Paper:

Comparative Analysis of Relevance for SVM-Based Interactive Document Retrieval

Hiroshi Murata*, Takashi Onoda*, and Seiji Yamada**

*Central Research Institute of Electric Power Industry (CRIEPI)
2-11-1 Iwado kita, Komae-shi, Tokyo 201-8511, Japan
E-mail: murata@criepi.denken.or.jp
**National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
[Received July 26, 2012; accepted December 20, 2012]

Support Vector Machines (SVMs) were applied to interactive document retrieval that uses active learning. In such a retrieval system, the degree of relevance is evaluated by using a signed distance from the optimal hyperplane. It is not clear, however, how the signed distance in SVMs has characteristics of vector space model. We therefore formulated the degree of relevance by using the signed distance in SVMs and comparatively analyzed it with a conventional Rocchio-based method. Although vector normalization has been utilized as preprocessing for document retrieval, few studies explained why vector normalization was effective. Based on our comparative analysis, we theoretically show the effectiveness of normalizing document vectors in SVM-based interactive document retrieval. We then propose a cosine kernel that is suitable for SVM-based interactive document retrieval. The effectiveness of the method was compared experimentally with conventional relevance feedback for Boolean, Term Frequency and Term Frequency-Inverse Document Frequency representations of document vectors. Experimental results for a Text REtrieval Conference data set showed that the cosine kernel is effective for all document representations, especially Term Frequency representation.

Keywords: interactive document retrieval, support vector machines, relevance feedback, kernel method

1. Introduction

The amount of text data is rapidly increasing with the development of information technology, and document retrieval is expected to become more sophisticated. The task of finding relevant documents is known as document retrieval. It can also be defined as a task for finding as many relevant documents as possible, even if there is a cognitive load imposed on the user. In interactive document retrieval including relevance feedback [1], users need to judge whether a document is relevant or irrelevant to their interest, and this judgment imposes a significant cognitive load on the user. This interactive document retrieval is the focus of our study. Document retrieval systems that use information from interactive user feedback have been studied in many ways [2, 3].

In most frameworks for information retrieval, a Vector Space Model (VSM) is used in which a document is described by using a high-dimensional vector [4]. An information retrieval system that uses a VSM computes the degree of relevance between a query vector and document vectors by using the cosine similarity of the two vectors. It then provides a list of retrieved documents to the user.

Since it is generally rare that a user describes a query precisely on the first attempt, an interactive approach has been proposed that modifies a query vector with a user's evaluation of documents in a list of retrieved documents. This method is called *relevance feedback* [1] and is used widely in information retrieval systems. With this method, a user directly evaluates whether a document in a list of retrieved documents is relevant, and the system modifies the query vector based on the user's evaluation. One conventional way to modify a query vector is through a simple learning rule that reduces the difference between the query vector and the documents evaluated as relevant by the user.

One relevance feedback method based on a VSM uses the Rocchio algorithm [1]. In this algorithm, new query vector Q_{m+1} is calculated using the following equation:

$$Q_{m+1} = Q_m + \beta \sum_{\mathbf{x} \in R_r^m} \mathbf{x} - \gamma \sum_{\mathbf{x} \in R_n^m} \mathbf{x}. \quad \ldots\ldots \quad (1)$$

Here, $\mathbf{x}$ represents document vectors and R_r^m and R_n^m are document sets that are determined as relevant and irrelevant, respectively, when feedback is obtained m times. β and γ are parameters that adjust the relative impact of relevance and irrelevance, respectively. The Rocchio algorithm evaluates the degree of relevance by using cosine similarity between new query vector Q_{m+1} and document vectors.

Another approach has been proposed in which relevant and irrelevant document vectors, respectively, are classified as positive and negative examples for a target concept based on classification learning [5]. Some studies have proposed that Support Vector Machines (SVMs) [6],

which have excellent ability for classifying input data into two classes, be applied to classification learning for relevance feedback. Drucker et al. applied SVMs to relevance feedback and confirmed its effectiveness in cases where the document database has few relevant documents [7]. They experimented with differences between document vectors such as the rate of change of relevant documents in the document database. In such a SVM system, the degree of relevance is evaluated by using a signed distance from the optimal hyperplane. It is not clear how the signed distance in SVMs has characteristics of VSM. We thus formulated the degree of relevance using the signed distance in SVMs and comparatively analyzed it with the Rocchio algorithm. Tong and Koller studied active learning for text classification based on SVMs [8]. They proposed a sample selection method for effectively cutting version space based on a margin and showed its effectiveness experimentally. They discussed only the application of text classification, however, and did not inspect the method from the viewpoint of interactive document retrieval.

