Association for Information Systems
AIS Electronic Library (AISeL)

December 2004

Quantitative Analysis of Website Based on Web Graph Theory

Youwen Wu
Fudan University

Youwei Wang
Fudan University

Weihui Dai
Fudan University

Follow this and additional works at: http://aisel.aisnet.org/pacis2004

Quantitative Analysis of Website Based on Web Graph Theory

Youwen Wu
Department of Information
Management and
Information System
Fudan University, Shanghai,
P.R.China
032025064@fudan.edu.cn

Youwei Wang
Department of Information
Management and
Information System
Fudan University, Shanghai,
P.R.China
ywwang@fudan.edu.cn

Weihui Dai
Department of Information
Management and
Information System
Fudan University, Shanghai,
P.R.China
whdai@vip.sina.com

ABSTRACT

Quantitative analysis of website has been applied widely in the structure design of the websites. Web metrics such as Web graph theory, Web page significance, Webpage Search and Retrieval and Usage Characterization have been researched for many years, aiming at providing good guidance for the structure design of website. But all the metrics need to be practiced to promote the theories themselves, and in fact this kind of practice or research is not fully developed. To supply this lack, this paper showed the guidance for such kind of research, demonstrated the method and process of using web graph metric to analyze hyperlink structure, and provided a case study based on the web graph theory. Base on the application and analysis, strategies for design and improvement of the theory are put forward as the main purpose of this paper.

Keywords: Web Metrics, Web Graph, Hyperlink Structure Analysis

1 Introduction

The research of web metrics has generated many useful methods and theories. They all aim at solving the practical problems in the application of websites, such as the navigation, the accessibility, and the information retrieval structure. Many predecessors have researched extensively on this topic. (Bray 1996) studied the global measurements of Web size, hit counts, and click through rates. (Pitkow 1997) studied the metric of information retrieval on the World Wide Web. Besides these, many other topics on web metrics have also been studied.

Before introducing the web graph theory, we first retrospect the web metrics that were brought till now. Based on the methods and measurements used, web metrics can be classified into the following categories:

1) *Web Graph Theory. The website was defined as a set of nodes/pages linked by the hyperlinks. Many graph-based metrics can be used to analyze the web structure under this premise.*
2) *Web Page Significance*. The information contained in the websites has different importance levels to the users. Page significance research mainly focused on the quality and relevance of web pages considering the requirements of users.
3) *Web Page Retrieval/Page Ranking*. These are metrics to evaluate the service performance of information retrieval and web search. Other theories like the web page similarity also belong to this frame.
4) *User Habit Research*. With the development of artificial intelligent and multi-agent theory, we can customize the website for each user based on his habits. Some application of this theory has been developed (Wang Youwei 2003).

Although a lot metrics have been studied, few researches engaged in the implementation of these metrics in actual websites. In this paper, we will explore this kind of research based on the web graph theory.

Web graph metrics grasp the character of the websites that is similar to net, making use of the graph theory in the mathematics to analyze the terms of website such as structure, hyperlinks, and hierarchy. According to this theory, all the pages are deemed as nodes in the graph, and all the hyperlinks between pages are deemed as arcs. The premise mainly attempt to achieve better analysis for the website design. The analyzing objects are the whole graph structure and the hyperlinks. The research of (Botafogo et.al1992) may be the prior holistic research on web graph theory. Based on the predecessors' research, in this paper we will conduct more practical research, make some improvement of the theory, and present the guidance for using these metrics.

The remaining part of this paper is organized as follows: section 2 describes the basics of web graph theory, including the fundamental theory and basic quantitative standards; a case study will be presented in section 3. The computing method of those quantitative terms will also be discussed in this section; Section 4 is based on the computing of section 3 and shows how to use the result to guide the design of websites. Some improvement of the theory is also presented in this section. Section 5 concludes the paper with remarks and future directions.

