An One Class Classification Approach to Non-relevance Feedback Document Retrieval

Takashi Onoda[1], Hiroshi Murata[1], and Seiji Yamada[2]

[1] Central Research Institute of Electric Power Industry,
System Engineering Research Laboratory, 2-11-1 Iwado Kita,
Komae-shi, Tokyo 201-8511 Japan
{onoda, murata}@criepi.denken.or.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi,
Chiyoda-ku, Tokyo 101-8430 Japan
seiji@nii.ac.jp http://research.nii.ac.jp/~seiji/index-e.html

Abstract. This paper reports a new document retrieval method using non-relevant documents. From a large data set of documents, we need to find documents that relate to human interesting in as few iterations of human testing or checking as possible. In each iteration a comparatively small batch of documents is evaluated for relating to the human interesting. The relevance feedback needs a set of relevant and non-relevant documents to work usefully. However, the initial retrieved documents, which are displayed to a user, sometimes don't include relevant documents. In order to solve this problem, we propose a new feedback method using information of non-relevant documents only. We named this method *non-relevance feedback document retrieval*. The non-relevance feedback document retrieval is based on One-class Support Vector Machine. Our experimental results show that this method can retrieve relevant documents using information of non-relevant documents only.

1 Introduction

As the Internet technology progresses, accessible information by end users is explosively increasing. In this situation, we can now easily access a huge document database through the Web. However it is hard for a user to retrieve relevant documents from which he/she can obtain useful information, and a lot of studies have been done in information retrieval, especially document retrieval [1]. Many works for such document retrieval have been reported in TREC (Text Retrieval Conference) [2] for English documents, IREX (Information Retrieval and Extraction Exercise) [3] and NTCIR (NII-NACSIS Test Collection for Information Retrieval System) [4] for Japanese documents.

In most frameworks for information retrieval, a vector space model in which a document is described with a high-dimensional vector is used [5]. An information retrieval system using a vector space model computes the similarity between a query vector and document vectors by the cosine of the two vectors and indicates a user a list of retrieved documents.

L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 1216–1225, 2005.

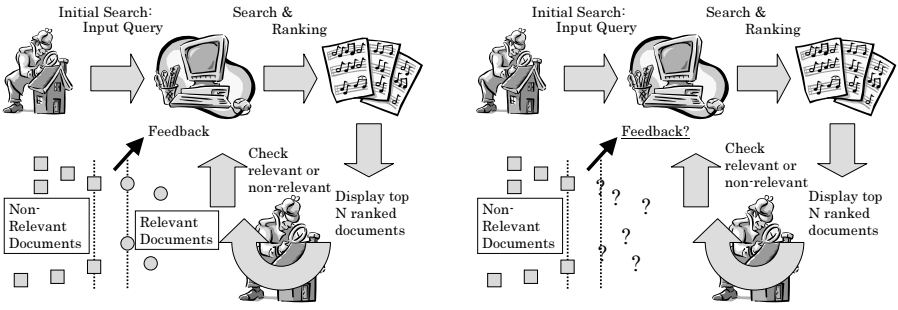

Fig. 1. Outline of the relevance feedback documents retrieval(left side) and Image of a problem in the relevance feedback documents retrieval(right side): The gray arrow parts are made iteratively to retrieve useful documents for the user. This iteration is called feedback iteration in the information retrieval research area. But if the evaluation of the user has only non-relevant documents, ordinary relevance feedback methods can not feed back the information of useful retrieval.

In general, since a user hardly describes a precise query in the first trial, interactive approach to modify the query vector using evaluation of the documents on a list of retrieved documents by a user. This method is called *relevance feedback* [6] and 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 a system modifies the query vector using the user evaluation. A traditional way to modify a query vector is a simple learning rule to reduce the difference between the query vector and documents evaluated as relevant by a user (see Figure 1 left side).

In another approach, relevant and irrelevant document vectors are considered as positive and negative examples, and relevance feedback is transposed to a binary class classification problem [7]. For the binary class classification problem, Support Vector Machines (which are called SVMs) have shown the excellent ability. And some studies applied SVM to the text classification problems [8] and the information retrieval problems [9]. Recently, we have proposed a relevance feedback framework with SVM as *active learning* and shown the usefulness of our proposed method experimentally [10].