Although vector normalization has been utilized as preprocessing for document retrieval [9], few studies have explained why vector normalization was effective. From the results of that analysis, we theoretically show the effectiveness of normalizing the document vector in SVM-based interactive document retrieval. We then propose a cosine kernel that expresses distance in SVMs by using cosine similarity. The cosine kernel is equivalent to a normalized linear kernel. A normalized linear kernel was proposed and its effectiveness evaluated experimentally by Hotta [10]. It was only derived, however, and no theoretical discussion was provided. In contrast, we derive the cosine kernel through comparative analysis of relevance evaluation and we show the reason for its effectiveness.

We also conducted various experiments on large document data sets with different document representations, such as Boolean, Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TFIDF). Results confirmed the effectiveness of our proposed method.

In Section 2 of this paper, we describe an information retrieval system that uses SVM-based relevance feedback. We propose the cosine kernel in Section 3. To evaluate the effectiveness of our approach, we present results of experiments performed with a Text REtrieval Conference (TREC) data set in Sections 4 and 5. We present conclusions in Section 6.

Basic ideas and contributions developed in this paper were published in [11].

2. Comparative Analysis of SVM-Based and Rocchio-Based Relevance Feedback

2.1. SVM-Based Relevance Feedback

The concept of relevance feedback based on SVMs is shown in **Fig. 1**. The white and black circles represent documents that have been labeled as relevant and irrelevant, respectively. Let us consider labeled document vectors $\mathbf{x}_i$ as having class labels $y_i = \pm 1$. A linear classifier for unlabeled document vector $\mathbf{x}$ is in the form

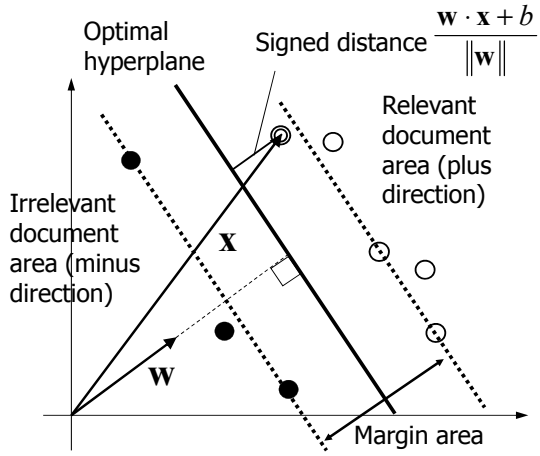

Fig. 1. Relevance feedback based on SVMs.

$$f(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x} + b. \quad \ldots\ldots\ldots\ldots (2)$$

In this framework, labeled document vectors and unlabeled document vectors correspond to document vectors judged as relevant or irrelevant and those not judged, respectively. The signed distance between labeled documents and the hyperplane (margin) is

$$\min_i \frac{\mathbf{w}\cdot\mathbf{x}_i + b}{||\mathbf{w}||}. \quad \ldots\ldots\ldots\ldots (3)$$

The minimum distance between labeled documents and the hyperplane is

$$\min_i \frac{|\mathbf{w}\cdot\mathbf{x}_i + b|}{||\mathbf{w}||}. \quad \ldots\ldots\ldots\ldots (4)$$

Here, we add constraint $\min_i |\mathbf{w}\cdot\mathbf{x}_i + b| = 1$. Eq. (4) can thus be rewritten as $1/||\mathbf{w}||$.

The hyperplane having the maximum margin is found by solving quadratic programming problem $\tau(\mathbf{w}) = ||\mathbf{w}||^2$, which is subject to inequality constraints $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1\ (i = 1,\ldots,\ell)$.

We construct a Lagrangian for solving the above equation as follows:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{\ell} \alpha_i (y_i((\mathbf{x}_i\cdot\mathbf{w}) + b) - 1) \quad (5)$$

where $\alpha_i \geq 0$ represents Lagrange multipliers. Minimization over $\mathbf{w}$ and b is achieved by the following differentiation:

$$\frac{\partial L}{\partial b} = 0, \quad \frac{\partial L}{\partial \mathbf{w}} = 0.$$

We therefore have the conditions

$$\sum_{i=1}^{\ell} \alpha_i y_i = 0 \quad \ldots\ldots\ldots\ldots (6)$$

$$\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \mathbf{x}_i. \quad \ldots\ldots\ldots\ldots (7)$$

By using the previous results, we obtain

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j) \quad \ldots \quad (8)$$

subject to

$$\alpha_i \geq 0 \, (i = 1, \ldots, \ell), \quad \sum_{i=1}^{\ell} \alpha_i y_i = 0. \quad \ldots \quad (9)$$

Document vectors $\mathbf{x}_i$ at $\alpha_i = 0$ have no affect on optimizing the hyperplane. Only document vectors $\mathbf{x}_i$ that are shown by circles on dashed lines in **Fig. 1** decide the optimal hyperplane. These documents are obtained under condition $\alpha > 0$. These $\alpha_i > 0$ data are called "support vectors." The distance between the document and the optimal hyperplane is defined as the degree of relevance in SVM-based relevance feedback.

