2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Analysis and Improvement of Ranking Algorithm for Web Mining System on the Hierarchical Web Environment

Heebyung Yoon[1], Kil-Seup Lee[1], and Hwa-Soo Kim[2]

[1] National Defense University, Dept. of Computer & Information Science,
205 Susaek-Dong, Eunpyeong-Gu, Seoul, Korea
{hbyoon, gislee}@kndu.ac.kr
[2] Ajou Univ., Graduate School of Information & Communication Technology,
7th FL, Daewoo Foundation Bldg, 526, Namdaemoon-ro 5ga, Chung-gu, Seoul, Korea
ajhskim@ajou.ac.kr

***Abstract* - The variety of document ranking algorithms have developed to provide efficient mining results for user's query on the web environment. The typical ranking algorithms are the Vector-Space Model based on the text, PageRank and HITS algorithms based on the hyperlink structures and other several improvement algorithms. All these are for the user's convenience and preference. However, these algorithms are usually developed on the horizontal and non-hierarchial web environments and are not suitable for the hierarchial web environments such as enterprise and defense networks. Thus, we must consider the special environment factors in order to improve the ranking algorithms. In this paper, we analyze the several typical algorithms used by hyperlink structures on the web environment. We, then suggest a configuration of the hierarchical web environment and also give the relations between agents of the web mining system. Next, we propose an improved ranking algorithm suitable to this kind of special environments. The proposed algorithm is considered both the hyperlink structures of the documents and the location of the user of the hierarchical web.**

I INTRODUCTION

Many researchers have studied on the various document ranking algorithms, which endow priorities on web documents, to provide more proper information to users of web-based internet. These algorithms are classified into several ways according to the experts of web. The most widely-used way is to classify into text-based and hyperlink-based one. Moreover, the latter can be classified into two ways whether it is dependent on query or not.

In general, these document ranking algorithms are usually developed by considering user's preference and convenience. The major points developing a general ranking algorithm, which is suggested in [1], are reputation of the source, page updating frequency, the popularity, degree of authority, degree of hubness, and speed access, and so on. Most of them are related with contents of web documents directly. The last, speed access, is done with only the location of web documents.

The facts above explains followings. Most of ranking algorithms have developed in the environment of horizontal and non-hierarchical web. However, they have not included the environment of hierarchical web with complex functional structure such as enterprise or defense networks. We need more efficient algorithm for the users of hierarchical networks. It means that the location of a document in a hierarchical network can be used as a weighting factor to the priorities of web documents for efficiency, which is separated from the web contents.

For developing an improved ranking algorithm for user's convenience, we organize this paper as follows: In chapter 2, we survey the related works on ranking algorithms. In chapter 3, we explain the hierarchical web-based environment as mentioned above and suggest the whole functionality of web mining system for the hierarchical web. In chapter 4, we propose an improved ranking algorithm on the hierarchical web environment.

II RELATED WORKS

In web mining system, endowing priorities to documents means to consider user's preference and convenience. Almost ranking algorithms have usually used to analyse the contents of documents or hyperlink structures. Among the algorithms, the typical examples are Vector-Space Model[2] based on the text, HITS[3] and PageRank[4] based on the hyperlink structures.

Here hyperlink algorithms can be classified into the query-dependent and the query-independent[5]. The query-dependent cases are HITS, HITS extensions[6],[7] which use the weighted version of the update rule comparing query term with anchor text, and SALSA[8] which uses two matrices as the stochastic approach for link analysis. And the query-independent ones are PageRank, Focused PageRank[9],[10] which compute a relative ranking of pages focusing on a specific topic, and so on. PagaRank algorithm executes approximative and iterative computation to decide the rank for each page, not to do the rank for whole site because of the actual size of site. The equation for the PageRank of page A is given by[4]

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \qquad (1)$$

where $PR(T_i)$ is the PageRank of pages T_i which links to page A, $C(T_i)$ is the number of outbound links on page T_i and d is a damping factor which can be set between 0 and 1.

The Focused PageRank is not an existing horizontal search like the PageRank algorithm used by Google[4], but a vertical (focused) search. However, this algorithm also does not use the location of document, in other words user's location, even though it considers the contents of a document. Recently, an approach has tried by Yoon et al.[11] develop a web mining system considering the location of a document in a hierarchical web environment such as enterprise network. In this paper, four agents cooperating and fourteen functional modules for web mining are suggested as the major components of an efficient web mining system. Moreover, it proposes the methods of the crawling, indexing and merging with respect to a entire network structure. However, the paper does not mentioned about the ranking algorithm which provides documents which are collected by robot agents, to users in an efficient way.

Ⅲ HIERARCHICAL WEB ENVIRONMENT

The conceptual configuration for a hierarchical web environment such as enterprise network or defense network is given as follows. The hierarchical network has a central core node such as headquarters at the center, several regional core nodes executing major tasks within the region under the central core node, functional core nodes doing core functions under the regional node, and branch nodes doing separate functions such as education, management, personnel, and so on. The hierarchical network forms a tree structure.

