# Document Re-ranking Based on Automatically Acquired Key Terms in Chinese Information Retrieval

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

## Abstract

For Information Retrieval, users are more concerned about the precision of top ranking documents in most practical situations. In this paper, we propose a method to improve the precision of top *N* ranking documents by reordering the retrieved documents from the initial retrieval. To reorder documents, we first automatically extract *Global Key Terms* from document set, then use extracted *Global Key Terms* to identify *Local Key Terms* in a single document or query topic, finally we make use of *Local Key Terms* in query and documents to reorder the initial ranking documents. The experiment with NTCIR3 CLIR dataset shows that an average 10%-11% improvement and 2%-5% improvement in precision can be achieved at top 10 and 100 ranking documents level respectively.

## 1 Introduction

Information retrieval (IR) is used to retrieve relevant documents from a large document set for a given query where the query is a simple description by natural language. In most practical situations, users concern more on the precision of top ranking documents than recall because users want to acquire relevant information from the top ranking documents.

Traditionally, IR system uses a one-stage or a two-stage mechanism to retrieve relevant documents from document set. For one stage mechanism, IR system only does an initial retrieval. For two-stage mechanism, besides the initial retrieval, IR system will make use of the initial ranking documents to automatically do query expansion to form a new query and then use the new query to retrieve again to get the final ranking documents. The effectiveness of query expansion mainly depends on the precision of top *N* (*N*<50) ranking documents in initial retrieval because almost all proposed automatic query expansion algorithms make use of the information in the top *N* retrieved. Figure 1 demonstrates the general processes of a two-stage IR system.

In this paper, we propose a method to improve the precision of top *N* ranking documents by reordering the initially retrieved documents in the initial retrieval. To reorder documents, we first automatically extract *Global Key Terms* from the document set, then use the extracted *Global Key Terms* to identify *Local Key Terms* in a single document or query topic, finally we make use of the *Local Key Terms* in queries and documents to reorder the initial ranking documents.

Although our method is general and can apply to any languages, in this paper we'll only focus on the research on Chinese IR system.

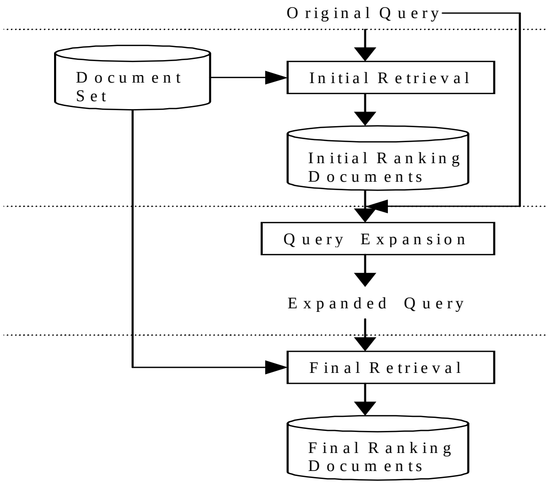

**Fig**. 1 Traditional Process of two-stages IR

The rest of this paper is organized as following. In section 2, we give an overall introduction of our proposed method. In section 3, we talk about what are *Global Key Terms* and what are *Local Key Terms* and how to acquire them. In section 4, we describe how these terms apply to Chinese IR system to improve the precision and quality of IR system. In section 5, we evaluate the performance of our proposed method and give some result analysis. In section 6, we present the conclusion and some future work.

