Proceedings of the 23rd IASTED International Multi-Conference
ARTIFICIAL INTELLIGENCE AND APPLICATIONS
February 14-16, 2005, Innsbruck, Austria

NON-RELEVANCE FEEDBACK FOR DOCUMENT RETRIEVAL

Hiroshi Murata and Takashi Onoda
System Engineering Research Laboratory
Central Research Institute of Electric Power Industry
Tokyo, Japan
email: {murata, onoda}@criepi.denken.or.jp

Seiji Yamada
National Institute of Informatics
Tokyo, Japan
email: seiji@nii.ac.jp

ABSTRACT
This paper reports a new document retrieval method which utilizes non-relevant documents. From a large data set of documents, we need to be able to find documents that relate to the subject of interest in as few iterations of testing or checking by a user as possible. In each iteration, a comparatively small batch of documents is evaluated to establish there relevance to the subject of interest. This method is called *relevance feedback*, and it requires a set of relevant and non-relevant documents. However, the documents initially presented for checking by a user do not always include relevant documents. Accordingly, we propose a feedback method using information on non-relevant documents only. We name this method *non-relevance feedback*. Non-relevance feedback selects a set of documents which are discriminated not non-relevant area and near the discriminant function based on learning result by one-class Support Vector Machine (one-class SVM). Results from experiments show that this method is able to retrieve a relevant document from a set of non-relevant documents effectively.

KEY WORDS
relevance feedback, document retrieval, non-relevant documents, classification learning, one-class SVM

1 Introduction

With the continued progression of Internet technology, the amount of information accessible by end users is increasing explosively. In this situation, it is now possible to access a huge document database through the web. However, it is difficult for a user to retrieve relevant documents from which he/she can obtain useful information, and therefore, many studies have been done on information retrieval, particularly document retrieval [11]. Various studies on such document retrieval have been reported in TREC (Text Retrieval Conference) [10] for English documents, and IREX (Information Retrieval and Extraction Exercise) [2] and NTCIR (NII-NACSIS Test Collection for Information Retrieval System) [3] for Japanese documents.

In most frameworks for information retrieval, a vector space model is used in which a document is described with a high-dimensional vector [7]. An information retrieval system using a vector space model computes the degree of similarity between a query vector and document vectors by using the cosine of the two vectors, and then indicates to the user a list of retrieved documents.

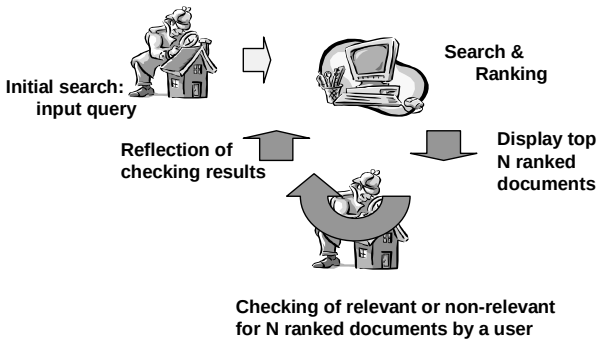

Figure 1. Relevance Feedback

In general, since a user rarely describes a query precisely in the first trial, an interactive approach to modifying the query vector on the basis of an evaluation by the user of documents in a list of retrieved documents, has been proposed. This method is called *relevance feedback* [6] and is used widely in information retrieval systems. In this method, a user directly evaluates whether a document in a list of retrieved documents is relevant or non-relevant, and the system modifies the query vector on the basis of the user's evaluation. A conventional way to modify a query vector is through a simple learning rule which reduces the difference between the query vector and the documents evaluated as relevant by a user. A conceptual diagram of relevance feedback is shown in Figure 1.

Another approach has been proposed in which classification learning treats relevant and non-relevant document vectors as positive and negative examples for a target concept [4]. Some studies have proposed that a Support Vector Machine (SVM) with excellent ability to classify examples into two classes be applied to the classification learning of relevance feedback [1].

We have proposed a relevance feedback framework with an SVM for active learning. In contrast to a conventional relevance feedback system which indicates a list of the most relevant documents to a user, our system provides a list of the most relevant documents which are difficult for the SVM to classify [5].

