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
† ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan

Email: 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, which are 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 fill out 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 the information needs by a few keywords. (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].)

Kitani et al. [2] considered that queries vary with respect to the amount of knowledge about concerned fields and compared the number of keywords used in a query 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 [2]. The results showed that it is not easy for a user to retrieve the exact information he/she requires, as adequate key-words for representing 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 information needs, one of the following two cases is conceivable.

[Case1] Documents irrelevant to information needs are retrieved.

[Case2] 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 information needs. In the former case, the user must execute “AND retrieval” for excluding irrelevant documents, and in the latter case, the user 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., [3], [4], [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.

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 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 new terms contained in retrieved documents without terms that the user does not select.

We introduce our method, and evaluate this system by having users answer a questionnaire.

2 Method of information retrieval support

2.1 Outline of retrieval process

The outline of the 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 term extraction

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

[Hypothesis 1] Terms contained many times in documents relevant to 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 the user's 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 the information needs, the term is not useful for excluding documents irrelevant to the user's 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, the term is irrelevant to the query.

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:

$$W(w,s) = tf(w,s) \times \log(|S| / df(w)) \times \log(dt(w) / tf(w,s)) \times \log(|S| - n) \quad (1)$$

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

df(w): frequency of documents containing term *w* in set *S* of documents assigned high ranks among the retrieval results,

dt(w): frequency of term *w* contained in set *S* of documents assigned high ranks among the retrieval results,

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 with a value calculated by the following expression:

$$\frac{\textit{frequency of KATAKANA terms}}{\textit{frequency of compound terms}} \quad (2)$$

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

$$\frac{\textit{frequency of compound terms}}{\textit{frequency of KATAKANA terms}} \quad (3)$$

[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:

$$Q \wedge (W_1 \vee W_2 \vee W_3 \vee ... \vee W_n) \quad (4)$$

Q: query inputted by a user.

W_1, W_2,…,W_n: 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 the next section. If the user could select many terms relevant to his/her information needs, as a result of retrieval by using the expanded query, documents containing many terms relevant to the 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:

$$W(w,s) = tf(w,s) \times \log(|S| / df(w)) \quad (5)$$

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

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

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 without 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 information needs, the extracted terms may not be relevant. Instead, this system adopts documents without 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 is able to retrieve documents without terms that the user does not select from the suggested terms.

$$Q \wedge not\ (T_1 \vee T_2 \vee T_3 \vee \ldots \vee T_m) \quad (6)$$

Q: query inputted by the user,

T_1, T_2, …, T_m: terms not selected by the user among terms suggested by the system.

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 (http://openlab.ring.gr.jp/namazu/), 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 (http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/juman.html), version 3.5 as a morphological analyzer. We use NTCIR Test-collection-1 (Constructed from about 330000 abstracts of papers) as a set of documents. We show an example of executing the system in Figure 1.

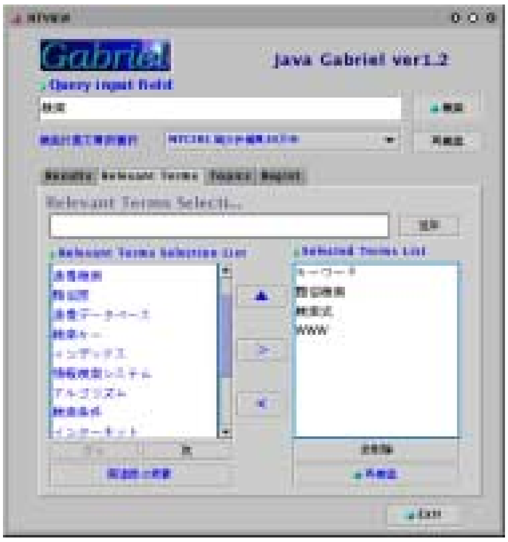

Figure 1. An example of terms suggested by the system

4 Experiments for evaluation

We consider that it is inappropriate to evaluate this retrieval support system by precision or recall. The reason why is that this system aims to retrieve adequate documents for a user by interacting several times with the user. It is therefore not necessary for the user to retrieve his/her adequate documents by only an initial query inputted by the user. We illustrate the experiments evaluating this system in the next section.

4.1 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.

[Step 1] 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.

[Step 2] The subjects select the predetermined number of documents relevant to the topics.

[Step 3] 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 (constructed from about 330000 abstracts of papers). The subjects selected 7～10 documents relevant to each topic by performing retrieval in NTCIR Test-collection-1.

4.2 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 2 shows the result.

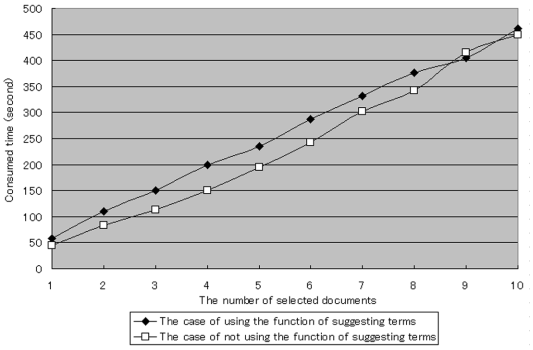

Figure 2. The average time when the subjects could select relevant documents

4.3 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 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 by selecting suggested terms. We consider that this system is useful for users by this evaluation.

5 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 distributing a questionnaire to subjects. From the results of the questionnaire, we consider that this system is useful for users.

Reference

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

2 Kitani, T., Takaki, T., Kihara, M. and Sekine, M. Information Retrieval Using a Full-text and Extracted Keywords. In *IPSJ SIG Notes 96-NL-115*, 1996, pp. 129 – 134, (in Japanese).

3 Kawano, H. and Hasegawa, T. Data mining technology for WWW resource retrieval. In *IPSJ SIG Notes 96-DBS-108*, 1996, pp. 33 – 40, (in Japanese).

4 Kambayashi, T., Shimizu, S., Sato, S. and Francis, P. Keyword Extraction for World-Wide Information Discovery. In *IPSJ SIG Notes 97-NL-118*, 1997, pp. 79 – 84, (in Japanese).

5 Sunayama, W., Ohsawa, Y. and Yachida, M. A Search Interface with Supplying Search Keywords by Using Structure of User Interest. *Journal of Artificial Intelligence*, 2000, 15(6), pp. 1117 – 1124, (in Japanese).

6 Miyata, Y., Furuhashi, T. and Uchikawa, Y. Query Expansion for Information Retrieval Support System Using Fuzzy Abductive Inference. *T.IEE Japan*, 1999, 119 - C(5), pp. 632 – 637, (in Japanese).