

Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages

Kazunari Sugiyama
Nara Institute of Science
and Technology
8916-5 Takayama, Ikoma,
Nara 630-0192, Japan
kazuna-s@is.aist-
nara.ac.jp

Kenji Hatano
Nara Institute of Science
and Technology
8916-5 Takayama, Ikoma,
Nara 630-0192, Japan
hatano@is.aist-
nara.ac.jp

Masatoshi Yoshikawa
Nagoya University
Furo, Chikusa, Nagoya,
Aichi 464-8601, Japan
yosikawa@itc.nagoya-
u.ac.jp

Shunsuke Uemura
Nara Institute of Science
and Technology
8916-5 Takayama, Ikoma,
Nara 630-0192, Japan
uemura@is.aist-
nara.ac.jp

ABSTRACT

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 the contents of their hyperlinked neighboring pages. In this paper, we first propose several approaches to refining 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 approaches. Experimental results show that, generally, more accurate feature vectors of a target Web page can be generated in the case of utilizing the contents of its hyperlinked neighboring pages at levels up to second in the backward direction from the target page.

Categories and Subject Descriptors

H.3.1 [Content Analysis and Indexing]: Indexing methods; H.5 [Information Interfaces and Presentation]: Hypertext/Hypermedia

General Terms

Algorithms, Performance, Experimentation

Keywords

WWW, Information retrieval, TF-IDF scheme, Hyperlink

1. INTRODUCTION

The WWW (World Wide Web) is a useful resource for users to obtain a great variety of information. Three billion Web pages are the lower bound that comes from the coverage of search engines [3], and it is obvious that the number of Web pages continues to grow. Therefore, it is getting more and 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 effectively, and they are classified into two generations based on their indexing techniques [6]. In the first-generation search engines developed in the early stages of the Web, only terms included in Web pages were utilized as indexes. Therefore, the traditional document retrieval technique was merely applied to Web page search. However, Web pages have peculiar features such as hyperlink structures or are in numbers too great to search effectively. Consequently, users are not satisfied with ease of use and retrieval accuracy of the search engines because such features of Web pages are not exploited in the first-generation search engines.

To deal with these problems, in the second-generation search engines, the hyperlink structures of Web pages are taken into account. For example, the approaches called PageRank [18] and HITS (Hypertext Induced Topic Search) [14] are applied to Google¹ [5] and the search engine of the CLEVER project [13], respectively. 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 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. Taking these points into account, Davison [10] concentrated on textual content and showed that Web pages are significantly more likely to be related topically to pages to which they are linked. Based on this finding, his research group has released the search engine “Teoma²” that does context-sensitive HITS on the lines of the CLEVER project. This search engine uses the concept of “Subject-Specific Popularity [1],” which ranks a site based on the number of same-subject pages that reference it, not just general popularity, to determine a site’s level of authority. However, the problem of Web pages irrelevant to a user’s query often being ranked highly still remains. Hence, 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. In order to achieve this purpose, we have proposed some methods for improving a feature vector for a target Web page [23]. Our proposed methods, however, also have a problem in that only Web pages out-linked from a target Web page are used in order to generate feature vector of the target Web page. Since Web pages usually has their in- and out- linked pages, in the case of generating more accurate feature vector of a Web page, it is necessary to use both its in- and out- linked pages. Therefore, in this paper, we first propose three approaches to refining the tf-idf scheme [22] for a target

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

HT’03, August 26–30, 2003, Nottingham, United Kingdom.
Copyright 2003 ACM 1-58113-704-4/03/0007 ...\$5.00.

¹<http://www.google.com/>

²<http://www.teoma.com/>

Web page using both its in- and out-linked pages in order to represent the contents of the target Web page more accurately. Then, we compare retrieval accuracy of our proposed approaches using the refined feature vector. Our approach is novel in refining tf-idf based feature vectors of target Web pages by reflecting the contents of their hyperlinked neighboring Web pages.

The remainder of this paper is organized as follows: In Section 2, we review related work using hyperlink structures of the Web. In Section 3, we propose novel approaches to refining feature vectors for Web pages by using their hyperlinked neighboring pages. In Section 4, we present the experimental results for evaluating our proposed methods. Finally, we conclude the paper with a summary and directions for future work in Section 5.

2. RELATED WORK

Hyperlink structures are one of the features of Web pages. Users can navigate the huge Web space easily through this hyperlink structures; therefore, many researches on Web IR have been focusing on the Web’s hyperlink structures. In this section, we review related work of IR systems using the Web’s hyperlink structures, especially the systems based on the concept of “optimal document granularity” and the two most popular Web page weighting algorithms, *HITS* and *PageRank*.

2.1 IR Systems based on the Concept of “Optimal Document Granularity”

With respect to this research area, we refer to the following works. Tajima et al. [25] presented a technique which uses “cuts” (results of Web structure analysis) as retrieval units for the Web. Moreover, they extended to rank search results involved multiple keywords by (1) finding minimal subgraphs of links and Web pages including all keywords; and (2) computing the score of each subgraph based on locality of the keywords within it [24]. Following these works, Li et al. [16] introduced the concept of “information unit,” which can be regarded as a logical document consisting of multiple physical Web pages as one atomic retrieval unit, and proposed a novel framework for document retrieval by information units. However, these approaches require considerable processing time to analyze hyperlink structures and to discover the semantics of Web pages. In addition, while these systems can find retrieval units exactly, it often arises that they find retrieval units irrelevant to the user specified query terms. As for these works, we do not believe that users could understand the search results intuitively, because these systems return the search results where the multiple query keywords disperse in several hyperlinked Web pages.

2.2 HITS Algorithm

Kleinberg [14] originally developed HITS algorithm applied to the Web search engine in the CLEVER project [13]. This algorithm depends on the query and considers the set of pages S that point to or are redirected by pages in the answer. Web pages that have many links pointing to them in S are called *authorities*, while Web pages that have many outgoing links are called *hubs*. That is to say, 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 page p . These scores are defined such that the following equations are satisfied for all pages p :

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

where $H(p)$ and $A(p)$ are normalized for all Web pages. These

scores can be determined through an iterative algorithm, and they converge to the principal eigenvector of the link matrix of S .

