Vocabulary-based Re-ranking for Geographic and Temporal Searching at NTCIR GeoTime Task

Kazuaki Kishida
School of Library and Information Science, Keio University
2-15-45 Mita, Minato-ku,Tokyo,
JAPAN
+81-3-5418-6739

kishida@slis.keio.ac.jp

ABSTRACT

This paper reports on experiments in the NTCIR-8 GeoTime task performed by research group at School of Library and Information Science in Keio University (KOLIS), which tried to explore techniques for searching a Japanese document collection for requests on geographic and temporal information. A special component of re-ranking for enhancing performance of geographic and temporal searches was added to the KOLIS system, in which standard BM25 and probabilistic pseudo-relevance feedback (PRF) were implemented. That is, at the first stage, a list of documents relevant to a given topic was specified by standard IR techniques, and at the second stage, the list was re-ranked after scores of documents which included geographic and temporal terms were increased. More specifically, the different number of geographic and temporal terms appearing in each document was counted using a special dictionary including only such kind of terms, and its document score was modified based on the number. In this experiment of Japanese monolingual (JA-JA) retrieval and English to Japanese bilingual (EN-JA) retrieval, the search runs using jointly the re-ranking and PRF showed the highest performance, followed by the re-ranking only runs, and PRF only runs, in this order. This result indicates that the simple re-ranking technique has an effect on enhancement of geographic and temporal searches. In comparison of performance between JA-JA and EN-JA searches, performance of the bilingual searches was just slightly inferior to that of monolingual searches. The experiment by KOLIS group adopts a simple query translation approach using public machine translation services of two Internet search engines, it turned out that translations obtained from the search engines worked well (the translations were segmented into a set of terms using the same indexing method applied to the Japanese document set. The method was a hybrid approach concatenating two results from longest matching with a Japanese dictionary and decomposing sentences into character-based overlapped bi-grams).

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval – retrieval models, search process.

General Terms

Experimentation; Performance; Measurement

Keywords

Japanese Monolingual Information Retrieval; Cross-lingual Information Retrieval

1. INTRODUCTION

The GeoTime task at NTCIR-8 challenges to enhance effectiveness of searching for geographic or temporal information [1]. In this paper, techniques used for the task by a research group at School of Library and Information Science in Keio University (KOLIS) are described, and its results of Japanese monolingual search runs (JA-JA runs) and English to Japanese bilingual search runs (EN-JA runs) are discussed (i.e., the KOLIS group employs only Japanese document sets).

Although the KOLIS system basically consists of only a plain search engine based on a standard BM25, a special component of re-ranking documents that are selected according to BM25 scoring was newly incorporated for effective searches of geographic or temporal information. In the re-ranking component, a dictionary including geographic and temporal terms is used, and the different number of such terms appearing in each document is counted. After that, each document score computed by BM25 is modified based on the number of geographic and temporal terms so that scores of documents having more geographic and temporal terms increases. The mechanism is simple, but if our experiment can show that the re-ranking has a positive effect, its result would be useful because this vocabulary-based re-ranking technique is easy to incorporate into a system.

2. RE-RANKING BASED ON SPECIFIC SUB-VOCABULARY

2.1 Two-stage Searching

As described above, in the NTCIR-8 GeoTime task, documents including geographic or temporal information relevant to the search topics have to be ranked higher in output lists. For example, the topic “GeoTime-0001” is that “the user wants to know when and in what city the children's author Astrid Lindgren died” in which date and place information is assumed to be asked by the end-user. Since the target documents in Japanese used for the NTCIR GeoTime task is a large set of news articles (The Mainichi 2002-2005 news collection), two-stage searching would be a realistic strategy, i.e., at the first stage, a subset of documents relevant to the general subject of each topic is specified, and at the second stage, documents including geographic and/or temporal information are identified from the subset.

While it is enough to apply a conventional IR technique at the first stage, a novel method would be useful for enhanced processing at the second stage. For example, machine learning approaches may be useful for finding documents with geographical and/or temporal information, similarly in applications such as opinion analysis. Another strategy is to provide a higher weight to documents including a specific vocabulary of geographic or temporal concepts or events. This means that the IR system re-ranks documents obtained at the first stage according to a weighting formula. The vocabulary-based re-ranking is simpler than machine learning approach. Therefore, its performance may be relatively lower, but vocabulary-based re-ranking has an advantage in term of implementation. This paper adopts the simple vocabulary-based re-ranking approach.

