A Practical Passage-based Approach for Chinese Document Retrieval

Szu-Yuan Chi[1], Chung-Li Hsiao[1], Lee-Feng Chien[1,2]

1. Department of Information Management, National Taiwan University
2. Institute of Information Science, Academia Sinica, Taipei, Taiwan

{r93036, r93042}@im.ntu.edu.tw and *lfchien@iis.sinica.edu.tw*

Abstract. TF*IDF-based methods, as they are easily to implement, are widely accepted in information retrieval industry. It is interesting to investigate a feasible and practical technique to improve the retrieval performance of these conventional IR methods. In this paper, we are going to introduce a good alternative approach that uses passage-based ranking as the second stage of the retrieval process in them.

1. Introduction

Previous research has shown that passage-level evidences can bring additional benefits to document retrieval when documents are long or span different subject areas. Callan (1994) addressed that the types of passages explored by researchers can be grouped into three classes: discourse, semantic, and window. Discourse passages are based upon textual discourse units (e.g. sentences, paragraphs and sections). Semantic passages are based upon the subject or content of the text. Window passages are based upon a number of words. Kaszkiel & Zobel (1997, 2001) also investigated the effectiveness on using passages for information retrieval. Their research showed that passages can be used in different ways. One is to provide a good basis for a question-and-answer style of information retrieval. Another approach is to use passages as proxies for documents, and documents are ranked according to similarities computed for their passages. It also addressed that fixed-length arbitrary passages of 150 words or more and starting at small interval so that the passages heavily overlap can give substantial empowerments in effectiveness, particularly for collections of long documents.

Recently Liu (2002) further demonstrated that passages can be used effectively in a language modeling framework. They found passage retrieval based on language models can provide more reliable performance than retrieval based on full documents. The previous works have proven that passage-based retrieval can get better performance than document-based ones in different applications. It's unfortunately that conventional passage-based retrieval most calculates documents ranks with the aids of passage-level indexing and single-stage processing. Although this can reduce the computing cost of passage-based retrieval, it is not flexible to consider the contextual effects of matched query terms in a passage and determine an appropriate weighting scheme through the access of indices. In some cases, it needs to analyze the content of document, for example, to determine if the occurrences of the matched query terms are appearing in a critical passage. Our research on passage retrieval is just at the beginning. Our long term research goal is to adopt sophisticated text analysis in combination with index-based ranking schemes to reach a balance between retrieval speed and accuracy.

The purpose of this paper aims at developing a practical approach to improving conventional TF*IDF IR methods, without the involvement of using some sophisticated techniques such as query expansion and ontology-based ranking. The goal is not trivial. As users' queries are often short, only a few conventional IR systems provide query expansion.

It's getting popular to improve performance by using a 2-stage strategy in retrieval task (Kwok et la., 1998). Nevertheless, most of the researches used pseudo-relevance feedback as the 2nd stage. Unlike them, the proposed approach is a two-stage retrieval process that uses passage-level analysis as the second stage in the retrieval process of conventional TF*IDF-based methods. The first stage utilizes an Okapi-based ranking algorithm to retrieve top-n relevant documents as an initial set. The passage features are then used in the 2nd stage to try to filter out irrelevant ones from the set. The proposed approach has been tested with the Chinese monolingual IR task of NTCIR-4 (Kando, 2004). The obtained preliminary result shows that the Okapi-based approach can be improved using the two-stage

process and passage-based ranking. Though the archived performance is close to NTCIR4 participants' average, the proposed approach is believed easier to be applied in commercial applications.

2. Related Work on NCTIR-4 Experiments

In this section, we will review the performance of Chinese-to-Chinese (C-C) monolingual runs achieved by NTCIR-4 participants. Table 1 shows the obtained average, median, maximum and minimum values of MAP by type of run based on rigid relevance. We use following notations:

C-C-T: all C-C <TITLE>-only runs (T-runs)
C-C-D: all C-C <DESC>-only runs (D-runs)

Table 2 shows the top 5 groups tanked according to MAP values of D-runs based on rigid relevance. I2R-C-C-D-01 which was based on ontological query expansion achieved the best performance. The research work of the top three groups is briefly summarized.

