Comparison of Learning Performance and Retrieval Performance for Support Vector Machines based Relevance Feedback Document Retrieval

Takashi Onoda, Hiroshi Murata
Central Research Institute of
Electric Power Industry
2-11-1, Iwado Kita, Komae-shi,
Tokyo 201-8511 JAPAN
{onoda, murata}@criepi.denken.or.jp

Seiji Yamada
National Institute of Informatics
2-1-1, Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430 JAPAN
seiji@nii.ac.jp

Abstract

This paper presents a learning performance and a retrieval performance of an interactive document retrieval method, which is based on Support Vector Machine(SVM). Some works have been done to apply classification learning like SVM to relevance feedback and obtained successful results. However they did not fully utilize characteristic of example distribution in document retrieval. We propose heuristics to bias document showing in order to take good learning performance and good retrieval performance of relevance feedback. This paper introduces two evaluation crietria. One criterion measures the learning performance and the other measures the retrieval performance. We compared a SVM-based system with our heuristic with conventional systems like Rocchio-based system and a SVM-based system without our heuristic by the introduced crietria. We could confirm the learning performance of our system outperformed other ones.

1. Introduction

As 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 WWW. 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 [11]. Active works for such document retrieval have been reported in TREC(Text Retrieval Conference) [9] for English documents, 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(which is called VSM) in which a document is described with a high-dimensional vector is used [7]. An information retrieval system using a vector space model computes the similarity between a query vector and document vectors by cosine of the two vectors and indicates a user a list of retrieved documents.

In general, since a user hardly describes a precise query in the first trial, interactive approach to modify the query vector by evaluation of the user on documents in a list of retrieved documents. This method is called *relevance feedback* [6] and used widely in information retrieval systems. In this method, a user directly evaluates whether a document is relevant or irrelevant in a list of retrieved documents, 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.

In another approach, relevant and irrelevant document vectors are considered as positive and negative examples, and relevance feedback is transposed to a binary classification problem [4]. For the binary 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[1]. Now, we are interested in how to evaliate the peformance of relevance feedback document retrieval methods. We think two perfomances should be evaluated for relevance feedback document retrieval methods. They are the learning performance and the retrieval performace. This paper introduces two evaluation crietria. One criterion measures the learning performance and the other measures the retrieval performance. We compared a SVM-based system with our heuristic with conventional systems like Rocchio-based system and a SVM-based system without our heuristic by the introduced crietria.

0-7695-3028-1/07 $25.00 © 2007 IEEE
DOI 10.1109/WI-IATW.2007.34

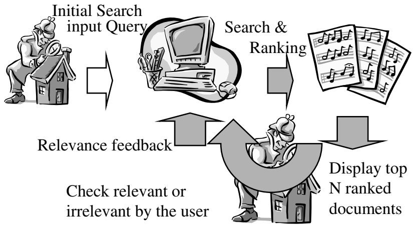

Figure 1. Image of the interactive document retrieval with the relevance feedback

2. SVM based Interactive Document Retrieval

In this section, we describe the information retrieval system using relevance feedback with SVM.

Figure 1 shows the concept of the interactive document retrieval with the relevance feedback. In Figure 1, the iterative procedure is the gray arrows parts. The SVMs have a great ability to discriminate even if the training data is small. Consequently, we have proposed to apply SVMs as the classifier in the relevance feedback method. The retrieval steps of proposed method perform as following procedure:

Step 1: Preparation 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 and displayed by using cosine distance between the request query vector and each document vector for the first feedback iteration.

Step 2: Judgment of documents by a user

A user then evaluates and classifies these N displayed documents into relevance or non-relevance. After the user's evaluation, the relevant documents have relevance label and the non-relevant documents have non-relevance label. For example, the relevant documents have "+1" label and the non-relevant documents have "-1" label generally, after the user's judgment.

Step 3: Determination of the optimal hyper-plane

The optimal hyper-plane for classifying relevant and non-relevant documents is generated by using a SVM which is learned by labeled documents(see the left side of Figure 2).

Step 4: Selection of documents

Non-checked documents, which are not checked by the user, are mapped into the feature space that has the optimal hyper-plane as a discriminant function. The SVM learned at the previous step classifies the non-checked documents as relevant or non-relevant. Then the system selects the documents based on the distance from the optimal hyper-plane and the distribution of the relevant documents and non-relevant documents. The feature of this distribution can be used as a prior knowledge. The detail of the selection rules are described in the next section. If the number of feedback iterations is more than m, then go to next step. Otherwise, according to the selection rule, N documents are displayed to the user to check these documents by the user and return to Step 2. The m is a maximal number of feedback iterations and is given by the user.

