On a Retrieval Support System by suggesting terms to a user

Hiroyuki SAKAI Kiyonori OHTAKE† Shigeru MASUYAMA
Department of Knowledge-based Information Engineering, Toyohashi University of Technology,
Toyohashi 441-8580, Japan
sakai@smlab.tutkie.tut.ac.jp, kohtake@slt.atr.co.jp, masuyama@tutkie.tut.ac.jp

Abstract

We propose a support method for information retrieval. This method automatically suggests terms relevant to a query, to a user, and when the user can not select adequate terms from among the suggested terms, the method suggests new terms contained in retrieved documents. We implemented an information retrieval system based on our support method and evaluated it by having users answer a questionnaire. From the results of evaluation experiments, we consider that this system is useful for users who have insufficient knowledge about the fields concerned.
Keywords: *information retrieval, human-machine interaction, user support*

1 Introduction

The recent rapid progress in computers and Internet technology has enabled us to access enormous amounts of information easily. Accordingly, document retrieval techniques to obtain necessary information quickly have become more and more important.

Most information retrieval systems currently use keywords inputted by users as queries. However, it is not easy for a user to retrieve the exact information he/she requires. In particular, it is difficult for the user to represent his/her information needs by a few keywords[1] .

Kitani et al.[4] considered that queries vary with respect to the amount of knowledge about concerning fields and compared the number of keywords against two cases : (1) Users have sufficient knowledge about concerned fields. (2) Users have insufficient knowledge about them. Kitani et al. reported that the number of keywords contained in queries made by users having sufficient knowledge about concerned fields is greater than the number of keywords contained in queries made by users with little knowledge about them[4]. The results showed that it is not easy for a user to retrieve the exact information he/she requires, as adequate keywords for representing his/her information needs are hard to find when the user has insufficient knowledge about concerned fields.

If the number of keywords is insufficient for informing the retrieval system of the user's information needs, one of the following two cases is conceivable.

Case 1 Documents irrelevant to the user's information needs are retrieved.

Case 2 A part of the required documents are retrieved.

To cope with this problem, one effective approach is to expand a query by adding terms relevant to the query when the keywords inputted by the user are insufficient for informing the retrieval system of his/her information needs. In Case 1, the user must execute "AND retrieval" for excluding irrelevant documents, and in Case 2, he/she must execute "OR retrieval" by adding new keywords. There are a number of related studies on the extraction of terms to expand a query, see e.g.,[2, 3, 5, 6].

We propose a user support method for information retrieval that suggests to users terms relevant to queries. Our method has the following three features.

- The system automatically suggests to users terms relevant to queries, which are useful for excluding retrieved documents irrelevant to the user's information needs.

The system extracts terms contained in documents that are assigned high ranks among the retrieval results. Accordingly, our method can be applied to search engines with a function for ranking retrieved documents.

- The user selects adequate terms relevant to his/her information needs from the suggested terms and the system performs retrieval by using the query expanded by adding the selected terms.

Documents containing many terms that are selected by the user are assigned high ranks by the system.

- Even if the user can not select adequate terms from the suggested terms, our system suggests

†Currently with ATR Spoken Language Translation Research Laboratories.

[1] It is said that the average number of keywords inputted by a user to Excite (http://www.excite.com), one of the more popular retrieval sites on WWW, is 2.35 [1]

new terms contained in retrieved documents having no terms that the user does not select.

We participate in the Task J-J of NTCIR-2 for evaluation and introduce our method and show results of this task in this paper. Moreover, we consider that it is inappropriate to evaluate this retrieval support system by precision or recall, and we evaluate this system by having users answer a questionnaire.

2 Method of information retrieval support

2.1 Outline of retrieval process

The outline of retrieval process by our retrieval support method is as follows.

Step 1 A user inputs a query and the system retrieves across given documents by using the query. If the system retrieves adequate documents, the process ends. Otherwise, go to Step 2.

Step 2 The system suggests to the user terms extracted from the documents assigned high ranks among the retrieval results.

Step 3 The user selects adequate terms relevant to his/her information needs from the suggested terms.

Step 4 The system expands the query by adding the selected terms, and performs retrieval by using the expanded query.

Step 5 Return to Step 2.

2.2 Method of terms extraction

Our method of terms extraction is based on the following two hypotheses.

Hypothesis 1 Terms contained many times in documents relevant to the user's information needs are relevant to the query.