Table 1 MAP of overall C-C runs

	Average	Median	Min	Max
C-C-T	0.1943	0.1881	0.1327	0.3146
C-C-D	0.1826	0.1741	0.1251	0.3255

Table 2 Top-ranked 5 groups (C-C, Rigid, D-runs)

Run-ID	Mean Average Precision
I2R-C-C-D-01	0.3255
OKI-C-C-D-04	0.2274
Pircs-C-C-D-02	0.2150
RCUNA-C-C-D-01	0.2087
UniNE-C-C-D-03	0.2011

I2R: Using knowledge ontology. (Yang et al., 2004)

The I2R group has built knowledge ontology for query terms by using a search engine on the Internet with manual verification. Firstly, they automatically extract terms from documents and use them to build indexes; secondly, they use short terms in the query and documents to do initial retrieval; thirdly, they build ontology for the query to do query expansion and implement second retrieval. Finally, they use long terms to reorder the top N retrieved documents. The knowledge ontology appears to include narrower terms, related terms and so on. They combine information from the ontology with that from pseudo-relevance-feedback to expand query terms.

OKI: (Nakagawa and Kitamur, 2004)

As widely known, pseudo-relevance-feedback (PRF) of blind feedback brings us improvement or retrieval performance. Some research groups, however, challenge to use **non-standard PRF** methods. For example, the OKI group adopts Ponte's ration method (Ponte, 1998).

PIRCS: (Kwok et al, 2004)

For PIRCS group, Chinese monolingual retrieval was performed as before (Kwok, 2002): based on combination of retrieval lists using bigram + unigram, and short word + character indexing.

3. Two-stage Document Retrieval

In our research, we pursue a simpler approach that can achieve acceptable performance. We propose a two-stage passage-based document retrieval approach. The retrieval process performed at the first stage is similar to that in conventional n-gram-based Chinese IR systems. That is, all unique character unigrams and bigrams in a document except some stop characters will be extracted to form the term vector of the document and a TF*IDF-based weight value is assigned to each term as its significance value. In addition, an Okapi-based similarity estimation function (Robertson et al., 1994)

is adopted to estimate the relevance score between the input query and each indexed document. As any two feature dimensions of a vector-space model are assumed independent, at the first stage it doesn't consider the contextual effects between the query terms and the document of concern.

The second-stage process is proposed as an additional retrieval process performing detailed passage analysis. As shown in previous research (Callan, 1994; Kaszkiel & Zobel, 1997), passage-level evidences can bring additional benefits to document retrieval when documents are long or span different subject areas. The additional process re-examines the occurrences of the query terms and analyzes their presences at the passage level of a document. The addition of the second-stage process is tried to investigate if there is a simple approach to improve conventional document retrieval methods, without the involvement of sophisticated techniques, such as query expansion and ontology-based ranking.

An overview of the two-stage document retrieval approach is shown in Figure 1, in which some techniques and strategies which are planned to be tested are listed.

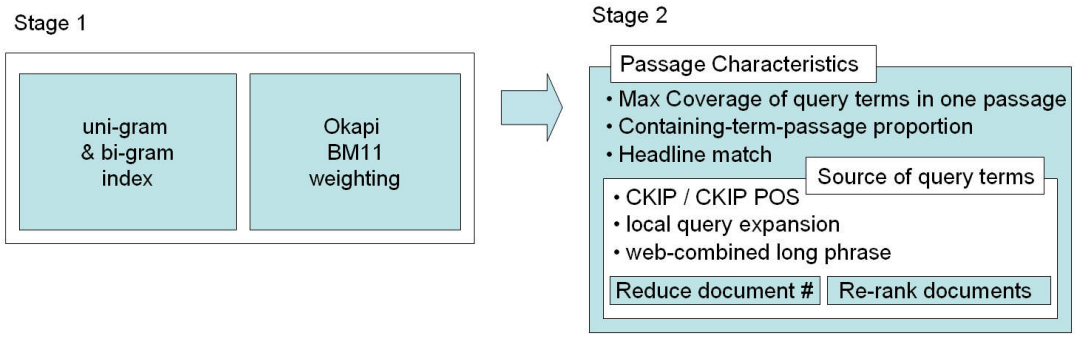

Figure 1: An overview of the two-stage document retrieval approach.

