Proc. of The Third Intl. Conf. on Advances in Computing, Electronics and Communication - ACEC 2015
Copyright © Institute of Research Engineers and Doctors, USA .All rights reserved.
ISBN: 978-1-63248-064-4 doi: 10.15224/ 978-1-63248-064-4-10

Selecting Negative Training Documents for Better Learning

Abdulmohsen Algarni

Abstract*—In general, there are two types of feedback documents: positive feedback documents and negative feedback documents. These types share some knowledge because they retrieved using the same query. It is clear that all feedback documents contain some noise knowledge that affects the quality of the extracted features. The amount of noise is different from document to another. Therefore, the number of feedback documents affects the amount of extracting noise features. Then, using all feedback documents can increase the number of extracted noise features. However, we believe that negative feedback documents contain more noise than positive feedback documents. In this paper, we introduce a methodology to select some negative feedback documents to extract high-quality features and to reduce the amount of noises features.

***Keywords*— knowledge extraction, feature selection , relevance feedback.**

I. Introduction

Relevance feedback has been used widely in the area of information retrieval. It has been shown to be effective with different kinds of retrieval models [20], [9], [18], [21], [28]. The idea of relevance feedback is to involve the user in the retrieval process so as to improve the final result set.

The popular term-based IR models include the Rocchio algorithm [20], [6], Probabilistic models and Okapi BM25 [16], [7] (more details about Rocchio algorithm and BM25 can be found in Section 5.2), and language models, including modelbased methods and relevance models [15], [9], [28], [14], [27]. Generally, in the vector space model, terms have been extracted from feedback by using the Rocchio algorithm. Those term are used to form a new query vector by maximizing its similarity to relevant documents and minimizing its similarity to non-relevant documents [20]. In the language modelling approaches, the key elements are the probabilities of word sequences, which include both words and phrases (or sentences).They are often approximated by n-gram models [26], such as Unigram, Bigram or Trigram, for considering term dependencies.

Some kinds of retrieval models also used pseudo relevance feedback [13], [12] especially when there are no relevance judgments available. In pseudo it has assumed that a small number of top-ranked documents in the initial retrieval results are relevant and then relevance feedback is applied. However, this kind of feedback suffers from similarity (all eggs in one basket) and uncertainty [8]. close this file and download the file for “MSW A4 format”.

Many researchers believe that there are plenty negative information available and negative documents are very useful because they can help users to search for accurate information [27]. However, whether negative feedback can indeed largely improve filtering accuracy is still an open question. The existing methods of using negative feedback for IF can be grouped into two approaches. The first approach is to revise terms that appear in both positive samples and negative samples (e.g., Rocchio based models and SVM [17] based filtering models). This heuristics is obvious when people assume that terms are isolated atoms. The second approach is based on how often terms appear or do not appear in positive samples and negative samples (e.g., probabilistic models [2], and BM25 [17]). However, using only positive feedback could help isolate most documents irrelevant to user needs. For example, if the user provides feedback documents about “apple,” using positive feedback documents could help make a design appropriate for whether the user means the fruit apple or Apple the accompany. Thus, using all negative feedback documents can affect the quality of extracted knowledge. Therefore, we introduce a methodology to select useful negative feedback documents to extract high-quality features and reduce the amount of noise features in the extracted knowledge.

II. Text Representations

There will be a large number of terms extracted from text using data mining methods. The high dimensionality of the feature space leads to computational complexity and over fitting problems. The simple way to reduce the dimensionality is the filtering approach, which filters irrelevant terms based on the measures derived from the statistical information. Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. Feature selection can significantly improve the comprehensibility of the results by reducing the dimensionality.

Many types of text representations have previously been proposed. A well-known one was the bag of words that used keywords (terms) as elements in the vector of the feature space. In [10], the *tf**idf* weighting scheme was used for text representation in Rocchio classifiers. Enhanced from *tf**idf*, the global IDF and entropy weighting scheme proposed by Dumais [4] improved performance by an average of 30%. Various weighting schemes for the bag of words representation approach were given in [1]. The problem of the bag of words approach was how to select a limited

Abdulmohsen Algarni

College: College of computer Science, King Khalid University.
Country: Saudi Arbia

number of feature terms in order to increase the system's efficiency and avoid over fitting [22]. In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on [22].

