Experiments on the TREC-9 Filtering Track

Keiichiro Hoashi† Kazunori Matsumoto† Naomi Inoue† Kazuo Hashimoto†
Takashi Hasegawa‡ Katsuhiko Shirai‡

KDD R&D Laboratories, Inc.†
2-1-15 Ohara Kamifukuoka, Saitama 356-8502, Japan

School of Science and Engineering, Waseda University‡
3-4-1 Okubo Shinjuku, Tokyo 169-8555, Japan

1 Introduction

KDD R&D Laboratories has been participating in previous TREC conferences with the cooperation of students from Waseda University. This year, KDD R&D Laboratories and Waseda University are officially participating as a joint research team.

We have focused our experiments for TREC-9 on the adaptive filtering experiments of the Filtering Track. Our goal was to evaluate the filtering method using a non-relevant information profile. We have also made experiments of a new feedback method to increase the accuracy of pseudo feedback. In this paper, we will describe our filtering methods, and present results of our evaluations.

2 Filtering methods

In this section, we will describe the filtering methods used in our experiments, and present some results from previous TREC experiments for background.

2.1 Profile updating using word contribution

Query expansion method using word contribution was applied to the profile updating process of our filtering system. Word contribution (WC) is a measure to express the influence of a word to query-document similarity. WC is defined by the following formula:

$$Cont(w, q, d) = Sim(q, d) - Sim(q'(w), d'(w)) \quad (1)$$

where $Cont(w, q, d)$ is the contribution of the word w in the similarity between query q and document d, $Sim(q, d)$ is the similarity between q and d, $q'(w)$ is query q excluding word w, and $d'(w)$ is document d excluding word w. In other words, the contribution of word w is the difference between the similarity of q and d, and the similarity of q and d when word w is assumed to be nonexistent in both data. Therefore, there are words which have positive contribution, and words which have negative contribution. Words with positive contribution raise similarity, and words with negative contribution lower similarity.

Analysis on WC[3] show that words with either highly positive or negative contribution are few, and that most words have contribution near zero. This means that most words do not have a significant influence on query-document similarity. As obvious from the definition of word contribution, words with highly positive contribution are words which cooccur in the query and document. Such words can be considered as informative words of document relevance to the query. On the contrary, words with highly negative contribution can be considered as words which discriminate relevant documents from other non-relevant documents contained in the data collection.

In the query expansion method based on WC, words used for QE were extracted only from relevant documents. In the profile updating method based on WC[1], information from all selected documents were used, regardless of their relevance to the profile.

First, the word contribution of all words in the selected document are calculated. From each selected document d, N words with the lowest contribution are extracted. Next, a score for each extracted word w is calculated by the following formula:

$$Score(w) = wgt \times Cont(w, p, d) \tag{2}$$

where wgt is a parameter with a negative value (since the contribution of the extracted word is also negative), and $Cont(w,p,d)$ is the WC of word w to the similarity of profile p and document d. On this procedure, the calculated score is regarded as the TF (term frequency) element of the word. Finally, all extracted words and their weights are added to the profile, unless the calculated weight of the word is negative.

A Rocchio-like algorithm[6] is applied here to add information from non-relevant documents to the profile. When the selected document d is relevant to the profile, the weight of word w is added to the element of the profile vector which expresses w. When d is non-relevant, the weight is subtracted from the element of the profile vector. Seperate parameters (wgt) are used for the calculation of $Score(w)$ described in Formula (2), depending on the relevance of d. wgt_{relR} is the parameter for words extracted from relevant documents, and wgt_{nrelR} is the parameter for words extracted from non-relevant documents.

Elements of the profile vector with negative weights are not used for similarity calculation, but all weights are accumulated for profile updating on upcoming documents. Therefore, the weights of words which appear in both relevant and non-relevant documents are restrained, thus emphasizing words which only appear in relevant documents.

2.2 Filtering method using non-relevant information profile

To improve filtering performance without sacrificing retrieval of relevant documents, it is necessary to reduce non-relevant document selection. However, the analysis on results of the experiments described in the previous section showed that this is difficult when filtering is based on only the similarity between the profile and incoming documents, as in most existing filtering systems.

In order to reduce retrieval of non-relevant documents, we have proposed the use of a profile which expresses the features of non-relevant documents[4]. By calculating the similarity between this *non-relevant information profile* and incoming documents which have passed the initial profile, and rejecting documents which have high similarity to the non-relevant information profile, it is possible to avoid selection of documents highly similar to past retrieved non-relevant documents. By rejecting such documents, improvement of filtering performance is expected.

The process flow of filtering with the non-relevant information profile is illustrated in Figure 1, where d is the selected document, p_R is the initial profile, p_N is the non-relevant information profile, and $Sim(p,d)$ is the similarity between profile p and document d.