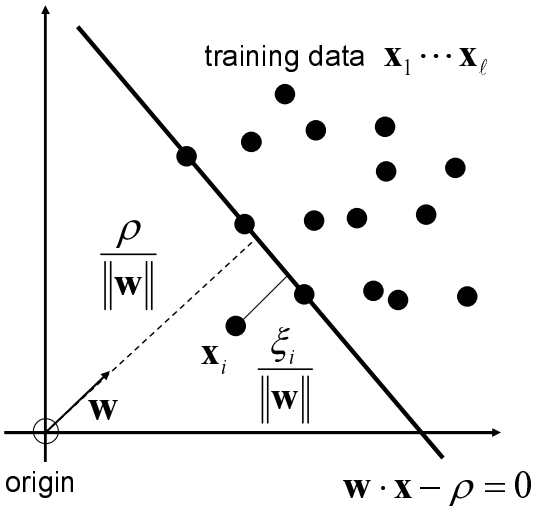

Figure 2. One-class SVM

In relevance feedback, however, the user evaluates many documents until a relevant document is obtained. If all documents are judged by the user as non-relevant, classification learning cannot be applied to relevance feedback.

Classification learning which deals with one class has been developed recently. Accordingly, we propose a framework for relevance feedback based on such classification learning, using only information on non-relevant documents. We call the feedback method which uses only the non-relevant documents, *non-relevance feedback*. We use a one-class SVM [9] in our approach to classification learning.

In the remainder of this paper, we explain the one-class SVM algorithm in the second section, and our document retrieval method with a one-class SVM for non-relevance feedback in the third section. As discussed in the fourth section, in order to evaluate the effectiveness of our approach, we carried out experiments using a TREC data set obtained from the Los Angeles Times, and we present the experimental results. Finally, we conclude our work and discuss the remaining problems in the fifth section.

2 One-class SVM

Let the training data be $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$, $\mathbf{x} \in \mathbf{R}$. A one-class SVM [9] returns a function f that takes the value +1 in a small region that captures most of the training data points, and -1 elsewhere. This strategy is to separate the data from the origin with a maximum margin. To separate the data set from the origin, we solve the following quadratic program:

$$\begin{aligned} \min \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ \text{subject to} \quad & (\mathbf{w} \cdot \mathbf{x}_i) \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

Here, $\nu \in (0, 1)$ is a parameter whose meaning is the fraction of the outliers.

A conceptual diagram of the one-class SVM is shown in Figure 2.

Since the nonzero slack variables ξ_i are penalized in the objective function, we can expect that if $\mathbf{w}$ and ρ solve this problem, then the decision function

$$f(\mathbf{x}) = \operatorname{sgn}\left((\mathbf{w} \cdot \mathbf{x}) - \rho\right) \tag{2}$$

will be positive for most examples of $\mathbf{x}_i$ contained in the training set. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on.

Using multipliers $\alpha_i, \beta_i \geq 0$, we introduce a Lagrangian

$$\begin{aligned} & L(\mathbf{w}, \vec{\xi}, \rho, \vec{\alpha}, \vec{\beta}) = \\ & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ & - \sum_i \alpha_i((\mathbf{w} \cdot \mathbf{x}_i) - \rho + \xi_i) - \sum_i \beta_i \xi_i \end{aligned} \tag{3}$$

and set the derivatives with respect to the primal variables $\mathbf{w}, \xi_i$ and ρ to be equal to zero, yielding

$$\mathbf{w} = \sum_i \alpha_i \mathbf{x}_i, \tag{4}$$

$$\alpha_i = \frac{1}{\nu\ell} - \beta_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{5}$$

In (4), all patterns $\{\mathbf{x}_i : i \in [\ell], \alpha_i > 0\}$ are called support vectors. The support vector expansion transforms the decision function (2)

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i \mathbf{x}_i \cdot \mathbf{x} - \rho\right). \tag{6}$$

Substituting (4) and (5) into (3), we obtain the dual problem:

$$\min_{\vec{\alpha}} \quad \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \mathbf{x}_i \cdot \mathbf{x}_j \tag{7}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\nu\ell}, \quad \sum_i \alpha_i = 1. \tag{8}$$

One can show that at the optimum condition, the two inequality constraints (1) become equalities if α_i and β_i are nonzero, namely, if $0 < \alpha \leq 1/(\nu\ell)$. Therefore, we can recover ρ by exploiting the fact that for any such α_i, the corresponding pattern $\mathbf{x}_i$ satisfies

$$\rho = (\mathbf{w} \cdot \mathbf{x}_i) = \sum_j \alpha_j \mathbf{x}_j \cdot \mathbf{x}_i. \tag{9}$$

3 Non-relevance Feedback

In this section, we describe a method of document retrieval which uses a one-class SVM for non-relevance feedback.

In relevance feedback, a user has the option of labeling some of the top ranked documents according to whether they are relevant or non-relevant. The labeled documents, along with the original request, are then input to a supervised learning procedure to produce a new classifier. The new classifier is used to produce a new ranking, which retrieves more relevant documents at higher ranks than the original ranking. Non-relevance feedback is used when a user classifies all of the initial top ranked documents as non-relevant.