The choice of a representation on what one was regarded as meaningful units of text and meaningful natural language rules for the combination of these units was reported in [22]. With respect to the representation of the content of documents, some research works have used phrases or n-grams rather than individual words [5]. In [23], a phrase-based text representation for Web document management was proposed that used rule-based Natural Language Processing (NLP) and Context Free Grammar techniques. Recently a concept-based model that analyzes terms on the sentence and document levels was introduced in [24], while the existing concept-based text mining approaches usually relied upon their employed NLP techniques.

The main problem is that the training document contains different knowledge in different documents. For example it contains long documents and short documents both documents are contain user information need. What makes documents interested to the user can be one paragraph only.

Therefore, the problem is that these traditional steps cannot remove all the noise from the data. Based on that observation, we propose a new methodology that can be used to reduce the noise in the training documents. As we know that the length of a document is different from one document to another, the question here can be whether the user is interested in all the paragraphs in the document or not. We believe that not all training documents useful to extract knowledge from. And we believe that short documents is more important than long documents. Because short documents contains less noises data.

III. Methodology

A. *Document Selection*

The most frequently used collection for experiments in information filtering is the Reuters dataset. During the past decade, several versions of Reuters corpora have been released. The most recent version of this popular data collection, the Reuters Corpus Volume 1 (RCV 1), was selected for this experiment [25]. The RCV1 dataset contains approximately 100 topics. The documents on each topic are divided into two groups: training and testing. The training group of documents are divided into two groups: positive training documents and negative training documents. There is some overlap knowledge among feedback documents. Most importantly, some overlap knowledge is noise knowledge. Based on this knowledge, the training documents can be grouped into three groups. The first group is all positive feedback documents. The second group is negative feedback training documents that contain unique knowledge. The third group is negative feedback documents that contain large amounts of noise and some overlap knowledge with positive feedback documents. The knowledge extracted from the third group of documents affects the quality of extracted knowledge from the feedback document and uses the system to make wrong decisions in the ranking algorithm. To isolate the three groups of documents, we introduce the following steps:

- Extract knowledge from all feedback documents.
- Rank the training documents using extracted knowledge and the following function:

$$Rank(d)=\sum_{t\epsilon T} w(t)\ \tau(t,d)$$

$$rank(d) = \sum_{t\in T} w(t)\tau(t,d)$$

Where $w(t) = w(t, D)$; and $\tau(t, d) = 1$ if $t \in d$; ; otherwise $\tau(t, d) = 0$.

- Select all positive feedback documents as the first group.
- Select bottom n negative feedback training documents as the second group.

$$n = \frac{|D^+|}{2}$$

- Used ***n*** negative feedback documents and positive feedback documents to extract knowledge.

The algorithm takes time to rank and select training documents *$O(mlog^m)$.*

B. *Data set*

Reuters Corpus Volume 1 (RCV1) was used to test the effectiveness of the proposed model. RCV1 corpus consists of all and only English language stories produced by Reuter's journalists between August 20, 1996, and August 19, 1997 with total 806,791 documents. The document collection is divided into training sets and test sets. TREC (2002) has developed and provided 100 topics for the filtering track aiming at building a robust filtering system. The topics are of two types: 1) A first set of 50 topics are developed by the assessors of the National Institute of Standards and Technology (NIST)(i.e., assessor topics); the relevance judgments have been made by assessor of NIST. 2) A second set of 50 topics have been constructed artificially from intersections of pairs of Reuters categories (i.e., intersection topics) [25]. For that reason we use the 50 assessor topics in this paper where the result is more reliable. Figure 1 shows the number of training documents for each topic in the positive and negative feedback training documents. RCV1 collection is marked in XML. To avoid bias in experiments, all of the meta-data information in the collection has been ignored. The documents are treated as plain text documents by preprocessing the documents. The tasks of removing stop-words according to a given stop-words list and stemming term by applying the Porter Stemming algorithm are conducted [11].

