# I2R at NTCIR5

Yang Lingpeng, Ji Donghong
Institute for Infocomm Research
21, Heng Mui Keng Terrace
Singapore 119613
{lpyang, dhji}@i2r.a-star.edu.sg

## Abstract

*The $I^2R$ group participated in the cross-language retrieval task (CLIR) at the fifth NTCIR workshop (NTCIR 5). In this paper, we describe our approach on Single Language Information Retrieval (SLIR) on Chinese language. We use bi-grams as index units and use OKAPI BM25 as retrieval model. The initial retrieved documents are re-ranked before they are used to do standard query expansion.*

*Our document re-ranking method bases on term distribution, which integrates the information from relative document frequency, document position and term length. One advantage is that the term weighting scheme is based on both local and global distributions, which can ensure more meaningful terms for document re-ranking. Another advantage is that we don't need to pre-specify the number of pseudo-relevance documents*

*Experiences show our method achieves 0.5047, 0.5441 mean average precision on T-only run (Title based) at rigid, relax relevant judgment and 0.4826, 0.5249 mean average precision on D-only run (short description based) at rigid, relax relevant judgment in SLIR on Chinese Language.*

**Keywords:** Document Re-ranking, Term Extraction, Chinese Information Retrieval, Query Expansion

## 1. Introduction

At NTCIR 5, we participated in the Cross Lingual Information Retrieval (CLIR) where the query and document set are Chinese language. Readers are referred to [3] to get the information about NTCIR5 and the task description in detail. We submitted two compulsory runs: a T-only run which uses field TITLE (noun or noun phrases about topic) as query and a D-only run which uses field DESC (a short description of topic) as query.

In NTCIR5, we use completely automatic methods. We use OKAPI BM25 as retrieval model and use bi-grams as index units. The initial retrieved results are re-ranked before standard query expansion.

The rest of this paper is organized as following. In section 2, we describe the pre-processing on documents and queries. In section 3, we describe the retrieval model used in our system. In section 4, we describe our document re-ranking method. In section 5, we describe how to do query expansion in our system. In section 6, we evaluate the performance of our proposed method on NTCIR5 and give out some result analysis. In section 7, we present the conclusion and some future work.

## 2. Pre-Processing

Before the normal Chinese IR process, all documents and queries are pre-processed as:

- All documents and queries are converted from BIG-5 code based to GB2312 code based so that we can save indexes space without losing too much precision. The BIG5 to GB2312 mapping is a many to one mapping because there are 13060 Chinese Characters in BIG5 representation but only 6763 Chinese Characters can be represented in GB2312 code. For those BIG5 Chinese Characters which have no mapping in GB2312 code, we assign 0xFEFE (first byte and second byte are 0xFE) as their mapping code in GB2312.

## 3. Retrieval Model

We use bi-grams as index units and use OKAPI BM25 as retrieval model.

For the BM25 model, the relevance between the document and the query is defined in (1)-(3).

$$\sum_{t \in q} w_t \frac{(k_1 + 1)tf_d(t)}{K + tf_d(t)} \frac{(k_3 + 1)tf_q(t)}{k_3 + tf_q(t)} \quad (1)$$

$$\mathrm{w_t} = \log \frac{(N - df(t) + 0.5)}{df(t) + 0.5} \quad (2)$$

$$K = k_1 \times ((1 - b) + b \times \frac{dl}{avdl} \quad (3)$$

where $\mathrm{w_t}$, defined in (2), is the Robertson/Spark Jones weight of $t$. $k_1$, $b$ and $k_3$ are parameters. $k_1$ and $b$ are set as 1.2 and 0.75 respectively by default, and $k_3$ is set as 7. $dl$ and $avdl$ are respectively the document length and average document length measured by the number of the bi-grams.

## 4. Document Re-ranking

How to further improve the rankings of the relevant documents after an initial search has been extensively studied in information retrieval. Such studies include two main streams: automatic query expansion and automatic document re-ranking. While the assumption behind automatic query expansion is that the high ranked documents are likely to be relevant so that the terms in these documents can be used to augment the original query to a more accurate one, document re-ranking is a method to improve the rankings by re-ordering the position of initial retrieved documents without doing a second search. After document re-ranking, it's expected that more relevant documents appear in higher rankings, from which automatic query expansion can benefit.

Our proposed document re-ranking approach bases on term distribution, which integrates the information from relative document frequency, document position and term length.

We use the same term extraction method introduced in NTCIR4 [4] to extract key terms from top 1000 retrieved documents.