3.1 The Stage One Process

As described previously the purpose of the first-stage processing is to form uni-gram and bi-gram feature vectors for input query and all documents, and an Okapi similarity estimation algorithm, i.e., BM11 defined below, is adopted to retrieve and rank these documents. For more information about BM11 can be referred to (Robertson et al., 1994).

$$\text{(BM11)} \qquad w = \frac{tf}{(\frac{k_1 \times d}{\Delta} + tf)} \times \log \frac{N - n + 0.5}{n + 0.5} \times \frac{qtf}{(k_3 + qtf)} + k_2 \times nq \frac{(\Delta - d)}{(\Delta + d)}$$

N: Number of items (documents) in the collection
n: Collection frequency: number of items containing a specific term
tf: Frequency of occurrence of the term within a specific document
qtf: Frequency of occurrence of the term within a specific query
d: Document length arbitrary units
Δ : Average document length
k_i: Constants used in various BM functions

To realize the achieved performance, our research was conducted based on the Chinese monolingual IR task of NTCIR4 (Kishida, et al, 2004). We tested the Okapi-based approach (the first-stage processing) and the obtained MAP (Mean Average Precision) value was about 0.18, which is close to NTCIR4 participants' average and is thus taken as the baseline. The obtained MAP value for each test topic is shown in Figure 2; and as in Figure 3, it was found that for most test topics the obtained recall rates of top 10000 retrieved documents are very high and almost close to 1. The second-stage process is, therefore base on the answer set (retrieved documents) to re-rank some relevant documents to higher position.

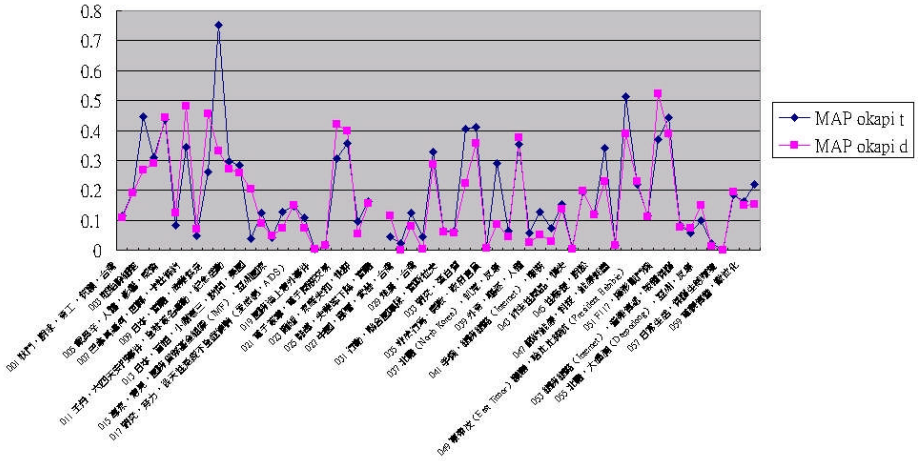

Figure 2: Obtained MAP values of top 10,000 retrieved documents using Okapi BM11 in NTCIR4, in which *t* means the result of title run and *d* is that of description run.

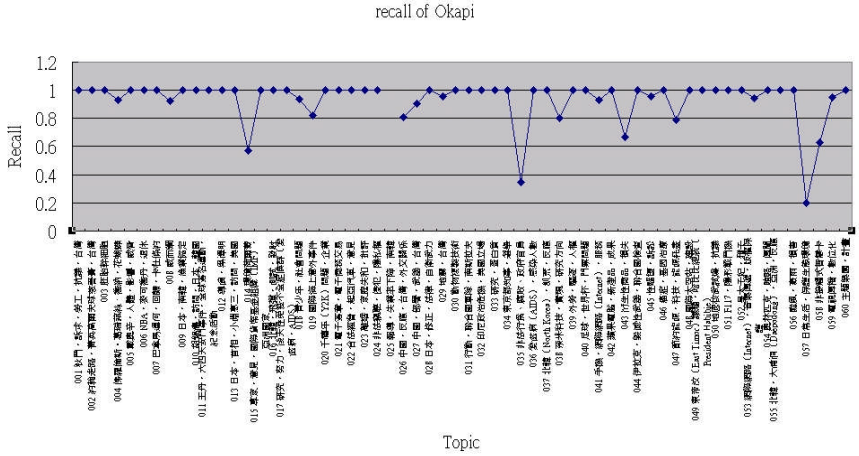

Figure 3: Obtained recall rates of top 10,000 retrieved documents using Okapi BM11 in NTCIR4, which are with respect to each test topic.