IV. Evaluation

SEEK
DIGITAL LIBRARY

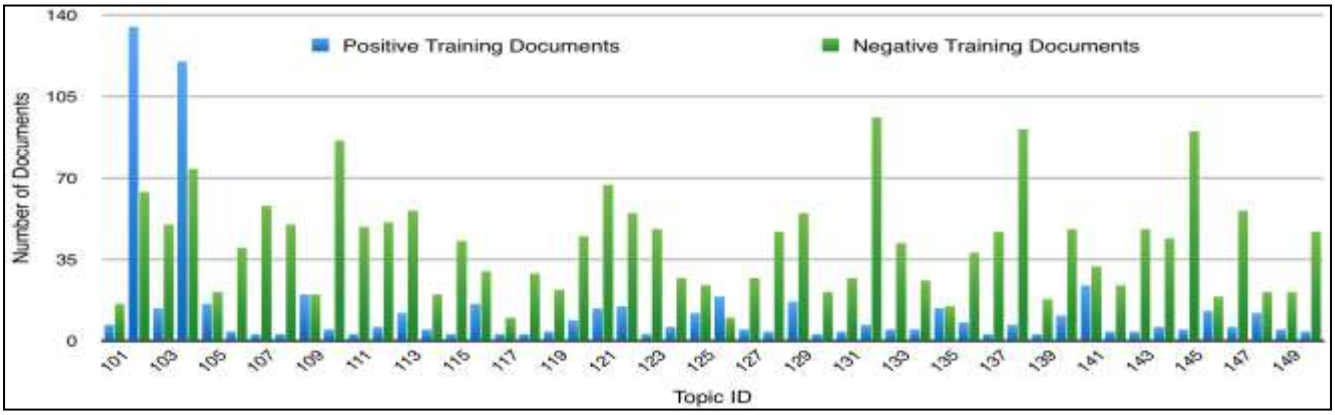

Fig. 1 : Number of positive and negative training documents in in the data set.

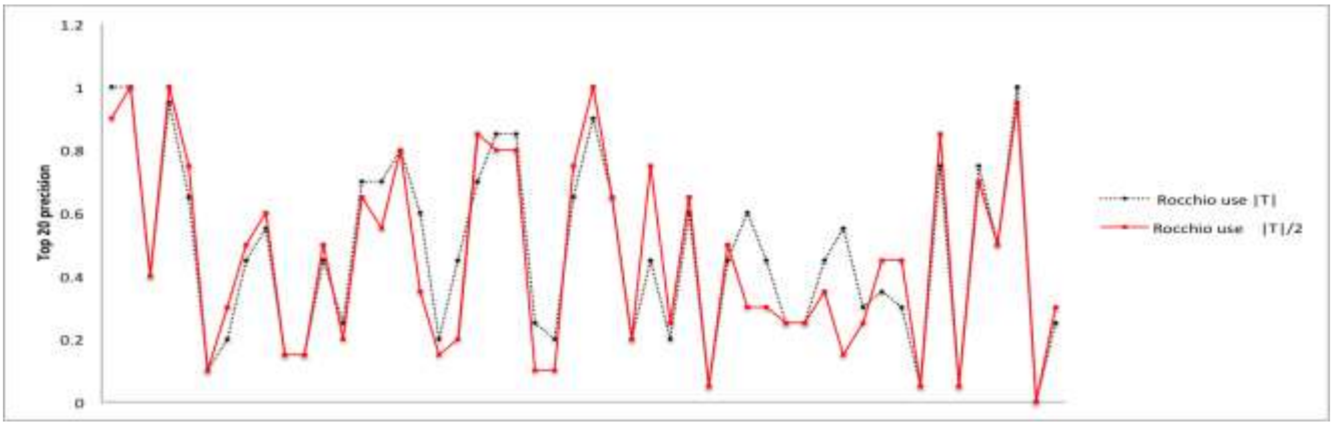

Fig. 3: Comparison of original Rocchio and the proposed method in each topic.

