Ricoh at CLEF 2003

Yuichi Kojima, Hideo Itoh, Hiroko Mano and Yasushi Ogawa
Software R&D Group, RICOH CO., Ltd.
1-1-17 Koishikawa, Bunkyo-ku, Tokyo 112-0002, JAPAN
{ykoji,mano,hideo,yogawa}@src.ricoh.co.jp

Abstract

This paper describes the participation of RICOH in the Monolingual Information Retrieval tasks of the Cross-Language Evaluation Forum (CLEF) 2003. We used our system with same kind of stemmer, same options and different parameters for 5 European languages to compare each result. Total performance of the system was reasonable. For French, German and Italian, we found some problems.

1 Introduction

For CLEF 2003 monolingual information retrieval task, RICOH submitted runs for French, German, Italian, Spanish and Dutch. We have worked on English and Japanese text retrieval in past few years [2,3,4,5]. CLEF 2003 experiments were our first trials for European languages. Our main focuses at the experiments were:

1) to test our approach based on the probabilistic model in European languages

2) to find language-specific problems

This paper is organized as follows: Section 2 introduces outline of our system, section 3 shows modifications for the experiments, section 4 describes the results, and section 5 reports some conclusions.

2 System descriptions

Before describing our approach to European languages, we give the system description as background. The basic features of the system are as follows:

□Effective document ranking based on the probabilistic model [8] with query expansion using pseudo-relevance feedback [2]

□Scalable and efficient indexing and search based on the inverted file module [4]

This system was also used for TREC and NTCIR experiments and showed its effectiveness.

In the following sections, we explain the processing flow of the system [5].

2.1 Query term extraction

We used “title” and “description” fields of each topic. Input topic string is transformed into a sequence of stemmed tokens using the tokenizer and the stemmer. Stop words are eliminated using a stopword dictionary. Two kinds of terms are extracted from stemmed tokens for initial retrieval. “single term” is each stemmed token and “phrasal term” consists of two adjacent tokens in the stemmed 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:

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

$$s_{q,d} = \sum_{t \in q} \frac{f_{t,d}}{K + f_{t,d}} \cdot \frac{w_t}{k_4^{'} \cdot N + 1}$$

$$K = k_1 \left((1-b) + b \frac{l_d}{l_{ave}} \right)$$

where 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, 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 initial retrieval, top 10 documents are assumed to be relevant (pseudo-relevant) to the query and selected as a “seed” of query expansion. Candidates of expansion terms are extracted from the seed documents by the same way as in 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 top ranked terms are selected as expansion terms. The weight is re-calculated as $w2_t$ with the Robertson/Sparck-Jones formula [7]

$$RSV_t = w2_t \cdot \left({r_t}/{R} - {n_t}/{N} \right)$$

where R is the number of relevant documents, r_t is the number of relevant documents containing the term t and α is a parameter.

The weight of initial query term is re-calculated with the same formula as above, but with a different α value and an additional adjustment to make the weight higher than expansion terms.

2.4 Final retrieval

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

3 Experiments

There are four items in the system that need adjustments depending on the language, 1) tokenizer, 2) stemmer, 3) stopword dictionary and 4) training data. We used the same tokenizer originally developed for English for all target languages. The others are as follows.

3.1 Stemming

We used Snowball stemmers [1] for all target languages because 1) we didn’t have stemmers for European languages except for English 2) we aren’t familiar these languages to develop stemmers and 3) unlike the earlier result [9], Snowball stemmers showed their reasonable efficiencies for preparatory experiments. Table 1 shows the results using CLEF 2002 data with and without stemming.

Table 1: Average precision with and without stemming using title and description queries

	French	German	Italian	Spanish	Dutch
with stemming	0.4334	0.3701	0.4000	0.4936	0.4187
without stemming	0.3841	0.3392	0.3899	0.4468	0.4023

3.2 Stopword dictionary

We didn’t use stopword dictionaries because we didn’t have them.

3.3 Training

We trained the system by selecting the best parameter-set from 500 candidate parameter-sets for each language to get the highest average precision score.

There was a bug in our training scripts. The system was trained using CLEF 2002 queries and CLEF *2003* data collections, instead of CLEF *2002* data collections. This mismatch resulted in extra noise documents in retrieved documents and made tuning performance rather worse.

