WEB MINING–BASED OBJECTIVE METRICS FOR MEASURING WEB SITE NAVIGABILITY

Design Science Track

Xiao Fang*
College of Business Administration
University of Toledo
Toledo, Ohio
xiao.fang@utoledo.edu

Michael Chau*
School of Business
University of Hong Kong
Pokfulam, Hong Kong
mchau@business.hku.hk

Paul J. Hu*
School of Accounting and Information Systems
University of Utah
Salt Lake City, Utah
actph@business.utah.edu

Zhuo Yang*
School of Accounting and Information Systems
University of Utah
Salt Lake City, Utah
actpzy@business.utah.edu

Olivia R. Liu Sheng*
School of Accounting and Information Systems
University of Utah
Salt Lake City, Utah
actos@business.utah.edu

Abstract

Web site design is critical to the success of electronic commerce and digital government. Effective design requires appropriate evaluation methods and measurement metrics. The current research examines Web site navigability, a fundamental structural aspect of Web site design. We define Web site navigability as the extent to which a visitor can use a Web site's hyperlink structure to locate target contents successfully in an easy and efficient manner. We propose a systematic Web site navigability evaluation method built on Web mining techniques. To complement the subjective self-reported metrics commonly used by previous research, we develop three objective metrics for measuring Web site navigability on the basis of the Law of Surfing. We illustrate the use of the proposed methods and measurement metrics with two large Web sites.

Keywords: Web site design, Web log mining, Web site navigability

Introduction

Effective Web site design is critical to the success of electronic commerce and digital government. Substantial efforts have been undertaken to evaluate and improve Web site designs, generating a cumulative set of guidance, principles, assessment methods, and measurement metrics. Navigation support represents a fundamental aspect of Web site design (Nielson 2000) that guides and facilitates individuals' visits throughout a Web site to locate target contents in an effective and efficient fashion. According to Palmer (2002), navigation support is indispensable to Web site success.

* Contributed equally to this paper

An easy-to-navigate Web site enables visitors to access target contents by following explicitly designed user paths. A site's structure can greatly affect the overall accessibility and performance of a Web site. However, effective structure design remains challenging. As Lee et al. (2002) have noted, many Internet users are seriously concerned about issues pertaining to Web site accessibility and ease of operations.

Measurement is a necessary precursor to effective Web site structure design. A review of relevant previous research suggests the dominant use of subjective, self-reported metrics to measure the effectiveness of Web site structure. According to Trice and Treacy (1986), such subjective measures may not provide the accuracy necessary to reveal actual usability or the usage of a computer-based system. Ettema (1985) and Straub et al. (1995) report that subjective, user-reported measures are weakly correlated with objective metrics for information system usage, such as system-generated logs and statistics.

The current research examines Web site navigability, or the extent to which a visitor can follow a Web site's hyperlink structure to locate target contents successfully in an easy and efficient manner. In line with design science research guidelines (Hevner et al. 2004), our objectives are twofold: (1) to design a systematic Web mining–based method for evaluating Web site navigability; and (2) to develop three metrics that can measure the navigability of a Web site automatically and objectively by analyzing Web logs that reveal visitors' access patterns and navigation behaviors on the Web site. We take a Web mining approach to aggregate and analyze voluminous user behavioral data recorded in Web logs and thereby mitigate processing capability constraints while addressing privacy concerns commonly associated with the use of objective measures (Straub et al. 1995). In the following section, we describe the proposed Web mining–based method for evaluating Web site navigability.

A Web Mining–Based Web site Navigability Evaluation Method

We propose a method for evaluating Web site navigability that builds on the analysis of Web logs and takes into consideration both Web site structure and visitors' access/surfing patterns. A Web log refers to a collection of records that objectively document visitors' surfing behaviors on a Web site (Cooley et al. 1999; Fang and Sheng 2004). We focus on prominent, frequent access patterns and extract them from Web logs rather than directly examining all records in logs. Focusing on visitors' prominent access patterns is advantageous in several ways. First, infrequent visiting behaviors recorded in log records can distract or even mask analysis results and therefore should be filtered. By concentrating on frequent patterns, we uncover prominent surfing behaviors likely to be observed in future visits. Thus, our analysis of Web site visiting patterns not only reveals the navigability of the design but also points to areas in which the current structure design could be improved. In addition, the prominent surfing patterns extracted can be applied easily to predict the performance of a Web site and identify desirable enhancements to be implemented.