Several researchers have extended the above original HITS algorithm. For instance, the *ARC* algorithm of Chakrabarti et al. [8] enhanced the HITS algorithm with textual analysis. ARC computes a distance-2 neighborhood graph and weights edges. The weight of each edge is based on the match between the query terms and the text surrounding the hyperlink in the source document. In [4], Bharat et al. introduced additional heuristics to HITS algorithm by giving a document an authority weight of $1/k$ if the document is in a group of k documents on a first host which link to a single document on a second host, and a hub weight of $1/l$ if there are l links from the document on a first host to a set of documents on a second host. However, the major problem in this technique is that the Web pages that the root document point 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, most of which are not very relevant to the query. In order to solve this problem, Li et al. [15] proposed a new weighted HITS-based method that assigns appropriate weights to in-links of root Web pages and combined content analysis with HITS-based algorithm. In addition, Chakrabarti et al. [7, 9] considered not merely the text of Web page but also the Document Object Model (DOM) within the HTML. This improves on intermediate work in the CLEVER project [13] that broke hubs into pieces at logical HTML boundaries.

2.3 PageRank Algorithm

PageRank 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 Web page can be computed. This value is then used as a part of the ranking mechanism used by Google [5]. 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 the 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 0.15-0.2 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. To yield more accurate search results, Rafiei et al. [20] proposed using the set of Web pages that contain some terms as a bias set for influencing the PageRank computation, with the goal of returning terms for which a given page has a high reputation. Furthermore, Haveliwala [11] proposed computing a set of PageRank vectors to capture more accurately the notion of importance with respect to a particular topic.

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 [14]

and PageRank [5] also have problems: (1) the weight for a Web page is merely defined; and (2) the relativity of contents among hyperlinked Web pages is not considered.

On the basis of these problems, the feature vector of a Web page should be generated by using the contents of its hyperlinked neighboring pages in order to represent the contents of Web pages more accurately. We, therefore, propose refining the tf-idf scheme for a target Web page by using the contents of its hyperlinked neighboring pages. Unlike researches described in the previous section, our approach is novel in refining tf-idf based feature vector of a target Web page by reflecting the contents of its hyperlinked neighboring Web pages. Our approach is query-independent, and link-based computations are performed offline as well as PageRank algorithm. At query time, we only compute the similarity between the refined feature vector and user specified query. Therefore, the query-time costs are not much greater than HITS algorithm whose query-time costs depends on the query.

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 ($p_{i1}, p_{i2}, \dots, p_{iN_i}$) in the i^{th} level from p_{tgt} . Moreover, we denote the feature vector $\mathbf{w}^{p_{tgt}}$ of p_{tgt} as follows:

$$\mathbf{w}^{p_{tgt}} = (w_{t_1}^{p_{tgt}}, w_{t_2}^{p_{tgt}}, \dots, w_{t_m}^{p_{tgt}}), \quad (1)$$

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

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

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 $\mathbf{w}^{p_{tgt}}$ as the ‘‘initial feature vector.’’ Subsequently, we denote the refined feature vector $\mathbf{w}'^{p_{tgt}}$ as follows:

$$\mathbf{w}'^{p_{tgt}} = (w_{t_1}'^{p_{tgt}}, w_{t_2}'^{p_{tgt}}, \dots, w_{t_m}'^{p_{tgt}}),$$

and refer to this $\mathbf{w}'^{p_{tgt}}$ as the ‘‘refined feature vector.’’ In this paper, we propose three approaches to refining the ‘‘initial feature vector’’ based on the tf-idf scheme defined by Equation (2) as follows:

- Method I** the approach relies on 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} ,
- Method II** the approach relies on 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} ,
- Method III** the approach relies on 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} .

Method I

In this approach, 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 Web pages 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 $\mathbf{w}^{p_{tgt}}$ and feature vector of in- and out-linked pages of

p_{tgt} in the vector space on each element of initial feature vector $\mathbf{w}^{p_{tgt}}$. For example, Figure 1(a) shows that $\mathbf{w}'^{p_{tgt}}$ 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 $\mathbf{w}^{p_{tgt}}$. 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. In addition, Figure 1(b) shows that refined feature vector $\mathbf{w}'^{p_{tgt}}$ is generated by reflecting each feature vector of in- and out-linked pages of p_{tgt} on the initial feature vector $\mathbf{w}^{p_{tgt}}$. In this approach, each element $w_{t_k}'^{p_{tgt}}$ of $\mathbf{w}'^{p_{tgt}}$ is defined as follows:

$$\begin{aligned} w_{t_k}'^{p_{tgt}} &= w_{t_k}^{p_{tgt}} \\ &+ \frac{1}{Dim} \left(\sum_{i=1}^{L_{(in)}} \sum_{j=1}^{N_{i(in)}} \frac{w_{t_k}^{p_{ij(in)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(in)}})} \right) \\ &+ \frac{1}{Dim} \left(\sum_{i=1}^{L_{(out)}} \sum_{j=1}^{N_{i(out)}} \frac{w_{t_k}^{p_{ij(out)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(out)}})} \right). \quad (3) \end{aligned}$$

Equation (3) shows that the product of $w_{t_k}^{p_{ij(in)}}$ (weight of term t_k in Web page $p_{ij(in)}$) and the reciprocal of $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(in)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{p_{ij(in)}}$ 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 $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(out)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{p_{ij(out)}}$ in the vector space) is added to $w_{t_k}^{p_{tgt}}$ (weight of term t_k in p_{tgt} , computed by Equation (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} . If the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{p_{ij(in)}}$, $\mathbf{w}^{p_{ij(out)}}$ in the vector space is very close, the values of the second and third terms of Equation (3) 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. $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(in)}})$ and $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(out)}})$ are defined the following equations, respectively:

$$\begin{aligned} dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(in)}}) &= \sqrt{\sum_{k=1}^m (w_{t_k}^{p_{tgt}} - w_{t_k}^{p_{ij(in)}})^2}, \\ dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{p_{ij(out)}}) &= \sqrt{\sum_{k=1}^m (w_{t_k}^{p_{tgt}} - w_{t_k}^{p_{ij(out)}})^2}. \end{aligned}$$

Method II

In this approach, 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 $\mathbf{w}'^{p_{tgt}}$ by reflecting centroid vectors of clusters generated from $G_{i(in)}$ and $G_{i(out)}$ on initial feature vector $\mathbf{w}^{p_{tgt}}$. This approach 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 $\mathbf{w}^{p_{tgt}}$ and the centroid vectors of the clusters in the vector space on each element of the initial feature vector $\mathbf{w}^{p_{tgt}}$. In other words, we first create Web page groups $G_{i(in)}$ and $G_{i(out)}$ defined as follows:

$$G_{i(in)} = \{p_{i1(in)}, p_{i2(in)}, \dots, p_{iN_i(in)}\}, \quad (4)$$

$$\begin{aligned} G_{i(out)} &= \{p_{i1(out)}, p_{i2(out)}, \dots, p_{iN_i(out)}\}, \quad (5) \\ &(i = 1, 2, \dots, L), \end{aligned}$$

