performance over baseline run, under various methods of comparison.

1. INTRODUCTION

Enterprise Search track, organized by NIST, is aimed at studying information search problems faced by employees (mostly) in an organization. In 2007, two tasks were proposed in this track:

- Document Search
- Expert (People) Search.

These tasks simulated problems faced by Science Communicators of the (CSIRO) organization. As their work task, communicators are required to create an *overview* page for a given topic. Systems' task is therefore defined as to retrieve documents that would help such science communicators in creating such an overview page. Sample topic, composed by the communicators, is as follows:

> **query**: hairpin RNAi/gene silencing
> **narr**: Information to help scientists find out more about hairpin RNAi technology. Specific contacts to obtain vectors.
> **page**: CSIRO197-05231046
> **page**: CSIRO139-13111797

As observed, along with *query* and *narrative*, each topic consists of list of documents (under *page* fields), tagged relevant by topic creators. These documents could be used for relevance feedback runs.

To appear in 16th Text Retrieval Conference proceedings, Gaithersburg Maryland, USA,2007.

Our interest in this year's participation is to study the term expansion from the relevant documents provided. We propose a method to rank sentences in the relevant documents and then extract terms from top ranked sentences of each relevant document.

2. TERM EXPANSION

We hypothesize that terms extracted from summary of a document would enable us to capture the intent of an user (who judged the relevance of the document). We propose a method of sentence selection that would identify sentences closer to topic model than general English model. We built topic relevance model using the top 20 documents retrieved for initial query, as proposed by Lavrenko and Croft [3].

Using both, relevance and generic collection models, we rank the sentences within each document based on Kullback-Leibler Divergence (KLD) measure. KLD value of a sentence is computed as follows:

$$KLD(rm, cm) = \sum_{t \epsilon S} p_{rm}(t) * log \frac{p_{rm}(t)}{p_{cm}t} \quad (1)$$

where p_{rm} is the probability of a term in relevance model, p_{cm} is the probability of the term in general collection model and t represents terms belonging to a sentence S. KLD measure in our method would indicate the "goodness" of a sentence towards topic model than collection model . Each sentence would obtain "goodness" value displaying its closeness towards language model built from relevant documents as supposed to general English language model. Büttcher.et.al [1] used similar measure to re-rank documents initially retrieved for a given query.

We then rank sentences based on their KLD values and extract top ranking sentence from each related article. We pool these top ranked sentence extracted from each related article and extract terms for query expansion.

3. EVALUATION

Documents retrieved were evaluated on a three point scale: non-relevant (0), relevant (1) and highly relevant (2), where highly relevance are documents that would help the science communicators create an *overview* page. The following runs are were compared in our experiments:

- uwBase - baseline Okapi BM25 ranking
- uwKLd - pseudo-relevance feedback run, with term extracted based in KL Divergence [2]

- uwRF - terms extracted from top ranked sentences, selected from documents under *pages* fields

Run	MAP	P@5	P@10	ndcg1000
uwbase	0.3878	0.6520	0.5780	0.7071
uwKld	0.3877	0.6160	0.5380	0.7335
uwRF	0.4299	0.7040	0.6220	0.7557

Table 1: Standard Measures: Comparison of systems (without altering any runs)

In Table 1, we present the standard results of evaluation for all three runs. Only documents judged with relevance value of 2 are considered relevant in these evaluation. However, as mentioned in guidelines, it would be unfair to compare relevance feedback runs with pseudo-relevant feedback and/or baseline runs directly — relevance feedback runs might just rank those documents provided under *pages* fields than identifying different relevant documents. NIST proposed different means to compare relevant feedback runs directly with baseline and pseudo-relevant feedback runs.

3.1 Promotion Evaluation

In this mode of comparison, all documents provided by topic creators under *pages* fields are ranked at the top of each system's run. Through this method of comparison, we could investigate the extent to which any relevance feedback method would actually find relevant documents instead of just *promoting* already known relevant documents. Evaluation results are presented in Table 2.

Run	MAP	P@5	P@10	ndcg1000
uwbase.p	0.4698	0.8720	0.7080	0.7506
uwKld.p	0.4663	0.8360	0.6860	0.7697
uwRF.p	0.5004	0.8920	0.7160	0.7871

Table 2: Promotion Evaluation: Comparison of systems in which all documents provided in *pages* fields are placed at the top of each run.

3.2 Residual Collection

In residual collection method of evaluation, all provided relevant documents(i.e documents from *pages* fields) are removed from evaluation process. Systems are then compared based on their ability on retrieving all possible relevant documents from residual collection. Table 3 presents evaluation results for all three systems under this approach.

Run	MAP	P@5	P@10	ndcg1000
uwbase.rc	0.3607	0.6040	0.5280	0.6466
uwKld.rc	0.3567	0.5680	0.4900	0.6596
uwRF.rc	0.3946	0.6360	0.5600	0.6910

Table 3: Residual Collection Evaluation: Comparison of systems in which documents provided under *pages* fields are not included in evaluation.

4. DISCUSSION AND FUTURE WORK

As observed from evaluation results, relevance feedback run improves over baseline run and performs better than pseudo-relevant feedback runs in all three methods of comparison (although not statistically significant). We would like to now compare different sentence selection methods with our sentence selection method. We would also attempt to replicate similar experiments on previous Enterprise Search test data to test the efficiency of our term extraction method on tasks such as Email search, Known-item search.

5. REFERENCES

[1] S. Büttcher, C. L. A. Clarke, and P. C. K. Yeung. Index pruning and result reranking: Effects on adhoc retrieval and named page finding. In *15th Text REtrieval Conference*, Gaithesburg, USA, (2006).

[2] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Trans. Inf. Syst.*, 19(1):1–27, 2001.

[3] V. Lavrenko, M. Choquette, and W. B. Croft. Cross-lingual relevance models. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 175–182, New York, NY, USA, 2002. ACM Press.