A Method for Statistical Diagnosing a Web Site Structures based on Graph Theory

Hideaki SANO*, Masaaki KANAKUBO** and Chiaki HISHINUMA**

**Toshiba Information Systems Technology Co. Ltd., 7-1, Nisshin-cho,*
Kawasaki-ku, Kawasaki 210-0024, Japan

***Tokyo University of Technology, 1404-1 katakura-cho, Hachioji-city,*
Tokyo 192-0982, Japan

email:kanakubo@cs.teu.ac.jp, hisinuma@cs.teu.ac.jp

Abstract: We propose a new method for statistical diagnosing a web site. In this paper, we focus to hyperlinks of a web site. We applied the way of web structure mining to the diagnosis of a web site. The pages and hyperlinks of a web site may be viewed as nodes and edges in a directed graph. We defined new two original criteria for diagnosis, that is, averaged number of clicks and reachable rate. The former means average value of the number of clicks to reach each page containing a web site from the top page. The later means affinities between the structure of a web site and the strongly connected directed graph. Our new tool simply diagnoses the whole web site only by hyperlinks. In the test, we input 140 URLs of company .etc to this diagnosing tool. It has shown that the average number of clicks is only 2.43 and about 48% pages of all are reachable from a certain arbitrary page, and that this rate reaches about 87% by back button. By the two criteria, our proposed tool enables us to examine extremely easily the accessibility of a web site.

Ⅰ. INTRODUCTION

Recently, the internet has become very popular. So, the web sites as customer centers have become much more important for companies. Also the web sites will cut costs in the running of the office. Therefore, especially for companies, one of the first problems is to construct handy web sites. And the web site structure plays a very important role in setting the accessibility of a web site. However, there is no method for diagnosing of a web site structure.

In this paper, diagnosing of a web site means evaluating of the user-ability or accessibility of the web site.

Up to now, many automatic diagnosing systems for web sites have been developed. However, these systems diagnose each pages of the web site, for instance, file size or character size instead of diagnosing the web site structure.

On the other hand, recently, many researches on the web structure mining have been done to discover new knowledge from a great deal of web sites. For instance, it is proved that there is a strongly connected structure at the center of huge web site by a research on about two hundreds million pages. And, many researches on links of web sites have been done to discover web communities, which mean groups of sites concerning the same matter. However, we had never seen the application of graphical structure of hyperlinks to diagnosing the user-ability of a web site.

In this paper, we propose a new method for statistical diagnosing a web site based on graph theory. The feature of the proposed method is to place more emphasis on the accessibilities of each page containing the site from the top page.

We defined new two original criteria for diagnosis, that is, averaged number of clicks and reachable rate. The former means average value of the number of clicks to reach each page containing a web site from the top page. The later means affinities between the structure of a web site and the strongly connected directed graph. The proposed method cans easily diagnosing the user-ability of a web site. The system developed by the proposed method is implemented on our web server.

Ⅱ. THE OUTLINE OF THE PROPOSED METHOD

A. Basic concept of the assessment for a web site

Current automatic diagnostic systems only diagnose file size and the size of characters of each web page. Thus, their diagnostic range is extremely restricted. Contrary to this, our new tool diagnoses the whole web site. And it simply diagnoses only by links of the pages belonging to a web site. For this purpose, we defined new two original criteria for diagnosis, that is, reachable rate and averaged number of clicks. And we calculated these criteria by Dijkstra algorithm that is amazingly efficient shortest path algorithm.

B. Averaged number of clicks

If it requires a lot of times clicks to reach the target page, it takes very long time and the web page is hard to use. Contrary to this, a web page in which the user can reach the target page by few clicks is good page. In current researches the numbers of clicks often used as a measurement for the distance between a certain page and the other. However, these researches focused on the distance between two pages belonging different sites to each other. From this reason, we introduced averaged number of clicks as a new criterion. The averaged number of clicks is defined as follows.

$$Ave. = \frac{\sum_{i=1}^{n} Ti}{n+1} \ldots\ldots\ldots\ldots \quad (1)$$

where, n is the number of pages belonging to the web site, Ti is the number of necessary clicks to reach i*th* page from the top page and Ave. is the number of clicks.