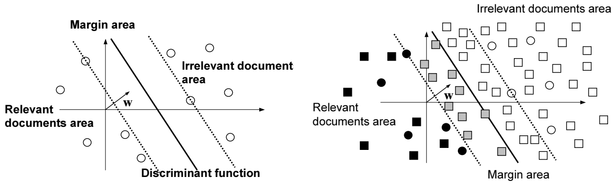

Figure 2. The discriminant function and the displayed documents

Step 5: Display of the final retrieved documents

The all documents, which include the checked documents, are mapped into the feature space that has the optimal hyper-plane as the discriminant function. The all documents are ranked by the distance between the documents and the hyper-plane which is the discriminant function determined by SVM. According to this rank, the order of the displayed documents is determined.(see the right side Figure 2). According to this order, the system displays top H ranked documents to the user as the final retrieval results.

3. Selection Rules of Displayed Documents

In this section, we discuss two selection rules for displayed documents, which are displayed to the user and used for the judgment by the user. In this paper, we compare the effectiveness of the document retrieval and the learning performance among the following two selection rules for displayed documents.

3.1. Proposed Selection Rule

The all documents are mapped into the feature space. The learned SVM classifies the documents as relevant or non-relevant. The documents, which are discriminated relevant and in the margin area of SVM, are selected. The top N ranked documents, which are ranked using the distance from the optimal hyperplane, are displayed to the user as

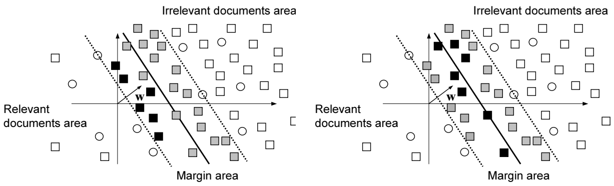

Figure 3. The general selection rule and the proposed selection rule

the information retrieval results of the system(see the left side of Figure 3). This rule is expected to achieve the most effective retrieval performance. This rule is our proposed one for the relevance feedback document retrieval. We call the system based on this rule "SVM-A".

3.2. General Selection Rule

The all documents are mapped into the feature space. The learned SVM classifies the documents as relevant or non-relevant. The documents, which are on the optimal hyperplane or near the optimal hyperplane of SVM, are selected. The system chooses the top N ranked documents in these selected documents to display to the user as the information retrieval results of the system(see the right side of Figure 3). This rule is expected to simply achieve the best learning performance from an active learning point of view. We call the system based on this rule "SVM-S".

4. Experiments

4.1. Experimental setting

We made experiments for evaluating the learning performance and the retrieval performance of our proposed interactive document retrieval with SVM in section 2. The document data set we used is a set of articles in ad hoc task which was widely used in the document retrieval conference 6th, 7th and 8th TREC[9]. The data set has about 530 thousands news paper articles. Each TREC provides 50 retrieval problems and the information of relevant documents for each retrieval problem. Hereafter, we call the retrieval problem "topic". In our experiments, 150 topics are tested. Each topic has three tags, which consist of a title tag, a description tag, and a narrative tag. The title tag has 2 or 3 terms to describe the topic. The description tag introduces the topic. The narrative tag reports the topic. Our experiments used 2 or 3 terms of the title tag as a query. And our experiments also removed the stopword and made stemming for documents and queries.

We used TFIDF[11], which is one of the most popular methods in information retrieval to generate document feature vectors.

The size N of retrieved and displayed documents in **Step 1** in section 3 was set as 10, 20, 40. The feedback iterations m were 1, 2, 3, 4, 5. This situation means that a user evaluates from 10 documents to 200 documents by getting final retrieval results.

In our experiments, we used the linear kernel for SVM learning, and found a discriminant function for the SVM classifier in this feature space. The VSM of documents is high dimensional space. Therefore, in order to classify the labeled documents into relevant or irrelevant, we do not need to use the kernel trick and the regularization parameter.

For comparison with our approach, three information retrieval methods were adopted. The first is an information retrieval method that uses the selection rule 2, which is described in section 4. The second is an information retrieval method using conventional Rocchio-based relevance feedback[6] which is widely used in information retrieval research. The third is Okapi system based on BM25[5]. The Okapi system uses a probabilistic model of documents and shows the good performance for the document retrieval.

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

$$Q_{i+1} = Q_i + \alpha \sum_{x \in R_r} x - \beta \sum_{x \in R_n} x, \qquad (1)$$

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

In order to compare the learning performance of our proposed method with the other methods, we evaluated the following criterion.

P30: Precision within the top 30 documents, which is a ratio of relevant documents in the top 30 documents.

Therefore, after our relevance feedback document retrieval procedure, whcih dicribed in section 2, we set the number of final displayed documents H to 30. And in order to compare the retrieval performance of our proposed method with the other methods, we evaluated the following criterion P.

$$P = \frac{N_{rel}}{N_{dis}}$$

where, N_{rel} denotes the number of relevance documents in the all displayed documents and N_{dis} denotes the number of the all displayed documents.

