Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 117–122

ISBN 978-83-60810-22-4

Problem of website structure discovery and quality valuation

Dmitrij Żatuchin
Institute of Informatics, Wrocław University of Technology, Wybrzeże Wyspiańskiego 27,
50-370 Wrocław, Poland,
Email: Dmitrij.Zatuchin@pwr.wroc.pl

***Abstract*—Navigation as a part of an interface was always an important issue of a design process. Because information architecture and the navigation of current websites are very complex, especially of e-commerce websites or information portals, it is very hard to analyze or redesign a structure in a manual way. In order to solve the problem of automation of website structure analysis, there should be defined its model. Also, during the study of a subject it was found, that there is a lack of a quality estimator, which allows to valuate in a vary moments the quality of the structure. Observation of a structure quality gives possibility to analyze and decide when the structure should be changed basing on decision rules or calculated thresholds for analyzed amount of time.**

The main aim of this study is to describe a model for website structure representation, derive the quality estimator, define and solve the problem of website structure discovery and quality valuation utilizing the proposed metric.

Finally, experiment with utilization of proposed methods is presented.

I. Introduction

THE main task of the website is to provide the content and functionalities of the system. Such functionalities usually are placed to the sub-pages of the website. The main task of the navigation is to service efficiently and effectively users' requests, which are provided through these functionalities. Website usability is the measure of a success which users experience while interacting with the system. This is the extent to which users can achieve the desired objectives during their visit. Some of the website usability factors include: the compatibility of the site layout, ease of use of a search engine, adequate links that provide instant access to information, a site map which serves as a table of content for the whole site, legible fonts and appropriate use of colors to highlight and organize information or functionalities. All these factors contribute to the ease of use and make a visit on the website useful and enjoyable [1]. In 2001 Donahue [2] pointed, that difficult navigation with the limited flexibility constitutes the major problem of the usability. Therefore, a good solution for the navigation should be provided [3]. According to the National Institute of Standards and Technology [4] ease of navigation is essential for users at all levels of proficiency in using computers in order to navigate and obtain the desired information on the websites. It can be stated, that the number of planned functionalities at the design stage influences on the final number of pages and links [5]. Also it may be stated, that the increase of the quality of the interface is possible by increasing the quality of the structure of the navigation.

Users may come to the website in many ways: through the main page, as a result of the reference link from other websites, from the search results or an advertisement. Moreover, users have different goals and their objectives are very diverse. Navigation should support those differences through a variety of solutions: a hierarchical, task-oriented, chronological, alphabetical, and based on the popularity of information architecture [1]. It is required in order to avoid situation, that user will be trapped inside one page or reach the orphaned page [6]. On the static website or dynamic one with a limited number of links the navigation structure has a large impact on the quality of delivered content. Therefore, evaluation and improvement of the website structure becomes a key issue what repeatedly was underlined by researchers and experts of Human-Computer Interaction field of studies [7], [8], [9], [10], [11], [12], [13], [14], [15].

II. Website Structure Discovery

For the problem of quality valuation of the website structure there are some estimators in the literature of subject. Unfortunately, none of them treat the structure as a network and take into account such information as connections between pages, popularity of single pages, their position or utilization of edges connecting these pages among the website. Few proposed [16], [17] treat the website as a tree and in order to estimate aptitude of a structure they utilize knowledge such as click number into a page and distance from a root item. What is important is that distance from a root items is treated like in height in tree structures, what is not true for the website.

The motivation for solving the problem of a structure discovery with maximum data utilization and the problem of a structure quality measurement comes from a need to increase the efficiency of the interface redesign process. It can be reached by including into the process an automatic analysis of usage data and generating recommendations to change

The research in this paper has been partially supported by: European Union in the scope of the European Regional Development Fund program no. POIG.01.03.01-00-008/08 and European Social Fund – fellowship "Mloda Kadra".

existing website structures into modified ones – adapted to users' needs.