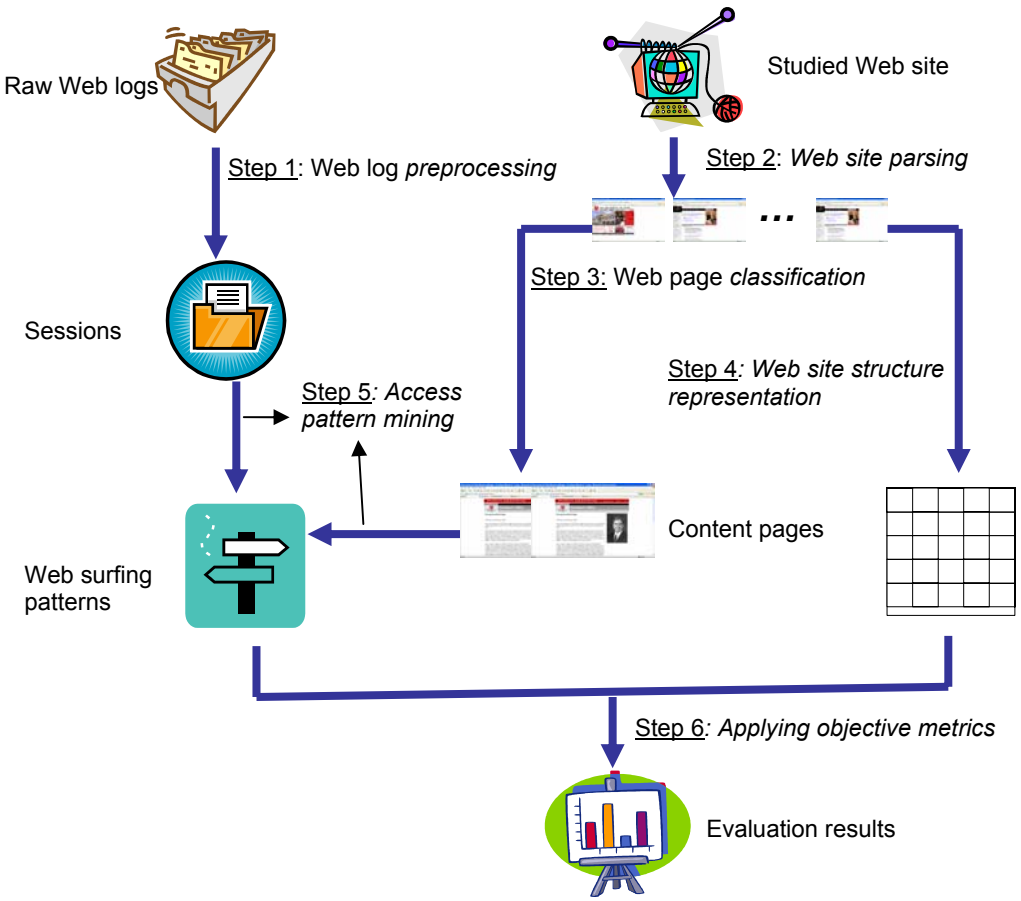


Figure 1: Proposed Web Mining–Based Method for Evaluating Web site Navigability

To derive the navigational structure of a Web site, we parse the site and represent its navigational structure with a distance matrix. The indexes of the matrix are the pages of the Web site, and its elements denote the distance between any two pages, as measured by the number of clicks required to move between them. Based on the Law of Surfing (Huberman et al. 1998), which depicts the probability of surfing distance during a Web site visit, we propose three objective metrics for measuring Web site navigability: efficiency, load, and power. Our metrics take as inputs the navigational structure of a Web site and visitors' access patterns exhibited on that Web site. Figure 1 summarizes the proposed method for analyzing and evaluating Web site navigability on the basis of our proposed metrics. We detail the proposed method in the following subsections.

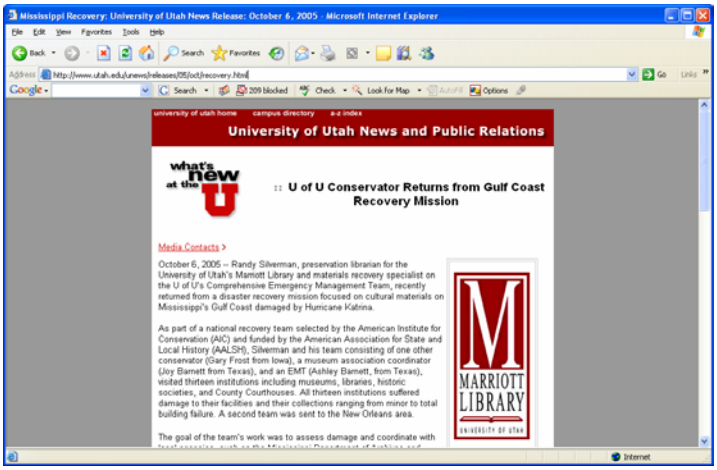


Figure 2: A Sample Web page

Web log Preprocessing

A visitor's request for a Web page like the example one shown in Figure 2 generates three Web log records (see Table 1). The first record documents the user's request to this particular Web page, and the remaining records are created for requests made to obtain the images inlined on the requested page. A typical log record consists of an IP address, time, and URL (describing by whom and at what time a page was requested), the status of the request (success or failure), and the size of the content transmitted to the user.

Table 1: Sample Web log Records

IP Address	Time	Method	URL	Protocol	Status	Size
128.196.xxx.xxx[1]	[07/Oct/2005:05:38:33 -0700]	GET	/unews/releases/05/oct/recovery.html	HTTP 1.0	200	4134
128.196.xxx.xxx	[07/Oct/2005:05:38:33 -0700]	GET	/unews/uwhatsnew.gif	HTTP 1.0	200	765
128.196.xxx.xxx	[07/Oct/2005:05:38:33 -0700]	GET	/unews/library_sm.gif	HTTP 1.0	200	864

Web logs must be preprocessed to be analyzable (Cooley et al. 1999). Generally preprocessing includes cleaning and session identification. To clean Web logs, we remove erroneous and accessory log records. Erroneous log records are generated when access requests fail for various reasons and can be identified by a status code of 400 or above, whereas accessory log records document requests associated with a user request for a particular Web page, such as requests for pictures or images inlined on the user-requested page.

The resultant (cleaned) Web logs then may be analyzed to identify visit sessions, each of which constitutes a basic processing unit for the discovery of interesting, prominent access patterns. Consistent with Cooley et al. (1999), we define a visit session as a sequence of Web page accesses during a single visit to a Web site. Because of the stateless connections between the client and the web server, Web log records have no visit session designations. Therefore, to reconstruct visit sessions from Web log records, we use the common timeout method, according to which a visit session is considered terminated when the time elapsed between page requests exceeds a specified limit. The timeout method is fairly effective and accurate for identifying visit sessions; according to Spiliopoulou et al. (2003), more than 90% of genuine visit sessions can be identified by the timeout method. We group the cleaned log records into different visit sessions, using the rule of thumb that stipulates that the duration of a visit session cannot exceed 30 minutes (Spiliopoulou et al. 2003).