## 2 Overview of Document Reordering in Chinese IR

For Chinese IR, many retrieval models, indexing strategies and query expansion strategies have been studied and successfully used in IR. Chinese Character, bi-gram, n-gram (n>2) and word are the most used indexing units. (Li. P. 1999) gives out many research results on the effectiveness of single Chinese Character as indexing unit and how to improve the effectiveness of single Chinese Character as indexing unit. (K.L. Kwok. 1997) compares three kinds of indexing units (single Character, bigram and short-words) and their effectiveness. It reports that single character indexing is good but not sufficiently competitive, while bi-gram indexing works surprisingly well and it's as good as short-word indexing in precision. (J.Y. Nie, J. Gao, J. Zhang and M. Zhou. 2000) suggests that word indexing and bi-gram indexing can achieve comparable performance but if we consider the time and space factors, then it is preferable to use words (and characters) as indexes. It also suggests that a combination of the longest-matching algorithm with single character is a good method for Chinese and if there is unknown word detection, the performance can be further improved. Many other papers in literature (Palmer, D. and Burger, J, 1997; Chien, L.F, 1995) give similar conclusions. Although there are still different voices on if bi-gram or word is the best indexing unit, bi-gram and word are both considered as the most important top two indexing units in Chinese IR and they are used in many reported Chinese IR systems and experiences.

There are mainly two kinds of retrieval models: Vector Space Model (G. Salton and M. McGill, 1983) and Probabilistic Retrieval (N. Fuhr, 1992). They are both used in a lot of experiments and applications.

For query expansion, almost all of the proposed strategies make use of the top *N* documents in initial ranking documents in the initial retrieval. Generally, query expansion strategy selects *M* indexing units (*M*<50) from the top *N* (N<25) documents in initial ranking documents according to some kind of measure and add these *M* indexing units to original query to form a new query. In such process of query expansion, it's supposed that the top *N* documents are related with original query, but in practice, such an assumption is not always true. The Okapi approach (S.E. Roberson and S.Walker, 2001) supposes that the top *R* documents are related with query and it selects *N* indexing unit from the top *R* documents to form a new query, for example, *R*=10 and *N*=25. (M. Mitra., Amit. S. and Chris. B, 1998) did an experiment on different query topics and it is reported the effectiveness of query expansion mainly depends on the precision of the top *N* ranking documents. If the top *N* ranking documents are highly related with the original query, then query expansion can improve the final result. But if the top *N* documents are less related with the original query, query expansion cannot improve the final result or even reduces the precision of final result. These researches conclude that whether query expansion is successful or not mainly depends on the quality of top *N* ranking documents in the initial retrieval.

The precision of top *N* documents in the initial ranking documents depends on indexing unit and retrieval models and mainly depends on indexing unit. As discussed above, bi-gram and word both are the most effective indexing units in Chinese IR.

Other effort has been done to improve the precision of top *N* documents. (Qu. Y, 2002) proposed a method to re-rank initial relevant documents by using individual thesaurus but the thesaurus must be constructed manually and depends on each query topic.

In this paper, we propose a new method to improve the precision of top *N* ranking documents in initial ranking documents by reordering the top *M* (*M* > *N* and *M* < 1000) ranking documents in initially retrieved documents. To reorder documents, we try to find long terms (more than 2 Chinese characters) that generally represent some complete concepts in query and documents, then we make use of these long terms to re-weight the top *M* documents in initial ranking documents and reorder them by re-weighted value. We adopt a two-stage approach to acquire such kinds of long terms. Firstly, we acquire *Global Key Term*s from the whole document set; secondly, we use *Global Key Term*s to acquire *Local Key Term*s in a query or a document. After we have acquired *Local Key Terms*, we use them to re-weight the top *M* documents in initial ranking documents. Figure 2 demonstrates the processes of an IR system that integrates with this new method.