2.2. Comparative Analysis of Relevance Feedback

The signed distance between the optimal hyperplane and an unlabeled document vector, which is shown by a double circle in **Fig. 1**, is expressed as

$$\frac{\mathbf{w} \cdot \mathbf{x} + b}{||\mathbf{w}||} = \frac{||\mathbf{w}||||\mathbf{x}|| \cos\theta_w + b}{||\mathbf{w}||} = ||\mathbf{x}|| \cos\theta_w + \frac{b}{||\mathbf{w}||} \quad \ldots \quad (10)$$

where θ_w is the angle between $\mathbf{w}$ and $\mathbf{x}$.

Labels for relevant and irrelevant documents are thus $y_i = 1$ and $y_i = -1$, respectively, and

$$\mathbf{w} = \sum_j \alpha_j \mathbf{x}_j - \sum_k \alpha_k \mathbf{x}_k \quad \ldots \quad (11)$$

where j and k are indices of relevant and irrelevant documents, respectively.

The query update equation of the Rocchio-based method, as shown in Eq. (1), is transformed as follows:

$$Q_{m+1} = Q_0 + \sum_j \beta \mathbf{x}_j - \sum_k \gamma \mathbf{x}_k \quad \ldots \quad (12)$$

where Q_0 is an initial query vector that is derived from a user's initial input query.

A comparison of Eqs. (11) and (12) suggests that the vector $\mathbf{w}$ equation of SVM-based relevance feedback is equivalent to the query update equation of the Rocchio-based method when the initial query vector is a zero vector.

The degree of relevance determines the documents to be initially retrieved. The cosine similarity of the initial query is used in a SVM-based relevance feedback and the Rocchio-based method for determining the degree of relevance. Initially retrieved documents therefore contain words in the initial query and there is negligible difference between Eqs. (11) and (12).

Moreover, vector $\mathbf{w}$ in Eq. (11) does not include documents corresponding to nonsupport vectors with $\alpha_i = 0$, and the update in Eq. (12) includes all document vectors containing nonsupport vectors. This shows that the SVM-based method weights all documents including $\alpha_i = 0$, and the Rocchio-based method weights relevant and irrelevant documents with constant values β and γ, respectively.

The above discussion suggests that the equation of vector $\mathbf{w}$ for SVM-based relevance feedback is equivalent to the query update equation of Rocchio-based relevance feedback, and vector $\mathbf{w}$ is very similar to the updated query vector of Rocchio-based relevance feedback.

Furthermore, in Eq. (11), the SVM-based method decides the weights of individual document vectors as α_i. This weight α_i is calculated to minimize structural risk under zero empirical risk. In contrast, in the Rocchio-based method, individual document vectors have the same weights as β and γ. These parameters are labeled through trial and error.

In one study of text classification applying a Rocchio-based method, document weights were automatically decided by maximizing the break-even point at which recall is equal to precision [12]. This method needs training documents and evaluation documents, so applying it to relevance feedback is difficult.

A study was also conducted to automatically determine document weights for relevance feedback [13]. The study proposed a method to determine feedback weight α for balancing queries and feedback documents in a Kullback-Leibler divergence retrieval model by machine learning. In this study, weight α was determined through logistic regression by using heuristically selected features. This method needs training queries, however, and α changes with these queries. The SVM-based method automatically decides weights by using feedback results. This method is therefore more feasible.

2.3. Improvement of Document Representation Based on Comparative Analysis

In the previous section, we showed that $\mathbf{w}$ of relevance feedback based on SVMs is equivalent to the query update equation of the Rocchio-based method. The degree of relevance in SVM-based method is evaluated as $||\mathbf{x}|| \cos\theta_w$ from Eq. (10) because unique values of $\mathbf{w}$ and b are determined from labeled document vectors. The degree of relevance in the Rocchio-based method is evaluated, however, as $\cos\theta_q$, which denotes the cosine similarity of angle θ_q between the document vector and the query vector.

This comparison suggests that the degree of relevance in the SVM-based method increases with target document vector $||\mathbf{x}||$. This means that the degree of relevance increases with a large $||\mathbf{x}||$ rather than a small θ_w. A document that includes many words therefore has high relevance, and a document that includes fewer words has low relevance. To avoid this problem, we define the degree of relevance in the SVM-based method as genuine cosine similarity $\cos\theta_w$. A simple method for achieving this is to normalize a document vector.

