Implicit Measures of Lostness and Success in Web Navigation

Jacek Gwizdka [1] [2], Ian Spence [3]

[2)] School of Communication, Information, and Library Studies,
Rutgers University,
4 Huntington St, New Brunswick, NJ 08901, USA
iwc2006@gwizdka.com

[3)] Department of Psychology
University of Toronto
100 St. George St, Toronto, Ontario M5S 3G3, Canada
spence@psych.utoronto.ca

This is the authors' version of the manuscript accepted for publication in Interacting with Computers.

[1] Corresponding author

Implicit Measures of Lostness and Success in Web Navigation

Abstract

In two studies, we investigated the ability of a variety of structural and temporal measures computed from a web navigation path to predict lostness and task success. The user's task was to find requested target information on specified websites. The web navigation measures were based on counts of visits to web pages and other statistical properties of the web usage graph (such as compactness, stratum, and similarity to the optimal path). Subjective lostness was best predicted by similarity to the optimal path and time on task. The best overall predictor of success on individual tasks was similarity to the optimal path, but other predictors were sometimes superior depending on the particular web navigation task. These measures can be used to diagnose user navigational problems and to help identify problems in web site design.

Keywords

Web navigation; Web navigation graph; Navigation path similarity; Implicit measures; Lostness; Compactness; Stratum; User studies.

Classification

2240.000 Empirical data; 4200.000 Metrics; 7230.000 User studies; 7520.000 Web & hypertext navigation

1. Introduction and Motivation

Navigating large, complex websites is frequently difficult. The task places a heavy cognitive burden on users who often become lost or disoriented. Indeed, cognitive overload and lostness have long been recognized as major barriers experienced by users in hypermedia navigation (Conklin, 1987). Disoriented searchers seem to have difficulty forming a cognitive model of the information structure (Kim & Hirtle, 1995; Dieberger, 1997; Boechler, 2001). Since the structure of the information space is usually not transparent it is often difficult for users to navigate in a goal-directed way (Dieberger, 1995). Users can become lost because of the non-linear nature of hypertext systems (Chen & Macredie, 2002) and, if there is considerable cross-referencing among pages, looping behavior may result (Boechler, 2001). However, despite a long history of research on hypertext and more recent studies in the area of web navigation, relatively little is known about the statistical relationships among web navigation patterns, lostness, and success on information-seeking tasks. In our view, a better understanding of the characteristics of successful and unsuccessful navigation will be assisted by the computation of structural and temporal measures that quantify different aspects of navigation behaviors. An important potential benefit of our structural approach is that it may be possible to suggest strategies for improving the user experience that do not depend on an analysis of the content of web pages

We present the results of two exploratory studies that examine the relationships among structural measures that characterize web navigation paths, lostness, and task success. Related previous work on the structural aspects of web navigation is discussed in section two and our approach is described in section three. The empirical studies designed to evaluate the various measures—old and new—are presented in sections four to six. The paper ends with a discussion of the results, conclusions, and future directions.

2. Related Work

2.1 Web-navigation Graph

Hypertext is traditionally conceptualized (and visualized) using the traditional node-and-link model to present both the structure and the use of the web. Graphs that represent user navigation

on the web are called *web-navigation graphs*, or *web-usage graphs*. Visited web pages are represented by graph nodes and traversed links are represented by the edges of the graph. The focus of this paper is on quantifying the structural and temporal aspects of a user's navigational history. While clearly also important, we do not study the influence of the content of web pages on the user's navigational choices (such influence was studied by, for example, Chi et. al., 2003; Berendt, 2002).

Recently, the structural properties of web-navigation graphs have been correlated with user task outcomes. McEneaney (McEneaney, 2001) demonstrated that learning task success (using a task which required a broad exploration of on-line material) was correlated with shallow and broad hierarchical navigation (reflected in high compactness[1] of the navigation graph), while task failure was related to a linear style of navigation (high stratum) (Botafogo et al., 1992). Shih and his colleagues (Shih et al., 2004) used web-based courseware to study navigation behavior, finding that the navigation paths of people who had greater prior experience with web-based instructional tools were more linear and less dispersed (high stratum & low compactness). They also found that stratum and compactness for those more experienced people differed according to the task phase: (1) exploration, (2) resolution, and (3) completion. Significant relationships among stratum, compactness and navigation task outcomes have not been found in all studies. For example, (Herder, 2003) reported that no correlation was found between these two graph measures and user disorientation. However, the tasks employed in that study were of mixed type; some were open-ended while others were goal-oriented. It is thus natural to enquire how the observed relations among stratum, compactness, lostness, and navigation task success change for different types of information-seeking tasks.

2.2 Navigation Styles

In a study reported in (Juvina & van Oostendorp, 2004; Herder & Juvina, 2004), compactness, stratum, path density, and average connected distance were used to characterize user navigation styles. The navigation styles were second-order constructs derived from observable and measurable user behavior. Factor analysis was used to create aggregate measures and two

[1] Compactness and stratum are defined in section 3.1.2.

navigation styles, *flimsy navigation* and *laborious navigation,* were proposed. These two styles accounted for 27% and 23% of the total variance, respectively In the flimsy style, users often returned to previously visited pages, including the start page. The preferred return mechanism was the web browser's back button, rather than via direct links. In contrast, the laborious navigation style involved thorough exploration of the website making extensive use of the navigational mechanisms provided by the site. On page revisits, users typically followed a different link and thus explored a different branch of the website. It seems likely that this variation in navigation styles might be helpful in predicting navigation task outcomes.

