Improvement in TF-IDF scheme for Web Pages and its Retrieval Accuracy

Kazunari SUGIYAMA[♡] Kenji HATANO[◇]
Masatoshi YOSHIKAWA[♠] Shunsuke UEMURA[◇]

In IR (information retrieval) systems based on the vector space model, the tf-idf scheme is widely used to characterize documents. However, in the case of documents with hyperlink structures such as Web pages, it is necessary to develop a technique for representing the contents of Web pages more accurately by exploiting that of their hyperlinked neighboring pages. In this paper, we first propose some methods for improving the tf-idf scheme for a target Web page by using the contents of its hyperlinked neighboring pages, and then compare retrieval accuracy of our proposed methods.

1. Introduction

The WWW (World Wide Web) is a useful resource for users to obtain a great variety of information. However, since the number of Web pages continues to grow, it is getting more and more difficult for users to find relevant information on the WWW. Under these circumstances, search engines are one of the most popular methods for finding valuable information effectively. Recently, in order to obtain more higher retrieval accuracy, the hyperlink structures of Web pages are taken into account in IR systems. For example, IR systems based on the concept of "Optimal Document Granularity" using the hyperlink structures of Web pages [1, 2, 3] are proposed. However, as for these works, we do not believe that users could understand the search results intuitively because the multiple query keywords disperse in several hyperlinked Web pages. In addition, although HITS (Hypertext Induced Topic Search) [4] and PageRank [5] achives higher retrieval accuracy using the hyperlink structures of Web pages, these algorithms have shortcomings in that (1) the weight for a Web page is merely defined; and (2) the relativity of contents among hyperlinked Web pages is not considered. As a result, the problem of Web pages irrelevant to user's query often being ranked higher still remains. Therefore, in order to provide users with relevant Web pages, it is necessary to develop a technique for representing the contents of Web pages more accurately. To achieve this purpose, we have proposed some methods for improving a feature vector for target Web page [6]. Our proposed methods, however, also have a problem in that only Web pages out-linked from a target Web page are exploited to generate the feature vector of target Web page. Therefore, in this paper, we propose three methods for improving the tf-idf scheme [7] for a target Web page using both its in- and out-linked pages in order to represent the contents of the target Web page more accurately. Our method is novel in improving tf-idf based feature vector of target Web page by reflecting the contents of its hyperlinked neighboring Web pages.

2. Proposed Method

On the basis of the problems of IR systems described in Section 1, in order to represent the contents of Web pages more accurately, the feature vector of a Web page should be generated by using the contents of its hyperlinked neighboring pages. We, therefore, propose improving the tf-idf scheme for a target Web page by using the contents of its hyperlinked neighboring pages. Unlike the works described in Section 1, our method is novel in improving tf-idf based feature vector of a target Web page by reflecting the contents of its hyperlinked neighboring Web pages.

In the following discussion, let a target page be p_{tgt}. Then, we define i as the number of the shortest directed path from p_{tgt}. Let us assume that there are N_i Web pages $(p_{i_1}, p_{i_2}, \cdots, p_{i_{N_i}})$ in the i^{th} level from p_{tgt}. Moreover, we denote the feature vector $\boldsymbol{w}^{p_{tgt}}$ of p_{tgt} as follows:

$$\boldsymbol{w}^{p_{tgt}} = (w_{t_1}^{p_{tgt}}, w_{t_2}^{p_{tgt}}, \cdots, w_{t_m}^{p_{tgt}}), \tag{1}$$

where m is the number of unique terms in the Web page collection, and $t_k (k = 1, 2, \cdots, m)$ denotes the each term. Using the tf-idf scheme, we also define the each element $w_{t_k}^{p_{tgt}}$ of $\boldsymbol{w}^{p_{tgt}}$ as follows:

$$w_{t_k}^{p_{tgt}} = \frac{tf(t_k, p_{tgt})}{\sum_{s=1}^{m} tf(t_s, p_{tgt})} \cdot \log \frac{N_{web}}{df(t_k)}, \tag{2}$$

where $tf(t_k, p_{tgt})$ is the frequency of term t_k in the target page p_{tgt}, N_{web} is the total number of Web pages in the collection, and $df(t_k)$ is the number of Web pages in which term t_k appears. Below, we refer to $\boldsymbol{w}^{p_{tgt}}$ as the "*initial feature vector*." Subsequently, we denote the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ as follows:

$$\boldsymbol{w}'^{p_{tgt}} = (w_{t_1}'^{p_{tgt}}, w_{t_2}'^{p_{tgt}}, \cdots, w_{t_m}'^{p_{tgt}}), \tag{3}$$

and refer to this $\boldsymbol{w}'^{p_{tgt}}$ as the "*improved feature vector*." As we describe below, we propose three methods for improving the "initial feature vector" based on the tf-idf scheme defined by Equation (2).

