Web Page Scoring Based on Link Analysis of Web Page Sets

Hitoshi Nakakubo[†], Shinsuke Nakajima[‡], Kenji Hatano[⋆],
Jun Miyazaki[‡], and Shunsuke Uemura[*]

[†] Innovative Technology Development Center, U-TEC Corporation
21–1 Omori, Nara, Nara 630–8131, Japan
hitoshi.nakakubo@u-tc.co.jp
[‡] Graduate School of Information Science, Nara Institute of Science and Technology
Keihanna Science City, Ikoma, Nara 630–0192, Japan
{shin|miyazaki}@is.naist.jp
[⋆] Faculty of Culture and Information Science, Doshisha University
1-3 Tatara-Miyakodani, Kyotanabe, Kyoto 610-0394, Japan
khatano@mail.doshisha.ac.jp
[*] Faculty of Informatics, Nara Sangyo University
3–12–1 Tateno-Kita, Sango, Ikoma, Nara 636-8503, Japan
UemuraShunsuke@nara-su.ac.jp

Abstract

We propose a new Web page scoring method based on the link analysis among sets of Web pages. Conventional link analyses such as PageRank and HITS calculate importance degree of each Web page; however, the authors of Web pages often create multiple pages to describe a specific topic. The importance degrees of such multiple Web pages cannot be derived by the conventional link analyses accurately. To cope with this problem, we need to treat the Web pages with the same contents edited by the same author as a Web page set (WPS). After constructing the link structure among WPSs, we calculate their importance degrees by using conventional link analysis schemes. In this paper, we compared our approach with the conventional method by using the NTCIR test collection, and found that our approach was better than the conventional method in terms of both WRR and DCG evaluation measures.

1 Introduction

Recently, Web search engines have become crucial tools for finding information on the Internet. However, users are not always satisfied with the search results produced by conventional search engines. For example, the Web pages returned from search engines may not contain needed information, even though they surely include query keywords. Therefore, we need to develop a novel search engine which can solve this problem and to satisfy users' information needs.

Link analysis of Web pages is one of the most effective methods to improve the retrieval accuracy of Web search engines. In particular, PageRank and HITS are popular algorithms of the link analysis for Web search engines [2, 8]. The PageRank algorithm produces importance scores of each Web page by analyzing the Web link structure based on a random walk model, while the HITS algorithm calculates the worth of each Web page based on the concepts of *Authorities* and *Hubs*. Both methods analyze the Web link structure, and then, calculate the importance degrees of individual Web pages.

However, Web content about an specific topic created by a single author is often spread over multiple pages, not one. Moreover, some Web contents are designed to be viewed over multiple Web pages in order. Conventional PageRank and HITS algorithms may not correctly generate the importance degrees of the Web pages in such Web page organization, because these algorithms were originally proposed to evaluate individual Web pages, not sets of pages.

In this paper, therefore, we propose a new Web page scoring method based on the link analysis among sets of Web pages with the same contents, called the *Web Page Set*, or *WPS* for short, not among Web pages individually. We believe that our method can accurately calculate the importance degrees of Web page sets, so that it becomes a key

technology to build useful Web search engines.

The remainder of this paper is organized as follows. We first describe related work in Section 2 and explain our proposed method in Section 3. Section 4 shows experimental results to evaluate our method. Finally, Section 5 concludes the paper.

2 Related Work

There has been a lot of research which try to improve the retrieval accuracy of Web search engines by treating multiple Web pages as a set.

Sugiyama et al. tried to improve the retrieval accuracy by modifying the feature vector of a Web page [12]. In order to modify the feature vector, their method first calculates the center of the feature vectors of the multiple Web pages linked from a Web page. The Web pages are, then, clustered based on their feature vectors, taking inter-document similarity into account. However, it takes a long time to calculate the feature vectors.

Masada et al. also tried to improve the retrieval accuracy by using the Web pages belonging to the same Web site [10]. Similar to Sugiyama's approach [12], this method also modifies the feature vector of a Web page by using the clustering result of the multiple Web pages in the same Web sites. However, the modification of the feature vectors is calculated only with hyperlink structures.