2 Web Graph Algorithm

Web graph algorithm is the core concept of the web graph theory. It's a definition based on the premise of web graph theory. Under such premise, all websites can be seen as a graph. For example, for a website with 4 pages and hyperlinks { $1\rightarrow2, 1\rightarrow3, 1\rightarrow4, 2\rightarrow3, 3\rightarrow4$}, it's web graph is presented in Fig.1. Then the distance matrix should be defined. If a website has N nodes, it can be represented as an N*N *distance matrix* M. Where, the M_{ij} is the number of links that have to be followed to reach node j starting form node i or simply the distance from i to j. Otherwise, if nodes i and j are unconnected in the graph, M_{ij} is set to a suitable predefined constant K (here k=4). Table.1 shows an example hyperlinks matrix. The k is usually set a value bigger than the biggest distance among the node reachable to each other. The matrix is named *converted distance matrix* (Dhyani, D 2002)
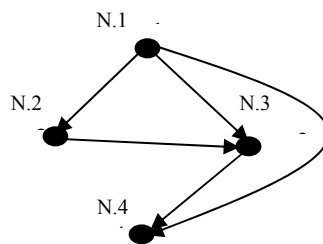

Fig.1 Web graph presentation of the web site

Other measurements are defined based on the web graph. Local Metrics such as centrality describe the status of the single node in the graph; while Global Metrics function as the quantitative definition of the performance of the whole website.

Table.1 Converted distance matrix of the website and the terms of centrality

	N.1	N.2	N.3	N.4	COD	ROC

N.1					
	0	1	1	1	3	10.33
N.2	4	0	1	2	7	4.43
N.3	4	4	0	1	9	3.44
N.4	4	4	4	0	12	2.58
CID	12	9	6	4	31	
RIC	2.58	3.44	5.17	7.75		

2.1 Centrality Metrics

Nodes are the basic elements of a web graph. Centrality defines the relationship between one node and the others. In the graph, M_{ij} is the distance from node i to node j ; the sum of row i is defined as the *converted out distance(COD)* of the node i. Similarly, the *converted in distance(CID)* is defined as the sum of the column i.

$$COD_i = \sum_i M_{ij}$$

$$CID_j = \sum_j M_{ji}$$

The *Converted Distance (CD)* of a web graph is defined as the sum of all entries in the matrix.

$$CD = \sum_i \sum_j M_{ij}$$

Naturally, the smaller the CID/COD, the more central status the node in the web graph. Further, when the number of the nodes is very different in two graphs, the value of the COD/CID can't be used in comparing two websites. So other relative evaluation methods for obliterating the influence of the size of the website is developed for measuring the centrality of a node. The Relative Out Centrality (ROC) metric for node i is defined as the COD of node i dividing CD; the Relative In Centrality (RIC) metric has similar definition.

$$ROC_i = CD\Big/COD_i$$

$$RIC_j = CD\Big/CID_j$$

Examples for computing the values of these metrics can refer to Table.1. Two concepts should be defined here: *index nodes* and *reference nodes*. As the name implies, index nodes are nodes that can be used as an index or guide to many other nodes, and reference nodes are nodes referred by other nodes or saying reachable by other nodes. Following (Botafogo and Shneiderman 1991), we use the definitions as follows:

• Let μ be the mean of the CODs of the nodes in the web graph and let μ' be the mean of the CIDs of the nodes. Note that μ = μ' since every link that leaves a node has to reach another node. For this reason we will use μ for both means.
• Let σ be the standard deviation of the CODs of the nodes.
• Let σ' be the standard deviation of the CIDs of the nodes.
• Let ξ be a threshold value. The value is equal to 3σ or 3σ'
• An index node is a node whose COD is greater than μ + τ.
• A reference node is a node whose CID is greater than μ + τ.

2.2 Global Metrics

When the website is considered as a whole, some global metrics are developed to measure its characteristics. Based on the premise of the web graph, two global metrics were defined to analyze the hierarchical and cross-referencing links. As a global metric, Compactness was designed to indicate the extent of the cross-referencing. Another global metric discussed is Stratum, which is used to capture the linear ordering of the web graph.