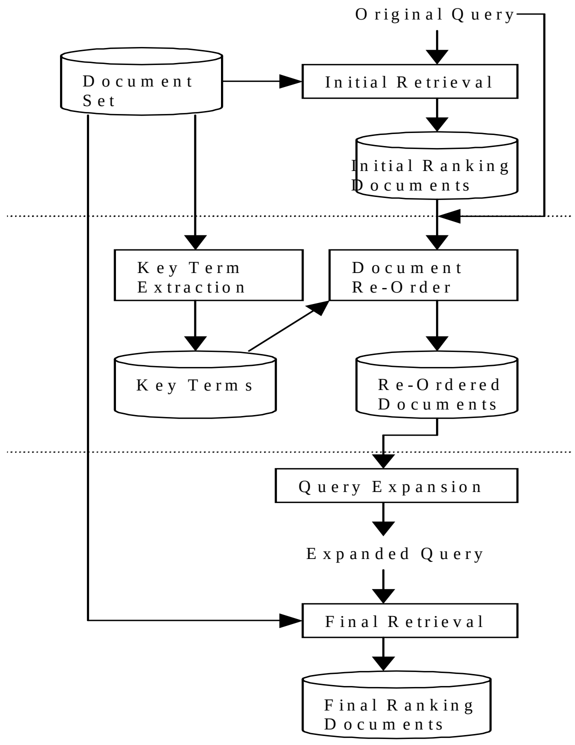

**Fig**. 2 Enhanced Process of IR

## 3 Global/Local Key Term Extraction

The *Global* /*Local Key Term* extraction concerns the problem of what is a *key term*. Intuitively, *key terms* in a document are some conceptual terms that are prominent in document and play main roles in discriminating the document from other documents. In other words, a *key term* in a document can represent part of the content of the document. Generally, from the point of the view of conventional linguistic studies, *Key Term*s may be some NPs, NP-Phrases or some kind of VPs, adjectives that can represent some specific concepts in document content representation.

We define two kinds of *Key Terms*: *Global Key Terms* which are acquired from the whole document set and *Local Key Terms* which are acquired from a single document or a query.

We adopt a two-stage approach to automatically acquire *Global Key Term*s and *Local Key Term*s. In the first stage, we acquire *Global Key Term*s from document set by using a seeding-and-expansion method. In the second stage, we make use of acquired *Global Key Term*s to find *Local Key Term*s in a single document or a query.

### 3.1 Global Key Terms

*Global Key Term*s are terms which are extracted from the whole document set and they can be regarded to represent the main concepts of document set.

Although the definition of *Global Key Term*s is difficult, we try to give some assumptions about a *Global Key Term*. Before we give these assumptions, we first give out the definition of *Seed* and *Key Term* in a document (or document cluster) *d*.

The concept *Seed* is given to reflect the prominence of a Chinese Character in a document (or document cluster) in some way.

Suppose *r* is the reference document set (reference document set including document set and other statistical large document collection), *d* is a document (or a document set), *w* is an individual Chinese Character in *d*, let $P_r(w)$ and $P_d(w)$ be the probability of *w* occurring in *r* and *d* respectively, we adopt 1), *relative probability* or *salience* of *w* in *d* with respect to *r* (Schutze. 1998), as the criteria for evaluation of *Seed*.

1) $P_d(w) / P_r(w)$

We call *w* a *Seed* if $P_d(w) / P_r(w) \geq \delta (\delta > 1)$.

Now we give out the assumptions about a *Key Terms* in document *d*.

i) a *Key Term* contains at least a *Seed*.
ii) a *Key Term* occurs at least *N (N>1)* times in *d*.
iii) the length of a *Key Term* is less than *L* (*L*<30).
iv) a *maximal character string* meeting i), ii) and iii) is a *Key Term*.
v) for a *Key Term*, a *real maximal substring* meeting i), ii) and iiI) without considering their occurrence in all those *Key Terms* containing it is also a *Key Terms*.

Here a *maximal character string* meeting i), ii) and iii) refers to a adjacent Chinese character string meeting i), ii) and iii) while no other longer Chinese character strings containing it meet i), ii) and iii). A *real maximal substring* meeting i), ii) and iii) refers to a real substring meeting i), ii), and iii) while no other longer real substrings containing it meet i), ii) and iii).

We use a kind of seeding-and-expansion-based statistical strategy to acquire *Key Term*s in document (or document cluster), in which we first identify seeds for a *Key Term* then expand from it to get the whole *Key Term*.