2.3 Navigation Sequence Similarity

Web pages arranged in the order of user visits form a *navigation sequence* or *surfing path*. This path has been found to be useful in a variety of applications. For example, (Pitkow & Pirolli, 1999) used a longest repeated sequence algorithm to predict user surfing paths. They also used the similarity of the navigation sequence to the most common sequence to facilitate efficient caching of web pages. Path similarity has also been used in the analysis and clustering of user navigation behavior: (Wang & Zaïane, 2002) employed a sequence alignment algorithm to cluster user web navigation sessions. We use a similar algorithm to assess similarity between the user navigation path and the optimal navigation path.

2.4 Implicit Measures of User Behavior

Implicit measures of user behavior may be used to predict subjective user preferences. This approach has a long history in Information Retrieval (IR), where relevance feedback is used to indicate a user's information interests and preferences (Kelly & Teevan, 2003; Oard & Kim, 2001). While older systems were based on explicit feedback (Spink & Losee, 1996), more recent work in IR employs implicit measures. Implicit measures are observable measures of user behavior that can be used to infer or predict user attitudes, interests, preferences, or user performance on a task. In the context of web search and navigation, implicit measures can be used to predict user satisfaction, user lostness, or task success. In a recent paper, (Fox et al., 2005) used implicit measures of user interest and satisfaction on web search tasks. They found that time on a web page, clickthrough, and what a user did after visiting a search result or how a user ended a search session were good predictors of user satisfaction. Some implicit measures

seem to be sensitive to task context. For example, several studies found document reading time (e.g. time on a web page) to be a good indicator of document relevance to the user (Masahiro & Yoichi, 1994; Oard et al., 2001), but other studies have not confirmed this finding (Kelly & Belkin, 2001; Kellar et al., 2004). Furthermore, (Herder & Juvina, 2004) found that time spent on a web page was a good indicator of user lostness on web navigation tasks. While these results may seem to be contradictory, they likely demonstrate that the nature of the user task and environment, as well as the website information architecture and its content may have large effects on predictive models. These results show that the usefulness of one-measure-models is questionable. Establishing reliable and generalizable relationships between implicit measures and task outcomes holds the best promise for building predictive models that could be used in real-time, in a variety of contexts.

2.5 Getting Lost in Hypertext

Getting lost, or *disoriented*, is known to be one of the most important problems in hypertext navigation, yet there have been but a few attempts to assess and quantify lostness. Smith (1996) proposed an objective measure of lostness based the ratios of visited and optimal node counts as shown in equation (2) (section 3.1.1). Larson and Czerwinski (1998) compared user performance on information search in three different hypertext hierarchies: 8x8x8 (eight links at three hierarchy levels), 32x16 (thirty two links at the top level, and sixteen at the bottom level), and 16x32. Using Smith's measure Larson and Czerwinski showed that users were more lost on a hypertext with the 8x8x8 hierarchy than on either the 16x32 or the 32x16 hierarchies. Otter & Johnson (2000) described two measures designed to assess user lostness. The first of their measures combines previous work by (Smith, 1996) with the effects of different types of links. Their second measure is concerned with the accuracy of users' mental models of websites. The authors suggested that to capture lostness in hypertext, a battery of measures was needed, Herder's (Herder, 2003) work supported this viewpoint. (Ahuja & Webster, 2001) conducted an experiment demonstrating that user perceived disorientation in web navigation (assessed by a questionnaire developed by the authors) is only weakly related to user behavior, and that perceived disorientation is a better predictor of performance (time) than user behavior (such as the number of visited web pages, and page revisits). In a study that examined perceived user disorientation in hypermedia (Herder, 2003) found that perceived disorientation (measured using

the Ahuja & Webster's instrument) was correlated with a combined page return rate (average rate of revisits to pages which were visited at least twice) with median page view times, but not with the page revisitation ratio. Thus, in contrast to Ahuja & Webster, Herder's work demonstrated that user lostness was correlated with diverse measures of user behavior. These findings suggest that lostness is not a simple unidimensional construct. Since using questionnaire-based measures of lostness, as proposed by Ahuja & Webster (2001), is difficult, if not impossible, in most studies conducted in real-world contexts, the successful assessment of lostness based on observable real-time user behavior would be of great practical value. However, previous research has not yet provided the basis for deciding on whether lostness can be assessed in this way. To advance the field, the identification of measures that can predict lostness with accuracy over a wide variety of search tasks is desirable.

3. Research Objectives

The discovery of appropriate quantitative measures of navigational behavior is fundamental to advancing our understanding of the phenomenon of lostness in web navigation. While anecdotes and informal observations may be valuable and suggestive, we believe that an empirical comparison of quantitative measures of user behavior will provide an improved characterization of user navigation paths and that this may, in turn, inform the use of such measures in diagnosing user web navigation problems and in evaluating web sites. Previous research (Chi et. al., 2003; Berendt, 2002) has frequently focused on approaches and measures that are informed by the content of the web pages. Our study (along with, for example, McEneaney, 2001; Herder, 2003; Herder & Juvina, 2004) complements this work by investigating how navigational efficiency and success may be assessed by considering the clickstream alone, without reference to the meaning or content of the web pages and links.