3.2 Observation after Stage One Processing

As discussed conventional Okapi algorithms treat term as independent entities and ignore semantic or locality properties of terms in documents. According to the answer set, we observed a few interesting properties of passages that can be further analyzed.

We examined the answer sets with four types of relevance in answers: highly relevant (S), relevant (A), partial relevant (B), and irrelevant (C). There were three features to be observed. An interesting finding is that relevant documents (type S, A and B) have a higher chance to contain matched query terms in their passages than irrelevant documents. The finding attracted us to do more experiments to be introduced in the next sections.

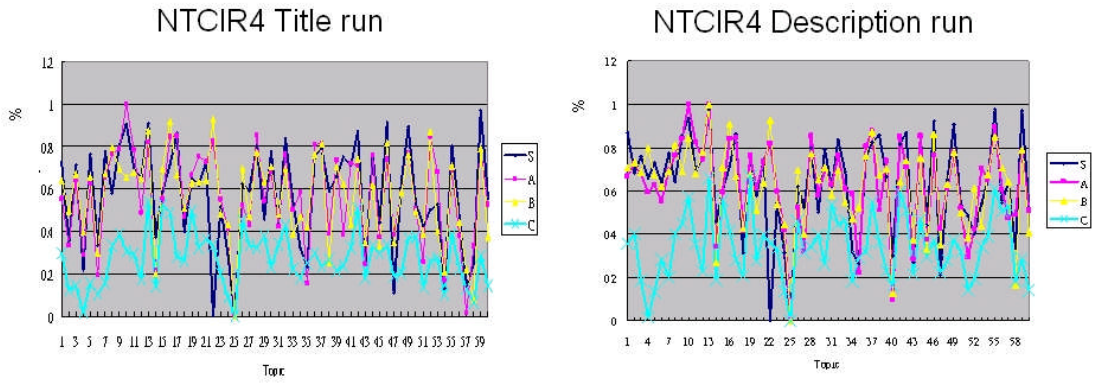

Figure 4: For most topics, relevant documents (S+A+B) get higher ratios of passages with matched query terms than irrelevant ones (C). Left: NTCIR4 title run. Right: NTCIR4 description run.

4. Passage-based Retrieval & Experiments

The second-stage process performs passage-based retrieval. The documents retrieved at the first stage will be segmented into passages via period markers in text as passage boundaries. In literature, passage is often defined as fix-length short article, and every passages are half-overlap with previous one. Passages are indexed and ranked, and the relevance of document is decided by its passage. If passage is taken as the unit to calculate scores and rank, the retrieval cost of passage become several times of that of document. In our work, we see a complete sentence separated by periods as a passage.

Query terms are main factors to effect retrieval outcome. Our Chinese segmentation tool is powered by the CKIP group, Academia Sinica. Because our test data set is a news story material, some news events use new words or longer terms that dictionary don't cover. We used the Web as the corpus to segment unknown words and extract longer terms. By sending all words in a query to Google, we can get a result that some words are frequently concatenated together in search result snippets. Adapting these combined phrases was found can remedy the lack of new terms in the dictionary.

The ranking policy is based on observation on relevant documents in which query terms are often concatenated in one paragraph. It is curious to know if using query term coverage in passages to re-score documents and change ranks of retrieval results could archive a better MAP value. We match query terms for each passage and increase the score of the document based on original score. Several passage-based ranking strategies were proposed and tested.

