Incremental Learning for Profile Training in Adaptive Document Filtering[*]

Liang Ma, Qunxiu Chen, Shaoping Ma, Min Zhang, Lianhong Cai
State Key Lab of Intelligent Tech. & Sys., CST Dept, Tsinghua University, Beijing 100084, China
Maliang00@mails.tsinghua.edu.cn

Abstract

In this paper, we describe our ideas and related experiments in TREC-11 Adaptive Filtering Track. In the track we focused much on a robust way for effective profile training. We developed an incremental learning method which selects pseudo positive documents in less bias from a few initial positive training documents. We also did some experiments with newly emerged information retrieval model, language model-based retrieval mechanism, to evaluate its performance when used in adaptive filtering task. Related experiment results show the incremental learning method can be helpful for profile training, while the new language model perform not well.

1. Introduction

In adaptive filtering, firstly we do profile training to get an initial profile, then based on this profile we do adaptive profile updating. Most of the research work now focus on the algorithms for adaptive profile updating because of its immediate effect to the performance. While even a perfect adaptive profile updating mechanism will suffer a poor result if starting updating from a biased initial profile. In fact, it is high potential to get a bias initial profile because of insufficient topic features provided by such few initial positive training documents.

A common method for profile training is like this. First use the query constructed from initial positive documents to score the training set, then get all the pseudo positive documents(used to expand initial profile vector) or setup initial profile threshold. In [1], initial profile threshold is set in a rank position of these scored documents. System in [3] select n documents (n=k*m, m is the number of initial positive documents) with the highest score value in the training set to be pseudo positive documents. Then these documents, together with initial positive documents, are used as positive documents to set initial profile vector and threshold.

It is a simple method together with some problems. By this way of one-step learning, the pseudo positive document and profile threshold are totally depend on these highly limited initial positive documents. Thus any bias in training from these initial training documents will lead to amplified bias in pseudo documents and initial profile. Also the fixed number of pseudo documents, usually set by experience in [3], is hard to be determined for various topics.

In TREC-11, we do further research work in profile training to find a better way for un-bias profile training. In the next section, the detail of profile training will be introduced. After that, experiment data and evaluation result, including our experiments on language model, will be listed. At the end of this paper there is a summary.

2. Incremental Learning in Profile Training

2.1 Feature selection for initial profile

We introduce a two–phase selection mechanism for feature selection from positive documents(in this

[*] Supported by the Chinese National Key Foundation Research & Development Plan (Grant G1998030509), Natural Science Foundation No.60223004, and National 863 High Technology Project No. 2001AA114082.

section, term 'document' are called as 'doc'). First we select key features by a step-extend morphological analysis, then get more extended features by incremental learning.

(1) Get key features for initial profile

We extract initial key terms from topic statement and 3 initial positive training documents.

For topic statement, we use a parser[6] and get terms(called TK term) step by step, something unlike the common way which simply extract all the terms once. The idea is explained as follow:

1. From title field, get all the words as key terms and add them to KeyTerm set.
2. From desc field, we only find the words which limit key terms in KeyTerm set and add them to KeyTerm set.
3. From narr field, we do the same process as step 2.

For 3 positive training docs, we statistic the terms in title and text field(after stopword removing) . The terms(called TK term) whose weight are higher than the double of the average term weight and occur in more than 1 doc are selected as key terms.

We combine the TK terms and DK terms to construct a basic profile. Here the DK terms use their statistic weight. We set the average weight of DK terms as the weight of TK terms. For terms from desc and title field, increase their weight with different weight plus(>1).

(2) Incremental learning for initial profile

We use an incremental learning mechanism for more extend features from pseudo positive docs. Different from the common ways only score training set once and select all the pseudo docs(easily cause the bias problem), in our mechanism training set be repeated scored more than once. After each scoring only a small number of pseudo documents who is highly relative to the exist positive docs are selected, and these docs are used to do limited feedback for new profile vector terms. By this step learning process we can decrease the potential bias pseudo positive docs. The detail of learning process is:

1. Define a set U for positive docs(including pseudo positive docs from learning). Here P_u is number of elements in U. Also we get initial profile by process described above.
2. Use current profile to score all the docs in training set, then sort them by their score in descend order. Set AVG_u as average score of all the docs in U and S_{min} as minimal score of docs in U.
3. Select new pseudo positive documents from scored docs and add them to U. Two rules for selecting them:

   Rule 1: If the first P_u docs are all in U, select the No.(P_u +1) doc.

   Rule 2: Else, select lower value among AVG_u and S_{min} as threshold. The docs whose score are higher than threshold are new pseudo docs.
4. Do feedback(for example, Rocchio method) to profile with new selected pseudo positive docs.
5. goto step 2 for next learning. Exit loop if :
   1. if no new pseudo positive docs can be got.
   2. if P_u >n or already kept learning for r times.

2.2 Setup Initial Threshold

With the key features and extend features we create the term vector of initial profile. The initial profile threshold should be set to value that can result in the highest value of T11F. In calculating the T11F, we count all the docs in U as positive docs.

3. Language Model in Adaptive Filtering

Besides the research work in profile training, we have the interest in how well the language model IR model perform in adaptive filtering. The Lemur [7] tool kits is chosen here to support our experiment. As a newly released IR tool kits, it provide a language model-based IR mechanism for relevance score and related feedback methods. We submitted one run simply with the default parameters based on this system. Thinking that other TREC team which also use it in adaptive filtering will deliver a detail report about it, we just run our experiment with the default parameters and make no deep analysis to it.

