YFilter at TREC-9

Yi Zhang    Jamie Callan
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15232, USA
{yiz,callan}@cs.cmu.edu

ABSTRACT

We built a filtering system YFILTER this year, which we used for experiments on profile updating and thresholds setting. Our focus is on using incremental Rocchio for introducing new query terms and term weighting. Although 1, 0.5, 0.25 is a widely used Rocchio ratio for query expansion based on relevance feedback, we found that the optimal setting for information filtering is corpus and profile dependent. In addition to a new Rocchio ratio, we tested a modified idf measure for term weighting (ydf) that is biased towards words with middle range term frequency.

Keywords
Information Filtering

1. INTRODUCTION

Given an initial description of a profile, a filtering system must sift through a stream of information and deliver the most relevant documents to the profile [19]. Filtering is more like a classification problem than a traditional search problem, because of the threshold, which makes it a binary decision process. Many text classification algorithms, such as SVM, Rocchio, Boosting and Naive Bayes, can be applied to filtering [1, 4, 5, 6, 13, 15...], especially for batch filtering and routing. However, as documents arrive sequentially over time, it is unrealistic to use time-consuming algorithms for online filtering. Our goal is to use a computation and storage efficient algorithm, thus making adaptive filtering possible in a normal environment, such as a PC, while still maintaining reasonably good accuracy. Our research interests caused us to adopt stricter constraints than those imposed by the Filtering task [19]. Our filtering system examines each document just once, accruing a small amount of statistical information in the process. The system does not accumulate or store batches of documents, so it requires only minimal storage and moderate computational resources.

A high performance information filtering system should be effective and efficient. Although this is a general requirement for all engineering systems, this year’s 5000 MESH profiles make it a clear requirement. A system must handle a large number of user profiles (such as 5000) and a large volume of information efficiently. Previous experiments on text classification suggest that the performance of most text classification algorithms is relatively similar. We hypothesize this will also be true for information filtering. Our starting point, therefore, is to find an algorithm that can be implemented quickly and then to refine it to perform well. We tried several methods that can be implemented incrementally for profile updating (including Rocchio, mutual information and tf.idf). Based on experiments with TREC-8 and TREC-6 filtering data, we used a refined version of the Rocchio algorithm for our final run on this year’s OHSUMED data.

2. SYSTEM DESCRIPTION

In order to achieve our goal, we built a new filtering system, called YFILTER. YFILTER is architecturally similar to InRoute [4] and consists of 3 major modules: YParser, YClipset and YLearner, which we used to support dealing with the input stream, filtering according to the profiles, and learning from relevance feedback, respectively. Different learning algorithms for profile updating can be implemented in the YLearner module.

YFILTER supports both structured Boolean and natural language descriptions of initial profiles. For natural language profiles, this system can automatically update the profile according to user relevance feedback.

YFILTER processes text first by removing useless symbols (such as punctuation, and special characters) and fields (such as the .P field), excluding the 418 highly frequent terms listed in the default INQUERY stop words list [3], and then stemming using the Porter stemmer [16].

3. ALGORITHM FOR TREC-9 FILTERING RUNS

3.1 Initial Profile Setting

In all of our experiments, the initial profile is a list of words from the Title and Description fields of the corresponding TREC topic. The weight of each word is its term frequency in the topic.
Report Documentation Page

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE
2006

2. REPORT TYPE

3. DATES COVERED
00-00-2006 to 00-00-2006

4. TITLE AND SUBTITLE
YFilter at TREC-9

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Carnegie Mellon University, School of Computer Science, Pittsburgh, PA, 15213

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR’S ACRONYM(S)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:

<table>
<thead>
<tr>
<th>a. REPORT</th>
<th>b. ABSTRACT</th>
<th>c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>unclassified</td>
<td>unclassified</td>
<td>unclassified</td>
</tr>
</tbody>
</table>

17. LIMITATION OF ABSTRACT

18. NUMBER OF PAGES
6

19a. NAME OF RESPONSIBLE PERSON

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18
for each document $d_i$
for each profile $p_k$
if $d_i$ is filtered to $p_k$ and has feedback //update profile
   update statistics;
   if feedback is relevant
      add all words in $d_i$ to $p_k$’s candidate term list
      calculate weight of each term in $p_k$’s candidate term list according to the Rocchio formula
      sort according to the weight, put the top words in the profile’s word list
else //automatically decrease threshold
   if (current delivery ratio $\varphi'$ < minimum delivery ratio $\varphi$)
      if (performance($p_k$) > 0) decrease_step = (Threshold - 0.4) * $\varphi$
      else decrease_step = (Threshold - 0.4) * ($\varphi' - \varphi$)
   threshold = max(threshold - decrease_step, 0.400)

Figure 1: An outline of profile updating algorithm

Given 2 initial relevant documents, we update the profile using the Rocchio algorithm [17], and then use the new profile from these 2 initial documents to score other documents without relevance feedback in OHSUMED87. The initial threshold is set to allow the top $\varphi(M + \delta)$ documents in the training dataset to pass. $\varphi$ is an estimate of the expected minimum delivery ratio we want to achieve. $\delta$ is used to make the initial threshold a little lower, so that we can filter more at the beginning, thus obtain more feedback for learning. We arbitrarily set $\delta = 2$ in our experiment.