3. Cosine Kernel Based on Comparative Analysis

We substituted kernel K for the dot product in Eq. (8). The kernel corresponds to a dot product in a feature space that is related to input space via nonlinear map Φ, which can have high dimensionality:

$$K(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}'). \quad \ldots\ldots\ldots (13)$$

By using K, the classifier takes the form (cf. Eqs. (2) and (7)):

$$\begin{aligned} f(\Phi(\mathbf{x})) &= \mathbf{w} \cdot \Phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{\ell} \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b \\ &= \sum_{i=1}^{\ell} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b. \quad \ldots\ldots (14) \end{aligned}$$

The dual optimization problem of Eq. (8) is rewritten to maximize

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j), \quad \ldots (15)$$

with the same constraints.

We consider that kernel $K(\mathbf{x},\mathbf{x}')$ has cosine similarity when the angle between the two vectors $\mathbf{x}$ and $\mathbf{x}'$ is θ such that

$$K(\mathbf{x},\mathbf{x}') = \cos\theta = \frac{\mathbf{x} \cdot \mathbf{x}'}{||\mathbf{x}||||\mathbf{x}'||}. \quad \ldots\ldots (16)$$

It can be seen from this equation that using cosine similarity as a kernel is equivalent to normalizing the document vector. We call this the "cosine kernel."

4. Experimental Evaluation

4.1. Experimental Setting

We proposed a cosine kernel that expresses distance in the SVMs by using cosine similarity. As seen from the comparative analysis developed in Sections 2 and 3, the cosine kernel theoretically has a function similar to normalizing document vectors in advance. Since the effectiveness of the cosine kernel had not been investigated, however, we needed to evaluate it through experiments with large data sets. This is the objective of these experiments.

We conducted various experiments on large document data sets and several document vector representations to examine the effect of the cosine kernel. Note that the purpose of these experiments is to experimentally investigate the effectiveness of our proposed method for various document vector representations rather than showing that our method outperforms state-of-the-art methods.

The document data set we used was a set of articles in an ad hoc task, which was widely used in the 6th, 7th and 8th Text REtrieval Conferences (TREC). The data set has about 530,000 newspaper articles. Each TREC provides 50 retrieval problems and information on relevant documents for each retrieval problem. Hereafter, we call the retrieval problem a "topic." In these experiments, 150 topics were tested. Each topic had three tags: a title tag, a description tag, and a narrative tag. The title tag had two or three terms to describe the topic. The description tag introduced the topic. The narrative tag reported the topic. An example of a topic is shown in **Fig. 2**. Our experiments used two or three terms from the title tag as a query. Experiments also removed stopwords and created stemming for documents and queries.

```
Number       301
Title        International Organized Crime
Description  Identify organizations that
             participate in international
             criminal activity, the
             activity, and, if possible,
             collaborating organizations
             and the countries involved.
Narrative    A relevant document must
             as a minimum identify the
             organization and the type
             of illegal activity (e.g.,
             Columbian cartel exporting
             cocaine). Vague references
             to international drug trade
             without identification of the
             organization(s) involved would
             not be relevant.
```

Fig. 2. Example of TREC topics.

We compared differences in document vector representations that were due to dependence on classification performance. We used Boolean, TF, and TFIDF methods to represent a document vector. The TFIDF used was

$$w(t,d) = \frac{\log(\mathrm{tf}(t,d)+1)}{\log(\mathrm{uniq}(d))} \log\frac{N}{\mathrm{df}(t)}. \quad \ldots (17)$$

In this equation,

- $w(t,d)$ is the weight of term t in document d.
- $\mathrm{tf}(t,d)$ is the frequency of term t in document d.
- N is the total number of documents in a data set.
- $\mathrm{df}(t)$ is the number of documents including term t.
- $\mathrm{uniq}(d)$ is the number of different terms in document d.

One document vector has about 760,000 dimensions.

The SVM algorithm was implemented by using LibSVM software. The selection rule deciding which documents to display was "all documents are mapped onto feature space." The learned SVM classified the documents as relevant or irrelevant. Displayed documents were selected from the relevant area of SVMs. Top S-ranked documents, which were ranked by using the distance from the optimal

hyperplane, were displayed to a user as results of the system's information retrieval.

To compare the *retrieval performance* of our proposed method with that of other methods, we evaluated the following criterion:

P: Precision of all displayed documents, where

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

Here, N_{rel} denotes the number of relevant documents among all displayed documents and N_{dis} denotes the total number of displayed documents.

To compare the *learning performance* of our proposed method with that of other methods, we evaluated the following criterion:

P30: Precision within the top 30 documents, which is the proportion of relevant documents in the top 30 documents [13].

