Applying Multiple Characteristics and Techniques in the NICT Information Retrieval System in NTCIR-5

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 information and uses numerous sophisticated techniques. Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be effective, are used in the system. Characteristics of newspapers such as locational information are used. 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 that have been statistically shown 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 decreases the scores for stop words that occur in many queries. It can be very useful for foreign languages for which we cannot determine stop words. We participated in three monolingual information retrieval tasks (Korean, Japanese, and English) and two bilingual information retrieval tasks (Japanese-English and English-Japanese) in NTCIR-5. We obtained high precision in all the tasks in which we participated compared to other participants. In particular, we obtained the best precision in the Korean title-based monolingual information retrieval and the Japanese-English bilingual information retrieval.

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

1 Introduction

Our information retrieval system takes advantage of numerous characteristics of information and uses numerous sophisticated techniques. Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be very effective, have been used in the system. We used characteristics of newspapers such as locational information. Our system is very effective in retrieval from collections of newspaper articles, such as the document set for NTCIR-5. 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 that have been statistically shown 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 decreases the scores for stop words that occur in many queries. In NTCIR-5, we applied the system to the three tasks of monolingual information retrieval, JJ, KK, and EE, and to the two tasks of bilingual information retrieval[1], JE and EJ. JJ, KK, and EE stand for Japanese, Korean, and

[1] Bilingual information retrieval is also called *cross-lingual information retrieval*.

English monolingual information retrieval. JE stands for Japanese-English bilingual information retrieval. The source language (used in queries) is Japanese and the target language (used in documents) is English. EJ stands for Japanese-English bilingual information retrieval. Our system obtained high precision compared to those of other participants in all the tasks in which we participated. In particular, we obtained the best precision in the Korean title-based monolingual information retrieval and the Japanese-English bilingual 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 the retrieval.

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

where $Score(d,q)$ is the score of a document d against a query q, t indicates a term that appears in the query, $tf(d,t)$ is the frequency of t in document d, $tf_q(q,t)$ is the frequency of t in a query q, $df(t)$ is the number of documents in which t appears, N is the total number of documents, $length(d)$ is the length of document d, Δ is the average length of the documents, and k_t and k_q are experimentally determined constants.

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

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

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

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, when 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. The total number of queries is Nq, and $qf(t)$ is the number of queries in which t occurs. Terms that occur frequently in queries are words 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 decreases the scores for words that occur in many queries (i.e., stop words). It can be very useful for foreign languages for which we cannot determine stop words. When we use QIDF, we use 1 for k_{Nq}. When we do not use QIDF, we use 0 for k_{Nq}. We introduce the extended numerical terms $K_{location}$ and K_{detail} to improve the precision of results. The location of the term within the document determines $K_{location}$. If the term is in the title or at the beginning of the body of the document, it is given a higher weight. Information such as whether the term is a proper noun and/or a stop word determines K_{detail}. In the next section, we explain these extended numerical terms in detail.

3 Extended numerical terms

We use 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. Therefore, precision in information retrieval can be improved by assigning greater weight to terms from these locations. This is achieved by using $K_{location}$, which adjusts 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. We express $K_{location}$ as follows:

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

   where $P(d,t)$ 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

[2] This equation is BM11, which corresponds to BM25 when $b = 1$ [7].

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

appears is used to set this parameter. The terms $k_{location,1}$ and $k_{location,2}$ are experimentally determined constants.

2. Other information (K_{detail})

   The more detailed numerical term, K_{detail}, 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") or *mono* ("thing"). If the term is a proper noun, it is assigned a high weight. If it is a stop word, it is assigned a low weight. For simplicity, K_{detail} is expressed in the following way; 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., a description, then 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 that 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, if "enterprise amalgamation was input, "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 for 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, for "enterprise amalgamation realization", $n = 3$.

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 equations used for information retrieval and the use or non-use of a morphological analyzer.

   For "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 hoc normalization that 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

[4] This example is the literal English translation of a Japanese term, "*kigyou gappei seiritsu*". It means "realization of enterprise amalgamation".

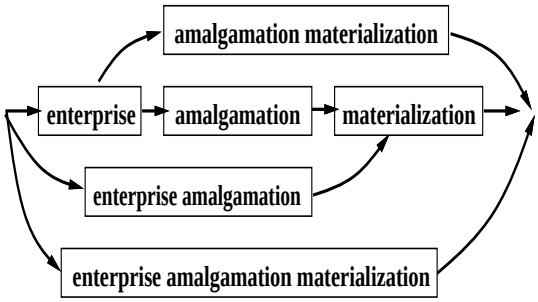

Figure 1. Example of lattice structure.

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 a weight, ${k_{down}}^{x-1}$, to such terms, where x is the number of shortest terms, and k_{down} is experimentally determined.

5 Dividing a query into short terms

We used morphological analyzers to divide queries into terms. We used ChaSen [2] for JJ and HAM5.0/KMA5.0 for KK. For 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 Eq. (2). To use 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 the term t, $Ratio\ D(t)$ is the proportion of all documents in which the term t appears, and $IDF_{orig}(t)$ is the original IDF term. This formula is based on Rocchio's formula [8]. We experimentally determine the constants k_{af} and k_r.

Term expansion is also used 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 that 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 using the following equation[5] and k_p is an experimentally determined constant.

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

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

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

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

7 Weighting the numbers counted in the automatic feedback process

We considered terms that occur in higher-ranked documents and are retrieved on the first retrieval to be 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 an experimentally determined constant. The frequency calculated by the above equation is used in calculating Eqs. (5) and (7).