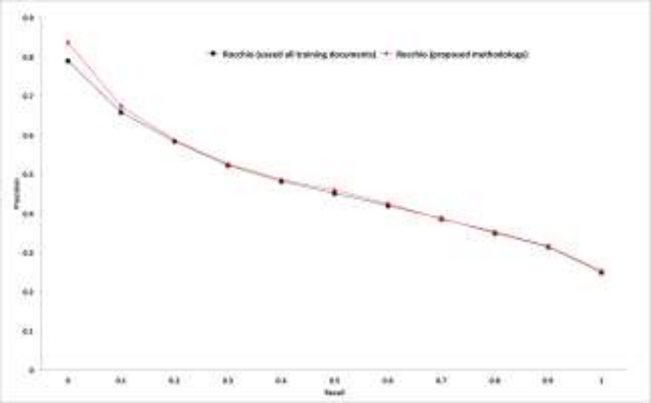

Fig. 2. Comparison of 11-pointers results of original Rocchio and the proposed Method.

In this paper, we conduct binary text classification to test the proposed approach. We use routing filtering to avoid the need for threshold tuning, which is beyond our research scope. The proposed model in this paper is assumed that not all training documents are useful to train the classifier. Therefore, we try to reduce the number of the training documents randomly and study the effect of that to the classifier result.

According to Buckley and others [3], 50 topics are adequate to make a stable, high quality experiment. This evaluation used the 50 expert-designed topics in Reuters Corpus Volume 1 (RCV1) [25]. RCV1 corpus consists of 806,791 documents produced by Reuter's journalists. The document collection is divided into training sets and test sets. These topics were developed by human assessors of the National Institute of Standards and Technology (NIST). The documents are treated as plain text documents by preprocessing the documents. The tasks of removing stop-words according to a given stop-words list and stemming term by applying the Porter Stemming algorithm are conducted

A. *Used Models and Setting*

The main models used in to conduct the result were the well-known term-based methods Rocchio. For each topic, we chose 150 terms in the positive documents, based on *tf*idf* values for all term-based models. The Rocchio algorithm [19] has been widely adopted in the areas of text categorization and information filtering. It can be used to build the profile for representing the concept of a topic which consists of a set of relevant (positive) and irrelevant (negative) documents. The Centroid $\vec{c}$ of a topic can be generated as follows:

$$\propto \frac{1}{|D^+|}\sum_{\vec{d}\in D^+}\frac{\vec{d}}{||\vec{d}||} - \beta\frac{1}{|D^-|}\sum_{\vec{d}\in D^-}\frac{\vec{d}}{||\vec{d}||}$$

Where we set $\alpha = \beta = 1.0$ in this paper.

B. *Evaluation Measures*

Precision p and recall r are suitable because the complete Classification is based on the positive class. In order to evaluate the effectiveness of the proposed method, we utilized a variety of existing methods; Mean Average Precision (MAP),breakeven points(b/p), the precision of top-20 returned documents, F-scores and recall at 11-points (IAP).

Table 1: Detailed Comparisons Of Original Rocchio And The Prposed Mehod.

	Top-20	**MAP**	**F`$_{\beta=1}$**	**b/p**	**IAP**
Rocchio	0.486	0.451	0.442	0.432	0.473
Rocchio*	0.495	0.459	0.449	0.436	0.841
%Change	**2%**	**1%**	**2%**	**2%**	**2%**

* The proposed method

These methods have been widely used to evaluate the performance of information filtering system. A statistical method, t-test, was also used to analyses the experimental results. The t-test assesses whether the means of two groups are statistically different from each other. If the p-value associated with t is significantly low (<0.05), there is evidence to reject the null hypothesis, and the difference in means across the paired observations is significant. In summary, the effectiveness is measured by five different means: the average precision of the top 20 documents, F1 measure, Mean Average Precision (MAP), the break-even point (b=p), and Interpolated Average Precision (IAP) on 11-points. The larger their values are, the better the system performs.