The performance of relevance feedback changes with the number of evaluated documents [14]. It was thus necessary to change the number of evaluated documents and the number of feedback iterations in the experiment. The size S of retrieved and displayed documents at each iteration was set at 10 or 20. Feedback iterations M were decided according to the total number of displayed documents. In these experiments, we set the total number of displayed documents at 100, which includes initial search documents. When size S of retrieved and displayed documents at each iteration was 10, feedback iterations M were between 1 and 9. When S was 20, feedback iterations M were between 1 and 4.

4.2. Experimental Results

Experimental results for P are shown in **Figs. 3** and **4**, and experimental results for P30 are shown in **Figs. 5** and **6**. Vertical axes show P or P30, respectively, and horizontal axes show the number of feedback iterations. In these figures, 0 as the number of feedback iterations indicates the performance of initial retrieval, bold lines indicate results of the cosine kernel, and solid lines indicate results of a linear SVM.

These figures suggest that the cosine kernel showed better performance, especially when used with the TF method (comparison of lines with crosses in figures).

We furthermore compared the proposed method with the Rocchio-based method that applies normalized document vectors. Parameters of the Rocchio algorithm were $\beta = 1.0, \gamma = 0.5$ in a previous work that compared performances of SVM-based and Rocchio-based methods [15]. Experimental results for P are shown in **Figs. 7** and **8** and the experimental results for P30 are shown in **Figs. 9** and **10**. These figures are results for TFIDF, which is the most popular document representation. Retrieval performance P of the SVM was better than that of Rocchio when the number of feedback documents was over 30 ($M = 3$ or more at $S = 10$ in **Fig. 3**, and $M = 2$ or more at $S = 20$ in **Fig. 4**). Similarly, learning performance P30 of SVM was better than that of Rocchio when $M = 2$ or more at $S = 10$ and $S = 20$ in **Figs. 5** and **6**, respectively.

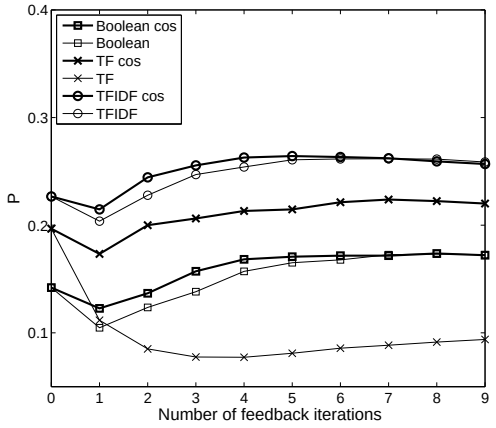

Fig. 3. Results of retrieval performance with criterion P (displayed documents $S = 10$).

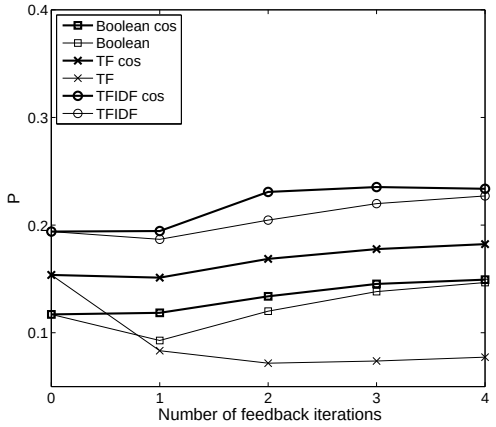

Fig. 4. Results of retrieval performance with criterion P (displayed documents $S = 20$).

5. Discussion

5.1. Effectiveness of Cosine Kernel

The degree of relevance in a SVM-based method is evaluated as $||\mathbf{x}||\cos\theta_w$ from Eq. (10). The degree of relevance in the Rocchio-based method, however, is evaluated as $\cos\theta_q$, which denotes the cosine similarity of angle θ_q between the document vector and the query vector. This comparison suggests that the degree of relevance in the SVM-based method increases with the length of document vector $||\mathbf{x}||$. This means that the degree of relevance increases with a large $||\mathbf{x}||$ rather than a small θ_w. To avoid this problem, we proposed the cosine kernel, which

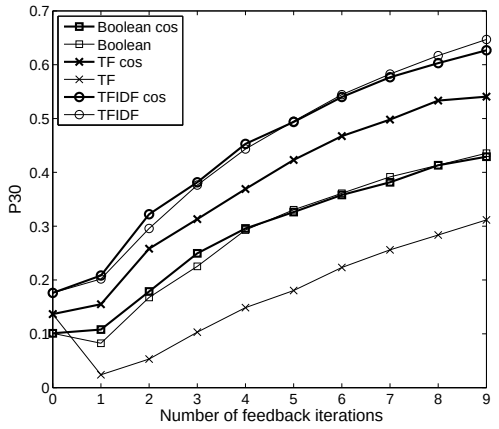

Fig. 5. Results of learning performance with criterion P30 (displayed documents $S = 10$).