4.1 Passage-based Ranking

1. Ranking with Average Term Coverage of Passages

In the first experiment, the relevance value of a document was re-scored as the weighted sum of its Okapi score (the value of BM11) and the average term coverage rate of the composed passages (namely Strategy I). The term coverage rate is the percentage of the unique segmented query terms appearing in a passage, and the average term coverage rate means the average value of all passages' term coverage values. This experiment was performed to see if the addition of passage ranking score can improve the Okapi result. Figure 5 and Table 3 summarize the result of the experiment, in which the obtained MAP values are depicted with the change of the weight w from 0 (only using Okapi score), 0.01 to 0.10. The best results were obtained at w= 0.03 and 0.04, which perform better than that only using the Okapi score.

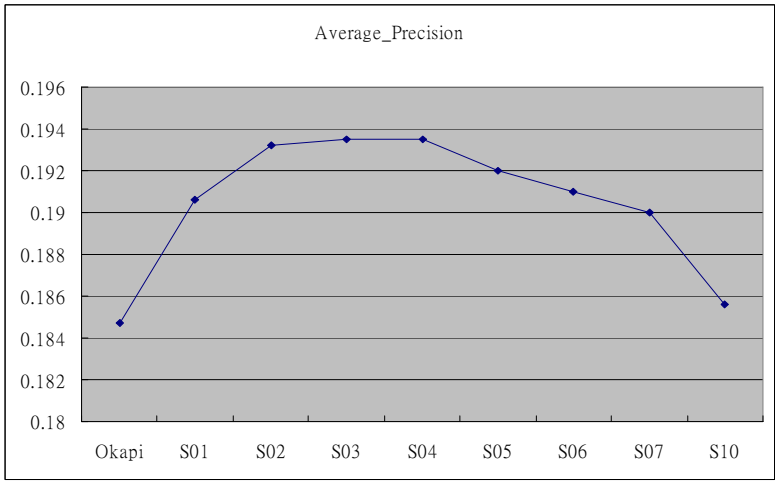

Figure 5: The MAP values obtained with Strategy I, which are depicted with respect to the increasing of the weight from 0.0 to 0.10.

Table 3 Selected 11-ponits precision rates obtained with Strategy I when w =0.01, 0.04 and 0.07

Recall	Precision						
	Okapi	Passage w=0.01	% change	Passage w=0.04	% change	Passage w=0.07	% change
0.00	0.4985	0.5199	+ 4.29	0.5360	+ 7.52	0.5319	+ 6.70
0.01	0.3729	0.3902	+ 4.64	0.3885	+ 4.18	0.3938	+ 5.60
0.20	0.2817	0.2999	+ 6.46	0.3094	+ 9.83	0.3063	+ 8.73
0.30	0.2496	0.2479	-0.68	0.2486	-0.40	0.2505	+ 0.36
0.40	0.2091	0.2147	+ 2.68	0.2193	+ 4.88	0.2074	-0.81
0.50	0.1742	0.1817	+ 4.30	0.1853	+ 6.37	0.1798	+ 3.21
0.60	0.1501	0.1530	+ 1.93	0.1574	+ 4.86	0.1487	-0.93
0.70	0.1169	0.1193	+ 2.05	0.1195	+ 2.22	0.1186	+ 1.45
0.80	0.0920	0.0944	+ 2.61	0.0928	+ 0.87	0.0923	+ 0.32
0.90	0.0680	0.0695	+ 2.20	0.0754	+ 10.88	0.0749	+ 10.14
1.00	0.0476	0.0477	+ 0.21	0.0519	+ 9.03	0.0511	+ 7.35
Avg Precision	0.1847	0.1906	+ 3.19 %	0.1935	+ 4.76 %	0.1900	+ 2.87 %

2. Ranking with Max Term Coverage of Passages

In the second experiment, the passage-based relevance score of a document is measured as the maximum term coverage of its composed passages, that is, for a document with three passages and the query with four segmented terms. If three passages contain 3, 1, 2 query terms respectively, then the max term coverage is 0.75 (3/4). That is quite simple to calculate. The relevance value of a document was then re-scored as the weighted summation of its Okapi score and the new passage-based relevance score (namely Strategy II). Figure 6 and Table 4 summarize the experimental result, in which the precision rates obtained with the addition of average term coverage to the Okapi are depicted with the change of the weight w from 0.0 to 1.0. The best result was obtained at w= 0.6, which perform slightly better than that using the Okapi score but worst than using Strategy I.