Web site Parsing

A Web site can be parsed by a web spider program that gathers and parses all its pages. Web spiders, or crawlers, are software programs that automatically collect pages from the Web by following hyperlinks exhaustively. In this study, we use SpidersRUs, developed by Chau et al. (2005), to gather Web pages by specifying the URL of the homepage of the studied Web site as the starting URL. When accessing a page, SpidersRUs downloads the page and extracts its hyperlinks, downloads the pages pointed to by the extracted hyperlinks, and then extracts their hyperlinks. The process continues until SpidersRUs has visited all pages on the Web site. Subsequently each page downloaded by SpidersRUs is parsed to generate useful features from the page, as we detail next.

Web page Classification

Each page downloaded and parsed in Web site parsing is classified as either a content page or an index page. An index page mostly consists of hyperlinks pointing to other Web pages and primarily serves navigational purposes. In contrast, a content page typically contains textual contents and other information potentially of interest to visitors; thus, content pages are generally the pages for which visitors search. The sheer volume of pages available on a Web site makes a manual classification approach prohibitively costly, if at all feasible. We therefore use an automatic classifier to differentiate index and content pages. Specifically, we use a classifier built upon Support Vector Machine (SVM),[2] a common computational method for searching for a hyperplane that best separates two classes

[1] Full, genuine IP addresses are not disclosed for privacy reasons.

[2] Specifically, our classifier is based on the SVM-light implementation (Joachims 1999).

(Vapnik 1998). We choose SVM primarily because of its effectiveness in text-based classification tasks (Yang and Liu, 1999).

Our goal is to classify Web pages on the basis of their structure rather than their content; therefore, we take a feature-based classification approach (Chau 2004) instead of a conventional keyword-based approach. Based on our observations and a review of related research (e.g. Gillenson et al. 2000), we select six fundamental page structure features pertaining to size and number of links: (1) number of outgoing links, (2) number of internal outlinks (links pointing to other pages in the same domain), (3) number of external outlinks (links pointing to other pages not in the same domain), (4) length of the document, (5) number of words in the document, and (6) number of anchor text words in the document. Each downloaded page possesses a specific value for each feature. We randomly selected a set of downloaded pages as the training sample and asked two domain experts, highly experienced in and knowledgeable about Web site design and administration, to classify them manually. These "manually tagged" pages then enable us to construct a SVM model. According to our empirical evaluation of several alternative kernels, the Radial Basis Function (RBF) kernel (g = 0.4) appears to achieve the best performance with an average accuracy of 88.2%; therefore, we use it to classify the downloaded pages as index versus content pages.

Web site Structure Representation

We represent the navigational structure of a Web site using a distance matrix of the pages on the Web site. The indexes of the matrix are Web pages, wherein the elements of the matrix denote the distance between any two pages. We consider a Web site a directed graph and denote pages as vertices and hyperlinks as edges. We measure the distance between two pages with the distance definition in graph theory. Specifically, the distance from vertex u to vertex v in a graph is the length of the shortest path from u to v (West 2001). Accordingly, the distance from page p_1 to page p_2 on a Web site can be measured by the length of the shortest path from p_1 to p_2, or the number of hyperlinks surfed or number of clicks required. When no path connects page p_1 to page p_2, the distance from p_1 to p_2 is considered ∞.

For a Web page p, we represent the set of Web pages pointed to by the hyperlinks on page p as $t(p)$. In this case, we can obtain $t(p)$ by parsing p. We denote $D(p,l)$ as a set of Web pages for which the distance from page p to another page in $D(p,l)$ is l clicks ($l = 1,2,\cdots$). We thus can derive the following equations:

$$D(p,1) = t(p), \text{ and} \tag{1}$$

$$D(p,2) = \bigcup_{\forall k \in D(p,1)} t(k) - D(p,1). \tag{2}$$

Generalizing equations (1) and (2), we have,

$$D(p,l) = \begin{cases} t(p) & \text{if } l = 1 \\ \bigcup_{\forall k \in D(p,l-1)} t(k) - \bigcup_{j=1}^{l-1} D(p,j) & \text{if } l > 1 \end{cases}. \tag{3}$$