3.2 Scoring Method

We used the BM25 tf.idf formula [18] for scoring. Idf is initialized based on OHSUMED87 and updated over time as documents are filtered.

3.3 Profile Updating

YFILTER has profile-specific anytime updating. That is, it updates a profile immediately whenever feedback, positive or negative, is available for that profile (Figure 1).

3.3.1 Threshold Updating

The TREC9 $T9P$ metric is defined as $R_1 / \max(MinD, (R_2 + N_2)) [19]$. $T9P$ demands a minimum number of documents (MinD) be delivered to each profile. MinD is set at 50 documents over the 4-year test period, thus approximately 1 per month. We set our delivery ratio $\varphi$ accordingly. Each negative feedback increases the threshold. Otherwise, the threshold is always decreasing according to $\varphi$ and the current profile’s performance. The magnitude of an increase to a threshold is limited by $maxstep$, which is empirically set to 0.005, so that a non-relevant document with an extremely high score will not push the threshold too high. Thus, we can avoid undue influence of an outlier. For measuring the performance of a profile, we arbitrarily used the utility F2 used in TREC-8 [9]. If a profile’s F2 utility is positive, we regard it as a good profile, and, therefore decrease its threshold comparatively faster (Figure 1). All of the parameters are set based on TREC8 and TREC6 data. The intuition is to filter more documents for good profiles, while keeping the delivery ratio for bad profiles at least meet the requirement set by the $T9P$ measure.

For the $T9U = \{2R_2 - N_2 \text{ if } (2R_2 - N_2) > MinU\}$ measure [19], if we filter by estimated probability of relevance based on the score of the current document only, the linear component of $T9U$ is equivalent to the retrieval rule:

Retrieve if $P(\text{rel} | \text{score}) > 0.33$.

Unfortunately, the sparsity of positive relevance feedback makes it hard to find the optimal threshold for most of the profiles, especially for online searching while doing information filtering. If we set the threshold to the one that yields the best utility over the
accumulated documents, while the current profile is built from the same documents, we expect that the resulting threshold is biased. We prefer the T9P measure to the T9U measure, because when there are no positive utilities, filtering no documents is usually the best strategy. Our profile updating method is T9P oriented, and our submission of a T9U oriented run is just a small change of T9P optimization to make the minimum delivery ratio $\varphi$ 50 times smaller and to decrease the threshold if the profile's precision is better than average.

3.3.2 Updating Terms and Term Weights

In all of our experiments, each time a positive relevance feedback arrives (including those in the training data), all words in that document are added to the profile’s candidate list of terms. Then the weight of each word in the candidate list is calculated according to the incremental Rocchio formula:

$$Rocchio = \alpha \cdot w_q + \beta \cdot w_{rel} - \gamma \cdot w_{non-rel}$$  \hspace{1cm} (1)

Where

$$w_q(t) = \max(\text{term frequency of word } t \text{ in original topics}, 0.5)$$

$$w_{rel}(t) = \frac{1}{\text{rel_set}(t)+1} \sum_{d \in \text{rel_set}(t)} tf_{rel_d} \cdot ydf_{rel_d}$$  \hspace{1cm} (2)

$$ydf_{rel_d} = -\log(idf_{rel_d}) (\log(1 - idf_{rel_d})) \log(1 - idf_{rel_d})$$  \hspace{1cm} (3)

$$idf_{rel_d} = \kappa \cdot \log((C_d + 0.5) / df_{rel_d}) / \log(C_d + 1.0)$$  \hspace{1cm} (4)

$$tf_{rel_d} = tf_{rel_d}(tf_{rel_d} + 0.5 \cdot (dl_d / \text{avg} \_ dl_d)$$  \hspace{1cm} (5)

The meanings of the above parameters are:

$tf_{rel_d}$: Number of times term $t$ occurs in document $d$

$C_d$: Number of documents in the profile

$dl_d$: Length of document $d$

$adv\_dl_d$: Average length of documents arrived before document $d$

$rel\_set(t)$: Relevant documents after word $t$ is added to the candidate list of the profile

$\kappa$: Parameter used to monitor the query zone

In order to learn faster, we set a bigger $\beta$ than usual in the relevance feedback formula to emphasize the importance of relevant documents, and changed $w_{rel(t)}$ to emphasize the difference between important words and noisy words.