To re-rank retrieved documents, we use the key terms in the top 1000 documents, and suppose that these key terms will contribute to the re-rankings. Here, we only focus on the terms which also occur in the queries, which means that we don't use any query expansion. So, the terms can also be referred to as query terms. To weigh a query term, we consider the following three factors.

i) Relative distribution: the ratio of document frequency of a term in the top 1000 retrieved document against the document frequency of the term in the whole document collection.

Intuitively, the more frequently a term occurs in the 1000 documents relative with the whole collection, the more important the term tends to be.

ii) Term length: the number of Chinese characters a term contains.

Intuitively, the longer a term is, the more contribution to the precision the term may have.

iii) Document ranking position: the serial number of a document in top 1000 documents.

Intuitively, the higher ranking a document is, the more important the terms in it tend to be.

With both the local and global information taken into consideration, the weight assigned to a key term $t$ is given by the following formula.

$$\sqrt{\frac{(\sum_{i=1}^{1000} df(t, d_i) \times f(i)) \big/ 1000}{DF(t, C) / R}} \times \sqrt{|t|} \quad (4)$$

$$df(t, d_i) = \begin{cases} 1 & t \in d_i \\ 0 & t \notin d_i \end{cases} \quad (5)$$

where $d_i$ is the i-th (i=1, …, 1000) document, $R$ is the number of total documents in the whole collection $C$, $df(t,d)$ and $df(t,C)$ are the document frequency in $d$ and $C$ respectively, $|t|$ is the length of the key term $t$. $f(i)$, defined in (6), is the weight given to $d_i$, which implies a downgraded document frequency.

$$f(i) = 1 + 1/\mathrm{sqrt}(i) \quad (6)$$

Intuitively, a term gets a lower document frequency if occurring in a lower-ranking document, and a higher one if occurring in a higher-ranking document. This is in contrast of the usual way for document frequency that a document gets 1 count no matter where the document is located in the list.

In the re-ranking phase, for each document $d$ in top 1000 retrieved documents, we first find out the query terms which occur in $d$ and sum the weight of these query terms, then we use the accumulated value and the initial similarity between $d$ and query $q$ to calculate a new ranking score; finally, we use the new ranking score to re-order the 1000 documents.

Figure 1 gives out the pseudo code for the re-ranking procedure.

```
Step 1: Acquiring terms in $q$ and their weights;
    1.1 Extract terms from each document $d$ in top $1000$ retrieved documents; in practice, term extraction from each document is done only once and this process can be considered as a part of indexing.
```

1.2 Collect terms that occur in $q$ and calculate their weights by formula (4) and (5);
Step 2: Re-order the 1000 documents;
2.1. For each document $d_i$ in the 1000 documents, calculate its new ranking value $S_i$ by formula (7), $R_i$ is its initial similarity value;

$$S_i = R_i \times (1 + \sum_{t_j \in q, d_i} W(t_j)) \qquad (7)$$

2.2: Re-order top 1000 retrieved documents by $\{S_1, \ldots, S_i, \ldots, S_{1000}\}$.

Fig. 1 The Procedure of Document Re-Ranking

## 5. Query Expansion

We use re-ranked retrieved documents to do query expansion. We use Robertson's RSV scheme [5] to select 200 bi-grams from top 20 re-ranked documents. We also make use of Rocchio's [2] formula, as improved by Salton and Buckley [1] to perform query expansion. The new query is retrieved again to get the final result.

## 6. Evaluation

We submitted two compulsory runs to NTCIR5: a T-only run which only uses field TITLE as query and a D-only run which only uses field DESC as query. Table 1 and Table 2 list statistical result of mean average precision (MAP) for 50 query topics on relax relevance judgment and rigid relevance judgment. Relax relevance judgment considers high relevant documents, relevant documents and partially relevant documents. Rigid relevance judgment only considers high relevant documents and relevant documents. In table 1 and 2, column [C-C-T] represents Chinese to Chinese T-only run, [C-C-D] represents Chinese to Chinese D-only run; Row [min] represents the minimum MAP among all participants, Row [max] represents the maximum MAP among all participants, Row [med] represents the medium MAP among all participants, Row [ave] represents the average MAP of all participants, and Row [$I^2R$] represents our group's MAP result.

From the statistical results, for T-only run, our group achieves 0.5047 and 0.5441 MAP on rigid and relax relevance judgment; for D-only run, our group achieves 0.4826 and 0.5249 MAP on rigid and relax relevance judgment.