Table 2 shows the results with and without training.

Table 2: Average precision with and without training using title and description queries

	French	German	Italian	Spanish	Dutch
without training	0.4334	0.3701	0.4000	0.4936	0.4187
with training using 2002 data	0.4493	0.3746	0.4088	0.5004	0.4371
with training using 2003 data	0.4493	0.3746	0.4018	0.4985	0.4371

4 Results

Table 3 shows summarize of our official results for CLEF 2003. Table 4 shows summarize of our additional results using parameters trained with CLEF *2002* (correct) data collections. The additional result for Dutch is same as formal one because data collection is same. The additional results for French and German are same as formal ones because new parameters selected by correct training scripts were unchanged from formal runs.

Table 3: formal runs for CLEF 2003

Language	Run	Relevant	Rel ret	Average Prec	R-precision	Query Expansion
French	rfrtdp03	946	927	0.4916	0.4697	NO
	rfrtde03	946	928	0.4901	0.4634	YES
German	rdetdp03	1825	1583	0.4425	0.4230	NO
	rdetde03	1825	1693	0.4736	0.4385	YES
Italian	rittdp03	809	761	0.5200	0.4954	NO
	rittde03	809	782	0.5296	0.4868	YES
Spanish	restdp03	2368	2206	0.4727	0.4605	NO
	restde03	2368	2248	0.5174	0.4806	YES
Dutch	rnltdp03	1577	1415	0.4439	0.4206	NO
	rnltde03	1577	1421	0.4719	0.4498	YES

Table 4: Additional runs for CLEF 2003

Language	Run	Relevant	Rel ret	Average Prec	R-precision	Query Expansion
French	rfrtdp03	946	927	0.4916	0.4697	NO
	rfrtde03	946	928	0.4901	0.4634	YES
German	rdetdp03	1825	1583	0.4425	0.4230	NO
	rdetde03	1825	1693	0.4736	0.4385	YES
Italian		809	767	0.5140	0.4874	NO
		809	779	0.5166	0.4829	YES
Spanish		2368	2207	0.4864	0.4719	NO
		2368	2285	0.5293	0.4906	YES
Dutch	rnltdp03	1577	1415	0.4439	0.4206	NO
	rnltde03Re	1577	1421	0.4719	0.4498	YES

5 Conclusions

Our approach was tested and its results were reasonable. According to “comparison to median by topic”, the Spanish result may be good, but the German and French results may not.

We compared results for each language under same conditions. The comparison brought us questions for each language.

- Why query expansion is not effective for French and Italian?
- Why retrieval of some queries failed badly in French and German?

It is likely that our query expansion doesn't work well with few relevant documents. There is a strong correlation between the effectiveness of our expansion and the number of relevant documents for each language. This correlation should be checked with each query.

We think that there are different kind of problems about failure of queries in French and German. For the German result, we expect that the main causes are that we have no German compound splitter. For the French result, we need time to analyze it.

References

[1] Snowball web site. At *http://snowball.tartarus.org/ visited 7th November 2002*.

[2] Y. Ogawa, H. Mano, M. Narita, and S. Honma. Structuring and expanding queries in the probabilistic model. In *The Eighth Text REtrieval Conference (TREC-8),* pages 541-548, 2000.

[3] M. Toyoda, M. Kitsuregawa, H. Mano, H. Itoh and Y. Ogawa. University of Tokyo/RICOH at NTCIR-3 Web Retrieval Task. At *http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings3/NTCIR3-WEB-ToyodaM.pdf*.

[4] Y. Ogawa and H. Mano. RICOH at NTCIR-2. In *Proceedings of the Second NTCIR Workshop Meeting,* pages 121-123, 2001.

[5] H. Itoh, H. Mano and Y. Ogawa. RICOH at TREC-10. In *The Tenth Text REtrieval Conference (TREC-2001),* pages 457-464, 2001.

[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. *In Proceedings of the 20th Annual International ACM SIGIR Conference (SIGIR '97)*, pages 16-24, 1997.

[9] A. MacFarlane. Pliers and snowball at CLEF 2002. In *Working Notes for the CLEF 2002 Workshop*, Rome, Italy, September 2002.