Compactness

Compactness express the readability and navigation of a website. Its value is between 0 and 1. If the value is 0 for a web graph's compactness, the website must be completely disconnected. If the value is 1, the website will be a fully connected website. The definition of Compactness is as follows:

$$
\begin{cases}
C_p = \dfrac{Max - \sum_i \sum_j M_{ij}}{Max - Min} \\
Max = (N^2 - N)K, Min = (N^2 - N)
\end{cases}
$$

Both 0 and 1 should be avoided for high readability and navigation when designing the website. Here it's a series web graphs with their compactness to show the direct meanings. From the fig 2, the relation between the shape of the graph and the value of the compactness is clearly presented. The Cp value of the completely connected graph is 1, while the completely disconnected graph is 0. The value of the first graph is between 0 and 1.
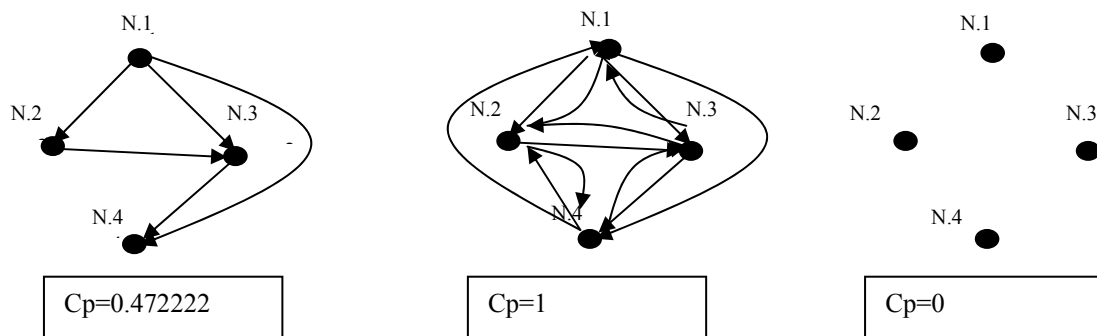

Fig.2 The Relationship between the shape and the value of Cp

Stratum

Stratum metric uses the different matrix as discussed above. In this matrix the K are all replaced by null based on the converted distance matrix. The sum of row i is called the status of the node i, while the sum of the column j is called the contra status of the node j. The prestige of a node is equal to the absolute difference between status and contra status of the node. Node that has a very low status and very low contra status is both hard to leave and hard to reach. The Stratum Metric is used to characterize the linearity in the structure of a web graph. The bigger the Stratum is, the more linear the web graph. The high linearity often induces tedious browsing (as in Table 2).

Table 2 the example based on fig.1 for computing the absolute prestige

	N.1	N.2	N.3	N.4	Status	Absolute prestige
N.1	0	1	1	1	3	3
N.2	∞	0	1	2	3	2
N.3	∞	4	0	1	5	3
N.4	∞	∞	∞	0	0	4

Contra status	0	5	2	4	11	**12**

Here is the definition of the Stratum:

$$\begin{cases} S = \dfrac{\sum_i (the\ absolute\ prestige\ of\ node\ i)}{LAP} \\[2mm] LAP = \dfrac{N^3}{4}, if\ n\ is\ even \\[2mm] LAP = \dfrac{(N^3 - N)}{4}, Othewise \end{cases}$$

