A COMPARATIVE APPROACH TO WEB EVALUATION AND WEBSITE EVALUATION METHODS

Dalal Ibrahem Zahran

Dept. of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
dzahran@kau.edu.sa

Hana Abdullah Al-Nuaim

Dept. of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia
hnuaim@kau.edu.sa

Malcolm John Rutter

School of Computing
Edinburgh Napier University
Scotland, UK
M.Rutter@napier.ac.uk

David Benyon

School of Computing
Edinburgh Napier University
Scotland, UK
D.Benyon@napier.ac.uk

Abstract

There is still a lack of an engineering approach for building Web systems, and the field of measuring the Web is not yet mature. In particular, there is an uncertainty in the selection of evaluation methods, and there are risks of standardizing inadequate evaluation practices. It is important to know whether we are evaluating the Web or specific website(s). We need a new categorization system, a different focus on evaluation methods, and an in-depth analysis that reveals the strengths and weaknesses of each method. As a contribution to the field of Web evaluation, this study proposes a novel approach to view and select evaluation methods based on the purpose and platforms of the evaluation. It has been shown that the choice of the appropriate evaluation method(s) depends greatly on the purpose of the evaluation.

Keywords: Web Evaluation Methods; Website Evaluation Methods; Web Engineering; Usability Evaluation Methods.

1. Introduction

Web development is a complex and challenging process that must deal with a large number of heterogeneous interacting components (Murugesan, 2008). Although the construction of Web applications has evolved some discipline, there is still a lack of an engineering approach for building Web systems, and the entire development process is still un-engineered (Ahmad et al., 2005).

An ad-hoc development approach to building complex Web systems quickly leads to poorly designed websites that may cause disasters to many organizations (Ahmad et al., 2005). Nielsen (2011) discovered that the same Web design mistakes occurred over and over again, leading him to publish a series of top-ten Web design mistakes based on testing widely used websites. Progressively, “Web Engineering” is emerging as a new discipline addressing the unique needs and challenges of Web systems and is officially defined as: "The application of systematic, disciplined and quantifiable approaches to development, operation, and maintenance of Web-based Information Systems" (Deshpande et al., 2002). The main topics of Web engineering include, but are not limited to, the following areas: Web development methodologies and models, Web system testing and validation, quality assessment, Web metrics and Web quality attributes disciplines, performance specification and evaluation, Web usability, and user-centric development (Kumar and Sangwan, 2011; Murugesan, 2008).

Unfortunately, evaluation of websites is too often neglected by many organizations, public or commercial, and many developers test systems only after they fail or after serious complications have occurred. Although testing a complex Web system is difficult and may be expensive, it shouldn't be delayed until the end of the development process or performed only after users report problems. The development of a Web system is not a one-off event; it’s rather a user-centered continuous process with an iterative life cycle of analysis, design, implementation, and testing (Murugesan, 2008). In this context, testing plays an important role in Web development, and therefore several methods have been proposed by scholars for evaluating websites. Yet, research that assesses evaluation methods has been in crisis for over a decade, with few publications and risks that inadequate evaluation practices are becoming standardized (Woolrych et al., 2011). In fact, the notion of website evaluation is often confused with Web evaluation in the literature. It is important to know the scope and purpose of evaluation: Are we evaluating the Web or specific website(s)? Also, is the goal to redesign the website, for example, or to obtain Web-ranking and traffic statistics? We need a different focus on evaluation methods and a new categorization system according to the purpose and platforms of evaluation.

Therefore, and to fill a gap in the literature of Web evaluation methods, the following are the objectives of this paper: (1) to distinguish between Web and website evaluation methods; (2) to identify the strengths and weaknesses of the respective approaches; and (3) to recommend the appropriate evaluation method(s) for assessing the Web/website based on the purpose of the evaluation.

2. Related Work

2.1. Web Metrics

Palmer (2002) focused on the need of metrics and emphasized that metrics help organizations generate more effective and successful websites. A survey by Hong (2007) on Korean organizations found that a key enabler of website success measurement is website metrics. These metrics play two important roles: They determine if a website performs to the expectations of the users and the business running the site, and they identify website design problems.

An earlier attempt to measure the Web was introduced in 1996 by Bray, who tried to answer questions such as the size of the Web, its connectivity, and the visibility of sites (Dhyani et al., 2002). Stolz et al. (2005) introduced a new metric assessing the success of information-driven websites that merged user behavior, site content, and structure while utilizing user feedback.

Calero et al. (2005) studied published Web metrics from 1992 to 2004. Using a three-dimensional Web quality model (WQM), they classified 385 Web metrics. The WQM defines a cube structure in which three aspects are considered when testing a website: Web features, life-cycle processes, and quality aspects. The results confirm that most metrics (48% of the metrics studied) are usability metrics, and 44% of them related to "presentation". In this respect, usability is a quality attribute that assesses how easy user interfaces are to use and also refers to methods for improving ease-of-use during the design process (Nielsen, 2012b). In the life cycle dimension, the majority of metrics are related to operation (43.2%) and maintenance processes (30%) (Figure 1). In addition, a large number of metrics are automated (67%).

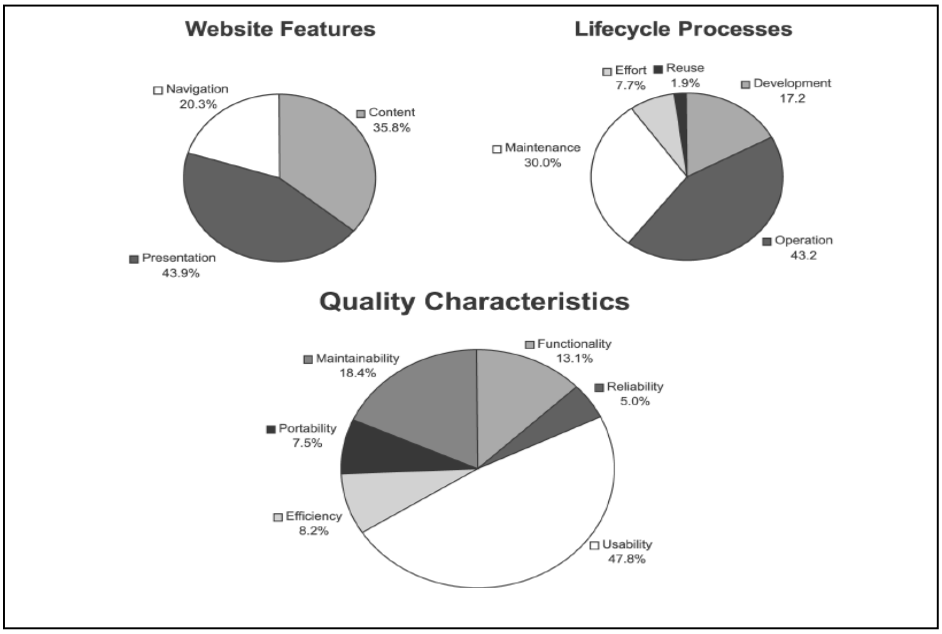

Figure 1. Metric Distribution across the Model Dimensions (Calero et al., 2005)