Method I

In this method, we reflect the contents of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction and levels up to ${L_{(out)}}^{th}$ in the forward direction from the target page p_{tgt}. Based on the ideas that (1) there are Web pages similar to the contents of p_{tgt} in the neighborhood of p_{tgt}; and (2) since on one hand such Web pages exist right near p_{tgt}, on the other hand they might exist far removed from p_{tgt} in the vector space, we reflect the distance between $\boldsymbol{w}^{p_{tgt}}$ and feature vector of in- and out-linked pages of p_{tgt} in the vector space on each element of initial feature vector $\boldsymbol{w}^{p_{tgt}}$. For example, Figure 1(a) shows that $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting the contents of all Web pages at levels up to second in the backward and forward directions from p_{tgt} on $\boldsymbol{w}^{p_{tgt}}$. In Figure 1(a), $p_{ij_{(in)}}$ and $p_{ij_{(out)}}$ correspond to the j^{th} page in the i^{th} level in the backward and forward directions from p_{tgt}, respectively. Additionally, Figure 1(b) shows that improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting each feature vector of in- and out-linked pages of p_{tgt} on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In this method, each element $w_{t_k}'^{p_{tgt}}$ of $\boldsymbol{w}'^{p_{tgt}}$ is defined as follows:

$$\begin{aligned} w_{t_k}'^{p_{tgt}} &= w_{t_k}^{p_{tgt}} \\ &+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}} \sum_{j=1}^{N_{i(in)}} \frac{w_{t_k}^{p_{ij(in)}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(in)}})}\right) \\ &+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}} \sum_{j=1}^{N_{i(out)}} \frac{w_{t_k}^{p_{ij(out)}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(out)}})}\right). \end{aligned} \tag{4}$$

If the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{p_{ij(in)}}, \boldsymbol{w}^{p_{ij(out)}}$ in the vector space is very close, the values of the second and third terms of Equation (4) can be dominant compared with the first term $w_{t_k}^{p_{tgt}}$. Therefore, in order to prevent this phenomenon, we also define Dim, which denotes the number of unique terms in the Web page collection. $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(in)}})$ and $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(out)}})$ are defined the following equations, respectively:

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(in)}}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_{tgt}} - w_{t_k}^{p_{ij(in)}})^2},$$

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij(out)}}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_{tgt}} - w_{t_k}^{p_{ij(out)}})^2}.$$

[♡] Student Member Graduate School of Information Science, Nara Institute of Science and Technology
kazuna-s@is.aist-nara.ac.jp

[◇] Regular Member Graduate School of Information Science, Nara Institute of Science and Technology
{hatano, uemura}@is.aist-nara.ac.jp

[♠] Director Information Technology Center, Nagoya University
yosikawa@itc.nagoya-u.ac.jp

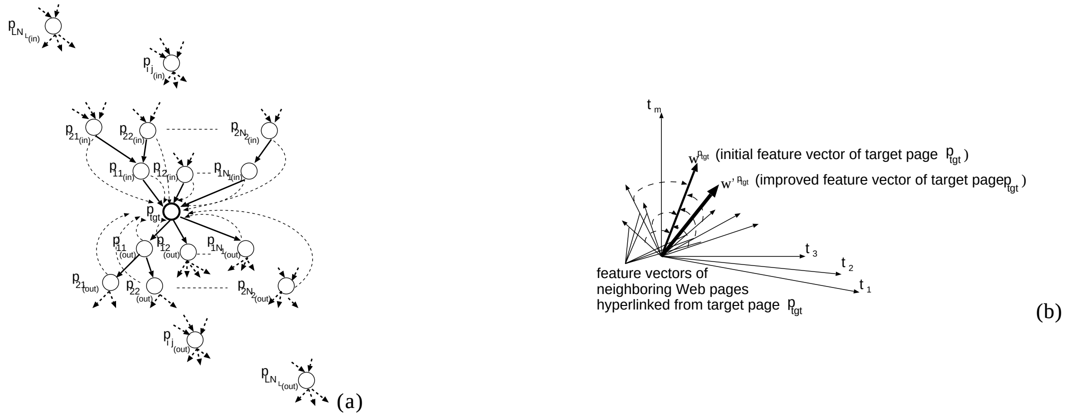

Fig.1 The improvement of a feature vector as performed by Method I [(a) in the Web space, (b) in the vector space].