4. Experiment

4.1 Performance of Incremental Learning

We use Reuter training corpus as training set to do incremental learning for 100 TREC topics, and let the positive training documents for batch filtering as relative documents for our test. The average T11F and T11U score of 100 topics are calculated for performance estimation. We do two training, one by fixed learning used in [3] with different fixed number (see figure 1) and one by incremental learning(see table 1)

In figure 1, the performance of fixed learning decrease as the number of pseudo positive document increase. Though there is a higher score for small fixed pseudo documents, it is no practical use for training because of low recall.

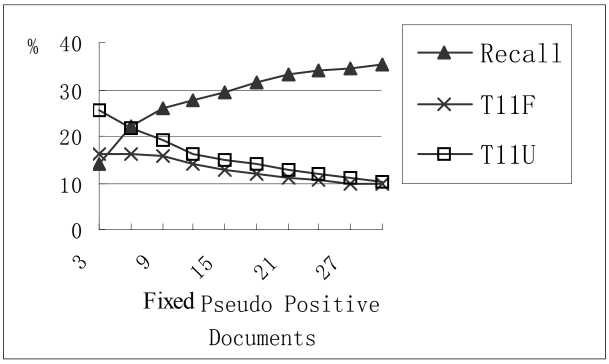

Figure 1. Fixed learning for pseudo positive documents

Table 1: Average Score of All Topics For two methods

Method	Parameters	Pseudo Documents Number	T11F	T11U
Incremental Learning	n=15; r=3	11	21.55%	27.16%
	n=20; r=4	13	20.63%	25.78%
Fixed Learning	n=15	15	12.85%	14.79%
	n=12	12	14.15%	16.31%
	n=9	9	15.74%	19.04%

Two evaluation results of incremental learning run with two set of typical parameters are listed in table 1. With both sets system reach to the similar performance(for example, pseudo documents number). In comparison with the old way, we also list another 3 results by fixed learning which has the approximate pseudo documents number with that in incremental learning. It is obviously that the new incremental learning get the high score both in T11U and T11F.

4.2 Runs Submitted and Evaluation Result

This year we submit 3 runs for adaptive filtering(algorithm for adaptive profile updating is ignored here). To compare the performance of new IR model , traditional Vector Space Model are also used for guideline, and all the runs are optimized by same criteria (T11F). Table 2 show the technology used in each runs.

Table 2: Technology Used in Each Run

Runs	IR model	Score Method	Feedback	Other Process
ThuT11af1	Vector Space Model	TF/IDF(bm25)	Improved Rocchio	Query Expansion

ThuT11af2	Vector Space Model	TF/IDF(bm25)	Improved Rocchio	
ThuT11af3	Language Model	SimpleKL	Mixture Feedback	Query Expansion

Table 3 list the evaluation result for each runs. For the 4 evaluation criteria, we calculate the average value for the first 50 and second 50 topics. Also the average of median value in all the TREC-11 runs is listed for comparison.

Table 3: Average Result of Evaluation for Each Runs

Runs	R101-R150				R151-R200			
	T11U	T11F	Set Precision	Set Recall	T11U	T11F	Set Precision	Set Recall
ThuT11af1	0.395	0.417	0.512	0.367	0.059	0.040	0.038	0.101
ThuT11af2	0.389	0.422	0.474	0.417	0.061	0.052	0.057	0.065
ThuT11af3	0.277	0.337	0.357	0.504	0.052	0.030	0.037	0.060
Avg median	0.381	0.306	0.395	0.286	0.257	0.02	0.031	0.021

From the date above we find the Language Model perform not as well as we expect. Compared to the traditional model, it does not show the predominance in adaptive filtering. We also notice that second 50 topics get better position in all Trec-11 runs than the first 50 topics do, indicating our incremental learning in profile training is more effective for second 50 topics.

5. Summary and Future Work

In TREC-2001 adaptive filtering track, we developed a general method for effective learning in profile training, and did some performance evaluation for language model. Though the language model applied to adaptive filtering wok not well as wish, the new incremental learning method demonstrate its advantage than the old ways. Knowing little on the effect that feedback method we used in incremental learning, we are going to do detailed analysis for different feedback methods in incremental learning.

Reference

[1] C. Zhai, P. Jansen, N. Roma, E. Stoica, D.A. Evans. Optimization in CLARIT TREC-8 Adaptive Filtering Trec 8. In Proceeding of eighth Text Retrieval Conference(TREC-8), NIST

[2] S. Robertson, D.A. Hull. The TREC-9 Filtering Track Final Report. In Proceeding of ninth Text Retrieval Conference(TREC-9), NIST

[3] L. Wu, X. Huang, J. Niu, Y. Guo, Y. Xia . FDU at TREC-10: Filtering, QA, Web and Video Tasks. In Proceeding of tenth Text Retrieval Conference(TREC-10), NIST

[4] A. Arampatzis Unbiased S-D Threshold Optimization, Initial Query Degradation, Decay, and Incrementality, for Adaptive Document Filtering. In Proceeding of tenth Text Retrieval Conference(TREC-10), NIST

[5] S. Robertson, I. Soboroff. The TREC 2001 Filtering Track Report. In Proceeding of tenth Text Retrieval Conference(TREC-10), NIST

[6] Dekang Lin. MiniParser. http://www.cs.ualberta.ca/~lindek/minipar.htm

[7] The Lemur Toolkit for Language Modeling and Information Retrieval. http://www-2.cs.cmu.edu/~lemur/