The initial retrieved documents, which are displayed to a user, sometimes don't include relevant documents. In this case, almost all relevance feedback document retrieval systems do not work well, because the systems need relevant and non-relevant documents to construct a binary class classification problem (see Figure 1 right side).

While a machine learning research field has some methods which can deal with one class classification problem. In the above document retrieval case, we can use non-relevant documents information only. Therefore, we consider this retrieval situation is as same as one class classification problems.

In this paper, we propose a framework of an interactive document retrieval using non-relevant documents information only. We call this interactive docu-

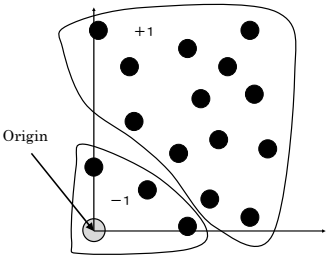

Fig. 2. One-Class SVM Classifier: the origin is the only original member of the second class

ment retrieval as *non-relevance feedback document retrieval*, because we can use non-relevant documents information only. Our proposed non-relevance document retrieval is based on One Class Support Vector Machine(One-Class SVM) [11]. One-Class SVM can generate a discriminant hyperplane that can separate the non-relevant documents which are evaluated by a user. Our proposed method can display documents, which may be relevant documents for the user, using the discriminant hyperplane.

In the remaining parts of this paper, we explain the One-Class SVM algorithm in the next section briefly, and propose our document retrieval method based on One-Class SVM in the third section. In the fourth section, in order to evaluate the effectiveness of our approach, we made experiments using a TREC data set of Los Angels Times and discuss the experimental results. Finally we conclude our work and discuss our future work in the fifth section.

2 One-Class Support Vector Machine

Schölkopf et al. suggested a method of adapting the SVM methodology to one-class classification problem. Essentially, after transforming the feature via a kernel, they treat the origin as the only member of the second class. The using "relaxation parameters" they separate the image of the one class from the origin. Then the standard two-class SVM techniques are employed.

One-class SVM [11] returns a function f that takes the value +1 in a *small* region capturing most of the training data points, and -1 elsewhere.

The algorithm can be summarized as mapping the data into a feature space H using an appropriate kernel function, and then trying to separate the mapped vectors from the origin with maximum margin (see Figure 2).

Let the training data be

$$\mathbf{x}_1, \ldots, \mathbf{x}_\ell \tag{1}$$

belonging to one class X, where X is a compact subset of R^N and ℓ is the number of observations. Let $\Phi : X \rightarrow H$ be a kernel map which transforms the training examples to feature space. The dot product in the image of Φ can be computed by evaluating some simple kernel

$$k(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})) \tag{2}$$

such as the linear kernel, which is used in our experiment,

$$k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}. \tag{3}$$

The strategy is to map the data into the feature space corresponding to the kernel, and to separate them from the origin with maximum margin. Then, to separate the data set from the origin, one needs to solve the following quadratic program:

$$\begin{aligned} \min_{\mathbf{w} \in H, \xi \in R^\ell \rho \in R^N} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu\ell}\sum_i \xi_i - \rho \\ \text{subject to} \quad & (\mathbf{w} \cdot \Phi(\mathbf{x}_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0. \end{aligned} \tag{4}$$

Here, $\nu \in (0, 1)$ is an upper bound on the fraction of outliers, and a lower bound on the fraction of Support Vectors (SVs).

Since 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

$$f(\mathbf{x}) = \operatorname{sgn}\left((\mathbf{w} \cdot \Phi(\mathbf{x})) - \rho\right) \tag{5}$$

will be positive for most examples $\mathbf{x}_i$ contained in the training set, while the SV type regularization term $\|w\|$ will still be small. The actual trade-off between these two is controlled by ν. For a new point $\mathbf{x}$, the value $f(\mathbf{x})$ is determined by evaluating which side of the hyperplane it falls on, in feature space.

In our research we used the LIBSVM. This is an integrated tool for support vector classification and regression which can handle one-class SVM using the Schölkopf etc algorithms. The LIBSVM is available at
`http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

3 Non-relevance Feedback Document Retrieval

In this section, we describe our proposed method of document retrieval based on Non-relevant documents using One-class SVM.

