Applying Multiple Characteristics and Techniques to Obtain High Levels of Performance in Information Retrieval at NTCIR-4

Masaki Murata
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
murata@nict.go.jp

Qing Ma
Ryukoku University
Otsu, Shiga, 520-2194, Japan.
qma@math.ryukoku.ac.jp

Hitoshi Isahara
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan
isahara@nict.go.jp

Abstract

Our information retrieval system takes advantage of numerous characteristics of the information and applies numerous sophisticated techniques. Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be effective, have been applied in the system. Characteristics of newspapers such as locational information were applied. We present our application of Fujita's method, where longer terms are used in retrieval by the system but de-emphasized relative to the emphasis on the shortest terms; this allows us to use both compound and single-word terms. The statistical test used in expanding queries through an automatic feedback process is described. The method gives us terms which have been statistically confirmed to be related to the top-ranked documents that were obtained in the first retrieval. We also used a numerical term QIDF, which is an IDF term for queries. It has a function to decrease the scores for stop words that occur in many queries. It can be very useful for foreign languages for which we cannot examine stop words. We participated in three tasks (Korean, Japanese, and English) of monolingual information retrieval at NTCIR 4. We obtained relatively higher precisions in all the tasks in which we participated. In particular, we obtained the best precision in Korean description-based monolingual information retrieval.

Keywords: *Monolingual IR, Locational information, De-emphasis of longer terms, Statistical test, QIDF*

1 Introduction

Our information retrieval system has taken advantage of numerous characteristics of the information and applied numerous sophisticated techniques. Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be very effective, have been applied in the system. We used such characteristics of newspapers as locational information. This method is very effective in retrieval from collections of newspaper articles, such as the document set for NTCIR 4. We applied Fujita's method, where longer terms are used in retrieval by the system but are assigned lower weights than the shortest terms; this allows us to use compound terms as well as single-word terms. We also used a statistical test in expanding queries through an automatic feedback process. This method gives us terms which have been statistically confirmed to be related to the top-ranked documents that were obtained in the first retrieval. We also used a numerical term QIDF, which is an IDF term for queries. It has a function to decrease the scores for stop words that occurs in many queries. We applied the system to the three tasks of monolingual information retrieval at NTCIR 4, referred to as JJ, KK, and EE.[1] Our system obtained relatively higher precisions in all the tasks in which we participated. In particular, we obtained the best precision in Korean description-based monolingual information retrieval.

[1] JJ means Japanese monolingual information retrieval, KK means Korean monolingual information retrieval, and EE means English monolingual information retrieval.

2 Outline of our system

Our system uses Robertson's 2-Poisson model[6], which is a probabilistic approach. In Robertson's method, each document's score is calculated by using the following equation.[2] The documents that obtain high scores are then output as the results of retrieval. $Score(d,q)$ below is the score of a document d against a query q.

$$Score(d,q) = \sum_{\substack{\text{term } t \\ \text{in } q}} \left(\frac{tf(d,t)}{tf(d,t) + k_t \frac{length(d)}{\Delta}} \times log\frac{N}{df(t)} \times \frac{tf_q(q,t)}{tf_q(q,t) + kq} \right) \quad (1)$$

where t indicates a term that appears in a query. $tf(d,t)$ is the frequency of t in a document d, $tf_q(q,t)$ is the frequency of t in a query q, $df(t)$ is the number of the documents in which t appears, and N is the total number of documents, $length(d)$ is the length of a document d, and Δ is the average length of the documents. k_t and k_q are constants which are set according to the results of experiments.

In this equation, we call $\frac{tf(d,t)}{tf(d,t) + k_t \frac{length(d)}{\Delta}}$ the TF term, (abbr. $TF(d,t)$), $log\frac{N}{df(t)}$ the IDF term, (abbr. $IDF(t)$), and $\frac{tf_q(q,t)}{tf_q(q,t)+kq}$the TF$_q$ term (abbr. $TF_q(q,t)$).

