Systems and Computers in Japan, Vol. 36, No. 14, 2005
Translated from Denshi Joho Tsushin Gakkai Ronbunshi, Vol. J87-D-II, No. 2, February 2004, pp. 113–125

Improvement in TF-IDF Scheme for Web Pages Based on the Contents of Their Hyperlinked Neighboring Pages

Kazunari Sugiyama,[1] Kenji Hatano,[1] Masatoshi Yoshikawa,[2] and Shunsuke Uemura[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, 630-0192 Japan

[2]Information Technology Center, Nagoya University, Nagoya, 464-8601 Japan

SUMMARY

The TF-IDF scheme is widely used to characterize documents in an information retrieval (IR) system based on the vector space model. However, for documents having a hyperlink structure such as Web pages, the Web page contents can be characterized more accurately by using the contents of hyperlinked neighboring pages. Therefore, in this paper, we propose several techniques for using the contents of hyperlinked neighboring pages to improve the TF-IDF scheme for Web pages and then verify the effectiveness of our techniques. © 2005 Wiley Periodicals, Inc. Syst Comp Jpn, 36(14): 56–68, 2005; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/scj.20189

Key words: WWW; information retrieval; TF-IDF scheme; hyperlink.

1. Introduction

The World Wide Web (WWW) is a useful resource for users to obtain a great variety of information. Today, search engines store more than eight billion Web pages [1], and that number will clearly continue to grow in the future. Therefore, it will be increasingly more difficult for users to find relevant information on the WWW. Under these circumstances, Web search engines are one of the most popular methods for finding valuable information efficiently, and they are classified into two generations based on their indexing techniques [2]. A Web page is a semistructured document that can be represented by a tree structure. Using this characteristic, in the first-generation search engines developed in the early stages of the Web, only words surrounded by title tags that existed in a portion near the root of a Web page were exploited as the indices of that page. Therefore, with this kind of characterization scheme, users were not satisfied with their retrieval accuracy. To deal with this problem, in the second-generation search engines the hyperlink structures of Web pages are taken into consideration. For example, PageRank [3] is the algorithm applied to the search engine Google (http://www.google.com), and HITS (Hypertext Induced Topic Search) [5] is the algorithm applied to the search engine in the CLEVER project [6]. In these algorithms, weighting Web pages based on hyperlink structures achieves higher retrieval accuracy compared with the first-generation search engines. However, these algorithms have shortcomings that (1) the weight for a Web page is merely defined and (2) the relativity of contents among hyperlinked Web pages is not taken into consideration. Consequently, there still remains the problem that Web pages irrelevant to a user's query are often ranked highly.

Taking these points into account, in order to provide users with Web pages relevant to a user's query, it is necessary to develop a technique for representing the contents of Web pages more accurately. Therefore, feature vectors

should be calculated so that both Web pages linked to the target Web page (in-linked pages) and from the target Web page (out-linked pages) are taken into consideration. In this paper, we first propose several approaches to improving feature vectors of a target Web page created in advance based on the TF-IDF scheme [7] using both its in- and out-linked pages. Then we verify the effectiveness of our proposed approaches. Compared with the second-generation search engines, our approach is novel in characterizing Web pages more accurately by reflecting the contents of their hyperlinked neighboring Web pages.

The rest of this paper is organized as follows. In Section 2, we review related work using hyperlink structures of the WWW. In Section 3, we propose novel methods for improving feature vectors of Web page by using their hyperlinked neighboring pages. In Section 4, we present the experimental results for evaluating our proposed methods and discuss those results. In Section 5, we conclude the paper with a summary and directions for future work.

2. Related Work

Hyperlink structures are one of the features of the WWW. Users can navigate the huge Web space easily through the hyperlink structures; therefore, a great deal of research related to Web information retrieval has focused on the hyperlink structure of the WWW. In this section, we review related work of IR systems using the hyperlink structure of the WWW, especially IR systems based on the concept of "optimal document granularity" and the two most popular Web page weighting algorithms, *HITS* and *PageRank*.

2.1. Information retrieval systems based on the concept of "optimal document granularity"

With respect to this research area, we refer to the following works. Tajima and colleagues [8] proposed a technique that uses a concept called "cuts" (results of Web structure analysis) as retrieval units for the WWW. Moreover, they extended to rank search results that include multiple keywords by (1) finding minimal subgraphs of links and Web pages including all keywords from the hyperlink structure of the WWW and (2) calculating similarities for keywords related to each minimal subgraph based on the locality of the keywords within it [9]. Following these works, Li and colleagues [10] introduced the concept of an "information unit" as one minimal retrieval unit for a document that consists of multiple Web pages and proposed a novel framework for Web page retrieval based on these "information units." However, these approaches require considerable processing time to analyze hyperlink structures and to discover the semantics of Web pages. In addition, it is often the case that they find retrieval units that are irrelevant to the query terms specified by the user. As for these works, we do not believe that users could understand the search results intuitively, since the multiple query keywords are scattered among several hyperlinked Web pages.

2.2. HITS algorithm

The HITS algorithm [5] is applied to the search engine in the CLEVER project [6]. This algorithm depends on the query, and considers the set of pages S that point to or are redirected to, pages in the answer. In S, pages that have many links pointing to them are called "authorities," while Web pages that have many outgoing links are called "hubs." In other words, better authorities come from incoming edges from good hubs, and better hubs come from outgoing edges to good authorities. Let $H(p)$ and $A(p)$ be the hub and authority score of Web page p, respectively. These scores are defined so that the following equations are satisfied for all pages p:

$$H(p) = \sum_{u \in S | p \to u} A(u), \quad A(p) = \sum_{v \in S | v \to p} H(v)$$

where $H(p)$ and $A(p)$ are normalized for all pages p. These scores can be determined through an iterative algorithm, and they converge to the principal eigenvector of the link matrix of S. However, the major problem in this algorithm is that the Web pages that the root document points to get the largest authority scores because the hub score of the root page dominates all the others when a Web page has few in-links but a large number of out-links. In order to resolve this problem, several extended HITS algorithms have also been proposed [11–15].

2.3. PageRank algorithm

The PageRank algorithm simulates a user navigating randomly in the Web who jumps to a random page with probability d or follows a random hyperlink with probability $1 - d$. It is further assumed that this user never returns to a previously visited page following an already traversed hyperlink backwards. This process can be modeled with a Markov chain, from which the stationary probability of being in each page can be computed. This value is then used as a part of the ranking mechanism used by Google. Let $C(a)$ be the number of outgoing links from Web page a, and suppose that Web pages p_1 to p_n point to Web page a. Then, the PageRank *PR(a)* of a is defined as

$$PR(a) = d + (1 - d) \sum_{i=1}^{n} \frac{PR(p_i)}{C(p_i)}$$

where the value of d is empirically set to about 0.15 to 0.2 by the system. The weights of other Web pages are normalized by the number of links in the Web page. PageRank can be computed using an iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web. The major problems of this algorithm are that (1) the contents of Web pages are not analyzed, so the "importance" of a given Web page is independent of the query, and (2) specific famous Web sites tend to be ranked more highly. Therefore, to yield more accurate search results, some algorithms that extend this algorithm have been proposed [16, 17].