Table 1. Evaluation of Learning Performance

The number of displayed documents is 10

m	SVM-A	SVM-S	Rocchio	Okapi
1	0.333	0.333	0.401	0.467
2	0.476	0.454	0.459	0.493
3	0.551	0.548	0.485	0.511
4	0.610	0.614	0.511	0.515
5	0.680	0.664	0.541	0.525

The number of displayed documents is 20

m	SVM-A	SVM-S	Rocchio	Okapi
1	0.410	0.410	0.436	0.489
2	0.553	0.538	0.520	0.529
3	0.665	0.642	0.572	0.545
4	0.732	0.711	0.606	0.548
5	0.771	0.758	0.628	0.555

Table 2. Evaluation of Retrieval Performance

The number of displayed documents is 10

m	SVM-A	SVM-S	Rocchio	Okapi
1	0.341	0.341	0.383	0.500
2	0.372	0.318	0.393	0.471
3	0.381	0.319	0.384	0.439
4	0.376	0.319	0.366	0.403
5	0.372	0.319	0.353	0.376

The number of displayed documents is 20

m	SVM-A	SVM-S	Rocchio	Okapi
1	0.293	0.293	0.324	0.412
2	0.312	0.279	0.333	0.380
3	0.318	0.280	0.312	0.343
4	0.311	0.275	0.291	0.315
5	0.311	0.267	0.272	0.290

4.2. Experimental results

Table 1 and 2 show the experimental results. Table 1 shows the values of P30 to evaluate the learning performance when the number of displayed documents at each iteration is 10 and 20. Table 2 shows the values of P to evaluate the retrieval performance when the number of displayed documents at each iteration is 20. Each value is an average value for topics in the table 1 and 2.

From table 1, we can understand that our proposed interactive document retrieval method shows better learning performance than the other methods, when the number of iterations is over 2. However, the learning performance of our proposed method could not win Rocchio-based method and Okapi after the first feedback iteration. In the first feedback iteration, our proposed method can not use the good selection rule for displayed documents. Therefore, our proposed method could not win them after just first feedback.

From table 2, we can understand that our proposed interactive document retrieval method shows better retrieval performance than the other methods, when the number of iterations is over 4. However, the learning performance of our proposed method could not win Rocchio-based method and Okapi when the number of iterations 1, 2, 3. The value of our criterion to evaluate the retrieval performance depends on the first retrieval performance. Okapi shows much better retrieval performance than our proposed method in the initial retrieval. Therefore, our proposed method could not win Okapi when the number of iterations 1, 2, 3.

5 Conclusion

In this paper, we proposed an interactive document retrieval based on SVM using the special selection rule, which can select the effective documents for the retrieval system and the user. In our experiments, our proposed interactive document retrieval method showed better learning performance than the two conventional methods and one probabilistic method.

The proposed selection rule 1, where the documents that are near the bound of the relevant documents area and in the margin area of SVM, are displayed to a user, show better performance of document retrieval. Generally, data mining applications based SVM to drug discovery[10], bioinformatics and so on are discussed from active learning point of view. In the discussion, the learning performance of computers is very important. In the interactive documents retrieval between human and computers, however, the retrieval performance is more important than the learning performance of computers. Therefore, interactive documents retrieval systems should display the documents which are interesting for users at each iteration and keep the learning performance of the systems. In this paper, we showed the selection rule of documents is very important to display the interesting documents to users and keep the learning performance in the interactive documents retrieval.

References

[1] H. Drucker, B. Shahrary, and D. C. Gibbon. Relevance feedback using support vector machines. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 122–129, 2001.

[2] IREX. http://cs.nyu.edu/cs/projects/proteus/irex/.

[3] NTCIR. http://www.rd.nacsis.ac.jp/˜ntcadm/.

[4] M. Okabe and S. Yamada. Interactive document retrieval with relational learning. In *Proceedings of the 16th ACM Symposium on Applied Computing*, pages 27–31, 2001.

[5] S. Robertson and S. Walker. Okapi/keenbow at trec-8. In *Proceedings of TREC 8*, pages 151–162, 2000.

[6] G. Salton, editor. *Relevance feedback in information retrieval*, pages 313–323. Englewood Cliffs, N.J.: Prentice Hall, 1971.

[7] G. Salton and J. McGill. *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[8] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. In *Journal of Machine Learning Research*, volume 2, pages 45–66, 2001.

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

[10] M. Warmuth, G. Rätsch, M. Mathieson, J. Liao, and C. Lemmen. Active learning in the drug discovery process. In *Advances in Neural information processings systems*, volume 14, 2002.

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