Some researchers argue that words with middle range term frequencies are more informative than rare words and high frequency words [22], so we introduced a new parameter $ydf$ that favors those words. Ideally $ydf$ is a convex function that reaches its maximum at the optimal query expansion zone. We arbitrarily calculate $ydf$ based on $idf$ using Formula 3 & 4. Setting a different $\kappa$ can move the query zone. Figure 2 shows the effect of setting $\kappa$ to 0.75 and to 1.0. We first introduced it while testing mutual information, because mutual information favors rare words [23]. We also found it helpful for Rocchio.

**Figure 2:** Using $ydf$ to favor words with middle range term frequency: the relationship between document frequency and $ydf$

In order to avoid adding too many noisy words, especially at the early stage, the number of words added to a profile by its $t$th updating is at most $7 / \gamma \cdot \gamma$ increases as the number of relevant documents increases. We also set the maximum number of words for each profile to 60, because our experiments on query expansion for routing task shows that adding more words than that does not improve the performance. Generally speaking, for the number ($N_{kt}$) of key words in a profile ($P_k$) after the $t$th updating, we have:

$$N_{kt} \leq \max(N_{kt-1} + 7 / \lambda, 60)$$

$$\lambda = \sqrt{R_e + 1}$$  \hspace{1cm} if current document that triggers profile updating is relevant, otherwise $7 / \lambda$ was set to 0. $R_e$ is the number of relevant documents that have seen for this profile.

4. ANALYSIS OF RESULT
### 4.1 Profile Updating

In order to optimize for the measures used for TREC-9, we set the minimum delivery ratio \( \phi = \min(D/(6000 * \min(D + 50000)), \) where 50000 is the approximate number of documents in OHSUMED87. We have not performed a thorough study of the relationship between \( \phi \) and the actual delivery ratio based on our algorithm. For the T9P-oriented run, the delivery ratio is about 1/5000 on OHSU topics, and 1/4400 on MESH topics. This is not surprising, because we are expecting a higher delivery ratio for better profiles (Figure 1). So for MESH topics, which contain more relevant documents on average, the actual delivery ratio is higher than our target. But for the same topic, a higher delivery ratio usually results in lower precision.

In our experiments with TREC-6 and TREC-8, we found that a higher \( \beta \) in the Rocchio formula results in higher performance on both corpora. We guessed that this is also true for other corpora if we have enough number of relevance feedback and the positive feedback itself best represents the corresponding topic. As a result we set \((\alpha, \beta, \gamma) = (1,3.5,2)\) for the final run on TREC-9. After submitting our results, we did the same experiment on OHSU topics and the first 50 MESH terms to measure the relation between \( \beta \) and precision. The result confirmed our hypothesis (Figure 3), and we did not observe any decrease on precision as \( \beta \) increases. In fact our setting of the Rocchio ratio was a little conservative, although it is already quite different from the widely used \((1,0.5,0.25)\) setting.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Optimized for</th>
<th>Average utility/precision</th>
<th>Percent of profiles better than or equal to median</th>
<th>Average median utility</th>
</tr>
</thead>
<tbody>
<tr>
<td>MESH</td>
<td>T9P</td>
<td>0.359</td>
<td>0.50</td>
<td>0.41</td>
</tr>
<tr>
<td>MESH</td>
<td>T9P SAMPLE</td>
<td>0.363</td>
<td>0.55</td>
<td>0.40</td>
</tr>
<tr>
<td>MESH</td>
<td>T9U SAMPLE</td>
<td>26.7</td>
<td>0.53</td>
<td>34.10</td>
</tr>
<tr>
<td>OHSU</td>
<td>T9P</td>
<td>0.267</td>
<td>0.68</td>
<td>0.24</td>
</tr>
<tr>
<td>OHSU</td>
<td>T9U</td>
<td>9.3</td>
<td>0.97</td>
<td>-3.75</td>
</tr>
<tr>
<td>OHSU</td>
<td>T9P</td>
<td>0.279</td>
<td>0.83</td>
<td>0.24</td>
</tr>
<tr>
<td>OHSU</td>
<td>T9U</td>
<td>10.1</td>
<td>0.97</td>
<td>-3.75</td>
</tr>
</tbody>
</table>