A. Model of a website structure

The mere idea of using graphs to model the structure of the website is widely used by practitioners and researchers. First, Perkowitz [18] used to describe links on a single webpage with the graph structures. In Garofalakis' studies [19] model of a website was described as a tree, because of optimization simplification. Yen et al. [20] defined the environment consisting of three layers of evaluation and expansion of website projects using graphs for modeling. In 2008 Yang defined the conceptual model for the structure [21] thus using ontological paradigm. Another way to model the website is to use Markov chains [22]. Also there can be found works inspired by PageRank algorithm [23], where web mining is used in order to optimize website structure [24].

The proposed model of a structure is understood as a set of unique pages which may be reached with an internal linking through navigation elements (Fig. 1b). The structure is mapped by a set of nodes and connections (Fig. 1a,c). The first node in a structure is called a root node, which is accessible through the default domain or IP address in a browser and returns the main page of a website. A navigation element may occur on one or many pages, but each time it may consist of a different set of pages and links. Connections between pages may be directed or undirected (Fig.1c).

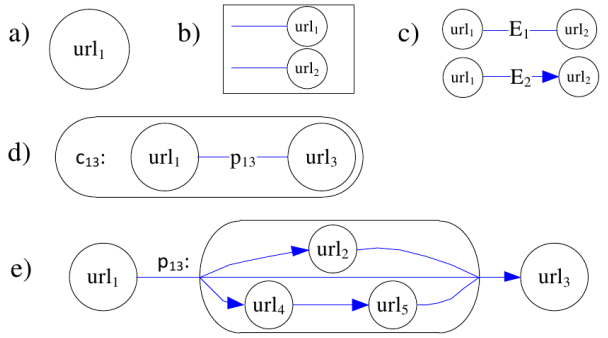

Fig. 1. Elements of website structure: a) a subsite, b) a navigation element, c) connection between two sites, d) path, e) connection.

Such defined website l is modeled as a graph structure in equation (1) of N nodes (pages) and M edges (links), where URL_l is a finite set of pages (1) which are in one domain in the range of l website and E_l (1) is a finite set of links which does not contain loops or reverse connections, where e_{li} is an edge from url_{lm} to url_{ln} to .

$$GI_l = (URL_l, E_l, P_l, C_l),\ URL_l = \{url_{li} \mid i \in [1,N]\}$$
$$E_l = \{e_{li} \mid e(m,n), m \neq n, m,n \in [1,N], i \in [1,M]\} \quad (1)$$

Basic elements of a structure model paths P (Fig. 1d) and connections C (Fig. 1e).

Path is a set of edges, which are the passage from one node to another. Paths are directed – if between nodes there is one edge, or transitive – if on a path there is more than one intermediary node. Set of paths P_l is defined in (2).

$$P_l = \{p_{lij}, i,j \in [1,N] \mid \{e_{li} \prec e_{lj} \mid \exists p_{liz} \wedge \exists p_{lzj}, z \in [1,N]\}\} \quad (2)$$

A connection between two nodes exists if it is possible to pass from one node to another. The set C_l contains all connections in a graph, see equation (3).

$$C_l = \{c_{lij}, i,j \in [1,N] \mid \forall c_{lij} \exists p_{lij} \rightarrow url_{li} \prec url_{lj}\} \quad (3)$$

The length of from to is an average length of all paths between nodes (4).

$$D(c_{lij}) = \overline{p_{liJ}} \quad (4)$$

Nodes and edges have additional characteristics and indicators, which will be used in a proposal of formulation of a quality estimator.

