RICOH at NTCIR-2

Yasushi OGAWA Hiroko MANO
Software Research Center, RICOH Co., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN
{yogawa, mano}@src.ricoh.co.jp

Abstract

At NTCIR-2, RICOH submitted eight runs for the Japanese IR task. Of the eight runs, four runs use the title field only and the other four use the description field only.

RICOH's system is built on our English text retrieval system and augmented to handle Japanese text. The system features (1) hybrid retrieval using a combination of n-gram indexing and word-based document ranking; (2) word-based and n-gram-based query expansion; (3) a modified version of the Okapi's probabilistic model.

Keywords: *n-gram indexing, hybrid approach, query expansion, Okapi probabilistic model.*

1 Introduction

This is RICOH's first participation in NTCIR, and we submitted eight runs for the Japanese IR task.

Our system adopts the probabilistic model with automatic query expansion [5], and retrieval is outlined as follows:

1. Query construction

   Each topic is morphologically analyzed using a Japanese parser developed by our research group and predefined stop-words are filtered out from the words obtained.

2. Initial retrieval

   The selected words are assigned weights, and scores are computed for each document containing any of the selected words using their weights.

3. Query expansion

   The initial query is automatically expanded; the terms except stop-words in the top-ranked documents in the initial retrieval are evaluated and the terms ranked the highest are added to the original query.

4. Final retrieval

   Document scores are computed using the term weights assigned during query expansion, and the final results are determined.

The system features:

- Hybrid retrieval

  We combined n-gram (n successive characters) indexing with word-based document ranking to achieve high retrieval effectiveness and efficiency.

- Word-based and n-gram-based query expansion

  We added query expansion to the hybrid retrieval method, to further improve retrieval effectiveness. We tried both word-based expansion and n-gram-based expansion.

- Modified Okapi model

  We modified the Okapi's probabilistic model to ease parameter tuning and improve the quality of term selection in expansion.

2 Hybrid retrieval

In Japanese, word boundaries are not indicated by punctuation marks. In order to apply word indexing to Japanese text, morphological analysis is required to identify words in the text. This analysis, however, poses some problems; indexing speed is degraded, the dictionary requires constant updating, indexes must be reconstructed following updates, and so on. To avoid these problems, n-gram indexing, which uses n-grams as indexing units [1][2][6], has been proposed and widely used in Japanese text retrieval systems.

For document ranking, however, using n-grams also as scoring units tends to result in low retrieval effectiveness [3][4], though n-gram-based ranking is certainly fast since a simple lookup is all that

is needed to obtain frequency statistics of each n-gram. The concern about the retrieval effectiveness led to proposals of a combination of a *word-based ranking method on n-gram indexing*, sometimes called the *hybrid* method [1][3][4] and this is the approach we employed for the NTCIR-2 experiments.

In the hybrid method, a query processing incorporates a morphological analysis, but the problems associated with morphological analysis are minimal in retrieval; the analysis takes place only during query construction – the amount of text is limited, so a lower processing speed does not matter for queries, and re-indexing after dictionary updates is not necessary since the indexed text is not parsed.

3 Query expansion

In query expansion, again, there is a similar issue as to what should constitute an expansion *term*.

A straightforward way is to use words as terms, as in document ranking. The word-based query expansion requires morphological analysis to identify words in each of the retrieved documents. Unlike in query processing, however, the amount of text that needs to be morphologically analyzed is much larger, resulting in a long processing time. Also, obtaining necessary frequency statistics of each word is costly since these statistics have to be calculated using the data in the n-gram index. Consequently, word-based query expansion requires a considerable time.

Another possible term unit is an n-gram, the unit used for indexing. The n-gram-based query expansion, with which n-grams are used as expansion terms, eliminates much of the computation that is required for the word-based method, hence faster expansion is possible. On the other hand, since n-grams do not reflect semantics as words do, retrieval effectiveness may suffer.

We implemented both word-based and n-gram-based query expansion and compared the performance in terms of speed and retrieval effectiveness.

4 Retrieval model

To enhance the Okapi's probabilistic model, the following modifications were added.

4.1 Initial weight

Okapi's weighting formula has a problem of possible negative term weights [7]. For easier parameter setting, we changed the formula so that term weights are always positive for initial retrieval [5]. The weight of each term is calculated by using the formula

$$w_t = \log\left(k_4' \cdot \frac{N}{n_t} + 1\right),$$

where N is the number of documents in the collection, n_t is the document frequency of the term t, and $k_4'(\geq 0)$ is a tuning parameter.

Note that the term weights, with our formula, never get negative. By keeping the term weights positive, the quality of retrieval is maintained even in the worst case.