C. The reachable rate

If there is a directed edge between two arbitrary nodes of a directed graph, this directed graph is denoted by “strongly connected”. The pages and hyperlinks of a web site may be viewed as nodes and edges in a directed graph. If the structure of a web site is strongly connected, the web site is convenient because it enables us to reach arbitrary page from current page. Actually, quite a few web sites may have a strongly connected directed graph. So, we defined the reachable rate that means affinities between the structure of a web site and the strongly connected directed graph.

If a web site has n pages and the number of pages to which can reach from i*th* page is Ri, the reachable rate of i*th* page is defined as Ri / (n-1). And the reachable rate of the entire site R*rate* is defined as follows.

$$Rrate = \frac{\sum_{i=1}^{n} Ri}{n} \ldots\ldots\ldots\ldots \quad (2)$$

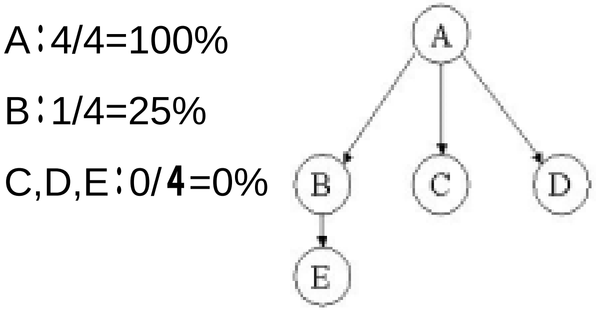

Fig.1 Example of the reachable rate.

Fig.1 shows an example of the reachable rates. Since the users of the site can reach every page of the site from the top page A, the reachable rate of page A is 100%. Since from page B the users can reach only page E, the reachable rate of page B is 25%. And since from page C or D or E, the user can reach no page, the reachable rates of these pages are 0%. Finally, the reachable rate of the entire site is 25% as shown in Fig.1. If the reachable rate of the entire site is 100%, the web site’s structure is equal to the strongly connected.

D. The reachable rate considering the back button

The users of internet can return to the previous page by clicking the back button of the browser. Then, we defined the reachable rate considering the back button. This rate is defined by the same expression that used in definition of the plain reachable rate. However, as a rule, only when the user visits a page having no next link, he/she can back to the previous page using the back button. Using the buck button, the user can back only to the previous page.

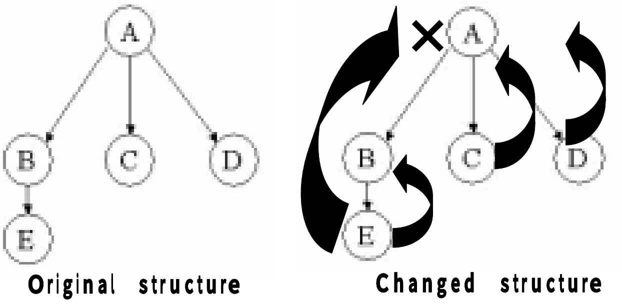

Fig.2 Example of change of a web structure considering the buck button.

Fig.2 shows an example of change of a web site structure considering back button. The reason we introduced this rule is that if the user can use back button at any page, the reachable rate of the entire site become 100% by returning to the top page. In the case of the structure shown in Fig.3, since the reachable rates of page C, page D and page E rise to 25%, respectively, the reachable rate of the entire site also rise to 40%.

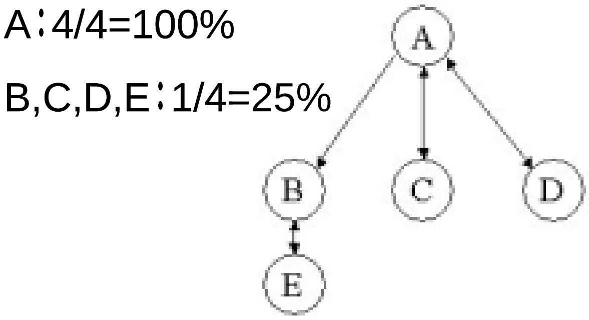

Fig.3 Example of the reachable rate considering back button.