Tajima et al. developed a Web search engine which derives search results based on the notion of a *minimal graph* [13]. The minimal graph consists of the multiple Web pages which contain all of the query keywords provided by users. Li et al. also proposed the concept of a *information unit* which is similar to the minimal graph [9]. These studies share a common idea that search results are expressed as sets of Web pages.

As mentioned above, these conventional schemes utilize the link structure among Web pages to modify their feature vectors and to arrange a unit of search results in order to improve the retrieval accuracy of Web search engines. Most people do not doubt that the link analysis methods such as PageRank and HITS algorithms are effective for searching Web pages related to information needs. To solve the problem as described in Section 1, we also try to adopt a link analysis approach. Namely, our method is referred to as a refined version of link analysis because its scores are calculated based not on Web pages but on Web page sets.

3 Scoring Based on a Link Analysis of Web Page Sets

Conventional link analysis methods are considered to be insufficient for calculating the importance degrees of the Web pages correctly when one specific topic is described over several Web pages. Therefore, we first extract Web page sets, each of which represents one specific topic, before applying a conventional link analysis method. We reconstruct the link structure of the simplified graph which consists of the obtained Web page sets as nodes and existing links as edges. With the reconstructed link structure, the importance degrees of Web page sets are calculated.

3.1 Definition of a Web Page Set

Here, we define a *Web Page Set*, or *WPS* for short, which is a unit of our link analysis, as *a set of Web pages that contains only one identical topic written by a unique author*. This definition is based on the following empirical observations.

- *A Web page set is created by a unique author.*
  Topic and quality of a Web page depend on its author. Even if the same topic is described by two authors, the topic and the level of details of the content are totally different.

  For example, a Web content regarding "Database" described by an expert of database systems tends to be more technical and more academic than that by a novice, even if both contents deal with the same topic. If a Web search engine equally treats such Web pages with diverse contents and quality, users cannot accurately retrieve search results because the importance degrees of Web pages are either overestimated or underestimated. For this reason, we treat a set of Web pages created by a unique author as a unit when performing link analysis.

  In addition, our method divides a Web page set into multiple sets according to a boundary of the Web sites. This is based on the assumption that Web pages are managed by each Web site. In this paper, a Web site is regarded as a set of Web page sets which has loosely unified contents and designs as well as one entry page. The concept of a entry page is basically similar to that of a top page; however, the entry page is defined as a unique entry point of a logical domain of Web space, while the top page is the entry point of a physical Web server. Further discussion appears in section 3.2.

- *A Web page set has one identical topic.*
  Web pages created by a unique author do not always have one identical feature. For example, suppose that an author writes Web pages on two topics related to "Database" and "Sports". The levels of author's understanding on both topics are not necessarily the same. In this case, their importance degrees of the Web pages should be calculated individually. Therefore, we also regard a set of Web pages on one identical topic as a unit for link analysis.

Only the set of Web pages that meets the above concepts are called a *Web Page Set*. The way to derive Web page sets is discussed in Section 3.2.

3.2 Scoring based on WPSs

Figure 1 indicates the procedure of our method. The detailed explanation is given below.

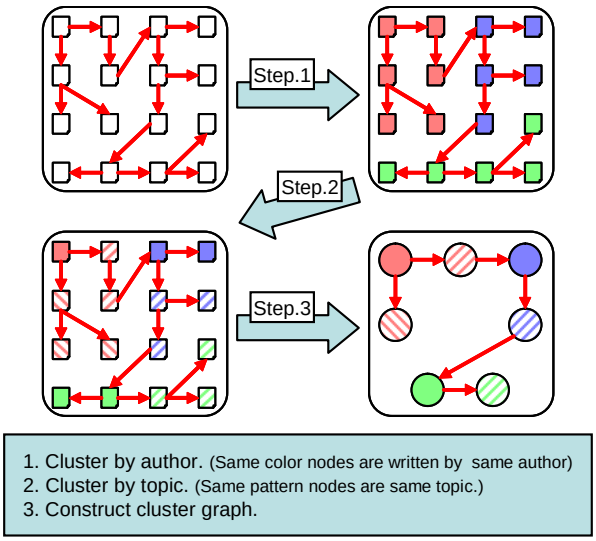

Figure 1. Procedure of Our Method

Step 1. *Determining the boundary of Web sites*
We adopt the method proposed by Ayan et al. [1] to determine the boundaries of Web sites. Their method extracts logical domains, which is identified by semantic relation of Web pages, as opposed to physical domains, which is identified by domain name of Web servers. The followings are the procedures of extracting logical domains of Web pages:

Determining an entry page: The entry page is a representative of WPSs. In order to determine the entry page, their method adds a point to each Web page if the Web page meets ten given requirements like URL strings, titles of the Web pages, anchor texts, and the number of links. The Web page with the highest point is then nominated as a candidate for the entry page.

Determining a boundary of a logical domain: Their method treats the Web pages in a physical domain as a directory tree. Candidates for the entry page and its descendants are regarded as belonging to the same logical domain. This is called the *Path-based Boundary Definition Approach* in [1]. If the number of Web pages included in a logical domain is fewer than a given threshold, the domain is merged into a parent logical domain in the tree. In this paper, we set the threshold to 10 pages.

Step 2. *Determining a field of Web pages*
We make use of a clustering method to determine a field of Web pages based on their contents. The obtained clusters are identical to WPSs. In order to extract the feature vectors of Web pages, our method analyzes them by ChaSen[1], which is a morphological analysis tool, and then, applies TF-IDF as a term-weighting strategy to them. In addition, we adopt Ward's method which is one of the accurate classifiers to cluster the Web pages [6]. For the present, the number of extracted clusters is assumed to be one tenth of the number of Web pages in each logical domain.

Step 3. *Constructing a link structure among WPSs*
Now, we derived WPSs by the above Step 1 and 2. Each of the obtained WPSs is regarded as one. We then construct the link structure of the simplified graph which consists of the WPSs and existing links as follows.

Step 3-1. Delete all links among Web pages within the same WPSs.

Step 3-2. Modify initial and final points of links so that all of the links point to WPSs.

Step 3-3. Delete all duplicate links between any two WPSs.

Step 4. *Calculating importance degrees of WPSs based on PageRank*
The importance degrees of WPSs are calculated by applying PageRank to the link structure obtained in Step 3. We should finally compute the importance degrees of individual Web pages in a WPS; however, we currently assign the same importance degree, i.e., page score, as the WPS to each Web page in it.

Since WPSs are derived by using TF-IDF and clustering methods after constructing logical domains, the contents of each WPS has almost the same topic written by a unique author. Therefore, the quality of the contents in each WPS is stable. However, the results of ranking WPSs by PageRank do not expose which WPSs have better quality but more relevant pages. The judgment of the quality is left to users, because each user has each own familiarity to topics. A user can determine the quality of a WPS by browsing at least one page in the WPS. If the page in the WPS is not sufficient for their needs, the user can proceed the next WPS without seeing the rest of the pages in the WPS.

[1] http://chasen-legacy.sourceforge.jp/

4 Experimental Evaluation

4.1 Experiments

In order to verify the effectiveness of our link analysis approach based on the WPSs, we compared our approach with a link analysis approach based on each Web page. In this experiment, we implemented our approach based on a link structure constructed from WPSs as well as a conventional Web search engine using a link structure constructed from Web pages as done by Google[2]. These two Web search engines were implemented using gram-based indices [11], so that they perform high recall level characteristically. Link structures of Web pages and WPSs in these Web search engines were constructed by using the PageRank algorithm [2].

In the following experimental evaluation, we used large-scale Web page collection, called NW100G-01, which is a test collection of the NTCIR (NII Test Collection for IR Systems) Project[3]. The NTCIR offers a large-scale test collection used to evaluate information access technologies including information retrieval, question answering, text summarization, and so on. The data size of NW100G-01 we used is about 100GB, and it contains about 11 million Web pages. We also used search topics and relevance judgment, called NTCIR-4 WEB Info 1[4]. The NTCIR provides four relevance levels (highly relevant, relevant, partially relevant and irrelevant) for the relevance judgment of contents of the Web pages. Thus, we can evaluate our approach and the conventional one based on two evaluation measures, Weighted Reciprocal Rank (WRR) [3] and Discounted Cumulative Gain (DCG) [4, 5]. Both were used to evaluate Web search engines in the 4th NTCIR Workshop[5]. These kinds of new evaluation measures have also been used in recent IR researches. For example, XCG (eXtended Cumulated Gain) evaluation measure [7] which is similar to DCG is used in the INEX (INitiative for the Evaluation of XML Retrieval) Project[6], which aims to provide a test collection to evaluate retrieval methods using uniform scoring procedures.