Table 1 Statistics on Rigid Judgment

| | C-C-T | C-C-D |
|---|---|---|
| min | 0. 0086 | 0. 0061 |
| max | 0. 5047 | 0. 4826 |
| med | 0. 3069 | 0. 3223 |
| ave | 0. 2874 | 0. 2986 |
| I2R | 0. 5047 | 0. 4826 |

Table 2 Statistics on Relax Judgment

| | C-C-T | C-C-D |
|---|---|---|
| min | 0. 0112 | 0. 0113 |
| max | 0. 5441 | 0. 5249 |
| med | 0. 3576 | 0. 3839 |
| ave | 0. 3319 | 0. 3523 |
| I2R | 0. 5441 | 0. 5249 |

Comparing our retrieved results with official released judgments, we find we get poor results on several individual query topics, such as topic 20, topic 26 and topic 31. To demonstrate what problems we encountered, we list these query topics as following:

```
<TOPIC>
<NUM>020</NUM>
<SLANG>KR</SLANG>
<TLANG>CH</TLANG>
<TITLE>变性，青蛙，鱼</TITLE>
</TOPIC>

<TOPIC>
<NUM>026</NUM>
<SLANG>KR</SLANG>
<TLANG>CH</TLANG>
<TITLE>捐献，百万富翁，遗产</TITLE>
</TOPIC>

<TOPIC>
<NUM>031</NUM>
<SLANG>KR</SLANG>
<TLANG>CH</TLANG>
<TITLE>细微尘粒，心脏疾病</TITLE>
</TOPIC>
```

Analyzing the content of official released relevant documents, we find the problem is mostly caused by the limitation of only using bi-grams as index units. For example, 鱼(fish) in topic 20 cannot be indexed by bi-gram. Another example is, while 捐献(donation) in topic 26 seldom occurs in relevant documents, 捐赠

(donation) occurs in most relevant documents but it doesn't occur in query topic. Such problem may be solved by using both single Chinese Characters as index units and using bi-grams as index units.

Table 3-4 list the statistical results on different retrieval stages. In table 3-4, row [INI] represents the initial results; row [DR] represents the re-ranking result based on initial retrieval results; row [DR + QE] represents the second round retrieval results after query expansion based on re-ranked documents in initial retrieval.

Table 3 Statistics on T-run

| | Rigid | Relax |
|---|---|---|
| INI | 0. 2928 | 0. 3551 |
| DR | 0. 3473 | 0. 4251 |
| DR+QE | 0. 5047 | 0. 5441 |

Table 4 Statistics on D-run

| | Rigid | Relax |
|---|---|---|
| INI | 0. 2685 | 0. 3326 |
| DR | 0. 3187 | 0. 3725 |
| DR+QE | 0. 4826 | 0. 5249 |

Our experiments show that using bi-grams as index units can produce comparable results but it's not enough. Our experiments also show document re-ranking can improve the effectiveness of retrieved documents and can help query expansion to produce better results.

## 7. Conclusion

In this paper, we introduce our approach for Chinese IR and our experience in participating in SLIR in NTCIR5. Our system achieves 0.5047 and 0.5441 MAP on rigid and relax relevance judgment for T-only run and 0.4826 and 0.5249 MAP on rigid and relax relevance judgment for D-only run.

Our experimental results show that proper document re-ranking can improve the precision of top retrieved documents and further improve the effectiveness of query expansion.

Our experimental results also show that using bi-grams as index units is not enough. Combing single Chinese characters as index units and bi-grams as index units may produce better results. In future, we'll do some experiments to find a way to combine different index units.

## References

[1]G.Salton, C. Buckley. *Improving Retrieval Performance by relevance feedback*. J. Am. Soc. Inf. Sci. 41, 288-297. 1990.

[2] J. Rocchio. *Relevance Feedback in Information Retrieval*. In the SMART Retrieval System – Experiments in Automatic Document Processing. G.Salton, Ed., Prentice Hall, Englewood Cliffs, N.J. 1971

[3] K. Kishida. K. H. Chen, S. Lee, K. Kuriyama, N. Kando, H. H. Chen, S. H .Myaeng. *Overview of CLIR task at the fifth NTCIR Workshop*. In Working Notes of the Fifth NTCIR Workshop Meeting, 2005

[4] L.P. Yang, D.H. Ji, L. Tang. *Chinese Information Retrieval Based on Terms and Ontology*. In the Fourth NTCIR Workshop.

[5] S. E., Robertson. *On Term Selection for Query Expansion*. Journal of Documentation 46. Dec 1990, pp 359-364.

[6] S.E. Robertson, S. Walker, and M. Sparck Jones, *Okapi at TREC-3*. Proc. of Third Text Retrieval Conference (TREC-3), 1995.