In our system, several terms are added to extend this equation, and the method for doing this is expressed by the following equation.

$$Score(d.q) = \left\{ \sum_{\substack{\text{term } t \\ \text{in } q}} (TF(d,t) \times IDF(t) \times TF_q(q,t) \times K_{location}(d,t) \times K_{detail} \times \left(log\frac{Nq}{qf(t)} \right)^{k_{Nq}} \right) + \frac{length(d)}{length(d) + \Delta} \right\} \quad (2)$$

The TF, IDF and TF$_q$ terms in this equation are identical to those in Eq. (1). The value of the term $\frac{length}{length+\Delta}$ increases with the length of the document. This term is introduced because, in a case where all of the other information is exactly the same, a longer document is more likely to include content that is relevant as a response to the query. Nq is the total number of queries and $qf(t)$ is the number of queries in which t occurs. Those terms which occur more frequently in queries are more likely to be such as *bunsho* "document" and *mono* "thing". We use $log\frac{Nq}{qf(t)}$ to decrease the scores for stop words. We refer to this numerical term as QIDF, because it is an IDF term for queries. It has a function to decrease the scores for words that occurs in many queries (i.e. stop words). It can be very useful for foreign languages for which we cannot examine stop words. $K_{location}$ and K_{detail} are extended numerical terms that are introduced to improve the precision of results. $K_{location}$ uses the location of the term within the document. If the term is in the title or at the beginning of the body of the document, it is given a higher weighting. K_{detail} uses information such as whether the term is a proper noun and/or a stop word. In the next section, we explain these extended numerical terms in detail.

[2] This equation is BM11, which corresponds to BM25 in the case where $b = 1$ [7].

3 Extended numerical terms

We use the two extended numerical terms $K_{location}$ and K_{detail} in Eq. (2). In this section, they are explained in detail.

1. Locational information ($K_{location}$)[3]

   The title or first sentence of the body of a document in a newspaper will generally indicate the subject. So, precision in information retrieval can be improved by assigning greater weight to terms from these locations. This is achieved by $K_{location}$, which is used to adjust the weight of a term according to whether or not it appears at the beginning of a document. A term in the title or at the beginning of the body of a document, is assigned a higher weight. A term elsewhere is given a lower weight. $K_{location}$ is expressed as follows:

$$K_{location}(d,t) = \begin{cases} k_{location,1} & \text{(when a term } t \text{ occurs in the title of a document } d\text{)}, \\ 1 + k_{location,2} \frac{(length(d) - 2 * P(d,t))}{length(d)} & \\ \text{(otherwise)} & \end{cases} \quad (3)$$

   $P(d,t)$ is the location of a term t in the document d. When a term appears more than once in a document, the location in which it first appears is used to set this parameter. $k_{location,1}$ and $k_{location,2}$ are constants to which values are assigned according to the results of experiments.

2. Other information (K_{detail})

   K_{detail} is a more detailed numerical term that uses different information, such as whether or not a term is a proper noun and whether or not it is a stop word such as *bunsho* "document" and *mono* "thing". If a term is a proper noun, it is assigned a high weight. If a term is a stop word,

[3] This method was developed by Murata et. al. [3].

it is assigned a low weight. K_{detail} is expressed in the following way for simplicity; the variables for the document and term, d and t, have been omitted:

$$K_{detail} = K_{descr} \times K_{proper} \times K_{num} \quad (4)$$

The terms in this equation are explained below.

- K_{descr}
  When a term is obtained from the title of a query, i.e. DESCRIPTION, K_{descr} = $k_{descr}(\geq 1)$. Otherwise, K_{descr} = 1. This is because we can assume that terms obtained from the description of the query are important.
- K_{proper}
  When a term is a proper noun, K_{proper} = $k_{proper}(\geq 1)$. Otherwise K_{proper} = 1. This is because terms that are proper nouns are important.
- K_{num}
  When a term is numeric, K_{num} = $k_{num}(\leq 1)$. Otherwise, K_{num} = 1. A term which consists solely of numerals will not contain much relevant information, and thus lacks importance for the query.

4 How terms are extracted

We are only able to use Eq. (2) in information retrieval after we have extracted terms from the query. This section describes how this is achieved. We considered several methods of term extraction as listed below.

1. Using only the shortest terms

   This is the simplest method. In this method, the query sentence is divided into short terms by using a morphological analyzer or similar tool. All of the short terms are used in the retrieval process. The method used to divide the query sentence into short terms is described in Section 5.