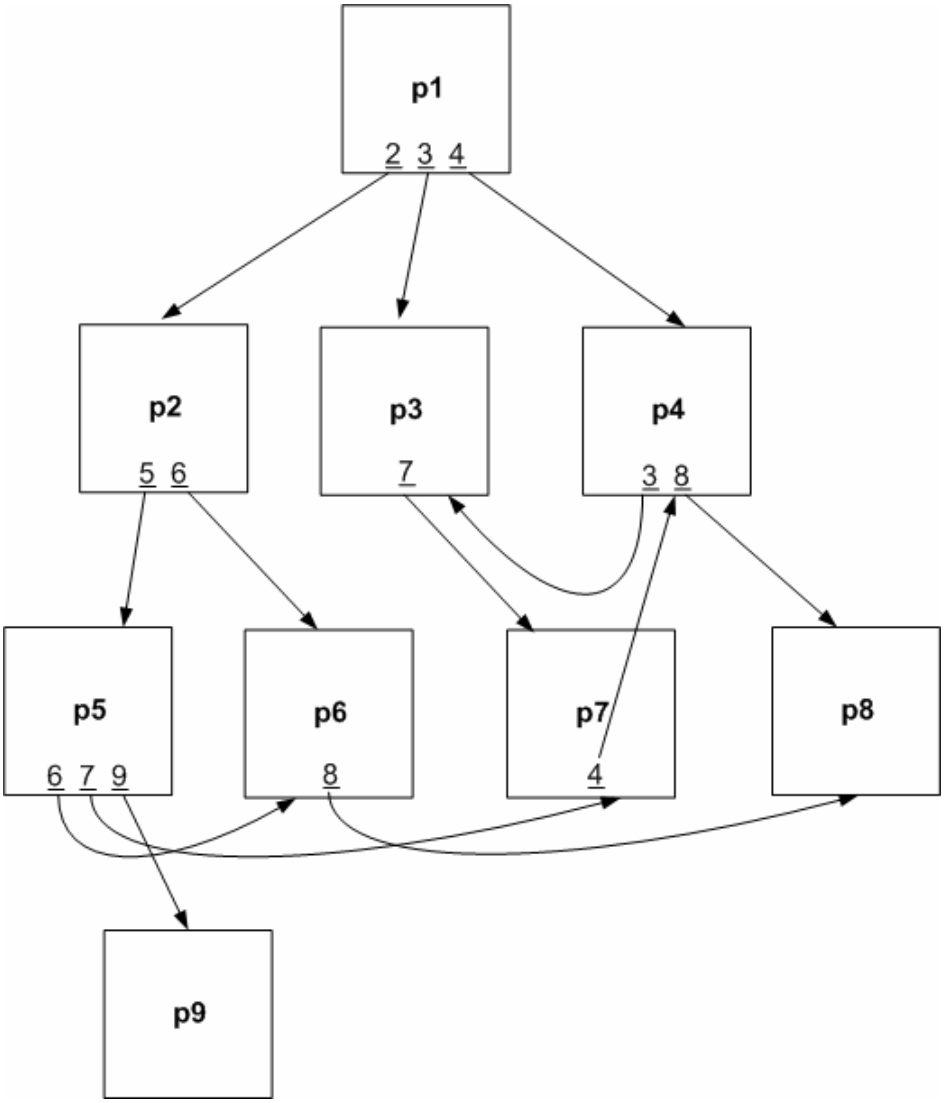


Figure 3: Representation and Analysis of Web site Structure: An Illustrative Example

We illustrate $D(p,l)$ using a sample Web site (shown in Figure 3) that consists of nine Web pages (i.e., $p_1, p_2, \cdots, p_9$) and eight hyperlinks (i.e., $1,2,\cdots,8$) that point to Web pages $p_1, p_2, \cdots, p_8$. We formulate the example as follows:

$$\begin{aligned} D(p_1,1) &= t(p_1) \\ &= \{p_2, p_3, p_4\} \end{aligned}$$

As shown, pages $p_2, p_3,$ and p_4 are one click away from page p_1. According to equation (3),

$$\begin{aligned} D(p_1,2) &= \bigcup_{\forall k \in D(p_1,1)} t(k) - D(p_1,1) \\ &= t(p_2) \cup t(p_3) \cup t(p_4) - D(p_1,1). \\ &= \{p_5, p_6, p_7, p_8\} \end{aligned}$$

Therefore, pages $p_5, p_6, p_7,$ and p_8 are two clicks away from page p_1. Similarly,

$$\begin{aligned} D(p_1,3) &= \bigcup_{\forall k \in D(p_1,2)} t(k) - [D(p_1,1) \cup D(p_1,2)] \\ &= t(p_5) \cup t(p_6) \cup t(p_7) \cup t(p_8) - [D(p_1,1) \cup D(p_1,2)] \\ &= \{p_4, p_6, p_7, p_8, p_9\} - [\{p_2, p_3, p_4\} \cup \{p_5, p_6, p_7, p_8\}] \\ &= \{p_9\}. \end{aligned}$$

Hence, page p_9 is three clicks away from page p_1.

$$D(p_1,4) = \bigcup_{\forall k \in D(p_1,3)} t(k) - [D(p_1,1) \cup D(p_1,2) \cup D(p_1,3)] = \phi.$$

Because $D(p_1,4)$ is ϕ, it is not necessary to continue the analysis; that is, $D(p_1,l) = \phi$ for all $l \geq 4$. In Figure 4, we summarize the overall procedure of obtaining $D(p,l)$ for all pages on a Web site. Deriving a distance matrix of Web pages from $D(p,l)$ is trivial and therefore not discussed here.

Input: a set of Web pages in a Web site.
Output: $D(p,l)$ for each Web page p.

for each Web page p,
  finding $t(p)$ by parsing Web page p
end for
for each Web page p
  $i = 1$
  $D(p,1) = t(p)$
  while $D(p,i) \neq \phi$
    $D(p,i+1) = \phi$
    for each $k \in D(p,i)$
      for each $h \in t(k)$
        if $h \notin D(p,j)$ for all $j \leq i$
          $D(p,i+1) = D(p,i+1) \cup \{h\}$
        end if
      end for
    end for
    $i = i + 1$
  end while
end for

Figure 4: An Algorithm for Finding $D(p,l)$

Access Pattern Mining

Prominent page access patterns are critical to Web site structure design and therefore must be extracted from Web logs. Commercial Web log analysis tools (e.g. WebTrends) offer basic utilities for identifying simple surfing patterns, such as the most frequently visited pages. These simplistic patterns can be derived from the statistics generated for individual Web pages. When visiting a Web site, a visitor typically accesses a set of pages sequentially, but statistics about the individual pages do not effectively reflect the visitor's access behavior because

they do not include important information about one's surfing behaviors, such as pages accessed together in a visit session or the exact page access sequence. We target Web surfing patterns — that is, prominent consecutive page access sequences — that can be identified by mining the Web logs.