Our first goal is thus to review and select appropriate structural and temporal measures that characterize user navigation in information-seeking tasks on websites. We present the measures that we have selected for review later in this section. Our second goal is to improve understanding of the commonalities and differences among these measures, and to determine whether we can identify navigation styles similar to those suggested by (Juvina & van Oostendorp, 2004; Herder & Juvina, 2004). Our third goal is to determine which measures are

the best predictors of (i) becoming lost on a website, and (ii) of task success (i.e. success in finding information). All measures are based on observable user behavior. In one of the studies reported in this paper, we used these measures to predict lostness, which was assessed in a post-task evaluation of the information-seeking session (details are given in the Methodology section). We attempted to determine whether lostness and task success can be predicted using measures such as:

- the time spent on the navigation task and the speed of clicking;
- the number of visited pages, the number of re-visited pages, and the ratio of revisited to visited pages;
- the "shape" of the web navigation graph;
- the similarity of the user navigation path to the optimal path.

3.1 The measures

3.1.1 Page-count measures

The logged time-stamped URLs were used to calculate: (i) the number of web pages visited in a session (N); the number of unique web pages visited (U); the number of web pages on the optimal path (O) (section 4.3.3); and (ii) the time spent on each web page (Time_per_page); and the total time on each question (Total_time). The measures were obtained directly from the recorded web session logs by a Python script running on a user computer. Two derived measures, "Revisits" (Tauscher & Greenberg, 1997) and "Objective lostness" (Lostness_obj) (Smith, 1996), were calculated using the ratio of visited and optimal node counts as shown below:

$$\text{Revisits} = 1 - U/N \,, \qquad (1)$$

$$\text{Lostness\_obj} = \sqrt{(U/N-1)^2 + (O/U-1)^2} \,, \qquad (2)$$

3.1.2 Global properties of the web navigation graph

If we consider the individual web pages visited by searcher to be the nodes of a graph and the links followed by the searcher to be the edges of the graph, we can compute global properties

such as stratum and compactness (Botafogo et al., 1992). Stratum and compactness were used to characterize searcher's behavior on similar web navigation tasks by McEneaney (2001), Shih et al. (2004), Herder (2003) and Herder & Juvina (2004) thus a comparison with our results is possible.

Compactness is a measure of the connectedness of a graph. It varies between zero and one; it is close to zero for sparsely linked graphs and close to one for highly connected graphs.

Stratum measures how close the navigation path is to a linear ordering. This statistic is based on the notion of the status, contra-status and prestige of the nodes in the path. Nodes that are hard to reach, but from which other nodes can be easily reached, have high status. Nodes that are easy to reach, but from which it is hard to reach other nodes, have high contra-status. A node's prestige is the difference between its status and contra-status. The stratum measure is defined as the sum of the absolute prestige of all nodes divided by the maximum possible value of prestige for a fully linear ordering. Like compactness, stratum varies between zero and one. A value close to zero indicates a less linear navigation path; a value close to one indicates a more nearly linear navigation path.

3.1.3 Similarity to the optimal path

These measures assessed the similarity between the user's path and the optimal path. The optimal path is the shortest path leading to the web page containing the sought-for information. For factual information tasks (section 4.1) the optimal path exists by definition. In our studies, we also assumed that the optimal path was unique.

Two similarity measures were calculated based on a well-known dynamic programming procedure by Needleman and Wunsch (1970). The method uses a global sequence alignment algorithm with a non-zero gap cost and an arbitrary distance function. The non-zero gap cost was used to apply a penalty for diversion in web navigation. Our use of the N-W algorithm assumed that (i) nodes were uniquely identified by webpage URLs composed of three parts: <host>,<path>,<query>, and (ii) the distance between two nodes was calculated based on similarity between their three-part URLs, where matching was done between each corresponding URL part (e.g., between <paths>).

Two measures—LCSMax and LCSlenMax—were derived using the N-W algorithm. LCSMax is a measure of similarity between the user path and the optimal path, normalized to the maximum possible score (an ideal match) for a path length equal to the user path length. LCSlenMax is the length of the longest common subsequence between the user path and the optimal path divided by the length of the user path. If no gaps in the matching paths were permitted, LCSlenMax would be equal to the number of visited pages on the optimal path. Since the N-W algorithm uses a non-zero gap cost, the common subsequence may contain pages which appear only on one of the paths, thus interpretation of the LCSlenMax measure is not as straightforward.

Table 1. Characteristics of the navigation measures used.

Measure Group	Symbol	Short Description	Description / Equation	Lower Bound	Upper Bound
Web page count metrics	O	Optimal path length	section 3.1.1	1	-
	U	Number of unique pages	section 3.1.1	0	-
	N	Number of total pages	section 3.1.1	1	-
Web page revisit metrics	Revisits	Ratio of revisited pages to all	section 3.1.1 / (1)	0	1
	Lostness_Obj	Objective Lostness (Smith)	section 3.1.1 / (2)	0	$+\infty$
Web graph metrics	Compactness	Graph connectedness	section 3.1.2	0	1
	Stratum	Graph linearity	section 3.1.2	0	1
Similarity to optimal navigation path	LCSMax		section 3.1.3	-1	1
	LCSlenMax		section 3.1.3	0	1

Our approach is to build predictive models using empirical data collected from web navigation sessions using large, complex, natural web sites. In the studies presented below, the user navigation tasks were specified before the search began.