2.2 Re-ranking Procedure

In this experiment, a geographical dictionary in the ChaSen system which is a well-known Japanese morphological analyzer [2] was employed, i.e., Noun.place file of IPADIC ver 2.6.3 was incorporated into our KOLIS system as a special dictionary, and ‘geographic terms’ were operationally defined as those appearing in the Noun.place file. At the second stage, after the different numbers of the ‘geographical terms’ that occur actually in 1000 documents specified at the first stage were counted, each document score was modified such that

$$v'_i = v_i \times \left(1.0 + 0.5 \times \frac{x_i}{\max_{k=1,\ldots,1000} x_k} \right) \qquad (1)$$

where

v_i : Original document score of *i*-th document in the output,

v'_i : Modified document score of *i*-th document in the output, and

x_i : The different number of geographic terms in *i*-th document in the output.

The final output was sorted in descending order of the modified document score. It should be noted Japanese representations indicating a specific year (00 to 99 years, i.e., “00 年” to “99 年”) and a month (January to December, i.e., “1 月” to “12 月”) were compulsorily added into the Noun.place in this experiment. Therefore, the re-ranking was executed based on not only geographic terms but also temporal representations.

3. IR System

3.1 Indexing

In this experiment, Japanese texts of search topics and of documents were segmented based on a hybrid indexing technique, in which all word segments identified by

- character-based overlapped bi-gram technique, and
- longest matching with a machine-readable dictionary

were adopted as index terms.

For example, suppose that the dictionary contains entries “日本国 (Japan country)” and “憲法 (the constitution)” and the text is “日本国憲法(the Constitution of Japan)”. In this case, the KOLIS system registers a set of segmented strings, “日本国” and “憲法” obtained from matching with the dictionary and “日本”, “本国”, “国憲” and “憲法” generated as bi-grams. Note that “憲法” is doubly extracted and its term frequency becomes two.

By using jointly the two indexing techniques, it is expected to specify exhaustively strings to be needed for searching. Actually, noun files of IPADIC ver. 2.6.3 [2] were used as the dictionary for this processing expect for Noun.number file. Strings extracted by matching with the dictionary and bi-grams were recorded in an index file implemented as a B-tree.

3.2 Document Scoring and Pseudo-relevance Feedback

In this experiment, a standard BM25 [3] was used for computing document scores at the first stage, and a standard pseudo-relevance feedback (PRF) technique based on a probabilistic term weighting [4] was applied in some search runs. More specifically, in our system, 10 new terms which have the highest term weights among those in top-ranked 30 documents are added to the set of original query terms.

3.3 Bilingual Searching

For English to Japanese (EN-JA) bilingual searching, the text of each search topic was simply entered into machine translation (MT) systems provided by Yahoo! Japan [5] and Excite Japan [6]. Translation results from both the MT services for each search topic were straightforwardly concatenated and treated as a set of sentences representing the topic in Japanese. After that, search runs were executed in the same manner with JA-JA monolingual searches. The reason why two MT services were jointly used is the almost same with that for adopting the hybrid indexing technique mentioned above, i.e., augmenting probability of obtaining correct translations is the main reason.

4. Experiment

4.1 Submitted Runs

Table 1 shows an outline of nine search runs submitted officially to NTCIR GeoTime organizers (it should be noted that each search run was executed for 25 search topics prepared by the organizers, respectively. However, after submission, topic 17 was removed by the organizers for a reason [1]). The baseline searches are KOLIS-JA-JA-D-01 (for Japanese monolingual search) and KOLIS-EN-JA-D-01 (for English-Japanese bilingual search) in which any re-ranking and PRF were not applied.

Re-ranking and PRF techniques were additionally applied to the baseline searches as follows.

- KOLIS-JA-JA-D-02 and KOLIS-EN-JA-D-02: PRF was added.
- KOLIS-JA-JA-D-03 and KOLIS-EN-JA-D-03: Re-ranking was added.
- KOLIS-JA-JA-D-03 and KOLIS-EN-JA-D-03: both PRF and Re-ranking were added. More precisely, after identifying 1000 documents by PRF, re-ranking technique was applied in these runs.

In the above search runs, only <description> filed was used as search topic. On the other hand, KOLIS-JA-JA-DN-05 search run was executed based on both <description> and <narrative> fields including rich terms on each topics. Note that KOLIS-JA-JA-DN-05 was executed with no re-ranking and no PRF.