Fig. 3 describes the procedure to extract *Key Terms* from a document (or document cluster) *d*.

```
let $F_d(t)$ represents the frequency of $t$ in $d$;
let $N$ is a given threshold ($N$>1);
$K$ = {};
collect $Seeds$ in $d$ into $S$;
for all $c \in S$
{
    let $Q = \{t$: $t$ contains $c$ and $F_d(t) \geq N\}$;
    while $Q \neq NIL$
    {
      $max\text{-}t$ $\leftarrow$ the longest string in $Q$;
      $K \leftarrow K + \{ max\text{-}t \}$;
      Remove $max\text{-}t$ from $Q$;
      for all other $t$ in $Q$
      {
          if $t$ is a substring of $max\text{-}t$
          {   $F_d(t) \leftarrow F_d(t) - F_d(max\text{-}t)$;
              if $F_d(t)<N$
                 removing $t$ from $Q$;
          }
      }
    }
}
return $K$ as $Key\ Terms$ in document $d$;
```

**Fig. 3** Key Term Extraction from document *d*

To acquire *Global Key Term*s, we first roughly cluster the whole document set *r* into *K (K<2000)* document clusters, then we regard each document cluster as a large document and apply our proposed *Key Term* Extraction algorithm (see Fig. 3) on each document cluster and respectively get *Key Terms* in each document cluster. All these *Key Terms* from document clusters form *Global Key Terms*.

There are many document clustering approaches to cluster document set. K-Means and hierarchical clustering are the two usually used approaches. In our algorithm, we don't need to use complicated clustering approaches because we only need to roughly cluster document set *r* into *K* document clusters. Here we use a simple K-Means approach to cluster document set. Firstly, we pick up randomly 10**K* documents from document set *r;* secondly, we use K-Means approach to cluster these 10**K* documents into *K* document clusters; finally, we insert every other document into one of the *K* document clusters. Fig. 4 describes the general process to cluster document set *r* into *K* document clusters.

```
let $K$ is the number of documnet clusters to get;
$T \leftarrow 10*K$ documents randomly pickuped from $r$;
cluster $T$ into $K$ clusters $\{K_j\}$ by using K-Means;
for any document $d$ in $\{r\text{-}T\}$
{
   $K_i \leftarrow$ document cluster which has the maximal similarity with d;
   insert $d$ to document cluster $K_i$;
}
return $K$ document clusters $\{K_j|1<=j<=K\}$;
```

**Fig. 4** Cluster document set *r* into *K* clusters

Fig. 5 describes the procedure to acquire *Global Key Terms* from document set *r*.

```
roughly cluster document set $r$ to $K$ document clusters $\{K_j|1<=j<=K\}$ (See Fig. 4);
$G$ = {};
for each $K_j$
{
    extract $Key\ Terms$ $g$ from $K_j$ ; (See Fig. 3)
    $G \leftarrow G + g$;
}
return $G$ as $Global\ Key\ Terms$ in document set r;
```

**Fig. 5** Global Key Terms Acquisition

In the processing of *Global Key Terms* acquisition, the frequency of each *Global Key Term* is also recorded for further use in identifying *Local Key Terms* - terms in a single document or query.

## 3.2 Local Key Terms

Unlike *Global Key Terms*, *Local Key Terms* are not extracted by using *Key Term* extraction algorithm from single document or query, they are identified based on *Global Key Terms* and their frequencies.