In relevance feedback document retrieval, the 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 given to a supervised learning procedure to produce a new classifier. The new classifier is used to produce a new ranking, which retrievals more relevant documents at higher ranks than the original did (see Figure 1 left side) [10].

The initial retrieved documents, which are displayed to a user, sometimes don't include relevant documents. In this case, almost all relevance feedback document retrieval systems do not contribute to efficient document retrieval, because the systems need relevant and non-relevant documents to construct a binary class classification problem (see Figure 1 right side).

The One-Class SVM can generate discriminant hyperplane for the one class using one class training data. Consequently, we propose to apply One-Class SVM in a *non-relevance feedback document retrieval method*. The retrieval steps of proposed method perform as follows:

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

Step 2: Judgment of documents
The user then classifiers these N documents into relevant or non-relevant. If the user labels all N documents non-relevant, the documents are labeled "+1" and go to the next step. If the user classifies the N documents into relevant documents and non-relevant documents, the non-relevant documents are labeled "+1" and relevant documents are labeled "-1" and then our previous proposed relevant feedback method is adopted [10].

Step 3: Determination of non-relevant documents area based on non-relevant documents
The discriminant hyperplane for classifying non-relevant documents area is generated by using One-Class SVM. In order to generate the hyperplane, the One-Class SVM learns labeled non-relevant documents which are evaluated in the previous step (see Figure 3 left side).

Step 4: Classification of all documents and Selection of retrieved documents
The One-class SVM learned by previous step can classifies the whole documents as non-relevant or not non-relevant. The documents which are discriminated in *not non-relevant are* are newly selected. From the selected documents, the top N ranked documents, which are ranked in the order of the distance from the non-relevant documents area, are shown to user as the document retrieval results of the system (see Figure 3 right side). These N documents have high existence probability of initial keywords. Then return to Step 2.

The feature of our One-Class SVM based non-relevant feedback document retrieval is the selection of displayed documents to a user in Step 4. Our proposed method selects the documents which are discriminated as *not non-relevant* and near the discriminant hyperplane between non-relevant documents and not non-relevant documents. Generally if the system got the opposite information from a user, the system should select the information, which is far from the opposite information area, for displaying to the user. However, in our case, the classified non-relevant documents by the user includes a request query vector of the user. Therefore, if we select the documents, which are far from the non-relevant documents area, the documents may not include the request query of the user.

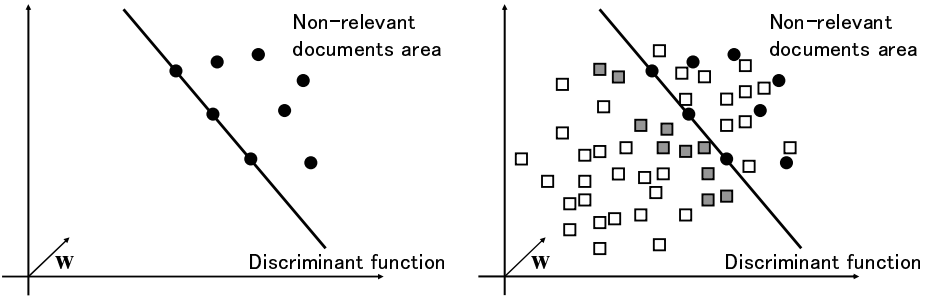

Fig. 3. Generation of a hyperplane to discriminate non-relevant documents area and Mapped non-checked documents into the feature space: Circles denote documents which are checked non-relevant by a user. The solid line denotes the discriminant hyperplane. Boxes denote non-checked documents which are mapped into the feature space. Gray boxes denotes the displayed documents to a user in the next iteration. These documents are in the *not non-relevant document area* and near the discriminant hyperplane.

Our selected documents (see Figure 3 right side) is expected that the probability of the relevant documents for the user is high, because the documents are not non-relevant and may include the query vector of the user.