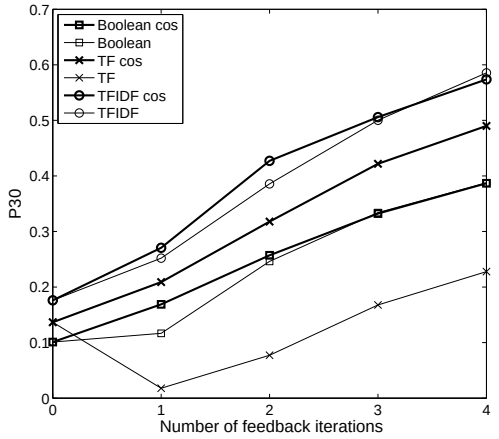

Fig. 6. Results of learning performance with criterion P30 (displayed documents $S = 20$).

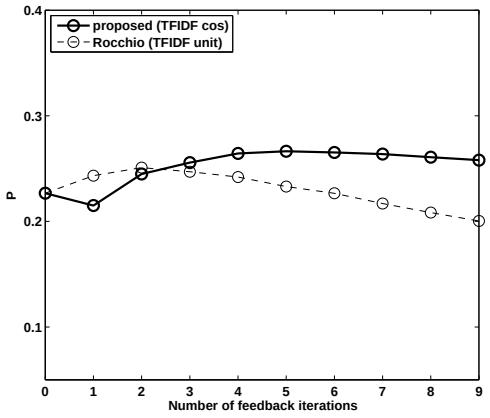

Fig. 7. Comparison of results of retrieval performance with criterion P between proposed method and Rocchio-based method (displayed documents $S = 10$).

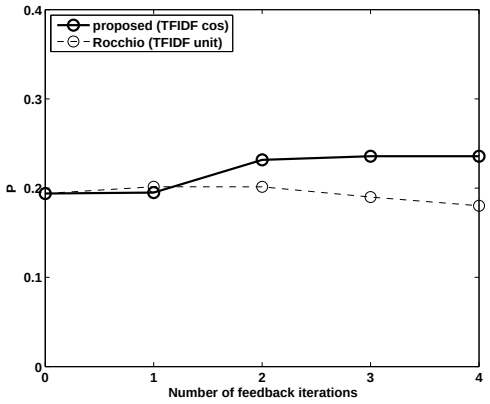

Fig. 8. Comparison of results of retrieval performance with criterion P between proposed method and Rocchio-based method (displayed documents $S = 20$).

is equivalent to normalizing the document vector. **Figs. 3** to **6** suggest that TF representation greatly improves document retrieval. TF representation is strongly affected by vector length. A comparison of displayed vector lengths $||\mathbf{x}||$ with the SVM-based method when $m = 1, N = 10$ is shown in **Table 1**. The TF representation has some very large vectors compared with other representations. We consider these large vectors to have a negative effect on retrieval accuracy. The cosine kernel evaluates the angle between vectors without length, so the cosine kernel improves retrieval performance.

TFIDF is the most popular representation of document vectors in VSM. Calculating TFIDF requires the frequency of a term in a document (TF) and the number of documents that include a term (DF). TF can be calculated from one document. The calculation of DF, however, assumes that a whole document set is known. This assumption required for DF calculation is not satisfied in the retrieval of Web pages and microblogs [16, 17]. A whole document set is rarely obtained in such document retrieval. TF is therefore an important representation of documents, especially for document retrieval on the Web.

In contrast, Boolean and TFIDF approaches are effective in the early stage of feedback iteration, although the improvement in results gradually decreases as iteration progresses. For this reason, the display order is not changed significantly by $||\mathbf{x}||$ because the classification accuracy of the discriminant hyperplane with SVMs gradually increases due to an iterative feedback process and $||\mathbf{x}||$ does not change greatly, as **Table 1** shows.

The high performance in the early stage of feedback iteration suggests that a small number of user evaluations is useful. The evaluation of documents in relevance feedback is a task with a high cognitive load for the user. The proposed method, which obtains high retrieval performance by using a small number of evaluations, therefore has significant advantages in usability and practical use.