3. Proposed Method

As we described in Section 2.1, the IR systems based on the concept of "optimal document granularity" have a problem, in that the search results are incomprehensible for users. Moreover, HITS and PageRank algorithms also have problems: (1) the weights for a Web page are merely defined and (2) the relativity of contents among hyperlinked Web pages is not considered. Based on these problems, we believe that the feature vectors of Web pages should be generated by reflecting the contents of hyperlinked neighboring pages in order to represent the contents of Web pages more accurately. Therefore, in this paper, we propose a method for 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 Web page be p_{tgt}. Then we define i as the length of the shortest directed path from p_{tgt}. Let us assume that there are N_i Web pages $(p_{i_1}, p_{i_2}, \ldots, p_{i_{N_i}})$ 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 denotes the number of unique terms in the Web page collection, and t_k $(k = 1, 2, \ldots, m)$ denotes each term. Using the TF-IDF scheme, we also define each element $w_{t_k}^{p_{tgt}}$ of $\boldsymbol{w}^{p_{tgt}}$ as follows:

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

where $tf(t_k, p_{tgt})$ is the frequency of term t_k in the target Web 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 $\boldsymbol{w}'^{p_{tgt}}$ of $\boldsymbol{w}^{p_{tgt}}$ 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." In the following, we also define the distance $dis(\boldsymbol{a}, \boldsymbol{b})$ between two m-dimensional vectors $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)$ as

$$dis(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{k=1}^{m} (a_k - b_k)^2} \tag{4}$$

In this paper, we propose three methods for improving the initial feature vector $\boldsymbol{w}^{p_{tgt}}$ defined by Eq. (1) based on the TF-IDF scheme defined by Eq. (2). In the following, the forward direction means navigating to a page that is linked from p_{tgt} [out-linked page; downward from p_{tgt} in Figs. 1(a), 2(a), and 3(a)], and the backward direction means navigating to a page that is linked to p_{tgt} [in-linked page; upward from p_{tgt} in Figs. 1(a), 2(a), and 3(a)].

3.1. Method I

In this method, we reflect the feature vectors 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} on the feature vector of 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 the one hand such Web pages exist right near p_{tgt}, on the other hand they might also exist far removed from p_{tgt} in the vector space, we generate the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the distance between the initial feature vector $\boldsymbol{w}^{p_{tgt}}$ and the feature vectors of the 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, Fig. 1(a) shows that $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting the feature vectors of all Web pages at levels up to the second in the backward and forward directions from p_{tgt} on $\boldsymbol{w}^{p_{tgt}}$. In Fig. 1(a), $p_{ij_{(in)}}$ and $p_{ij_{(out)}}$ correspond to the j-th pages in the i-th levels in the backward and forward directions from p_{tgt}, respectively. In addition, Fig. 1(b) shows that improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting each feature vector of in- and out-linked pages 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 by preliminary experiments as follows:

$$w_{t_k}'^{p_{tgt}} = w_{t_k}^{p_{tgt}} + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}} \sum_{j=1}^{N_{i_{(in)}}} \frac{w_{t_k}^{p_{ij_{(in)}}}}{N_{i_{(in)}} \cdot dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(in)}}})}\right) + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}} \sum_{j=1}^{N_{i_{(out)}}} \frac{w_{t_k}^{p_{ij_{(out)}}}}{N_{i_{(out)}} \cdot dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(out)}}})}\right) \tag{5}$$

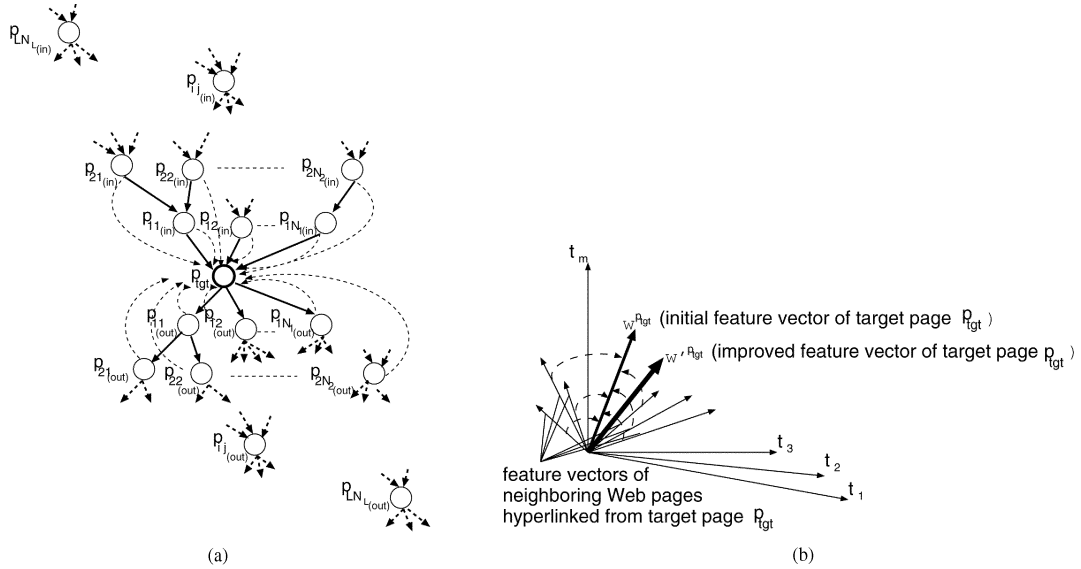

Fig. 1. The improvement of a feature vector as performed by Method I [(a) in the Web space, (b) in the vector space].

Equation (5) shows that the product of $w_{t_k}^{p_{tgt}}$ (weight of term t_k in Web page $p_{ij_{(in)}}$) and the reciprocal of $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(in)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{p_{ij_{(in)}}}$ in the vector space), and similarly, the product of $w_{t_k}^{p_{ij_{(out)}}}$ (weight of term t_k in Web page $p_{ij_{(out)}}$) and the reciprocal of $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{p_{ij_{(out)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{p_{ij_{(out)}}}$ in the vector space) is added to $w_{t_k}^{p_{tgt}}$ [weight of term t_k in p_{tgt}, computed by TF-IDF scheme defined by Eq. (2)] with respect to 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}. *Dim*, which denotes the number of unique terms, is introduced to prevent the second and third terms of Eq. (5) from being dominant over the weight $w_{t_k}^{p_{tgt}}$ of the original index term.