2. Using all term patterns

   The first method produces terms that are too short. For example, "enterprise" and "amalgamation" would be used separately while "enterprise amalgamation" would not be used. We felt that "enterprise amalgamation" should be used with the two short terms. Therefore, we decided to use both short and long terms. We call this the "all term-patterns method". For example, when "enterprise amalgamation realization"[4] was input, we used "enterprise", "amalgamation", "realization", "enterprise amalgamation", "amalgamation realization", and "enterprise amalgamation realization" as terms in information retrieval. We felt that this method would be effective because it makes use of all term patterns. We also felt, however, that having only the three terms "enterprise", "amalgamation", and "realization" derived from "... enterprise ... amalgamation ... realization ...", while six terms are derived from "enterprise amalgamation realization" would lack balance. We examined several methods of normalization in preliminary experiments, then decided to divide the weight of each term by $\sqrt{\frac{n(n+1)}{2}}$, where n is the number of successive words. For example, in the case of "enterprise amalgamation realization", $n = 3$.

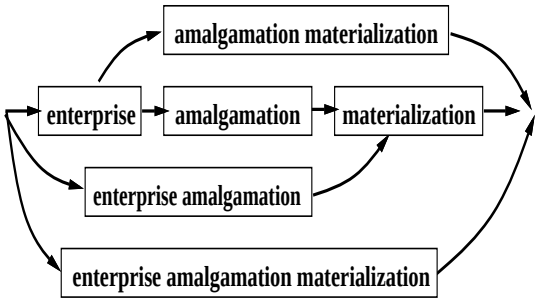

Figure 1. An example of a lattice structure

3. Using a lattice

   Although the above method effectively uses all patterns of terms, it needs to be normalized by using the ad hoc equation $\sqrt{\frac{n(n+1)}{2}}$. We thus considered a method in which all term patterns are stored in a lattice. We used the patterns in the path with the highest score on Eq. (2). The method is thus almost the same as Ozawa's [5]. The differences are in the fundamental equation used for information retrieval, and the use or non-use of a morphological analyzer.

   In the case of "enterprise amalgamation realization", for example, we obtain the lattice shown in Fig. 1. The score for each of the four paths shown in this figure is calculated by using Eq. (2), and the terms along the highest-scoring path are used. This method does not require the ad

[4] This example is not a term in English and is the English translation of a Japanese term "*kigyou* (enterprise) *gappei* (amalgamation) *seiritsu* (realization)". Its meaning is "realization of enterprise amalgamation".

hoc normalization which the method of using all term patterns requires.

4. Using de-emphasis of longer terms ("down-weighting") [1]

   Fujita proposed this method at the IREX contest [9]. It is similar to the all-term-patterns method, but the method of normalization is different. The weights of the shortest terms are kept constant while the weights of the longer terms are decreased. We decided to apply the weight ${k_{down}}^{x-1}$ to such terms, where x is the number of shortest terms and k_{down} was set according to the results of experiments.

5 Dividing the query sentence into short terms

We used morphological analyzers to divide the queries into terms. We used ChaSen [2] for JJ and HAM5.0/KMA5.0 for KK. In EE, we used the OAK system for stemming terms in sentences.

6 Automatic feedback

Automatic feedback is also used in our system. An element of automatic feedback is included in our system via the IDF term of equation (2). In applying automatic feedback, we substitute the following equation for the original IDF term.

$$IDF(t) = \{E(t) + k_{af} \times (Ratio\ C(t) - Ratio\ D(t))\} \times IDF_{orig}(t) \quad (5)$$

$$E(t) = \begin{cases} 1 & \text{(when a term t is in a query)} \\ 0 & \text{(otherwise)} \end{cases} \quad (6)$$

where $Ratio\ C(t)$ is the proportion of the top k_r documents retrieved in the first round of retrieval that include a term t. $Ratio\ D(t)$ is the proportion of all documents in which the term t appears. $IDF_{orig}(t)$ is the original IDF term. This formula is based on Rocchio's formula [8]. k_{af} and k_r are constants, which are set according to the results of experiments.