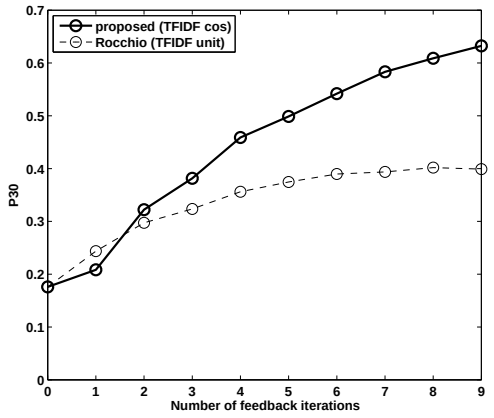

Fig. 9. Comparison of results of learning performance with criterion P30 between proposed method and Rocchio-based method (displayed documents $S = 10$).

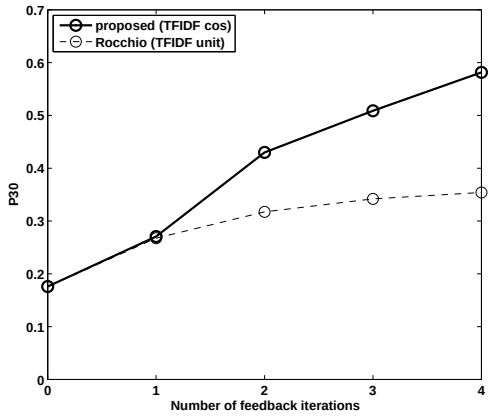

Fig. 10. Comparison of results of learning performance with criterion P30 between proposed method and Rocchio-based method (displayed documents $S = 20$).

5.2. Performance Comparison Between Proposed Method and Rocchio Method

In order to verify the effectiveness of the proposed method, we compared the proposed method with a Rocchio-based method that applies normalized document vectors.

Figures 7 and **8** show that the proposed method had better retrieval performance than the Rocchio-based method when feedback was iterated. **Figs. 9** and **10** show learning performance trends. The cause of low performance for the first feedback iteration is that active learning did not appear at that point because the system did not select training data actively.

As **Figs. 7** and **8** show, the decay rate of the precision of the Rocchio-based method was larger than that of the proposed method. Drucker et al. also reported that the precision of the Rocchio-based method decreased as feedback was iterated [7]. A similar effect appeared in our experiments.

Table 1. Comparison of displayed vector lengths $||\mathbf{x}||$ with the SVM-based method when $m = 1, N = 10$.

	mean	median	min	max	SD
Boolean	23.0	16.1	2.6	118.7	20.1
TF	248.4	77.7	3.5	12258.6	724.2
TFIDF	29.1	16.2	3.2	228.4	32.9

SD: Standard Deviation

6. Conclusions

The degree of relevance in an SVM-based retrieval system was evaluated by using the signed distance from the optimal hyperplane. It was not clear, however, how the signed distance in SVMs had characteristics of vector space model. In this paper, we formulated the degree of relevance by using the signed distance to SVMs and analyzed it comparatively with a conventional Rocchio-based method. We showed that the degree of relevance is expressed as $||\mathbf{x}|| \cos\theta_w$, where θ_w was the angle between document vector $\mathbf{x}$ and vector $\mathbf{w}$. The equation of vector $\mathbf{w}$ for SVM-based relevance feedback was equivalent to the query update equation of Rocchio-based relevance feedback, and vector $\mathbf{w}$ was very similar to the updated query vector of Rocchio-based relevance feedback.

We then proposed a cosine kernel that denotes cosine similarity, suitable for SVM-based interactive document retrieval, from comparative analysis. Since the effectiveness of the cosine kernel had not been investigated, however, we needed to evaluate it through experiments with large data sets. We thus conducted various experiments on large TREC data sets to examine the effect of the cosine kernel. We compared differences in document vector representation that were due to dependence on classification performance. We used Boolean, TF, and TFIDF methods to represent a document vector. In particular for TF representation, our proposed approach was demonstrated experimentally to show better performance.

We furthermore compared the proposed method with a conventional Rocchio-based method that applies normalized document vectors. The proposed method had better retrieval performance than the Rocchio-based method when feedback was iterated.

In future work, we will study extending our comparative analysis to interactive information retrieval with other classification learning excluding SVMs.