C. *Results And Discussion*

Table I shows the results of the comparison between the proposed models and the traditional Rocchio models assessment of all topics. It is clear that not all negative training documents are useful for extracting knowledge. Additionally, Figure 2 shows the comparison results for 11 points in all the negative feedback documents and selected negative training documents. Figure 2 proves that not all negative training documents are useful and that the proposed model produces good results for selecting useful negative feedback documents. Figure 3 illustrates the selection of some negative feedback decreases for some topics. The figure shows different affects among topics, but in the overall results, the proposed method results in significant improvement.

V. Conclusion

We assumed that not all training documents are useful in the training process. The major reason is that some documents have more noise data than useful data. Using these documents causes the extracted features to contain many noise features. Negative feedback documents contain more noise than positive feedback demounts, primarily because negative feedback can be everything but positive. Based on these results, a clustering method can be used to group training documents into three groups. As shown in the results section, reducing the number of training documents by selecting some negative training documents improves the results. In the future we are planning to improve the selection process and test the model in more than one model.

References

[1] K. Aas and L. Eikvil. Text categorisation: A survey., 1999.

[2] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.

[3] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In Research and development in information retrieval; SIGIR ’00:, pages 33–40. ACM, 2000.

[4] S. T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, & Computers, 23:229–236, 1991.

[5] G. Ifrim, G. Bakir, and G. Weikum. Fast logistic regression for text categorization with variable-length n-grams. In Knowledge discovery and data mining; KDD ’08:, pages 354–362. ACM, 2008.

[6] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In ICML ’97: Proceedings of the Fourteenth International Conference on Machine Learning, pages 143–151, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[7] K. S. Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments – part 1. Inf. Process. Manage., 36:779–808, 2000.

[8] K. Y. Lanbo Zhang Yi Zhang, Jadiel de Arma. UCSC at Relevance Feedback Track. In Text Retrieval Conference (TREC 2009), 2009.

[9] V. Lavrenko and W. B. Croft. Relevance based language models. In Research and development in information retrieval; SIGIR ’01, pages 120–127. ACM, 2001.

[10] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In Proceedings of the 18th international joint conference on Artificial intelligence, pages 587–592. Morgan Kaufmann Publishers Inc., 2003.

[11] B. Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer, January 2007.

[12] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In Information and knowledge management; CIKM ’09:, pages 255–264. ACM, 2009.

[13] C. D. Manning, P. Raghavan, and H. Schtze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[14] D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In Research and development in information retrieval; SIGIR ’07, pages 311–318. ACM, 2007.

[15] J. M. Ponte. A language modeling approach to information retrieval. Master’s thesis, 1998.

[16] C. J. V. Rijsbergen. Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 1979.

[17] S. E. Robertson and I. Soboroff. The trec 2002 filtering track report. In TREC, 2002.

[18] S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. pages 143–160, 1988.

[19] J. Rocchio. Relevance feedback in information retrieval, volume In The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, 1971.

[20] J. J. Rocchio. Relevance feedback in information retrieval. In The SMART Retrieval System, pages 313–323. Prentice Hall, 1971.

[21] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. pages 355–364, 1997.

[22] F. Sebastiani. Machine learning in automated text categorization. ACM Comput. Surv., 34(1):1–47, 2002.

[23] R. Sharma and S. Raman. Phrase-based text representation for managing the web documents. In Information Technology: Coding

and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference on, pages 165 – 169, 2003.

[24] S. Shehata, F. Karray, and M. Kamel. A concept-based model for enhancing text categorization. In Knowledge discovery and data mining; KDD ’07, pages 629–637. ACM, 2007.

[25] I. Soboroff and S. Robertson. Building a filtering test collection for trec 2002. In SIGIR, pages 243–250. ACM, 2003.

[26] F. Song and W. B. Croft. A general language model for information retrieval. In Information and knowledge management; CIKM ’99:, pages 316–321. ACM, 1999.

[27] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In Research and development in information retrieval; SIGIR ’08, pages 219–226. ACM, 2008.

[28] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In Information and knowledge management; CIKM ’01:, pages 403–410. ACM, 2001

SEEK
DIGITAL LIBRARY