A Survey of Metrics for Assessing the Navigational Quality of a Website Based on the Structure of Website.

Sumit Chhabra
Dept of Computer Science and Applications
Khalsa College for Women
Amritsar (India)
sumitchhabra_12@yahoo.com

Mini Singh Ahuja
Dept of Computer science and Engg
GNDU Regional Campus
Gurdaspur (India)
minianhadh@yahoo.co.in

Pritika Mehra
Dept of computer science and applications
Khalsa College for Women
Amritsar (India)
mehra_priti@yahoo.com

Abstract: **The World Wide Web, abbreviated as WWW and commonly known as the Web, is a system of interlinked hypertext documents accessed via the Internet. With a web browser, one can view web pages that may contain text, images, videos, and other multimedia and navigate between them by using hyperlinks. Navigation is the process through which the users can achieve their purposes in using Web site, such as to find the information that they need or to complete the transactions that they want to do. A navigability measurement metrics is a well-defined mathematical formula that maps Web sites or portals to a numerical system that indicates the navigability of the Web sites/portals. In this paper, we have presented a brief study of Web metrics, such as average out-going links from a page, fan outs of a Web site compactness, stratum, lostness etc to determine the navigational efficiency or quality of a website.**

Keywords—**naviagtion, website structural complexity , lostness or disorientation, URL (Uniform Resource Locator), hypertext, hyperlink.**

I. INTRODUCTION

A. *WWW*

The concept of WWW was given in 1989 by Tim Berners-Lee while at CERN (the European Laboratory for Particle Physics). The WWW system is based on client/server architecture. A web browser works as an HTTP client because it sends requests to an HTTP server which is called Web server. Upon receiving HTTP request, a Web server executes the associated action and sends a Web document or a file to the browser client that sent the request. The documents that the browsers display are hypertext documents. Hypertext is text with pointers to other text. The browsers lets a user deal with the pointers in a transparent way -- select the pointer, and the user is presented with the text that is pointed to. Hypermedia is a superset of hypertext - it is a medium

with pointers to other media. This means that browsers might display an image, sound or animation on the clicking of a pointer.
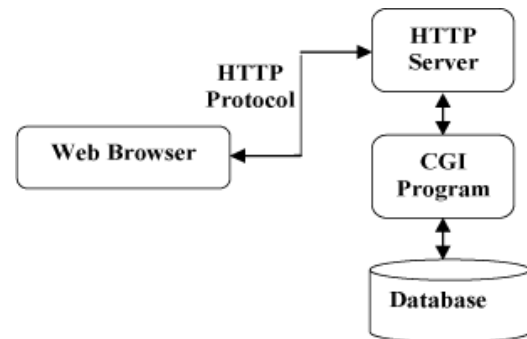
Figure 1: Working of WWW

Web pages are identified by a Uniform Resource Identifier (URI), also known as uniform Resource Locator (URL) or simply a Web address. Its general structure is as follows:

http://host:port/path?query#fragment

The host part is either a numerical Internet (IP) address or a registered, logical name like 'www.foo.com' that is resolved to an IP address by a domain name server. It identifies the Web server to which a connection should be made. Optionally a port can be specified through which the connection should take place. The path specifies the location of the file on the web server; this may be a directory path and a filename, but many dynamic web servers use alternative database or script commands that are not human legible. The query string contains data to be passed to web applications such as CGI programs .The fragment when used with HTTP specifies a location on the Web page requested,

typically a marker to which the Web browser should automatically scroll.

Web pages are written in the Hypertext Markup Language (HTML). Within running text, markup tags are used for structuring and laying out text and images, inserting links and scripts, and other non-textual items. The actual contents of a Web page are preceded by a header, which provides meta-information, such as the page title, author, language, and the date of creation. Current web sites often make use of external cascading style sheets, which specify the manner in which all tags should be rendered and improves page layout consistency. Interactive screen elements can be integrated using JavaScript and applets-code that is executed on the client side.

Individual HTML files with unique addresses are called Web pages, and a collection of Web pages and related files (such as graphics files, scripted programs, and other resources) sharing a set of similar addresses is called a Web site. The main or introductory page of a Web site is usually called the site's home page. Users may access any page by typing in the appropriate address, search for pages related to a topic of interest by using a search engine, or move quickly between pages by clicking on hyperlinks incorporated into them.