III. Experiments

A. User interface

Fig.4 shows the window of the proposed diagnosing system. The user can input the URL of the top page of the site that the user wants to diagnosis into the upper text box. Then he/she push the execution button, the average number of clicks, the reachable rate, the reachable rate considering the back button and total page number appear the next text box, respectively. And the entire structure will have shown graphically in the lower large box. This window can be shown by following URL.

http://www.teu.ac.jp/hishi/

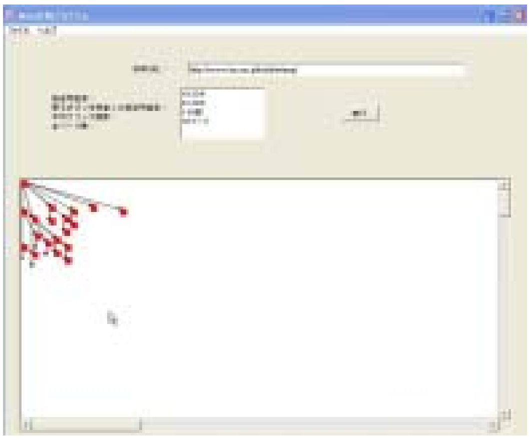

Fig.4 The window of the proposed diagnosing system.

B. The way of the test

In the test, we input 140 URLs of companies etc to this diagnosing tool. There are seven different categories, that is, software makers, electrical machinery stores, universities, cosmetics companies, call centers, transportation companies and semiconductor manufacture companies making up these companies etc. Each category includes 20 companies etc, respectively. Each web site diagnosed by four criteria, that is, the number of pages, averaged number of clicks, reachable rate and the reachable rate considering the back button.

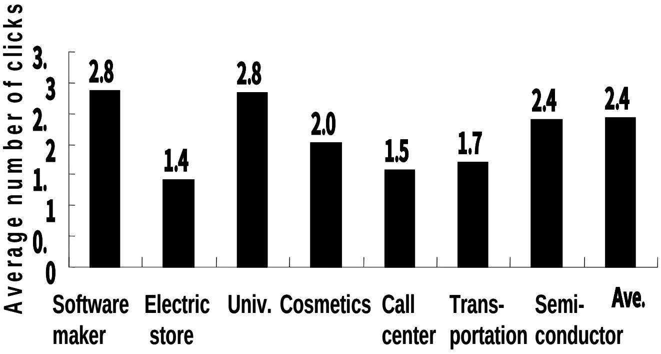

Fig.5 Average number of clicks of each category.

C. Averaged number of clicks results

Fig.5 shows average numbers of clicks of each category. The test showed that the average number of clicks is only 2.43 clicks to go to other pages from the top page of the web sites. The reference mentioned that the desirable number of layers of links is under four. It is clarified that most sites has the number of layers to fill this demand. Our tool acquired the standard number of layer of normal site.

D. Reachable rates results

Fig.6 shows average reachable rates and the average reachable rates considering the back button of 140 web sites. Fig.7 shows the same rates of each category. It is clarified that the average reachable rate remains 48%, but the rate reaches about 87% by back button. Therefore, there is no problem in practice. Fig.7 shows that there is no significant difference of reachable rates according to the category. Our tool also acquired the standard reachable rates of normal site.

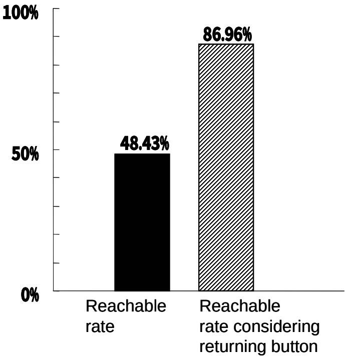

Fig.6 Average reachable rates and the average reachable rates considering the back button of 140 web sites.

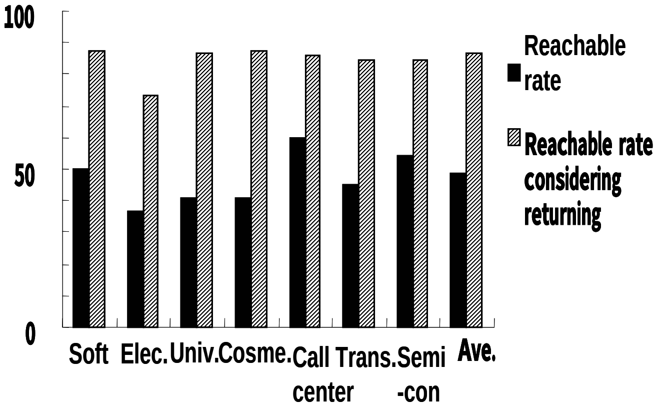

Fig.7 Two kinds of reachable rates of each category.