3.2. 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 target page p_{tgt}. Then, we generate improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting centroid vectors of the clusters generated from $G_{i_{(in)}}$ and $G_{i_{(out)}}$ on the initial feature vector

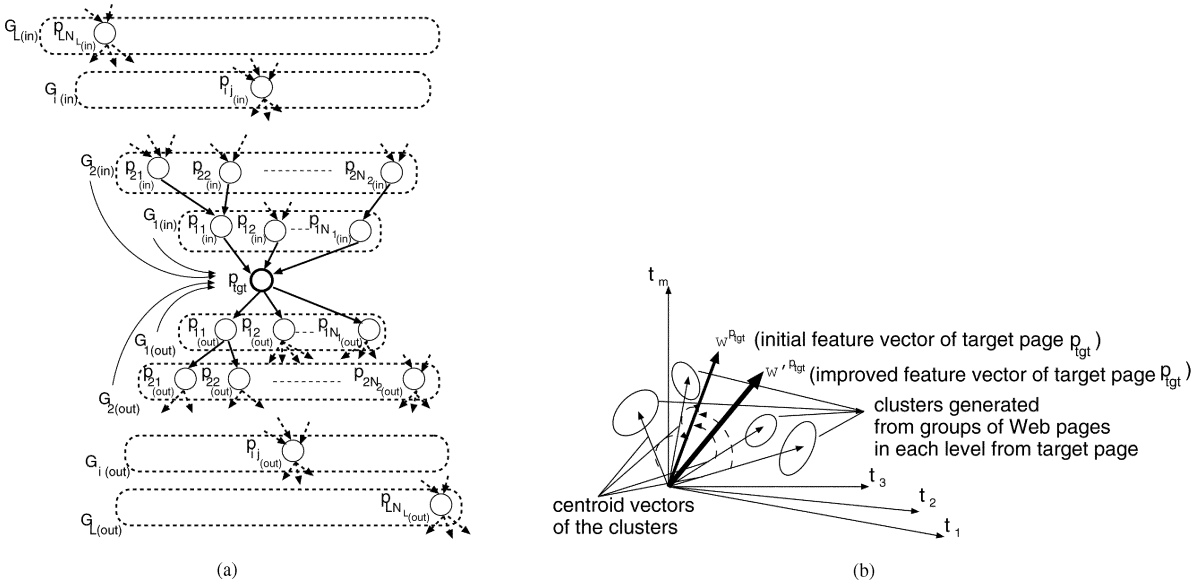

Fig. 2. The improvement of a feature vector as performed by Method II [(a) in the Web space, (b) in the vector space].

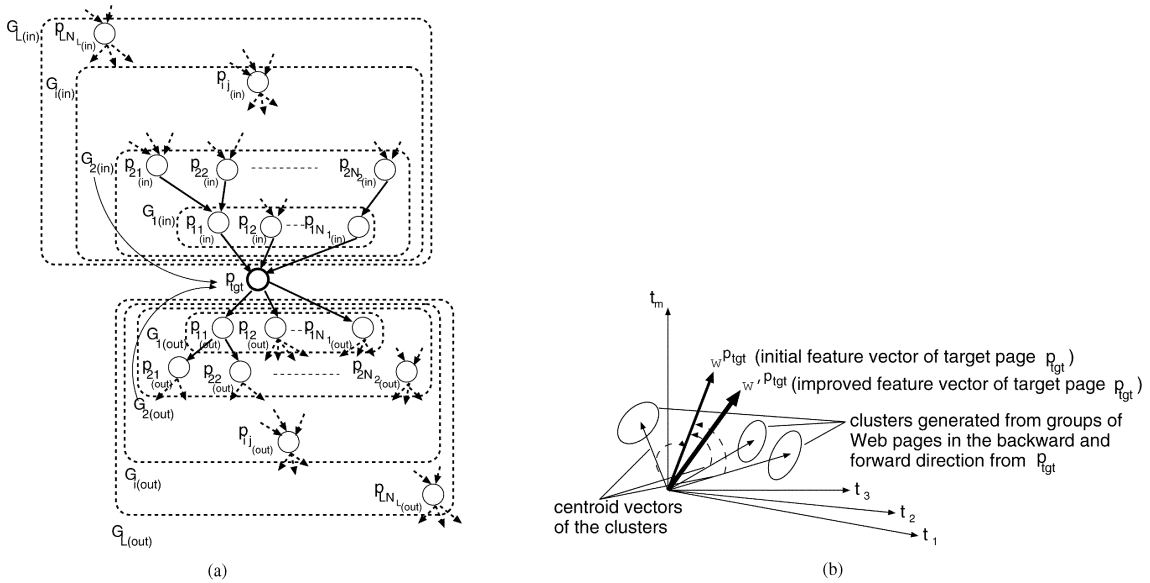

Fig. 3. The improvement of a feature vector as performed by Method III [(a) in the Web space, (b) in the vector space].

$\boldsymbol{w}^{p_{tgt}}$. This method is based on the idea that Web pages at each level in the backward and forward directions from p_{tgt} are classified into several topics in 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 construct Web page groups $G_{i_{(in)}}$ and $G_{i_{(out)}}$, which are defined by

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

$$G_{i_{(out)}} = \{p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}}\} \quad (7)$$

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

Then, we produce K clusters in each Web page group $G_{i_{(in)}}$ and $G_{i_{(out)}}$ by means of the K-means algorithm [18]. The centroid vectors $\boldsymbol{w}^{g_{ic_{(in)}}}$ and $\boldsymbol{w}^{g_{ic_{(out)}}}$ (c = 1, 2, . . . , K) are produced in $G_{i_{(in)}}$ and $G_{i_{(out)}}$, respectively. We generate the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the distance between each centroid vectors $\boldsymbol{w}^{g_{ic_{(in)}}}$ and $\boldsymbol{w}^{g_{ic_{(out)}}}$ and the initial feature vector $\boldsymbol{w}^{p_{tgt}}$ on each element of $\boldsymbol{w}^{p_{tgt}}$.