B. Web Metrics and Quality Determination

Quality is defined as the level to which a product conforms to its requirements. Quality is by nature a multi-faceted concept that means different things to different people. The concept of quality depends highly on the entity of interest, the viewpoint on that entity, and the quality attributes of that entity. A good quality car for a family would probably be one that has enough room for all the members of the family, has an economical engine and is safe in the event of a collision with another car. On the other hand, a good quality car to a race-car driver would be one that is lightweight, has high acceleration, good brakes etc.

Software quality attributes are a high-level a set of attributes of a software product by which its quality is described and evaluated. A software quality attribute may be refined into multiple levels of sub-attributes. High-level quality attributes are at a high level of abstraction or generalization that can usually be broken down into sub-attributes. For example, the attribute reliability can be broken down into the sub-attributes: Maturity, Fault-Tolerance, and Recoverability. Achieving these more specific sub-attributes will mean achieving the overall attribute of reliability. The sub-attributes can occasionally be again divided into sub-sub-attributes. These forms a tree of attributes starting from the most general (and high-level) attribute at the root and the most specific (and low-level) attributes at the bottom. Quality assessment process can be subjective or objective concept. Subjective quality measurements emphasize the necessity of design guidelines and a development culture that encourages simplicity, intuitiveness and understandability of software designs to humans.

In objective quality assessment, metrics are used as the primary tool for assessing quality which quantifies some characteristic or attribute of a computer software entity. For example, a FileSize metric, which is the total number of characters in the source files of a program and is used to determine the measure of a particular program, such as 3K byte. Metrics refer to standards of measurement and measures attributes of different types of entities. Website is an entity. Therefore, web metrics are standardized ways of measuring attributes related to the Website for e.g. design, usability, navigation etc. Navigation plays a crucial role in the design of website structure because it determines the path to be traveled to reach a required web page

II. NAVIGATIONAL METRICS

Generally speaking, navigation is the process of determining a path to be travelled through a chosen environment. Navigation comes from two Latin words: *navis* (ship) and *agrere* (to drive). According to the Merriam-Webster Dictionary, the general meaning of "navigation" is "to steer a course through a medium, to get around, move, to make one's way over or through and to operate or control the course of." Therefore, navigation is the action or process of determining the position and directing the course to be travelled through a given environment. In the environment of a Web site, navigation is the process through which the users achieve their purposes in using Web site, such as to find the information that they need or to complete the transactions that they want to do. Therefore, a navigability measurement metric is a well-defined mathematical formula that maps Web sites or portals to a numerical system that indicates the navigability of the Web sites/portals. Typical examples of such Web site complexity metrics are average out-going links from a page, the average fan outs of a Web site, measures of lostness, mental models and so forth.

A website can be modeled as a graph consisting of a pair <G, S >, where G= (V, E) is a directed graph representing the website; V is the set of nodes representing web pages; E is the set of edges representing links between web pages; and S is the start node of the graph, i.e. the home page of the website.
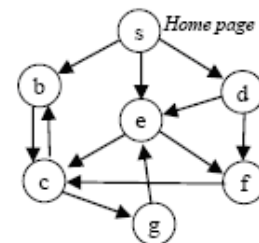

Figure 2: Web Graph

Structural complexity emerges from the relationships among the pages of the website. The most basic and important relationship is that a page is linked to another through hyperlinks. The hyperlinks between web pages of a website form the navigational paths through which users browse the website to find the information that they want. The more complex that the web pages are inter-linked, the more likely that a user becomes lost in the information ocean, and hence, the more difficult to navigate. Therefore, structural complexity metrics can be used to quantify the navigational quality of website.

A. Centrality Metrics

According to Botafogo et al. [7] centrality metrics defines the relationship between one node and the others. To formalize the notion of centrality, the sum of distances from a node to all other nodes in the web graph is used. A matrix called distance matrix is used to store the entries of distances of each node to every other node of the graph. When a node does not reach another node in the website, the entry in the distance matrix is infinite. Because of the many infinite entries in distance matrix, it won't bet easy to distinguish which node is more central. Therefore, the converted distance matrix is defined as follows:

Let C be the converted distance matrix and M the distance matrix .Thus

$$C_{ij} = \begin{cases} M_{ij} & \text{if } M_{ij} \neq \infty; \\ K & \text{otherwise} \end{cases}$$

The value of K is usually set to a value bigger than the biggest distance among the node reachable to each other. K is called conversion constant.