4 Experiments

4.1 Experimental Setting

We made experiments for evaluating the effectiveness of our interactive document retrieval based on non-relevant documents using One-Class SVM described in section 3. The document data set we used is a set of articles in the Los Angels Times which is widely used in the document retrieval conference TREC [2]. The data set has about 130 thousands articles. The average number of words in a article is 526. This data set includes not only queries but also the relevant documents to each query. Thus we used the queries for the experiments. We used three topics for experiments and show these topics in Table 1. These topics do not have relevant documents in top 20 ranked documents in the order of the cosine distance between the query vector and document vectors. Our experiments set the size of N of displayed documents presented in Step 1 in the section 3 to 10 or 20.

Table 1. Topics, query words and the number of relevant documents in the Los Angels Times 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 [1], which is one of the most popular methods in information retrieval to generate document feature vectors, and the concrete equation [12] of a weight of a term t in a document d w_t^d are in the following.

$$w_t^d = L \times t \times u \tag{6}$$

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

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

The notations in these equation denote as follows:

- w_t^d is a weight of a term t in a document d,
- $tf(t,d)$ is a 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.

In our experiments, we used the linear kernel for One-class SVM learning, and found a discriminant function for the One-class SVM classifier in the feature space. The vector space model of documents is high dimensional space. Moreover, the number of the documents which are evaluated by a user is small. Therefore, we do not need to use the kernel trick and the parameter ν (see section 2) is set adequately small value ($\nu = 0.01$). The small ν means hard margin in the One-Class SVM and it is important to make hard margin in our problem.

For comparison with our approach, two information retrieval methods were used. The first is an information retrieval method that does not use a feedback, namely documents are retrieved using the rank in vector space model (VSM). The second is an information retrieval method using the 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 by evaluation of a user using the following equation.

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \tag{7}$$

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

4.2 Experimental Results

Here, we describe the relationships between the performances of the proposed method and the number of feedback iterations. Table 2 left side gave the number

Table 2. The number of retrieved relevant documents at each iteration: the number of displayed documents is 10 at each iteration

	# of displayed doc. is 10			# of displayed doc. is 20		
topic 306	# of retrieved relev. doc.			# of retrieved relev. doc.		
# of iter.	Proposed	VSM	Rocchio	Proposed	VSM	Rocchio
1	1	0	0	1	1	0
2	–	0	0	–	–	0
3	–	1	0	–	–	0
4	–	–	0	–	–	0
5	–	–	0	–	–	0
topic 343	# of retrieved relev. doc.			# of retrieved relevant doc.		
# of iter.	Proposed	VSM	Rocchio	Proposed	VSM	Rocchio
1	0	0	0	1	0	0
2	1	0	0	–	0	0
3	–	0	0	–	0	0
4	–	0	0	–	1	0
5	–	0	0	–	–	0
topic 383	# of retrieved relev. doc.			# of retrieved relevant doc.		
# of iter.	Proposed	VSM	Rocchio	Proposed	VSM	Rocchio
1	0	0	0	1	0	0
2	1	0	0	–	1	0
3	–	0	0	–	–	0
4	–	1	0	–	–	0
5	–	–	0	–	–	0

of retrieved relevant documents at each feedback iteration. At each feedback iteration, the system displays ten higher ranked *not non-relevant* documents, which are near the discriminant hyperplane, for our proposed method. We also show the retrieved documents of the Rocchio-based method at each feedback iteration for comparing to the proposed method in table 2 left side.

We can see from this table that our non-relevance feedback approach gives the higher performance in terms of the number of iteration for retrieving relevant document. On the other hand, the Rocchio-based feedback method cannot search a relevant document in all cases. The vector space model without feedback is better than the Rocchio-based feedback. After all, we can believe that the proposed method can make an effective document retrieval using only non-relevant documents, and the Rocchio-based feedback method can not work well when the system can receive the only non-relevant documents information.

Table 2 right side gave the number of retrieved relevant documents at each feedback iteration. At each feedback iteration, the system displays twenty higher ranked *not non-relevant* documents, which are near the discriminant hyperplane, for our proposed method. We also show the retrieved documents of the Rocchio-based method at each feedback iteration for comparing to the proposed method in table 2 right side.