Dominic and Jati (2010) evaluated the quality of Malaysian University websites based on 11 quality criteria, such as load time, frequency of update, accessibility errors, and broken links, using the following Web diagnostic tools: Websiteoptimization (online performance and speed analyzer), Checklink validator, HTML validator, link popularity tool, and accessibility testing software. From the viewpoint of Treiblmaier and Pinterits (2010), there are two basic criteria for describing websites: "What is presented?" (Content) and "How is it presented?" (Design). The dimension "Ease of Use" contains navigation/organization and usability, the "Usefulness" dimension includes information or site content quality, while the third dimension is "Enjoyment" (Figure 2).

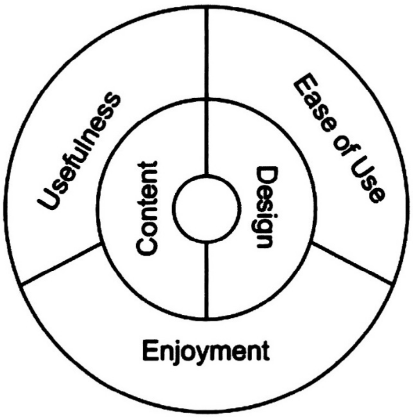

Figure 2. Framework for Web Metrics (Treiblmaier and Pinterits, 2010)

2.2. Trends and Existing Evaluation Approaches

Reviewing previous studies on existing evaluation methods reveals the following problems:

a) Researchers in the field use the terms "Web evaluation methods" (WEMs) and "website evaluation methods" (WSEMs) interchangeably. That is, they do not differentiate between diverse platforms of assessment methods; neither do they consider the purpose of the evaluation. For example, some studies evaluate the Web as a whole phenomenon for the purpose of site ranking or the connectivity and visibility of sites, such as Dhyani et al. (2002) and Stolz et al. (2005). Others assess specific websites against certain attributes aiming to discover the usability problems of the site, such as the studies of Calero et al. (2005), Dominic and Jati (2010) and Treiblmaier and Pinterits (2010).

b) Researchers in the field seldom classify evaluation methods. Nielsen and Mack (1994) classified usability evaluation methods (UEMs) into four categories: automatic (software evaluation), empirical (user testing), formal (evaluation models), and informal (expert evaluation), and later Ivory and Hearst (2001) categorized them into five categories: testing, inspection, inquiry, analytical modeling, and simulation. Recent attempts by Fernandez et al. (2011) adopted the same taxonomy as Ivory and Hearst. Unfortunately, those classifications of evaluation methods are few, old, and missing newer approaches, as neither of these taxonomies reflects, for example, Web analytics or link analysis aspects of UEMs.

c) Researchers in the field often applied the method(s) on different websites but seldom analyzed them or identified their strengths and weaknesses. For instance, link analysis methods have been used widely, but very few authors, such as Jalal et al. (2010), Noruzi (2006), and Shekofteh et al. (2010), evaluate them. Also, Fernandez et al. (2011) and Hasan (2009) indicated that there is little detail about the benefits and drawbacks of each method. Woolrych et al. (2011) warned that research that assesses UEMs has been in crisis for over a decade because of fewer publications. There are also risks that inadequate evaluation practices are becoming prematurely standardized.

d) Few compare evaluation methods or look at a combination of them. Summarizing the knowledge on UEMs over the last 14 years (1996 till 2009), Fernandez, et al. (2011) confirmed that studies often compare a limited number of evaluation methods. Also, Woolrych et al. (2011) argue that very few comparative studies investigate evaluation methods. Reviewing studies from 1995 till 2006, Chiou et al. (2010) stated that there was very limited research exploring the strategies of website evaluation.

A sample of studies using or comparing evaluation methods (explained in the next section) is presented in Table 1. Most of the research uses one or a few techniques only, and the literature is lacking the identification and classification of WEMs. It is worth noting that user testing and heuristics evaluation are traditional methods defined earlier by Nielsen (1993), whereas webometrics is a relatively new and evolving approach.

Table 1. Web Evaluation Methods

Authors	User Testing	Heuristics Evaluation	Automatic Tools	Analytics Tools	Google Analytics	Alexa	PageRank	Webometrics
Brajnik(2004a; 2004b; 2008); Ivory & Chevalier (2002); Dingli & Mifsud (2011); Dominic et al. (2010); Berntzen & Olsen (2009); Olsen et al. (2009); Ataloglou & Economides (2009)			√					
Palmer (2002)				√				
Hasan et al. (2009)					√			
Cho & Adams (2005)							√	
Noruzi (2005; 2006); Björneborn (2004); Jeyshankar & Babu (2009); Holmberg & Thelwall (2009); Li (2003); Thelwall & Zuccala (2008); Boell et al. (2008); Petricek et al. (2006); Shekofteh et al. (2010); Aminpour et al. (2009)								√
Nielsen (1993); Stone et al. (2005); Folmer & Bosch (2004); Lárusdóttir (2009)	√	√						
Prom (2007)				√	√			
Fang (2007)				√	√		√	
Scowen (2007)			√			√	√	
Matera et al. (2006)	√	√	√	√				
Hasan (2009)	√	√	√	√	√			

3. Classification of Evaluation Methods

The development of a Web system is a continuous process with an iterative life cycle of analysis, design, implementation, and testing (Murugesan, 2008). In the process of analyzing websites, Stolz et al. (2005) distinguished between three basic measurements: Web structure measurement (organization and navigability/links), Web content measurement, and Web usage measurement (as page view, sessions, frequency, unique users, and duration). Another view by Hasan (2009) categorized the assessment pattern into user, evaluator, and tool-based UEMs. But what we need really is a different focus on evaluation methods and a new categorization system according to the purpose and platforms of evaluation. Therefore, we propose a distinction between Web and website evaluation methods. We also stress the need for a more systematic identification of those methods.