A hierarchical web structure and its configuration with multi-agents for web mining system is depicted in Figure 1. Here, four agents are shown for a web mining system. 1) User Interface Agent (UIA) analyses user's request information and provides the result of a retrieval to the users. 2) Merge Agent (MA) merges the indexed databases of sub-websites with union function and provides the index words according to the condition of user's request. 3) Index Agent (IA) stores the retrieved documents after indexing to a database and ranks them. 4) Robot Agent (RA) retrieves and collects web documents from web sites using crawlers.

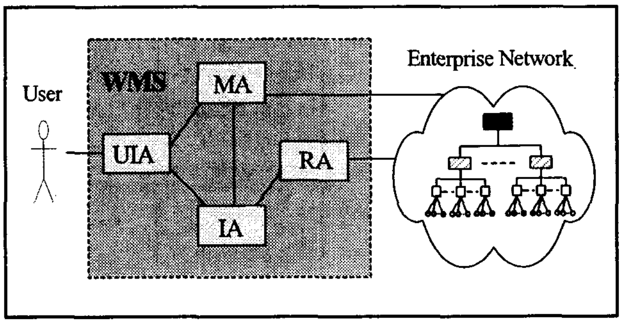

Figure 1. Overall web mining system on the hierarchical and functional web environment

Figure 2 shows the relation between agents of the web mining system and the hierarchical web environment. In this figure, a crawling strategy and merging method are shown for a enterprise network with four layer using some arrow symbols. And three different types of databases are deployed in each layer. A Merged Index DB is just union of Index DBs, Index DB has index words with weighting value according to the order of document identifier, and Ranked DB has document identifier and it score considering the factors of a hierarchical web environment.

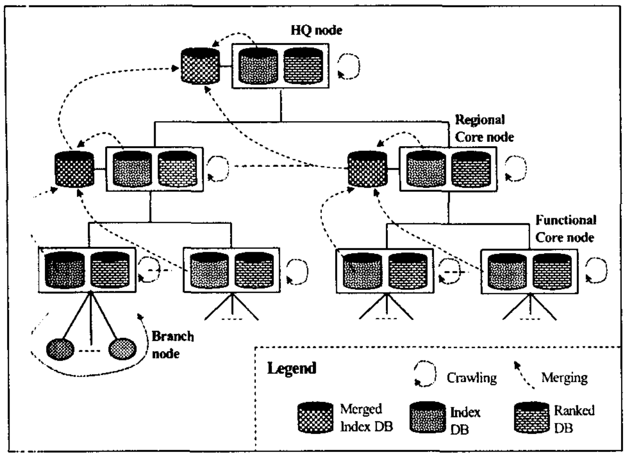

Figure 2. The relation between agents of the web mining system and the hierarchical web environment

IV IMPROVED RANKING ALGORITHM

In this paper we suggest an improved ranking algorithm which applies a new weighting factor of environment considering the location of a document to the existing PageRank algorithm[4], i.e., this algorithm uses the hyperlink structure and the actual location of a document for ranking documents. Figure 3 explains how our algorithm is derived.

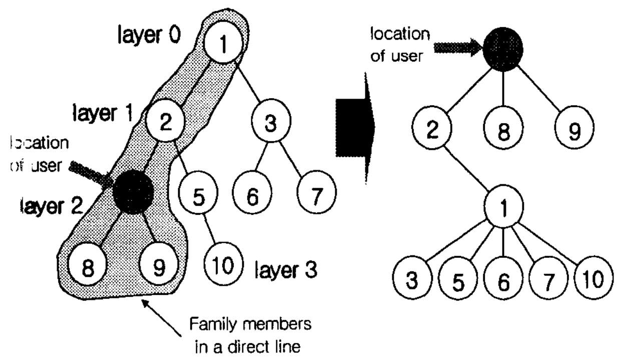

Figure 3. Rearrangement of the whole graph centered on node 4 for computing the weighting value which is considered the hierarchical environment factor

The left diagram of Figure 3 shows four layers from layer 0 to layer 3. Supposed for a user to retrieve some information at node 4 in layer 2, the user may expect to show up documents in his/her node first. Of course, if we consider the importance of documents, the documents from node 1, or node 5, or other node shall be the first of the list of documents. Hence our algorithm improved is to combine the location of documents and the contents of a document in the existing PageRank algorithm. In this case, if we rearrange the hierarchy tree into a new tree with the root of node 4 and a new relation between parents and children. The right part of Figure 3 shows the result of rearrangement as a new graph.