The relevance feedback based on an SVM assumes both relevant and non-relevant documents which a user has judged. Namely, an SVM, which is a binary classifier, needs both relevant and non-relevant documents. The feedback including only non-relevant documents is not applicable for a two-class SVM classifier.

Non-relevant documents, however, are obtained more easily than relevant documents. In the early stage of relevance feedback, the documents which are retrieved by the system are frequently non-relevant. Using the information on the non-relevant documents may improve the efficiency of document retrieval. In this paper, We propose an efficient retrieval method which uses information on the non-relevant documents only by applying a one-class SVM.

As mentioned in section 2, a one-class SVM clarifies the area of a given class. The area of non-relevant documents in the multidimensional vector space is clarified by a one-class SVM. Therefore, if documents which do not belong to the area of the non-relevant documents are presented, there is a high possibility that a user will judge these documents to be relevant.

The retrieval steps of the proposed method are performed as follows:

Step 1: Preparation of documents for the first feedback
The conventional information retrieval system based on a vector space model displays the top N ranked documents along with a request query to a user. In our method, for the first feedback iteration, the top N ranked documents are selected by using the cosine distance between the request query vector and each document vector.

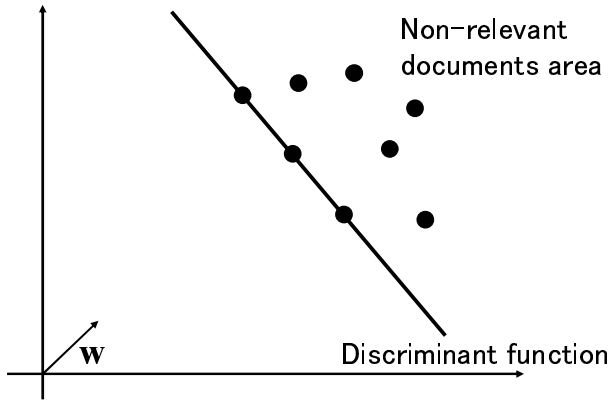

Figure 3. Determination of non-relevant documents area: Circles denote documents which are checked non-relevant by a user. Solid line denotes the discriminant function.

Step 2: Judgment of documents
The user then classifies these N documents as relevant or non-relevant. In cases that the user labels all documents non-relevant, non-relevance feedback is used. Then go to the next step. If these documents are labeled both relevant and non-relevant, then skip to **Step 5**.

Step 3: Determination of non-relevant documents area
The non-relevant documents area is determined by using a one-class SVM which is learned by non-relevant documents only. (see Figure 3).

Step 4: Discrimination of all documents and information retrieval
The one-class SVM learned in the previous step classifies all documents. The documents which are discriminated as being in the "not non-relevant area" are newly selected. From the newly selected documents, the top N ranked documents, which are ranked in order of their distance from the non-relevant documents area, are presented to the user as the information retrieval results of the system (see Figure 4). Then return to **Step 2**.

Step 5: Shift to Relevance feedback
If documents are obtained both relevant and non-relevant, usual relevance feedback is applied.

The non-relevance feedback is intended to present the relevant documents quickly. As mentioned in **Step 4**, the selected documents are discriminated "not non-relevant" and are near the discriminant function. The reason is that we consider these selected documents are not non-relevant and include given queries because non-relevant documents include given queries in this case. If we select a document far from the discriminant function, the document has no relation to the given queries.

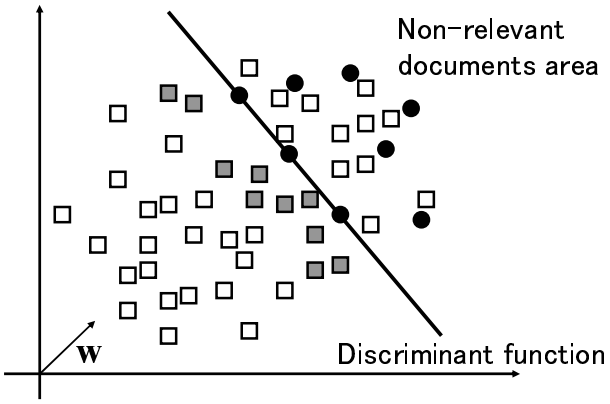

Figure 4. Mapped non-checked documents in the feature space: Boxes denote non-checked documents which are mapped into the feature space. We show the documents which are represented by gray boxes to a user for the next iteration. These documents are top N ranked in the "not non-relevant document area" and are near the discriminant function.