B. Node characteristics and indicators

Every node in the structure has a distance from the root node, which is calculated according to the distance function in equation (5). For the root node the distance equals 1. Pages are usually visited by users in such way, that nodes on lower levels are rarely visited. The form of equation was derived as a result of analysis of statistics gathered by Google Analytics of approximately 500 000 different users visiting 12 websites from different categories and left by them paths of visits.

$$d(url_{li}) = pow(\min|P(url_{l1}, url_{li})| + 1, \tfrac{1}{\sqrt{2}}) \quad (5)$$

Another characteristic for a node is *input-output* defined in equation (6), which is the sum of number of nodes, which edges point to the processed node and number of outgoing edges from the processed node (Fig. 2a).

$$io(url_{li}) = |\{e_{lj} \mid \exists e(i,n) \vee \exists e(m,i); i \in [1,N], j \in [1,M]\}| \quad (6)$$

Node indicators determine the characteristics based on usage data. These are:

- *OccU(url_{li}, [t, t+τ_k])* – specifies the number of edges used in all paths of all users of the website during the interval of time *[t, t+$_\tau$k]*;
- *PopU(url_{li}, [t, t+$_\tau$k])* – popularity of a node relatively to the entire structure of the website at the interval of time *[t, t+$_\tau$k]*, defined as equation (7);

$$PopU(url_{li}, [t, t+\tau_k]) = \frac{OccU_k(url_{li})}{N_l} \quad (7)$$

- *Acc(url_{li})* – availability of a node in the structure defined in (8), is an ease for url_{li} to be visited by users and depends on the distance from the root and *io* characteristic;

$$Acc(url_{li}) = \frac{io(url_{li})}{d(url_{li})} \quad (8)$$

C. Edge characteristics and indicators

Characteristics for edges are:

- *Out($_{el}$i)* – specifies the number of all edges that come from the same node as the beginning of the edge. This indicator is illustrated in Fig. 2b and is defined as (9);

$$out(e_{li}) = |\{e_{lj}|\ parent(e_{li}) = url_{lm} \wedge parent(e_{lj}) = url_{lm}\}| \quad (9)$$

- *Reach(e_{li})* characteristic defined as (10) returns the number of nodes that can be achieved through following the edge (Fig. 2c). If a node, which indicates an edge, is a leaf, then the value of the characteristic equals 0.

$$reach(e_{li}) = |\{url_{lj}|\ e_{li} \in p_{lxj}; x, j \in [1, N]\}| \quad (10)$$

Indicators are defined as follow:

- *OccE(e_{li}, [t, t+τ_k])* – specifies the number of occurrences of edges in all paths in the interval of time *[t, t+$_\tau$k]*;
- The indicator *r* in (11) is an information about traffic between two nodes;

$$r(m,n) = \sum_{i=1}^{|p_{lij}|} OccE(e_l, [t, t+\tau_k]), e_l \in p_{lij} \quad (11)$$

- *PopE(e_{li}, [t, t+τ_k])* is the edge's popularity (12) – a number of occurrence of an edge in all paths obtained from usage data in the interval of time *[t, t+$_\tau$k]*;

$$PopE(e_{li}, [t, t+\tau_k]) = \frac{OccE(e_{li}, [t, t+\tau_k])}{|P_l|} \quad (12)$$

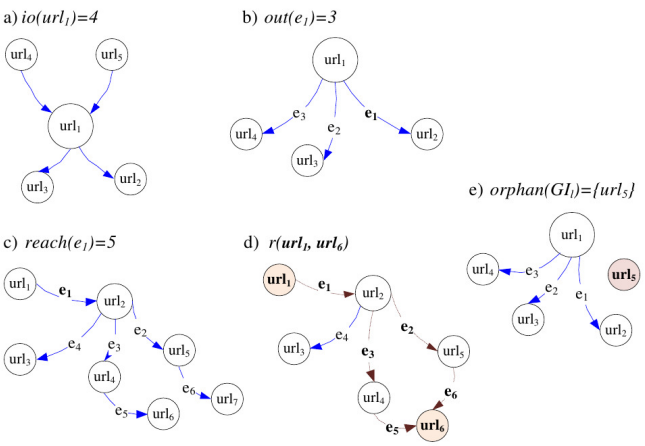

Fig. 2. Illustration of some characteristics of a model in a website structure.

D. Problem of a website structure discovery

Structure discovery is a required procedure for quality valuation task.