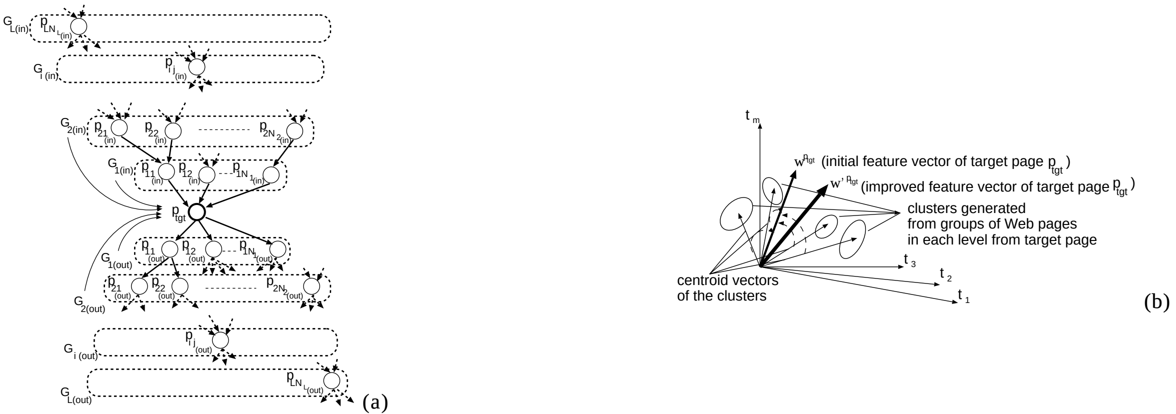

Fig.2 The improvement of a feature vector as performed by Method II [(a) in the Web space, (b) in the vector space].

Method II

In this method, we first construct Web page groups $G_{i_{(in)}}$ at each level up to $L_{(in)}{}^{th}$ in the backward direction, and $G_{i_{(out)}}$ at each level up to $L_{(out)}{}^{th}$ in the forward direction from the target page p_{tgt}. Then, we generate $\boldsymbol{w'}^{p_{tgt}}$ by reflecting centroid vectors of clusters generated from $G_{i_{(in)}}$ and $G_{i_{(out)}}$ on initial feature vector $\boldsymbol{w}^{p_{tgt}}$. This method is based on the idea that Web pages at each level in the backward and forward directions from p_{tgt} is classified into some topics in the each level. In addition, we reflect the distance between $\boldsymbol{w}^{p_{tgt}}$ and the centroid vectors of the clusters in the vector space on each element of the initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In other words, we first create Web page groups $G_{i_{(in)}}$ and $G_{i_{(out)}}$ defined as follows:

$$G_{i_{(in)}} = \{p_{i1_{(in)}}, p_{i2_{(in)}}, \cdots, p_{iN_{i_{(in)}}}\}, \tag{5}$$

$$G_{i_{(out)}} = \{p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}}\}, \tag{6}$$