We denote a Web log that includes *k* visit sessions as $S = \{s_i\}, i = 1,2,...,k$, where $k \geq 1$. A visit session s_i is represented by *m* Web pages, sequentially accessed in that session, $s_i =< p_1, p_2, ..., p_m >$, where $p_i, i = 1,2,...,m$ represents a Web page and $m \geq 1$. For each session s_i in S, non-content pages are removed from s_i. A consecutive sequence $< a_1, a_2, ..., a_n >$ represents the sequence of content pages accessed consecutively, and $n \geq 1$. A consecutive sequence $< a_1, a_2, ..., a_n >$ is contained in a visit session $s_i =< p_1, p_2, ..., p_m >$ if and only if there exists l, $1 \leq l \leq m - n + 1$, such that $a_i = p_{l+i-1}, i = 1,2,...,n$. For the Web log listed in Table 2, a consecutive sequence $< p_1, p_4 >$ is contained in visit sessions s_2 and s_4 but not in sessions s_1, s_3, or s_5. The visiting rate $v(c)$ of a consecutive sequence c is calculated as the ratio between the number of visit sessions that contain this particular sequence and the total number of visit sessions. For the Web log summarized in Table 2, $v(< p_1, p_4 >) = 2/5 = 0.4$. The key consecutive sequences are those whose visiting rate is greater than a pre-specified threshold. Given a user-specified threshold 0.35, $< p_1, p_4 >$ is a key consecutive sequence.

Table 2: An Illustration of Key Consecutive Sequences

Session	Content Pages Sequentially Visited in Session
s_1	$< p_1, p_3, p_4 >$
s_2	$< p_1, p_4, p_5 >$
s_3	$< p_4, p_5 >$
s_4	$< p_1, p_4, p_2 >$
s_5	$< p_1, p_5 >$

Extracting all key consecutive sequences from a Web log is a nontrivial task. The search space is vast because of the exponential nature of the number of potential key consecutive sequences and the sheer volume of records in a Web log, which can easily number in the millions. We adapt the AprioriAll algorithm, developed by Agrawal and Srikant (1995), as an efficient algorithm for discovering large sequences from a customer transaction database, to extract key consecutive sequences. Specifically we modify this algorithm by incrementing the count of a consecutive sequence only if the content pages contained in the consecutive sequence appear in a visit session consecutively. Details of the AprioriAll algorithm are available in Agrawal and Srikant (1995).

Applying Objective Metrics

Using the identified prominent page access patterns and the Web site's structure as inputs, we apply objective metrics to evaluate the navigability of the Web site under study. We describe these proposed metrics in the following section.

Proposed Objective Metrics for Measuring Web site Navigability

As previous research has shown, a visitor's surfing behavior on a Web site is not completely random but rather exhibits some regularity. Huberman et al. (1998) propose the Law of Surfing, which characterizes the probability $p(l)$ of surfing l hyperlinks within a Web site visit according to the following equation:

$$p(l) = \sqrt{\frac{\lambda}{2\pi l^3}} \exp\left[\frac{-\lambda(l-\mu)^2}{2\mu^2 l}\right], \qquad l \geq 1. \tag{4}$$

The mean and variance of l are μ and $\frac{\mu^3}{\lambda}$, respectively. According to Huberman et al. (1998), $p(l)$ exhibits a good fit with real-world surfing data, as manifested by a mean of 3.86 and a variance of 6.08. On the basis of the $p(l)$ measure proposed by Huberman et al. (1998), we define $G(l)$ as the probability of surfing at least l hyperlinks within a Web site visit:

$$G(l) = \sum_{\forall x \geq l} p(x), \qquad l \geq 1. \tag{5}$$

We thus can derive the following equation:

$$G(l) = \begin{cases} 1 & \text{if } l = 1 \\ G(l-1) - p(l-1) & \text{if } l > 1 \end{cases}. \tag{6}$$

Using equations (4) and (6), we can derive $G(l)$ from $l = 1$.

In this study, we propose three objective metrics for measuring Web site navigability: efficiency, load, and power. *Efficiency* refers to the ease with which the visitor can locate target contents, whereas *load* measures the level of difficulty the visitor encounters in deciding on a navigation direction in the course of locating the target contents. In addition, *power* reveals the probability that the visitor successfully will locate target contents. Broadly, target contents refer to the key consecutive sequences described previously.

Let *C* be a set of *n* key consecutive sequences extracted from a Web log, $C = \{c_i\}, i = 1,2,\cdots,n.$ The efficiency of locating a key consecutive sequence $c_i =< p_1, p_2, \cdots, p_m >$, $E(c_i)$, is calculated as follows:

$$E(c_i) = \frac{mN - \min(d(h, p_1), N) - \sum_{j=1}^{m-1} \min(d(p_j, p_{j+1}), N)}{mN - m}. \tag{7}$$