For instance, Fig. 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 the second in the backward and forward directions from p_{tgt}, and generate an improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the centroid vectors of each cluster produced in each of these Web page groups on initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In addition, Fig. 2(b) shows that improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting the centroid vectors of each cluster on initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In this method, each element $w_{t_k}^{'p_{tgt}}$ of $\boldsymbol{w}'^{p_{tgt}}$ is defined by preliminary experiments as follows:

$$w_{t_k}^{'p_{tgt}} = w_{t_k}^{p_{tgt}} + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(in)}}\sum_{c=1}^{K}\frac{w_{t_k}^{g_{ic_{(in)}}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})}\right) + \frac{1}{Dim}\left(\sum_{i=1}^{L_{(out)}}\sum_{c=1}^{K}\frac{w_{t_k}^{g_{ic_{(out)}}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})}\right) \quad (8)$$

Equation (8) shows that the product of $w_{t_k}^{g_{ic_{(in)}}}$ (weight of term t_k in centroid vector $\boldsymbol{w}^{g_{ic_{(in)}}}$ of cluster c constructed from $G_{i_{(in)}}$) and the reciprocal of the distance $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(in)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{g_{ic_{(in)}}}$ in the vector space), and similarly, the product of $w_{t_k}^{g_{ic_{(out)}}}$ (weight of term t_k in centroid vector $\boldsymbol{w}^{g_{ic_{(out)}}}$ of cluster c constructed from $G_{i_{(out)}}$) and the reciprocal of the distance $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{ic_{(out)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{g_{ic_{(out)}}}$ in the vector space) are added to $w_{t_k}^{p_{tgt}}$ (weight of term t_k in p_{tgt}, computed by TF-IDF scheme defined by Eq. (2)] with respect to all centroid vectors constructed 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}. In addition, Dim, which denotes the number of unique terms, is introduced to prevent the second and third terms of Eq. (8) from being dominant over the weight $w_{t_k}^{p_{tgt}}$ of the original index term.

3.3. 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 target page p_{tgt} are composed of several 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 then generate the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting centroid vectors of the clusters on 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 first create Web page groups $G_{i_{(in)}}$ and $G_{i_{(out)}}$ defined by

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

$$\begin{aligned} G_{i_{(out)}} &= \{p_{11_{(out)}}, p_{12_{(out)}}, \cdots, p_{1N_{1_{(out)}}}, \\ &\quad p_{21_{(out)}}, p_{22_{(out)}}, \cdots, p_{2N_{2_{(out)}}}, \\ &\quad p_{i1_{(out)}}, p_{i2_{(out)}}, \cdots, p_{iN_{i_{(out)}}}\} \\ &\quad (i = 1, 2, \cdots, L) \end{aligned} \tag{10}$$

Then, we produce K clusters to create 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, . . . , K) are produced in $G_{i_{(in)}}$ and $G_{i_{(out)}}$, respectively. Then we generate the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the distance between each centroid vector $\boldsymbol{w}^{g_{c_{(in)}}}$ and $\boldsymbol{w}^{g_{c_{(out)}}}$ (c = 1, 2, . . . , K) and initial feature vector $\boldsymbol{w}^{p_{tgt}}$ on each element of $\boldsymbol{w}^{p_{tgt}}$.

For example, Fig. 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 directions from p_{tgt}, and generate improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ by reflecting the centroid vectors of clusters produced in these Web page groups on initial feature vector $\boldsymbol{w}^{p_{tgt}}$. Furthermore, Fig. 3(b) shows that improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ is generated by reflecting centroid vectors of each cluster on initial feature vector $\boldsymbol{w}^{p_{tgt}}$. In this method, each element $w_{t_k}'^{p_{tgt}}$ of $\boldsymbol{w}'^{p_{tgt}}$ is defined by preliminary experiments as follows:

$$\begin{aligned} w_{t_k}'^{p_{tgt}} &= w_{t_k}^{p_{tgt}} \\ &\quad + \frac{1}{Dim}\left(\sum_{c=1}^{K} \frac{w_{t_k}^{g_{c_{(in)}}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(in)}}})}\right) \\ &\quad + \frac{1}{Dim}\left(\sum_{c=1}^{K} \frac{w_{t_k}^{g_{c_{(out)}}}}{dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(out)}}})}\right) \end{aligned} \tag{11}$$

Equation (11) shows that the product of the weight $w_{t_k}^{g_{c_{(in)}}}$ (weight of term t_k in centroid vector $\boldsymbol{w}^{g_{c_{(in)}}}$ of cluster c generated from $G_{i_{(in)}}$) and the reciprocal of $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(in)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{g_{c_{(in)}}}$ in the vector space), and similarly the product of element $w_{t_k}^{g_{c_{(out)}}}$ (weight of term t_k in centroid vector $\boldsymbol{w}^{g_{c_{(out)}}}$ of cluster c generated from $G_{i_{(out)}}$) and the reciprocal of $dis(\boldsymbol{w}^{p_{tgt}}, \boldsymbol{w}^{g_{c_{(out)}}})$ (the distance between $\boldsymbol{w}^{p_{tgt}}$ and $\boldsymbol{w}^{g_{c_{(out)}}}$ in the vector space) are added to $w_{t_k}^{p_{tgt}}$ [weight of term t_k in p_{tgt}, computed by TF-IDF scheme defined by Eq. (2)], with respect to the number of clusters K. As mentioned in Methods I and II, in order to prevent the value of the second and third terms of Eq. (11) from being dominant compared with the original term weight $w_{t_k}^{p_{tgt}}$, we introduce Dim, which denotes the number of unique terms.

4. Experiments

4.1. Experimental setup

We implemented the three methods described in Section 3 using Perl on a workstation (CPU: UltraSparc-II 480 MHz×4; memory: 2 Gbytes; OS: Solaris 8) and conducted experiments in order to verify the retrieval accuracy of each method using the TREC (text retrieval conference) WT10g test collection [19]. This test collection was created by Commonwealth Scientific & Industrial Research Organization (CSIRO; http://www.csiro.au/) in Australia in 1997 using a portion of the Web pages collected by the Internet Archive (http://www.archive.org). This collection consists of approximately 1.69 million Web pages (10 Gbytes) and information on in- and out-linked pages for each Web page in the test collection, sets of query topics, and set of relevant documents. The set of query topics was created for 50 topics based on the query log of Excite (http://www.excite.com/), and is composed of a title field describing the search request, description field describing a sentence that satisfies the queries, and a narrative field describing the criteria for judging relevant documents, as shown in Fig. 4.