References:

[1] G. Salton, (Ed.), "The SMART Retrieval System – Experiments in Automatic Document Processing," Prentice Hall, Englewood, Cliffs, New Jersey, 1971.
[2] P. Ingwersen, "Information Retrieval Interaction," Taylor Graham, 1992.
[3] J. Koenemann and N. J. Belkin, "A case for interaction: a study of interactive information retrieval behavior and effectiveness," In Proc. of 27th Annual SIGCHI Conf. on Human factors in Computing Systems, pp. 205-212, 1996.
[4] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

[5] M. Okabe and S. Yamada, "Learning filtering rulesets for ranking refinement in relevance feedback," Knowledge-Based Systems, Vol.18, pp. 117-124, April 2005.
[6] V. Vapnik, "Statistical Learning Theory," John Wiley and Sons Inc., 1998.
[7] H. Drucker, B. Shahrary, and D. C. Gibbon, "Support vector machines: relevance feedback and information retrieval," Information Processing & Management, Vol.38, pp. 305-323, May 2002.
[8] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," J. of Machine Learning Research, Vol.2, pp. 45-66, 2002.
[9] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," In Information Processing and Management, pp. 513-523, 1988.
[10] K. Hotta, "Local normalized linear summation kernel for fast and robust recognition," Pattern Recognition, Vol.43, pp. 906-913, March 2010.
[11] H. Murata, T. Onoda, and S. Yamada, "Comparative Analysis of Relevance Evaluation for Interactive Document Retrieval Based on SVMs (in Japanese)," J. of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol.23, No.6, pp. 853-862, 2011.
[12] A. Moschitti, "A Study on Optimal Parameter Tuning for Rocchio Text Classifier," In Proc. of the 25th European Conf. on Information Retrieval Research (ECIR '03), pp. 420-435, 2003.
[13] Y. Lv and C. Zhai, "Adaptive Relevance Feedback in Information Retrieval," In Proc. of the 18th ACM Conf. on Int. Knowledge Management, pp. 255-264, 2009.
[14] J. Montgomery, L. Si, J. Callan, and D. A. Evans, "Effect of varying number of documents in blind feedback: analysis of the 2003 NRRC RIA workshop " bf numdocs "experiment suite," In Proc. of 27th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 476-477, 2004.
[15] T. Onoda, H. Murata, and S. Yamada, "SVM-based interactive document retrieval with active learning," New Generation Computing, Vol.26, pp. 49-61, November 2007.
[16] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst, and A. C. König, "BLEWS: Using Blogs to Provide Context for News Articles," In Proc. of Int. Conf. on Weblogs and Social Media, 2008.
[17] M. Klein and M. L. Nelson, "Correlation of Term Count and Document Frequency for Google N-Grams," In Proc. of the 31th European Conf. on IR Research on Advances in Information Retrieval (ECIR '09), pp. 620-627, 2009.

Name:
Hiroshi Murata

Affiliation:
Senior Research Scientist, Central Research Institute of Electric Power Industry

Address:
2-11-1 Iwado kita, Komae-shi, Tokyo 201-8511, Japan

Brief Biographical History:
1993- Researcher, Central Research Institute of Electric Power Industry
2012- Senior Research Scientist, Central Research Institute of Electric Power Industry

Main Works:
• H. Murata and T. Onoda, "Applying Kernel Based Subspace Classification to Non-intrusive Monitoring for Household Electric Appliances," Artificial Neural Networks – ICANN 2001, pp. 692-698, 2001.

Membership in Academic Societies:
• The Japanese Society for Artificial Intelligence (JSAI)
• The Institute of Electrical Engineers of Japan (IEEJ)

Name:
Takashi Onoda

Affiliation:
Central Research Institute of Electric Power Industry, Tokyo Institute of Technology

Address:
2-1-1 Iwado Kita, Komae-shi, Tokyo 201-8511, Japan

Brief Biographical History:
1988- Researcher, Central Research Institute of Electric Power Industry
1997-1998 Visiting Researcher, GMD FIRST
2003- Senior Research Scientist, Central Research Institute of Electric Power Industry
2007- Visiting Professor, Tokyo Institute of Technology
2012- Deputy Associate Vice President, Central Research Institute of Electric Power Industry

Main Works:
• G. Raetsch, T. Onoda, and K.-R. Mueller, "Soft margins for AdaBoost," Machine Learning, Vol.42, pp. 287-320 2001.

Membership in Academic Societies:
• The Japanese Society for Artificial Intelligence (JSAI)

Name:
Seiji Yamada

Affiliation:
National Institute of Informatics, Sokendai, Tokyo Institute of Technology

Address:
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan

Brief Biographical History:
1989- Research Associate, Osaka University
1991- Lecturer, Osaka University
1996- Associate Professor, Tokyo Institute of Technology
2002- Professor, National Institute of Informatics

Main Works:
• K. Kobayashi, S. Yamada, S. Nakagawa, and Y. Saito, "Rebo: A Pet-like Strokable Remote Control," J. of Advanced Computational Intelligence and Intelligent Informatics, Vol.16, No.7, pp. 771-783, 2012.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE)
• Association for the Advancement of Artificial Intelligence (AAAI)
• Association for Computing Machinery (ACM)
• The Japanese Society for Artificial Intelligence (JSAI)
• Information Processing Society of Japan (IPSJ)
• Japan Human Interface Society (HIS)