4 Experiments

4.1 Experimental setting

We conducted experiments for evaluating the utility of our method, as reported in section 3. The document data set we used is a set of articles in the Los Angeles Times which has been widely used in the document retrieval conference, TREC [10]. The data set has approximately 130,000 articles. The average number of words in an article is 526. This data set includes not only queries but also the documents relevant to each query. We used three topics for the experiments as shown in Table 1. These topics have no relevant documents in the top 30 ranked documents for retrieval using an initial query vector.

Table 1. Topics used for experiments

topic	query words	# of relevant doc.
306	Africa, civilian, death	34
343	police, death	88
383	mental, ill, drug	55

We used TFIDF [11], which is one of the most popular methods in information retrieval, to generate the document feature vectors, and the concrete equation [8] of a weight of a term t in a document d w_t^d as shown in the following.

$$w_t^d = L \times t \times u \tag{10}$$

$$L = \frac{1+\log(tf(t,d))}{1+\log(\text{average of } tf(t,d) \text{ in } d)} \text{ (TF)}$$

$$t = \log(\frac{n+1}{df(t)}) \text{ (IDF)}$$

$$u = \frac{1}{0.8+0.2\frac{uniq(d)}{\text{average of } uniq(d)}} \text{ (normalization)}$$

The notations in these equation are as follows:

- w_t^d is the weight of a term t in a document d,
- $tf(t,d)$ is the frequency of a term t in a document d,
- n is the total number of documents in a data set,
- $df(t)$ is the number of documents including a term t,
- $uniq(d)$ is the number of different terms in a document d.

The sizes N of the groups of retrieved and displayed results developed in **Step 1** in section 3 were set as 10 and 20. In our experiments, we used the linear kernel for a one-class SVM learning, and found a discriminant function for the one-class SVM classifier in the original feature space. The vector space model of the documents is a high-dimensional space. Moreover, the documents which are labeled by a user are small in number. Therefore, the parameter ν (see section 2) is set to have an adequately small value ($\nu = 0.01$). The small ν means a hard margin in the SVM.

For comparison with our approach, two information retrieval methods were used. The first is an information retrieval method that does not use feedback, namely, documents are retrieved using the ranking in vector space model. The second is an information retrieval method using conventional Rocchio-based relevance feedback [6] which is widely used in information retrieval research.

The Rocchio-based relevance feedback modifies a query vector Q_i on the basis of the evaluation of a user, using the following equation.

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \tag{11}$$

where R_r is a set of documents which were evaluated as relevant by a user at the ith feedback, and R_n is a set of documents which were evaluated as non-relevant at the ith feedback. α and β are weights for the relevant and non-relevant documents, respectively. In this experiment, we set $\alpha = 1.0$ and $\beta = 0.5$ which are decided experimentally.

Table 2. Number of retrieved relevant documents as a function of the number of iterations (number of presented documents is 20).

topic 306	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	1	0
2 (60)	–	–	0
3 (80)	–	–	0
4 (100)	–	–	0
5 (120)	–	–	0

topic 343	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	0	0
2 (60)	–	0	0
3 (80)	–	0	0
4 (100)	–	1	0
5 (120)	–	–	0

topic 383	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (40)	1	0	0
2 (60)	–	1	0
3 (80)	–	–	0
4 (100)	–	–	0
5 (120)	–	–	0

*: Number in parentheses is the number of presented documents at this point.

Table 3. Number of retrieved relevant documents as a function of the number of iterations (number of presented documents is 10).

topic 306	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	1	0	0
2 (30)	–	0	0
3 (40)	–	1	0
4 (50)	–	–	0
5 (60)	–	–	0

topic 343	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	0	0	0
2 (30)	1	0	0
3 (40)	–	0	0
4 (50)	–	0	0
5 (60)	–	0	0

topic 383	# of retrieved relevant doc.		
# of iterations*	One-class	VSM	Rocchio
1 (20)	0	0	0
2 (30)	1	0	0
3 (40)	–	0	0
4 (50)	–	1	0
5 (60)	–	–	0

*: Number in parentheses is the number of presented documents at this point.

4.2 Experimental results

In this experiment, we evaluate how many relevant documents are presented for each feedback iteration. Here, we describe the relationships of the number of feedback iterations with the number of retrieved relevant documents for the proposed method (One-class), for the retrieval using the initial query vector only (VSM) and for the Rocchio-based feedback (Rocchio). Table 2 shows that the number of presented documents is 20, and Table 3 shows that the number of presented documents is 10.

In Table 2, 1 iteration means that a user has judged the twenty documents and is shown the next twenty documents. Therefore, the user has seen forty documents at this point. In Table 3, 1 iteration means that a user has judged the ten documents and has seen twenty documents altogether.

