Notice of Violation of IEEE Publication Principles

"A Study of Relevance Feedback on Retrieved Documents in Vector-space-modeled System"
by Ruiling Zhang, Hongsheng Xu, Yanfei Li
in the Proceedings of the International Conference on Computer Science and Information Technology, August 2008, pp. 973-977

After careful and considered review of the content and authorship of this paper by a duly constituted expert committee, this paper has been found to be in violation of IEEE's Publication Principles.

This paper contains significant portions of original text from the paper cited below. The original text was copied with insufficient attribution (including appropriate references to the original author(s) and/or paper title) and without permission.

Due to the nature of this violation, reasonable effort should be made to remove all past references to this paper, and future references should be made to the following article:

"A Study of Relevance Feedback on Retrieval Documents in a Vector-space-modeled System"
by Weiping Chang
in his doctoral dissertation, submitted to National Central University of Taiwan, July 2007

A study of relevance feedback on retrieved documents in vector-space-modeled system

ZHANG RUILING
XU HONGSHENG
LI YANFEI
Academy of Information Technology
Luoyang Normal University
Luoyang Henan 471022 CHINA
xhs_ls@sina.com

Abstract

Relevance feedback is one of the techniques applied in a vector-space-modeled Information Retrieval system to enhance retrieval effectiveness. The feedback process usually has the user rate the documents retrieved as relevant or non-relevant. Most past studies apply the information of document relevance to the modification of the vector that is used to manifest the user's information interest. In this study, we have identified additional information obtained from relevance feedback that was not fully studied in the past from the rated relevant/non-relevant documents for application. The information pertains to is about the situations of term appearances in the relevant/non-relevant documents. We have developed a method together with an IR system to demonstrate the application of the information of term appearance situation. Experiments have also been conducted to study its effect. The experimental results preliminarily show that the information of the term appearance situation could be extracted and appropriately applied to enhance retrieval effectiveness.

1. Introduction

Due to the explosive growth of information on the Internet, contemporary Internet users may face the situation of information-overload. Studies on vector-space-modeled information retrieval (IR) systems have applied many techniques to limit the amount and increase the relevance of information retrieved.

Our interest here is that could the information of "term appearance situations" (abbreviated as tas later) just mentioned be extracted and effectively applied to enhance information retrieval. Considering that terms with different tas could be of different usefulness and importance in the manifestation of the user's interest and disinterest, the following two statements are worthy of study.

Terms belonging to tas 1 and 3 could have different power in the expression of the user's interest.Terms belonging to tas 2 could be of great importance in the expression of the user's disinterest.

Therefore, the primary purpose of this research is, first, to develop a method together with an IR system to demonstrate the application of the terms belonging to tas 1 and 3 in the expression of the user's interest and the application of the terms belonging to tas 2 in the expression of the user's disinterest and, second, to study the effect of the method of application on the enhancement of information retrieval.

2. Relevance Feedback

Our interest here focuses on the extraction and application of the valuable information residing in the rated relevant/non-relevant documents.

2.1. Conventional relevance feedback

For instance, Hoeber et al. developed a web search system which allowed the user to interactively re-sort the search results based on the frequencies of the selected terms within the document surrogates, as well as to add remove terms from the query, generating a new set of search results [1]. Li et al. proposed an approach towards intelligent information retrieval by providing clustered web pages and mined concepts based on results of search engines [2].

2.2. New relevance feedback

In this study, we have identified some other information "tas" as aforementioned. Based on the ideas of past studies that have achieved successful

978-0-7695-3308-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ICCSIT.2008.40

performance, our propositions for the application of tas are as follows.

(1) The terms belonging to tas 1 and 3 could be used with different weight to show their relative importance (termed as 'sensitivity' in the paper) in the vector to express the user's interest.

(2) The terms belonging to tas 2 could be used in the vector to express the user's disinterest. It could be used to provide key words in the query string with "NOT" operator.

Therefore, the extraction and the application of tas could be located as complementary. In this study, we experiment with the effect of the propositions mentioned. According to the propositions, the design of the system for the expression of the user's information requirement can be specified as follows.

(1)Maintain a positive user profile to contain terms appeared in relevant documents and a negative user profile to contain terms appeared in non-relevant documents to support the extraction of term appearance situations.

(2)Maintain a positive user profile weighted by term's frequency and term's relative importance together to express the user's interest and use this expression to provide key words for the generation of query string and basis for similarity comparison between the user's interest and the retrieved document.

(3)Maintain a negative user profile to contain terms appeared in non-relevant document only and weighted by term's frequency to express the user's disinterest and use this expression to provide key words for the generation of query string with "NOT" operator.

3. The Experimental Vector-Space-Modeled System

Based on the system design initiated in the previous section, we have developed an IR system, EIRS (Experimental Informational Retrieval System), to extract and apply the information of tas.