Based on the previous discussion of classifying the assessment approaches to Web or website evaluation methods and extending Stolz et al. and Hasan's work, the following taxonomy of evaluation method is proposed:

1. Website evaluation methods (WSEMs):
    A. User-based usability evaluation methods
    B. Evaluator-based usability evaluation methods
    C. Automatic website evaluation tools (Bobby, LIFT, etc.)
2. Web evaluation methods (WEMs):
    A. Web analytics tools: (Google analytics, Alexa)
    B. Link analysis methods:
        i. PageRank
        ii. Webometrics methods.

3.1. Website Evaluation Methods (WSEMs)

The WSEMs measure a limited number of websites, manually or automatically, based on assigned criteria to achieve a high-quality website. Manual evaluation includes experts or real user testing, while automatic assessments employ different software-testing tools. The output of such an evaluation is a list of usability problems and recommendations to improve the tested website.

3.1.1. User-based Usability Evaluation Methods

The whole process of design for usability, user testing, and redesign is called User-centered Design (Folmer and Bosch, 2004; Nielsen, 1993). The term "usability evaluation" is used to describe the entire test, including planning and conducting the evaluation and presenting the results. The goal of a usability evaluation is to measure the usability of the system and identify usability problems that can lead to user confusion, errors, or dissatisfaction (Lárusdóttir, 2009). The user evaluation approach includes a set of methods that employs representative users to execute some tasks on a selected system. The users' performance and satisfaction with the interface are then recorded. The most common, valuable, and useful method in this category is user testing. Suggested techniques during a user-testing session include the think-aloud method, field observation, questionnaires, and interviews (Hasan, 2009):

User Testing

According to Stone et al. (2005), when users use a system, they work towards accomplishing specific goals in their minds. A goal is an abstract end result indicating what is to be achieved, and it can be attained in numerous ways. Consequently, each goal breaks down into tasks specifying what a person has to do, and then each task decomposes into an individual step that needs to be undertaken. In fact, user testing must be a sampling process, and users should be able to do basic tasks correctly and quickly. To select tested tasks, the examiner begins by exploring all the tasks within the website then narrowing them down to those that are the most important to users. A good task is one that discovers a usability problem or one that reveals an error that is difficult to recover from. The next step is how to present selected tasks to the participants, and one way to do this is to use a “scenario” in which the task is embedded in a realistic story. A good scenario is short, in the users' words, and directly linked to the user's everyday tasks and concerns. It does not give the steps for doing the task, since the point of the test is to see if a user can figure out the required steps alone.

It is important to test users individually and let them solve problems on their own. Actually, the purpose of a usability study is to test the system and not the users, and this aspect must be explicitly explained to tested users (Nielsen, 1993; Stone et al., 2005). The following metrics can be collected from user testing: time for users to learn a specific function, speed of task performance, type and rate of users' errors, user retention of commands over time, and user satisfaction (Abras et al., 2004). Moreover, how many participants to include in a user testing is a major issue in the usability field. Usually, three to five participants are needed to see all the potential usability problems (Nielsen, 1993; Stone et al., 2005). Nielsen confirmed that the best results come from the first five users and that roughly 85% of the usability problems in a product are detected with five participants.

The Think-aloud Method

Lárusdóttir (2009) and Nielsen (1993) regard thinking aloud as the single most valuable usability evaluation method, and Nielsen (2012a) still holds the same opinion, as he titled his article, "Thinking Aloud: The #1 Usability Tool." Basically, this method involves an end user using the system while thinking out loud. By verbalizing their thoughts, the test users enable us to understand how they view or interpret the system and what parts of the dialogue cause problems. Its strength lies in the wealth of collected qualitative data that can be obtained from a small number of users. The users' comments can be included in the test report to make it more informative. However, to some extent, thinking aloud seems an unnatural setting for users, and sometimes it may give a false impression of the actual cause of usability problems if too much weight is given to the users' justifications (Nielsen, 1993).

3.1.2. Evaluator-based Usability Evaluation Methods

Evaluators or experts inspect the interface and assess system usability using interface guidelines, design standards, users' tasks, or their own knowledge, depending on the method, to find possible user problems (Lárusdóttir, 2009). The inspectors can be usability specialists or designers and engineers with special expertise (Matera et al., 2006). In this category, there are many inspection methods, such as cognitive walkthrough, guideline reviews, standard inspection, and heuristic evaluation (Hasan, 2009).

Heuristic Evaluation

Heuristic evaluation is a very efficient usability engineering method, and it is especially valuable when time and resources are scarce. A number of evaluators assess the application and judge whether it conforms to a list of usability principles, namely "heuristics" (Hasan, 2009). There are two sets of guidelines that are widely used in heuristic evaluation, Nielsen's (1993) heuristics being the most common, followed by Gerhardt-Powals' (1996) (Lárusdóttir, 2009). Nielsen's heuristics are part of the so-called "discount usability methods" which are easy, fast, and inexpensive. During the heuristic evaluation, each evaluator goes individually through the system interface at least twice, and the output of such evaluation is a list of usability problems with reference to the violated heuristics (Matera et al., 2006). In principle, heuristic evaluation can be conducted by only one evaluator, who can find 35% of total usability problems (Nielsen, 1993), but another view by Matera et al. (2006) believes that better results are obtained by having five evaluators and certainly not fewer than three for reasonable results.

3.1.3. Automatic Website Evaluation Tools

Automatic evaluation tools are software that automates the collection of interface usage data and identify potential Web problems. The first study of automatic tools was conducted by Ivory and Chevalier (2002), who concluded that more research was needed to validate the embedded guidelines and to make the tools usable. Thus Web professionals cannot rely on them alone to improve websites. Brajnik (2004b) mentioned several kinds of Web-testing tools: accessibility tools such as Bobby, usability tools such as LIFT, performance tools such as TOPAZ, security tools such as WebCPO, and classifying website tools such as WebTango. He stated that the adoption of tools is still limited due to the absence of established methods for comparing them and also suggested that the effectiveness of automatic tools has to be itself evaluated (2004a). In fact there are many automated tools available as either Web-based services or desktop applications. A recent popular free Web-based accessibility tool is Cynthia Says (http://www.cynthiasays.com/) which is a product from HiSoftware that allows you to enter the URL to be analyzed in to the sight and get a report on how it complies with Section 508 standards and/or the Web Content Accessibility Guidelines (WCAG). Table 2 shows some studies that use different kinds of automatic website evaluation tools.