Fig.6 describes the procedure of *Local Key Terms* acquisition from a single document or query *d*.

```
Given threshold $M$ ($M$>10), $N$ ($N$>100) and document $d$;
$L$ = {};
collect Global Key Terms occurred in $d$ and their frequency in document set $r$ into $S$ = <$c$, $tf$>;
for all <$c$, $tf$>$\in S$
{
    if $tf < M$
        remove <$c$, $tf$> from $S$;
};
for all <$c$, $tf$>$\in S$
{
    if $c = c_1c_2$ and <$c_1$, $tf_1$>$\in S$ and <$c_2$, $tf_2$>$\in S$
        if ($tf_1 > tf*N$ and $tf_2 >> tf*N$)
            remove <$c$, $tf$> from $S$;
};
while $S \neq NIL$
{
    let $Q$ = {<$t$, $tf$>: $t$ is the longest string is $S$};
    find <max-c,max-tf> in $Q$ where max-tf has the maximum value;
    remove <max-t, max-tf> from $S$;
    if max-t occurs in $d$
    { $L \leftarrow L$ + max-t;
      remove all occurrance of max_t in $d$;
      for all <$b$, tf-b>$\in S$ where $b$ is a substring of max-t;
          if tf-b < max-tf remove <$b$,tf-b> from $S$;
    }
};
return $L$ as Local Key Terms in document $d$;
```

**Fig. 6** Local Key Terms Acquisition

Following are some examples of *Global Key Terms* and *Local Key Terms* in a query.

**Example**:

Query: 查询故宫博物院所举办之千禧汉代文物大展相关内容

(Find information of the exhibition "Art and Culture of the Han Dynasty" in the National Palace Museum)

*Global Key Terms* occurred in Query and their frequencies in document set:

查询 (Cha2 Xun2)– 4948
故宫 (Gu4 Gong1)– 3456
故宫博物院(Gu4 Gong1 Bo2 Wu4 Yuan4)– 727
博物院(Bo2 Wu4 Yuan4) – 772
院所(Yuan4 Suo3) – 2991
举办(Zhu3 Ban4) – 38698
千禧(Qian1 Xi3)– 11510
汉代(Han4 Dai4) – 411
汉代文物(Han4 Dai4 Wen3 Wu4) - 173
汉代文物大展(Han4 Dai4 Wen3 Wu4 Da4 Zhan3) – 133
文物(Wen3 Wu4) – 7088
文物大展(Wen3 Wu4 Da4 Zhan3) – 158
大展(Da4 Zhan3) – 2270
相关(Xiang3 Guan3) – 67990
相关内容(Xiang3 Guan3 Nei3 Rong2) – 148
内容(Nei3 Rong2) – 31165

*Local Key Terms* in Query:

汉代文物大展(Han4 Dai4 Wen3 Wu4 Da4 Zhan3)
汉代文物(Han4 Dai4 Wen3 Wu4)
文物(Wen3 Wu4)
大展(Da4 Zhan3)
故宫博物院(Gu4 Gong1 Bo2 Wu4 Yuan4)
博物院(Bo2 Wu4 Yuan4)
故宫(Gu4 Gong1)
相关(Xiang3 Guan3)
内容(Nei3 Rong2)
举办(Zhu3 Ban4)
千禧(Qian1 Xi3)
查询(Cha2 Xun2)

From the example, we can see the difference between *Global Key Terms* and *Local Key Terms*. For example, 院所(Yuan4 Suo3) and 文物大展(Wen3 Wu4 Da4 Zhan3) are *Global Key Terms*, but they are not the *Local Key Terms* of query.

# 4 Document Reordering

After we have acquired *Global Key Terms* in document set and *Local Key Terms* in every document and query, we make use of them to reorder the top *M* (*M*<=1000) documents in initial ranking documents. Suppose *q* is a query, Fig. 7 is the algorithm to reorder top *M* documents in initial ranking documents where *w(t)* is the weight assigned to *Local Key Term t*. *w(t)* can be assigned different value by different measures. For example,

i) *w(t)* = the length of *t;*
ii) *w(t)* = the number of Chinese Characters in *t;*
iii) *w(t)* = square root of the length of *t;*

iv) *w(t)* = square root of the number of Chinese Characters in *t;* (default)

```
for each document $d$ in top $M$ ranking documents
    $sim$ ← similary value between $d$ and $q$;
    $w$ ← 0;
    for each Local Key Term $t$ in query $q$;
    { if $t$ is a Local Key Term of $d$
        $w$ ← $w$ + weight($t$) };
    if ($w > 0$)
    { $sim$ ← $sim * w$;
      set $sim$ as the new similary between $d$ and $q$ };
  reorder top $M$ documents by their new similarity values with query $q$;
```