For given: URL address of the *l* website; usage data collected for the website *l*; time *k*;

Determine: website structure based on navigation patterns modeled as a graph GI_{lk}.

There are several limitations: orphan pages (Fig. 1e) must be excluded; all internal content resources are excluded (i.e. mp3, avi, flv, swf etc. files), non-significant nodes in folders (e.g. /css, /media, /js, /admin) are filtered; limited response time for server; threshold of usage for every node (i.e. less than 1.5%) are ignored.

The proposed solution combines two approaches found in the literature:

- The structure discovered on the basis of usage. Such approach includes such nodes and edges, which are present in navigation patterns.
- The website crawled with a robot.

The structure is fully discovered after the results of two approaches are merged. It will effect with the structure containing detail information of both the navigation and the usage of nodes and edges. The solution of this task consists of three subtasks:

- The scanning task of structure;
- The process of exclusion of orphan nodes;
- The task of sampling and applying data obtained from statistics module.

An algorithm for website discovery is proposed (Fig.3). The scanning task is solved by using the method of multithreaded searching by a limited number of crawlers. It is assumed that the tested website has a system for monitoring the usage data. Such an external system through a connector does exchange data with the website.

Orphan nodes in the structure of the website are frequently the result of a human error. These are not nodes intently hidden from the user in the navigation, because such a node is discovered during obtaining the usage data.

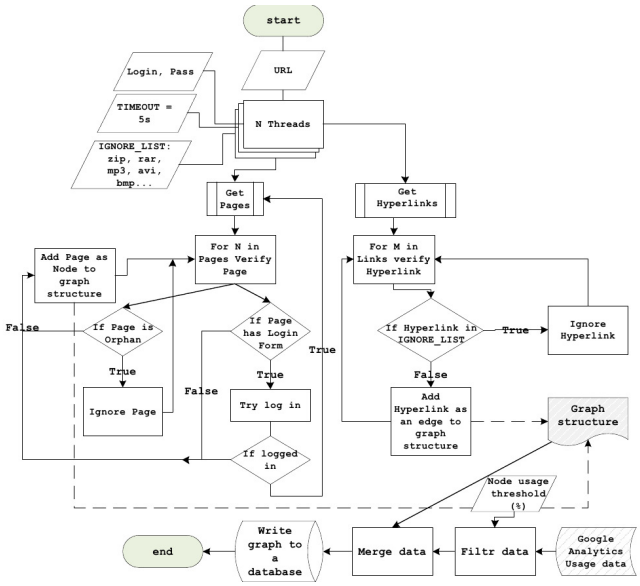

Fig. 3. An algorithm for discovering the website structure.

The node may contain links to external websites, but there is no reference to it from the navigation. Orphan exclusion is done in order to reduce the number of nodes that adversely affect the value of the quality estimator of the structure. Finding orphan nodes may help to understand which pages have not been connected with the others. This task is solved by using a recursive analysis of all the pages that contain at least one working internal link in order to detect all isolated nodes.

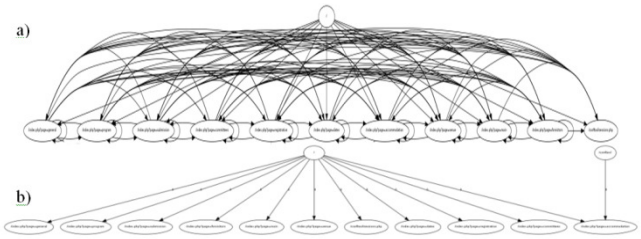

Fig. 4. A conference website icss.pwr.wroc.pl, 13 nodes and 124 edges. a) Graph structure before MST processing; b) after processing with MST algorithm.

E. Examples of discovered website structures

Using the proposed method of website representation it is possible to map simple structure with no connections between nodes. Such structure is constructed as a tree (Fig.4b, Fig.6). Because most of websites are graphs, in order to simplify the structure for analysis in terms of distribution of information in the nodes of the graph, a Minimum Spanning Tree (MST) algorithm is applied.