Hypothesis 2 Useful terms for excluding documents irrelevant to the user's information needs from retrieved documents are dispersed in the documents set relevant to his/her information needs.

We consider that even if a term is contained in documents relevant to the user's information needs, if the term is not dispersed in the documents set relevant to his/her information needs, the term is not useful for excluding documents irrelevant to his/her information needs from retrieved documents. This is because, even if a term not dispersed in the documents set is important with respect to a document in the documents set, the term may be irrelevant to a query which retrieves the documents set. Our method of extracting terms is as follows:

Step 1 The system retrieves documents by using a query inputted by a user.

Step 2 The system extracts terms from a set S of documents assigned high ranks among the retrieval results. Here, only KATAKANA terms, where all characters used are KATAKANA, compound terms, place names, and organization names are treated as terms.

Step 3 The weight value of term w contained in document s is calculated by the following expression:

$$\begin{aligned} W(w,s) &= tf(w,s) \times \log(|S|/df(w)) \\ &\quad \times \log(dt(w)/tf(w,s)) \times \log(|S|-n) \end{aligned}$$

$tf(w,s)$**:** frequency of term w contained in document s,

S**:** the set of documents assigned high ranks among the retrieval results,

$df(w)$**:** frequency of documents containing term w in set S,

$dt(w)$**:** frequency of term w contained in set S,

n**:** rank of document s,

This expression modifies the $tf \cdot idf$ method to increase the weight values of the terms appearing many times in the documents assigned high ranks among the retrieval results and dispersed in the documents set.

Step 4 The weight value of term w is $\max_{s \in S} W(w,s)$.

Step 5 The system compares the frequency of KATAKANA terms with that of compound terms in the retrieved document set.

Step 5.1 When the frequency of KATAKANA terms is greater than that of compound terms in the retrieved document set, the weight value of each KATAKANA term is multiplied by a value calculated using the following expression:

$$\frac{frequency\ of\ KATAKANA\ terms}{frequency\ of\ compound\ terms}$$

Step 5.2 Otherwise, the weight value of each compound term is multiplied by a value calculated using the following expression:

$$\frac{frequency\ of\ compound\ terms}{frequency\ of\ KATAKANA\ terms}$$

Step 6 The system suggests to the user the terms of weight values associated with them in decreasing order from the largest.

2.3 Query expansion technique

A query inputted by a user is expanded by adding terms selected by the user from suggested terms. The expanded query is as follows:

$$Q \wedge (W_1 \vee W_2 \vee W_3 \vee \ldots \vee W_n)$$

Q: query inputted by a user.

$W1, W2, \ldots, Wn$: terms selected by a user from terms suggested by the system.

The expanded query can retrieve documents containing at least a term selected by the user in the documents retrieved by using the query inputted by the user. Among the resulting documents of retrieval by using the expanded query, documents containing many terms selected by the user have high ranks assigned by the system when a ranking process is applied to the documents. Such a process is described in detail in the next subsection. If the user could select many terms relevant to his/her information needs, at a result of retrieval by using the expanded query, documents containing many terms relevant to his/her information needs are assigned high ranks by the system.

2.4 The ranking process of retrieved documents

The ranking process of the retrieved documents of our system is done by calculating the similarity of a document and the query. We adopt the inner product of a document vector and a query vector for the calculation. The document vector and the query vector are made of elements that are weight values of the terms defined below.
The query vector: the weight value of a term contained in the query is 1; otherwise 0.
The document vector: the weight value of a term contained in a document is calculated by the following expression:

$$W(w,s) = tf(w,s) \times \log(|S|/df(w))$$

$tf(w,s)$: frequency of term w contained in document s,

S: the set of retrieved documents,

$df(w)$: frequency of documents containing term w in set S,

2.5 Countermeasure when a user can not select adequate terms

If a user cannot select adequate terms relevant to his/her information needs from suggested terms, the system automatically suggests new terms. The new terms are extracted from retrieved documents having no terms that the user does not select from the suggested terms. If terms that the user can select do not exist, adequate documents for his/her information needs may not exist in the documents with high ranks among the retrieval results. Therefore, it becomes necessary to change the documents from which the system extracts terms. The system judges that terms that the user does not select are not relevant to his/her information needs. If the system extracts terms from documents that containing terms not relevant to the user's information needs, the extracted terms may be irrelevant to his/her information needs. Instead, this system adopts documents having no terms that the user does not select from the suggested terms as documents from which it will extract new terms. Applying this method prevents the system from suggesting terms irrelevant to the user's information needs. The query is as follows, which enables the system to retrieve documents having no terms that the user does not select from the suggested terms.

$$Q \wedge not\ (\ T_1 \vee T_2 \vee T_3 \vee \ldots \vee T_m\ )$$

Q: query inputted by the user

$T1, T2, \ldots, Tm$: terms not selected by the user among those suggested by the system.