**Fig. 7** Process of Document Reordering

## 5 Experience & Evaluation

We make use of the Chinese document set CIRB011 (132,173 documents) and CIRB20 (249,508 documents) and D-run type query topic set (42 topics) of CLIR in NTCIR3 (see http://research.nii.ac.jp/ntcir-ws3/work-en.html for more information) to evaluate our proposed method. We use vector space model as our retrieval model and use cosine to measure the similarity between document and query. For indexing units, we use bigrams and words respectively. To measure the effectiveness of IR, we use the same two kinds of relevant measures: relax-relevant and rigid-relevant. A document is rigid-relevant if it's highly relevant or relevant with a query, and a document is relax-relevant if it is high relevant or relevant or partially relevant with a query. We also use PreAt10 and PreAt100 to represent the precision of top 10 ranking documents and top 100 ranking documents.

When we use our proposed method and algorithm to extract *Global Key Terms* from document set *r*, we set all kinds of algorithm parameters as following:

- 10000 documents from *r* to do initial document clustering*;* (Fig. 4)
- 1000 document clusters; (Fig. 4)
- maximal length of *Key Terms*:30; (Fig. 3)
- minimal occurrence of *Key Terms*:2; (Fig. 3)
- minimum salience of *seed*:2; (Fig. 3)
- reorder the top 1000 documents;
- We also set *M* =10, *N= 100* for the algorithm to acquire *Local Key Terms*. (Fig. 6)

Table 1 lists the normal results and enhanced results based on bigram indexing. The enhanced results are acquired by using our method to enhance the effectiveness. PreAt10 is the average precision of 42 queries in precision of top 10 ranking documents, while PreAt100 is the average precision of 42 queries in precision of top 100 ranking documents. Column 2 (normal) displays the precision of normal retrieval, column 3 (Enhanced) displays the precision of using our proposed approach, and column 4 (ratio) displays the ratio of column 3 (enhanced) compared with column 2 (normal). Table 2 lists the normal results and our enhanced results based on word indexing.

| | Normal | Enhanced | Ratio |
|---|---|---|---|
| PreAt10(Relax) | 0.3642 | 0.4052 | 1.11258 |
| PreAt100(Relax) | 0.1886 | 0.1926 | 1.02121 |
| PreAt10(Rigid) | 0.2595 | 0.2871 | 1.10636 |
| PreAt100(Rigid) | 0.1278 | 0.133 | 1.04069 |

Table 1 Precision (bigram as indexing unit)

| | Normal | Enhanced | Ratio |
|---|---|---|---|
| PreAt10(Relax) | 0.3761 | 0.4119 | 1.09519 |
| PreAt100(Relax) | 0.1983 | 0.2074 | 1.04589 |
| PreAt10(Rigid) | 0.269 | 0.2952 | 1.0974 |
| PreAt100(Rigid) | 0.1381 | 0.1419 | 1.02752 |

Table 2 Precision (word as indexing unit)

From table 1, we can see that compared with bigrams as indexing units, our proposed method can improve PreAt10 by 11% from 0.3642 to 0.4052 in relax relevant measure and improve 11% from 0.2595 to 0.2871 in rigid relevant measure. Even in PreAt100 level, our method can improve 2% and 4% in relax relevant and rigid relevant measure. Fig. 8 displays the PreAt10 values of each query in relax relevant measure based on bigram indexing where the red lines represent the precision enhanced with our method while the black lines represent the normal precision. Among the 42 query topics, there are only 5 queries whose enhanced precisions are worse than normal precisions, the precisions of other 37 queries are all improved.

From table 2, using words as indexing units (we use a dictionary which contains 80000 Chinese items to segment Chinese document and query), our method can improve PreAt10 by 10% from 0.3761 to 0.4119 in relax relevant measure and improve 10% from 0.269 to 0.2952 in rigid relevant measure. Even in PreAt100 level, our method can improve 3% and 5% in rigid and relax relevant measure.

