Support Vector Machines based Active Learning for the 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 describes an application of SVM(Support Vector Machines) to interactive document retrieval using active learning. 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 according to distribution of examples in document retrieval. This heuristic is executed by selecting examples to show a user in neighbors of positive support vectors, and it improves learning efficiency. We implemented a SVM-based interactive document retrieval system using our proposed heuristic, and compare it with conventional systems like Rocchio-based system and a SVM-based system without the heuristic. We conducted systematic experiments using large data sets including over 500,000 paper articles and confirmed 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 [10]. Active works for such document retrieval have been reported in TREC(Text Retrieval Conference) [8] 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 [6]. 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* [5] 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 [7] and the information retrieval problems[1]. Now, we are interested in what is the most useful selection rule for displayed documents at each iteration to users. In this paper, we adopt several selection rules of displayed documents at each iteration, and then show the comparison results of the effectiveness for the document retrieval in these several selection rules.

2. Active Learning with SVM in Interactive Document Retrieval

In this section, we describe the information retrieval system using relevance feedback with SVM from an active

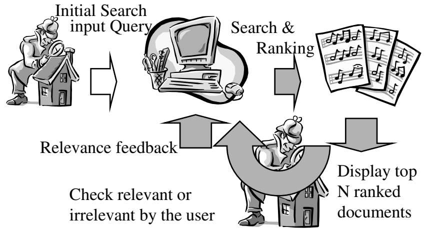

Figure 1. Image of the interactive document retrieval with the relevance feedback

learning point of view.

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

The documents, which are retrieved in the Step1, are mapped into the feature space that has the optimal hyper-plane as a discriminant function. The SVM learned by 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. The top N ranked documents, which are ranked using the distance from the optimal hyper-plane, are displayed to the user as the information retrieval results of the system. If the number of feedback iterations is more than m, then go to next step. Otherwise, return to Step 2. The m is a maximal number of feedback iterations and is given by the user.

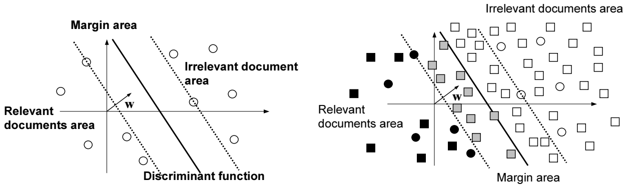

Figure 2. The discriminant function and the displayed documents

Step 5: Display of the final retrieved documents

The all documents 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).

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 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.

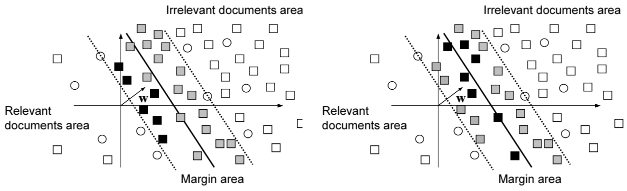

Figure 3. The general selection rule and the proposed selection rule

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 achieve the best learning performance from an active learning point of view generally.

4. Experiments

4.1. Experimental setting

We made experiments for evaluating the effectiveness of our proposed interactive document retrieval with active learning of 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[8]. 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[10], 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 or 20. The feedback iterations m were 5 or 2. This situation means that a user evaluates 50 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, two 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[5] 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.

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

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 usefulness of our proposed method with the other methods, we evaluated the following criteria:

P10: Precision within the top 10 documents, which is a ratio of relevant documents in the top 10 documents.

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

MAP: The average value of all precisions at every relevant document for a topic. When a relevant document is not retrieved, this value is 0.

R05P: Recall when precision first becomes less than 0.5 from the top document. 10 documents need to be checked at least.

P10 and P30 evaluate the precision of the document retrieval method and are the user oriented measures for the effectiveness of the document retrieval method. MAP and R05P evaluate the recall of the document retrieval method and are the system oriented measures for the learning performance of the document retrieval method.

4.2. Experimental results

Table 1 and 2 show the experimental results. Table 1 shows the experimental results when the number of displayed documents at each iteration is 10. Table 2 shows

Table 1. Experimental Results (A)

The number of displayed documents is 10

F#	P10			P30		
	S-A	S-N	Ro	S-A	S-N	Ro
1	0.368	0.169	0.301	0.247	0.140	0.214
2	0.436	0.291	0.282	0.317	0.231	0.218
3	0.407	0.304	0.255	0.301	0.261	0.200
4	0.350	0.319	0.223	0.268	0.261	0.181
5	0.357	0.303	0.225	0.275	0.233	0.170

F#	MAP			R05P		
	S-A	S-N	Ro	S-A	S-N	Ro
1	0.151	0.080	0.146	0.139	0.079	0.129
2	0.182	0.137	0.144	0.153	0.097	0.113
3	0.172	0.142	0.131	0.126	0.092	0.102
4	0.154	0.146	0.121	0.112	0.095	0.085
5	0.150	0.135	0.112	0.103	0.078	0.079

Table 2. Experimental Results (B)

(b) The number of displayed documents is 20

F#	P10			P30		
	S-A	S-N	Ro	S-A	S-N	Ro
1	0.404	0.189	0.266	0.285	0.141	0.201
2	0.427	0.314	0.271	0.310	0.253	0.206

F#	MAP			R05P		
	S-A	S-N	Ro	S-A	S-N	Ro
1	0.177	0.087	0.130	0.158	0.075	0.119
2	0.182	0.150	0.125	0.138	0.113	0.096

the experimental results when the number of displayed documents at each iteration is 20. In these tables, F# denotes the number of feedback iterations. S-A, S-N, and Ro are the compared interactive document retrieval methods. S-A denotes our proposed interactive document retrieval method based on the selection rule 1 using SVM. S-N denotes the general interactive document retrieval method based on the selection rule 2 using SVM. Ro denotes Rocchio-based interactive document retrieval method. Each value is an average value for topics in the table 1 and 2. The value with underline denotes the best performance.

From these tables, we can understand that our proposed interactive document retrieval method shows better performance than the other methods for all criteria. Therefore, our proposed interactive document retrieval method achieves more effective retrieval performance and better learning performance than the others.

When the number of feedback iterations is small, we can understand that Rocchio-based interactive document retrieval method shows better retrieval performance than the interactive document retrieval based on the selection rule 2 from the tables. Especially, Rocchio-based interactive document retrieval method shows much better retrieval performance than the interactive document retrieval based on the selection rule 2 when the number of iterations is one. The interactive document retrieval based on the selection rule 2 and our proposed method use SVM. But our proposed method always shows better retrieval performance than the others. Therefore, we understand the proposed selection rule 1 is very useful for the interactive document retrieval.

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 more effective document retrieval and better learning performance than the two conventional methods using some kind of criteria.

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

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

[7] 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.

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

[9] 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.

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