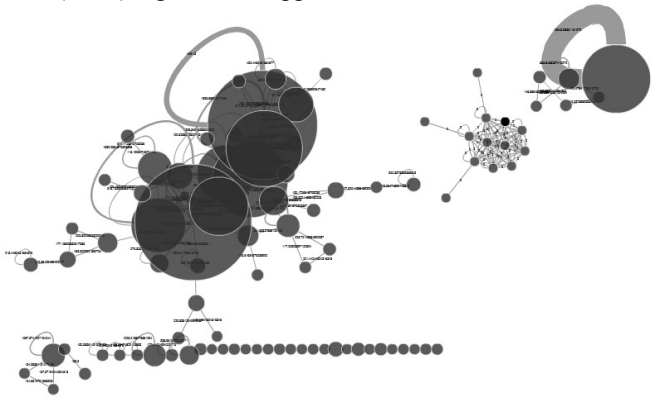

Fig. 5. Student's portal edukacja.pwr.wroc.pl, 41 nodes, 142 edges with applied usage data discovered between 2010-08-01 and 2011-05-05.

Structures of e-commerce websites are very complex, manual modification of the structure is done by an information architect and it takes long time to process it manually. Such a structure can be optimized using only the recommendations or intuitive knowledge. Application of MST is used to solve this problem. Developed method of structure discovery is regulated with a filtering parameter which allows filtering nodes to address name or percent of usage. It is reasonable to set usage data parameter to a high value (>2%) for complex structures.

The spectrum of possible application of described method is shown on figures (Fig.4ab, Fig.5, Fig.6, Fig.7).

III. Problem of Quality Valuation of Website Structure

The problem is: for given graph GI_l and history usage data in the interval of time $[t, t+_{\tau}k]$ determine the quality of the website structure. Note that for various periods of time τ the value of the quality estimator will be different. Assuming that the location of each node and the way of connections between them designates how users operate the website, the quality estimator should be dependent on the characteristics of the nodes and edges. Therefore there are defined quality criteria: number of node usage; number of edge usage; node position in the structure; the importance of a node; the importance of an edge.

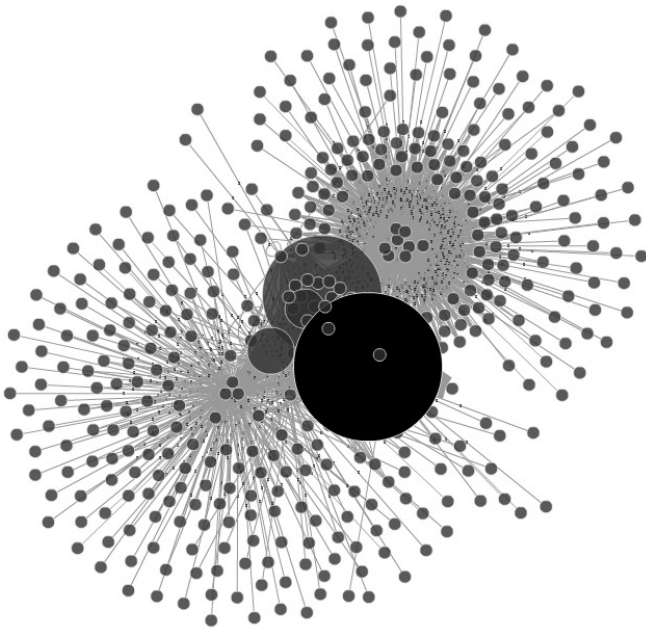

Fig. 6. E-commerce website structure with 638 nodes with usage data discovered between 2009-05-05 and 2011-05-05.