Table 1. Search runs submitted from KOLIS group

Run Type	ID	Topic field	Re-Ranking	PRF
JA-JA	KOLIS-JA-JA-D-01	D	No	No
	KOLIS-JA-JA-D-02	D	No	Yes
	KOLIS-JA-JA-D-03	D	Yes	No
	KOLIS-JA-JA-D-04	D	Yes	Yes
	KOLIS-JA-JA-DN-05	D and N	No	No
EN-JA	KOLIS-EN-JA-D-01	D	No	No
	KOLIS-EN-JA-D-02	D	No	Yes
	KOLIS-EN-JA-D-03	D	Yes	No
	KOLIS-EN-JA-D-04	D	Yes	Yes

Table 2 Performance of search runs (official results)

Run Type	ID	mean AP	mean Q	mean nDCG
JA-JA	KOLIS-JA-JA-D-01	0.2878	0.3327	0.4982
	KOLIS-JA-JA-D-02	0.3008	0.3378	0.5036
	KOLIS-JA-JA-D-03	0.3139	0.3459	0.5063
	KOLIS-JA-JA-D-04	**0.3250**	**0.3544**	**0.5159**
	KOLIS-JA-JA-DN-05	0.3027	0.3392	0.5095
EN-JA	KOLIS-EN-JA-D-01	0.2773	0.3232	0.4729
	KOLIS-EN-JA-D-02	0.2870	0.3277	0.4765
	KOLIS-EN-JA-D-03	0.2918	0.3329	0.4817
	KOLIS-EN-JA-D-04	**0.3145**	**0.3468**	**0.4956**

4.2 Results and Discussions

4.2.1 Basic statistics

The Japanese document collection of NTCIR-8 GeoTime task (i.e., The Mainichi 2002-2005 news collection) consist of 377,941 records, from which 1,940,553 distinct terms were identified by the indexing method of KOLIS system. The average document length was 375.3679.

4.2.2 Effect of re-ranking and PRF

Table 2 shows scores of mean average precision (AP), mean Q-measure (Q) and mean nDCG for 24 search topics (these scores are officially provided by the NTCIR GeoTime organizers). As the scores indicate, it turns out that search runs using jointly re-ranking and PRF achieved the highest performance in both JA-JA and EN-JA tasks (i.e., KOLIS-JA-JA-D-04 and KOLIS-EN-JA-D-04).

In comparison between KOLIS-JA-JA-D-02 and KOLIS-JA-JA-D-03, the latter run outperforms the former, which means that re-ranking based on geographic and temporal vocabulary bring more improvement of search performance than PRF. The same tendency was observed in KOLIS-EN-JA-D-02 and KOLIS-EN-JA-D-03 of bilingual searches. While the difference is not so large, this result shows that the re-ranking technique examined in this experiment has an effect to some degree.

Therefore, it can be concluded from the experiment that search performance is better in the following order;

Baseline < PRF < Re-ranking < (PRF + Re-ranking).

4.2.3 Topic-by-Topic analysis

Figure 1 is a plot of AP scores of baseline (KOLIS-JA-JA-D-01) and ‘PRF+Re-ranking’ (KOLIS-JA-JA-D-04) runs by each topic. If a marker which represents a pair of AP scores of two runs for a topic is located above the diagonal line, the ‘PRF+Re-ranking’ run outperforms the baseline run for the topic (vice versa). As Figure 1 shows, ‘PRF+Re-ranking’ is more effective in many topics, but reduced search performance in a few topic, i.e., ‘PRF+Re-ranking’ did not always improve performance.

Similarly, Figure 2 is a plot of re-ranking and baseline, and Figure 3 is a plot of re-ranking and PRF. The same tendency in Figure 1 is also observed in Figure 2.