4. Methodology

We conducted two question-driven, web-based, information-finding studies in a controlled experimental setting. Both studies used the same type of user task and the same apparatus. We first present the common elements shared by the two studies and then describe each study separately.

4.1 Navigation Task

A *factual information* task was used. According to Morrison et al. (Morrison et al., 2001) these types of tasks belong to the most common type on the Web and account for 25% of web search activity. A factual task is defined as an information finding task where the user seeks a specific piece of data (e.g., the name of a person or an organization, product information, a numerical value; a date; an address; etc.).

4.2 Procedures and Apparatus

Participants of the two studies performed question-driven, information-seeking tasks using two large Canadian government websites (a Government of Canada home page and a Health Canada home page). The websites were selected based on their complexity (over 10,000 web pages), and familiarity of their content to Canadian citizens and residents. The studies were conducted in a university laboratory using a PC running the Microsoft Windows 2000 operating system. Participants in each study were asked to perform a series of factual information finding tasks. Within each study, the tasks were the same for all participants and were presented in the same order. By keeping the order of tasks constant, any differential effects of learning website content were the same for all participants. At the beginning of each task, the participant started at the home page. Participants were instructed to find a single web page containing information that was specified by each task. The information sought was intended to be representative of the set of common possible information questions that citizens or residents of Canada might ordinarily ask when visiting these sites. While the formulation of the particular tasks was necessarily subjective, the list was arrived at after discussion among four members of the Engineering Psychology Lab at the University of Toronto, all of whom were either native Canadians or had lived in Toronto for many years. The tasks varied in difficulty, with the range of difficulty intended to be similar to that experienced by a typical user of the sites. Participants were asked to navigate in a single browser window, without using a search engine. The URLs of all visited web pages were logged to a local file, along with timestamps, and the screen coordinates of the link, or the button, or the graphic, that was clicked.

Here are two examples of the tasks (see Appendix A for the complete list):

- "Find a listing of addresses for passport offices in Ontario" (Government of Canada site)
- "Find a page that describes how to deal with stress for women" (Health Canada site)

4.3 Navigation Task Outcomes

Task success and user lostness were the two major outcomes (dependent variables) in both studies. Task success was defined as finding a web page with the information specified in a question. Lostness can be defined as an objective property (Smith, 1996) or, alternatively, it can be viewed as the subjective feeling of user disorientation on the web navigation task. We describe how these measures were operationalized in the descriptions of each study below.

5. Talk-Aloud Web Navigation Study (TA Study)

The talk-aloud method is a widely used method of studying cognitive processes, such as problem solving, learning, decision making, human-computer interaction, and cognitive task analysis. Participants carry out a task, while verbalizing their thoughts (Newell & Simon, 1972; Ericsson & Simon, 1980; Russo, Johnson, & Stephens, 1989). Most studies have found that the talk-aloud method does not alter task outcomes although it may increase the time spent on task (see Krahmer & Ummelen, 2004 for a review). We used the talk-aloud procedure to infer whether participants felt that they were becoming lost as they worked on the task.

Fourteen adults (six females and eight males; average age group 24-30) took part in 14 individual information-seeking sessions. Participants had an average of 9 years of Internet use experience. Their current average daily use of the Internet was ranged from 1 to 4 hours.

Each participant was asked to perform ten search tasks. Participants were asked to talk aloud while they were navigating the websites. Participants received the following instructions:

> *"While navigating the website please speak your thoughts out loud. This may feel a bit unnatural at first but please feel assured that we are not judging you but the usability of the website. Take your time and be thorough but still try to be efficient."*

There was no time limit on finding an answer to each of the ten questions. All sessions were recorded using the Camtasia screen cam software for capturing computer screens (along with

ambient sound). In addition, each participant's talk-aloud session was recorded on a tape recorder. Participants were paid $20 for their time.

5.1 Navigation Task Outcomes

5.1.1 Task success

Task success was scored true or false. Task success was also evaluated subjectively after the session. The participant provided a self-assessment of success and this judgment was verified by the experimenter, who checked the content of the final webpage visited by the participant. In cases of disagreement, if the participant declared success but the final page did not contain the required information, the task success score was adjusted by the experimenter to false.

5.1.2 Subjective evaluation of lostness

Lostness can be considered as an objective measure or as a subjective measure. The first approach was proposed by (Smith, 1996) who calculated lostness from observable user actions, such as the number of pages visited, the number of unique pages visited and the minimal (optimal) number of pages that need to be visited to complete the task. The second approach was advocated by (Ahuja et al., 2001) who measured perceived lostness by a post-task questionnaire. Similarly, Czerwinski et. al., (2001) demonstrated that subjectively estimated time on task (relative subjective duration, or RSD) is related to the user's success on the task; the time spent on failed tasks tended to be overestimated while the time spent on successful tasks tended to be underestimated. RSD is not a direct measure of lostness, but it would seem to be strongly correlated. Both of these subjective measures require the participants to make a judgment after the task has been completed.

