Chinese Information Retrieval Based on Terms and Ontology

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

The I^2R group participated in the cross-language retrieval task (CLIR) at the fourth NTCIR workshop (NTCIR 4). In this paper, we describe our approach on Single Language Information Retrieval (SLIR) on Chinese language. Firstly, we automatically extract terms (short-terms and long terms) from document set and use them to build indexes; secondly, we use short terms in query and documents to do initial search to get initial ranking documents; thirdly, we make use of ontology knowledge, long terms in query, relevant terms of terms in query, and top N documents in initial ranking documents to do query expansion to get a new query; fourthly, we use the new query to search again to get final ranking documents; finally, we use term coverage and event detection to adjust final ranking documents to reorder top N documents in final ranking documents. Experiences show our method achieves 31.46%, 37.99% mean average precision on T-only run (Title based) at rigid, relax relevant judgment and 32.55%, 38.80% mean average precision on D-only run (short description based) at rigid, relax relevant judgment in SLIR on Chinese Language.

Keywords: Ontology, Term Extraction, Chinese Information Retrieval, Query Expansion, Event Detection, Term Coverage

1. Introduction

At NTCIR 4, we participated in the Cross Lingual Information Retrieval (CLIR) where the query and document set are Chinese language. Readers are referred to [2] to get the information about NTCIR4 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.

For Chinese Information Retrieval, 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. 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 are done in [5]. K.L. Kwok. compared three kinds of indexing units (single Character, bigram and short-words) and their effectiveness in [4]. 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 in [3] suggest 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 ([8, 9]) give similar conclusions. 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 in NTCIR tracks.

There are mainly two kinds of retrieval models for Chinese Information Retrieval: Vector Space Model [10] and Probabilistic Retrieval [7]. They are both mainly used the experiences in NTCIR.

For query expansion, almost all of the proposed strategies make use of the top N documents in initial ranking documents in the initial search. 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.

In NTCIR4, we use the similar approach we used in NTCIT 3 as our basic approach [1] and add many new mechanisms to improve the effectiveness of our approach. Firstly, we automatically extract terms (short-terms and long terms) from document set and use them to build indexes; secondly, we use short terms in query and documents to do initial search to get initial ranking documents; thirdly, we make use of ontology knowledge, long terms in query, relevant terms (co-occurrence terms) of terms in query, and top N documents in initial ranking documents to do query expansion to get a new query; fourthly, we use the new query to search again to get final ranking documents; finally, we use term coverage and event detection to micro-adjust final ranking documents to reorder top documents in final ranking documents. Figure 1 demonstrates the processes of our Chinese IR system.

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 how to automatically extract terms from document set. In section 4, we describe the retrieval model and weighting scheme used in our system. In section 5, we describe how to do query expansion in our system. In section 6, we describe how to refine the final ranking documents by using term coverage and event detection. In section 7, we evaluate the performance of our proposed method on NTCIR 4 and give out some result analysis. In section 8, 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 (especial for bigram based indexing) 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.
- All kinds of data formats about date are unified as xxxx 年(year)xx 月(month)xx 日(day).