3.1. System framework

Figure 1 shows the system framework of EIRS. It contains five main modules denoted by the solid line rectangles. Two modules denoted by the dotted line rectangles outside the basic system are used to support the experimental process. The functional summary of each module is described as follows.

User Input and Feedback: This module supports the input of one example document originally, the input of the retrieved documents and the user's rating of relevance and irrelevance feedback.

Learning Agent: The main function of the learning agent is to learn the user's information requirement from the example documents provided by the user, the documents retrieved by EIRS and the relevance feedback given by the user for the documents retrieved.

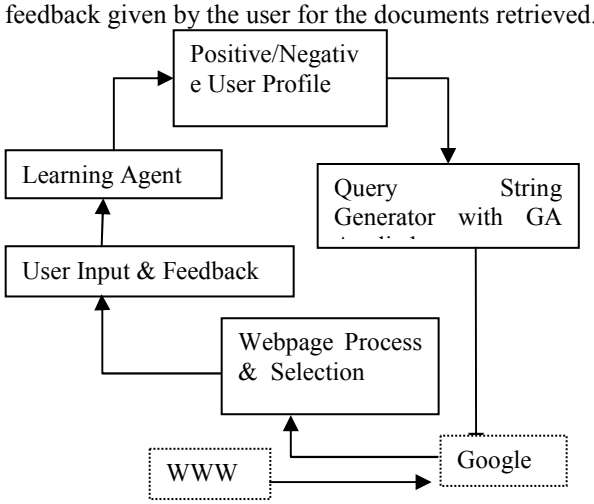

Figure 1. EIRS System framework

The positive user profile is a database table with three attributes – term, frequency and sensitivity.

Term frequency is determined by Formula and Table 1. After the user evaluates 10 webpages, sensitivity value is generated according to *tas* in relevant/non-relevant documents as shown in Table 2.

Table 1. Term frequency adjustment value

User's feedback on a retrieved document	C value in formula(1)	Positive/Negative user profile to be inferenced
Very Relevant	1.2	Positive user profile
Relevant	1	Positive user profile
In-Between	0	N/A
Non-relevant	1	Negative user profile
Very Non-relevant	1.2	Negative user profile

Table 2. Strategy of adjusting sensitivity from feedback documents

Conditions	Sensitivity value or action
The term appeared in the relevant document and not appeared in the non-relevant document	1.2

The term appeared in the relevant document and appeared in the non-relevant document	1 Remove the negative term from negative user profile

After the user inputs one example document, EIRS generates query string and retrieves 10 webpages for the user to evaluate. The user evaluates one webpage and providesnon-relevant feedback to EIRS. NFrequency, adjusted non-interested term frequency, is determined by Formula and Table 1. Before the user profile is generated, NFrequency in negative user profile has no initial value. Hence, NFrequency is determined by *FR*, term frequency in a retrieved document, and adjustment value C. After the user evaluates 10 webpages, negative user profile is generated and sorted by NFrequency in descending order. Next, EIRS generates the optimal query string and provides 10 documents for the user to evaluate.

Like most searchers setting the constant to a value for their experiments after pretest [3], C value in Table 1 and sensitivity value in Table 2 are decided after our preliminary tests.

(1) Frequency = Frequency ＋ C × *FR*

Where Frequency: initial frequency/adjusted term frequency in positive user profile

C: adjustment value based on the user's feedback for a retrieved document as Table 1.

FR: term frequency in a retrieved document

(2) NFrequency = NFrequency ＋ C × *FR*

Where NFrequency: adjusted term frequency in negative user profile.

C: adjustment value based on the user's feedback for a retrieved document as Table 1.

FR: term frequency in a retrieved document

(3) Frequency = Frequency × Sensitivity

Where Frequency: adjusted term frequency in positive user profile. Sensitivity: the sensitivity value based on the strategy of adjusting sensitivity as Table 2.

Query String Generator with GA Applied: The main function of this module is to generate a query string from the key words selected by the GA Agent according to the chromosome generated. In this module, each chromosome consists of 20 bits, 16 bits for positive keywords from positive user profile connected by AND operator and 4 bits for negative keywords from negative user profile connected by NOT operator. Each bit represents one keyword selected or not selected.

Webpage Processing and Selecting: The main function of this module is to compare the similarity between the document retrieved and the positive user profile first, then pass the similarity value to the GA in another module, and finally generate 10 most relevant documents to the user.