[5] In this study, we used the summation from 0 to k, but the summation from 0 to $k-1$ could also be used. When the summation from 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 documents, and it is eliminated. When the summation from 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 statistical test was developed by Murata, Utiyama, et al. in NTCIR-2 [4].

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	0.3622	0.4612	0.3613	0.4615
S2	JJ	D	2	n	y	y	n	5	0.7	0.3180	0.4240	0.3162	0.4154
S3	JJ	DN	3	n	y	y	y	5	0.7	0.3810	0.4873	0.3930	0.4953
S4	JJ	TDNC	4	n	y	y	y	5	0.7	0.3828	0.4786	0.3896	0.4894
S5	JJ	D	5	n	y	y	y	5	0.7	0.3108	0.4137	0.3086	0.4056
S6	KK	T	1	n	y	y	n	5	0.7	0.4764	0.5320	0.4912	0.5441
S7	KK	D	2	n	y	y	n	5	0.7	0.4771	0.5268	0.4897	0.5449
S8	KK	DN	3	n	y	y	y	5	0.7	0.4845	0.5560	0.5089	0.5771
S9	KK	TDNC	4	n	y	y	y	5	0.7	0.4874	0.5491	0.5159	0.5799
S10	KK	D	5	n	y	y	y	5	0.7	0.4718	0.5372	0.4936	0.5571
S11	EJ	T	1	n	y	y	n	5	0.7	0.2537	0.3271	0.2458	0.3210
S12	EJ	D	2	n	y	y	n	5	0.7	0.2752	0.3625	0.2663	0.3590
S13	EJ	DN	3	n	y	y	y	5	0.7	0.3063	0.4130	0.3139	0.4076
S14	EJ	TDNC	4	n	y	y	y	5	0.7	0.3056	0.4119	0.3006	0.4000
S15	EJ	D	5	n	y	y	y	5	0.7	0.2614	0.3592	0.2644	0.3601
S16	EE	T	1	n	y	y	n	5	0.7	0.4104	0.4738	0.3983	0.4552
S17	EE	D	2	n	y	y	n	5	0.7	0.4303	0.4658	0.4265	0.4587
S18	EE	DN	3	n	y	y	y	5	0.7	0.4479	0.5157	0.4509	0.5115
S19	EE	TDNC	4	n	y	y	y	5	0.7	0.4670	0.5339	0.4592	0.5235
S20	EE	D	5	n	y	y	y	5	0.7	0.4455	0.4806	0.4314	0.4729
S21	JE	T	1	n	y	y	n	5	0.7	0.3473	0.4019	0.3386	0.3808
S22	JE	D	2	n	y	y	n	5	0.7	0.3744	0.4011	0.3621	0.3852
S23	JE	DN	3	n	y	y	y	5	0.7	0.4253	0.4790	0.4153	0.4567
S24	JE	TDNC	4	n	y	y	y	5	0.7	0.4184	0.4728	0.4114	0.4601
S25	JE	D	5	n	y	y	y	5	0.7	0.4120	0.4363	0.3967	0.4234

8 How to handle bilingual information retrieval

We used high-level, commercially available software to translate a query into the target language[7], and then used the translated query for information retrieval in the target language. We did not translate the documents.

9 Experiments

The name of our team is NICT. Our experimental results are given in Table 1. "Query" indicates the parts of the query definition that provided input to our system. "T" indicates the title, "D" the description, "N" the narrative, and "C" the concept field of the query. The "ID" column indicates the system identifiers in the NTCIR-5 contest.[8] An entry of "–" in the ID column indicates a system that was not submitted for the formal run of the NTCIR-5 contest. The values of k_r and k_{af} are as given in the table. Entries in the columns marked "dw", "af", and "L" indicate the application of the longer-term de-emphasis method, automatic feedback method, and the use of QIDF and locational information. Use of a given method is indicated by a "y", and non-use by an "n". When we do not apply de-emphasis, we extract terms according to the shortest-terms method.[9] The other parameters were 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. In this table, we show the results for JJ, KK, and EJ only because we did not get the results for EE and JE from the CLIR task organizers.

The experimental results indicate the following:

- Using QIDF was not effective in JJ, KK, and EJ (compare "S2" and "S5", "S7" and "S10", and "S12" and "S15") and effective in EE and JE (compare "S17" and "S20", and "S22" and "S25").

- Using "DN" or "TDNC" gave the highest precision of our systems.

[7] The target language is the language used in the documents.

[8] We could submit up to five systems for each task of NTCIR-5.

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

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.

10 Conclusion

Multiple characteristics of information and many sophisticated techniques are used in our information retrieval system. The techniques include Robertson's 2-Poisson model and Rocchio's formula, both of which are known to be very effective. We used characteristics of newspapers such as locational information. We used Fujita's de-emphasis (down-weighting) method, which provides a reasonable way of using compound terms in retrieval. We also used a statistical test in expanding queries through automatic feedback. We used a numerical term, QIDF, which is an IDF term for queries. It decreases the scores for stop words that occur in many queries. It can be very useful for foreign languages for which we cannot determine stop words. We participated in three monolingual information retrieval tasks (Korean, Japanese, and English) in NTCIR-5 and described the results. We obtained relatively high precision compared to other participants in all the tasks in which we participated. In particular, we obtained the best precision in the Korean title-based monolingual information retrieval and the Japanese-English bilingual information retrieval.

Acknowledgements

We thank Prof. Satoshi Sekine for developing the OAK system that we used to obtain the stems of words in English sentences. We also thank Prof. Dosam Hwang for information on the Korean morphological analyzer. We are grateful to all of the organizers of NTCIR-5, who gave us a chance to participate in the NTCIR-5 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.