Stop words were eliminated from all Web pages in the collection based on the stopword list (ftp://ftp.cs.cornell.edu/pub/smart/English.stop) and stemming was performed using the Porter Stemmer [20] (http://www.tartarus.org/%7Emartin/PorterStemmer/). We formulated query vector represented as shown in Eq. (12) using the terms included in the "title" field for each of the 50 sets of query topics. Figure 4 shows one of the examples of a query topic.

$$\boldsymbol{Q} = (q_{t_1}, q_{t_2}, \cdots, q_{t_m}) \tag{12}$$

In Eq. (12), t_k (k = 1, 2, . . . , m) denotes an index term, and each element q_{t_k} is defined by

$$q_{t_k} = \left(0.5 + \frac{0.5 \cdot Qf(t_k)}{\sum_{k=1}^{m} Qf(t_k)}\right) \cdot \log \frac{N_{web}}{df(t_k)} \quad (k = 1, 2, \cdots, m) \tag{13}$$

where $Qf(t_k)$, N_{web}, and $df(t_k)$ denote the number of index terms t_k contained in query vector $\boldsymbol{Q}$, the total number of

```
<num> Number: 462
<title> real estate and new jersey

<desc> Description:
Find documents that contain residential
real estate listings within New Jersey.

<narr> Narrative:
Documents containing realtor data such
as point of contact, address, web site
or email address considered as a
real estate listing are relevant.
Listings of commercial real estate for
sale or auction are not relevant.
```

Fig. 4. An example of topic descriptions in the WT10g test collection.

Web pages in the test collection, and the number of Web pages in which term t_k appears, respectively. As reported in Ref. 21, Eq. (13) is the element of a query vector that brings the best retrieval accuracy. We then compute the similarity $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$ between the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ described in Section 3 and query vector $\boldsymbol{Q}$ using the equation

$$sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q}) = \frac{\boldsymbol{w}'^{p_{tgt}} \cdot \boldsymbol{Q}}{|\boldsymbol{w}'^{p_{tgt}}| \cdot |\boldsymbol{Q}|} \tag{14}$$

We evaluate our experimental results in Section 4.3 using average precision based on the value of $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$. In addition, we compute average precision $\bar{P}$ based on the equation

$$\bar{P} = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{Rel_{q_i}}{R_{q_i}} \tag{15}$$

where q_i is the i-th ($i = 1, 2, \ldots, N_q$) query, R_{q_i} is the number of relevant documents for q_i, and Rel_{q_i} is the number of the top R_{q_i} relevant documents that system returns. In this paper, with regard to the set of 50 retrieval tasks (N_q = 50), we apply Eq. (15) to evaluate the top 1000 Web pages that our proposed methods return based on the value of $sim(\boldsymbol{w}'^{p_{tgt}}, \boldsymbol{Q})$.

4.2. Experimental methods

In order to verify the effectiveness of the three proposed methods described in Section 3, we generated the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ for initial feature vector $\boldsymbol{w}^{p_{tgt}}$ of target page p_{tgt} with respect to the following cases:

[Method I]

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

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

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

[Method II]

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

(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 initial feature vector $\boldsymbol{w}^{p_{tgt}}$

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

[Method III]

(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 initial feature vector $\boldsymbol{w}^{p_{tgt}}$

(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 initial feature vector $\boldsymbol{w}^{p_{tgt}}$

(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 initial feature vector $\boldsymbol{w}^{p_{tgt}}$.

We generated the improved feature vector $\boldsymbol{w}'^{p_{tgt}}$ and conducted experiments to compare retrieval accuracies of our proposed methods by varying $L_{(in)}$ and $L_{(out)}$ for Methods I, II, and III so that $1 \le L_{(in)} \le 5$ for (a), $1 \le L_{(out)} \le 5$ for (b), and $1 \le L_{(in)}, L_{(out)} \le 5$ for (c) and also by varying the number of clusters K for Methods II and III so that $1 \le K \le 5$.

4.3. Experimental results and discussion

Figure 5 illustrates the average precision when the values of $L_{(in)}$, $L_{(out)}$, or $[L_{(in)}, L_{(out)}]$ vary in Method I (a),

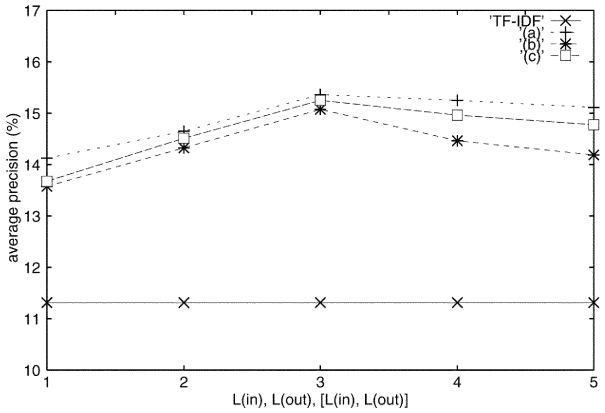

Fig. 5. Average precision based on Method I.

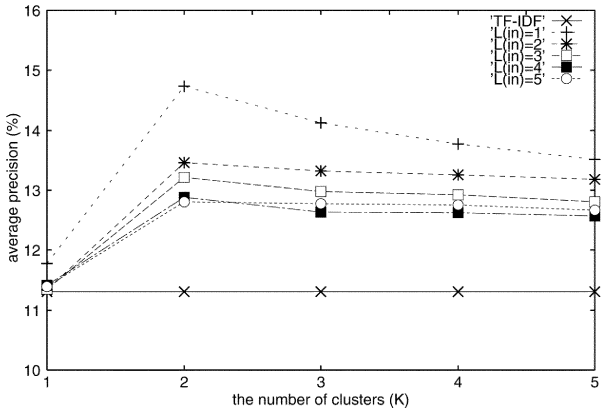

Fig. 7. Average precision based on Method II(a).

(b), or (c), respectively. Figures 7 to 9 illustrate the average precision when the number of clusters K varies in Method II (a), (b), and (c), respectively. Figures 10 to 12 illustrate the average precision when the number of clusters K varies in Method III (a), (b), and (c), respectively. In order to facilitate the comparison of the average precision obtained by our proposed methods and TF-IDF scheme, we also show the retrieval accuracy obtained by our proposed methods and TF-IDF scheme in each figure. The precision of the TF-IDF scheme does not depend on the values of $L_{(in)}$, $L_{(out)}$, and the number of clusters K. Therefore, the retrieval accuracy obtained by the TF-IDF scheme is fixed for these values.

We can observe the following findings in each method [22]. In Method I, as shown in Fig. 5, although the similarity between the target page p_{tgt} and Web pages at levels up to third ($L_{(in)} \leq 3$) in the backward direction from p_{tgt} and Web pages at levels up to third ($L_{(out)} \leq 3$) in the

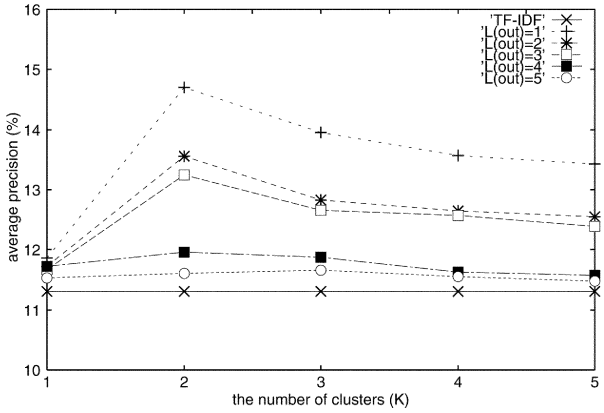

Fig. 8. Average precision based on Method II(b).