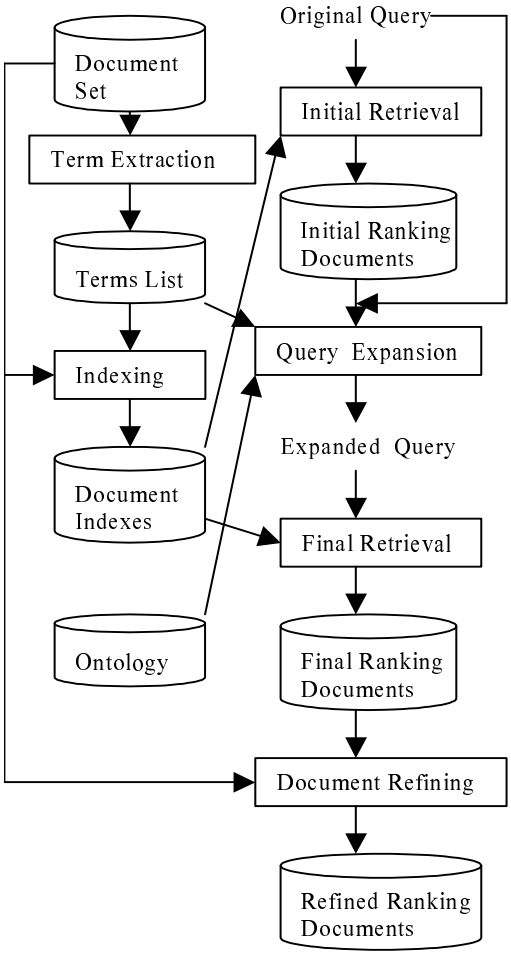

Fig. 1 Process of IR

3. Term Extraction

Although bigram and word are considered the best indexing unit in Chinese IR, we still use automatically extracted terms as indexing units in our NTCIR4 track. The basic method of term extraction is the same we used in NTCIR3. We still use a seeding-and-expansion mechanism to extract terms from documents (or document clusters). Readers are referred to [1] for more information of our term extraction algorithm.

To acquire terms, 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 term extraction

algorithm [1] on each document cluster and respectively get terms in each document cluster. All these terms from different document clusters form the whole terms list. We regard a term whose length is less than 4 Chinese Characters as a short term, and a term whose length is equal or greater than 4 Chinese Characters as a long term. Following is some examples of short terms and long terms.

(1) Short Terms
劳工 (Laborer)
抗议 (protest)
劳委会 (Council of Labor Affairs)
诉求 (Appeal)

(2) Long Terms
约翰走路 (Jonnie Walker)
高尔夫球 (Golf)
老虎伍兹 (Tiger Woods)
胚胎干细胞 (embryonic stem cells)

There are many document clustering approaches to cluster document set. K-Means and hierarchical clustering are the two often 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. 2 describes the process to cluster document set *r* into *K* document clusters.

let *K* is the number of document clusters to get;
T←10**K* documents randomly selected from *r;*
cluster *T* into *K* clusters $\{K_j\}$ by using K-Means;
for any document *d* in {*r*-*T*}
{
K_i← 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. 2 Cluster document set *r* into *K* clusters

In the processing of our CLIR task, we roughly cluster the whole document set (CIRB011: 132,173 documents and CIRB020: 249,508 documents) into 2000 document clusters, then we extract terms from these 2000 document clusters. All of the terms extracted from 2000 document clusters form the whole terms list. Every term in terms list is called a global term because it's extracted based on the whole document set *r*. The term list is considered as an automatically acquired word dictionary, it's used to find terms in a single document or query. To find terms in a single document or query, we make use of a variant of word segmentation method to segment document and query into terms. Unlike traditional word segmentation method, a global term and its sub-string may all be considered as a term in document. For example, if $g = cd$ is a global term where *c* and *d* are also global terms, then *g*, *c* and *d* are all considered as terms in document. Terms acquired from single document are regarded as local terms.

Local terms in documents are used as indexing unit to build index.

4. Retrieval Model and Weighting Scheme

There are mainly two kinds of retrieval models for Chinese Information Retrieval: Vector Space Model and Probabilistic Retrieval. We use Vector Space Model to represent documents and queries. Each document or query is represented as a vector in vector space where each dimension of vector is the weight given to some terms in document or query. The weight of term *t* in document *d* is given by the following TF/IDF weight scheme:

$$w(t, d)=\log(T(t, d)+1) * \log(N/D(t)+1)$$

where, $w(t, d)$ is the weigh given to *t* in *d*, $T(t, d)$ is the frequency of *t* in *d*, *N* is the number of documents in document set, $D(t)$ is the number of documents in document set which contain *t*.

The weight of term *t* in query *q* is given by the following weight scheme:

$$w(t, q) = T(t, q)$$

where, $w(t, q)$ is the weigh given to *t* in *q*, $T(t, q)$ is the frequency of *t* in *q*.

In the initial search, only short terms are used to construct document vectors and query vectors.

The similarity (distance) between a document *d* and a query *q* is calculated by distance between document vector and query vector by cosine measure.

5. Query Expansion

Query expansion is considered as an important supplement to improve the precision of IR. Almost all of the proposed query expansion strategies make use of the top *N* documents in initial ranking documents in the initial search.