Figure 2 and 3 illustrate the comparisons of the retrieval accuracies calculated by using WRR and DCG evaluation measures. Solid and broken lines indicate the retrieval accuracy of our approach and that of the conventional one in these figures, respectively.

According to these results, we found that we could obtain the best retrieval accuracy when using our approach.

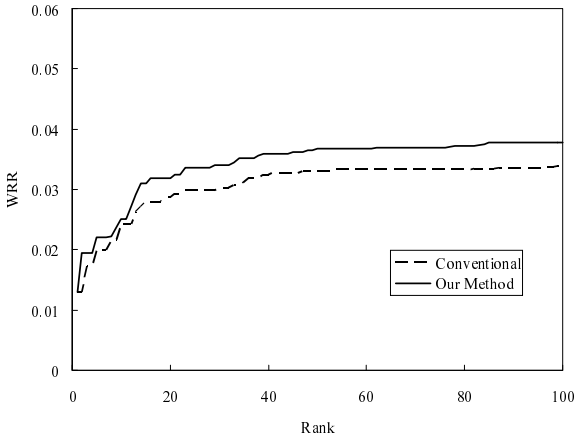

Figure 2. Retrieval Accuracy (WRR)

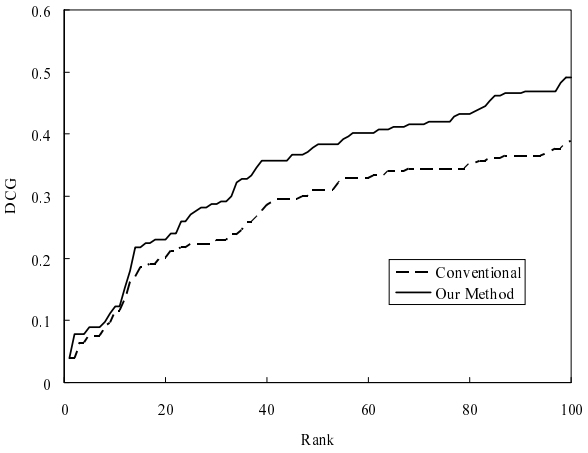

Figure 3. Retrieval Accuracy (DCG)

This is because Figure 2 shows that the rank of the first relevant Web page in our approach is higher than that in the conventional one, and also, Figure 3 reveals that the ranks of relevant Web page in our approach are higher than that in the conventional one.

That is to say, these results show that our approach is more effective than the conventional one in terms of retrieval accuracy. Therefore, we assert that the link analysis approach for Web page retrieval should target not individual Web pages but WPSs, because it allows for more accurate calculation of importance degrees of Web pages.

[2] http://www.google.com/

[3] http://research.nii.ac.jp/ntcir/.

[4] http://research.nii.ac.jp/ntcir/permission/ntcir-4/perm-en-WEB.html

[5] http://research.nii.ac.jp/ntcir/ntcir-ws4/work-en.html.

[6] http://inex.is.informatik.uni-duisburg.de/

4.2 Discussion

The experimental results showed that our method improves the accuracy of Web page retrieval with respect to both WRR and DCG evaluation measures, even when setting the minimum number of Web pages in a WPS, i.e., the threshold of WPSs, to 10. This means that the concept of PageRank based on WPSs has a potential as an accurate Web search method.

However, the current setting of the threshold of WPSs does not seem to be optimal. If the threshold decreases, the behavior of our method approaches the ordinary PageRank, which does not lead to desired results.