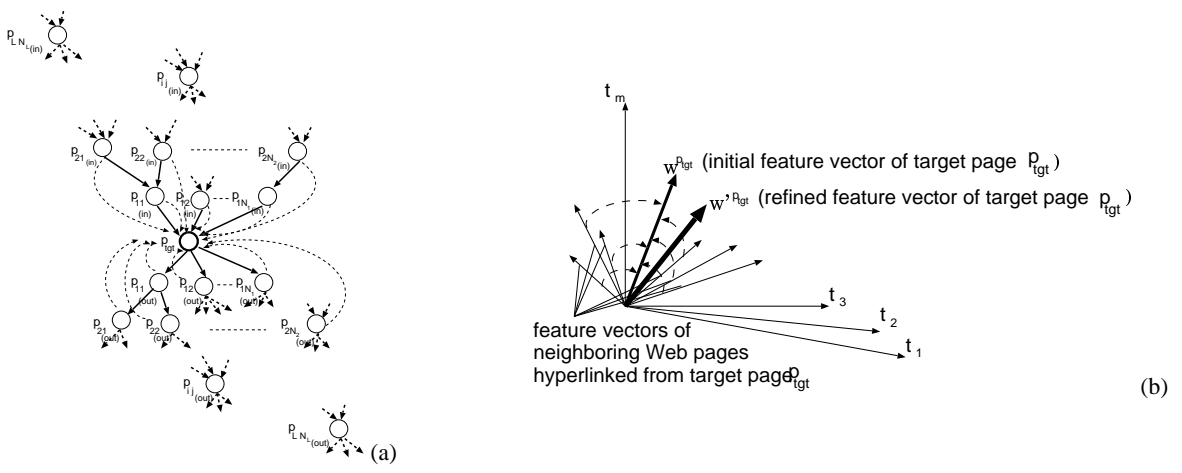


Figure 1: The refinement of a feature vector as performed by Method I [(a) in the Web space, (b) in the vector space].

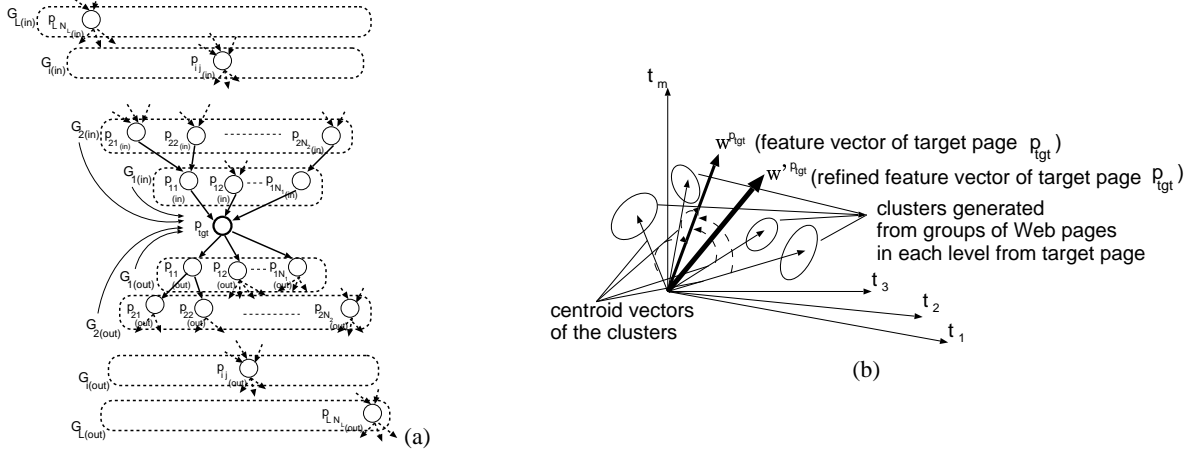


Figure 2: The refinement of a feature vector as performed by Method II [(a) in the Web space, (b) in the vector space].

and then produce K clusters in each Web page group $G_{i(in)}$ and $G_{i(out)}$ by means of the K -means algorithm [17]. The centroid vectors $\mathbf{w}^{g_{ic(in)}}$ and $\mathbf{w}^{g_{ic(out)}}$ ($c = 1, 2, \dots, K$) are produced in $G_{i(in)}$ and $G_{i(out)}$, respectively. We generate a refined feature vector $\mathbf{w}^{p_{tgt}}$ by reflecting the distance between each centroid vector, $\mathbf{w}^{g_{ic(in)}}$, $\mathbf{w}^{g_{ic(out)}}$ and the initial feature vector $\mathbf{w}^{p_{tgt}}$ on $\mathbf{w}^{p_{tgt}}$. 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 directions from p_{tgt} , and generate a refined feature vector $\mathbf{w}^{p_{tgt}}$ 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 $\mathbf{w}^{p_{tgt}}$. Moreover, Figure 2(b) shows that refined feature vector $\mathbf{w}'^{p_{tgt}}$ is generated by reflecting centroid vectors of each cluster on $\mathbf{w}^{p_{tgt}}$. In this approach, we define each element $w_{t_k}^{p_{tgt}}$ of $\mathbf{w}^{p_{tgt}}$ as follows:

$$w_{t_k}^{p_{tgt}} = w_{t_k}^{p_{tgt}} + \frac{1}{Dim} \left(\sum_{i=1}^{L(in)} \sum_{c=1}^K \frac{w_{t_k}^{g_{ic(in)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(in)}})} \right) + \frac{1}{Dim} \left(\sum_{i=1}^{L(out)} \sum_{c=1}^K \frac{w_{t_k}^{g_{ic(out)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(out)}})} \right). \quad (6)$$

Equation (6) shows that the product of $w_{t_k}^{g_{ic(in)}}$ (weight of term t_k in centroid vector $\mathbf{w}^{g_{ic(in)}}$ of cluster c constructed from $G_{i(in)}$) and the reciprocal of $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(in)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{g_{ic(in)}}$ in the vector space), and similarly, the product of $w_{t_k}^{g_{ic(out)}}$ (weight of term t_k in centroid vector $\mathbf{w}^{g_{ic(out)}}$ of cluster c constructed from $G_{i(out)}$) and the reciprocal of $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(out)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{g_{ic(out)}}$ in the vector space) are added to $w_{t_k}^{p_{tgt}}$ (weight of term t_k in p_{tgt} , computed by Equation (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, 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 (6). $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(in)}})$ and $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{ic(out)}})$ are defined as follows:

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

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

Method III

This approach 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 $\mathbf{w}^{p_{tgt}}$ by reflecting centroid vectors of the clusters on the initial feature vector $\mathbf{w}^{p_{tgt}}$. Furthermore, we reflect the distance between $\mathbf{w}^{p_{tgt}}$ and the centroid vector of the cluster in the vector space on each element of $\mathbf{w}^{p_{tgt}}$; in other words, we create Web page groups $G_{i(in)}$ and $G_{i(out)}$ as defined by Equation (7) and (8), respectively,

$$G_{i(in)} = \{p_{11(iin)}, p_{12(iin)}, \dots, p_{1N_{1(iin)}}, p_{21(iin)}, p_{22(iin)}, \dots, p_{2N_{2(iin)}}, p_{i1(iin)}, p_{i2(iin)}, \dots, p_{iN_{i(in)}}\}, \quad (7)$$

$$G_{i(out)} = \{p_{11(out)}, p_{12(out)}, \dots, p_{1N_{1(out)}}, p_{21(out)}, p_{22(out)}, \dots, p_{2N_{2(out)}}, p_{i1(out)}, p_{i2(out)}, \dots, p_{iN_{i(out)}}\}, \quad (8)$$

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

and produce K clusters in $G_{i(in)}$ and $G_{i(out)}$ by means of the K -means algorithm. The centroid vectors $\mathbf{w}^{g_{c(in)}}$ and $\mathbf{w}^{g_{c(out)}}$ ($c = 1, 2, \dots, K$) are produced in $G_{i(in)}$ and $G_{i(out)}$, respectively. Then, we generate refined feature vector $\mathbf{w}'^{p_{tgt}}$ by reflecting the distance between each centroid vector $\mathbf{w}^{g_{c(in)}}$, $\mathbf{w}^{g_{c(out)}}$ ($c = 1, 2, \dots, K$) and initial feature vector $\mathbf{w}^{p_{tgt}}$ on $\mathbf{w}^{p_{tgt}}$. 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 directions from p_{tgt} , and generate refined feature vector $\mathbf{w}'^{p_{tgt}}$ by reflecting the centroid vectors of clusters produced in Web page group $G_{2(in)}$ and $G_{2(out)}$ on the initial feature vector $\mathbf{w}^{p_{tgt}}$. Furthermore, Figure 3(b) shows that refined feature vector $\mathbf{w}'^{p_{tgt}}$ is generated by reflecting centroid vectors of each cluster on the initial feature vector $\mathbf{w}^{p_{tgt}}$. In this approach, each element $w_{t_k}^{p_{tgt}}$ of $\mathbf{w}^{p_{tgt}}$ is defined as follows:

$$w_{t_k}^{p_{tgt}} = w_{t_k}^{p_{tgt}} + \frac{1}{Dim} \left(\sum_{c=1}^K \frac{w_{t_k}^{g_{c(in)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(in)}})} \right) + \frac{1}{Dim} \left(\sum_{c=1}^K \frac{w_{t_k}^{g_{c(out)}}}{dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(out)}})} \right). \quad (9)$$

Equation (9) shows that the product of $w_{t_k}^{g_{c(in)}}$ (weight of term t_k in centroid vector $\mathbf{w}^{g_{c(in)}}$ of cluster c generated from $G_{i(in)}$) and the reciprocal of $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(in)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{g_{c(in)}}$ in the vector space), and similarly the product of element $w_{t_k}^{g_{c(out)}}$ (weight of term t_k in centroid vector $\mathbf{w}^{g_{c(out)}}$ of cluster c generated from $G_{i(out)}$) and the reciprocal of $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(out)}})$ (the distance between $\mathbf{w}^{p_{tgt}}$ and $\mathbf{w}^{g_{c(out)}}$ in the vector space) are added to $w_{t_k}^{p_{tgt}}$ (weight of term t_k in p_{tgt} computed by Equation (2)), with respect to the number of clusters K . As mentioned in Method I and II, in order to prevent the value of the second and third terms of equation (9) 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 $dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(in)}})$ and

$dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(out)}})$ as follows:

$$dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(in)}}) = \sqrt{\sum_{k=1}^m (w_{t_k}^{p_{tgt}} - w_{t_k}^{g_{c(in)}})^2}, \quad (10)$$

$$dis(\mathbf{w}^{p_{tgt}}, \mathbf{w}^{g_{c(out)}}) = \sqrt{\sum_{k=1}^m (w_{t_k}^{p_{tgt}} - w_{t_k}^{g_{c(out)}})^2}. \quad (11)$$

4. EXPERIMENT

4.1 Experimental Setup

We conducted experiments in order to verify the retrieval accuracy of our three approaches described in Section 3. They were implemented using Perl on a workstation (CPU: UltraSparc-II 480 MHz×4, Memory: 2 GBytes, OS: Solaris 8), and the experiments were conducted using the TREC WT10g test collection [12], which contains about 1.69 million Web pages. Stop words were eliminated from all Web pages in the collection based on the stopword list³ and stemming was performed using Porter Stemmer⁴ [19]. We formulated query vector \mathbf{Q} using the terms included in the “title” field in Topics from 451 to 500 at the TREC WT10g test collection. This query vector \mathbf{Q} is denoted as follows:

$$\mathbf{Q} = (q_{t_1}, q_{t_2}, \dots, q_{t_m}), \quad (12)$$

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

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

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 [21], Equation (13) is the element of a query vector that brings the best search result. We then compute the similarity $sim(\mathbf{w}'^{p_{tgt}}, \mathbf{Q})$ between refined feature vector $\mathbf{w}'^{p_{tgt}}$ and query vector \mathbf{Q} . The $sim(\mathbf{w}'^{p_{tgt}}, \mathbf{Q})$ is defined as follows:

$$sim(\mathbf{w}'^{p_{tgt}}, \mathbf{Q}) = \frac{\mathbf{w}'^{p_{tgt}} \cdot \mathbf{Q}}{|\mathbf{w}'^{p_{tgt}}| \cdot |\mathbf{Q}|}. \quad (14)$$

Based on the value of $sim(\mathbf{w}'^{p_{tgt}}, \mathbf{Q})$, we evaluate retrieval accuracy using “precision at 11 standard recall levels” described in [26, 2].