The metrics that can be used to define the notion of centrality are:

Converted out Distance (COD): The converted out distance (COD) for a node i is the sum of all entries in row i in the converted distance matrix (C). Formally,

$$COD_i = \sum_j C_{ij}$$

Converted in Distance (CID): The converted in distance (CID) for a node i is the sum of all entries in column i in the converted distance matrix. Formally,

$$CID_i = \sum_j C_{ji}.$$

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

$$CD = \sum_i \sum_j C_{ij}.$$

A central node is one whose distance to all the other nodes in the hypertext is small. As that distance grows, nodes become less central. Consequently, the smaller the COD the more central is the node. For a single Web graph, the COD is a good indication of the node centrality as compared with another node, but this number indicates little when two different Web graphs are compared .So to overcome this problem, two metrics can be used. Relative out Centrality (ROC) metric for node i is defined as the COD of node i dividing CD. The higher the ROC metric of a node, the more central is the node (the inverse of the COD).

$$ROC_i = CD/COD_i.$$

Similarly, the relative in centrality (RIC) metric can defined as

$$RIC_i = CD/CID_i.$$

B. Compactness (Cp)

Botafogo et al. [7] also stated that the readability and navigation of a website can be expressed through Compactness metric. The compactness indicates the interconnectedness of a hypertext. Its value lies between 0 and 1.If the value of compactness is 0 for a web graph, then the corresponding website is completely disconnected. If the value is 1, then the website will be a fully connected website.

Formally the compactness is defined as:

$$Cp = \left(\text{Max} - \sum_i \sum_j C_{ij}\right)/(\text{Max} - \text{Min}).$$

Cij is the distance from node i to node j. If node i is not connected to node j the distance between them is infinite, then the distance Cij is set equal to a constant K. This constant is called the conversion constant. Max is a parameter that depends only on the number of nodes in the graph and the converted distance. It is the distance from every node to every other node when the graph is completely disconnected. Max = $(n^2 - n)$ C, where n is the number of nodes in the hypertext and C is the maximum value an entry in the converted distance matrix can assume. Usually, C = K, where K is the conversion constant .The distance from a node to itself is always zero. Min is defined in a similar way as Max, but in this case the graph is completely connected. Min = $(n^2 - n)$. When the graph is completely connected, the distance of a node to any other node is equal to 1.
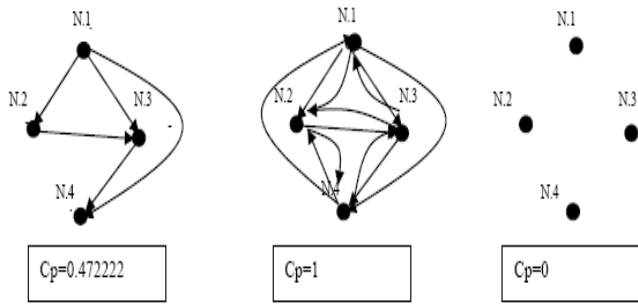
Figure 3: Relation between the shape of the graph and the value of the compactness

Both 0 and 1 should be avoided for high readability and navigation when designing the website. A too high compactness means that each node has many links and that consequently there are potentially many cycles. Traversing many cycles can disorient users. On the other hand, a too low compactness indicates insufficient links and that possibly parts of the hypertext are disconnected. Figure 3 shows the relation between the shape of the graph and the value of the compactness is. 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.

C. Stratum (St)

The Stratum Metric is used to characterize the linearity in the structure of a web graph. Like the Compactness metric, stratum values range from zero to one, with St=1 indicating a strictly linear sequence of nodes (that allows one and only one path within the network), while St=0 when a network is fully connected (every node is connected to every other node).