Tf-idf developed by Gerard Salton [4] is the most common term weighting scheme. However, in this study, we have adopted the Nick's term weighting scheme to avoid the requirement of providing a sufficient number of examples before the learning agent can begin the process of searching for new information in Salton's weighting scheme. The representation of the Dictionary was an $N \times 3$ matrix, where N is the number of keywords. The Dictionary's keywords were sorted according to their weight, which was given by the following formula (1) [5]:

$$w_i = \frac{\left(\frac{freq_i}{freq_{\max}}\right)}{\sqrt{\sum_{j=1}^{N}\left(\frac{freq_j}{freq_{\max}}\right)^2}} \quad (1)$$

Where

$freq_i$: the frequency of the keyword i in all texts in which it appear;

$freq_{\max}$: the maximum keyword frequency of all keywords in the Dictionary;

N: the number of keywords in the Dictionary.

In vector space model, cosine angle between two documents represented as vectors is the most popular approach to compare the similarity between two documents. The formula(2) is as follows:

$$sim(D,Q)=\cos(\theta)=\frac{\vec{D}\cdot\vec{Q}}{\left|\vec{D}\right|\left|\vec{Q}\right|}=\frac{D\bullet Q}{|D|\times|Q|}=\frac{\sum_{i=1}^{n} w_{Di}\times w_{Qi}}{\sqrt{\sum_{i=1}^{n} w_{Di}^2}\times\sqrt{\sum_{i=1}^{n} w_{Qi}^2}} \quad (2)$$

Where Sim (Q, Di) = similarity between Document Di and Query Q

D — Document; Q — Query

wDi — weight of term i in document D

wQi — weight of term i in query Q

If the document and the user profile are very similar, the angle should be very small.

3.2. System flow of EIRS

The system flows of EIRS are as follows.

Step 1: Input one example document in the User Input and Feedback module. The positive user profile with frequency as the term weight constructed.

Step 2: Generate query strings, return 10 webpages for users to evaluate.

Step 3: Evaluate the 10 webpages, and provide relevant/non-relevant feedback to the EIRS system in the User Input and Feedback module.

Step 4: Learn user's interest and disinterest from user's relevant and non-relevant feedback, and weight the terms.

Step 5: Create the positive user profile and negative user profile with frequency and sensitivity together as term weight revised. Positive user profile is sorted by adjusted term frequency in descending order and negative user profile is sorted by adjusted non-interested term frequency in descending order.

Step 6: Select top 16 positive terms from positive user profile by Frequency order. Select top 4 negative terms from negative user profile by NFrequency order. These 20 selected terms passed to GA to generate the optimal query string, select 10 documents most close to the user profile, and provide these 10 documents to the user for relevance feedback.

Step 7: Evaluate the 10 documents as (very) relevant or (very) non-relevant and feedback to the EIRS.

Step 8: Learn user's interest and disinterest from user's relevant and non-relevant feedback, and re-weight the terms.

Step 9: Modify positive user profile and negative user profile. Positive user profile is sorted by adjusted term frequency in descending order and negative user profile is sorted by adjusted non-interested term frequency in descending order.

Step 10: Select top 16 positive terms from positive user profile by Frequency order.Select top 4 negative terms from negative user profile by NFrequency order. These 20 selected terms passed to GA to generate the optimal query string, select 10 documents most close to the user profile, and provide these 10 documents to the user for relevance feedback.

Step 11: Evaluate the 10 documents as (very) relevant or (very) non-relevant and feedback to the EIRS.

The algorithm of modifying positive user profile and negative user profile according to the user's relevance feedback is as follows.

Input: positive/negative user profile, retrieved documents.

Output: Modified positive user profile with sensitivity set for each term, modified negative user profile.

```
1. For I = 1 to 10
2.  Case (the user's relevance rating on the retrieved document)
3.  Case 1: Very Relevant
4.      Frequency = Frequency + 1.2 × FR
5.  Case 2: Relevant
6.      Frequency = Frequency + FR
7.  Case 3: In-Between
8.    No action
9.  Case 4: Non-relevant
10.    NFrequency = NFrequency + FR
11. Case 5: Very Non-relevant
12      NFrequency = NFrequency + 1.2 × FR
13  End Case
14. Next I
15. If (the Term appeared in relevant documents but not appeared in the non-relevant documents)
16. Then
17.  Set Sensitivity = 1.2
18.  Frequency = Frequency × Sensitivity
19.   Else (the Term appeared in relevant documents and also appeared in the non-relevant documents)
20.  Set Sensitivity = 1
21.  Remove the Term from the negative user profile
22.  Endif.
```

4. Experiments and Results

To study the effect of the extraction and application of the information of *tas*, we have designed and conducted 2 experiments. We have selected twenty persons possessing a minimum of a bachelor's degree and five years of web search experience as the test.

4.1. Experimental process

The experimental processes and the formation of the user profiles are as follows.

(1) The user inputs one example document to EIRS.

(2) EIRS processes the input from step 1 and output the URLs of the retrieved documents.