We obtained a simple measure of subjective lostness based on the participant's verbal behavior throughout the session. Moreover, we did so without explicitly asking for subjective judgments of lostness. We used an independent rater to rate lostness in a post-task examination of the user's behavior. Participants occasionally expressed feelings of being lost (e.g. "I'm not in the right place", "I'm not sure what to do now"). Later, a trained human rater watched the audio-video record of information finding sessions and assessed, every 30 seconds, how lost the participant

appeared to be. Lostness was rated on a 4-point scale: 1-“Definitely Not Lost”, 2-“Probably Not Lost”, 3-“Probably Lost”, 4-“Definitely Lost”. Average values of subjective lostness were then calculated for each participant, for all questions. The reliability of the rater was verified by a second judge, who rated all tasks performed by a randomly selected study participant. Inter-rater reliability was checked by calculating intraclass correlation coefficient for the average values of subjective lostness. The obtained intraclass correlation coefficient was 0.94 and we concluded that ratings assigned by the principal rater were highly reliable. The average values of the subjective lostness are denoted by Lostness_R, which could range from 1 to 4..

Lostness_R = 1.3 Lostness_R = 2.8

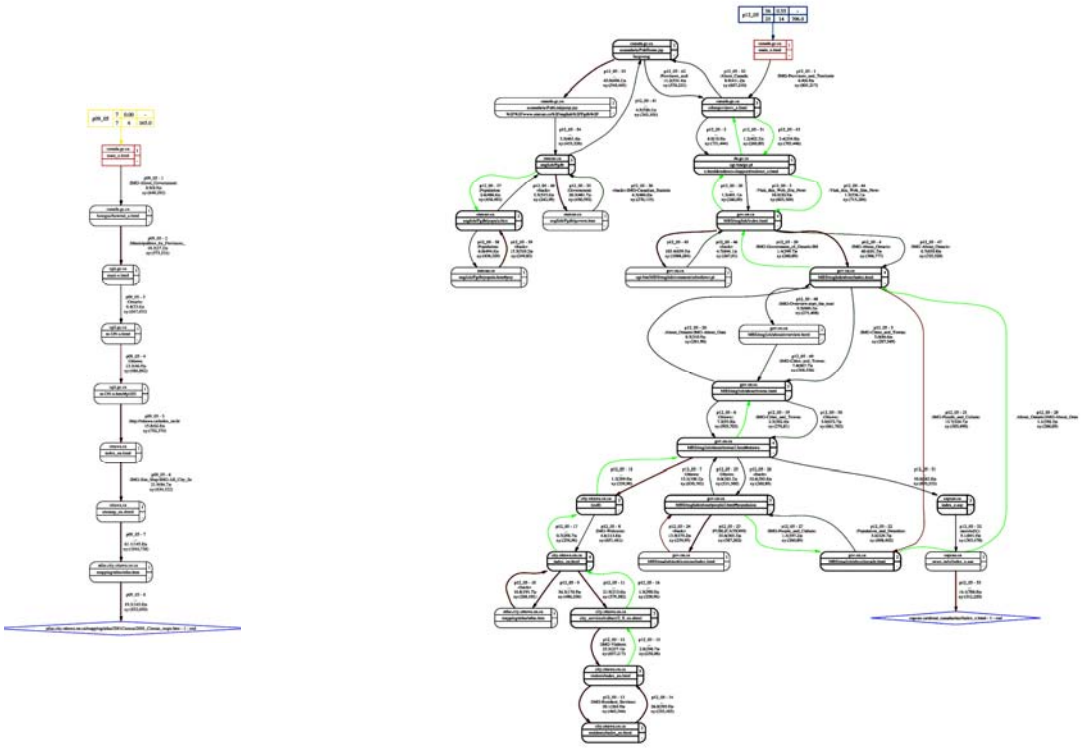


Figure 1. Differently shaped navigation paths with different subjective lostness ratings.

Figure 1 shows navigation graphs for two different participants on the same task. Nodes in these navigation graphs (rounded rectangles) represent visited web pages, while directed edges (lines ending with arrows) represent user traversal between web pages (by clicking on links or on the back button). The graphs were created by processing logged URL information using a version of Graphviz (North & Koutsofios, 1994; Gansner & North, 1999) with Pathalizer (Open Source, 2005) (the software was modified by us to meet our requirements). The visual representation

uses annotations and color-coding. The annotations provide temporal information and information about the link or button that has been clicked. Color-coding denotes speed of clicking (i.e. light green means quick clicking < 4s; darker green means between 4-8s per click, black means medium speed (8-13s per click); orange means slow clicking (13-20s); and red means a very slow clicking rate (>20s). The representation here is necessarily small and cannot show the level of detail in our originals; however, the reader should note that the shape of the navigation graph may differ considerably and that the shape seems to be related to feelings of lostness. In general, navigation paths that have a simple linear shape are associated with low values of our subjective lostness measure.

5.2 Results

Our overall goal was to build statistical models that would predict lostness and task success.