Here give a series graph with their value of S show the relationship between the shape and the value of S. The S value of the completely linear graph (2nd graph) is 1, while the completely nonlinear graph(3rd graph) is 0. The value of the first graph (the shape is not either a linear or nonlinear graph) is between 0 and 1.
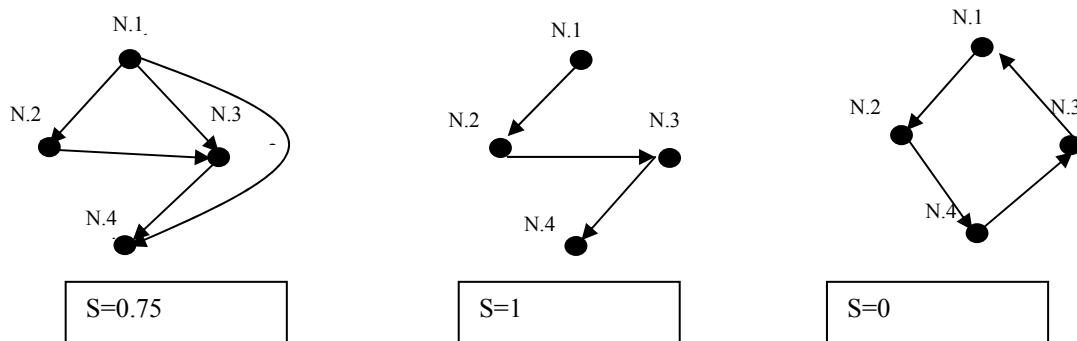

Fig.3 The Relationship between the shape and the value of S

Our definition uses the converted distance. Other Global Metrics by citing the measurements of *degree distribution* have also been studied (Kleinberg. J 1999; Kumar. R 1999) and been confirmed by experiments.

2.3 Local Metrics and Other Measurements

Local Metrics are designed to further indicate special nodes in the web graph. Some researchers call it *node metrics*. The meanings of local metrics can be illustrated through instances. For example, nodes that are too deep in the web graph are hard to reach. This might indicate that links are missing or that information contained in this node is very specialized. Generally, Local Metrics contains *depth* and *imbalance*. Other measurements focused on the links such as the *accessibility* has also been studied. The accessibility is defined as the *attractiveness* or *availability* (Yen B. P 2001). In this paper, the case study mainly focuses on the general design metrics such as centrality and global metrics. A new kind of *Classification* of the research on local metrics is proposed here. The definitions of local metrics and other measurements will be just generalized into two categories but won't be studied in the following case.

Local Metrics focused on Nodes
Depth. The first page of the website is defined as the root in the web graph. The depth of a node in a web graph is just its distance from the root. When the depth of a node is too big, the node is hard for the reader to reach and this leads to a less important status in the website. If the depth is so big that the pages seem unreachable, such nodes indicate possible bugs in the

design of the website. The extreme example of this situation is that the root cannot reach some nodes. Using a depth metric, the unreachable nodes can be located and verified.

Imbalance Metrics. Let T be a general tree. The root of T is the first page of the website. Assume that each node contains only one idea and the importance of each sub idea is equivalent, then the web graph should be a balanced tree. By definition the imbalance of a node is the indication of equilibrium of the T. Too much imbalance might indicate the bias of the designer or a poorly designed website. The further and specific definition of the imbalance metrics is studied (Botafogo et. al.1992). Fig.2 shows two simple examples, one is a balanced tree and the other is an imbalanced tree. In Fig.2, the dotted links represent the cross-references.
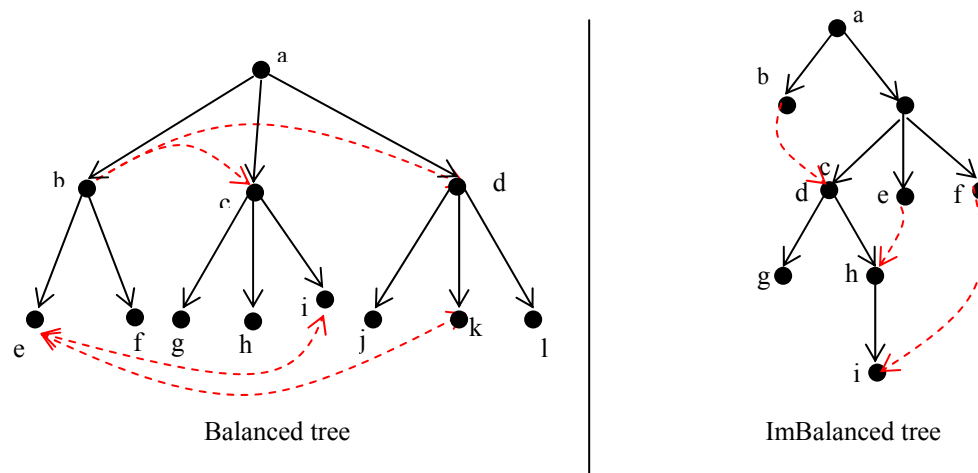
Balanced tree   ImBalanced tree