(3) The user browses the retrieved documents and ranks each as "Very Relevant", " Relevant", "In-Between", "Non-relevant" or "Very Non-relevant" and inputs the retrieved documents with ranks to EIRS.

(4) EIRS processes the input from step 3 and output the URLs of the retrieved documents.

(5) The user browses the retrieved documents and ranks each as "Very Relevant", " Relevant", "In-Between", "Non-relevant" or "Very Non-relevant" and inputs the retrieved documents with ranks to EIRS.

(6) EIRS processes the input from step 5 and output the URLs of the retrieved documents.

(7) The user browses the retrieved documents and ranks each as "Very Relevant", "Relevant", "In-Between", "Non-relevant" or "Very Non-relevant" as the basis to calculate rate of correctness.

4.2. Experimental Results

There are two experiments in this study. The processes and the results of these experiments are described as follows.

Experiment 1: The purpose of this experiment is to explore the effect of amount of terms of user profile. The experimental variable is the amount of terms of user profile. Figure 2 is the experimental result. Axis X marks the value of the experimental variable; axis Y marks EIRS performance. The result shows that EIRS performance increases first as the amount of terms in user profile increases, then it begins to drop as the terms in user profile reaches a certain amount.

The result shows that EIRS performance has the best performance, 64%, when the amount of terms of user profile equals to 100. According to the result from this experiment, the quantity of 100 terms of the user profile will be set to the best system configuration.

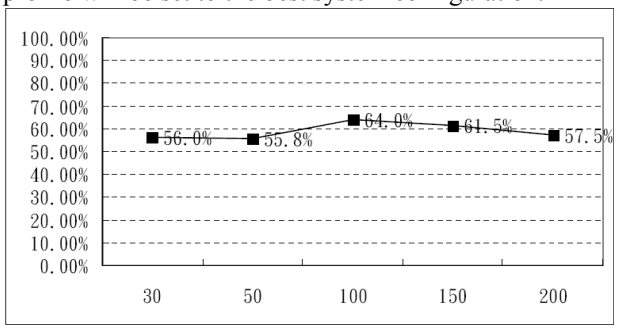

Figure2. The effect of amount of terms of user profile

Experiment 2: The purpose of this experiment is to explore the effect of the amount of key words represented by a chromosome. Figure 3 is the result of experiment 2. Axis X marks the value of the experimental variables; axis Y marks EIRS performance. The result shows that there is no obvious performance change when the amounts of key words represented by a chromosome are different. Consider the performance and system effectiveness, the quantity of 20key words are selected to be represented by a chromosome.

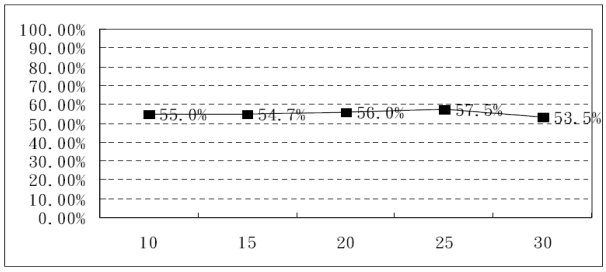

Figure3. The effect of the amount of key words represented by a chromosome

5. Conclusion

This study has identified the information of *tas* in the rated relevant/non-relevant documents. A method together with an IR system is developed to demonstrate the extraction and application of the information. It is complementary to the existing product instead of replacing. For instance, the information of *tas* 2 could be combined with Rocchio's formula to decrease the number of non-relevant documents retrieved. Furthermore, the information identified and extracted in this study also can be used in various feedback applications of IR systems.

Future work needs to be done to determine the appropriate value setting for the sensitivity. Factors to be considered could include the term appearance frequency and distribution under *tas*.

6. References

[1] Hoeber, O., and Yang, X., Interactive Web Information Retrieval Using WordBars, Proceedings of 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 875-882, Hong-Kong, December 2006.

[2] Li, F., Mehlitz, M., Feng, L., and Sheng, H., Web Pages Clustering and Concept Mining: An Approach Intelligent Information Retrieval, Technical Program of 2006 IEEE International Conferences on Cybernetics & Intelligent Systems (CIS) and Robotics, Automation & Mechatronics (RAM), Bangkok, Thailand, June 2006.

[3] Choi, J., Kim, M., and Raghavan, V., Adaptive relevance feedback method of extended Boolean model using hierarchical clustering techniques, Information Processing and management, Vol. 42, No. 2, March 2006, pp. 331-349.

[4] Salton, G., and Buckley, C., Term weighting approaches in automatic text retrieval, Information Processing and Management, Vol. 24, Nov. 1988, pp. 513-523.

[5] Nick, Z., and Themis, P., "Web Search Using a Genetic Algorithm", IEEE Internet Computing, Vol. 5, No. 2, March/April 2001, pp. 18-26.