$$(i = 1, 2, \cdots, L),$$

and then produce K clusters in each Web page group $G_{i_{(in)}}$ and $G_{i_{(out)}}$ by means of the K-means algorithm [9]. The centroid vectors $\boldsymbol{w}^{g_{ic_{(in)}}}$ and $\boldsymbol{w}^{g_{ic_{(out)}}}$ $(c = 1, 2, \cdots, K)$ are produced in $G_{i_{(in)}}$ and $G_{i_{(out)}}$, respectively. We generate an improved feature vector $\boldsymbol{w'}^{p_{tgt}}$ by reflecting the distance between each centroid vector, $\boldsymbol{w}^{g_{ic_{(in)}}}$, $\boldsymbol{w}^{g_{ic_{(out)}}}$ and the initial feature vector $\boldsymbol{w}^{p_{tgt}}$ on $\boldsymbol{w}^{p_{tgt}}$. For instance, Figure 2(a) shows that we construct Web page groups $G_{1_{(in)}}, G_{2_{(in)}}$, $G_{1_{(out)}}$, and $G_{2_{(out)}}$ at each level up to second in the backward and forward direction from p_{tgt}, and generate an improved feature vector $\boldsymbol{w'}^{p_{tgt}}$ by reflecting the centroid vectors of each cluster produced in each Web page group $G_{1_{(in)}}, G_{2_{(in)}}$, $G_{1_{(out)}}$, and $G_{2_{(out)}}$ on $\boldsymbol{w}^{p_{tgt}}$. Moreover, Figure 2(b) shows that improved feature vector $\boldsymbol{w'}^{p_{tgt}}$ is generated by reflecting the centroid vectors of each cluster on $\boldsymbol{w}^{p_{tgt}}$. In this method, we define each element $w'^{p_{tgt}}_{t_k}$ of $\boldsymbol{w'}^{p_{tgt}}$ as follows:

$$\begin{aligned} w'^{p_{tgt}}_{t_k} &= w^{p_{tgt}}_{t_k} \\ &+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}}\sum_{c=1}^{K}\frac{w^{g_{ic_{(in)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})}\right) \\ &+ \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}}\sum_{c=1}^{K}\frac{w^{g_{ic_{(out)}}}_{t_k}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})}\right). \end{aligned} \tag{7}$$

We introduce Dim for the purpose of preventing the values of the second and third terms from dominating compared with the first term in Equation (7). $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})$ and $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})$ are defined as follows:

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}}) = \sqrt{\sum_{k=1}^{m}(w^{p_{tgt}}_{t_k} - w^{g_{ic_{(in)}}}_{t_k})^2},$$

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}}) = \sqrt{\sum_{k=1}^{m}(w^{p_{tgt}}_{t_k} - w^{g_{ic_{(out)}}}_{t_k})^2}.$$

Method III

This method is based on the idea that Web pages at levels up to $L_{(in)}{}^{th}$ in the backward direction and levels up to $L_{(out)}{}^{th}$ in the forward direction from the target page p_{tgt} is composed of some topics. According to this idea, we cluster the set of all Web pages at levels up to $L_{(in)}{}^{th}$ in the backward direction and levels up to $L_{(out)}{}^{th}$ in the forward direction from p_{tgt}, and generate $\boldsymbol{w'}^{p_{tgt}}$ by reflecting centroid vectors of the clusters on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$. Furthermore, we reflect the distance between $\boldsymbol{w}^{p_{tgt}}$ and the centroid vector of the cluster in the vector space on each element of $\boldsymbol{w}^{p_{tgt}}$; in other words, we create Web page groups $G_{i_{(in)}}$ and $G_{i_{(out)}}$ as defined by Equation (8) and (9), respectively,

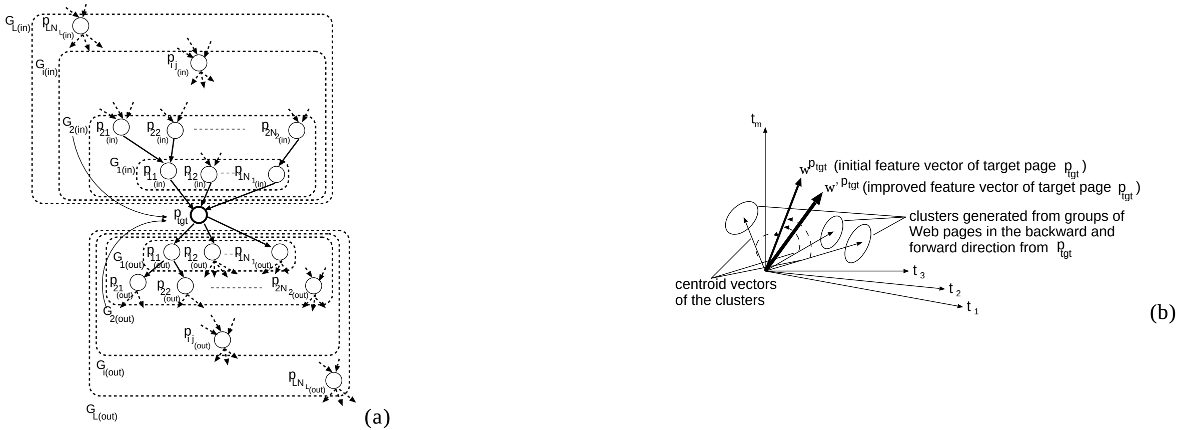

Fig.3 The improvement of a feature vector as performed by Method III [(a) in the Web space, (b) in the vector space].

