A new algorithm to measure relevance among Web pages

M. S. Sadi[1], M. M. H. Rahman[2] & S. Horiguchi[2]
[1]*Department of CSE, KUET, Khulna, Bangladesh*
[2]*GSIS, Tohoku University, Sendai, Japan*

Abstract

This paper proposes a new algorithm to measure relevance among Web pages (RWP) using a hybrid method of hyperlink analysis and content analysis. Here we used a new approach to Web searching where the input to the search process is not a set of query terms, rather the URL of a page, and the output is a set of related Web pages. A related Web page is one that addresses the same topic as the original page. Here, the proposed algorithm first uses only the connectivity information in the Web (i.e., the links between pages) and then the content of pages. To evaluate the performance, the algorithm is compared with existing algorithms. Experimental results show that RWP outperforms existing algorithms to find relevant Web pages. RWP increases the search efficiency effectively and enhances the application area of Web related research.
Keywords: Web mining, relevant pages, hyperlink analysis, content analysis.

1 Introduction

The World Wide Web is a rich source of information and continues to expand in size and complexity. How to efficiently and effectively retrieve required Web pages on the Web is becoming a challenge day by day [1–3]. There are many ways to find relevant pages. For example, as indicated in [4], Netscape uses Web page content analysis, usage pattern information, as well as linkage analysis to find relevant pages. Among the approaches of finding relevant pages, hyperlink analysis has its own advantages. Primarily, the hyperlink is one of the most obvious features of the Web and can be easily extracted by parsing the Web page codes. Most importantly, hyperlinks encode a considerable amount of latent human judgment in most cases [5, 6].

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060251

The page source for relevant page finding in [5] is derived directly from a set of parent pages of the given page. Kleinberg's HITS (Hyperlink-Induced Topic Search) algorithm [5] is applied directly to this page source and the top authority pages (e.g., 10 pages) with the highest authority weights are considered to be the relevant pages of the given page. Cocitation algorithm [3] is improved in two aspects: First, the page source is derived from both parent and child pages of the given page and the way of selecting pages for the page source is different from that of [5]. Second, the improved HITS algorithm [5], instead of HITS algorithm, is applied to this new page source. This algorithm is named Companion Algorithm [4]. The improved HITS algorithm reduces the influence of unrelated pages in relevant page finding. These algorithms focus on finding authority pages (as relevant pages) from the page source, rather than directly finding relevant pages from page similarities. Therefore, if the page source is not constructed properly, i.e., there are many unrelated pages in the page source, the topic drift problem [5] would arise and the selected relevant pages might not be actually related to the given page. Dean and Henzinger [4] proposed DH Algorithm to find relevant pages from page similarities. The page source of this algorithm, however, only consists of the sibling pages of the given page and many important semantically relevant pages might be neglected. Although the algorithm is simple and efficient, the deeper relationships among the pages could not be revealed. The relevance of Web pages was measured based on only hyperlink analysis in [3]. Combining content analysis and hyperlink analysis yields better results as shown in our proposed method.

In our research work, we have proposed an algorithm that uses page similarity as well as content analysis to find relevant pages. The page similarity analysis and definition are based on hyperlink information among the Web pages [7]. The hybrid method of hyperlink and content based similarity makes the result more appropriate than those were returned by only hyperlink analysis. The idea of content analysis is taken from keyword based clustering of Web documents [8]. RWP is a cocitation algorithm that extends the traditional cocitation concepts. It is intuitive and concise as the relevant pages returned by this algorithm contain those pages that address the same topic as the given page as well as semantically relevant to the given page.

The algorithm is based on the page cocitation analysis and content based clustering. Although this algorithm is simple and efficient, the deeper relationships among the pages cannot be revealed.

2 Problem statement

When hyperlink analysis is applied to the relevant page finding, its success depends on how to solve the following two problems [9, 10].

1. How to construct a page source that is related to the given page, and
2. How to establish effective algorithms to find relevant pages from the page source.