As illustrated in Figure 1, thresholds $Thres_R$ and $Thres_N$ are set for each profile. The similarity between p_N and documents which have passed p_R is calculated, and compared to

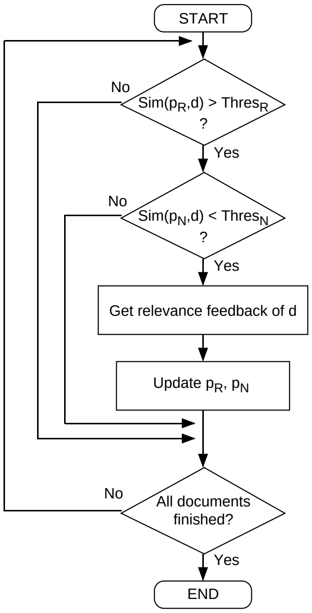
START
No
Sim(p_R,d) > $Thres_R$?
Yes
No
Sim(p_N,d) < $Thres_N$?
Yes
Get relevance feedback of d
Update p_R, p_N
No
All documents finished?
Yes
END

Figure 1: Filtering process with non-relevant information profile

$Thres_N$. If the similarity exceeds $Thres_N$, then the document is regarded as non-relevant, and, as a result, is rejected by p_N.

The method to build the non-relevant information profile is as the following:

Initial values of all elements in the non-relevant information profile are set to 0. For each selected document, N words are extracted and their weights are calculated based on WC. As in the original WC-based profile updating method, parameter wgt differs based on the relevance of the selected document. For the generation and updating of p_N, wgt_{relN} is the parameter for words extracted from relevant documents, and wgt_{nrelN} is the parameter for words extracted from non-relevant documents. To update the non-relevant information profile, the weights of words extracted from non-relevant documents are added, and weights of words extracted from relevant documents are subtracted from the regarding element of the profile vector. This is opposite from the updating of the initial profile, where the weights of words extracted from relevant documents were added to the regarding element of the profile vector, and the weights of words extracted from non-relevant documents were subtracted.

In addition to the updating of the non-relevant information profile, the initial profile p_R is also updated by the method described in Section 2.1.

2.3 Updating non-relevant profile with pseudo feedback

Results from the experiments described in the previous section show that there is a tradeoff between the strictness of $Thres_N$ and the performance of profile p_N. If $Thres_N$ is set at a low value, the number of documents blocked by p_N. This leads to the decrease of feedback information to the profile, which correlates to the performance of the filter itself. However, if $Thres_N$ is raised to increase feedback information, the number of documents rejected by p_N will also decrease, thus making the increase of feedback meaningless. To solve this problem, we propose the use of pseudo feedback[5] to increase feedback information.

Pseudo feedback is often used for QE in the text retrieval task, when the relevance of retrieved documents is uncertain. Generally, documents which are high-ranked on the initial search are assumed to be relevant. This assumation is sent back to the system, which utilizes this information to expand the query.

Our proposal is to assume documents that are blocked by p_N as non-relevant, and to send this information to the profile updating process. The documents regarded as non-relevant by pseudo feedback are handled as the same as documents which were actually regarded non-relevant from the original relevance feedback. This method allows $Thres_N$ to be strict without sacrificing feedback information.

2.4 Weighting pseudo feedback information

Our experiments with TREC-8 filtering data proved that pseudo feedback was effective. However, the number of relevant documents mistakenly rejected by the filtering system had increased by the implementation of pseudo feedback. This is caused by the inaccuracy of pseudo feedback information. Some relevant documents were mistakenly regarded as non-relevant in the pseudo feedback process, leading to mistaken feedback information to the profile.

In order to solve this problem, we propose the weighting of pseudo feedback information. Documents with high similarity to the non-relevant information profile have a higher probability to be actually non-relevant, compared to documents with low similarity to the non-relevant information profile. Our method applies a weight to the documents which pseudo feedback occurs from, based on the similarity between the document and the non-relevant information profile.

The weighting method is expressed by the following formula:

$$Value_{new}(w_i) = Value_{org}(w_i) \times \frac{sim_N - Thres_N}{1 - Thres_N} \tag{3}$$

where $Value_{org}(w_i)$ expresses the original feedback value for word w_i extracted by previously described methods, and $Value_{new}(w_i)$ expresses the feedback value weighted by our proposed method. In this formula, we multiply a weight to the originally extracted value. The weight is a normalized value of Sim_N, i.e., the similarity between the document and the non-relevant information profile. This method emphasizes pseudo feedback information extracted from documents which are assumed to have a high probability to be non-relevant to the profile, and reduce feedback information from "suspective" documents. Therefore, the improvement of the quality of pseudo feedback can be expected.

2.5 Additional System Details

Our system is based on the vector space model. The weighting calculation scheme is based on the TF*IDF based weighting formulas for the SMART system at TREC-7 [7], with minor customizations. The TF and IDF factors for our system are as the following:

- TF factor

$$\log(1 + tf) \tag{4}$$

- IDF factor

$$\log\left(\frac{M}{df}\right) \tag{5}$$

where tf is the term's frequency in the document, df is the number of documents that contain the term, and M is the total number of documents in the data collection. The document frequency data was generated from TREC CD-ROMs Vol 4 and 5. We have added 1 to the term frequency inside the logarithm of the TF factor because the tf value resulting from word contribution occasionally has values below 1, which results in a negative weight.