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 famous Okapi approach [11] 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 [6] did an experience 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 search.

Our system makes use of the information of top 20 initial ranking documents together with pre-built ontology knowledge, short terms and long terms in query and their relevant terms (co-occurred terms) in document set to do query expansion.

We only build ontology for some short terms by using search engine and manual verifying. Following is an example of the ontology about term 亚洲(Asia).

亚洲(Asia)
(1) 新加坡 (Singapore)
中国(China)
日本(Japan)
韩国(South Korea)
…
(2) 经济危机 (Economic Crisis)
1997 年 (Year 1997)
国际货币基金组织(IMF)
…

We acquire relevant terms of term *t* by their co-occurrence in documents and their Mutual Information. Following is an example of relevant terms.

Term: 地震 (Earthquake);
Relevant Terms:
强度 (Intensity)
伤亡人数 (Casualty)
救援 (Rescue)
…

Following is the method to expand a query *q*:

- All local long terms in *q* are added to new query with frequency in *q* as its weight;
- For each local term *t* in query *q*, O(*t*) is one of the ontology classes about *t*, all terms in O(*t*) will be added into *q* if at least two terms in O(*t*) occur in the top 20 ranking documents with 0.5 as weight.
- For each local term *t* in query *q*, R(*t*) is a relevant term of *t*, R(*t*) will be added into *q* with 0.5 as its weight if R(*t*) occurs at least in two documents among the top 20 ranking documents.

The original query plus new terms acquired by query expansion form a new query. This new query is used to search again to get final search result – final ranking documents.

6. Documents Refining

Document refining is used to micro-rank the top M (M<2000) document in the final ranking documents. The most important ranking methods we used are: ranking by term coverage and ranking by event detection.

The ranking by event detection method tries to find out if a given query is about an event. A query is about an event if the published dates of its relevant documents fall in a period of time. We make use of the date information between <DATE> and </DATE> in the top *N* final ranking documents to detect if the query is about an event. If a query is about an event, we try to find the possible date scope of the event and emphasis the documents that fall in the date scope by give more weights. Our method finds some queries in NTCIR 4 are about events. For example, query 2, 6, 10, 13, 29, 34, 40, 50 and 54. Following lists the content of query 2.

[Query 2]
<TITLE>约翰走路，菁英高尔夫球慈善赛，台湾</TITLE>
<DESC>查询 1999 年来台参加约翰走路菁英高尔夫慈善赛的国际高尔夫球星及相关活动的内容</DESC>
(<TITLE>Jonnie Walker, Charity Golf Tournament, Taiwan</TITLE>
<DESC>Find out who joined the Jonnie Walker Charity Golf Tournament in Taiwan in 1999 and the related activities</DESC>)

To query 2, our system finds out that most relevant documents are published between 1999

年 11 月 09 日 and 1999 年 11 月 14 日（November 9, 1999 and November 14, 1999）, so our system considers this query is about an event and give documents falling in 1999年11月09日 and 1999年11月14日（November 9, 1999 and November 14, 1999）more weight.

Another example is query 13. Following lists the content of query 13.

[Query 13]

<TITLE>日本，首相，小渊惠三，访问，美国</TITLE>

<DESC>查询日本首相小渊惠三访美相关内容</DESC>

<NARR>

<BACK>日本首相小渊惠三於 1998 年四月二十九日启程访美，这是日本首相十二年来首次正式访美。小渊惠三此次的美国行带给美国诸多见面礼，包括双方经济合作对策、国际金援以及美日防卫合作方针配套法案，使得美日安保体制进入新阶段。请查询此次访问活动的内容，包括日本所提出的支援、美日的合作协议等。</BACK>

<REL>相关资料为小渊惠三访美的内容报导，各国的看法或意见则为无关。</REL>

</NARR>

(<TITLE>Japan, Prime Minister, Keizo Obuchi, Visit, The U.S.</TITLE>

<DESC>Find articles pertaining to Japan Prime Minister Keizo Obuchi's visit to the U.S.</DESC>