Ideally, the page source is a page set from which the relevant pages are selected and should have the following properties.

1. The size of the page source (the number of pages in the page source) is relatively small.
2. The page source is rich in relevant pages.

The best relevant pages of the given page, based on the statement in [4], should be those that address the same topic as the original page and are semantically relevant to the original one.

3 Citation and cocitation analysis

Before describing RWP algorithm, the following terms should be defined.

Cocited: For a pair of document *p* and *q*, if they are both cited by the common document, documents *p* and *q* is said to be cocited. If a Web page *p* has a hyperlink to another page *q* then page *q* is said to be cocited by page *p*.

Cocitation degree: The number of documents that cite both *p* and *q* is referred to as the cocitation degree. The similarity between two documents is measured by their cocitation degree.

Backward cocitation degree: Two pages p_1 and p_2 are back cocited if they have a common parent page. The number of their common parent pages is their backward cocitation degree denoted as $b(p_1, p_2)$.

Forward cocitation degree: Two pages p_1 and p_2 are forward cocited if they have a common child page. The number of their common children is their forward cocitation degree, denoted as $f(p_1, p_2)$.

Intrinsic page: The pages are intrinsic pages if they have same domain. Consider two URLs such as www.kuet.ac.bd/student and www.kuet.ac.bd/teacher. We can say that these two pages are in the same domain www.kuet.ac.bd. Then we can conclude that those two pages are intrinsic pages.

Near duplicate page: Two pages are near duplicated pages if they have more than 10 hyperlinks and they have at least 95% (This is a user defined parameter) of their hyperlinks in common.

4 Relevance among Web Pages (RWP) algorithm

The steps of RWP algorithm can be outlined shortly in the following steps.

Step 1: Building the vicinity graph

When hyperlink analysis is applied to the relevant page finding, its success depends on constructing a page source that is related to the given page [9, 10]. Given a Web page u, its parents and child page could be easily obtained. After parent and child pages of u are obtained, it is possible to construct a new page source for u that is rich in related pages. Choose up to B arbitrary parents of u and choose first F children of u. For each of the parents' P, choose up to BF children (different from *u*) of P that surround the link from P to u. Thus we get the siblings of u as shown in Figure 1. For each of the children C,

choose up to FB parents (different from u) of C and a new set of parents FS is constructed. The page source structure constructed in the above way is termed as vicinity graph. The procedure of the building Vicinity graph is shown as follows:

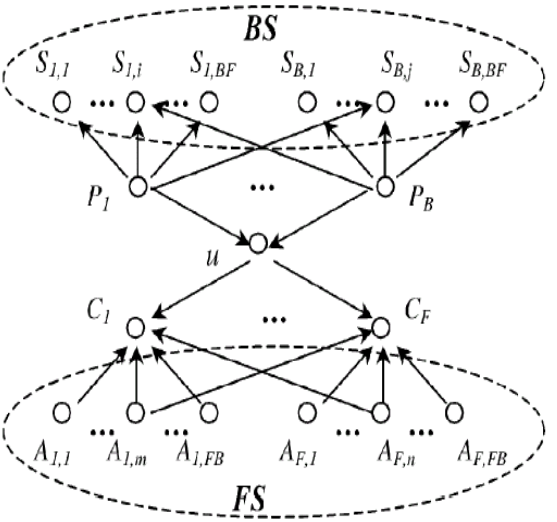

Figure 1: Page source structure for the RWP.

Procedure *BuildingVicinityGraph*

Input: B arbitrary parents of u.