Table 2. Examples of Automated Web Site Evaluation Studies

Name of the Study	Author / Year	Automatic tools
Assessing e-governance Maturity through Municipal Websites: Measurement Framework and Survey	(Rodríguez et al., 2009)	1-W3C validators 2-Xenu s\w (broken links) 3-Weight & image resolution 4-Source code analyzer
Quantitative Assessment of European Municipal Web Sites Development and Use of an Evaluation Toll	(Miranda, Sanguino, & Banegil 2009)	1-Google search engine 2-Link popularity check 3-Chronmeter (access speed)
Local E-government: Reconstructing Limassol's Municipality (Cyprus) Web Site to Provide Functional and Effective E-services	(Zevedeos, 2006)	1-WebXact (Bobby) 2-Lynx (accessibility) 3-Vischeck (color) 3-W3C Markup Validator 4-W3C CSS validator 5-W3C Link Checker
Performance Evaluation on Quality of Asian E-government Websites – an AHP Approach	(Dominic et al., 2010)	1-Website optimization (website performance and speed analyzer) 2-W3C checklink 3-Link popularity 4-Accessibility s\w Tawdis tester 5-Color-blind webpage filter
Evaluating Global E-government Sites: A View Using Web Diagnostic Tools	(Choudrie, Ghinea, & Weerakkody, 2004)	1-WebXact (accessibility, quality & privacy) 2-Netmechanic (Links) 3-W3C HTML validator 4-Vizcheck (color)

3.2. Web Evaluation Methods (WEMs)

The WEMs study the Web as a whole by calculating statistics about the detailed use of a site and providing Web-traffic data, visibility, connectivity, ranking, and the overall impact of a site on the Web.

3.2.1. Web Analytics Tools

Web analytics have been defined by the Web Analytics Association as "the measurement, collection, analysis and reporting of Internet data for the purpose of understanding and optimizing Web usage" (Fang, 2007). These tools automatically calculate statistics about the detailed use of a site helping, for example, in discovering navigation patterns corresponding to high Web usage or to the early leaving of a website (Matera et al., 2006). Originally, Web analytics is a business tool that started with some webmasters inserting counters on their home pages to monitor Web traffic. While most Web analytics studies target e-commerce, the method can be applied to any website (Prom, 2007). The two data collection methods for Web analytics are server-based log files (traffic data is collected in log files by Web servers) and client-based page-tagging (requiring the addition of JavaScript codes to webpages to capture information about visitors' sessions) (Hasan, 2009). The two well-known Web analytics tools are Google Analytics and Alexa.

Google Analytics

Google purchased a Web analytics company called Urchin software in 2005 and subsequently released Google Analytics to the public in 2006 (Fang, 2007; Hasan et al., 2009). The service is free for up to five million page views per month per account. Once signed up for Google Analytics, Google offers users code that must be inserted into each Web page to be tracked. Visual data results are displayed with a wealth of information as to where visitors came from, what pages they visited, how long they stayed on each page, how deep into the site they navigated, etc. (Fang, 2007).

Alexa

Alexa is a website metrics system owned by the Amazon Company that provides a downloadable toolbar for Internet Explorer users. It calculates traffic rank by analyzing the Web usage of Alexa toolbar users for three months or more as a combined measure of page views and reach (the number of visitors to the site). Although this information is useful, Alexa ranking is biased towards MS Windows and Internet Explorer users (Scowen, 2007).

3.2.2. Link Analysis Methods

Link analysis studies websites' topology, assuming that the quality of a Web page is dependent on its links. There are two important methods that use link analysis: PageRank and webometrics.

PageRank

A number of researchers investigated the Web link structure to improve search results and proposed ranking metrics. When Page and Brin designed the Google search engine, they considered links as positive referrals and created a system called PageRank. Google PageRank is a link analysis algorithm named after Larry Page that assigns a numerical weight to each hyperlink, and each page has a calculated PageRank based on the number and quality of links pointing to it (Scowen, 2007). Google takes 100 factors into consideration when determining the ranking of a page, but PageRank is the main factor in search-result ordering. The PageRank metric PR(p) defines the importance of page p to be the sum of the importance of the pages that point to p, and the PR(p) is high if many important pages point to p. The effectiveness of Google's search results and the adoption of PageRank by other search engines strongly indicate that it is an effective ranking metric for Web searches, but unfortunately it is heavily negatively biased against unpopular pages, especially those created recently (Cho and Adams, 2005).

Scowen (2007) tested e-learning websites against checklist guidelines then against five ranking systems: Google links search, Yahoo links, Delicious links, Google PageRank, and Alexa. The Google PageRank and Alexa were used to know their correlations with the usability of the website, although neither can be relied upon as a main indicator of popularity. He found that increased compliance with usability guidelines has a strong correlation with increased popularity of a website. Although Alexa is not a reliable indicator, it is at least consistent with other rankings. Thus, more usable websites achieve a higher PageRank and are also more popular in Alexa. Overall, the five ranking systems showed positive correlations to each other and to the usability of the sites.

Webometrics and the WIF Method

Björneborn (2004) has proposed webometrics as "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and infometric approaches." This means evaluation of websites can be conducted "webometrically" with the goal to validate links and furnish its acceptance as a useful metric to measure the Web. Webometrics assess the international visibility and impact of an institution or a country on the Web (Jeyshankar and Babu, 2009), but it is still a nascent field of research (Björneborn, 2004; Holmberg and Thelwall, 2009).

The Web Impact Factor (WIF) is the most important method in webometrics. In 1998, Peter Ingwersen proposed WIF through an analogy with the Journal Impact Factor (JIF) (Noruzi, 2005; Li, 2003) that represents the ratio of all citations to a journal to the total references published over a period of time (Dhyani et al., 2002). Since it is a snapshot of the Web and lacks peer review and quality control, the WIF is not exactly the equivalent of the JIF, but it was inspired by it (Thelwall and Zuccala, 2008). In this method, external inlinks are of more value and importance (Aminpour et al., 2009); the more people link to a website, the more WIF the site is getting and, in turn, the higher the impact factor, the higher the reputation and influence of a site (Jeyshankar and Babu, 2009; Shekofteh et al., 2010). Sometimes the WIF is wrongly compared to PageRank method. PageRank does not afford equal weight to links, and weightings vary depending on from where a link is coming (Boell et al., 2008).

Most of webometrics studies were performed on university sites such as the Cybermetrics Lab (2010), which has issued the "Ranking Web of World Universities" since 2004. A study by Thelwall and Zuccala (2008) measured the international interlinking to and from different European universities. Figure 3 shows European links from university networks with the width of arrows proportional to the number of pages between universities. Results show the dominance of the large, richer western European nations, especially the UK and Germany (de) as central actors on the Web and also strongly connected with each other. The importance of Switzerland (ch) is apparent, since it is connected strongly to the UK and Germany, weakly to seven countries, and medium to one country, France (fr). In turn, France is connected strongly to Germany, weakly to nine countries, and medium to four countries: Italy (it), Belgium (be), Switzerland (ch), and the Netherlands (nl). Poland (pl) is also well-connected and has a significant presence as a newcomer.