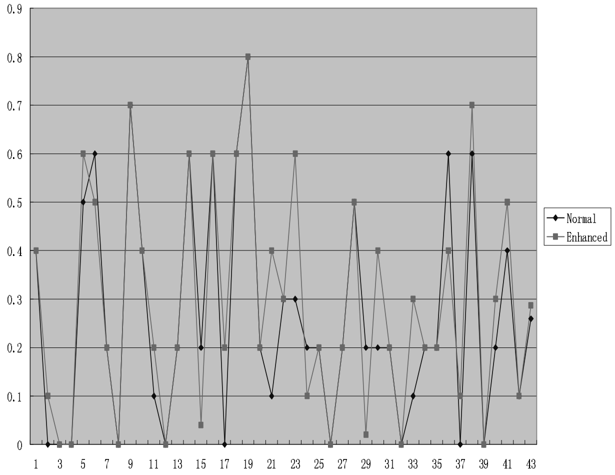

Fig. 8 PreAt10 of all queries in relax judgment

In our experiments, compared with the most important and effective Chinese indexing units: bigram and words, our proposed method improves the average precision of all queries in top 10 measure levels for about 10%. What lies behind our proposed method is that in most case, proper long terms may contain more information (position and Chinese Character dependence) and such information can help us to focus on relevant documents. Our experiment also shows improper long terms may decrease the precision of top documents. So it's very important to extract right and proper terms in documents and queries.

## 6 Conclusion

In this paper, we proposed a new method to improve the precision of top *N* initial ranking documents in Chinese IR. We try to find proper and important long terms in queries and documents, then we make use of these information to reweight the similarity between queries and documents and finally reorder the top *M* (*M*>*N*) documents by their new similarities with query. Our experiences based on bigram as indexing and word as indexing both show that our method can improve the performance of Chinese IR by 10%-11% at top 10 documents measure level and 2%-5% at top 100 documents document measure level. For the further work, we will try to improve the quality of *Global Key Terms* and *Local Key Terms*, and we will apply our method to English IR and other languages IR systems.

## References

Chien, L.F. *Fast and quasi-natural language search for gigabytes of Chinese texts*. In: Proc. 18th ACM SIGIR Conf. On R&D in IR. Fox, E., Ingwersen, P. & Fidel, R. (eds.) ACM: NY, NY. Pp.112-120.

G. Salton and M. McGill. *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.

H. Schutze. 1998. *The hypertext concordance: a better back-of-the-book index*. Proceedings of First Workshop on Computational Terminology. pp: 101-104.

J.Y. Nie, J. Gao, J. Zhang and M. Zhou. 2000. *On the Use of Words and N-grams for Chinese Information Retrieval*. In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages, IRAL-2000, pp. 141-148

K.L. Kwok. 1997. *Comparing Representation in Chinese Information Retrieval*. In Proceedings of the ACM SIGIR-97, pp. 34-41.

Li. P. 1999.*Research on Improvement of Single Chinese Character Indexing Method*, Journal of the China Society for Scientific and Technical Information, Vol. 18 No. 5.

M. Mitra., Amit. S. and Chris. B. *Improving Automatic Query Expansion*. In Proc. ACM SIGIR'98, Aug. 1998.

N. Fuhr. *Probabilistic Models in Information Retrieval*. The Computer Journal. 35(3), 1992.

Palmer, D. and Burger, J. *Chinese Word Segmentation and Information Retrieval*. AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, 1997

Qu Y., Xu G. and Wang J. *Rerank Method based on Individual Thesaurus*. In NTCIR Workshop 2.

S.E. Robertson and S. Walker. *Microsoft Cambridge at TREC-9: Filtering track*: In the Eight Text Retrieval Conference (TREC-8), pages 151-161, Gaithersburg, MD, 2001.