In equation (7), m represents the number of content pages in c_i, and N denotes a constant. A content page is considered most difficult to locate when it is at least N clicks away, whereas a page is least difficult to locate if it is only one click away. In the study, we set N to 10 because 97% of the visitors would not access or attempt to access pages located 10 or more clicks away; that is, $G(10) = 0.03$. Furthermore, $d(h, p_1)$ denotes the distance from homepage h to p_1, the first visited page in c_i. Thus, we assume a search for a key consecutive sequence starts at the homepage of a Web site, a reasonable assumption consistent with our preliminary analysis of actual Web site visit patterns. For example, in May 2005, we observed that more than 80% of visit sessions on the University of Utah's Web site started from its homepage. In equation (7), $d(p_j, p_{j+1})$ represents the distance between two subsequently visited pages in c_i, and $\min(x, y)$ is the smallest number in x and y. The value of $E(c_i)$ is within the range of [0,1], inclusively. In particular, the value of $E(c_i)$ equals 1 if and only if $d(h, p_1) = 1$ and $d(p_j, p_{j+1}) = 1$ for $j = 1,2,\cdots m-1$. On the other end, the value of $E(c_i)$ is 0 if and only if $d(h, p_1) \geq N$ and $d(p_j, p_{j+1}) \geq N$ for $j = 1,2,\cdots m-1$. The greater the value of $E(c_i)$, the higher is the efficiency of locating c_i. Because not all key sequences are equally significant, we introduce the weight $w(c_i)$ of c_i in *C*, calculated as follows:

$$w(c_i) = \frac{v(c_i)}{\sum_{\forall c \in C} v(c)}, \tag{8}$$

where $v(c_i)$ is the visiting rate of c_i.

The efficiency of a Web site can then be approximated by the efficiency $E(C)$ of locating key consecutive sequences in *C*:

$$E(C) = \sum_{i=1}^{n} w(c_i) E(c_i). \tag{9}$$

The cognitive load incurred by the visitor when determining the next move from the current page on the Web site correlates positively with the number of hyperlinks on that page. That is, the more hyperlinks a page has, the more stringent is the cognitive load required to decide a direction for subsequent navigation. Hence, the load of locating a key consecutive sequence c_i, $L(c_i)$, can be measured as the average number of hyperlinks on Web pages located in the path that leads to all the content pages in c_i from the homepage. Therefore, we measure the load of a Web site using the load $L(C)$ of locating key consecutive sequences in *C*:

$$L(C) = \sum_{i=1}^{n} w(c_i) L(c_i). \tag{10}$$

We also measure the power of locating a key consecutive sequence $c_i =< p_1, p_2, \cdots, p_m >, P(c_i)$, defined as the probability of locating all content pages in c_i. Assuming the search for c_i starts from the homepage *h* of a Web site, the distance from *h* to the first visited page p_1 in c_i is $d(h, p_1)$. Visitors who are willing to surf at least $d(h, p_1)$ hyperlinks can locate p_1. According to equation (5), the probability of surfing at least $d(h, p_1)$ hyperlinks is $G(d(h, p_1))$. That is, the probability of locating p_1 is $G(d(h, p_1))$. Similarly, the probability of locating p_j in c_i is $G(d(p_{j-1}, p_j))$, where $2 \le j \le m$. The power $R(c_i)$ of locating a key consecutive sequence c_i thus can be calculated as follows:

$$R(c_i) = G(d(h, p_1)) \prod_{j=2}^{m} G(d(p_{j-1}, p_j)). \tag{11}$$

We approximate the power of a Web site using the power $R(C)$ of locating key consecutive sequences in *C* as:

$$R(C) = \sum_{i=1}^{n} w(c_i) R(c_i). \tag{12}$$

Applications of the Proposed Method and Metrics: An Illustration

We apply the proposed method and metrics to evaluate the navigability of two large Web sites: the University of Utah's (hereafter referred to as utah.edu) and the State Government of Utah's Web portal (hereafter referred to as utah.gov).[3] We obtained Web logs for one month from each Web site and use them to assess Web site navigability. Following the method previously described, we cleaned the log records and reassembled the visit sessions. The

[3] The investigated Web sites can be accessed at http://www.utah.edu and http://www.utah.gov, respectively.

utah.edu logs contained 907,907 visit sessions, and the utah.gov logs had 259,479 visit sessions. We then parsed the Web pages and identified content pages on each Web site, following the procedure we described previously. In Table 3, we summarize some important descriptive statistics of the studied Web sites.

Table 3: Descriptive Statistics of utah.edu and utah.gov

Web site	No. of Content Pages	No. of Index Pages	Total No. of Pages	% Content Pages
utah.edu	5,246	463	5,709	92%
utah.gov	1,971	119	2,090	94%

We set the threshold to 0.2% and mined all key consecutive sequences from the preprocessed Web logs. The large number of visit sessions in the Web logs favors the use of a small threshold to identify a reasonable number of meaningful key consecutive sequences; a large threshold value would identify only few key consecutive sequences. We apply the proposed objective metrics to evaluate the identified key consecutive sequences. In Table 4, we summarize the key evaluation results obtained from utah.edu versus utah.gov in terms of these proposed metrics.

Table 4: Evaluation Results for utah.edu and utah.gov (Threshold = 0.2%)

Web site	Efficiency	Power	Load
utah.edu	0.97	0.97	153
utah.gov	0.79	0.70	103

According to our evaluation results, the efficiency of utah.edu is satisfactory, approaching the highest efficiency attainable by any Web site (i.e., 1); whereas the efficiency of utah.gov is substantially lower. The power of utah.edu is 0.97, which suggests an average probability of 0.97 that a visitor will successfully locate his or her target contents on this Web site. We observe a reasonable power level for utah.gov, on which a visitor on average has a 0.7 probability of successfully locating target contents. It is worth noting that both efficiency and power are calculated on the basis of the distance between Web pages and that the distance between any two Web pages equals the distance of the shortest path between them. As a result, the efficiency and power statistics in Table 4 represent the highest attainable or upper bounds of efficiency and power for these sites. Overall, utah.edu appears to be more navigable than utah.gov, according to their efficiency and power. Table 5 summarizes the average number of clicks required to locate a content page in each of the top-10 key consecutive sequences on utah.edu and utah.gov. In general, locating such a content page on utah.gov demands more clicks than are required to locate a content page in a top-10 key consecutive sequence on utah.edu.