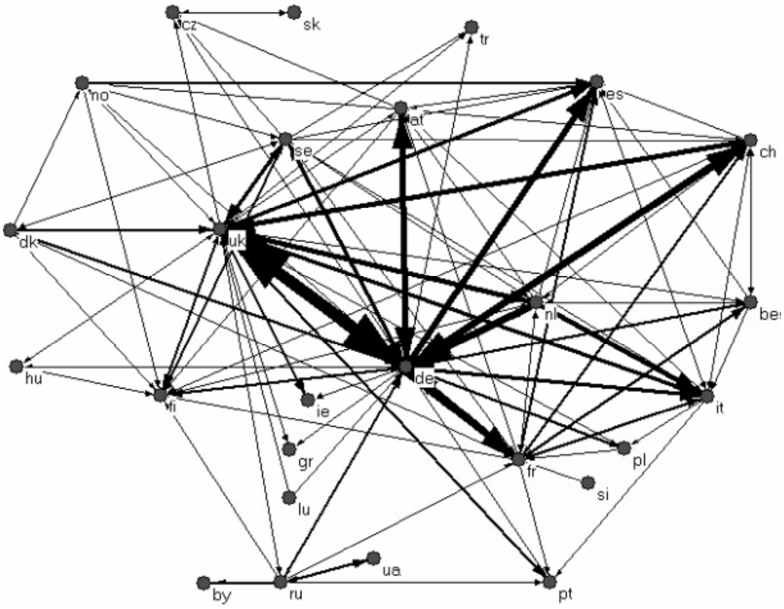

Figure 3. European Link Network. (Thelwall and Zuccala, 2008)

A few webometrics studies have been conducted on e-government, representing a new application of the WIF method. The first attempt to measure e-government webometrically was the study by Petricek et al. (2006), which compared the audit office sites in five countries and showed that the US and Canada emerge as the most connected sites, more than the UK, New Zealand, and Czech Republic.

4. Analysis of Evaluation Methods

This section examines existing evaluation methods individually, regardless of any proposed categorization in order to identify the strengths and weaknesses of each method.

Automatic website evaluation tools attract attention because they are fast, consistent, produce unbiased results, and obviate the shortage of experts and inconsistent results between them (Ataloglou and Economides, 2009; Dingli and Mifsud, 2011; Dominic et al., 2010). Also, these tools can offer an initial overview of the status of a website (Olsen et al., 2009). However, automation of website testing is an evolving method that cannot be considered efficient (Al-Juboori et al., 2011). Berntzen and Olsen (2009), Brajnik (2008), and Dingli and Mifsud (2011) concluded that automatic tools cannot replace human evaluators but should assist them. Ivory and Chevalier (2002) predicted that automation is a useful complement to standard evaluation techniques. Manual evaluations provide more details than automatic tests, which cannot capture the whole picture. Anything requiring assessment is likely to be poorly machine testable (Brajnik, 2004b).

Another concern is that the market forces can cause changes that threaten automatic tools' stability. For example, Bobby, an accessibility testing tool, was sold in 2004 to Watchfire, which provided the same free service in the WebXACT tool, but Watchfire was acquired by IBM in 2007. Bobby was then discontinued as a free tool, and currently it is included within the IBM Rational Policy Tester Accessibility Edition (Hasan, 2009). In fact, automatic tools are seldom used alone in website evaluation; also, very few studies compare the tools and validate their effectiveness (Al-Juboori et al., 2011; Brajnik, 2004a, 2004b). The most-used tools are Bobby, LIFT, W3C validators, and link-checker software. Most automatic tools focus on site accessibility rather than usability, and they are not considered efficient (Hasan, 2009; Scowen, 2007). Even the very few tools for usability often neglect structural and navigational problems (Matera et al., 2006). Further, information about LIFT is contradictory; some conceive LIFT as a test for accessibility and some as a usability tool. Also, features measured by LIFT are inconsistent with the USA Research Web Design and Usability Guidelines (Scowen, 2007).

On the other hand, Web analytics tools solve some problems in Web evaluation, since they might reduce the need for user testing, and often the data is collected automatically with high accuracy. They offer the possibility of analyzing a high number of visitors, thus increasing the reliability of the discovered errors; however, the inaccuracy of log files as a data source is acknowledged (Hasan, 2009). Another serious problem is the meaning of the collected information and how much it describes users' behavior (Matera et al., 2006). Palmer (2002) believes website traffic measures are used because they are easy to capture but are very often deemed to be inadequate and sometimes may generate conflicting results.

A Web analytics tool such as Alexa has some limitations; it is biased towards a sample of MS Windows and Internet Explorer users. The resulting statistics are unreliable since users of other operating systems or browsers are not recorded, and traffic from other Internet users is not counted (Scowen, 2007). Unfortunately, there are only a few studies that show the value of Google Analytics in assessing websites; Hasan (2009) developed a framework for evaluating three e-commerce sites in the kingdom of Jordan using heuristic evaluation, user testing, and Google Analytics. Jordanian companies took a long time to agree to participate in the research due to trust and security issues, since they were asked to add script code to their servers.

Noruzi (2006) considers the webometric method as an imperfect tool to measure the quality of websites. Questions are raised over the entire quantitative nature of the webometrics rankings (Björneborn, 2004). The tool used in the WIF analysis is not meant for the task, and search engines are designed for content retrieval, not link analysis; plus, they may create problems in drawing conclusions for the WIF since their coverage of the Web is incomplete. The lack of knowing why Web links are created is a major obstacle in the webometrics method; thus the motivations behind creating links raise questions of uncertainty (Noruzi, 2006). Also, some webometrics' studies found unexpected results and attributed them to the limitations of the WIF method. For example, a university with 993 links and 99 Web pages, by division, gets an impact factor of 10, whereas another one with 12,700 links and 87,700 Web pages obtains an impact factor below zero (Shekofteh et al., 2010).

Based on webometrics evaluation, university rankings have raised a large dispute, and several studies criticize them as merely a list of criteria that mirrors the superficial characteristics of universities. Noruzi (2006) argued that world university website ranking is dangerous and not meaningful because a high link rate may not always be associated with high quality. It is vulnerable to manipulation, since the WIF can be influenced by institutions that know how this method works. Shekofteh et al. (2010) concluded that the WIF alone is not a good measure for ranking universities, and Noruzi (2006) stated that with about 10 years of criticism, it seems that there is no obvious alternative yet. Webometrics is relatively a young field of research that needs different theories to be built, methods to be developed, and problems to be solved (Björneborn, 2004; Holmberg and Thelwall, 2009). Calculating the WIF for a website is easy, but what the figures mean is arguable. Thus, the researches on webometrics are in the process of developing and validating its methodologies.