The equation, which computes the weighting value according to the location of a document incorporating the case like Figure 3, is as follows:

$$W_i = 1 - \left| \frac{l_i - L_p}{L+1} \right| \qquad (2)$$

where W_i is a weighting value computed using the location of a document in node i, i.e., the relative location to a user, in a hierarchical web environment. ℓ_i is the location of the layer with the subject node i for ranking. For instance, the values of ℓ_i ranges from 0 to 3 in the case above. L_p is the location of user's layer, i.e., 2 in the case above. And L is the number of layers of the whole network, i.e., 4 in the case above. Thus, maximum value of W_i in equation (2) is 1. Meanwhile, an exception case exist in the right part of Figure 3 such that the nodes 3, 5, 6, 7, 10 are arranged under node 1, i.e., those nodes are not the direct children of node 4. The equation for the weighting value of them are as follows:

$$W_o = \min(W_i) - \frac{1}{L+1} \qquad (3)$$

where Wo means the weighting value of environment for a node which is not the direct child node, here o represents others.

If we compute the weighting value of environment for the node 4 in Figure 3 by applying equations (2) and (3) with respect to user's point of view, $W_4 = 1 - |(2-2)/(4+1)| = 1$. Similarly the values for other nodes can be done. For instance 4/5 for node 2, 3/5 for node 1, 4/5 for node 8 and 9, and 2/5 for other nodes. Finally, we derive the equation of the weighting value in a given hierarchical web environment after combining equation 1 and 2 (or 3) as follows:

$$PR(A)_{improved} = PR(A) \times W_i \qquad (4)$$

In Figure 4, we show the results of the comparison of weighting values with respect to hierarchical view (cases (a) and (b)) and functional view (cases (c) and (d)). In hierarchical cases of Figure 4, we can see that the difference between the weighting values became larger as the depth between the layers grew deeper. In functional cases, the weighting value is not dependent on the depths in the hierarchy, but the difference of weighting values became smaller as the layer goes lower.

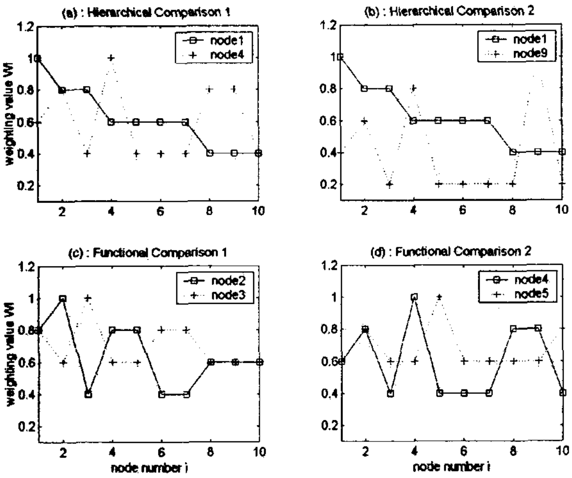

Figure 4. The comparison of weighting values for hierarchical (a and b) and functional cases (c and d)

V CONCLUSIONS

In this paper we propose an new improved ranking algorithm considering user's convenience in a hierarchical web based environment. For the algorithm, we suggest a configuration of the hierarchical web environment such as enterprise and defense networks. Moreover, we propose some equations using graph theory to incorporate user's real location for mining. Finally we show the characteristics of the weighting values as changing the node in each layer using diagrams. We believe that our algorithm to compute the weighting value in a hierarchical web environment is tried at first. And we expect that the algorithm provides more efficient result to the users in a hierarchical web environment.

REFERENCES

[1] Michelangelo Diligenti, Marco Gori, and Marco Maggini, "Web Page Scoring Systems for Horizontal and Vertical Search," 11th World Wide Web Conference, pp.508-516, 2002.

[2] Dik L. Lee, Huei Chuang, and Kent Seamons, "Document Ranking and the Vector-Space Model," IEEE Software, pp.67-75, 1997.

[3] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," Report RJ 10076, 1997.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank Citation Rankings: Bringing Order to the Web," Tech. Report, Computer Science Department, Stanford University, 1998.

[5] Monika R. Henzinger, "Hyperlink Analysis for the Web," IEEE Internet Computing, pp.5-50, 2001.

[6] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins, "The Web and Social Networks," IEEE Computer, pp.32-36, 2002.

[7] S.Chakrabarti et al., "Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text," 7th International World Wide Web Conference, Computer Networks and ISDN Systems, vol. 30, No.1-7, pp.65-74, 1998.

[8] R. Lempel and S. Moran, "SALSA: The Stochastic Approach for Link-Structure Analysis," ACM Transactions on Information Systems, Vol. 19, No. 2, pp.131-160, 2001.

[9] M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles and M. Gori, "Focused Crawling Using Context Graphs," 26th International Conference on Very Large Database, pp.527-534, 2000.

[10] Soumen Chakrabarti, Martin van den Berg, and Byron Dom, "Focused Crawling:A New Approach to Topic-Specific Web Resource Discovery," 8th World Wide Web Conference, pp.545-562, 1999.

[11] Heebyung Yoon, Kilsup Lee, and Hwa-Soo Kim, "Modeling a Multi-Agent based Web Mining System on the Hierarchical Web Environment," KNDU Technical Report, KNDU-TR-03-01, 2003.