Table 1: Submitted runs in TREC-9

#### 4.2 System Performance at Different Stages

![Figure 3: Precision and \( \beta \) setting](image)

Figure 3: Precision and \( \beta \) setting

![Figure 4: Filtering performances at different stages: F1 Metric](image)

Figure 4: Filtering performances at different stages: F1 Metric

![Figure 5: Filtering performances at different stages: Precision Metric](image)

Figure 5: Filtering performances at different stages: Precision Metric

Figure 4 and Figure 5 show our filtering performance using macro average precision and F1 metrics. The horizontal axis does not begin with 0 documents because at the early stage some profiles have no documents filtered to them, and thus the early stage is
not comparable with later stages. It is obvious that filtering performance is improving during the whole process in both measurements. This indicates that the filtering system is learning while filtering. Notice that the performance in Figure 4 and Figure 5 is accumulated performance, so the actual snap shot of performance at difference stages is higher than that was showed here.

4.3 Overall Performance in TREC-9

Our results on OHSU are satisfying, while our results on MESH topics are not so good (Table 1). Possible reasons are:

1. Many parameters are set from experiments on TREC-6 and TREC-8. OHSU topics are more like TREC-6 and TREC-8 topics in query length and in number of relevant documents per profile. For example, we set the maximum number of words for each profile to 60, while this should be different from query to query, based on the number of original query terms. Another example is the Rocchio ratio setting. Further experiments show that a higher $\beta$ setting, such as 9.7, has a very significant improvement on MESH topics (Figure 3). We believe for better performance, $\beta$ should be set higher.

2. 7 groups submitted OHSU results, while only 4 groups submitted their MESH results.

3. In order to maintain a certain retrieval rate, we introduced a minimum delivery ratio $\varphi$ for the threshold setting. But the actual delivery ratio is higher than $\varphi$, especially for good profiles, where the goodness is measure by the TREC-8 F2 metric. We have more good profiles in the MESH runs, so the actual delivery ratio is higher, thus a lower precision.

4.4 Defects and Explanation

Considering that there are too many non-relevant documents for MESH topics, we did not update the profile every time the system encountered a non-relevant document. While the Rocchio accumulator is inside the profile-updating module, thus the filtering system did not accumulate information for non-relevant documents. The effect is equal to $\gamma=0$ in Formula 1. In fact we want to use 2 for $\gamma$ based on our experiments on TREC-6 and TREC-8. After submitting the official result, the bug was fixed, which improved recall from 0.363 to 0.376, and improved precision from 0.267 to 0.271 for OHSU. Further experiments show that when $\beta$ is set bigger, such as 9, the change of $\gamma$ has no significant impact. One possible explanation is that non-relevant documents are heterogeneous while relevant documents are homogeneous.

5. CONCLUSIONS

We evaluated our new information filtering system YFILTER by participating in the TREC-9 Adaptive Filtering task. It takes about 10 minutes for YFilter to filter 4 years of MEDLINE dataset for 63 OHSU topics. The experimental results compared favorably with results from other filtering systems. In order to maintain a certain retrieval rate, we introduced a minimum delivery ratio $\varphi$ for threshold setting, and automatically decreased the profile threshold if its delivery ratio is below that. We found the difference between the actual delivery ratio and $\varphi$ is reasonable on TREC-6, TREC-8 and TREC-9 data, but further theoretical analysis is needed for a more justifiable threshold setting algorithm. We used incremental Rocchio with a quite different Rocchio ratio setting plus a $ydf$ measure that favors middle range term frequency words for adding new terms and term weighting. According to further tests on OHSUMED data after submission of our runs, we find that although our new ratio setting has improved the system performance significantly, it is not optimal for TREC-9. Further experiments show that the optimal Rocchio ratio is corpus and profile dependant. For profiles with more relevance judgments, a higher $\beta$ is much better, so we think a possible solution is to learn $\beta$ adaptively. As for the TREC-9 run, we found that Rocchio ratios and the $\varphi$ setting have a significant influence on system performance. We also believe that the idea of introducing $ydf$ will improve performance as well, but our function of mapping from $df$ to $ydf$ is arbitrary, thus the improvement on TREC-9 runs is not obvious. Further experiments can be focused on finding more principled or more accurate functions for $ydf$ calculation.

6. ACKNOWLEDGEMENTS

This material is based on work supported by National Science Foundation grant IIS-9873009 and Air Force Research Laboratory contract F30602-98-C-0110. Any opinions, findings, conclusions or recommendations expressed in this paper are the authors’, and do not necessarily reflect those of the sponsors.

7. REFERENCE

[22] V. Rijsbergen. 1979 Information Retrieval