The system extracts new terms from documents retrieved by using this query.

3 Implementation of the system

We implemented an information retrieval system based on our user support method. This system is implemented on Linux using JAVA. Our method can be applied to search engines having a function that ranks retrieved documents. We use Namazu[2], a search engine distributed as a free software application. The system performs retrieval by employing Namazu and ranks retrieved documents by using the method shown before. The system extracts terms from the top 100 ranked documents where the ranks are assigned by the system to retrieved documents. We employ JUMAN[3] Version 3.5 as a morphological analyzer. We show example of executing the system in Figures 1 and 2.
Figure 1 shows results of retrieval by using a keyword "検索 " (retrieval) as a query.
Figure 2 shows suggested terms to a keyword "検索" (retrieval) on a left column list and shows terms selected by a user on a right column list. If the user can select terms in the suggested terms, he should push a "再検索" (re-retrieval) button. The system retrieves by using a query expanded by adding the selected terms. If the user can not select the terms, he should push a "関連語の更新" (update of relevant terms) button, and the system suggests new terms.

4 Results of Task J-J

Results of Task J-J, average precisions over topics, is as follows. Table 1 shows the Recall - Precision Averages, Table 2 shows average precisions for 5 documents, 10 documents, ..., 1000 documents retrieved, where Level 1 is a case of S- or A-judgments which are rated as "Relevant", and Level 2 is a case of S-

[2] http://openlab.ring.gr.jp/namazu/
[3] http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html

Figure 1. An example of retrieved document

Figure 2. An example of terms suggested by the system

or A- or B-judgments which are rated as “Relevant”. The rank is 80 in 90 systems among those participated in Task J-J in NTCIR, a workshop for evaluation of information retrieval system, by an evaluation using Relevance Judgments Level 1. The rank is 77 in 90 systems by an evaluation using Relevance Judgments Level 2. The rank is 11 in 12 interactive systems.

Recall	Precision(Level 1)	Precision(Level 2)
at 0.00	0.6279	0.7019
at 0.10	0.4722	0.5178
at 0.20	0.3874	0.3961
at 0.30	0.2979	0.2890
at 0.40	0.2193	0.2066
at 0.50	0.1792	0.1709
at 0.60	0.1345	0.1213
at 0.70	0.0748	0.0773
at 0.80	0.0527	0.0424
at 0.90	0.0101	0.0015
at 1.00	0.0042	0.0000

Table 1. Interpolated Recall - Precision Averages

5 Discussion

The reason why the results of Task J-J were not satisfactory is that selected terms from suggested terms by a user is not relevant to topics given as his/her information needs because he is unfamiliar with contents of the topics. The user must judge if the suggested terms are relevant to the topic or not. Thus, if the user is unfamiliar with contents of the topic, he may not be able to judge if the suggested terms are relevant to the topic

X documents	Precision(Level 1)	Precision(Level 2)
At 5 docs	0.4286	0.5265
At 10 docs	0.4408	0.5449
At 15 docs	0.4150	0.5170
At 20 docs	0.3704	0.4704
At 30 docs	0.3272	0.4272
At 100 docs	0.1667	0.2263
At 200 docs	0.0984	0.1393
At 500 docs	0.0437	0.0332
At 1000 docs	0.0229	0.0332

Table 2. Average precisions at 5, 10, . . . , 1000 documents retrieved

or not and it is not easy for him to select terms relevant to the topic from the suggested terms.

We consider that it is inappropriate to evaluate this retrieval support system by precision or recall. The reason why it is inappropriate to evaluate this system by precision or recall is that this system aims to retrieve adequate documents for a user by interacting several times. It is not necessary for the user to retrieve the adequate documents by using an initial query inputted by the user.

Thus we perform original experiments for evaluation of this system. We illustrate the experiments for evaluation in the next section.