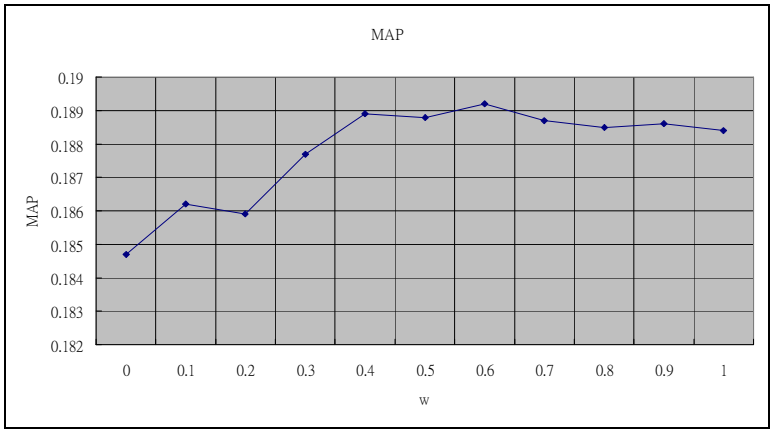

Figure 6: The MAP values obtained with Strategy II, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 4: The MAP values obtained with Strategy II when w =0.4, 0.5 and 0.6

	Okapi	Passage w = 0.4	% change	Passage w = 0.5	% change	Passage w = 0.6	% change
Avg. Precision	0.1847	0.1889	+ 2.27 %	0.1888	+ 2.22 %	0.1892	**+ 2.44 %**

3. Ranking with Average Term Coverage of the Top Three Passages

In the third experiment, the passage-based relevance score of a document is measured as the average term coverage of the top three passages. The relevance value of a document was then re-scored as the weighted sum of its Okapi score and the new passage-based relevance score (namely Strategy III). Figure 7 and Table 5 summarize the experimental result, in which the MAP values obtained with the addition of average term coverage to the Okapi are depicted with the change of the weight w from 0.0 to 1.0. The best results were obtained at w= 1.0, which perform better than that only using the Okapi score.

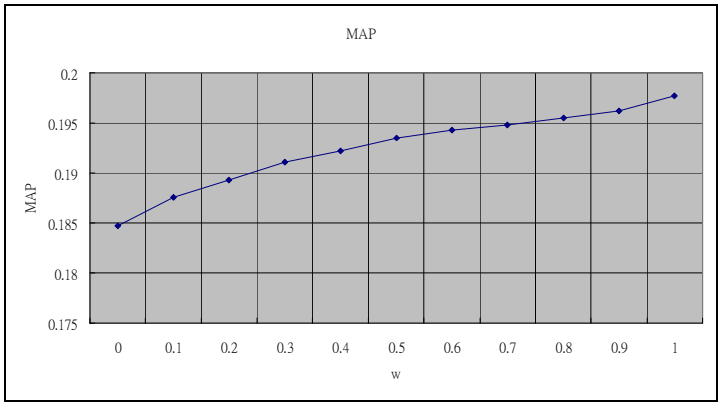

Figure 7: The MAP values obtained with Strategy III, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 5: The MAP values obtained with Strategy III when w =0.8, 0.9 and 1.

	Okapi	Passage w= 0.8	% change	Passage w = 0.9	% change	Passage w= 1	% change
Avg Precision	0.1847	0.1955	+ 5.85 %	0.1962	+ 6.23 %	0.1977	**+ 7.04 %**

4. Ranking with Percentage of Passages Containing Query Terms

In the forth experiment, the passage-based relevance score of a document is measured as the percentage of passages containing query terms. For a document with three passages and the query with four segmented terms, if two of the three passages contain at least one query term, then the percentage of passages containing query terms is 0.66 (2/3). The relevance value of a document was also re-scored as the weighted sum of its Okapi score and the new passage-based relevance score (namely Strategy IV). Figure 8 and Table 6 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.0 to 1.0. The best results were obtained at w= 0.3 and 0.2, which perform better than that only using the Okapi score.