Term expansion is also applied in our system. All of the terms in the top k_r documents from the first round of retrieval are tested against a binominal distribution; those terms which satisfy the test condition are introduced as terms. That is, the terms 'Terms', as defined below, are added to the set of terms.

$$Terms = \{t | P(t) \geq k_p\} \quad (7)$$

where $P(t)$ is calculated by the following equation[5] and k_p is a constant that is set based on experimental results.

$$P(t) = \sum_{r=0}^{k} C(n,r) p(u)^r (1 - p(u))^{n-r} \quad (8)$$

where $C(x, y)$ is the number of combinations when we select y items from x items, n is equal to k_r, k is the number of times the term t occurs in the top k_r documents, and $p(t)$ is calculated by

$$p(t) = \frac{\text{freq}(t)}{N} \quad (9)$$

where freq(t) is the number of documents where the term t appears and N is the number of all documents.[6]

7 Weighting of the numbers counted in the automatic feedback process

We considered that terms that occur in higher-ranked documents and are retrieved on the first retrieval are more important than those in documents of lower rank and those retrieved later on. Thus, when counting the frequency with which a term t occurs in a document d that has a rank of $Rank(d)$, the system applies the following factor $AFW(t, d)$ to the frequency.

$$AFW(t,d) = (k_{afw} + 1) - 2 \times k_{afw} \frac{Rank(d) - 1}{k_r - 1} \quad (10)$$

where k_{afw} is a constant that is set according to the results of experiments. The frequency calculated by the above equation is used in calculating Equations (5) and (7).

8 Experiments

The name of our team is CRL.[7] The experimental results are given in Table 1. "Query" indicates the parts of the query definition that provided inputs to our system. "T" indicates the title, "D" indicates

[5] In this study, we used the summation of 0 to k, but the summation of 0 to $k-1$ could also be used. When the summation of 0 to k is used, an expression having a lower value for $P(t)$ is judged to be an expression that occurs in the top documents less often than the average occurrence in the top ducuments and it is eliminated. When the summation of 0 to $k-1$ is used, an expression having a higher value for $P(t)$ is judged to be an expression that occurs in the top documents more often than the average occurrence and the expressions other than such an expression are eliminated.

[6] This method of term expansion using a statisctical test was developed by Murata, Utiyama, and et. al. in NTCIR 2 [4].

[7] CRL is an abbreviation of Communications Research Laboratory, which is the previous name of our institute, National Institute of Information and Communications Technology.

Table 1. Experimental results

	Task	Query	ID	Parameters						R-precision		Ave. precision	
				dw	af	L	qidf	k_r	k_{af}	rigid	relaxed	rigid	relaxed
S1	JJ	T	1	n	y	y	n	5	0.7	0.3730	0.4764	0.3524	0.4638
S2	JJ	D	2	n	y	y	y	5	0.7	0.3829	0.4758	0.3612	0.4665
S3	JJ	TDNC	3	n	y	y	y	5	0.7	0.4025	0.5106	0.3803	0.4955
S4	JJ	T	4	n	y	n	n	5	0.7	0.3754	0.4787	0.3491	0.4604
S5	JJ	D	5	n	y	n	y	5	0.7	0.3840	0.4770	0.3518	0.4595
S6	KK	T	1	n	y	y	y	5	0.7	0.4716	0.5105	0.4797	0.5230
S7	KK	D	2	n	y	y	y	5	0.7	0.4693	0.4982	0.4685	0.5097
S8	KK	TDNC	3	n	y	y	y	5	0.7	0.5369	0.5624	0.5322	0.5700
S9	KK	T	4	n	y	n	y	5	0.7	0.4716	0.5038	0.4755	0.5164
S10	KK	D	5	n	y	n	y	5	0.7	0.4586	0.4869	0.4551	0.4962
S11	EE	T	1	n	y	n	y	5	0.7	0.2742	0.3535	0.2362	0.3107
S12	EE	D	2	n	y	n	y	5	0.7	0.3271	0.4124	0.2997	0.3842
S13	EE	TDNC	3	n	y	y	y	5	0.7	0.3640	0.4427	0.3425	0.4240
S14	EE	T	4	n	y	n	y	5	0.7	0.2779	0.3561	0.2356	0.3111
S15	EE	D	5	n	y	n	y	5	0.7	0.3158	0.4018	0.2898	0.3731

the description, "N" indicates the narrative, and "C" indicates the concept field of the query. The column "ID" indicates the system identifiers in the NTCIR 4 contest.[8] "–" in "ID" indicates a system which was not submitted for the formal run of the NTCIR 4 contest. The values of k_r and k_{af} are as given in Table 1. Entries in the columns marked "dw", "af" and "L" indicate the application of the longer-term de-emphasis method, automatic feedback method, the use of QIDF and the use of locational information, respectively. Use of the given method is indicated by a "y", with non-use indicated by "n". When we do not apply de-emphasis, we extract terms according to the shortest-terms method.[9] The other parameters are set as follows: $k_{location,1} = 1.2$, $k_{location,2} = 0.1$, $k_{category} = 0.1$, $k_t = 1$, $k_q = \infty$, $k_p = 0.9$, $k_{afw} = 0.5$, k_{descr} = 1, k_{proper} = 1, and k_{num} = 1.