<NARR>

<BACK>Japanese Prime Minister Keizo Obuchi started his visit to the U.S. on April 29th, 1998. This was the first official visit of a Japanese Prime Minister to the U.S. in 12 years. Prime Minister Keizo Obuchi brought "gifts" to the U.S. in this visit including cooperative guidelines of economic policies, international financial support and a cooperative defense plan to bring a new stage to the security system between Japan and the U.S. Please query the contents of activities of this visit including Japan's offering of support and the cooperative agreement between Japan and the U.S., etc.</BACK>

<REL>Documents about Prime Minister Keizo Obuchi's visit to the U.S. are relevant. Viewpoints and opinions of other countries are not relevant.</REL>

</NARR>)

To query 13, although the background of query tells us that the visit is around April 29th, 1998, our system finds out that most relevant documents are published between 1999年04月27日 and 1999年05月06日（April 27, 1999 and May 6, 1999）not around April 27th, 1998, and our system considers this query is about an event and give documents falling in 1999年04月27日 and 1999年05月06日（April 27, 1999 and May 6, 1999）more weight.

The ranking by term coverage method tries to give more weight to documents that cover more terms of query. Intuitively to say, suppose *c*, *d* and *e* are terms of query *q*, then document *f* is more likely to be relevant with *q* than document *g* if document *f* contains term *c*, *d*, and *e*, but document *g* only contains *c* and *e*.

The weight given for term coverage can be the number of terms coved or the total length of terms coved or other measures.

7. Evaluation

We submitted two compulsory runs to NTCIR4: a T-only run which only uses field TITLE as query and a D-only run which only uses field DESC as query. There are 13 groups who submitted T-only run and D-only run which use Chinese language as query and document set. There are 60 query topics but only 59 query topics are evaluated. Table 1 and Table 2 list statistical result of mean average precision (MAP) for 59 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 13 participants, Row [max] represents the maximum MAP among 13 participants, Row [med] represents the medium MAP among 13 participants, Row [ave] represents the average MAP of 13 participants, and Row [I^2R] represents our group's MAP result.

Table 1 Statistics on Rigid Judgment

	C-C-T	C-C-D
min	0.1327	0.1251
max	0.3146	0.3255
med	0.1881	0.1741
ave	0.1943	0.1826
I2R	0.3146	0.3255

Table 2 Statistics on Relax Judgment

	C-C-T	C-C-D
min	0. 1638	0. 1548
max	0. 3799	0. 388
med	0. 2356	0. 2219
ave	0. 2378	0. 2328
I2R	0. 3799	0. 388

From the statistical results, for T-only run, our group achieves 0.3146 and 0.3799 MAP on rigid and relax relevance judgment; for D-only run, our group achieves 0.3255 and 0.388 MAP on rigid and relax relevance judgment.

Table 3 T-only run at initial retrieval

Topic	PreAt10	PreAt100	PreAt1000
1	0. 3	0. 09	0. 031
2	0. 3	0. 17	0. 022
3	0. 7	0. 18	0. 022
4	0. 7	0. 22	0. 022
5	0. 2	0. 13	0. 013
6	0. 1	0. 15	0. 026
7	0. 3	0. 16	0. 016
8	0. 1	0. 13	0. 053
9	0. 2	0. 04	0. 004
10	0. 5	0. 08	0. 008
11	0. 4	0. 24	0. 045
12	0. 3	0. 12	0. 013
13	0. 1	0. 02	0. 014
14	0. 3	0. 03	0. 006
15	0. 4	0. 16	0. 043
16	0. 1	0. 21	0. 049
17	0. 1	0. 3	0. 067
18	0. 1	0. 11	0. 043
19	0. 1	0. 2	0. 024
20	0	0. 1	0. 016
21	0. 4	0. 12	0. 021
22	0. 5	0. 07	0. 007
23	0. 3	0. 16	0. 03
24	0. 5	0. 17	0. 041
26	0. 1	0. 08	0. 013
27	0. 2	0. 12	0. 036
28	0. 2	0. 08	0. 012
29	0	0. 03	0. 23
30	0. 4	0. 42	0. 063

Although we get the best MAP results on both T-only run and D-only run on rigid relevance judgment and relax relevance judgment, we get poor results on several individual query topics. To find out the problem, we compare our final submitted results with our initial search results without query expansion. Table 3 lists the T-only run result of our initial retrieval on relax relevance judgment.