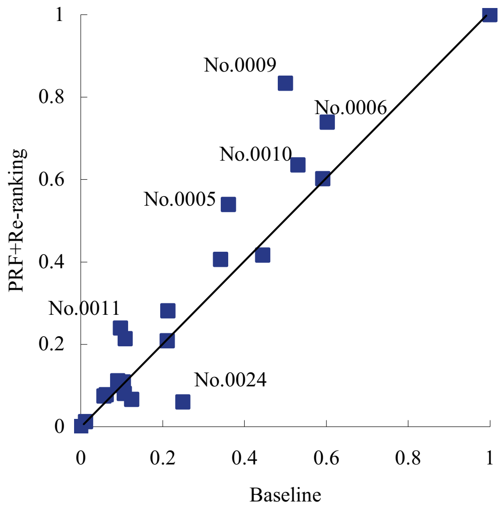

Figure 1 Plot of AP scores of baseline and ‘PRF+Re-ranking’ runs by topic

4.2.4 Performance of bilingual IR

In comparison of performance between bilingual (EN-JA) and monolingual (JA-JA) searches, monolingual runs outperforms naturally bilingual ones (see Table 2). However, the difference is so small. For example, mean AP of the best bilingual runs (KOLIS-EN-JA-D-04) reaches to about 97% of the best monolingual runs (KOLIS-JA-JA-D-04), i.e., the scores are 0.3145 and 0.3250, respectively. Intuitively, this means that translation services provided by search engines [5, 6] worked well in this experiment. However, it should be noted that there are some topics for which bilingual runs are more effective. It is not easy to asnwer a question why bilingual searches were superior or inferior, and more sophisticated analysis is needed for it.

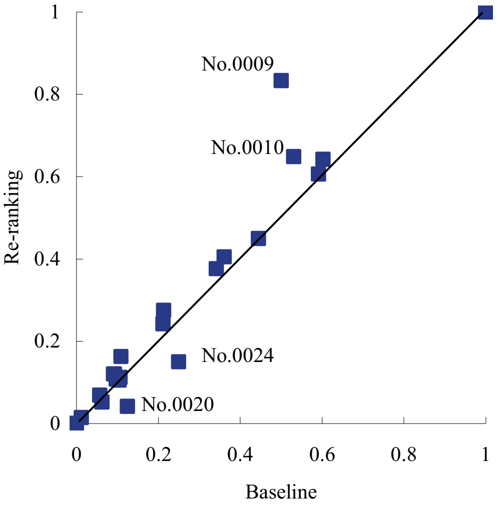

Figure 2 Plot of AP scores of baseline and ‘Re-ranking’ runs by topic

5. CONCLUDING REMARKS

It was shown by the experiment that vocabulary-based re-ranking for geographic and temporal searches discussed in this paper improves search performance slightly. However, the degree of the improvement was neither large nor statistically significant. The reason why the re-ranking is effective should be explored by deeper analysis in future researches. If the reason is clarified, more sophisticated techniques may be derived from it.

6. ACKNOWLEDGMENTS

The author thanks to the organizers for planning and managing of the NTCIR-8 GeoTime task.

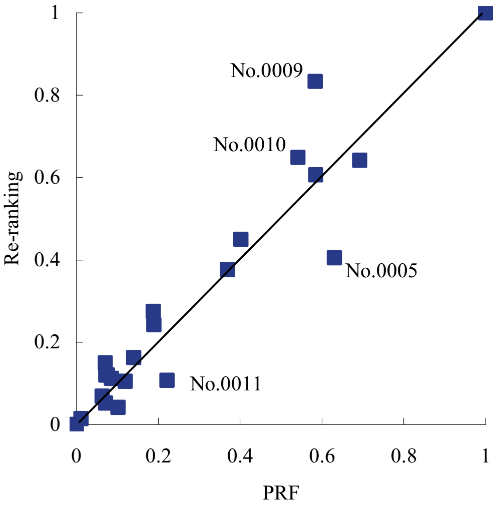

Figure 3 Plot of AP scores of ‘PRF’ and ‘Re-ranking’ runs by topic

7. REFERENCES

[1] Gey, F., Larson, R., Kando, N., Machado, J., and Sakai, T. 2010. NTCIR-GeoTime overview: evaluating geographic and temporal search. In Proceedings of NTCIR-8 (Tokyo, Japan, June 15-18, 2010).

[2] http://chasen.aist-nara.ac.jp/chasen/distribution.html.en

[3] Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M. 1995. Okapi at TREC-3. In Overview of the Third Text REtrieval Conference (TREC-3). National Institute of Standards and Technology, Gaithersburg.

[4] Robertson, S. E. and Sparck Jones, K. 1994. Simple, Proven Approaches to Text Retrieval, Technical Report No.356, Computer Laboratory, University of Cambridge.

[5] http://honyaku.yahoo.co.jp/

[6] http://www.excite.co.jp/world/