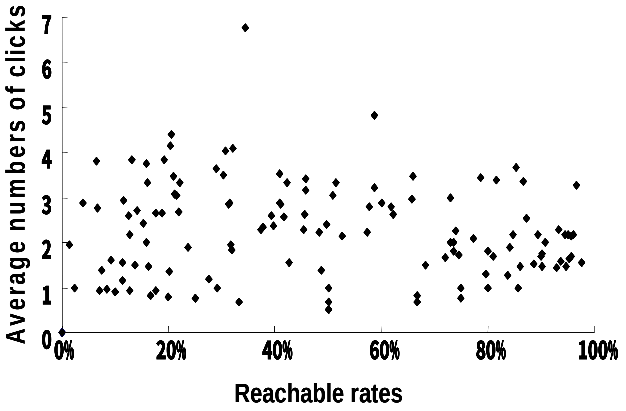

Fig.8 The relation between the two criteria.

E. *The relation between the two criteria*

Fig.8 shows the relation between the reachable rate and the averaged number of clicks. It has turned out that the two criteria are orthogonal for each other. It is clarified that web sites can be diagnosed by two different criteria.

IV. CONCLUSION

We propose a new method for statistical diagnosing a web site. We applied the way of web structure mining to the diagnosis of a web site. We defined new two original criteria for diagnosis, that is, averaged number of clicks and reachable rate. The former means average value of the number of clicks to reach each page containing a web site from the top page. The later means affinities between the structure of a web site and the strongly connected directed graph. Our new tool simply diagnoses the whole web site only by hyperlinks.

In the test, we input 140 URLs of company .etc to this diagnosing tool. It has shown that the averaged number of clicks is only 2.43 and about 48¥% pages of all are reachable from a certain arbitrary page, and that this rate reaches about 87¥% by back button. And it has turned out that the two criteria are orthogonal for each other. Our proposed tool acquired standard these criteria of normal site. By the two criteria, our proposed tool enables us to examine extremely easily the accessibility of a web site.

REFERENCES

[1] Hishinuma, C.: Web Contact using Internet -Contact Center Management; vol.5, Sept 2004, p.25, LCA Communications, (2004).

[2] Ishida, Y.: Web Usability & Accessibility Guidelines; Mainichi Communications Inc., p.271 (2003).

[3] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Steta, R., Tomkins, A., Wiener, J.: Graph Structure in The Web; The 9th International World Wide Web Conference,(2000).

[4] Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A.: Trawling the Web for Emerging Cyber Communities; The 8th International World Wide Web Conference,1999.

[5] Page, L., Brin, S., Motwani, R., Winograd, T.:The PageRank Citation Ranking; Bringing Order to the Web,1998.

[6] Haveliwala, H.: Efficient Computation of PageRank; Stanford Technical Report,1999.

[7] Hirabayashi, M., Ohtsuki, K., Kiyomitu, H., Morishita, J., Kitamura, S., Kinugawa, T.:Semantic Web Page Scoring Based on Intention of Hyperlinks; Transactions of Information Processing Society of Japan, Vol.43, No.SIG 12, pp.92-101 (2002).

[8]http://www.symphonic.co.jp/web/design/index.html

[9] http://www.syntax.co.jp/webchecks/index.html

[10] Matsuo, Y., Ohsawa, Y., Ishizuka, M.:A New Definition of Subjective Distance between Web Pages; Transactions of Information Processing Society of Japan, Vol.44, No.1, pp.88-94 (2003).

[11] Adamic, L.A.:The Small World Web; Proc. ECDL'99,pp.443-452 (1999).

[12] Chakrabarti, S., van den Berg, M., Dom, B.:Focused crawling: a new approach to topic-specific Web resource discovery; Proc. 8th WWW Conf.,1999.