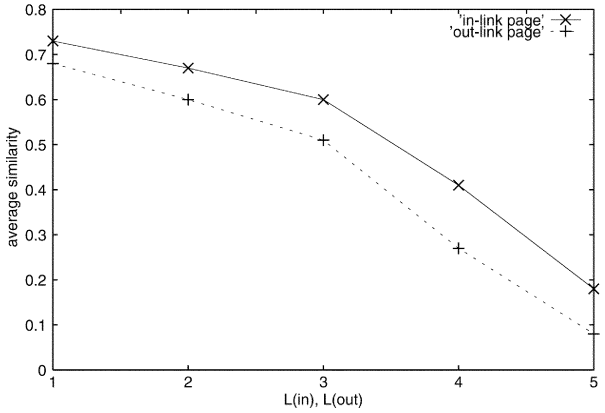

Fig. 6. Distribution of average similarity.

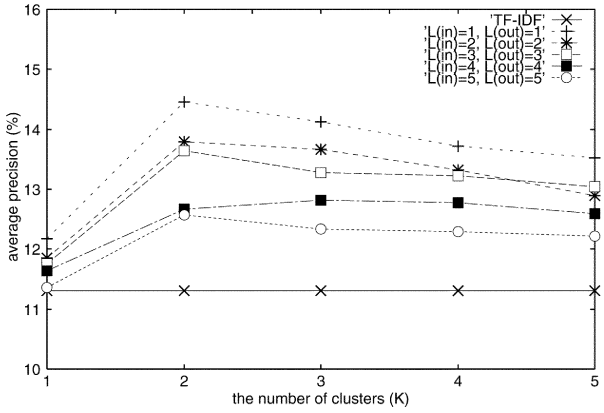

Fig. 9. Average precision based on Method II(c).

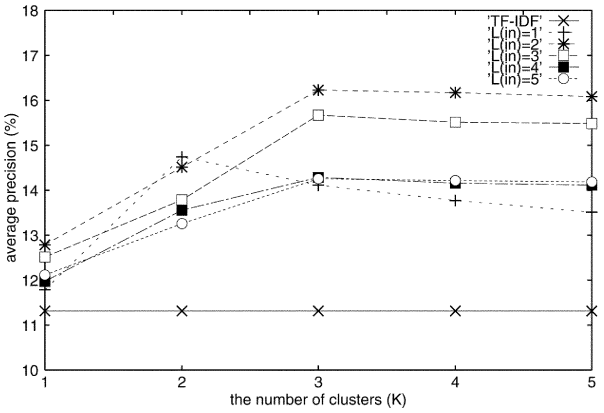

Fig. 10. Average precision based on Method III(a).

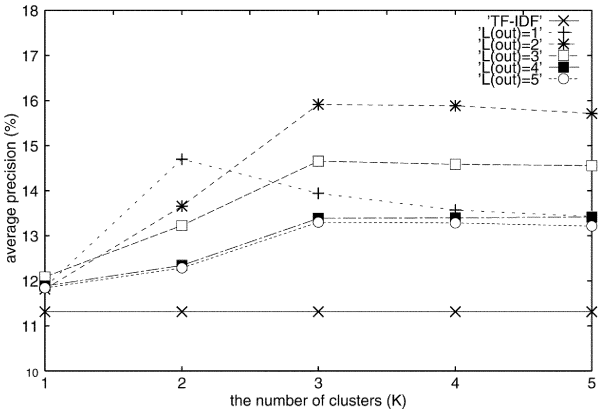

Fig. 11. Average precision based on Method III(b).

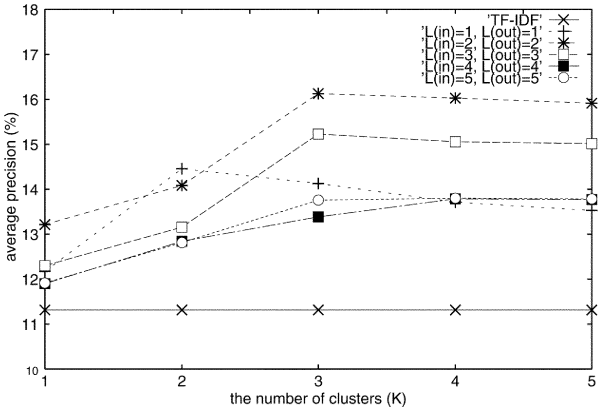

Fig. 12. Average precision based on Method III(c).

forward direction from p_{tgt} is high, these neighboring Web pages of p_{tgt} contribute in representing the contents of p_{tgt} more accurately. However, the similarity between the target page p_{tgt} and Web pages from fourth ($L_{(in)} \geq 4$) level in the backward direction from p_{tgt} and Web pages from fourth ($L_{(out)} \geq 4$) level in the forward direction from p_{tgt} is low; therefore we cannot observe the effect of representing the contents of p_{tgt} more accurately. Here, in Fig. 6, we show the distribution of the average similarity between Web pages in the WT10g test collection and Web pages in the backward and forward directions from each Web page in the collection. The average similarity between a Web page p in the collection and Web pages at levels up to third ($L_{(in)} \leq 3$) in the backward direction from p, and Web pages at levels up to third ($L_{(out)} \leq 3$) in the forward direction from p is relatively high. However, the average similarity between a Web page p in the collection and Web pages from fourth ($L_{(in)} \geq 4$) level in the backward direction from p and Web pages from fourth ($L_{(out)} \geq 4$) level in the forward direction from p is low. We consider that these facts affect the average precision. In addition, as Fig. 5 shows, in the case of $L_{(in)} \geq 4$ or $L_{(out)} \geq 4$, the average precision tends to decline. Therefore, it is appropriate that we examined the average precision in the range of $1 \leq L_{(in)} \leq 5$ or $1 \leq L_{(out)} \leq 5$.