With each term weighted according to the above formula, the ranking score for each document is calculated using the formula

$$s_{q,d} = \sum_{t \in q} \frac{f_{t,d}}{K + f_{t,d}} \cdot \frac{w_t}{k_4' \cdot N + 1},$$
$$K = k_1\left((1-b) + b\frac{l_d}{l_{ave}}\right),$$

where $f_{t,d}$ is the in-document frequency of the term in the document d, l_d is the document length (the number of characters), l_{ave} is the average document length, and k_1 and b are parameters.

4.2 Feedback weight and term selection

In query expansion, all terms except stop-words in the top-ranked documents are ranked according to its TSV (term selection value), and the terms ranked the highest are added to the initial query.

For each term collected, a new weight is assigned based on the feedback from the retrieved documents. The term re-weighting formula is similar to the Okapi's formula but reflects the change in the initial weighting mentioned above.

$$\begin{aligned} w_t = &\frac{k_5}{k_5 + \sqrt{R}} \log\left(k_4' \frac{N}{N - n_t} + \frac{n_t}{N - n_t}\right) \\ &+ \frac{\sqrt{R}}{k_5 + \sqrt{R}} \log \frac{r_t + 0.5}{R - r_t + 0.5} \\ &- \frac{k_6}{k_6 + \sqrt{S}} \log \frac{n_t}{N - n_t} \\ &- \frac{\sqrt{S}}{k_6 + \sqrt{S}} \log \frac{s_t + 0.5}{S - s_t + 0.5}, \end{aligned}$$

where R is the number of relevant documents, r_t is the number of relevant documents containing term t, S is the number of non-relevant documents, s_t is the number of non-relevant documents containing term t, and k_5 and k_6 are parameters. In our experiments, the top 10 documents were assumed to be relevant, and we did not use any non-relevant documents.

Table 1. Average precision

	title	desc
no-exp	0.2299	0.3243
no-exp(+pair)	0.2345	0.3175
exp(word)	0.2609	0.3669
exp(n-gram)	0.2695	0.3610

Table 2. Expansion time (sec)

	title	desc
exp(word)	68.43	67.89
exp(n-gram)	1.47	4.05

Then, TSV v_t is calculated using the formula

$$v_t = \left(\frac{\sum_{d \in R} \frac{f_{t,d}}{K + f_{t,d}}}{R} - \beta \cdot \frac{\sum_{d \in S} \frac{f_{t,d}}{K + f_{t,d}}}{S} \right) w_t,$$

where β is a parameter. We selected, for expansion, 30 terms with the highest TSVs.

Note that this formula is different from the Okapi's in that ours uses not just document frequency but also in-document frequency to reduce too specific terms.

5 Results

We submitted eight runs; the following four kinds of runs were produced for both the title and the description fields.

[**no-exp**] the initial retrieval result is used as the final result without query expansion.

[**no-exp(+pair)**] the adjacent word pairs in topics are added to queries. The word pairs are represented using the proximity operators and down-weighted as in our TREC experiments [5].

[**exp(word)**] the word-based query expansion was applied.

[**exp(n-gram)**] the n-gram-based query expansion was applied.

Average precisions are shown in Table 1. Comparing results without expansion, the addition of word pairs increased average precision in the title case, but decreased in the description case. Both the word-based method and the n-gram-based method of query expansion were quite effective, and there was not significant difference between the two methods.

Processing time used in query expansion was measured on SUN Ultra2 workstation with local disk, and average time per query is shown in Table 2. In the n-gram-based expansion, the processing time was reduced to less than 7% of that of in the word-based expansion. Considering there was no significant difference in precision between the two methods, the n-gram-based method looks promising.

References

[1] H. Fujii and W. B. Croft. A comparison of indexing techniques for Japanese text retrieval. In *Proc. of 16th ACM SIGIR Conf.*, pages 237–246, 1993.

[2] G. Jones, T. Sakai, et al. Experiments in Japanese text retrieval and routing using the NEAT system. In *Proc. of 21st ACM SIGIR Conf.*, pages 197–205, 1998.

[3] N. Kando, K. Kageura, M. Yoshioka, and K. Oyama. Phrase processing methods for Japanese text retrieval. *ACM SIGIR Forum*, 32(2):23–28, 1998.

[4] Y. Ogawa. Pseudo-frequency method: an efficient document ranking retrieval method for n-gram indexing. In *Proc. of 23rd ACM SIGIR Conf.*, pages 321–323, 2000.

[5] Y. Ogawa, H. Mano, M. Narita, and S. Honma. Structuring and expanding queries in the probabilistic model. In *Proc. of 8th TREC*, pages 541–548, 2000.

[6] Y. Ogawa and T. Matsuda. Optimizing query evaluation in n-gram indexing. In *Proc. of 21st ACM SIGIR Conf.*, pages 367–368, 1998.

[7] S. Robertson and S. Walker. On relevance weights with little relevance information. In *Proc. of 20th ACM SIGIR Conf.*, pages 16–24, 1997.