$$
\begin{aligned}
G_{i_{(in)}} = \{ & p_{11_{(in)}}, p_{12_{(in)}}, \cdots, p_{1N_{1_{(in)}}}, \\
& p_{21_{(in)}}, p_{22_{(in)}}, \cdots, p_{2N_{2_{(in)}}}, \\
& p_{i1_{(in)}}, p_{i2_{(in)}}, \cdots, p_{iN_{i_{(in)}}} \}, \qquad (8) \\
G_{i_{(out)}} = \{ & p_{11_{(out)}}, p_{12_{(out)}}, \cdots, p_{1N_{1_{(out)}}}, \\
& p_{21_{(out)}}, p_{22_{(out)}}, \cdots, p_{2N_{2_{(out)}}}, \\
& p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}} \}, \qquad (9) \\
& (i = 1, 2, \cdots, L),
\end{aligned}
$$

and produce K clusters in $G_{i_{(in)}}$ and $G_{i_{(out)}}$ by means of the K-means algorithm. The centroid vectors $\boldsymbol{w}^{g_{c(in)}}$ and $\boldsymbol{w}^{g_{c(out)}}$ $(c = 1, 2, \cdots K)$ are produced in $G_{i_{(in)}}$ and $G_{i_{(out)}}$, respectively. Then, we generate improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the distance between each centroid vector $\boldsymbol{w}^{g_{c(in)}}$, $\boldsymbol{w}^{g_{c(out)}}$ $(c = 1, 2, \cdots, K)$ and initial feature vector $\boldsymbol{w}^{p_{tgt}}$ on $\boldsymbol{w}^{p_{tgt}}$. For instance, Figure 3(a) shows that we construct Web page groups $G_{2_{(in)}}$ and $G_{2_{(out)}}$ at levels up to second in the backward and forward direction from p_{tgt}, and generate improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the centroid vectors of clusters produced in Web page group $G_{2_{(in)}}$ and $G_{2_{(out)}}$ on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$. Furthermore, Figure 3(b) shows that improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting centroid vectors of each cluster on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In this method, each element $w_{t_k}'^{p_{tgt}}$ of $\boldsymbol{w}'^{p_{tgt}}$ is defined as follows:

$$
\begin{aligned}
w_{t_k}'^{p_{tgt}} &= w_{t_k}^{p_{tgt}} \\
&+ \frac{1}{Dim}\left(\sum_{c=1}^{K} \frac{w_{t_k}^{g_{c(in)}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(in)}})}\right) \\
&+ \frac{1}{Dim}\left(\sum_{c=1}^{K} \frac{w_{t_k}^{g_{c(out)}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(out)}})}\right). \qquad (10)
\end{aligned}
$$

As mentioned in Method I and II, in order to prevent the value of the second and third term of equation (10) from becoming dominant compared with the original term weight $w_{t_k}^{p_{tgt}}$, we introduce *Dim*, which denotes the number of unique terms in the Web page collection. We also define $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(in)}})$ and $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(out)}})$ as follows:

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(in)}}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_{tgt}} - w_{t_k}^{g_{c(in)}})^2},$$

$$dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c(out)}}) = \sqrt{\sum_{k=1}^{m}(w_{t_k}^{p_{tgt}} - w_{t_k}^{g_{c(out)}})^2}.$$