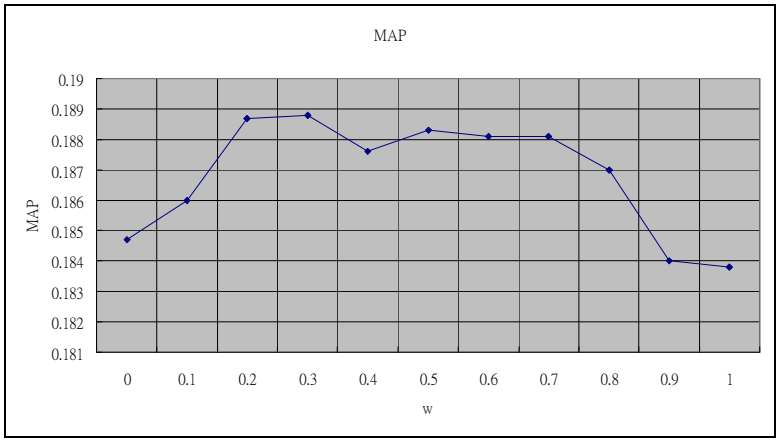

Figure 8: The MAP values obtained with Strategy IV, which are depicted with respect to the increasing of the weight from 0 to 1.

Table 6: The MAP values obtained with Strategy IV when w =0.2, 0.3 and 0.5.

	Okapi	Passage w=0.2	% change	Passage w=0.3	% change	Passage w=0.5	% change
Avg. Precision	0.1847	0.1887	+ 2.22 %	0.1888	**+ 2.22 %**	0.1883	+ 1.95 %

4.2 Ranking with "Headline" Matching

Our next group of experiments was performed to compare the results of "Headline" matching with the Okapi results and the combination of passage-based ranking scores. The NTCIR-4 document set is a news story set. News headlines normally contain keywords. We assume the query terms appearing in headlines are more critical. The relevance value of a document was re-scored as the weighted sum of its Okapi score and the headline-based relevance score (namely Strategy V). Critical keywords may appear in many sentences. If a document contains many occurrences of the critical keywords, the headline-matched query terms, the document will be assigned a higher score than that contains only other keywords. This strategy can help remove some news articles containing with query terms but irrelevant topics.

Figure 9 and Table 7 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.1 to 0.25. The best result was obtained at w= 0.13. As can be seen, the result is better than previous set of experiments.

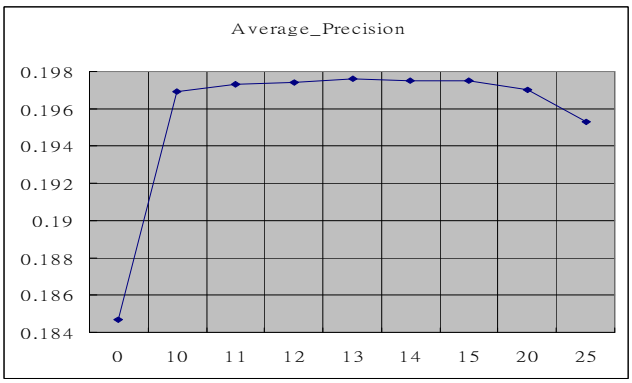

Figure 9: The MAP values obtained with Strategy V, which are depicted with respect to the increasing of the weight from 0.1 to 0.25.

Table 7: The MAP values obtained with Strategy V when w =0.1, 0.13 and 0.25.

	Okapi	HL w=0.1	% change	HL w=0.13	% change	HL w=0.25	% change
Avg Precision	0.1847	0.1969	+ 6.6 %	0.1976	+ **6.98** %	0.1953	+ 5.74 %

4.3 Combining Passage-based Ranking and "Headline" Matching

The last experiment was to compare the result of the combination of passage-based ranking and headline matching with the Okapi result. Figure 10 and Table 8 Figure 9 and Table 7 summarize the experimental result, in which the obtained MAP values are depicted with the change of the weight w from 0.01 to 0.02. The best result was obtained at w= 0.015. It is easy to see that the result is the best one that can be obtained in our experiment.