5.2.1 Prediction of User Lostness

Linear regression was used to help discover which measures best predicted subjective lostness (n=140). Due to exploratory nature of our study, we adopted a very conservative approach to regression model building and retained only variables that had $p < .001$. LCSMax and Total_time were found to the best predictors of subjective lostness and these two variables together accounted for over 51% of the total variance in the fitted regression model. The fitted model with standardized parameter estimates was:

$$Lostness\_R\_Predicted = -.46*LCSMax + 0.35*Total\_time, \quad (3)$$

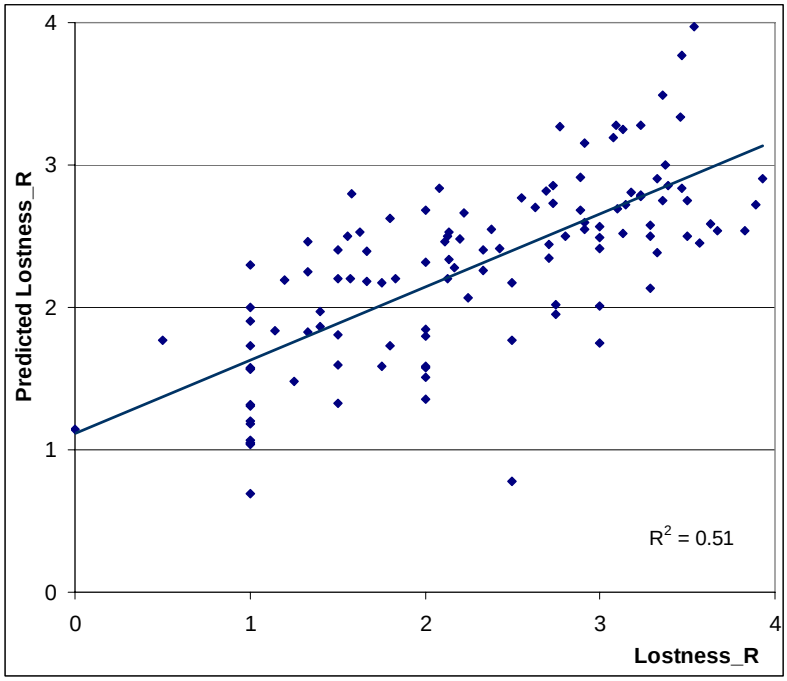


Figure 2. Predicted vs. actual Lostness_R for the TA study.

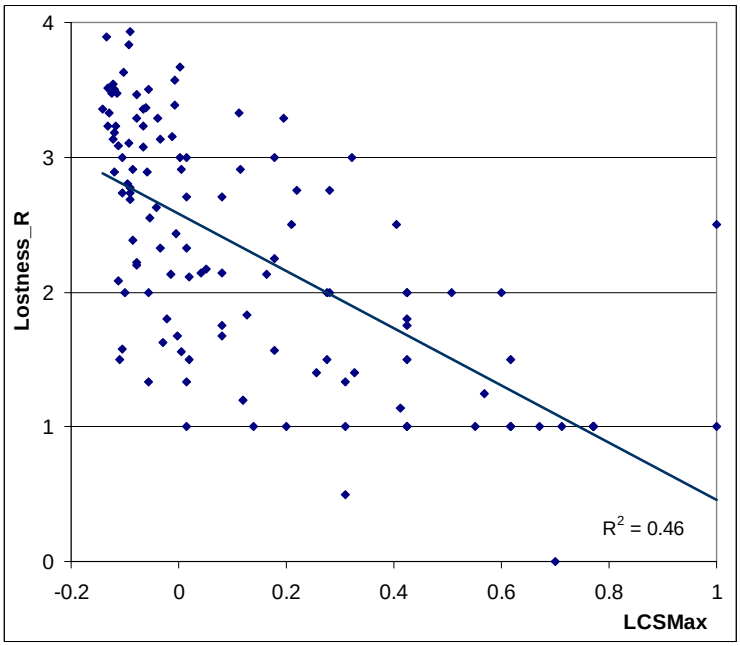


Figure 3. Relationship between subjective lostness (Lostness_R) and similarity to the optimal path (LCSlenMax) for the TA study.

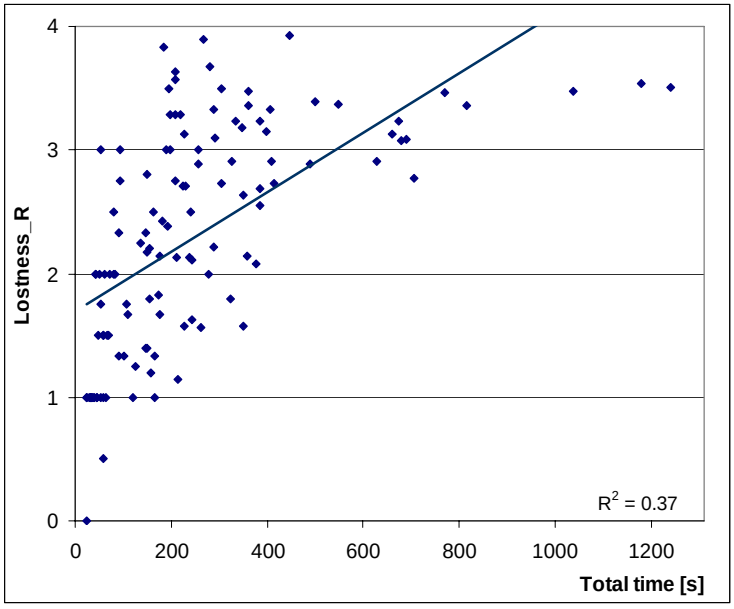


Figure 4. Relationship between subjective lostness (Lostness_R) and total time on task (Tot_time) for the TA study.