In Method II, as shown in Figs. 7 to 9, the farther the distance from target Web page p_{tgt}, in other words, the larger the values of $L_{(in)}$ or $L_{(out)}$ are, the smaller the gap between the graph of the average precision obtained by our proposed methods and the graph of the average precision obtained by the TF-IDF scheme. In other words, the degree of improvement in retrieval accuracy is small compared with the TF-IDF scheme. Thus, we found that with regard to the contents of Web pages, there is strong similarity between the feature vector of target page p_{tgt}, and the centroid vector generated by the Web page groups at each level up to first in the backward ($L_{(in)} = 1$) and forward ($L_{(out)} = 1$) directions from p_{tgt}. However, we also found that similarity between the feature vector of p_{tgt} and the centroid vector generated by the group of Web pages at each level from p_{tgt} decreases as the value of i, which denotes the length of the shortest directed path from p_{tgt} to its hyperlinked neighboring pages, increases.

In Method III, as shown in Figs. 10 to 12, the best retrieval accuracy is obtained when we generate improved feature vectors using each centroid vector of three clusters ($K = 3$) generated in a Web page group produced using all Web pages at levels up to second ($L_{(in)} \leq 2$, $L_{(out)} \leq 2$) in the backward and forward directions from the target page p_{tgt}. Since the best retrieval accuracy is obtained when the number of clusters is three ($K = 3$) in this method, we can infer that the topics of Web pages at levels up to second ($L_{(in)} \leq 2$, $L_{(out)} \leq 2$) in the backward and forward directions from p_{tgt} are usually composed of three topics.

Furthermore, we found the following relations between the number of clusters in Methods II and III. First, in Method II, as shown in Figs. 7 to 9, we observed that the average precision tends to decrease in the case of $K \geq 3$. On the other hand, in Method III, we obtain the same results as those of Method II, in the case of $L_{(in)} = 1$ or $L_{(out)} = 1$, as shown in Eqs. (8) and (11). However, in the case of $L_{(in)} \geq 2$ or $L_{(out)} \geq 2$, we observed that the average precision decreases gradually, when the number of clusters K is greater than 4. Therefore, we consider that it is valid that we examined the average precision in the range of $1 \leq K \leq 5$ in Methods II and III.

Table 1, which summarizes the results described above, illustrates the average precision when we generated the feature vector of Web page using TF-IDF scheme and the best average precision when we generated the feature vector of Web page using each of our proposed methods. In Method I, the best retrieval accuracy is obtained when we generate the feature vector of Web page by utilizing the contents of all Web pages at levels up to third ($L_{(in)} = 3$) in the backward direction from the target page p_{tgt}. In Method II, the best retrieval accuracy is obtained when we generate improved feature vectors using each centroid vector of two clusters ($K = 2$) generated in a Web page group produced using all Web pages at levels up to first ($L_{(in)} = 1$) in the backward direction from the target page p_{tgt}. Moreover, in Method III, we obtain the best retrieval accuracy when we generate improved feature vectors using each centroid vector of three clusters ($K = 3$) generated in a Web page group produced using all Web pages at levels up to second ($L_{(in)} = 2$) in the backward direction from the target page p_{tgt}. Furthermore, as shown in Table 1, in any case of Methods I, II, and III, the best retrieval accuracy is obtained in experiment (a), namely, in the case of using the contents of in-linked pages of a target page. We consider that this finding is obtained because we can easily reach Web pages similar to the target Web page when we follow the hyperlinks in the backward direction from the target page while we reach various Web pages that are not so similar to the contents of the target page when we follow the hyperlinks in the forward direction from the target page. In other words, Web pages have a characteristic that the in-linked pages of a target Web page have Web pages relevant to the contents of the target page. As described in Section 2.2, the HITS algorithm defines Web pages that have many outgoing links as "hubs," and also defines the quality of a Web page as "authority" by considering the hubs as its in-linked pages of authority. In addition, focusing on the importance of in-linked pages, the tool that enables navigation in the backward direction from a target page is also developed [23]. We can consider that the results in Table 1 suggest the usefulness of in-linked pages even in our study.

Table 1. Comparison of the optimal search accuracy that was obtained using Methods I, II, and III

	% average precision	% improvement
TF-IDF	11.31	—
Method I ($L_{(in)} = 3$)	15.30	+3.99
Method II ($L_{(in)} = 1$, $K = 2$)	14.74	+3.43
Method III ($L_{(in)} = 2$, $K = 3$)	**16.23**	**+4.92**

Table 2. Average precision of WT10g by using link information

Group	Outline of each approach	% average precision
Ref. 27	Anchor text	20.00
	Anchor text + long query	18.38
Ref. 26	Content-link	16.31
	4gram content-link	17.94
Ref. 24	Cocitation top 10	16.30
	Cocitation top 50	13.37
	HITS	4.88
Ref. 28	Okapi + probabilistic augmentation	17.36
Ref. 29	Anchor text	12.50
	Variant of anchor text	12.88
Ref. 30	Back link frequency	10.62
Ref. 25	Modified HITS	5.91
	Modified HITS with weighted links	6.37

4.4. Discussion related to search accuracy

According to Ref. 19, in the TREC-9 Web Track, the outline of methods using link information by participants and average precision are shown in Table 2. The reader is referred to the papers in Table 2 for their detailed methods. It is seen that the average precision based on HITS in Ref. 24 is 4.88%, and the average precisions based on modified HITS in Ref. 25 are 5.91 and 6.37%. These are poor results. On the other hand, as shown in Table 1, our best average precision is 16.23%. Since this result is comparable to the results of Refs. 26 and 24, we believe that our proposed methods are effective enough to characterize Web pages more accurately and the results obtained by using them are sufficiently solid.

5. Conclusions

In this paper, in order to represent the contents of Web pages more accurately, we proposed three methods for improving the TF-IDF scheme for Web pages using the feature vectors of hyperlinked neighboring pages. Our method is innovative in improving the TF-IDF based feature vector of a target Web page by reflecting the contents of its hyperlinked neighboring Web pages, which had not been done up to now. Naturally, if our scheme is used together with HITS or PageRank, which had already been proposed, a further improvement in retrieval accuracy can be expected.

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 pages from the enormous WWW space. Therefore, we plan to address the technique to provide users with personalized information.