3. Experiment

3.1 Experimental Setup

The experiments to verify the retrieval accuracy were conducted using the TREC WT10g test collection [10]. Stop words were eliminated from all the Web pages in the collection and stemming was performed. We formed query vector $\boldsymbol{Q}$ using the terms contained in the “title” field in each Topics from 451 to 500 at the TREC WT10g test collection. This query vector $\boldsymbol{Q}$ is denoted as follows:

$$\boldsymbol{Q} = (q_{t_1}, q_{t_2}, \cdots, q_{t_m}), \qquad (11)$$

where m is the number of unique terms in the Web page collection, and $t_k (k = 1, 2, \cdots, m)$ denotes the each term. Each element q_{t_k} of $\boldsymbol{Q}$ is defined as follows:

$$q_{t_k} = \left(0.5 + \frac{0.5 \cdot Qf(t_k)}{\sum_{s=1}^{m} Qf(t_s)}\right) \cdot \log \frac{N_{web}}{df(t_k)} \quad (k = 1, 2, \cdots, m), \qquad (12)$$

where $Qf(t_k)$, N_{web}, and $df(t_k)$ is the number of index terms t_k, the total number of Web pages in the test collection, and the number of Web pages in which the term t_k appears, respectively. As reported in [11], Equation (12) is the element of a query vector that brings the best search result. We then compute the similarity $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ between improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ and query vector $\boldsymbol{Q}$. The $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ is defined as follows:

$$sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q}) = \frac{\boldsymbol{w}'^{p_{tgt}} \cdot \boldsymbol{Q}}{|\boldsymbol{w}'^{p_{tgt}}| \cdot |\boldsymbol{Q}|}. \qquad (13)$$

Based on the value of $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$, we evaluate retrieval accuracy using “precision at 11 standard recall levels” described in [12, 13].

3.2 Experimental Results

We generated improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ for initial feature vector $\boldsymbol{w}^{p_{tgt}}$ of target page p_{tgt} using Method I, II, and III described in Section 2 with respect to the following cases:

Method I

(MI-a) where the contents of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

(MI-b) where the contents of all Web pages at levels up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

(MI-c) where the contents of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction and levels up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

Method II

(MII-a) where the centroid vectors of clusters generated by the group of Web pages at each level up to ${L_{(in)}}^{th}$ in the backward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

(MII-b) where the centroid vectors of clusters generated by the group of Web pages at each level up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

(MII-c) where the centroid vectors of clusters generated by the group of Web pages at each level up to ${L_{(in)}}^{th}$ in the backward direction and each level up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

Method III

(MIII-a) where the centroid vectors of clusters generated by the group of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$, (MIII-b) where the centroid vectors of clusters generated by the group of all Web pages at levels up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$,

(MIII-c) where the centroid vectors of clusters generated by the group of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction and levels up to ${L_{(out)}}^{th}$ in the forward direction from p_{tgt} reflect on the initial feature vector $\boldsymbol{w}^{p_{tgt}}$.

In this paper, the space is limited, therefore, we show only the results that illustrate the comparison of the best retrieval accuracy obtained using the Method I, II, and III. The detailed results obtained using these methods and discussions on them are described in [8]. According to these results, we found that we could obtain the best retrieval accuracy in comparison with the tf-idf scheme, in the case of generating improved feature vector by creating a group of all Web pages at levels up to second in the backward direction from p_{tgt} and producing three clusters from the group in Method III (MIII-a). In addition, Figure 4 shows that, in any case of Method I, II, and III, the best retrieval accuracy is obtained using the contents of in-linked pages of a target page. Therefore, it is assumed that more accurate feature vectors of Web pages can be generated by assigning higher weight to in-linked pages rather than out-linked pages of a target page.

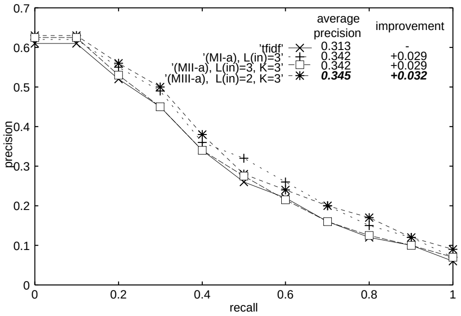

Fig.4 Comparison of the best search accuracy obtained using each Method I, II and III.