Matera et al. (2006) supported Nielsen (1993) in considering heuristic evaluation as a very efficient method when time and resources are scarce because experts can produce high-quality results in a limited time. But a negative aspect is its high dependence on skills and the experiences of the evaluators. They concluded that novice evaluators with no usability expertise are poor evaluators, usability experts are 1.8 times as good, while application domain and usability experts (double experts) are 2.7 as good. Another weakness of this method is the great subjectivity of the evaluation; there is a risk that the experts mistakenly consider some issues as problems but actually real users do not have trouble with them; this is often referred to as "false problems" (Lárusdóttir, 2009).

According to Nielsen (1993), user testing with the think-aloud technique finds more major Web problems than other evaluation methods but is poor in uncovering minor ones, and the situation is the opposite for the heuristic evaluation. Since they complement each other, he recommends first conducting a heuristic evaluation to find as many "obvious" usability problems then performing user testing to find the remaining problems. Likewise, Hasan (2009) reached the same conclusion of Nielsen and added that Google Analytics is a useful quick preliminary step to discover general usability problems. She found that user testing is good for identifying major usability problems in four areas: navigation, design, the purchasing process, and accessibility and customer service, while the heuristic evaluation identifies minor usability problems in eight areas: navigation, internal search, the site architecture, the content, the design, accessibility and customer service, inconsistency and missing capabilities, plus addressing security and privacy issues. Other Web experts recommend using several different evaluation techniques, since each one alone is not free of shortcomings (Ivory and Chevalier, 2002).

The overall recommendation by many researchers is to conduct heuristic evaluation and user testing to find most usability problems. Other evaluation methods are just useful complements offering the possibility of analyzing a high number of users as an initial preview of a website. Consequently, evaluations by experts or users are the mainstream approach, and probably the future trend is a mixture of automatic and manual website evaluations.

5. Selection of Appropriate Evaluation Method(s)

Kaur and Dani (2013) evaluated the state of navigability of Indian banking websites and found that Alexa and Google PageRank do not have significant correlations with navigability metrics, indicating that popularity and importance are not good indicators of website navigability; therefore, the traffic data and the back-links of the websites are not meaningful measures of site navigation assessment. Cho and Adams (2005) added that PageRank is not a metric of page quality. Further, Hong (2007) stated that most organizations use Web metrics to determine site traffic or popular content but seldom used them to improve navigation. Jalal et al. (2010) and Noruzi (2006) concluded that the webometric method is an imperfect tool to measure the quality of websites and that it reflects unreliable results in most cases.

The findings of these five studies support the argument that WEMs, such as the Web analytics tools and the link analysis methods, do not discover navigation problems accurately nor do they measure website quality. Further, it seems that WEMs are complementary approaches since they do not definitely discover usability problems of a site, rather they indicate their probability.

On the other hand, even though usability testing demonstrates how real users interact with a website and the exact problems they face, it cannot measure the success of a site or describe the interactions of large numbers of users with it (Hasan, 2009). This highlights the weakness that WSEMs, such as user, evaluator, or automatic evaluation methods, cannot provide traffic data, Web ranking of a site, or its online visibility among others.

Therefore, the choice of the appropriate evaluation method depends greatly on the purpose of the evaluation. If it is intended to redesign the website and wanted to discover most of its potential usability problems, then the best evaluation methods are user testing and expert evaluation, while an automatic tool or Google analytics is a useful complement in this situation. If the goal of the evaluation is to redesign a website then WSEM is the best approach, while WEMs are not useful enough in this circumstance. Similarly, if the goal is to clarify the extent of online correlation with other institutions/countries or to know the ranking of a website and how much traffic it attracted, then the best way is to use WEMs, link analysis methods, and Web analytics tools, respectively. Figure 4 shows how the purpose of Web evaluation determines the type of method; the dotted arrow is toward a complementary method.

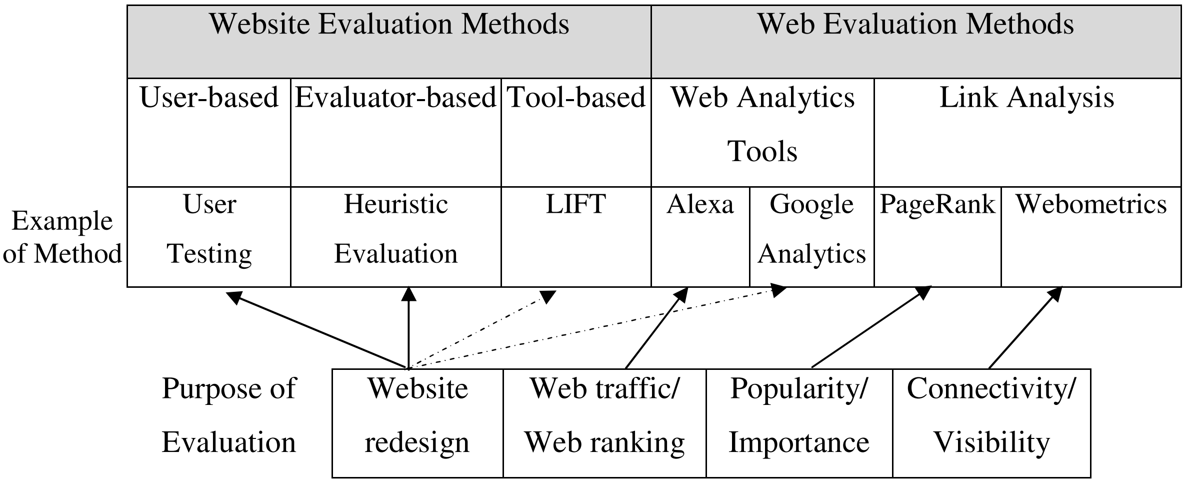

Figure 4. Purpose of Web Evaluation Determines the Appropriate Method Type

6. Conclusion

To address the challenge of developing complex Web systems, "Web Engineering" is an emerging discipline for the implementation of engineering principles to promote high quality websites that attract visitors. How to measure the Web has become a valuable area of ongoing research, but unfortunately the field is not yet mature; Web evaluation methods are scattered over the literature with a lack of studies that classify, compare, and determine the appropriate evaluation method(s).

Previous studies confused the term "Web evaluation methods" with "website evaluation methods," since they did not distinguish between diverse platforms of assessment methods and also did not address the purposes behind such evaluation. For example, some studies evaluated the Web in terms of ranking and connectivity of sites, while others assessed specific websites to discover their usability problems.