Fig.4. Web graphs in balance tree and imbalanced tree

Local Metrics focused on Links

The hyperlinks are also worth of research as the nodes. Many local metrics focused on links have been developed. *Accessibility* is local metric of this kind. By definition accessibility of a link is the attractiveness of the link in the web page where the link is located in. As the word attractiveness implied, it's the extent to which the users might notice the links that are highlighted or listed in a special area. The accessibility of links may be related to human factors. It's straightforward for the web designer to determine the accessibility of links by considering the highlighting or position or others. These metrics are related the web metrics on user habit research. The research on user interface also belongs to this kind. In (Nielsen 1999) study, it's obvious that the style of UI influence the users a lot. In other words, it greatly impacts the accessibility.

Table.3 All metrics researched in web graph theory

Category	Name of Metric	Main Functions
Centrality Metrics	**CID/COD**	Direct measure of single node
	RIC/ROC	Normalized measure of single node, to evaluate the central status
Global Metrics	**Compactness**	Evaluate the reachability of the whole site and indicate the extent of the cross-referencing
	Stratum	Capture the linear ordering of the web graph
Local Metrics	**Depth**	The distance from the root, evaluate the reachability of single node
	Imbalance	Gauge the balance of the information

	Accessibility	Evaluate the quality of the hyperlinks
	Other	Focused on nodes or links

This new classification of the local metrics gives us the guidance for further study. All the researches on local metrics are focused on either nodes or links. Local metrics make the optimization of website structure on microscope quite easy.

To make the whole system clear, Table.3 summarizes all the metrics in the web graph theory and give the main purpose of those metrics.

3. Case Study

In this section, we will test the theory of web metrics by study one academic web site. The domain name of this site is set anonymous for some reasons. The site belongs to a management school in a university. The main functions of this website are as follows: 1) the introduction for all departments and faculty; 2) marketing for all MBA project of this management school; 3) information release; 4) other functions. 484 pages/nodes are contained in this website. After parsing by program, we got 1796 hyperlinks between those nodes.

The computing processes and methods are described as followings:

Step 1: Based on the hyperlinks among the 484 pages, we get the adjacent matrix of the web graph. Then the adjacent matrix is transferred into the converted distance matrix. In the result largest distances 11, so we define the value of the K as 12. The matrix contains 484*484 items.

Step 2: Centrality metrics computing. Through some statistical tools of Microsoft Excel and SPSS 10.0, the data in converted distance matrix are transformed into the centrality metrics. The $CD = \sum_i \sum_j M_{ij}$ is equal to 2600778. Based on the definition of COD/CIDs and ROC/RICs the centrality metrics are easy to get. Fig.3 and Fig.4 shows the ROC and RIC series of this website by boxplot. Both ROC/RIC data are divided into 8 groups, every group contains 80(one group contains 84) nodes. The analysis of the box plots will be expatiated in the following section.

Step 3: Statistic data. Based on the definition in 2.1, the statistic terms of ROC/RIC are got. The mean of the ROC is equal to μ=484.5104, the Std. Deviation of the ROC is equal to σ =15.498, so the $\mu + \tau$ =531.0044. Also as for the RIC, the statistic terms are as followings: μ=508.879, σ=147.0667, $\mu + \tau$=950.0791.Table.3 and Table.4 show the details of the statistic result of ROC/RIC.