4.2 Experimental Results

Method I

We generated refined feature vector $\mathbf{w}'^{p_{tgt}}$ for initial feature vector $\mathbf{w}^{p_{tgt}}$ of target page p_{tgt} with respect to the following three cases:

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

³ftp://ftp.cs.cornell.edu/pub/smart/english.stop

⁴http://www.tartarus.org/%7Emartin/PorterStemmer/

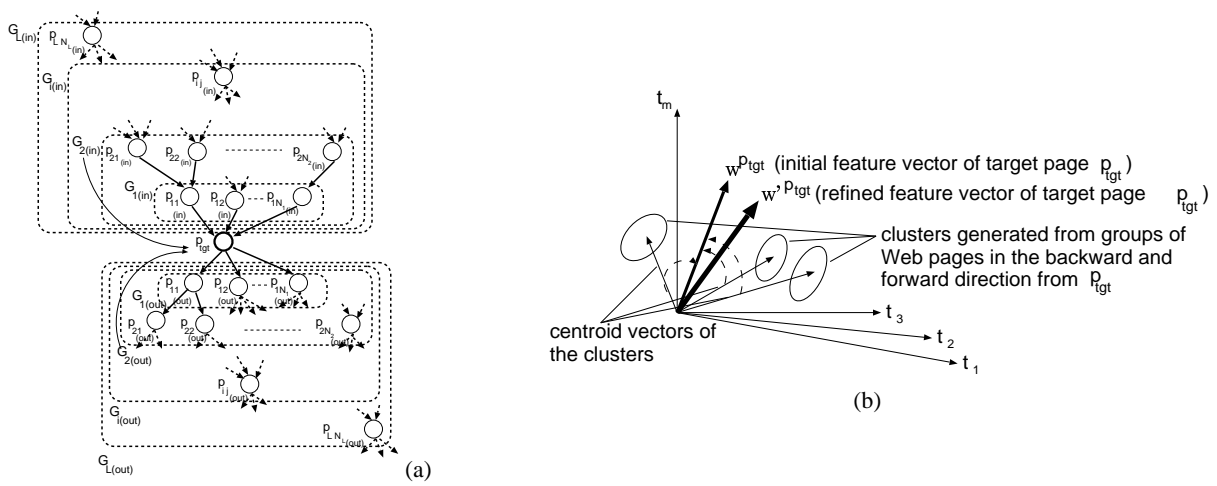


Figure 3: The refinement of a feature vector as performed by Method III [(a) in the Web space, (b) in the vector space].

(MI-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 the initial feature vector $w^{p_{tgt}}$.

Using these refined feature vectors, we verified the retrieval accuracy. Figure 4 shows the experimental results of (MI-a), (MI-b), and (MI-c). In these figures, the value of $L_{(in)}$ and $L_{(out)}$ represents the maximum value of i , which denotes the number of the shortest directed path from p_{tgt} as Figure 1(a) illustrates. As shown in (MI-a) and (MI-b), we found that we can not obtain much improvement in retrieval accuracy compared with the tf-idf scheme when using only in-linked pages or only out-linked pages of the target page p_{tgt} . In Figure 4(MI-a) and (MI-b), the average improvement ratio is +0.02 compared with the tf-idf scheme. On the other hand, in the case of reflecting the contents of all Web pages up to first in the backward and forward directions from p_{tgt} on $w^{p_{tgt}}$, as Figure 4(MI-c) shows, we can obtain slight improvement in retrieval accuracy (only +0.007 improvement). However, Figure 4(MI-c) illustrates that, when reflecting the contents of all Web pages at levels up to second or third in the backward and forward directions from p_{tgt} on $w^{p_{tgt}}$, we can obtain much better retrieval accuracy compared with the tf-idf scheme (+0.028, +0.029 improvement, respectively). Therefore, in general, we found that we can generate a more accurate feature vector of Web page by reflecting the contents of all Web pages at levels up to at least second in the backward and forward directions from the target page p_{tgt} .

Method II

We generated refined feature vector $w^{p_{tgt}}$ for initial feature vector $w^{p_{tgt}}$ of target page p_{tgt} with respect to the following three cases:

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

We then conducted experiments to verify the retrieval accuracy using these refined feature vectors. Figures 5, 6 and 7 show the experimental results of (MII-a), (MII-b), and (MII-c), respectively. In these figures, the value of K means the number of clusters produced by Web page groups denoted by Equation (4), and (5).

In regard to (MII-a), we could obtain the best retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to first in the backward direction from p_{tgt} , compared with the tf-idf scheme (Figure 5, $L_{(in)} = 1$, +0.027 improvement). However, we could not obtain notable retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to second (Figure 5, $L_{(in)} = 2$) or third (Figure 5, $L_{(in)} = 3$) in the backward direction from a p_{tgt} , compared with the tf-idf scheme (at best, +0.011, +0.018 improvement, respectively).

Regarding (MII-b), as well as (MII-a), we could obtain the best retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to first in the forward direction from p_{tgt} , compared with the tf-idf scheme (Figure 6, $L_{(out)} = 1$, +0.024 improvement). However, like (MII-a), we could not obtain notable retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to second (Figure 6, $L_{(out)} = 2$) or third (Figure 6, $L_{(out)} = 3$) in the forward direction from p_{tgt} (at best, +0.020, +0.016 improvement, respectively).

Furthermore, regarding (MII-c), just as with (MII-a) and (MII-b), we could obtain the best retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to first in the backward and forward directions from p_{tgt} , compared with the tf-idf scheme (Figure 7, $L_{(in)} = 1, L_{(out)} = 1$, +0.024 improvement). However, just as with (MII-a) and (MII-b), neither could we obtain remarkable retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at each level up to second (Figure 7, $L_{(in)} = 2, L_{(out)} = 2$, at best, +0.012 improvement) or third (Figure 7, $L_{(in)} = 3, L_{(out)} = 3$, at best, +0.012 improvement) in the backward and forward directions from p_{tgt} .

Based on the findings described above, we found that, with regard to the contents of Web page, there is strong similarity between the feature vector of the target page p_{tgt} and the centroid vector generated by Web page groups at each level up to first in the backward and forward 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} reduces as the value of i , which denotes the number of the shortest directed path from p_{tgt} , increases.

Method III

We generated refined feature vector for target page p_{tgt} with respect to the following three cases:

- (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 $\mathbf{w}^{p_{tgt}}$,
- (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 $\mathbf{w}^{p_{tgt}}$,
- (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 $\mathbf{w}^{p_{tgt}}$.

We then conducted experiments to verify the retrieval accuracy using these improved feature vectors. Figures 8, 9, and 10 show the experimental results of (MIII-a), (MIII-b), and (MIII-c), respectively. In these figures, the value of K represents the number of clusters generated by Web page groups denoted by Equation (7), and (8).