A novel approach to view evaluation methods is proposed, and a new categorization system has been suggested based on the purpose and platforms of evaluation. As a contribution to the field of Web evaluation, we have identified existing evaluation methods and accordingly classified them into two types: (1) website evaluation methods including user-based UEMs such as user testing and think aloud, evaluator-based UEMs such as heuristics evaluation, and automatic website evaluation tools and (2) Web evaluation methods including Web analytics tools (Google analytics, Alexa) and link analysis consisting of PageRank and webometrics methods.

Analyzing existing evaluation methods resulted in the following conclusions: First, standard evaluation techniques are user testing and heuristic evaluation. Second, tool-based evaluation methods offer a first insight into the status of a website. Automatic testing is a useful complementary tool but it is an evolving method with little evidence of its efficacy. Similarly, Web analytics tools provide some useful website traffic measures. However, the resulting statistics of Alexa, for example, are unreliable since it covers a limited number of Internet users. Also, Google Analytics is a quick preliminary step to discover usability problems, but its uses are limited due to trust and security issues. Third, link analysis methods try to validate links as a useful metric to measure the Web, but actually PageRank and webometrics methods can be regarded as indicators rather than definite conclusions on the visibility and impact of a website. For example, the WIF is partially successful; it does provide some useful information such as the relationship and type of communication between universities/countries and also how a website is isolated or connected with others online. On the other hand, the method is not appropriate for the ranking of websites since it is not a suitable tool for assessing a website's quality.

The purpose of Web evaluation determines the appropriate method(s) to be used. If the purpose is to redesign the website, then the scope of evaluation is WSEM, and therefore, as stated by the literature, the best evaluation methods are user testing and expert evaluation, while automatic and Web analytics tools (complementary) could provide a first insight into the status of the website. Similarly, if Web ranking and traffic statistics are of interest, then the scope of evaluation is WEMs; thus the best way is to use a Web analytics tool such as Alexa.

References

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-Centered Design. Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, *37*(4), 445-56.

Ahmad, R., Li, Z., & Azam, F. (2005). Web Engineering: A New Emerging Discipline. IEEE 2005 International Conference on Emerging Technologies. September 17-18, Islamabad.

Al-Juboori, A. F. M. A., Na, Y., & Ko, F. (2011). Web Site Evaluation: Trends and Existing Approaches. In Networked Computing (INC), 2011 The 7th International Conference, IEEE, 155-160.

Aminpour, F., Kabiri, P., Otroj, Z., & Keshtkar, A. (2009). Webometric Analysis of Iranian Universities of Medical Sciences. Scientometrics, *80* (1), 253–264.

Ataloglou, M., & Economides, A. (2009). Evaluating European Ministries' Websites. International Journal of Public Information Systems, *5*(3), 147–177.

Berntzen, L., & Olsen, M. (2009). Benchmarking E-government- A Comparative Review of Three International Benchmarking Studies. In Digital Society, 2009. ICDS'09. Third International Conference on IEEE, 77-82.

Björneborn, L. (2004). Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach. Royal School of Library and Information Science, Denmark. Retrieved from
http://vip.db.dk/lb/phd/phd-thesis.pdf

Boell, S., Wilson, C., & Cole, F. (2008). A Webometric Analysis of Australian Universities Using Staff and Size Dependent Web Impact Factors (WIF). University of New South Wales, School of Information Systems, Technology and Management (SISTM), Sydney, Australia

Brajnik, G. (2004a). Comparing Accessibility Evaluation Tools: A Method for Tool Effectiveness. Universal Access in the Information Society, *3*(3-4), Springer Verlag, 252-263

Brajnik, G. (2004b). Using Automatic Tools in Accessibility and Usability Assurance. 8th International ERCIM UI4ALL Workshop, June 2004, Vienna, Austria. Springer, Berlin, 219-234

Brajnik, G. (2008). Beyond Conformance: The Role of Accessibility Evaluation Methods. Keynote paper, 2nd International Workshop on Web Usability and Accessibility IWWUA08, September, 2008, Auckland, New Zealand. Retrieved from http://www.cast.org/learningtools/Bobby/index.html

Calero, C., Ruiz, J., & Piattini, M. (2005). Classifying Web Metrics Using the Web Quality Model. Online Information Review, Emerald Group, *29* (3), 227-248

Chiou, W. C., Lin, C. C., & Perng, C. (2010). A Strategic Framework for Website Evaluation based on a Review of the Literature from 1995–2006. Information & Management, 47(5), 282-290.

Cho, J., & Adams, R. (2005). Page Quality: In Search of an Unbiased Web Ranking. In Proceedings of the 2005 ACM SIGMOD conference, Baltimore, Maryland, 551-562

Choudrie, J., Ghinea, G., & Weerakkody, V. (2004). Evaluating Global e-Government Sites: A View using Web Diagnostic Tools. Brunel University, Uxbridge, UK. Electronic Journal of e-Government, 2 (2), 105-114

Cybermetrics Lab. (2010). Webometrics Ranking of World Universities. Retrieved from http://www.webometrics.info/

Deshpande, Y., Murugesan, S., Ginige A., Hansen, S., Schwbe, D., Gaedke, M., & White, B. (2002). Web Engineering. Journal of Web Engineering, Rinton Press, *1*(1), 003-017.

Dhyani, D., Ng, W., & Bhowmick, S. (2002). A Survey of Web Metrics. Nanyang Technological University. ACM Computing Surveys, *34*(4), 469–503.

Dingli, A., & Mifsud, J. (2011). USEFul: A Framework to Mainstream Web Site Usability through Automated Evaluation. International Journal of Human Computer Interaction (IJHCI), *2*(1), 10

Dominic, P., & Jati, H. (2010). Evaluation Method of Malaysian University Website: Quality Website Using Hybrid Method. In Information Technology (ITSim), 2010 International Symposium, IEEE, *1*, 1-6

Dominic, P., Jati, H., & Kannabiran, G. (2010). Performance Evaluation on Quality of Asian E-government Websites - an AHP Approach. International Journal of Business Information Systems, *6*(2), 219-239.

Fang, W. (2007). Using Google Analytics for Improving Library Website Content and Design: A Case Study. Library Philosophy and Practice 2007, 1-17

Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability Evaluation Methods for the Web: A Systematic Mapping Study. Information and Software Technology, *53*(8), 789-817.

