A Retrospective Study of Probabilistic Context-Based Retrieval

H. C. Wu R. W. P. Luk K. F. Wong[1] K. L. Kwok[2] W. J. Li

Department of Computing, The Hong Kong Polytechnic University
[1]Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong
[2]Department of Computer Science, Queen's College, City University of New York
{cshcwu, csrluk, cswjli}@comp.polyu.edu.hk, kfwong@se.cuhk.edu.hk, kwok@cs.qc.edu

ABSTRACT

We propose a novel probabilistic retrieval model which weights terms according to their contexts in documents. The term weighting function of our model is similar to the language model and the binary independence model. The retrospective experiments (i.e., relevance information is present) illustrate the potential of our probabilistic context-based retrieval where the precision at the top 30 documents is about 43% for TREC-6 data and 52% for TREC-7 data.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

General Terms

Experimentation, Performance

Keywords

Context, Retrospective Experiment

1. INTRODUCTION

We propose a novel model that utilizes the context information to compute the term weight at each location in the document of the matched search term (or query term). The term weight is calculated by multiplying probabilities similar to the well-known probabilistic models (i.e., binary independence model [1]) and language model (e.g., [2]). This paper focuses on whether the use of context information can enhance retrieval effectiveness in retrospective experiments (that use the statistics of relevance information similar to the *w4* term weight [1], the ratio of relevance odds and irrelevance odds). If there is a significant effectiveness enhancement, then the future research question (not addressed here) is how to obtain the effectiveness in predictive experiments that are close to our retrospective experiments. For valid comparisons, *w4* weight [1] is used because it requires relevance information and its effectiveness is known to be good.

There are research works (e.g., [3]) similar to ours in which the score of every location in the document of the search term contributes differently to the document similarity. By contrast, apart from incorporating the search term occurrences in the document for ranking, our score of every location in the document is determined by the terms located nearby the search term and by the relative location of these terms to the search term. Likewise, the passage retrieval (e.g., [4]) also treats the document as individual pieces. Unlike some passage retrieval, our model does not assign uniform scores to each search term occurrences in the passage. Finally, other studies (e.g., [5]) aggregate scores of the selected terms in the nearby context of the matched search term where the selected terms were found to have significant relations with the search term beforehand. By contrast, our model combines the scores of all terms in the context without identifying significant relations.

2. OUR MODEL

2.1 Overview

There are a variety of formulae for calculating the term weights [6]. One of the most widely used term weights is known as the inverse document frequency (IDF) which was proposed in [7] and combines with the term frequency (TF) to form the well-known TF-IDF function. In order to investigate the contribution of context information in term weighting, we consider the occurrences of the location-specific terms when calculating term weights.

Before further discussions, the meaning of a context should be clarified. Firstly, the position of each search term occurring in the document is located. Then the terms before and after the search term are considered to be the context of that search term. Hence, each context has a search term in the middle. This is different from passage retrieval in which the passages are defined by document structures or text blocks at regular intervals.

The first reason for having the context is that our model is simulating a user making relevance judgment using keywords in context [8] as in [9]. The other reason is that a term with its context has a definite meaning and it becomes apparent whether the matched term has a similar meaning to the search term's meaning. However, terms without contexts can be ambiguous.

2.2 Calculation of Term Weight

Each context $c(s_{i,k}, n)$ has a search term $s_{i,k}$ which is a query term and appears at the *k*-th location in the *i*-th document with context size *n*. The context $c(s_{i,k}, n)$ contains terms $\{t(i, k+p)\}$ which occur at the *p*-th location relative to $s_{i,k}$ in the *i*-th document where $p \in [-0.5n, 0.5n]$. We define the probabilities of which a context $c(s_{i,k}, n)$ is relevant and irrelevant by the following equations:

$$P(c(s_{i,k}, n) \mid relevant) = \prod_{p=-0.5n}^{0.5n} P(t(i, k+p) = \omega \mid relevant)$$

$$P(c(s_{i,k}, n) \mid irrelevant) = \prod_{p=-0.5n}^{0.5n} P(t(i, k+p) = \omega \mid irrelevant)$$

similar to the language model [2] where:

SIGIR'05, August 15─19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

$$P(\omega \mid relevant) = \frac{\#\ occurences\ of\ \ \omega\ in\ relevant\ documents}{\#\ occurences\ of\ \ all\ terms\ in\ relevant\ documents}$$

$$P(\omega \mid irrelevant) = \frac{\#\ occurences\ of\ \ \omega\ in\ irrelevant\ documents}{\#\ occurences\ of\ \ all\ terms\ in\ irrelevant\ documents}$$

are relative frequency estimates of the probabilities of the term ω which occurs in the relevant documents and irrelevant documents, respectively.