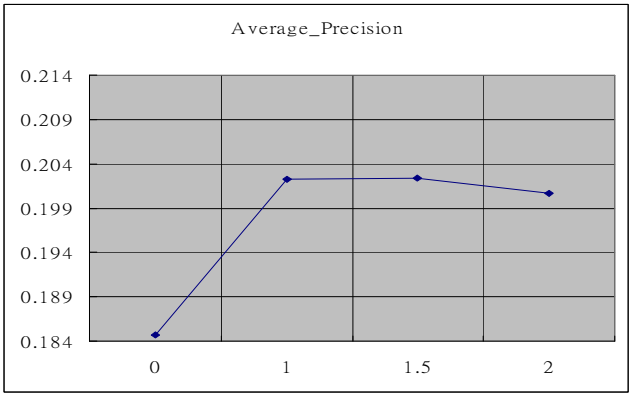

Figure 10: The MAP values obtained with Strategy VI, which are depicted with respect to the changes of the weights.

Table 7: Illustrated MAP values obtained with Strategy VI, the combination of Strategy I and V.

	Okapi	Passage w=0.01 HL 13%	% change	Passage w=0.015 HL 13%	% change	Passage w=0.02 HL 13%	% change
Avg Precision	0.1847	0.2022	+ 9.5 %	0.2024	+ **9.6 %**	0.2006	+ 8.6 %

5. Conclusion

In this paper, we have introduced a good approach to improving conventional TF*IDF methods. The approach is simple but practical. It combines the Okapi-based ranking algorithm with passage-based ranking strategies. The result also shows that using headline matching to determine critical keywords in queries is useful. A set of experiments have been conducted on the NTCIR-4 task for Chinese information retrieval. Although the proposed approach is simple, it is believed easily to be implemented and applied to the applications in industry.

6. Reference

1. Kaszkiel, M. and Zobel, J. (2001). Effective ranking with arbitrary passages. *Journal of the American Society For Information Science and Technology*, 52(4):344-364.

2. Callan, J.P. (1994). Passage-level evidence in document retrieval. In B.W. Croft & C.J. van Rijsbergen (Eds.), *Proceedings of the 17th annual international ACM-SIGIR conference on research and developments in information retrieval*, Dublin, Ireland, July (pp. 302-310), New York: ACM.

3. Kaszkiel, M. and Zobel, J. (1997). Passage retrieval revisited. In N. J. Belkin, D. Narasimhalu, & P. Willett (Eds.), *Proceedings of the 20th annual international ACM-SIGIR conference on research and development in information retrieval*, Philadelphia, PA (pp. 178-185).

4. Xiaoyong Liu, W. Bruce Croft. Language models for information retrieval: Passage retrieval based on language models. *Proceedings of the eleventh international conference on Information and knowledge management*, pp.375-382, November 2002

5. S. E. Robertson, S. Walker, S. Jones, M. M. HancockBeaulieu, and M. Gatford. *Okapi at trec*-3. In TREC-3, 1994.

6. K Kishida, K Chen, S Lee, K Kuriyama, N Kando, HH Chen, S.H. Myaeng, K. Eguchi. Overview of CLIR Task at the Forth NTCIR Workshop. *Proceedings of the 4th NTCIR*, 2004.

7. Noriko Kando, Overview of the Fourth NTCIR Workshop. *Proceedings of NTCIR-4 Workshop*, 2004.

8. Lingpeng Yang, Donghong Ji, and Li Tang. Chinese Information Retrieval Based on Terms and Ontology. *In: Proceedings of NTCIR-4 Workshop*, 2004.

9. Tetsuji Nakagawa and Mihoko Kitamura. NTCIR-4 CLIR Experiments at Oki. *In: Proceedings of NTCIR-4 Workshop*, 2004.

10. J. M. Ponte. A Language Modeling Approach to Information Retrieval. *Ph.D. Thesis, Graduate School of the University of Massachusetts Amherst*, 1998.

11. Kui-Lam Kwok, Norbert Dinstl and Sora Choi. NTCIR-4 Chinese, English, Korean Cross Language Retrieval Experiments Using PIRCS. *In: Proceedings of NTCIR-4 Workshop*, 2004.

12. K.L. Kwok. NTCIR-2 Chinese and cross language experiments using PIRCS. *In: Proceedings of NTCIR-2 Workshop*, 2002.