The further participants deviated from the optimal path, the more likely they were to be subjectively lost (Figure 3), and the more time participants spent on a task, the more likely they were to be subjectively lost (Figure 4).

It is worth noting that "objective" lostness, ratio of revisits and other variables were not retained in the "best" models. These variables, although occasionally significant at the .05 level when included in the regression models, contributed very little additional prediction variance.

5.2.2 Prediction of Task Success

Task success was defined as finding a web page with the information specified in a question and was rated on a binary scale (true/false). Consequently, we used logistic regression to find the best predictors of task success (n=140). The best regression model that explained the most variance included LCSMax as the only significant independent variable. The Wald chi-squared statistic associated with LCSMax was 10.3, p=.0013; 86.8% of predicted successes and failures agreed with the observed values. The R^2 of the model was .28. Thus, the similarity to the optimal path (LCSMax) was the best predictor of subjective task success.

6. Time-Limit Web Navigation Study (TL Study)

While the Talk-Aloud (TA) study was designed to shed light on the cognitive processes used by participants, we also wished to study the search behaviors of users who were not required to vocalize. Forty eight adults (29 females and 19 males; average age 20.5 years) took part in thirty eight individual information-seeking sessions. The participants used a computer for 21 hours per week on average, including 18 hours of the Internet use per week.

Each participant was asked to perform eight tasks. The tasks partially overlapped with those used in the TA study (see Appendix A). In contrast to the TA study, the time allowed for answering each question was limited to three minutes. If the requested information was not found in that time, the participant stopped and moved on to the next question. TL study participants were recruited from an undergraduate psychology class (PSY100) and received course credits for their time.

6.1 Navigation Task Outcomes

6.1.1 Task success

Task success was scored true or false. Task success was inferred from the time spent on each question. When participants spent more than three minutes and twenty seconds on a question, this was considered to be a failure (the time limit to answer each question was three minutes and we allowed a twenty second period of grace). The validity of this procedure was checked by examining a random sample (n=15) of the sessions that lasted less than three minutes and twenty seconds—in 95% of cases a page containing the appropriate information was the user's final selection.

6.2 Results

The data analysis was conducted with the two general objectives in mind: (i) to gain an improved understanding of the commonalities and differences among the various navigational measures; and (ii) to find predictors of lostness and task success.

6.2.1 Space of Web Measures – Second-Order Navigation Factors

To gain a better understanding of the structure underlying the space of the selected web measures, principal component analysis with varimax rotation was applied to the data from the TL study (n=384).

Table 2. Principal component loadings on the first three factors (after varimax rotation).

Variable	**Factor1**	**Factor2**	**Factor3**
Tot_time	**0.84**	0.40	0.16
U (unique pages)	**0.80**	0.03	-0.48
N (total pages)	**0.73**	0.38	-0.46
Lostness_Obj	**0.69**	0.41	-0.20
LCSMax	**-0.83**	-0.27	-0.01
LCSlenMax	**-0.85**	-0.33	0.11
Compact	0.20	**0.96**	-0.02
Revisits	0.45	**0.79**	-0.13
Stratum	-0.30	**-0.89**	0.12
Time_per_page	-0.08	-0.09	**0.96**

Figure 5 shows factor loadings of the navigation measures represented in the two-dimensional space defined by the first two (varimax rotated) factors, which we interpreted as follows:

Factor 1. *Navigational inefficiency*: characterized by a high number of visited pages, more time spend on the task, low task success, low similarity with optimal path, higher (objective) lostness. This factor explained 48% of the variance.

Factor 2. *Laborious navigation*: characterized by high compactness, high proportion of revisited web pages, and a low stratum (i.e., low linearity of the user path). This factor is very similar to a factor discussed by Juvina & Herder (2004). It also bears some similarity to a factor of the same name identified (Herder & Juvina, 2004). Laborious navigation explained 35% of the variance. All three components of this factor (compactness, stratum and ratio of revisits) are related to the shape of the navigation graph, that is, they are related to the user's navigation pattern.

Factor 3. *Navigation speed*. The remaining 17% of the variance was explained by the third factor, on which only one variable loaded highly (Time_per_page).

These three factors represent three aspects of user actions on the web navigation task: (1) total time and amount of clicking (total and "unnecessary" clicks), (2) user navigation patterns (e.g., forward paths, loops, rings, (Clark et al., 2006)), and (3) speed of clicking.