1. Let P_u be a set of parents pages of u.
2. $P_u = \{ p_i \mid p_i$ is a parents page of u without intrinsic and near duplicate pages, $i \in [1, B] \}$.
3. Let $S_i = \{ s_{i,k} \mid s_{i,k}$ is a child page of page p_i, $s_{i,k} \neq u$, $p_i \in P_{u,k} \in [1, BF]$, $i \in [1, B]$. Then step 1 and 2 produce the following set: $BS = \bigcup_{i=1}^{B} S_i$.
4. Choose first F children of u.
5. Let C_u be a set of child pages of u, $C_u = \{ C_i \mid C_i$ is a child page of u without intrinsic and near duplicate pages, $i \in [1, F] \}$. Let $A_i = \{ a_{i,k} \mid a_{i,k}$ is a parent page of page c_i, $a_{i,k}$ and u are neither intrinsic nor near-duplicate pages, $C_i \in C_u$, $k \in [1, FB]$, $i \in [1, F]$. Then, step 3 and 4 produce the following set: $FS = \bigcup_{i=1}^{F} A_i$.

Step 2: Merging intrinsic page

Merge intrinsic parent pages and child pages of u after building the vicinity graph as Figure 1. To merge two intrinsic pages, a node whose link is the union of the links of the two intrinsic pages replaces those two nodes.

Step 3: Merging near-duplicate page

Merge near-duplicate pages after merging the intrinsic pages. This process is same as merging intrinsic pages. To merge two near-duplicate pages, a node whose link is the union of the links of the two near-duplicate pages replaces those two nodes. This duplicate elimination phase is important because many pages are duplicate across hosts (e.g. mirror sites, different aliases for the same page), and it was observed that allowing them to remain separate can greatly distort the result.

Step 4: Calculation of cocitation degree and hence finding relevance

For a given selection threshold (δ), select pages from BS and FS such that their backward cocitation degree or forward cocitation degree with u are greater than or equal to δ. These selected pages are relevant pages of u, i.e. the relevant page set *RP* of u is constructed as:

$$RP = \{ P_i \in BS \text{ with } b(P_i, u) \geq \delta \text{ OR } P_i \in FS \text{ with } f(P_i, u) \geq \delta \}$$

RP is sorted according to cocitation degree and returned to display.

Step 5: Combining hyperlink analysis and content analysis

This idea is taken from keyword based clustering of Web documents [8]. The first step is to gather the keywords of the Web pages. For this purpose we read the content of HTML file and exclude the HTML tag and commented JavaScript code. We consider each word as a keyword and remove the words from each Web page which will not be considered as keywords.

The articles, pronouns, auxiliary verbs, conjunction, internet buzzwords like www, http etc. are not are considered as keywords. In this step we compare the keywords of u (input URL) with the pages that have same cocitation degree. The method of content analysis is that if a page has larger number of common keywords with another page with respect to other pages then the first one will be more relevant with the second one than others [8].

5 Experimental results

In our experiment, we selected an arbitrary Web page u as www.amardesh.com. For this URL, we selected 10 parents URL of u and for every parent we selected 6 children. In the same way, we selected 10 child pages of u and for every child we selected 6 parents. All the URLs are selected arbitrarily. Thus we formed the vicinity graph as Figure 1. Table 1 shows a list of parents of u that we choose and Table 2 shows their children. Every parent in Table 1 has 6 children in Table 2.

For each URL in Table 1, the index of its child pages is shown in the third column of the same table. For example, www.casio.com (index number 5 in Table 1) has child indexes 2, 15, 16, 17, 18, 19 which indicates that the child

pages of www.casio.com belongs to 2, 15, 16, 17, 18, 19 number of indexes in Table 2.

Table 2 shows the child pages of the URLs that are shown in Table 1. The index number of these child pages is used in Table 1 in the third column to show the set of child pages for each URL. For example, the parent page of www.ourbangla.com, www.mail.ac.bd, www.pager.com, www.builder.com, www.factor.com and www.source.com is www.casio.com and their index number in Table 2 are 2, 15, 16, 17, 18, and 19. So the fifth row of the Table 1 is formed using this information. In similar fashion, Table 1 and Table 2 are formed.

Table 1: Parent pages of u.