6 Experiments for evaluation

6.1 The method of the experiments

We give users topics and evaluate this system by having the users answer a questionnaire after retrieving documents relevant to the topics. We also hope that the time consumed for the retrieval is shortened if

this system is in fact useful. Therefore, we compare the time consumed for retrieval by using the function of suggesting terms relevant to the topics with the time consumed for retrieval by not using this function. The experiments for the evaluation are as follows.

- The subjects are given the topics and they perform retrieval using this system.

Half of the subjects are requested to retrieve documents relevant to the topics by using the function of suggesting terms, and the other half of the subjects are requested to retrieve documents relevant to the topics by not using the function.

- The subjects select the predetermined number of documents relevant to the topics.
- If the subjects can select the predetermined number of documents relevant to the topics, the experiments end.

We perform the evaluation by having the subjects answer a questionnaire after the retrieval and the time consumed for retrieval. Four subjects participated in these experiments for evaluation. We gave each subject six topics of NTCIR Test-collection-1. The subjects selected $7 \sim 10$ documents relevant to each topic by performing retrieval in NTCIR Test- collection-1.

6.2 The results of the experiments

We distributed the questionnaire to the subjects. The subjects evaluated the system by choosing one of four items, "1. This is very useful." "2. This is useful." "3. This is of little use." "4. This is of no use." As a result, all of the subjects selected "2. This is useful." We compared the average time taken when the subjects could select documents relevant to the topics by using the function of suggesting terms with the average time taken by not using this function. Figure 3 shows the result.

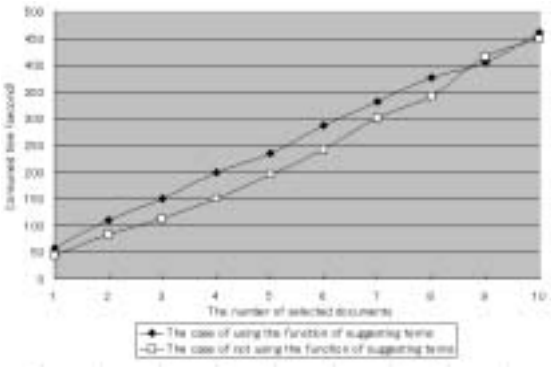

Figure 3. The time when the subjects can select relevant documents

7 Discussion on the experiments

We conclude that the time consumed for retrieval by using the function of suggesting terms is not different from the time consumed for retrieval by not using the function. We gave the subjects topics of NTCIR Test-collection-1 as their information needs. The information needs were therefore clearly stated, and it was easy for the subjects to represent queries. Even if a subject performed retrieval by not using the function of suggesting terms, he/she could end the task quickly if he/she could represent the query by using terms contained in the topics. Therefore, the system may not affect the time consumed for retrieval. However, each subject answered that this system is useful. The reason they gave is that, for example, even if a user performs retrieval by using a keyword that is inadequate, he/she can exclude retrieved documents irrelevant to his/her information needs by selecting suggested terms. We consider that this system is useful for users by this evaluation.

8 Conclusion

We proposed a user support method for information retrieval that suggests terms relevant to a query and implemented an information retrieval system based on our user support method. We consider that it is inappropriate to evaluate this retrieval support system by precision or recall. Therefore, we also evaluate this system by having subjects answer a questionnaire. From the results of the questionnaire, we consider that this system is useful for users.

References

[1] M. B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.

[2] T. Kambayashi, S. Shimizu, S. ya Sato, and P. Francis. Keyword extraction for world-wide information discovery. In *IPSJ SIG Notes 97-NL-118*, pages 79–84, 1997.

[3] H. Kawano and T. Hasegawa. Data mining technology for www resource retrieval. In *IPSJ SIG Notes 96-DBS-108*, pages 33–40, 1996.

[4] T. Kitani, T. Takaki, M. Kihara, and M. Sekine. Information retrieval using a full-text and extracted keywords. In *IPSJ SIG Notes 96-NL-115*, pages 129–134, 1996.

[5] Y. Miyata, T. Furuhashi, and Y. Uchikawa. Query expansion for information retrieval support system using fuzzy abductive inference. *T.IEE Japan*, 119-C(5):632–637, 2000.

[6] W. Sunayama, Y. Ohsawa, and M. Yachida. A search interface with supplying search keywords by using structure of user interest. *Journal of Artificial Intelligence*, 15(6):1117–1124, 2000.