Step 4: Global metrics. To benefit the computing of stratum, the K in the converted distance matrix should be replaced by null. In this case N is equal to 484, which is an even number. Here the $LAP = N^3 / 4$ is equal to 28344976. The final result: Compactness= 0.079520372, Stratum=0.006839. Table.5 and Table.6 show the more concrete process of the computing.
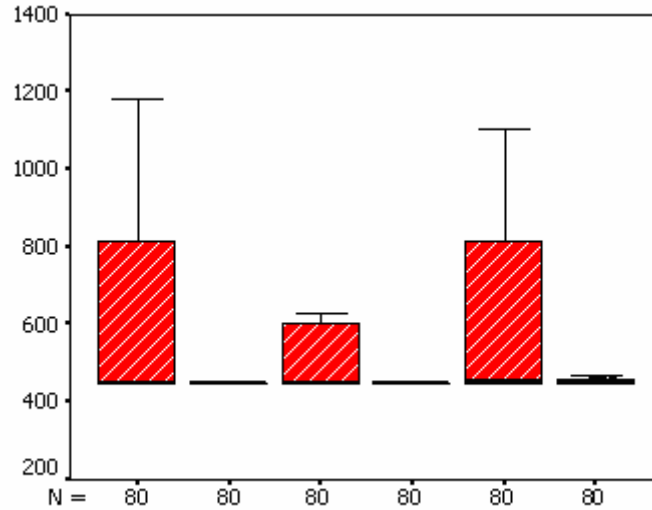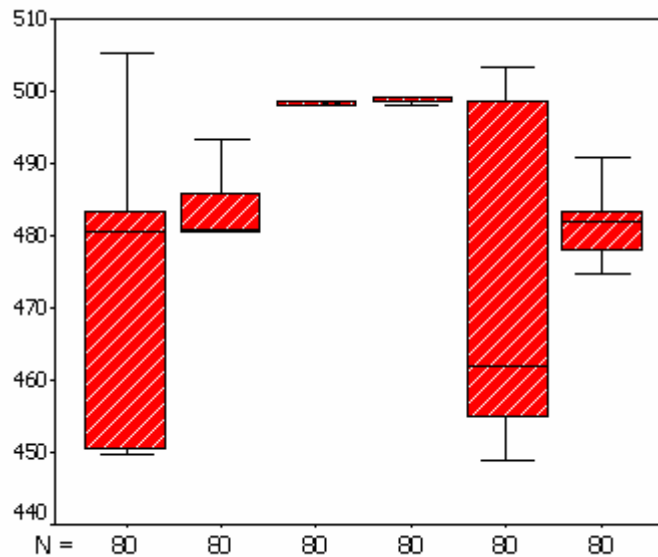
Fig.5 The distribution of the ROC

Fig.6 The distribution of the RIC

Table.4 The statistic terms of the ROC

	N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
ROC	484	448.72	1185.39	508.879	147.0667
	484	$\mu + 3\sigma$ (τ)		950.0791	

Table.5 The statistic terms of the RIC

	N	*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
RIC	484	448.72	505.4	484.5104	15.498
	484	$\mu + 3\sigma$ (τ)		531.0044	

Table.6 The computing process of Compactness

N	*Max=(N^2-N) K*	*Min= N^2-N*	*C*
484	2805264	233772	12
(Max - 2600778)/(Max - Min)	Compactness=0.079520372		

Table.7 The computing process of Stratum

N	LAP (even)=N^3/4	Total Absolute Prestige
484	28344976	193864
Stratum	0.006839	

Analysis

CID & COD Analysis. Through the distribution of CID and COD, some special phenomena were observed. No node has the obvious bigger value, so we can conclude that in the web no page has the high extent of cross-referencing. Through the data of COD, we can differentiate the leaf node clearly. Of course, in order to draw effective conclusion, analyzing the ROC and RIC is necessary.

RIC & ROC Analysis. By analyzing the distribution of the RIC data, it's observed that no node's value exceeds μ +3 σ (τ). This phenomenon is strange for a website. Based the premise of the RIC, there are no good in-landmarks. This is a normal problem in many website design. The suggestion for this problem will be proposed in the section 4. It's obvious that from the histogram the index nodes such as the main page have the lowest RIC among all the nodes. Therefore they are not easily reachable by any of the other nodes in the website. This phenomenon should be confirmed by the webmaster to see if that's the initial objective when designing the site. If the answer is not, further operations should be carried out to improve the website.