When the proposed method is used, a user can find a relevant document by seeing forty documents for every topic in Table 2. In other words, if the user judges the twenty documents which are retrieved using the initial query vector, the user can then find a relevant document in the next set of retrieved results. When the retrieval using the initial query vector is applied, the user can find a relevant document by seeing forty documents for topic 306, hundred documents for topic 343 and sixty documents for topic 383. However, the user cannot find a relevant document by seeing 120 documents for every topics, when the Rocchio-based method is used. We consider that these results are caused by the feedback method in equation (11). When the relevant documents exist, a useful query vector is created on the basis of the emphasis of the terms in the relevant documents. In the case of the non-relevant documents, however, only the minus term of equation (11) changes.

In Table 3 which shows that the number of presented documents is 10, a user can find a relevant document by judging ten documents for topic 306 when the proposed method is used. This shows that the early feedback on non-relevant documents is effective in the proposed method.

The precision-recall curve of topic 306 when a user has judged ten documents is shown in Figure 5.

This figure shows that the effectiveness of our method in retrieval is improved compared with other methods. With these results, we could confirm that our non-relevance feedback was a useful technique for improving the performance of information retrieval. This figure also shows that

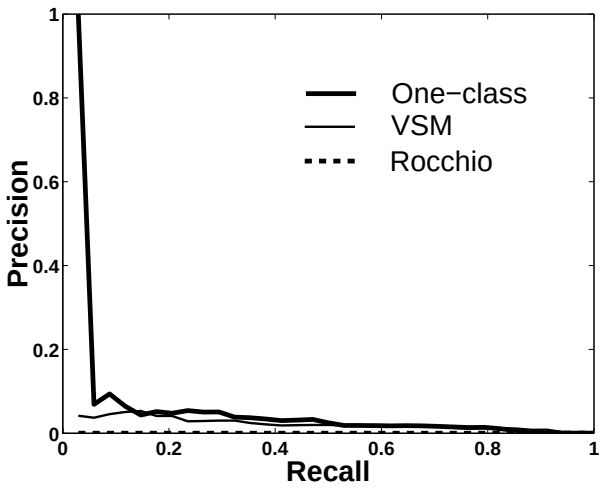

Figure 5. The precision-recall curve of topic 306 when a user has judged 10 documents

Rocchio-based feedback does not work only for the non-relevant documents. In Rocchio-based feedback, the actual relevant documents are found almost at the bottom of the list of retrieved documents. This is attributed to the fact that the non-relevant documents which have query words transform only the minus term of equation (11).

5 Conclusion

In this paper, we proposed the non-relevance feedback method which uses the one-class SVM for enhancement of the information retrieval efficiency. We compared non-relevance feedback with the retrieval using the initial query vector and the Rocchio-based feedback. Results of the experiment on a set of articles in the Los Angeles Times showed that the proposed method gave a better performance than the method it was compared with.

In the task of retrieving information at a user's request from large volumes of data, the information which is obtained at an early stage is often what the user does not want. Our future work will focus on the efficient use of such negative information, in various practical problems.

References

[1] H. Drucker, B. Shahrary, and D. C. Gibbon, Relevance Feedback using Support Vector Machines, *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001, 122–129.

[2] IREX Web page http://nlp.cs.nyu.edu/irex/index-e.html .

[3] NTCIR Web page http://research.nii.ac.jp/ntcir/index-en.html .

[4] M. Okabe and S. Yamada, Interactive Document Retrieval with Relational Learning, *Proceedings of the 16th ACM Symposium on Applied Computing*, 2001, 27–31.

[5] T. Onoda, H. Murata and S.Yamada, Interactive Document Retrieval with Active Learning, *Proceedings of International Workshop on Active Mining*, 2002, 126–131.

[6] J. Rocchio, Relevance Feedback Information Retrieval, in G. Salton (Ed.), *The Smart Retrieval System – Experiments in Automated Document Processing*, (Englewood Cliffs, N.J.: Prentice Hall, 1971) 313–323.

[7] G. Salton and J. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, 1983).

[8] R. Schapire, Y. Singer, and A. Singhal, Boosting and Rocchio Applied to Text Filtering, *Proceedings of the Twenty-First Annual International ACM SIGIR*, 1998, 215–223.

[9] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola and R. Williamson, Estimating the Support of a High-dimensional Distribution, *TR 87, Microsoft Research*, 1999.

[10] TREC Web page http://trec.nist.gov/.

[11] R. B. Yates and B. R. Neto, *Modern Information Retrieval* (Addison Wesley, 1999).