Exploration of Query Context for Information Retrieval

Keke Cai, Chun Chen*, Jiajun Bu, Peng Huang, Zhiming Kang
College of Computer Science, Zhejiang University
Hangzhou, 310027, China
*Corresponding Author, +86 571 8795 1431
{caikeke, chenc, bjj, huangp, kzm}@zju.edu.cn

ABSTRACT
A number of existing information retrieval systems propose the notion of query context to combine the knowledge of query and user into retrieval to reveal the most exact description of user's information needs. In this paper we interpret query context as a document consisting of sentences related to the current query. This kind of query context is used to re-estimate the relevance probabilities of top-ranked documents and then re-rank top-ranked documents. The experiments show that the proposed context-based approach for information retrieval can greatly improved relevance of search results.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models.*

General Terms
Algorithms, Design, Performance, Experimentation

Keywords
Query context, information retrieval, document re-ranking.

1. INTRODUCTION
The notion of query context has been widely mentioned in many recent studies of information retrieval. The purpose is to use a variety of knowledge involving query and user to explore the most exact understanding of user's information needs [1]. It has been argued that context-based information retrieval is promising to achieve considerable improvements on retrieval performance.

Different notions of query context have been used in many retrieval systems. In this paper, we interpret query context from a different perspective. Studies in [2] point out that user' information needs can be potentially described as an ideal document, from which one or more significant terms are generalized to form the query. Inspired by this idea, this paper defines query context in a more intuitive way, namely, as a document exhibiting most significant features about query.

Many previous applications of query context focus on query reformulation. Query-relevant terms extracted from query context are used to reformulate the initial query. By contrast, we do not apply query context in this way. We use the context information of query to re-estimate the relevance of the top-ranked documents for re-ranking. In the implementation, the matching probability is realized by measuring the similarity between query context and each top document in initial retrieval list. There are previous works on modeling the explicit context in the retrieval model, such as [7] that focus on modeling contexts of query terms by local surrounding terms in a document. In this paper, instead of using the simple definition of local context of query, we explore query context from different points of view.

Query context defined in this paper is in the form of document. As illustrated in [6], each document can be ultimately expressed as a grouping of sentences. From this perspective, our explanation of query context can be realized by identifying relevant sentences given a query. For this, three approaches for sentence retrieval are proposed, in which three kinds of context information are considered: query type context, query dependency context and query occurrence context. Query type context describes the necessary information types involved in relevant documents, query dependency context reveals the dependency relationships among query terms, and query occurrence context presents the relationships between query terms and other terms in occurrence. Sentences identified by these three sentence retrieval approaches are ultimately combined into a single ranked list, from which top-ranked sentences are selected and constitute the context of query.

2. INFORMATION RETRIEVAL MODEL BASED ON QUERY CONTEXT
Given a query, retrieval process of our proposed context-based retrieval model consists of the following steps:

(1) Implement the initial retrieval. In this paper, it is realized by using KL_divergence retrieval model [5].

(2) Construct query-relevant document collection, or *QRD*, by selecting the top-ranked documents in the initial retrieval list.

(3) Identify query-relevant sentences of documents in *QRD*. For this, three sentence retrieval approaches are implemented, i.e., query type based sentence retrieval, Markov Random Field (MRF) [3] based sentence retrieval and Bayesian network based sentence retrieval. a) Query type based sentence retrieval explores the associations between query and query type and explores sentences involved with information of the expected type; b) MRF based sentence retrieval utilizes association features between query terms in sentence to discover sentences reflecting exact relationships among query terms; c) Bayesian network based sentence retrieval utilizes the inference ability of Bayesian network to identify sentences semantically related to query. Results of each respective sentence retrieval approach are merged and top-ranked sentences are selected as query-relevant sentences, or query context.