Analyzing the data of ROC, there are several nodes that have the obvious biggest ROC value, exceeding μ +3 σ (τ). These nodes are good out-landmark in this website. It's clear that from the root of this website any node is easily reachable.

Generally speaking, through analyzing the data of RIC & ROC, referencing nodes should be extensively added. Also, in the use of the website it's hard to reach the nodes on other branch from a low lever node.

Compactness Analysis. Through analyzing the value *compactness*, it's discovered that the website is not very compact. Also, from the distance matrix, there are some problems of the link structure. The value of the compactness is small. The low compactness indicate the insufficient links and parts of the website are possibly disconnected. To improve the compactness, we suggest that the designer should optimize the structure of index and add more referencing hyperlinks among pages. Also, the value of the compactness is too small. There should also be some improvement in the future work for the webmasters.

Stratum Analysis. Through analyzing the value stratum, its low stratum also shows that it's a low linear website. Maybe somebody unfamiliar with this website cannot find which pages he should read first if he want to get some specific information. We suggest that web site designer should boost the value of the stratum and use more structural clue to solve this problem. In the definition of Stratum the LAP may be is not the best way for normalizing. The stratum has too small a value, which is not enough sensitivity for analyzing. Proposition for new definition of LAP will be presented in the section 4.

The analysis above serves as suggestions for improvement of website based on web graph metrics. In the next section, advanced strategies for site design and algorithm improvement will be discussed.

4. Algorithm Improvement

4.1 Strategies for Design

It's believed that the analysis of the web graph will yield many interesting results and that as more research is done new problems and ideas will come forth. To obtain good performance on web graph metrics, a list of design strategies are as follows:

- Hierarchical structure design is necessary. Through the case study we can conclude that this web site lack macro-design. So the compactness is low and through the specific analysis of many nodes they are almost isolated in the web graph.
- Adequate cross-references are necessary. Cross-references will help us browse the website easily. But the amount of the cross-references should be controlled on the appropriate level. To certify this strategy, a test was implemented. A web graph contains 25 nodes and 30 links. The compactness is 0.26. But when added 5 cross-referencing links the compactness change to 0.44. Table.8 shows the change.

Table.8 Improvement placed by cross-referencing links

Num. Of Nodes	Num. Of Links	Compactness
25	30	0.26
25	35	0.44

- Emphasis on the analysis of index nodes. From the case study, the effects of index nodes are not so obvious. But in fact the index nodes play important roles in the whole website. Through the specific analysis of index nodes we can make the index nodes exert their effects best.
- Gauge the importance of the information correctly. In order to obtain a good geometrical balance, and obtain the good balance of information structure, the correct evaluation is necessary.

All the strategies are for good performance on web graph metrics. Those strategies are not enough for our design. More strategies should be produced through the analysis in practice based on web graph theory and all other theories.

4.2 Improvement of the Algorithm

Through the analysis of the case, some shortcomings of the web graph theory are observed. Some improvements can be made according to our findings.

Improvement Ⅰ: the value of the K should be set equal to twice the biggest value in the original distance matrix. The statistic terms will be more sensitive inducing easier differentiation. We find that the distribution of ROC and RIC is in a more rational situation. The result of ROC is given in Table7. It's clear that Std. Deviation is bigger than the value in table.4. It will be easier to differentiate the different nodes because the mean and the std.deviation all show more sensitivity. Also the RIC statistics terms are more sensitive. But here this algorithm improvement is just proved by the two group of data (ROC & RIC), of course the data proved right. If the improvement should be completely defined, it should be proved by mathematic theory. So the improvement is a applied one, However, this suggestion need to be further certified.