Index#	URL	Child Index
1	www.toyota.com	1, 2, 3, 4, 5, 6
2	www.gd.com	1, 3, 5, 6, 7, 8
3	www.boy.com	5, 7, 8, 9, 10, 11
4	www.nybour.com	2, 4, 6, 12, 13, 14
5	www.casio.com	2, 15, 16, 17, 18, 19
6	www.fantasy.com	20, 21, 22, 23, 24, 25
7	www.toyota.com/search	16, 6, 2, 27, 28, 29
8	www.casio.com/take	8, 18, 29, 30, 31, 11
9	www.doggy.com	9, 30, 31, 32, 33, 34
10	www.nice.com	15, 16, 35, 36, 37, 38

Then we applied Extended Cocitation algorithm and DH algorithm [2] in this page source and found the result that is shown in Table 3. This algorithm measures cocitation degree with respect to given page www.amardesh.com. Best 10 relevant pages returned by Extended Cocitation algorithm and DH algorithm are shown in Table 3 and indexed in descending order of the cocitation degree.

From Table 3, it can be observed that Extended Cocitation algorithm calculates cocitation degree more effectively than DH algorithm. The highest cocitation degree measured by Extended Cocitation algorithm is 4, whereas the highest cocitation degree measured by DH algorithm is 3. Moreover, by Extended Cocitation algorithm it was found that there are three pages whose cocitation degree are equal to 3, but by DH algorithm here is only one page whose cocitation degree is equal to 3. Since, cocitation degree measures relevance among Web pages, so we can say that Extended Cocitation algorithm performs better than that of DH algorithm to do that job.

For the same cocitation degree, Extended Cocitation algorithm could not identify which page is more related to the given pages u . In Table 4 we see that this limitation of the Extended Cocitation algorithm can be overcome by our proposed *RWP* algorithm.

Table 2: Child pages of u.

Index#	URL (http ://)	Index#	URL (http ://)
1	www.prothom-alo.com	20	www.cyberspace.com
2	www.ourbangla.com	21	www.hara.com
3	www.yahoo.com	22	www.givson.com
4	www.espnstar.com	23	www.google.com
5	www.dailystar.com	24	www.altavista.com
6	www.yahoogreeting.com	25	www.latent.com
7	www.yahoofriend.com	26	www.friendfinder.com
8	www.yahoogroup.com	27	www.yahoomail.com
9	www.yahoonews.com	28	www.fieldserver.com
10	www.bbc.com	29	www.north.com
11	www.kuet.ac.bd	30	www.south.com
12	www.kuet.ac.bd/Teacher	31	www.india.com
13	www.buet.ac.bd	32	www.bangladesh.com
14	www.cuet.ac.bd	33	www.america.com
15	www.mail.ac.bd	34	www.korea.com
16	www.pager.com	35	www.Nepal.com
17	www.builder.com	36	www.sports.com
18	www.factor.com	37	www.Newsserver.com
19	www.source.com	38	www.microsoft.com

Table 3: Top 10 relevant pages returned by Extended Cocitation and DH algorithm.

Index#	URL(http ://)	Cocitation degree (returned by Extended Cocitation algorithm)	Cocitation degree (returned by DH algorithm)
1	www.ourbangla.com	4	3
2	www.yahoogroup.com	3	2
3	www.pager.com	3	2
4	www.khobor.com	3	2
5	www.espnstar.com	2	2
6	www.dailystar.com	2	2
7	www.yahoogreeting.com	2	2
8	www.yahoofriend.com	2	2
9	www.earth.com	2	1
10	www.india.com	2	1

Table 4: Top 10 relevant pages returned by the RWP algorithm.

Index #	URL (http ://)	Cocitation degree	Keywords similarity in percentage
1	www.ourbangla.com	4	70
2	www.yahoogroup.com	3	79
3	www.pager.com	3	65
4	www.khobor.com	3	45
5	www.espnstar.com	2	86
6	www.dailystar.com	2	73
7	www.yahoogreeting.com	2	60
8	www.yahoofriend.com	2	56
9	www.earth.com	2	32
10	www.india.com	2	27