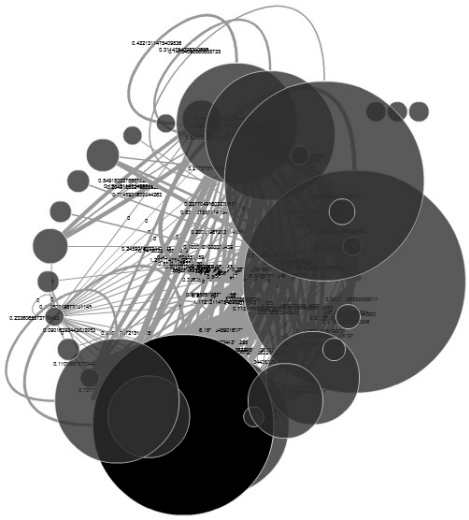

Fig. 7. Hair beauty website structure with 30 nodes and 177 edges and usage data discovered between 2010-10-01 and 2011-06-01.

Usage data results with information whether the originally designed structure is suitable for users or not. The quality measure should include popularity of nodes and edges. Therefore indicators of nodes and edges as defined in Section 2.B and 2.C are applied to develop the components of an estimator. Therefore, the node influence on the structure is defined as (13) and an indicator of the connectivity degree of an edge is defined as (14).

$$ImpU_k(url_{li}) = PopU_k(url_{li}) \cdot Acc(url_{li}) \quad (13)$$

$$CD_k(e_{li}) = PopE_k(e_{li}) \cdot \frac{reach(e_{li})/(N-1)}{out(e_{li})} \quad (14)$$

The quality estimator of the whole structure will be the sum of all values *ImpU* of the nodes proportional to the sum of edges values of *CD* and will be called Graph Energy and defined as (15).

$$En_k(GI_l) = \sum_{n=1}^{N} ImpU_k(url_{ln}) \cdot \sum_{m=1}^{M} CD_k(e_{lm}) \quad (15)$$

For the bare structure of the website it will be equal to 0 until the usage data will be applied. As the number of users during observed interval of time τ_k may vary and thus generate different traffic within nodes and edges, characteristics of the graph in time τ_{k+1} should be normalized proportionally to the value in τ_k period.

A. Complexity of a website structure discovery algorithm

The initial method in order to discover all connections within graph has a complexity of O*(N!)* and is a NP-hard, therefore the proposed method of website discovery explores only the paths which are actually used by users in selected interval of time. Computational complexity of method consists of the execution of subtasks:

- Scanning task (Fig. 3) is of O(*k(N+M)*), where *k* is the maximum depth of the graph.
- Usage data extraction is of O(*N+M*) complexity, because of need to obtain data of every node and every edge.
- Calculation of the graph characteristics: *io* has O(*N*), *out* is of O(*M*) complexity, *reach* has O(*M*), *d* is of O(*M+N*logN)*.

The total complexity of the quality estimation after simplification is *O(NlogN +M)*. The crucial for this method is number of nodes in the structure.

B. Experiments

To test proposed methods different websites were analyzed and monitored. In Fig.8 there are two of them – the conference website (*www.icss.pwr.wroc.pl*) and the website of hair beauty salon (*www.saloncesare.pl*). For both, the structure was scanned (Fig. 4a, Fig.7), then usage data was recorded and merged with the graph discovered by the crawler. The sampling of the Graph Energy was done daily (Fig.8), so the tendency could be better observed. There are two lines on Fig.8 – blue stands for the equation (15), and orange is a modified equation (15) with changed operation between *ImpU* and *CD* to the sum. Orange is less resistant for small changes, and blue is appropriate for long-term observations.

For the conference website, there were periods, where energy of the graph was lower or higher than average, and they are considered as potential points of change in the website structure, especially concerning the schedule of conference. Intuitively it is understood that the conference's structure should change in different periods of time i.e. paper submission, accommodation page before the conference start. As the graph energy is higher, the structure of the website is utilized better by users and they may reach goals more efficiently. Observation of changes in the quality of the structure contributes to detection of the moments in which there are derogations from certain value $En(\tau_k)$. For the hair beauty website the lower line shows the $En(\tau_k)$ (Fig.8b) and is inside lower and higher limit all the time with only twice short alarms detected. For the high and low limit Shewhart control chart [25] was utilized and for observing tendencies and making the decision of structure change Nine Nelson Rules [26] were applied.