Table.9 The statistic terms of the ROC and RIC under new K

	N	Minimum	Maximum	Mean	Std. Deviation
ROC	484	324.3	1587.67	756.4	193.25
	484	$\mu + 3\sigma(\tau)$		1336.15	
RIC	484	324.3	837.44	563.71	46.37
	484	$\mu + 3\sigma(\tau)$		702.82	

Improvement Ⅱ: The definition of the LAP should be modified, because the differences between different stratums will be very small. In this case the value is just 0.006839. Several tests were implemented. They all shows too small value, such as 0.00158, 0.00374, which are not so obvious for differentiations. it's too small to give our conveniences for analysis. Our research shows that this formula definition is not so good for website analysis. A new definition of the complete linear graph and the formula should be established. It's known that when the N is even the absolute prestige of complete linear graph is equal to $N^3/4$,while when N is odd the value is equal to $(N^3 - N)/4$. here the improvement to enlarge the data is to add a radical sign on the definition of LAP and the Stratum. So the definition of LAP can be improved by using following definitions:

$$\begin{cases} LAP = \dfrac{\sqrt{N^3}}{2} \text{ ,if } n \text{ is even} \\ LAP = \dfrac{\sqrt{N^3 - N}}{2} \text{ ,otherwise} \end{cases}$$

Accordingly, $S = \sqrt{\dfrac{\sum_i \left(the\ absoulute\ prestige\ of\ node\ i \right)}{LAP}}$

The definitions are based on the Number-Theory. By use of this new definition, the value of the Stratum can be almost controlled between 0.1 and 1. For instance the change process of LAP value 0.006839 to 0.005248 will be another process 0.0827 to 0.0724. Obviously the differences between two are enlarged. So the more trivial difference will also be distinguished under this new definition of LAP. It's more sensitive and beneficial for other analyzing. For further studying, we hope we can work our a completely new definition of linear graph. Because the old definition is not so practical for analyzing. Maybe the meaning of the Stratum is doubtful. We long for further certification.

5. Conclusion

In this paper, the theory of web graph was generalized and a case study was given. Mainly the paper engaged in the problem of improving the performance on web metrics on website design by give a series specific strategies. On the other hand, several new ideas such as the classification of the local metrics and the solution for some defects of the algorithm are proposed. However, many issues need to be solved in the further research. They are:

1) The parameters such as LAP in the algorithm need to be better normalized;
2) More practices are necessary to promote the research on applying web graph theory;
3) Converted distance based theory and out/in degree based theory need further research to integrate the advantages of the both.

The long term of the research aims at better design for website to promote the e-commerce.

More research work and experiment are needed to achieve this objective.

ACKNOWLEDGEMENTS

References

Botafogo R. A.and Shneiderman B., "Identifying Aggregates in Hypertext Structures," Proceedings of the third annual ACM conference on Hypertext, 1991, pp. 63-74."

Botafogo, R., Rivlin, E., and Shneiderman, B. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. ACM Trans. Inf. Syst. 10, 2 (Apr.), 1992. 142-180."

Bray, T. Measuring the web. In Proceedings of the 5th International World Wide Web Conferences (Paris, France, May, 1996.)."

Dhyani, D., Ng, W. K., and Bhowmick, S. S. A survey of web metrics. ACM Computing Surveys, 2002,34(4):469-503."

Kleinberg. J., Kumar. R., Raghavan. S., and Tomkins. A., The web as a graph: Measurements, models, and methods. In Proceedings of the 5th International Conference on Computing and Combinatorics(COCOON), 1999."

Kumar. R., Raghavan. P., Rajagopalan. S., and Tomkins. A., Trawling the web for emerging cyber-communities. In Proceedings of the 8th World Wide Web Conference, 1999."

Nielsen, J. 1999. User interface directions for the web. Communications of the ACM, 42(1): 65-72"

Pitkow, J. In search of reliable usage data on the WWW. In Proceedings of the 6th World Wide Web Conference, 1997."

Wang Youwei, Wang Dingwei. Research on design strategy of E-Supermarket Webpages. Journal of Systems Engineering (In Chinese). 2003.11, Vol.18, No.2, 104-108"

Yen B. P. -C., Fu K., Accessibility on Web Navigation, On International Conference on eCommerce Engineering(IceCE2001), 2001, Xi'an. "