Table 3 T-only run at initial retrieval (cont')

Topic	PreAt10	PreAt100	PreAt1000
41	0. 1	0. 08	0. 016
42	0. 2	0. 24	0. 027
43	0. 2	0. 05	0. 017
44	0. 2	0. 22	0. 065
45	0. 1	0. 14	0. 052
46	0. 4	0. 15	0. 02
47	0. 1	0. 06	0. 015
48	0. 4	0. 21	0. 021
49	0. 5	0. 17	0. 046
50	0	0. 14	0. 048
51	0. 4	0. 08	0. 013
52	0. 6	0. 06	0. 006
53	0. 3	0. 04	0. 01
54	0. 2	0. 2	0. 05
55	0. 1	0. 14	0. 027
56	0	0. 1	0. 023
57	0. 1	0. 11	0. 015
58	0. 5	0. 13	0. 018
59	0. 7	0. 31	0. 039
60	0. 2	0. 12	0. 038

Comparing our final ranking results with initial ranking results, we find that the final results of query 9, 18, 28, 33 and 58 are worse than initial results. More deeply analysis shows that the problems are mainly caused by improper relevant. For example, for query 33: <TITLE>研究，蛋白质</TITLE> (<TITLE>Research, Protein</TITLE>), we get some relevant terms as:

Term: 蛋白质 (Protein)
Relevant Terms:
营养 (nutrition)
食物 (food)

These relevant terms played negative roles to our retrieval results and reduced the precision of top *N* ranking documents.

Another example is query 58: <TITLE>非接触式智慧卡</TITLE> (<TITLE>Contactless SMART Card</TITLE>). After using query expansion, the precision of top 10 documents (PreAt10) becomes 0.0 from the original 0.5. Analysis shows improper relevant terms caused such problem. Following lists some improper relevant terms for this query:

电子收费 (Dian4 Zi3 Shou1 Fei4)

电子收费系统(Dian4 Zi3 Shou1 Fei4 Xi1 Tong3)

高速公路电子收费系统 (Gao1 Su4 Gong1 Lu4 Dian4 Zi3 Shou1 Fei4 Xi1 Tong3)

8. Conclusion

In this paper, we introduce our approach for Chinese IR and our experience in participating in SLIR in NTCIR4. Our system achieves 0.3146 and 0.3799 MAP on rigid and relax relevance judgment for T-only run and 0.3255 and 0.388 MAP on rigid and relax relevance judgment for D-only run.

Although our system gets significant results in T-only run and D-only run, we find a lot of difficulties to push our approach into actual applications. The most difficult problem is how to acquire proper relevant terms and ontology knowledge. Currently, we semi-manually build ontology by searching from Internet and manually verifying the result, but it's impossible to manually build a complete ontology for all short terms.

Our relevant terms are mainly based on the co-occurrence of documents. Experiences show some relevant terms acquired by this way are not actual relevant with the given term. One possible solution is to detect relevant terms by co-occurrences on paragraphs or on sentences.

In the future, we want to do some deep research on making use of the huge resources in Internet to automatically build ontology. And we also try to improve the effectiveness of relevant terms.

References

[1] D.H. Ji, L.P. Yang, Y. Nie. *Chinese Language IR* based on Term Extraction. In The Third NTCIR Workshop.

[2] K. Kishida, K.H. Chen, S. Lee, K. Kuriyama, et al. *Overview of CLIR Task at the Fourth NTCIR Workshop*. In The Fourth NTCIR Workshop.

[3] 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

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

[5] 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.

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

[7] N. Fuhr. *Probabilistic Models in Information Retrieval*. The Computer Journal. 35(3):243-254, 1992.

[8] Palmer, D. and Burger, J. *Chinese Word Segmentation and Information Retrieval.* AAAI Spring Symposium on Cross-Language Text and Speech Retrieval, Electronic Working Notes, 1997

[9] 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.

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

[11] S.E. Robertson and S. Walker. *Microsoft Cambridge at TREC-9: Filtering track*: In NIST Special Pub. 500-264: The Eight Text Retrieval Conference (TREC-8), pages 151-161, Gaithersburg, MD, 2001.