On the other hand, if the threshold increases, the number of in- and out-bound links of each WPS also grows. Since we assign the same score as the WPS to each page in it in our current implementation, all page scores may not be differentiated enough when the number of pages in a WPS is too large. In order to avoid this problem, PageRank can be applied inside of WPSs again, so that each page score can be distinguished.

5 Conclusion

In this paper, we proposed a new Web page scoring method based on a link analysis among WPSs, which are defined as sets of Web pages with the same content. We also compared our approach with the conventional scoring method based on link analysis like PageRank algorithm and tested them by using the NTCIR test collection, which is one of the most famous test collections for Web page retrieval. As a result, we found that our approach is more effective than the conventional method in terms of retrieval accuracy using WRR and DCG evaluation measures provided by the NTCIR test collection.

Though we obtained effective results in our approach, our current Web search engine does not take in account the differences of each Web page in the same WPS. It tentatively assigns the same score to each Web page. In order to retrieve Web pages accurately, Web search engines should recognize the differences of each Web page in the same WPS. Therefore, we have to implement an algorithm to recognize the differences. We also found out it is important to define what a WPS is. Retrieval accuracies depend on a method to determine WPSs in our Web search engine. Consequently, the remaining work also includes developing an efficient algorithm to cluster WPSs from large-scale Web page collections, and exploring how to derive an optimal WPS size.

Acknowledgment

This work was partly supported by MEXT (Grant-in-Aid for Scientific Research on Priority Areas #19024058), and JSPS (Grant-in-Aid for Scientific Research (A) #15200010)

References

[1] N. F. Ayan, W.-S. Li, and O. Kolak. Automating Extraction of Logical Domains in A Web Site. *Data & Knowledge Engineering*, 43(2):179–205, Nov. 2002.

[2] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proceedings of the 7th World-Wide Web Conference*, Apr. 1998. `http://www7.scu.edu.au/1921/com1921.htm`.

[3] K. Eguchi, K. Oyama, E. Ishida, N. Kando, and K. Kuriyama. Evaluation Methods for Web Retrieval Tasks Considering Hyperlink Structure. *IEICE Transactions on Information and Systems*, E86-D(9):1804–1813, Sep. 2003.

[4] K. Järvelin and J. Kekäläinen. IR Evaluation Methods for Retrieving Highly Relevant Documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48, July 2000.

[5] K. Järvelin and J. Kekäläinen. Cumulated Gain-based Evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, Oct. 2002.

[6] T. Kamishima. A Survey of Recent Clustering Methods for Data Mining (part 1) - Try Clustering! -. *Journal of the Japanese Society for Artificial Intelligence*, 18(1):59–65, Jan. 2003. (In Japanese).

[7] G. Kazai and M. Lalmas. eXtended Cumulated Gain Measures for the Evaluation of Content-Oriented XML Retrieval. *ACM Transactions on Information Systems*, 24(4):503–542, Oct. 2006.

[8] J. Kleinberg. Authoritative Sources in a Hyperlinked Environment. In *Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, Jan. 1998.

[9] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Retrieving and Organizing Web Pages by "Information Unit". In *Proceedings of the 10th International World Wide Web Conference*, pages 230–244, May 2001.

[10] T. Masada, A. Takasu, and J. Adachi. Link-Based Clustering for Finding Subrelevant Web Pages. In *Proceedings of the Third International Workshop on Web Document Analysis*, Aug. 2005.

[11] T. Sato, T. Satomoto, and K. Han. NTCIR-3 PAT Experiments at Osaka Kyoiku University: Long Gram-based Index and Essential Words. In *Proceedings of the Third NTCIR Workshop*, Jan. 2003.

[12] K. Sugiyama, K. Hatano, M. Yoshikawa, and S. Uemura. Refinement of TF-IDF Schemes for Web Pages using their Hyperlinked Neighboring Pages. In *Proceedings of the 14th Conference on Hypertext and Hypermedia*, pages 198–207, Aug. 2003.

[13] K. Tajima, K. Hatano, T. Matsukura, R. Sano, and K. Tanaka. Discovery and Retrieval of Logical Information Units in Web. In *Proceedings of the 1999 ACM Digital Library Workshop on Organizing Web Space*, pages 13–23, Aug. 1999.