3 Experiments

3.1 Conditions

As previously mentioned, we have focused our experiments on the adaptive task. Furthermore, we have only made experiments with the OHSUMED topic set.

Parameters for updating the initial profile (wgt_{relR}, wgt_{nrelR}) were fixed to -800 and -200, respectively. These values were derived from preliminary experiments on the original single-filter algorithm described in Section 2.1.

Parameters for the non-relevant information profile were set as the following: $wgt_{relN} = \{-200, -400, -800\}$, $wgt_{nrelN} = \{-100, -200, -400, -800\}$. The threshold for the initial profile $Thres_R$ was set at 0.1, and the threshold for the non-relevant information profile was set at 0.25. The thresholds were set at a moderate value in order to increase the retrieval of documents, so there will be sufficient data for analysis of the filtering process.

Using the parameters listed above, we ran experiments for the normal filtering method using the non-relevant information profile (*Normal*), the pseudo feedback method (*Pseudo*), and the method with weighting applied to pseudo feedback (*Weight*).

3.2 Results

Tables 1 to 3 show the average scaled utility (T9U) of the *Normal*, *Pseudo*, and *Weight* methods for each set of wgt_{nrel} parameters. The results officially submitted to TREC are written in bold font.

Table 1: Average scaled utility (T9U), *Normal*

	wgt_{nrelN}			
w_{nrelR}	-100	-200	-400	-800
-200	0.5570	0.5584	0.5637	**0.5662**
-400	0.5569	0.5574	0.5606	**0.5631**
-800	0.5551	0.5578	0.5591	0.5612
1-filter	0.5126			

The results in Tables 1 to 3 show that the non-relevant information profile was effective in improving filtering performance. However, we could not observe significant difference between the 3 methods using the non-relevant information profile, although the pseudo feedback weighting method had the best overall scaled utility.

Moreover, it can be observed that all methods with use of the non-relevant information profile achieved higher performance when the wgt_{nrelN} parameter was set at a higher absolute value than wgt_{nrelR}. This shows that the performance of the non-relevant information profile is better when feedback information from non-relevant documents are emphasized.

Table 2: Average scaled utility (T9U), *Pseudo*

	wgt_{nrelN}			
w_{nrelR}	-100	-200	-400	-800
-200	0.5567	0.5593	0.5622	**0.5655**
-400	0.5568	0.5585	0.5597	**0.5645**
-800	0.5560	0.5601	0.5593	0.5617
1-filter	0.5126			

Table 3: Average scaled utility (T9U), *Weight*

	wgt_{nrelN}			
w_{nrelR}	-100	-200	-400	-800
-200	0.5576	0.5576	0.5652	**0.5661**
-400	0.5578	0.5588	0.5623	**0.5650**
-800	0.5560	0.5594	0.5599	0.5635
1-filter	0.5126			

4 Discussion

Results from our experiments are somewhat similar to the results observed from our TREC-8 Filtering experiments, in which the system achieved higher (utility-wise) performance as the threshold became more strict. Therefore, we were refrained from exploring new research themes such as dynamic threshold adjustment, because the threshold will automatically converge to an extreme level if the threshold adjustment method was planned to be optimized based on utility. However, dynamic threshold adjustment is an obviously effective technique for achieving high filtering performance. We believe we have proved the effectiveness of the non-relevant information profile through our experiments, so our next step will be to implement threshold adjustment to our filtering system.

References

[1] K Hoashi, K Matsumoto, N Inoue, K Hashimoto: "Experiments on the TREC-8 Filtering Track", (to be published in *The 8th Text REtrieval Conference*"), 2000.

[2] K Hoashi, K Matsumoto, N Inoue, K Hashimoto: "TREC-7 Experiments: Query Expansion Method Based on Word Contribution", The 7th Text REtrieval Conference, NIST SP 500-242, pp 433-441, 1999.

[3] K Hoashi, K Matsumoto, N Inoue, K Hashimoto: "Query Expansion Method Based on Word Contribution", Proceedings of SIGIR'99, pp 303-304, 1999.

[4] K Hoashi, K Matsumoto, N Inoue, K Hashimoto: "Document Filtering Method Using Non-Relevant Information Profile", Proceedings of ACM-SIGIR 2000, pp 176-183, 2000.

[5] S Robertson, S Walker, S Jones, M Hancock-Beaulieu, and M Gatford, "Okapi at TREC-3", Overview of the Third Text REtrieval Conference, pp 109-125, 1994.

[6] J Rocchio: "Relevance Feedback in Information Retrieval", in "The SMART Retrieval System – Experiments in Automatic Document Processing", Prentice Hall Inc., pp 313-323, 1971.

[7] A Singhal, J Choi, D Hindle, D Lewis, and F Pereira: "AT&T at TREC-7", The Seventh Text REtrieval Conference, pp 239-251, 1999.