The following findings are indicated by the experimental results.

- The use of locational information as a characteristic of newspaper articles was often effective (compare "S1" and "S4", "S2" and "S5", "S6" and "S9", and "S7" and "S10" under "R-precision; rigid")

- Use of "TDNC" always obtained the highest precisions among our systems.

Although we did not check the effectiveness of the other methods (automatic feedback method etc.) applied in our system, they would be effective. Each method and technique may only make a small contribution to the overall effectiveness. However, using all of them makes for a better system.

In this paper, we could not show the results of more detailed experiments or the results of comparison experiments using a statistical test, because the schedule for writing is very tight and our system needs a lot of time. (Our system is very slow.) In future studies, we plan to make many kinds of experiments to confirm whether or not each of the many methods used in this system is effective. In the experiments, we will compare precisions for the case of using a method and for the case of not using the method. The results of these studies will be useful for improvement of information retrieval systems.

9 Conclusion

Multiple characteristics of information and many sophisticated techniques are applied in our information retrieval system. The techniques included Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be very effective. We used such characteristics of newspapers as locational information. We used Fujita's de-emphasis ("down-weighting") method, which provides a reasonable way of including compound terms as terms used in retrieval. We also used a statistical test in expanding the queries through automatic feedback. We also used a numerical term QIDF, which is an IDF term for queries. It has a function to decrease the scores for stop words that occur in many queries. It can be very useful for foreign languages for which we cannot examine stop words. We participated in three tasks of monolingual information retrieval (JJ, KK, and EE).

[8] We could submit up to five systems for each task of NTCIR 4.

[9] In previous work [3], we had confirmed that using all term patterns is not a good approach, while even the simple method of using only the shortest terms leads to good results.

Our system obtained relatively higher precisions in all the tasks in which we participated. In particular, we obtained the best precisions in Korean description-based monolingual information retrieval.

Acknowledgement

Our thanks go to Prof. Satoshi Sekine for developing the OAK system which we used to obtain the stems of words in English sentences. We thank Prof. Dosam Hwang for the information on the Korean morphological analyzer. We are grateful to all of the organizers of NTCIR 4, who gave us a chance to participate in the NTCIR 4 contest to improve and examine our information retrieval. We greatly appreciate the kindness of all those who helped us.

References

[1] S. Fujita. Notes on phrasal indexing JSCB evaluation experiments at IREX-IR. *Proceedings of the IREX Workshop*, pages 45–51, 1999.

[2] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, and M. Asahara. Japanese morphological analysis system ChaSen version 2.0 manual 2nd edition. 1999.

[3] M. Murata, K. Uchimoto, H. Ozaku, Q. Ma, M. Utiyama, and H. Isahara. Japanese probabilistic information retrieval using location and category information. *The Fifth International Workshop on Information Retrieval with Asian Languages*, pages 81–88, 2000.

[4] M. Murata, M. Utiyama, Q. Ma, H. Ozaku, and H. Isahara. CRL at NTCIR2. *Proceedings of the Second NTCIR Workshop Meeting on Evaluation of Chinese & Japanese Text Retrieval and Text Summarization*, pages 5–21–5–31, 2001.

[5] T. Ozawa, M. Yamamoto, H. Yamamoto, and K. Umemuru. Word detection using the similarity measurement in information retrieval. *Proc. of the 5th Conference on Applied Natural Language Processing*, pages 305–308, 1999. (in Japanese).

[6] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994.

[7] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *TREC-3*, 1994.

[8] J. J. Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Prentice Hall, Inc., 1971.

[9] S. Sekine and H. Isahara. IREX project overview. *Proceedings of the IREX Workshop*, pages 7–12, 1999.