Regarding (MIII-a), in comparison with the tf-idf scheme, we could obtain the best retrieval accuracy when we generated a refined feature vector that considered the contents of Web pages at levels up to second in the backward direction from p_{tgt} (Figure 8, $L_{(in)} = 2$, +0.032 improvement). Furthermore, when the Web page groups were produced from all Web pages at levels up to second (Figure 8, $L_{(in)} = 2$) or third (Figure 8, $L_{(in)} = 3$) in the backward direction from p_{tgt} , we could obtain the best retrieval accuracy as the value of K equaled three. We infer from this fact that topics of Web page that point to the target page p_{tgt} are generally composed of three topics.

In regard to (MIII-b), we could obtain the best retrieval accuracy when we generated a refined feature vector that considered the contents of all Web pages at levels up to second in the forward direction from p_{tgt} , compared with the tf-idf scheme (Figure 9, $L_{(out)} = 2$, +0.028 improvement). In addition, in the case of (MIII-b), we could obtain the best retrieval accuracy when the value of K equaled three in any case where $L_{(out)} = 1$, $L_{(out)} = 2$, $L_{(out)} = 3$. We can infer from this fact that topics of Web pages that we can follow from p_{tgt} is often composed of three topics.

Finally, in regard to (MIII-c), the best retrieval accuracy is obtained when we generated a refined feature vector that considered the contents of all Web pages at levels up to second in the backward and forward directions from p_{tgt} , compared with the tf-idf scheme (Figure 10, $L_{(in)} = 2, L_{(out)} = 2$, +0.026 improvement). Furthermore, in any case where [$L_{(in)} = 1, L_{(out)} = 1$], [$L_{(in)} = 2, L_{(out)} = 2$], [$L_{(in)} = 3, L_{(out)} = 3$], we could obtain the best retrieval accuracy when the value of K equaled three. Therefore, hyperlinked Web pages in the neighborhood of a target page tend to have about three topics.

As we described above, the best retrieval accuracy is obtained when we generate refined feature vectors considering the contents of all Web pages at levels up to second in the backward direction from p_{tgt} in (MIII-a), and considering the contents of all Web pages at levels up to second in the forward direction from p_{tgt} in (MIII-b). Therefore, we can infer that the result in (MIII-c) is the effect obtained by merging the best results of (MIII-a) and (MIII-b).

4.3 Summary of Experimental Results

In summary, Figure 11 illustrates the comparison of the best retrieval accuracy obtained using the Method I, II, and III described in the previous section. 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 refined 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. In addition, Figure 11 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.

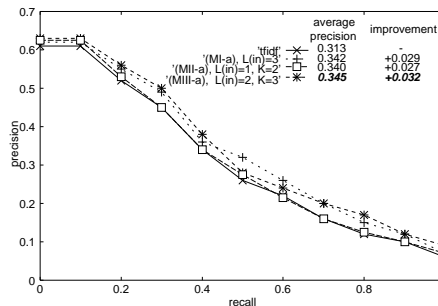


Figure 11: Comparison of the best search accuracy obtained using each Method I, II and III.

5. CONCLUSION

In this paper, in order to represent the contents of Web pages more accurately, we proposed three approaches to refining the tf-idf scheme for Web pages using their hyperlinked neighboring pages. Our approach is novel in refining the 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 approaches:

- the approach relies on 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 approach relies on 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 approach relies on 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 $L_{(in)}^{th}$ in the forward direction from the target page p_{tgt} ,

and evaluated retrieval accuracy of refined feature vector obtained from each approach using recall precision curves. Regarding Method I, we can generate a more accurate feature vector of Web page by utilizing the contents of all Web pages at levels up to at least second in the backward and forward directions from the target page p_{tgt} . In the case of Method II, we found that there is strong similarity between the feature vector of the target page p_{tgt} and the centroid vector generated by the group of Web pages at each level up to first in the backward and forward directions from p_{tgt} . On

the other hand, the similarity between p_{tgt} and the centroid vector generated by the group of Web pages at each level from p_{tgt} reduces as the number of the shortest directed path from p_{tgt} increases. With regard to Method III, a more accurate feature vector of Web page is generated when we use the contents of Web pages at levels up to second in the backward direction from p_{tgt} . Furthermore, compared with respective best retrieval accuracy obtained using these three approaches, 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 enormous Web space. Therefore, we plan to address the technique to provide users with personalized information.

6. REFERENCES

- [1] Ask Jeeves, Inc. Search with Authority: The Teoma Difference. <http://sp.teoma.com/docs/teoma/about/searchwithauthority.html>.
- [2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.
- [3] R. Baeza-Yates, F. Saint-Jean, and C. Castillo. Web Structure, Dynamics and Page Quality. In *Proc. of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002)*, pages 117–130, 2002.
- [4] K. Bharat and M. R. Henzinger. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. In *Proc. of the 21st Annual International ACM SIGIR Conference (SIGIR '98)*, pages 104–111, 1998.
- [5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of the 7th International World Wide Web Conference (WWW7)*, pages 107–117, 1998.
- [6] A. Broder and P. Raghavan. Combining Text- and Link-Based Information Retrieval on the Web. SIGIR'01 Pre-Conference Tutorials, Sep. 2001.
- [7] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. In *Proc. of the 10th International World Wide Web Conference (WWW10)*, pages 211–220, 2001.
- [8] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. In *Proc. of the 7th International World Wide Web Conference (WWW7)*, pages 65–74, 1998.
- [9] S. Chakrabarti, M. Joshi, and V. Tawde. Enhanced Topic Distillation using Text, Markup Tags, and Hyperlinks. In *Proc. of the 23rd Annual International ACM SIGIR Conference (SIGIR 2001)*, pages 208–216, 2001.
- [10] B. D. Davison. Topical Locality in the Web. In *Proc. of the 22nd Annual International ACM SIGIR Conference (SIGIR 2000)*, pages 272–279, 2000.
- [11] T. H. Haveliwala. Topic-Sensitive PageRank. In *Proc. of the 11th International World Wide Web Conference (WWW2002)*, pages 517–526, 2002.
- [12] D. Hawking. Overview of the TREC-9 Web Track. *NIST Special Publication 500-249: The Ninth Text REtrieval Conference (TREC-9)*, pages 87–102, 2001.
- [13] IBM Almaden Research Center. Clever Searching. <http://www.almaden.ibm.com/cs/k53/clever.html>.
- [14] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proc. of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 1998)*, pages 668–677, 1998.
- [15] L. Li, Y. Shang, and W. Zhang. Improvement of HITS-based Algorithms on Web Documents. In *Proc. of the 11th International World Wide Web Conference (WWW2002)*, pages 527–535, 2002.
- [16] W.-S. Li, K. S. 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)*, pages 230–244, 2001.
- [17] J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In *Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.
- [18] L. Page. The PageRank Citation Ranking: Bringing Order to the Web. <http://google.stanford.edu/%7Ebackrub/pageranksub.ps>, 1998.
- [19] M. F. Porter. An Algorithm for Suffix Stripping. *Program*, 14(3):130–137, 1988.
- [20] D. Rafiei and A. O. Mendelzon. What is this Page Known for? Computing Web Page Reputations. In *Proc. of the 9th International World Wide Web Conference*, pages 823–835, 2000.
- [21] G. Salton and C. Buckley. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24(5):513–523, 1988.
- [22] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.
- [23] 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)*, pages 891–901, 2002.
- [24] 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)*, pages 13–23, 1999.
- [25] 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)*, pages 235–244, 1998.
- [26] I. H. Witten and A. M. T. C. Bell. Managing Gigabytes. *Van Nostrand Reinhold*, pages 149–150, 1994.