We can observe from this table that our non-relevance feedback approach gives the higher performance in terms of the number of iteration for retrieving

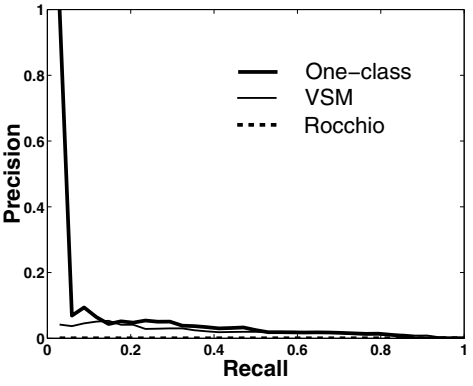

Fig. 4. The precision and recall curve of topic no. 304 at the second iteration

relevant documents, and the same experimental results as table 2 left side about the Rocchio-based method and VSM.

In table 2 left side, a user already have seen twenty documents at the first iteration. Before the fist iteration, the user have to see ten documents, which are retrieved results using the cosine distance between a query vector and document vectors in VSM. In table 2 right side, the user also have seen forty documents at the first iteration. Before the fist iteration, the user also have to see ten documents to evaluate the documents, which are retrieved results using the cosine distance between a query vector and document vectors in VSM.When we compare the experimental results of table 2 left side with the results of table 2 right side, we can observe that the small number of displayed documents makes more effective document retrieval performance than the large number of displayed documents. In table 2 left side, the user had to see thirty documents by finding the first relevant document about topic 343 and 383. In table 2 right side, the user had to see forty documents by finding the first relevant document about topic 343 and 383. Therefore, we believe that the early non-relevance feedback is useful for an interactive document retrieval.

We also show a precision and recall curve in figure 4. This figure is the precision and recall curve of topic no. 306 at the second iteration. From this figure, we can understand that all precision-recall curves are not good. However, our proposed approach is more efficient than the two other approaches.

5 Conclusion

In this paper, we proposed the non-relevance feedback document retrieval based on One-Class SVM using only non-relevant documents for a user. In our non-relevance feedback document retrieval, the system use only non-relevant documents information. One-Class SVM can generate a discriminant hyperplane of observed one class information, so our proposed method adopted One-Class SVM for non-relevance feedback document retrieval.

This paper compared our method with a conventional relevance feedback method and a vector space model without feedback. Experimental results on a set of articles in the Los Angels Times showed that the proposed method gave a consistently better performance than the compared method. Therefore we believe that our proposed One-Class SVM based approach is very useful for the document retrieval with only non-relevant documents information.

This paper proposed that the system should display the documents which are in the *not non-relevant* documents area and near the discriminant hyperplane of One-Class SVM at each feedback iteration. However, we do not discuss how the selection of documents influence both the effective learning and the performance of information retrieval theoretically. This point is our future work.

References

1. Yates, R.B., Neto, B.R.: Modern Information Retrieval. Addison Wesley (1999)
2. TREC Web page: (http://trec.nist.gov/)
3. IREX: (http://cs.nyu.edu/cs/projects/proteus/irex/)
4. NTCIR: (http://www.rd.nacsis.ac.jp/˜ntcadm/)
5. Salton, G., McGill, J.: Introduction to modern information retrieval. McGraw-Hill (1983)
6. Salton, G., ed. In: Relevance feedback in information retrieval. Englewood Cliffs, N.J.: Prentice Hall (1971) 313–323
7. Okabe, M., Yamada, S.: Interactive document retrieval with relational learning. In: Proceedings of the 16th ACM Symposium on Applied Computing. (2001) 27–31
8. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: Journal of Machine Learning Research. Volume 2. (2001) 45–66
9. Drucker, H., Shahrary, B., Gibbon, D.C.: Relevance feedback using support vector machines. In: Proceedings of the Eighteenth International Conference on Machine Learning. (2001) 122–129
10. Onoda, T., Murata, H., Yamada, S.: Relevance feedback with active learning for document retrieval. In: Proc. of IJCNN2003. (2003) 1757–1762
11. Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support for a high-dimensional distribution. Technical Report MSR-TR-99-87, Microsoft Research, One Microsoft Way Redmon WA 98052 (1999)
12. Schapire, R., Singer, Y., Singhal, A.: Boosting and rocchio applied to text filtering. In: Proceedings of the Twenty-First Annual International ACM SIGIR. (1998) 215–223