REFERENCES

1. Baeza-Yates R, Saint-Jean F, Castillo C. Web structure, dynamics and page quality. Proc 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002), p 117–130.
2. Broder A, Raghavan P. Combining text- and link-based information retrieval on the web. SIGIR'01 Pre-Conference Tutorials.
3. Page L. The PageRank citation ranking: Bringing order to the web. http://google.stanford.edu/~backrub/pageranksub.ps, 1998.
4. Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. Proc 7th International World Wide Web Conference (WWW7), p 107–117, 1998.
5. Kleinberg JM. Authoritative sources in a hyperlinked environment. Proc 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998), p 668–677.
6. IBM Almaden Research Center. Clever Searching. http://www.almaden.ibm.com/cs/k53/clever.html
7. Salton G, McGill MJ. Introduction to modern information retrieval. McGraw–Hill; 1983.
8. Tajima K, Mizuuchi Y, Kitagawa M, Tanaka K. Cut as a querying unit for WWW, netnews, e-mail. Proc 9th ACM Conference on Hypertext and Hypermedia (Hypertext '98), p 235–244.
9. Tajima K, Hatano K, Matsukura T, Sano R, Tanaka K. Discovery and retrieval of logical information units in web. Proc 1999 ACM Digital Libraries Workshop on Organizing Web Space (WOWS '99), p 13–23.
10. Li W-S, Selçuk Candan K, Vu Q, Agrawal D. Retrieving and organizing web pages by "Information Unit". Proc 10th International World Wide Web Conference (WWW10), p 230–244, 2001.
11. Chakrabarti S, Dom B, Raghavan P, Rajagopalan S, Gibson D, Kleinberg J. Automatic resource compilation by analyzing hyperlink structure and associated text. Proc 7th International World Wide Web Conference (WWW7), p 65–74, 1998.
12. Bharat K, Henzinger MR. Improved algorithms for topic distillation in a hyperlinked environment. Proc 21st Annual International ACM SIGIR Conference (SIGIR '98), p 104–111.
13. Li L, Shang Y, Zhang W. Improvement of HITS-based algorithms on web documents. Proc 11th International World Wide Web Conference (WWW2002), p 527–535.
14. Chakrabarti S. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. Proc 10th International World Wide Web Conference (WWW10), p 211–220, 2001.
15. Chakrabarti S, Joshi M, Tawde V. Enhanced topic distillation using text, markup tags, and hyperlinks. Proc 23rd Annual International ACM SIGIR Conference (SIGIR 2001), p 208–216.
16. Rafiei D, Mendelzon AO. What is this page known for? Computing web page reputations. Proc 9th International World Wide Web Conference (WWW9), p 823–835, 2000.
17. Haveliwala TH. Topic-sensitive PageRank. Proc 11th International World Wide Web Conference (WWW2002), p 517–526.
18. MacQueen J. Some methods for classification and analysis of multivariate observations. Proc 5th Berkeley Symposium on Mathematical Statistics and Probability, p 281–297, 1967.
19. Hawking D. Overview of the TREC-9 web track. NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9), p 87–102, 2001.
20. Porter MF. An algorithm for suffix stripping. Program 1988;14:130–137.
21. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manage 1988;24:513–523.
22. Sugiyama K, Hatano K, Yoshikawa M, Uemura S. Refinement of TF-IDF schemes for web pages using their hyperlinked neighboring pages. Proc 14th ACM Conference on HyperText and Hypermedia (HT '03), p 198–207.

23. Chakrabarti S, Gibson DA, McCurley KS. Surfing the web backwards. Proc 8th International World Wide Web Conference, p 1679–1693, 1999.
24. Kraaij W, Westerveld T. TNO/UT at TREC-9: How different are Web documents? http://trec.nist.gov/pubs/trec9/papers/tno-ut.pdf, 2000.
25. Crivellari F, Melucci M. Web document retrieval using passage retrieval, connectivity information, and automatic link weighting-TREC-9 report. http://trec.nist.gov/pubs/trec9/papers/jhuapl.pdf, 2000.
26. Clake CLA, Cormack GV, Kisman DIE, Lynam TR. Question answering by passage selection (MultiText experiments for TREC-9). http://trec.nist.gov/pubs/trec9/papers/mt9.pdf, 2000.
27. Fujita S. Reflections on "Aboutness" TREC-9 evaluation experiments at Justsystem. http://trec.nist.gov/pubs/trec9/papers/ jscbt9w_paper.pdf, 2000.
28. Savoy J, Rasolofo Y. Report on the TREC-9 experiment: Link-based retrieval and distributed collections. http://tree.nist.gov/pubs/trec9/papers/unine9.pdf, 2000.
29. Singhal A, Kaszkiel M. AT&T at TREC-9. http://trec.nist.gov/pubs/trec9/papers/att-trec9.pdf, 2000.
30. McNamee P, Mayfield J, Piatko C. The HAIRCUT system at TREC-9. http://trec.nist.gov/pubs/trec9/papers/jhuapl.pdf, 2000.

AUTHORS (from left to right)



Kazunari Sugiyama (student member) graduated in 1998 with a specialty in computer science from the Department of Engineering at Yokohama National University, completed the first half of his computer science doctoral course in 2000, and joined KDD Corp. (currently, KDDI) where he worked until 2001. In 2004, he completed the second half of his doctoral course at Nara Institute of Science and Technology, receiving a Ph.D. degree in engineering. He then joined Hitachi, Ltd., Software Division, and is engaged in research on information retrieval. He is a member of the Information Processing Society of Japan, the Japanese Society for Artificial Intelligence, the Association for Computing Machinery, the American Association for Artificial Intelligence, and IEEE.

Kenji Hatano (member) graduated in 1995 with a specialty in precision machinery from the Department of Engineering at Kobe University, completed the second half of his doctoral course in 1999, and became an assistant in the Graduate School of Information Science at Nara Institute of Science and Technology. He is engaged in research on XML databases and information retrieval. He holds a Ph.D. degree in engineering, and is a member of the Information Processing Society of Japan, the Association for Computing Machinery, and the IEEE Computer Society.

Masatoshi Yoshikawa (member) graduated in 1980 with a specialty in information science from the Department of Engineering at Kyoto University, completed the second half of his doctoral course in 1985, and became a lecturer at Kyoto Sangyo University. After serving as an assistant professor at Sangyo University and at Nara Institute of Science and Technology, he has been a professor in the Information Technology Center at Nagoya University since 2002. He is engaged in research on XML databases and multidimensional space indexes. He holds a Ph.D. degree in engineering, and is a member of the Information Processing Society of Japan, the Association for Computing Machinery, and the IEEE Computer Society.

Shunsuke Uemura (member, fellow) graduated in 1964 with a specialty in electronics from the Department of Engineering at Kyoto University, completed his master's course in 1966, and joined the Electrical Testing Institute of the Ministry of International Trade and Industry (MITI) National Research Institutes (currently, the Industrial Technology Research Institute). In 1988, he became a professor of mathematics and computer science in the Department of Engineering at Tokyo University of Agriculture and Technology. He has been a professor in the Graduate School of Information Science at Nara Institute of Science and Technology since 1993. In 1970–1971, he was a visiting researcher at Massachusetts Institute of Technology. He is engaged in research on database systems, natural language processing, and programming languages. He holds a Ph.D. degree in engineering, and is a fellow of IEEE and the Information Processing Society of Japan and a member of the Association for Computing Machinery.