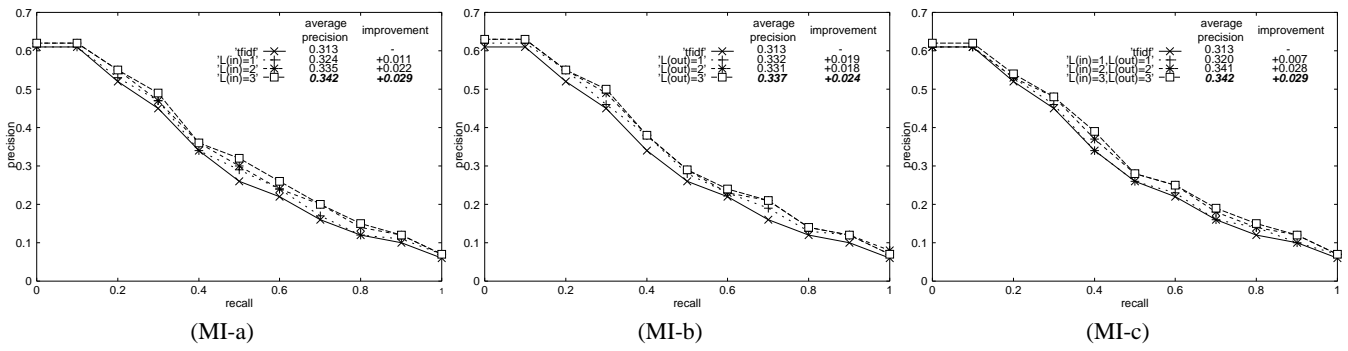


Figure 4: Comparison of search accuracy obtained using Method I [(MI-a): in-link pages only, (MI-b): out-link pages only, (MI-c): both in-link and out-link pages].

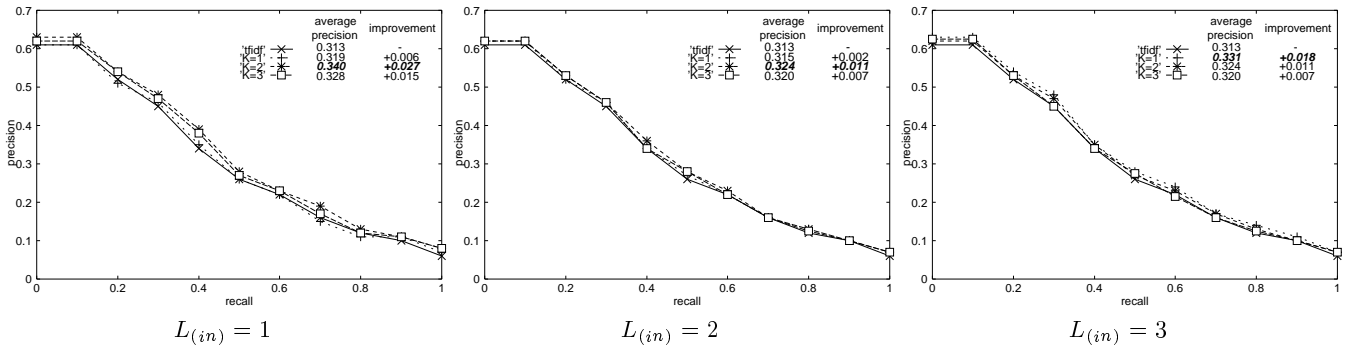


Figure 5: Comparison of search accuracy obtained using Method II [(MII-a): in-link pages only].

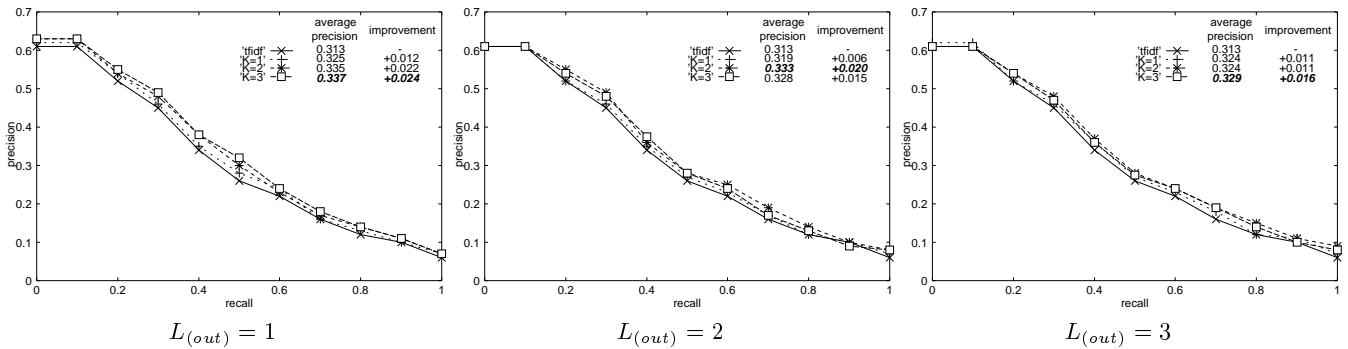


Figure 6: Comparison of search accuracy obtained using Method II [(MII-b): out-link pages only].