Calculation of the Stratum metric begins with the distance matrix and two concepts originally defined in social network theory: status and contrastatus. One common application of these terms is in social networks for establishing a system for seniority. In this kind of system, the status of an individual refers to the number of persons who are subordinate to that individual, while contrastatus refers to the amount of status weighing down on an individual from "above", and is calculated by summing numbers of superordinates (i.e. "bosses") for an individual. Prestige is defined as the status of an individual minus the contrastatus of that individual, and the absolute prestige for a network is calculated by summing the absolute values of finite prestige values for all the nodes in the network. Unlike the status and contrastatus measures, prestige ranges across both positive and negative values. Large positive values for individuals represents high level in the seniority order, prestige at or near 0 represents middle level and large negative values represents low level individuals in the seniority order. The absolute prestige (or the

absolute stratum) of the organization is defined as being the sum of the absolute values of the prestige of each individual. The status and contrastatus are calculated as COD and CID are, by summing across rows (status) and down columns (contrastatus), ignoring infinite values which indicate that nodes are unconnected. Moreover, in a distance matrix without infinite values, COD=status and CID=contrastatus. These stratum-related measures are absolute and therefore do not provide a suitable basis for comparisons across networks, which require a normalized measure. To overcome this problem, Botafogo et al. [7] define the linear absolute prestige (AP) of a network with n nodes, showing that

$$LAP = \begin{cases} \dfrac{n^3}{4}, & \text{if } n \text{ is even.} \\ \dfrac{n^3 - n}{4}, & \text{if } n \text{ is odd.} \end{cases}$$

And then go on to formally define the Stratum (St) of a network as

$$St = \frac{absolute\ prestige}{LAP}.$$

D. Mc. Cab's cyclomatic complexity metric

Sreedhar et al. in their paper "Measuring Quality of Web Site Navigation" [11] proposed that the structural complexity of website can be determined with Mc. Cab's cyclomatic complexity metric. This metric is used to know navigation path for a desired web page. The cyclomatic complexity metric is derived in graph theory using a tree graph. A tree graph can be constructed with home page as root and website by considering various hyperlinks in the website. The tree consists of various sub trees and leaf nodes. Each sub tree of the graph represents a web page which has further hyperlinks to the next web pages and leaf node represents a web page which does not have further links to any web pages. In tree graph, all web pages that do not have further links are represented with one leaf node and a sub tree at each level consists of links to the web pages to the next level. The cyclomatic complexity can be calculated using equation as follows:

$$CC_{Metric} = (e-n+d+1)/n$$

where e = number of web page links
   n = number of nodes in the graph
   d = number of leaf nodes in the graph.

E. Path length metric

Sreedhar et al. [11] also suggested path length as a measure of navigational efficiency. A path length is used to find average number of clicks per page. The path length of the tree is the sum of the depths of all nodes in the tree. It can be computed as a weighted sum, weighting each level with its number of nodes or each node by its level using equation.

$$\text{Path length} = \Sigma \; l_i * m_i$$

Where l_i is level number i,
   m_i is number of nodes at level i.

The average no. of clicks is computed using equation.

$$\text{Average no. of clicks} = \text{path length}/n$$

where n is the number of nodes in the tree .

F. Outgoing Links

According to Zhu et al. [2] the number of outgoing links of a Web page indicates how easy it is to get lost, since each outgoing link represents a choice for the next step in navigation. The following metric is defined as the total number of outgoing links within a Web site.

$$OutLinks(W) = \sum_{n \in Node(W)} Out(n)$$

where W is the Web site to be measured, Node(W) is the set of the pages of the Web site W, Out(n) is the number of different Web pages that the node n links to. The metric Outgoing Links catches the intuition that a small Web site, with fewer pages and links, is less complex than a large Web site that has hundreds even thousands of pages and links. However, for comparison purposes, it is desirable to know its relative complexity taking size into consideration. Dividing the overall complexity by the number of pages gives a normalized complexity defined as average number of out links.

$$AverageOutLinks(W) = \sum_{i=1}^{n} outlink(i) \Big/ n$$

G. Fan Out

The research on software measurement suggested that complexity increases with the square of connections (fanout) where fanout is number of the calls from a given module. So Zhu et al. [2] suggested that in a web site design, all pages are connected by hyperlinks. Thus the fan out metric for Web site structural complexity can be defined as:

$$FanOut(W) = \sum_{n \in Node(W)} Out(n)^2$$

H. Broken link calculation

According to Sreedahar et al. [1] navigation problem in a website structure is due to broken links. Broken links can be identified by the PowerMapper[13] tool. In a website, broken links are counted at various levels and broken error index is calculated based on percentage of broken links. The percentage of broken links is calculated using equation:

Percentage of Broken links (PBL) = (number of broken links / number of web pages)* 100

I. Lostness or Disorientation

Looking for information on the World Wide Web is often not as efficient and easy as users might like. If websites are not well structured, users can become disoriented or lost and will not be able to (efficiently) find the information they want. Being "lost in hyperspace" is actually one of the most common problems users experience while navigating and affects performance and user's satisfaction making users to become frustrated and lose interest . Lostness or the tendency to lose one's sense of location and direction in a website has been explored in different ways. One approach is to infer users' levels of lostness from analyzing user's actions when browsing. Smith [6] created metric to measure lostness. This metrics rely on the path taken by user to reach the information required within a Hypertext web document. The path measurements from which indicators of lostness and efficiency are derived are:

T the total number of nodes accessed;
D the number of different nodes accessed;
R the number of nodes which need to be visited to complete a task;
S the total number of nodes visited while searching;
N the number of different nodes visited while searching;
V the number of nodes visited while verifying.

T and D were intended to provide information on activity and certainty. It was considered that a lost user might be expected to access significantly more nodes than necessary (R). The final three path measures are designed to distinguish between searching and verification, based on the notion that one is not lost when verifying.

The assumption is that for a perfect search, a user will visit exactly the number of nodes required to complete an information retrieval task. Therefore:

T =D = R; S = T

N/S =1; R/N = 1

Smith's [6] measure of lostness (L) is calculated as:

$$L = (N/S - 1)^2 + (R/N - 1)^2$$

This formula calculates the degree of lostness on a scale from 0 to 1, where 0 indicates no lostness at all. From her usability study, Smith [6] concluded a user is lost when L is 0.42 or higher.

J. Link weighted lostness metric

Smith's metric of lostness does not take into account the fact that different types of links might have different degrees of influence on lostness. When a hypertext system is being designed, if the designer fails to take into account the different structural aspects of the hypertext, then it is likely that the user of such a system will experience disorientation. Therefore, Otter and Johnson [4] designed a new measure of disorientation called link weighted lostness .The link weighted lostness metric (LWLM) is based on Smith's [6] metric and is calculated in a very similar way:

$$LWLM = L/(LW/4)$$

where L is the Smith's lostness metric and LW is the Link weightings. Link weightings depend on types of links. According to De Rose's [12] taxonomy of links, there are four types of links:

1. Annotational links are given the lowest likelihood of inducing lostness because they link one node to one other node and the destination node is made explicit in the link label.
2. Sequential links are given the next lowest likelihood of inducing lostness. The reason for this is that once a user is in a path or sequence, nodes within that pathway will be chunked in the memory as a procedural unit, so that the exact location and content of each individual sequential node need not be remembered, thus reducing cognitive load.
3. Taxonomic links are given a medium likelihood of inducing lostness because they have multiple target locations and there is not the same path-like ordering imposed on them as exists for sequential links.
4. Associative links are given the highest likelihood of inducing lostness because they link one node to one other in a highly unpredictable and arbitrary manner.

If the subject when completing an information retrieval task, uses predominantly associative links (high likelihood of inducing lostness) which have a link weighting of 1, then L would be divided by a number close to a quarter (1/4) thus weighting L quite heavily. If the user used predominantly taxonomic links then L would be divided by a number close to a half (2/4). If the user used predominantly sequential links then L would be divided by a number close to three quarters (3/4). Lastly, if the user used predominantly annotational links (low likelihood of inducing lostness) then L would be divided by a number close to one (4/4) thus putting very little or no weighting on L.

LWLM increases as lostness increases. The only difference between this new metric and Smith's is that once Smith's value of L (lostness) has been calculated, then it is then divided by the total link weighting for all of the nodes visited by that user, per task. In order to retain the fact that the new values will fall between zero and one, the total link weighting is itself divided by four. Thus, if the value of L is zero (indicating that the subject, according to Smith's metric was not lost at all), then the link weightings automatically do not apply as zero divided by anything is still zero.