Utah.edu records a load of 153; on average, a visitor must choose from a total of 153 hyperlinks when deciding on the next move on the Web site. Our analysis indicates the load of utah.gov is 103; hence, this Web site has considerably fewer average hyperlinks on any page. Both utah.edu and utah.gov have relatively high load values, possibly because both sites have unique informational functions that require a large collection of interrelated content pages that point to one another. A related but subtly different explanation suggests that both Web sites have a considerable number of frequently visited index pages that contain navigation bars with a few dozen hyperlinks. For example, we find 147 hyperlinks in the hompage of utah.edu. Similarly, the homepage of utah.gov contains 117 hyperlinks. In Figure 5, we show a frequently visited index page that contains a total of 116 hyperlinks.

Table 5: Average Number of Clicks to Locate a Content Page in a Top-10 Key Consecutive Sequence in utah.edu and utah.gov

utah.edu		utah.gov	
Rank of Key Consecutive Sequences by Visiting Rate	**Average Number of Clicks to Locate Content Page**	**Rank of Key Consecutive Sequences by Visiting Rate**	**Average Number of Clicks to Locate Content Page**
1	1	1	1
2	1	2	1
3	1	3	3
4	2	4	1
5	1	5	3
6	1	6	1
7	1	7	2
8	1	8	3
9	1.5	9	1
10	1	10	2

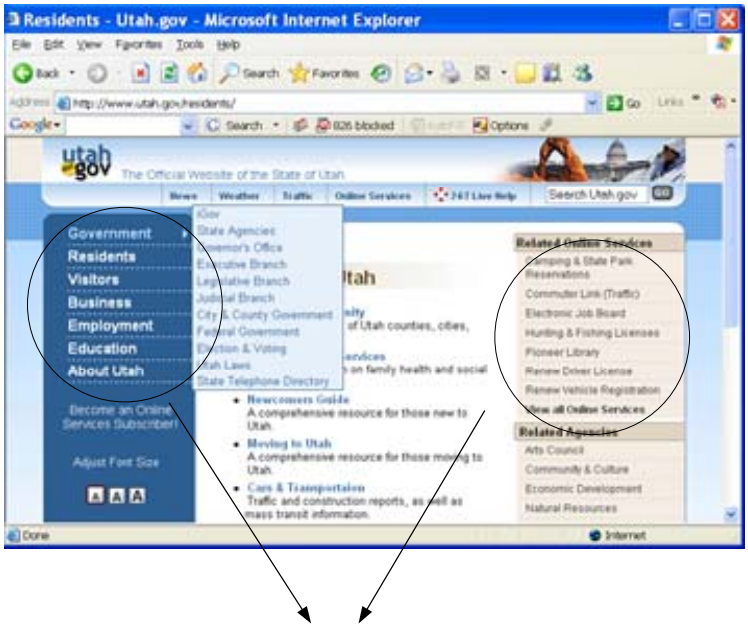
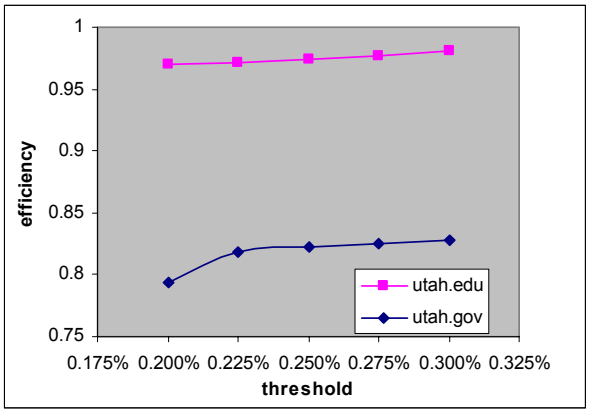
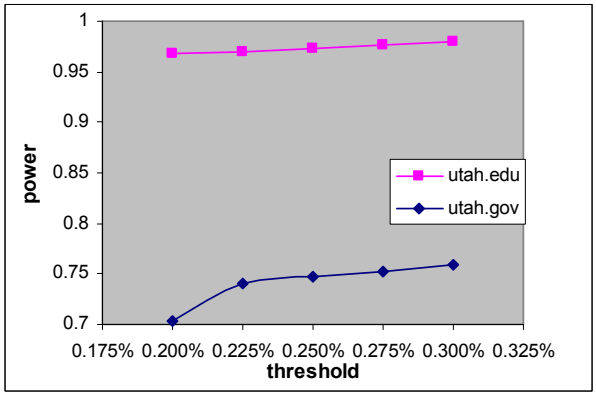


Figure 5: A Sample Index Page with Navigation Bars

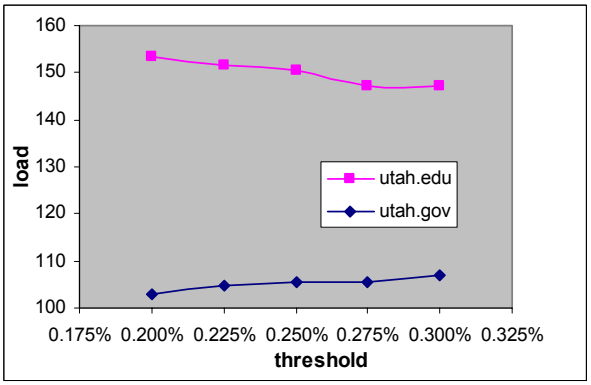
To examine the adequacy of the selected threshold value (i.e., 0.20%), we perform a series of sensitivity analyses over the range 0.2–0.3% in increments of 0.025%. As shown in Figure 6, the efficiency, power, and load of the respective Web sites appear generally stable over the range of threshold values. The observed robustness suggests that our choice of threshold value is reasonable and does not seem to affect the evaluation results (in terms of efficiency, power, and load) significantly.