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

(4) Re-rank documents in initial retrieval list. Given the query context *QC* identified above, each document *D* in initial retrieval list is re-ranked according to its translation probability to *QC*. The translation-based approach for document relevance evaluation is inspired by the idea of example-based bi-lingual translation [4]. Formally, the translation is formulated as:

$$relevance(D,QC) = \prod_{cs_i} \sum_{ds_j} P(cs_i \mid ds_j) * P(ds_j \mid D) \qquad (1)$$

where cs_i and ds_j are respectively sentence contained in *QC* and *D*. $P(cs_i|ds_j)$ is the translation probability from ds_j to cs_j and is measured by the similarity between ds_i and cs_j. Here, we adapt the measure of edit-distance to obtain this similarity. $P(ds_j|D)$ means the generation probability of ds_j given *D* and is defined as the product of the distribution probability of each term of ds_j in *D*.

3. EXPERIMENTS

Table 1. Performances of MRR and P@k on topics 301-350

	BASE	OCR
MRR	0.533	0.638
P@5	0.328	0.487
P@10	0.316	0.439
P@20	0.263	0.388

Table 2. Performances of MRR and P@k on topics 351-400

	BASE	OCR
MRR	0.556	0.698
P@5	0.404	0.532
P@10	0.374	0.503
P@20	0.318	0.461

Table 3. Performances of MRR and P@k on topics 401-450

	BASE	OCR
MRR	0.468	0.581
P@5	0.312	0.422
P@10	0.270	0.408
P@20	0.220	0.365

We use TREC disks 4 and 5 in the experiments and evaluate the proposed retrieval approach on ad hoc tasks for TREC6 with topics 301-350, TREC7 with topics 351-400 and TREC8 with topics 401-450. Only the title portions of these topics are used as experimental queries. Relevance of the retrieved documents is assessed by the relevance assessments provided by TREC. In our experiments, the initial (baseline) retrieval is realized by using KL_divergence retrieval approach with Dirichlet smoothing and pseudo-relevance feedback.

Table 1-3 shows the comparison results between the baseline retrieval (BASE) and the proposed context-based retrieval (QCR). QCR places more emphasis on precision in top-ranked documents, hence in our experiments the retrieval is mainly measured by using metrics MRR and P@*k*, for *k* = 5, 10, 20. MRR is defined as the reciprocal of the first relevant document's rank in the ranked list averaged over all queries. P@*k* is defined as the precision at the top *k* sentences averaged over all queries. As can be seen from Table 1-3, performances of QCR are consistently better than that of BASE. MRR are respectively improved by 19.7%, 25.5% and 24.1% and P@*k* achieves better value at each point of *k*. The comparison results prove the efficiency of sentence-based query context in the application of context-based retrieval and its particular performance in improving retrieval precision.

4. CONCLUSIONS AND FUTURE WORK

We propose a novel context-based retrieval approach for document re-ranking. Query context is interpreted by considering various features of query and can provide information for user's query disambiguation. Experiments prove the efficiency of the proposed retrieval approach. In this paper, we have only explored the edit distance for evaluating the similarity between two sentences. More sophisticated analysis is interesting to explore.

5. REFERENCES

[1] Allan, J. *et al.* Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002. *SIGIR Forum*, 37, 1 (2003), 31-47.

[2] Berger, A., Lafferty, J. Information Retrieval as Statistical Translation. In *Proceedings of SIGIR'99* (Berkeley, CA, USA, Aug. 15-19, 1999). ACM Press, New York, NY, 1999, 222-229.

[3] Dobrushin, R. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probability and its Applications*. 13 (1968) 197-224.

[4] Doi, T., Yamamoto, H., and Sumita, E. Example-based machine translation using efficient sentence retrieval based on edit-distance. *ACM Transactions on Asian Language Information Processing*, 4, 4 (Dec. 2005), 377-399.

[5] Lafferty, J., and Zhai, C. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of SIGIR'01*(New Orleans, LA, USA, Sep. 9-13, 2001). ACM Press, New York, NY, 2001, 111-119.

[6] Nallapati, N., and Allan, J. Capturing term dependencies using a sentence tree based language model. In *Proceedings of CIKM'02* (McLean, VA, USA, Nov. 4-9, 2002) ACM Press, New York, NY, 2002, 383-390.

[7] Wu, H., Luk, R., Wong, K., Kwok, K. Probabilistic document-context based relevance feedback with limited relevance judgments. In *Proceedings of CIKM'06* (Arlington, VA, USA, Nov. 6-11, 2006) ACM Press, New York, NY, 2002, 854-855