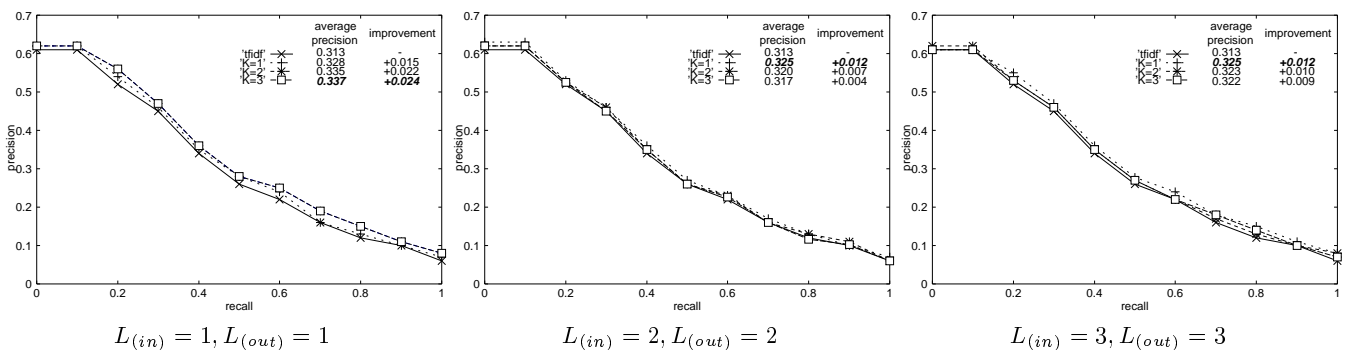


Figure 7: Comparison of search accuracy obtained using Method II [(MII-c): both in-link and out-link pages].

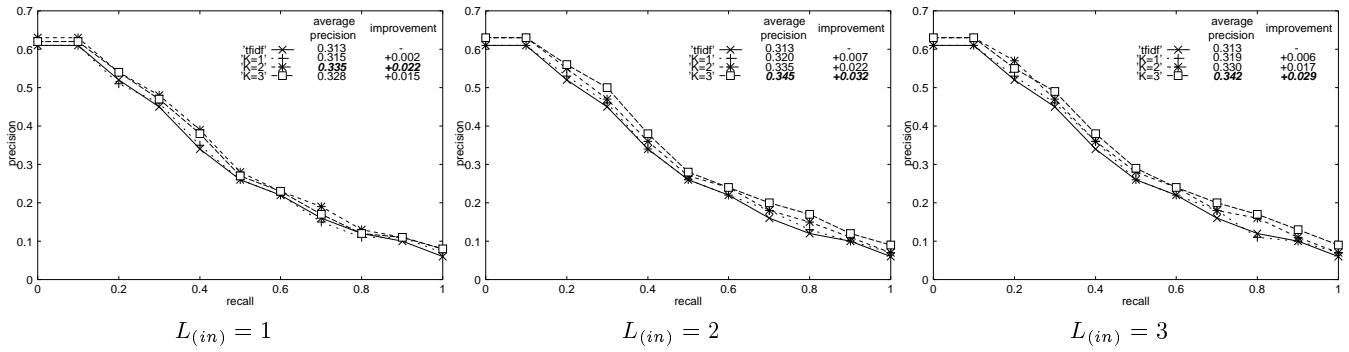


Figure 8: Comparison of search accuracy obtained using Method III [(MIII-a): in-link pages only].

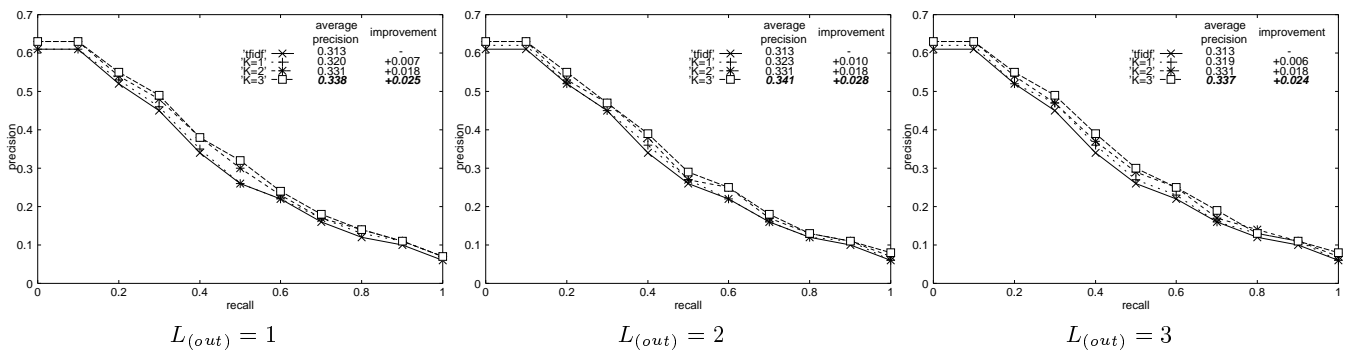


Figure 9: Comparison of search accuracy obtained using Method III [(MIII-b): out-link pages only].

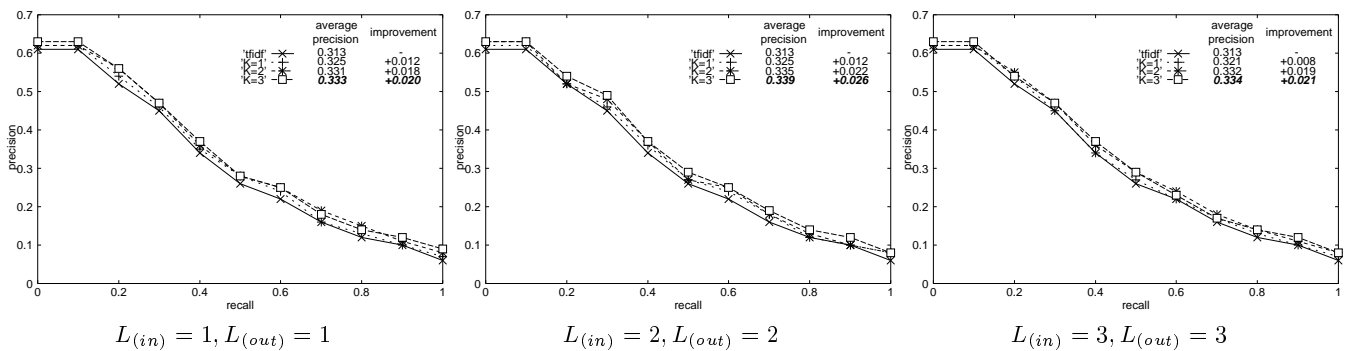


Figure 10: Comparison of search accuracy obtained using Method III [(MIII-c): both in-link and out-link pages].