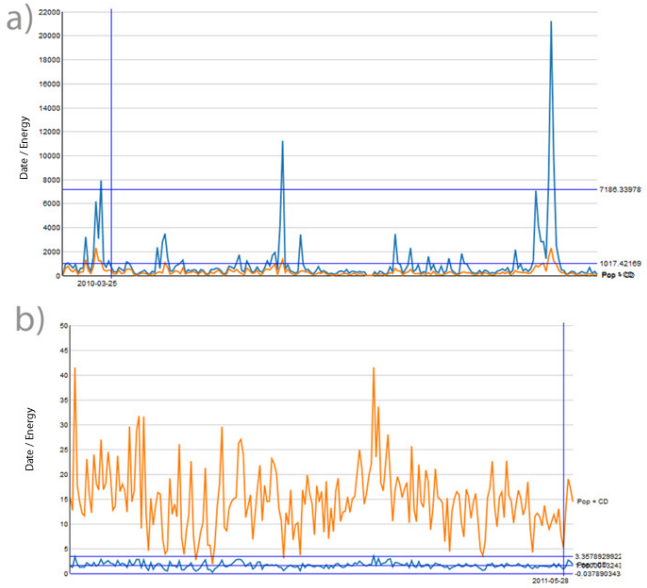

Fig. 8. Graph energy diagram of:
a) a conference website, b) a hair beauty salon website.

In order to compare results of two websites from the same categories (i.e. two beauty salons) the normalization of edge's and node's indicators should be done. More data on experiment and framework can be found in [27].

IV. SUMMARY

The main aim of this study was to state and propose a solution for structure discovering problem and quality valuation problem. A website structure quality estimator gives the ability to evaluate a website's navigation conformity to the way of how real users do use the website after its release to the general public. Besides consistency and repeatability, automation provides increased cost benefits to the developer. It improves the website redesign process and saves time.

In the future research there will be concerned following tasks. First, the problem of website structure optimization will be stated and the usage of quality metric based on Graph Energy will be proposed. Second, the problem of usage change detection will be stated and algorithms for it will be proposed. Third, the experimental framework will be upgraded with the new methods. Forth, the solution for discovering groups of users and recommending adapted structures to them individually will be proposed. The solution of all

discussed tasks will be an important step forward in studies on automated analysis, optimization problem and synthesis of website structures.

REFERENCES

[1] Douglas K. Van Duyne, James A. Landay, and Jason I. Hong, The Design od Sites. Upper Saddle River, NJ: Pearson Education, Inc., 2007.

[2] George M. Donahue, "Usability and the Bottom Line" IEEE Software 18 Issue 1, pp. 31-37, 2001.

[3] James Kalbach, Designing Web Navigation: Optimizing the User Experience, 1st ed.: O'Reilly Media, 2007.

[4] National Institute of Standards & Technology. (2002, May) WebSAT Evaluation Rules. [Online]. http://zing.ncsl.nist.gov/WebTools/WebSAT/websat_rules.html

[5] Witold Suryn, "Software Quality Engineering: The Leverage for Gaining Maturity" in Maturing Usability.: Springer-Verlag London Limited, 2008, vol. I, pp. 33-55.

[6] Sanjay J. Koyani, Robert W. Bailey, and Janice R. Nal, The Research-Based Web Design & Usability Guidelines: U.S. Department of Health and Human Services, 2006.

[7] Dave Gehrke and Efraim Turban, "Determinants of Successful Website Design: Relative Importance and Recommendations for Effectiveness" in Thirty-second Annual Hawaii International Conference on System Sciences-Volume 5, Maui, Hawaii, 1999, p. 5042.

