# Ricoh in the NTCIR-4 CLIR Tasks

Yuichi KOJIMA Hideo ITOH
Ubiquitous Solution Lab, Software R&D Group, Ricoh Co., LTD.
1-1-17 Koishikawa, Bunkyo-ku,Tokyo 112-0002, Japan
{ykoji, hideo}@src.ricoh.co.jp

## Abstract

*This paper describes Ricoh's participation in the NTCIR-4 CLIR tasks. We used the same approach as we took at the NTCIR-3 IR tasks for Japanese. We applied our system using a Traditional/Simplified Chinese converter and n-gram indexing for the Chinese IR task. The results show that our simple approach for Chinese IR can provide information retrieval for both Traditional and Simplified Chinese.*
**Keywords:** *Chinese, Japanese, n-gram, pseudo-relevance feedback.*

## 1 Introduction

For the NTCIR-4 SLIR sub-tasks, Ricoh submitted runs for Chinese and Japanese. We have worked on English and Japanese text retrieval for the past few years[2, 5, 9]. Recently, we have been enhancing our information retrieval(IR) system for some European languages[3]. The SLIR experiments contained our first trial in written Chinese. Our approach for Chinese is n-gram indexing and character normalization using a map from Traditional Chinese characters to Simplified Chinese characters. Therefore, our focus in the experiments was to test our approach for supporting these two Chinese language.

Section 2 outlines our system, Section 3 describes the modifications made for the experiments, Section 4 presents the results, and Section 5 gives some conclusions.

## 2 Description of the System

Before describing our approach, we brifly describe the system as background information. Its basic features are:

- Effective document ranking based on the probabilistic model[8] with query expansion using pseudo-relevance feedback[4]
- Scalable and efficient indexing and searching based on the inverted file module[5]

This system was also used for TREC, CLEF, and previous NTCIR experiments, where its effectiveness was shown. In the following sections, we explain the processing flow of the system[2].

### 2.1 Query Term Extraction

An input topic string is transformed into a sequence of normalized and stemmed tokens using a tokenizer, normalizer and stemmer. Stop words are identified using a stopword dictionary and eliminated. Two kinds of terms are extracted from normalized and stemmed words for initial retrieval: a "single term" is each word and a "phrasal term" consists of two adjacent words in the query string.

### 2.2 Initial Retrieval

Each query term is assigned a weight $w_t$, and documents are ranked according to the score $S_{q,d}$ as follows:

$$W_t = \log(k_4^{'} \bullet \frac{N}{n_t} + 1) \quad (1)$$

$$S_{q,d} = \Sigma_{t \in q} \frac{f_{t,d}}{K + f_{t,d}} \bullet \frac{w_t}{k_4^{'} \bullet N + 1} \quad (2)$$

$$K = k_1 \bullet ((1-b) + b \bullet \frac{l_d}{l_{ave}}) \quad (3)$$

where $N$ is the number of documents in the collection, $n_t$ is the document frequency of the term $t$, $f_{t,d}$ is the in-document frequency of the term $t$, $l_d$ is the document length, $l_{ave}$ is the average document length, and $k_4^{'}$, $k_1$ and $b$ are parameters. Weights for phrasal terms are set lower than those for single terms.

### 2.3 Query Expansion

As a result of the initial retrieval, the top ten documents are assumed to be relevant (pseudo-relevance) to the query and are selected as a "seed" for query expansion. Candidates for expansion terms are extracted from the seed documents in the same way as for the query term extraction mentioned above. Phrasal terms are not used for query expansion. The candidates

are ranked on the Robertson's Selection Value[6], or $RSV_t$ and the top-ranked terms are selected as expansion terms. The weight is re-calculated as $w2_t$ using the Robertson/Sparck-Jones formula[7]

$$RSV_t = w2_t \bullet (\frac{r_t}{R} - \frac{n_t}{N}) \quad (4)$$

$$w2_t = \alpha \bullet w_t + (1-\alpha) \bullet \log \frac{\frac{r_t+0.5}{R-r_t+0.5}}{\frac{n_t-r_t+0.5}{N-n_t-R+r_t+0.5}} \quad (5)$$

where $R$ is the number of relevant documents, $r_t$ is the number of relevant documents containing the term $t$, and $\alpha$ is a parameter. The weight of the initial query term is re-calculated using the same formula as above, but with a different $\alpha$ value and an additional adjustment to make the weight higher than the expansion terms.

### 2.4 Final Retrieval

Using the initial query and expansion terms, the ranking module performs a second retrieval to produce the final result.

## 3 Experiments

Five elements in the system must be adjusted depending on the language: (1) the tokenizer, (2) the normalizer, (3) the stemmer, (4) the stopword dictionary and (5) the parameter set. In addition, it is necessary to select (6) the method of deleting abnormal documents from the test collection. For the stemmer, we used the English Snowball stemmer[1] for words in Japanese and Chinese documents. The other elements are described below.

### 3.1 Tokenizer and Normalizer

Our system has a different behavior for each language.

For Japanese: (1) a Japanese string (consisting of Chinese, Hiragana, Katakana characters) is divided into Japanese words and (2) each word is normalized using the Japanese normalizing rule, the main part of the rule is the replacement of substrings of a word.