RWP algorithm not only uses linkage information among the pages but also uses the content of the pages and computes the keywords similarity to the page u (input URL). Then the relevant page is further sorted as keyword similarity. The highest similarity indicates that it is more relevant to a given page. In Table 4, the fourth column shows the keywords similarity (with respect to www.amardesh.com) of the Web pages that are shown in the second column (respectively). The pages www.yahoogroup.com, www.pager.com, and www.khobor.com in Table 4 have their same cocitation degrees (equals to 3) but their relevance can be varied if the keywords similarity is measured. And then it is found that www.yahoogroup.com has the highest similarity which is 79% whereas the rest two have the keywords similarity only 65% and 45% respectively. So, www.yahoogroup.com is more relevant to www.amardesh.com than other two pages and www.pager.com has more relevance than www.khobor.com with that page.

Therefore, we can say that RWP algorithm returns more relevant pages compared to Extended Cocitation algorithm and the order of returning relevant pages in RWP algorithm is more accurate than Extended Cocitation algorithm as well as some other dominant algorithms.

6 Conclusion

This research work proposes a new algorithm RWP (relevance among Web pages) to find relevant pages of a given page. This algorithm is based on the combined method of hyperlink analysis and content analysis among the page. It avoids some useful information from omission and prevents the results from distortion by malicious hyperlinks. This algorithm could identify the pages that

are relevant to the given page in a broad sense, as well as those pages that are semantically relevant to the given page. Experimental results show that RWP reveals deeper (mathematical) relationships among the pages and finds out relevant pages more precisely and effectively. The ideas in this work would also be helpful to other linkage-related analysis. Our algorithm can be extended to handle more than one input URL. In this case, the algorithms would compute pages that are related to all input URLs. This algorithm only deals with the "static" links among the pages and this work can be extended to deal with dynamic links.

Acknowledgments

The authors thank Md. Rashedul Hasan and Mohammad Riyadh Hossain of KUET for their help in this research work. This research was supported in part by the research grant (2005B2-46) of Sahi Glass Foundation.

References

[1] Scarpa, M., Puliafito, A., Villari, M., and Zaia, A., A Modeling Technique for the Performance Analysis of Web Searching Applications. *IEEE Trans. on Knowledge and Data Engineering*, 6(11), pp. 1339–1356, 2004.

[2] Diligenti, M., Gori, M., and Maggini, M., A Unified Probabilistic Framework for Web Page Scoring Systems. *IEEE Trans. on Knowledge and Data Engineering*, 16(1), pp. 4–16, 2004.

[3] Hou, J. and Zhang, Y., Effectively Finding Relevant Web Pages from Linkage Information. *IEEE Trans. on Knowledge and Data Engineering*, 15(4) pp. 940–951, 2003.

[4] Dean, J., and Henzinger, M., Finding Related Pages in the World Wide Web, *Proc. 8th WWW Conf.*, pp. 389–401, 1999.

[5] Kleinberg, J., Authoritative Sources in a Hyperlinked Environment, *J. ACM*, 46(5), pp. 604–632, 1999

[6] Abiteboul, S., Preda, M., and Cobena, G., Adaptive on-line page importance computation. *Proc. of 12th WWW Conf.*, Budapest, Hungary, pp.280–290, 2003.

[7] Bharat, K., Broder, A., Henzinger, M., Kumar, P., and Venkatasubra manian, S., The Connectivity Server: Fast Access to Linkage Information on the Web. *Proc. 7th WWW Conf.*, pp. 469–477, 1998.

[8] Sadi, M.S., Rahman, C.M., Hasan Babu., H.M., An Efficient and Coherent Method Using Data Mining to Cluster Web Documents. *Proc. 2nd ICECE*, Dhaka, Bangladesh, pp. 273–276, 2002.

[9] Bharat, K., and Henzinger, M., Improved Algorithms for Topic Distillation in a Hyperlinked Environment. *Proc. 21st Int'l ACM Conf. Research and Development in Information Retrieval*, pp. 104–111, 1998.

[10] Brin S., and Page, L., The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proc. 7th WWW Conf.*, pp. 14–18, 1998.