[8] Jakob Nielsen, Designing Web usability: The practice of simplicity. Indianapolis: New Riders Publishing, 1999.

[9] Ping Zhang and Gisela M. von Dran, "User Expectations and Rankings of Quality Factors" International Journal of Electronic Commerce Vol.6 No.2, pp. 9-33, 2002.

[10] Raquel Benbunan-Fich, "Using protocol analysis to evaluate the usability of a commercial web site" Journal Information and Management Vol. 39 Issue 2, 2001.

[11] Janna B. Arney and Paul J. Lazarony, "An Inclusive Guide To Assessing Web Site Effectiveness" Journal of College Teaching & Learning Vol.2, Number 1, pp. 27-36, 2005.

[12] Jinwoo Kim, Jungwon Lee, Kwanghee Han, and Moonkyu Lee, "Businesses as Buildings: Metrics for the Architectural Quality of Internet Businesses" Information Systems Research Vol. 13 No.3, pp. 239-254, 2002.

[13] Johnathan W. Palmer, "Web site usability, design, and performance metrics" Information Systems Research Vol.13 No.2, pp. 151-168, June 2002.

[14] Layla Hasan and Emad Abuelrub, "Assessing the Quality of Web Sites" INFOCOMP Journal of Computer Science Vol.7 No.4, 2008.

[15] Horton S. Lynch P.J., Web style guide: basic design principles for creating Web sites. NJ: Yale University Press, 2009. [Online]. http://info.med.yale.edu/caim/manual

[16] G. Sreedhar, A. A. Chari, and V. V. Venkata Ramana, "Measuring Quality Of Web Site Navigation" Journal of Theoretical and Applied Information Technology, pp. 80-86, 2010.

[17] Wen-long Lin and Ye-zheng Liu, "A Novel Website Structure Optimization Model for More Effective Web" in Workshop on Knowledge Discovery and Data Mining, Adelaide, 2008, pp. 36-41.

[18] Mike Perkowitz and Oren Entzioni, "Adaptive Sites: Automatically Learning from User Access Patterns" Washington, Technical report UW-CSE-97-03-01 1997.

[19] John Garofalakis, Panagiotis Kappos, and Dimitris Mourloukos, "Web Site Optimization Using Page Popularity" Web Software, pp. 22-29, July-August 1999.

[20] Benjamin Yen, Paul Hu, and May Wang, "Toward an analytical approach for effective Web site design: A framework for modeling, evaluation and enhancement" Electronic Commerce Research and Applications 6, pp. 159-170, 2007.

[21] Sheng-Yuan Yang, "An ontological website models-supported search agent for web services" Expert Systems with Applications 35, pp. 2056–2073, 2008.

[22] Nicoleta David and Liviu Stelian Begu, "A Website Structure Optimization Model" in ACS'10 Proceedings of the 10th WSEAS international conference on Applied computer science, Iwate, 2010, pp. 426-429.

[23] Serge Brin and Larry Page, "The anatomy of a large-scale hypertextual Web search engine" in Proceedings of the VII International World Wide Web Conference, in: Computer Networks and ISDN Systems vol. 30, 1998, pp. 107-117.

[24] Jonathan Jeffrey, Peter Karski, Björn Lohrmann, Keivan Kianmehr, and Reda Alhajj, "Optimizing Web Structures Using Web Mining Techniques" in Intelligent Data Engineering and Automated Learning - IDEAL 2007, vol. 4881, Birmingham, 2007, pp. 653-662.

[25] Michele Basseville and Igor V. Nikiforov, Detection of abrupt changes: Theory and Application. Englewood Cliffs, N.J: Prentice-Hall, 1993.

[26] Lloyd S. Nelson, "Technical Aids," Journal of Quality Technology, vol. 16, no. 4, pp. 238-239, 1984.

[27] Dmitrij Żatuchin, “Webgraph - system do analizy i syntezy struktur serwisów www” Interfejs użytkownika - Kansei w praktyce, pp. 72-87, Warsaw, June 2011.