The term weight of each search term in the document depends on its context and is defined by the odds:

$$\frac{P(relevant \mid c(s_{i,k},n))}{P(irrelevant \mid c(s_{i,k},n))} = \frac{P(c(s_{i,k},n) \mid relevant)}{P(c(s_{i,k},n) \mid irrelevant)} \times \frac{P(relevant)}{P(irrelevant)}$$

using Bayes' rule. Since *P*(*relevant*) and *P*(*irrelevant*) are constants, their ratio is a constant and hence this ratio can be discarded during ranking. Finally, the term weight $w(s_{i,k})$ for the search term $s_{i,k}$ is the log-odds ratio (similar to [1]):

$$w(s_{i,k}) = \log\left(\frac{P(c(s_{i,k},n) \mid relevant)}{P(c(s_{i,k},n) \mid irrelevant)}\right)$$

According to the disjunctive relevance decision principle [9], we define the similarity between query q and document d_i by picking the highest as the representative score:

$$sim(d_i, q) = \max_{s_{i,k} \in q} \{w(s_{i,k})\}$$

in order to avoid spurious matching of search terms.

3. EXPERIMENT

The experiment compares the context-based score mentioned above and the *w4* weighting using the TREC-6 and -7 collections for ad hoc retrieval with 50 title queries each. Initially, we used the BM11 weights of the 2-Poisson model [10] with blind feedback in a predictive experiment. The results in Table 1 are our baseline. We report the precision at top N documents (P@N), R-Precision and the mean average precision (MAP).

Table 1: Effectiveness of our BM11 weighting

	P@10	P@30	R-Precision	MAP
TREC-6	.416	.320	.288	.261
TREC-7	.402	.290	.244	.208

Table 2: Effectiveness of *w4* and our context-based weighting

	P@10		P@30		R-Precision		MAP	
T	*w4*	ours	*w4*	ours	*w4*	ours	*w4*	ours
6	.482	.540	.355	.437	.309	.397	.286	.363
7	.478	.644	.355	.529	.280	.388	.254	.350

Next, we modified the BM11 weight by substituting the *w4* weighting into the IDF factor. For our model, the best context size should be empirically determined and it is set to 100 here as in [11] for simplicity. Table 2 shows the results of the retrospective experiment that compares the *w4* weighting effectiveness and the context-based weighting effectiveness. When the relevance judgments are present, the performance is expected to be increased. As our proposed weighting scheme depends on location-specific information of search terms, every search term weight would be different for different contexts. The results showed that the precision of our weighting scheme is highly promising (c.f. [12]). The evidence suggests that context information is important for information retrieval.

4. FUTURE WORK

Our experiment results show that theoretically applying the context information could improve the performance for re-ranking documents. In this paper, experiments are done retrospectively and the results are promising. We would further investigate how to estimate the probabilities without relevance information for our model to operate in predictive experiments

5. ACKNOWLEDGEMENTS

This work is supported by PolyU 5183/03E and partially supported by the CUHK strategic grant initiative (account #4410001). The basic retrieval system was developed when Robert was on leave at UMASS.

6. REFERENCES

[1] S.E. Robertson and K. Sparck Jones, Relevance Weighting of Search Terms, *Journal of the American Society for Information Science*, 27, 129-146, 1976

[2] J. F. Ponte and W. B. Croft, A Language Modeling Approach for Information Retrieval, *Proc. ACM SIGIR Conference*, pp. 275-281, 1998

[3] O. de Kretser and A. Moffat, Effective Document Presentation with Locality-Based Similarity Heuristic, *Proc. ACM SIGIR Conference*, pp. 113-120, 1999

[4] M. Kaszkiel, J. Zobel and R. Sacks-Davis, Efficient Passage Ranking for Document Databases, *ACM TOIS*, 17(4), 406-439, 1999

[5] P. Bruza and D. Song, A comparison of various approaches for using probabilistic dependencies in language modeling, *Proc. ACM SIGIR Conference*, pp. 219-420, 2003

[6] J. Zobel and A. Moffat, Exploring the Similarity Space, *ACM SIGIR Forum*, 32(1), 18-34, 1998

[7] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, 28, 11-21, 1972

[8] J. Kupiec, J. Pedersen, and F. Chen, A Trainable Document Summarizer, *Proc. ACM SIGIR Conference*, pp. 68-73, 1995

[9] Y. K. Kwong, R.W.P. Luk, W. Lam, K.S. Ho and F.L. Chung, Passage-based retrieval based on parameterized fuzzy operators, *ACM SIGIR Workshop on Mathematical/Formal Methods for Information Retrieva*l, 2004

[10] S. E. Robertson and S. Walker, Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval, *Proc. ACM SIGIR Conference*, pp.232-241, 1992

[11] C.L.A. Clarke and E.L. Terra, Passage retrieval vs document retrieval for factoid question. *Proc. ACM SIGIR Conference*, pp. 427-428, 2003

[12] K. Sparck Jones, Summary Performance Comparisons TREC-2 Through TREC-7, *Proc. of TREC-7 Conference*, pp. B-1, 1998