(a)



(b)



(c)

Figure 6: (a) Efficiency (b) Power, and (c) Load of utah.edu and utah.gov

As shown in Figure 6, the navigability of utah.edu is consistently higher than that of utah.gov, as measured by efficiency and power. However, utah.gov offers a better load than does utah.edu. We recall that the navigability of a Web site generally is affected by its size (i.e., number of Web pages) and structure design (i.e., positioning of hyperlinks and their inter-linkages). Conceivably, visitors can locate target contents on a small Web site more easily than they can on a large Web site with substantially more Web pages. In this study, we focus on the structural aspect of Web site navigability but also need to consider Web site size to extend our proposed method and metrics for Web site navigability evaluations.

Conclusions and Ongoing Research

The current research examines Web site navigability with a particular focus on the structural aspect of Web site design. We propose a systematic method for evaluating Web site navigability and develop three objective metrics to measure it, and then illustrate the applications of the proposed method and metrics using two large, real-world Web sites. Thus, we make several important contributions to Web site design research. First, compared with conventional Web site design evaluations that use subjective, self-reported measurements, our proposed method, which uses objective metrics, is more cost effective (greatly reducing data collection costs), efficient (producing evaluation results in minutes), and flexible (offering baseline measurements that can be extended or refined). Second, we contribute to web metrics research. Most prior research on web metrics (e.g. Dhyani et al. 2002; Zhang et al. 2004) attempts to evaluate Web site structure design using a Web site's structural information and typically considers all the pages on a Web site equally important. However, a frequently visited Web page intuitively is more significant in

Web site structure design evaluations than those seldom requested by visitors. In response, we integrate content, structure, and usage information to evaluate Web site navigability in an innovative and comprehensive fashion.

Our ongoing research agenda includes designing and conducting controlled experiments to evaluate real-world Web sites with considerably different navigability in terms of efficiency, power, and load. These experiments will provide empirical validations of our proposed metrics and include navigation behaviors and task performance on representative Web sites from studied domains that may include education, government, and business. Another area worthy of exploration is determining how to apply our proposed metrics to identify probable limitations in the navigational structure of a Web site, which could improve existing Web site design considerably. In addition, the assumption that users search for target contents by starting at the homepage of a Web site needs to be relaxed to make the proposed metrics robustly applicable in different evaluation scenarios.

References

Agrawal, R. and Srikant R. "Mining Sequential Patterns," in *Proceedings of the 11th International Conference on Data Engineering*, Taipei, 1995, pp. 3-14.

Chau, M. "Applying Web Analysis in Web page Filtering," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, Tucson, Arizona, USA, June 7-11, 2004, p. 376.

Chau, M., Qin, J., Zhou, Y., Tseng, C., and Chen, H. "SpidersRUs: Automated Development of Vertical Search Engines in Different Domains and Languages," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, Denver, Colorado, USA, June 7-11, 2005, pp. 110-111.

Cooley, R., Mobasher, B., and Srivastava, J. "Data preparation for mining World Wide Web browsing patterns". *Knowledge and Information Systems* 1:1, 1999, pp. 1-27.

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

Ettema, J.S. "Explaining information system use with system-monitored versus self-reported measures" *Public Opinion Quarterly*, 49, 1985, 381-387.

Fang, X. and Liu Sheng, O. R. "LinkSelector: A Web Mining Approach to Hyperlink Selection for Web Portals," *ACM Transactions on Internet Technology*, 4(2), 2004, pp. 209-237.

Gillenson, M. L., Sherrell, D. L., and Chen, L., "A taxonomy of web site traversal patterns and structure," *Communication of the Association for Information Systems*, 3, June 2000.

Hevner, A., March, S., Park, J., and Ram, S., "Design science in information systems research," *MIS Quarterly*, 28(1), 2004, pp. 75-105.

Huberman, B. A., Pirolli, P., Pitkow, J.E. and Lukose, R. "Strong regularities in World Wide Web Surfing," *Science* 280, 3 (1998), 95-97.

Joachims, T. "Making large-Scale SVM Learning Practical," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, Y.R. "IMQ: A Methodology for Information Quality Assessment," *Information and Management* 40:2, December, 2002, pp. 133-146

Nielsen, J., *Designing Web Usability*, New Riders Publishing, 2000.

Palmer, J. W., "Web Site Usability, Design, and Performance Metrics", *Information Systems Research*, 13:2, Jun 2002.

Spiliopoulou M., Mobasher B., Berendt B., and Nakagawa M. "A framework for the evaluation of session reconstruction heuristics in web-usage analysis", *INFORMS Journal on Computing* 15, 2 (2003), pp. 171-190.

Straub, D., Limayem, M. and Karahanna-Evaristo. "Measuring System Usage: Implication for IS Theory Testing", *Management Science*, 41: 8, August 1995, pp.1328-1342

Trice, A. W. and M. E. Treacy, "Utilization as a dependent variable in MIS research", *Proceedings of the International Conference on Information Systems*, 1986

Vapnik, V. *Statistical Learning Theory*, Wiley, Chichester, GB, 1998.

West, D. B. *Introduction to Graph Theory*, 2nd edition. Prentice Hall, 2001.

Yang, Y. and Liu, X. "A Re-examination of Text Categorization Methods," *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR'99)*, 1999, pp. 42-49.

Zhang, Y., Zhu, H., and Greenwood, S., "Web site complexity metrics for measuring navigability," *Proceedings of the Fourth International Conference on Quality Software*, 2004, pp. 172-179.