For Chinese: (1) a Chinese string (string of Chinese characters) is divided into characters which are regarded as pseudo words and (2) each pseudo word is normalized using the rule for converting from Traditional Chinese characters to Simplified Chinese characters.

Since our system performs information retrieval for Chinese documents using a simple conversion rule, the system can provide four kinds of retrieval: (1) finding Traditional Chinese documents using a Traditional Chinese query, (2) finding Traditional Chinese documents using a Simplified Chinese query, (3) finding Simplified Chinese documents using a Simplified Chinese query and (4) finding Simplified Chinese documents using a Traditional Chinese query.

### 3.2 Stopword Dictionary

The stopword dictionary used in our system was used in the NTCIR-3 tasks for Japanese documents and has only one word U+7684 for Chinese documents. The tokenizer makes pseudo words from a Chinese string. So, we made a list of pseudo stop words which are list of characters by trial and error. The character U+7684 was effective and other characters such as U+6211, U+4F55 and U+4F60 were not so effective for current test collections.

### 3.3 Parameter Set

The parameter set is different for each language. We trained the system by selecting the best parameter set from among hundreds of candidate parameter sets for each language to get the highest average precision score with the test collection used for the NTCIR-3 CLIR tasks.

### 3.4 Methods of Deleting Abnormal Documents

The Japanese and Chinese test collections contained some abnormal documents. We used different methods of deleting them for each language because of the limited time (Table 1).

Table 1. Deletion methods

| Language | Method |
|---|---|
| Japanese | remove abnormal records and re-execute indexing process |
| Chinese | delete IDs of abnormal records from the output |

## 4 Results

Table 2 shows a comparison of performance with and witout Chinese character normalization. The result shows the effectiveness of the normalization using a map from Traditional Chinese characters to Simplified Chinese characters.

Table 3 gives a summary of our official results for the NTCIR-4 CLIR tasks. We cannot compare the average precision values across tasks. However, we can estimate our position in each task using the median and the maximum values of the average precisions. Although our performance of in the Chinese information retrieval experiment was reasonable, there is still room for improvement compared with the Japanese information retrieval experiment.

Table 2. Comparison of performance with and without Chinese character normalization

| Task | Eval | Average precision with normalization | Average precision without normalization |
|---|---|---|---|
| C-C-T | rigid | 0.2112 | 0.2101 |
| | relaxed | 0.2641 | 0.2588 |
| C-C-D | rigid | 0.2087 | 0.1912 |
| | relaxed | 0.2671 | 0.2432 |

Table 3. Evaluation results of the tasks

| Task | Eval | Av. precision | Median | Max. |
|---|---|---|---|---|
| C-C-T | rigid | 0.2112 | 0.1881 | 0.3146 |
| | relaxed | 0.2641 | 0.2356 | 0.3799 |
| C-C-D | rigid | 0.2087 | 0.1741 | 0.3255 |
| | relaxed | 0.2671 | 0.2219 | 0.3880 |
| J-J-T | rigid | 0.3720 | 0.3135 | 0.3890 |
| | relaxed | 0.4663 | 0.4112 | 0.4864 |
| J-J-D | rigid | 0.3680 | 0.3352 | 0.3804 |
| | relaxed | 0.4691 | 0.4295 | 0.4838 |

C-C-T: Chinese, title only
C-C-D: Chinese, description only
J-J-T: Japanese, title only
J-J-D: Japanese, description only

## 5 Conclusions

Our system can handle four cases of Chinese information retrieval. The experiments showed reasonable performance for one case: finding Traditional Chinese documents using a Traditional Chinese query. On the other hand, we are optimistic about one case. Finding Simplified Chinese documents using a Simplified Chinese query is similar to the first case, because the normalizer converts a Traditional Chinese character into a Simplified Chinese character. Future work will include checking the performance for other cases.

## References

[1] Snowball web site. http://snowball.tartarus.org/.

[2] H. Itoh, H. Mano, and Y. Ogawa. Ricoh at trec-10. 2001. Tenth Text Retrieval Conference (TREC-2001).

[3] Y. Kojima, H. Itoh, H. Mano, and Y. Ogawa. Ricoh at clef 2003. 2003. http://clef.iei.pi.cnr.it:2002/2003/WN_web/26.pdf.

[4] H. Ogawa, H. Mano, M. Narita, and S. Honma. Structuring and expanding queries in the probabilistic model. 2000. Eighth Text Retrieval Conference (TREC-8).

[5] Y. Ogawa and H. Mano. Ricoh at ntcir-2. 2001. Proceedings of the Second NTCIR Workshop Meeting.

[6] S. E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4):359–364, 1990.

[7] S. E. Robertson and K. Spark-Jones. Relevance weighting of search terms. *Journal of ASIS*, 27:129–146, 1976.

[8] S. E. Robertson and S. Walker. On relevance weights with little relevance information. 1997. 20th Annual International ACM SIGIR Conference (SIGIR '97).

[9] M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh, and Y. Ogawa. University of tokyo/ricoh at ntcir-3 web retrieval task. 2002. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-WEB-ToyodaM.pdf.