4. Conclusion

In this paper, in order to represent the contents of Web pages more accurately, we proposed three methods for improving tf-idf scheme for Web pages using their hyperlinked neighboring pages. Our approach is novel in improving tf-idf based feature vector of target Web page by reflecting the contents of its hyperlinked neighboring pages. Then, we conducted experiments with regard to the following three methods:

- the method for reflecting the contents of all Web pages at levels up to ${L_{(in)}}^{th}$ in the backward direction and levels up to ${L_{(out)}}^{th}$ in the forward direction from the target page p_{tgt},
- the method for reflecting the centroid vectors of clusters generated from Web page groups created at each level up to ${L_{(in)}}^{th}$ in the backward direction and each level up to ${L_{(out)}}^{th}$ in the forward direction from the target page p_{tgt},
- the method for reflecting the centroid vectors of clusters generated from Web page groups created at levels up to ${L_{(in)}}^{th}$ in the backward direction and levels up to ${L_{(out)}}^{th}$ in the forward direction from the target page p_{tgt},

and evaluated retrieval accuracy of improved feature vector obtained from each method using recall precision curves. Compared with respective best retrieval accuracy obtained using these three methods, we found that in-linked pages of a target page mainly affect for generating feature vector that represents the contents of the target page more accurately. Consequently, it is assumed that more accurate feature vector of Web pages can be generated by assigning higher weight to in-linked pages rather than out-linked pages of a target page. We plan to verify this assumption in future work.

In this paper, we used the K-means algorithm in order to classify the features of in- and out-linked pages of a target page. However, since we have to set the number of clusters initially in the K-means algorithm, we consider this algorithm to be inappropriate for classifying the features of Web pages that have various link environments. Therefore, in future work, we plan to devise some clustering methods appropriate for various link environments of Web pages. Moreover, in this paper, we focused on the hyperlink structures of the Web aiming at generating more accurate feature vectors of Web pages. However, in order to satisfy the user's actual information need, it is more important to find relevant Web page from the enourmous Web space. Therefore, we plan to address the technique to provide users with personalized information.

[References]

[1] K. Tajima, Y. Mizuuchi, M. Kitagawa, and K. Tanaka. Cut as a Querying Unit for WWW, Netnews, E-mail. In *Proc. of the 9th ACM Conference on Hypertext and Hypermedia (HYPERTEXT '98)*, pp. 235–244, 1998.

[2] K. Tajima, K. Hatano, T. Matsukura, R. Sano, and K. Tanaka. Discovery and Retrieval of Logical Information Units in Web. In *Proc. of the 1999 ACM Digital Libraries Workshop on Organizing Web Space (WOWS '99)*, pp. 13–23, 1999.

[3] W-S. Li, K. Selçuk Candan, Q. Vu, and D. Agrawal. Retrieving and Organizing Web Pages by "Information Unit". In *Proc. of the 10th International World Wide Web Conference (WWW10)*, pp. 230–244, 2001.

[4] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998)*, pp. 668–677, 1998.

[5] L. Page. The PageRank Citation Ranking: Bringing Order to the Web.
http://google.stanford.edu/%7Ebackrub/pageranksub.ps, 1998.

[6] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. A Method of Improving Feature Vector for Web Pages Reflecting the Contents of their Out-Linked Pages. In *Proc. of the 13th International Conference on Database and Expert Systems Applications (DEXA2002)*, pp. 891–901, 2002.

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

[8] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *Proc. of the 14th Conference on Hypertext and Hypermedia (HT'03) (to appear)*, 2003.

[9] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the 5th Berkeley Symposium on Mathmatical Statistics and Probability*, pp. 281–297, 1967.

[10] D. Hawking. Overview of the TREC-9 Web Track. *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pp. 87–102, 2001.

[11] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, Vol. 24(5), pp. 513–523, 1988.

[12] I. H. Witten and A. Moffatand T. C. Bell. Managing Gigabytes. *Van Nostrand Reinhold*, pp. 149–150, 1994.

[13] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

Kazunari SUGIYAMA

is currently a student of doctor course of Graduate School of Information Science, Nara Institute of Science and Technology. He has been working in information retrieval. He is a student member of ACM, IEICE, IPSJ, JSAI.

Kenji HATANO

is an assistant professor of Graduate School of Information Science, Nara Institute of Science and Technology. He has been working in XML database and information retrieval. He is a member of ACM, IPSJ.

Masatoshi YOSHIKAWA

is a professor of Information Technology Center, Nagoya University. He has been working in database system. He is a member of ACM, IEEE, IEICE, IPSJ.

Shunsuke UEMURA

is a professor of Graduate School of Information Science, Nara Institute of Science and Technology. He has been working in database system. He is a fellow of IEEE, IEICE, IPSJ.