Folmer, E., Bosch, J. (2004). Architecting for Usability: A Survey. Journal of Systems and Software, 2004, *70*(1), 61–78. Retrieved from http://dissertations.ub.rug.nl/FILES/faculties/science/2005/e.folmer/c2.pdf

Hasan, L. (2009). Usability Evaluation Framework for E-commerce Websites in Developing Countries. A Doctoral dissertation. Loughborough University

Hasan, L., Morris, A., & Probets, S. (2009). Using Google Analytics to Evaluate the Usability of E-commerce Sites. In Human Centered Design, Springer Berlin Heidelberg, 697–706

Holmberg, K., & Thelwall, M. (2009). Local Government Web Sites in Finland: A Geographic and Webometric Analysis. Scientometrics, *79*(1), 157–169.

Hong, I. (2007). A Survey of Web Site Success Metrics Used by Internet-dependent Organizations in Korea. Internet Research, Emerald Group, *17*(3), 272-290

Ivory, M. & Chevalier, A. (2002). A Study of Automated Web Site Evaluation Tools. University of Washington, Department of Computer Science2002

Ivory, M. Y., & Hearst, M. A. (2001). The State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys (CSUR), *33*(4), 470-516.

Jalal, S., Biswas, S. & Mukhopadhyay, P. (2010). Web Impact Factor and Link Analysis of Selected Indian Universities. Annals of Library and Information Studies, *57*, 109 – 121. Retrieved from http://eprints.rclis.org/16164/1/Annals-57-2.pdf

Jeyshankar, R., & Babu, R. (2009). Websites of Universities in Tamil Nadu: A Webometric Study. Annals of Library and Information Studies, *56*(2), 69-79. Retrieved from http://nopr.niscair.res.in/bitstream/123456789/5939/1/ALIS%2056(2)%2069-79.pdf

Kaur, A., & Dani, D. (2013). The Web Navigability Structure of E-Banking in India. International Journal of Information Technology and Computer Science *(IJITCS)*, 5(5), 29. Retrieved from http://www.mecs-press.org/ijitcs/ijitcs-v5-n5/IJITCS-V5-N5-4.pdf

Kumar, S. Sangwan, S. (2011). Adapting the Software Engineering Process to Web Engineering Process. International Journal of Computing and Business Research, *2*(1)

Lárusdóttir, M. (2009). Listen to Your Users: The Effect of Usability Evaluation on Software Development Practice. A Doctoral Dissertation, Department of Information Technology, UPPSALA University, Sweden

Li, X. (2003). A Review of the Development and Application of the Web Impact Factor. Online Information Review, Emerald Group, *27*(6), 407-417. Retrieved from http://www.emeraldinsight.com/researchregister

Matera, M., Rizzo, F., Carughi, G. (2006). Web Usability: Principles and Evaluation Methods. Department of Electronic Information, Milano, Italy

Miranda, F., Sanguino, R., & Banegil ,T. (2009). Quantitative Assessment of European Municipal Web Sites - Development and Use of an Evaluation Tool. Internet Research, 19(4), 425-441, Emerald Group Publishing Limited

Murugesan, S. (2008). Web Application Development: Challenges and the Role of Web Engineering. University of Western Sydney, Australia.

Nielsen, J. (1993). Usability Engineering. San Francisco, CA: Morgan Kaufmann

Nielsen, J. (2011). Top 10 Mistakes in Web Design. Retrieved from http://www.useit.com/alertbox/9605.html

Nielsen, J. (2012a). Thinking Aloud: The #1 Usability Tool. Retrieved from http://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/

Nielsen, J. (2012b). Usability 101: Introduction to Usability. Retrieved from http://www.nngroup.com/articles/usability-101-introduction-to-usability/

Nielsen, J. & Mack, R. L. (Eds.) (1994). Usability Inspection Methods. John Wiley & Sons, New York.

Noruzi, A. (2005). Web Impact Factors for Iranian Universities. Webology, *2*(1), Retrieved from http://webology.ir/2005/v2n1/a11.html

Noruzi, A. (2006). The Web Impact Factor: A Critical Review. The Electronic Library, *24*(4), 490-500.

Olsen, M., Nietzio, A., Snaprud, M., & Fardal, F. (2009). Benchmarking and Improving the Quality of Norwegian Municipality Web Sites. Automated Specification and Verification of Web Systems, 115

Palmer, J. (2002). Web Site Usability, Design, and Performance Metrics. Information Systems Research, *13*(2), 151-167

Petricek, V., Escher, T., Cox, I., & Margetts, H. (2006). The Web Structure of E-Government - Developing a Methodology for Quantitative Evaluation. In Proceedings of the 15th international conference on World Wide Web, ACM, 669-678

Prom, C. (2007). Understanding On-line Archival Use through Web Analytics. ICA-SUV Seminar, Dundee, Scotland. Retrieved from http://www.library.uiuc.edu/archives/workpap/PromSUV2007.pdf

Rodríguez, R., Estevez, E., Giulianelli, D, & Vera, P. (2009). Assessing e-Governance Maturity through Municipal Websites- Measurement Framework and Survey Results. 6th Workshop on Software Engineering, Argentinean Computer Science Conference, San Salvador de Jujuy, Argentina, Oct 2009.

Scowen,G. (2007). Increased Website Popularity through Compliance with Usability Guidelines. A Doctoral Dissertation. Information Science, University of Otago, New Zealand

Shekofteh, M., Shahbodaghi,A., Sajjadi, S., & Jambarsang, S. (2010). Investigating Web Impact Factors of Type 1, Type 2 and Type 3 Medical Universities in Iran. Journal of Paramedical Sciences (JPS), *1*(3).

Stolz, C., Viermetz, M., Skubacz, M., & Neuneier, R. (2005). Guidance Performance Indicator- Web Metrics for Information Driven Web Sites. IEEE/WIC/ACM International Conference on Web Intelligence 2005, 186–192

Stone, D., Jarrett, C., Wodroffe, M., & Minocha, S. (2005). User Interface Design and Evaluation. San Francisco, CA: Morgan Kaufmann

Thelwall, M., & Zuccala, A. (2008). A University-Centred European Union Link Analysis. Scientometrics, *75*(3), 407–420.

Treiblmaier, H., & Pinterits, A. (2010). Developing Metrics for Web Sites. Journal of Computer Information Systems, *50*(3).

Woolrych, A., Hornbæk, K., Frøkjær, E., & Cockton, G. (2011). Ingredients and Meals rather than Recipes: A Proposal for Research that does not treat Usability Evaluation Methods as Indivisible Wholes. International Journal of Human-Computer Interaction, *27*(10), 940-970.

Zevedeos, R. (2006). Local e-Government: Reconstructing Limassol's municipality (Cyprus) Web Site to Provide Functional and Effective E-services. Master Dissertation. University of Sheffield.