K. Mental Models

Otter and Johnson [4] offered another metric to explore lostness: the accuracy of the mental models users have about the hypertext's structure. Mental models are conceptual and operational mental representations that users develop while interacting with the system. These representations provide explanatory and predictive power, but often they are incomplete representations of how to work with the system or to navigate through the website. The main assumption is that if users have a poor mental model of the website, then it is likely that they will experience disorientation.

Otter and Johnson [4] asked users to draw paths (from memory) of the website from the start page to a certain end point. The end points were always in links not previously visited by the user. In this way, users could not produce the answer based on retrieval cues but using their mental models of the structure of the website. From these pathways it is possible to calculate the accuracy of

mental models for hypertext systems (AMMH):

$$AMMH = ⅓(C / AD + CCP / RD + LBE / RD)$$

Where:

AD = Number of nodes actually drawn

RD = Number of nodes required to be drawn (nodes in the optimal or shortest path)

C = Number of nodes that are correct (regardless of their placement)

CCP = Number of nodes that are both correct and correctly placed

LBE = Number of different levels of a hierarchy correctly drawn before the subject made an error, i.e. the degree of hierarchical depth achieved before an error occurs

AMMH has a minimum value of 0 (a very poor mental model) and a maximum of 1 (a very good mental model).

L. Revisits

The revisitation measure is used for calculating the probability that any URL visited is a repeat of a previous visit, using the ratio of the different pages visited to total pages visited. Catledge and Pitkow [5], calculated the probability that any URL visited is a repeat of a previous visit, using the following formula:

$$Revisits = \left(1 - \frac{DifferentURLsVisited}{TotalURLsVisited}\right) \times 100\%$$

According to Smith [6], a higher amount of revisitation indicates that a user is more likely to be lost.

CONCLUSION

In this paper we have discussed various metrics which can be used to measure the navigational efficiency of websites based on the structure of websites. Efficiency of websites can be determined by quantifying the structural characteristics of websites.

REFERENCES

[1] G. Sreedhar, A.A. Chari and V.V.VenkataRamana, "A Qualitative and Quantitative Frame Work for effective Website Design", *International Journal of Computer Applications*, pp: 71-79, 2010.

[2] H. Zhu and Y. Zhang, "Navigability Design and Measurement", *Encyclopedia of Portal Technologies and Applications*, pp: 642-646, 2007.

[3] J.S.Ahuja and J. Webster, "Perceived disorientation: an examination of a new measure to assess web design effectiveness". *Interacting with Computers*, pp.15-29, 2001.

[4] M.Otter and H. Johnson, "Lost in hyperspace: metrics and mental models", *Interacting with Computers*, pp. 1-40, 2000.

[5] L.D.Catledge and J.E Pitkow, "Characterising Browsing Strategies in the World-Wide Web", *Computer Networks and ISDN Systems*, pp. 1065-1073, 2001.

[6]P. A. Smith, "Towards a practical measure of hypertext usability" *Interacting with Computers*, pp. 365-381, 1996.

[7]R.A. Botafogo, E. Rivlin and B. Shneiderman, "Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics", *ACM Transactions on Information Systems, USA*, pp. 142-180, 1992

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

[9]Y. Zhang, H. Zhu and S. Greenwood, "Website complexity metrics for measuring navigability", *Proceedings of the Eighth International Conference on Quality Software*, pp.172-179, 2004.

[10] L. Tauscher and S. Greenberg, "How people revisit web pages: empirical findings and implications for the design of history systems". *International Journal of Human-Computer Studies*, pp. 97-137, 1997.

[11] G. Sreedhar, A.A. Chari and V.V.VenkataRamana," Measuring Quality of Web Site Navigation", *Journal of Theoretical and Applied Information Technology*, pp: 81-86.

[12] S .De Rose, "Expanding the notion of links", *Proceedings of the Hypertext Conference. Chapel Hill, North Carolina*, 1989.

[13] http://www.powermapper.com/

[14] G. Sreedhar, A.A. Chari and V.V.VenkataRamana, "Evaluating Qualitative Measures for Effective Website Design", *International Journal on Computer Science and Engineering*, pp.61-68, 2010.