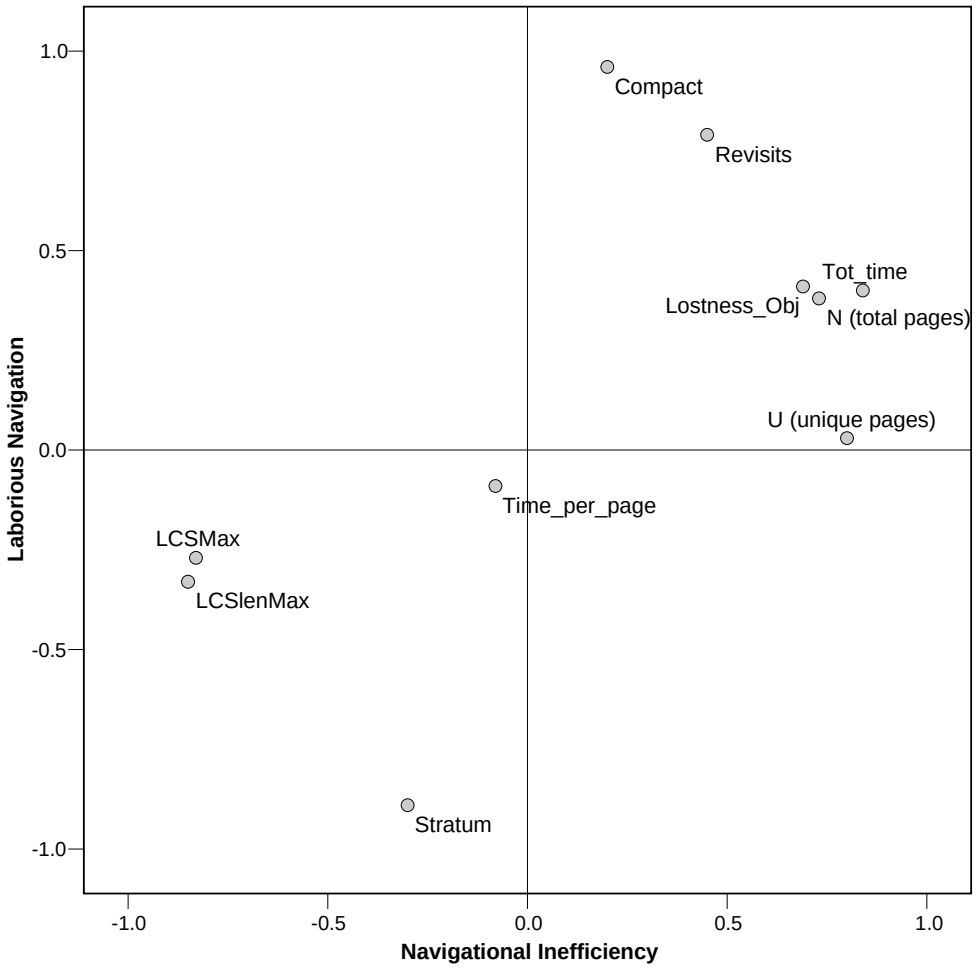


Figure 5. Web navigation measures represented in 2D space defined by factor loadings on the two extracted factors "Inefficiency" and "Laborious Navigation".

6.2.2 Prediction of Task Success

Task success was scored true or false, similarly as in the TA study. Thus logistic regression was used to find the best predictors of task success (n=384).

The Wald chi-squared statistic for the regression model (with the following predictors: LCSMax, Stratum, Compact) was 58.6, $p<.0001$; 81.3% of predicted successes and failures agreed with the observed values. The R^2 of the model was .30. LCSMax was the best predictor of task success,

the other two variables (Stratum and Compact) used in the model were not significant (p>.2). The Wald chi-squared statistic associated with LCSMax in this model was 32.6, p<.0001. Significant models were also obtained for LCSMax, stratum and compactness individually. The R^2 for these three models was .27, .17, and .15 respectively.

Thus, LCSMax was the best predictor of task success in both the TA study and the TL study. This result confirmed our a priori intuitions: the higher the similarity to an optimal path, the better the chances for success on an information finding task. Figure 6 presents the relationship between average level of task success and similarity to an optimal path calculated for each question in study TL. Figure 7 and Figure 8 present the relationship between average level of task success and, respectively, average level of stratum and compactness calculated for each question in study TL.

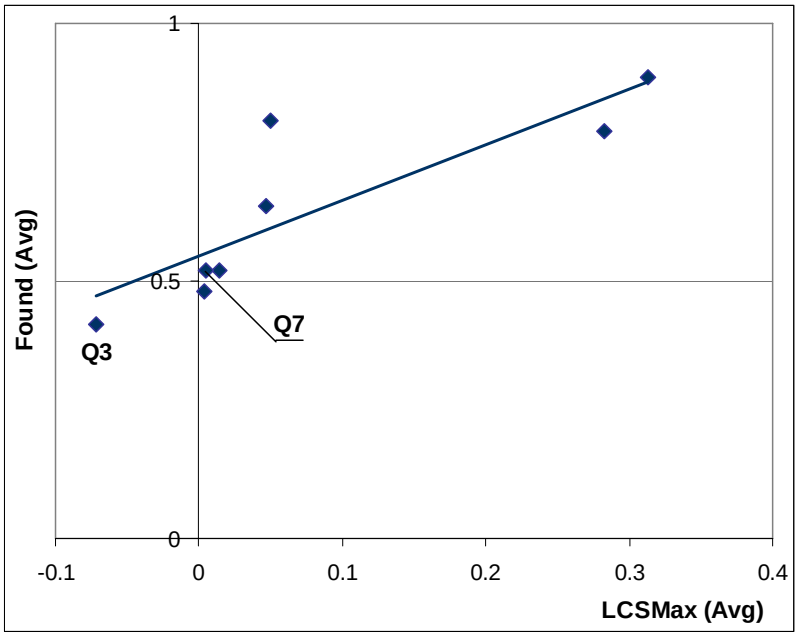


Figure 6. Relationship between task success (Found) and similarity to the optimal path (LCSMax) for average values calculated for each question in study TL. Points corresponding to questions Q3 and Q7 are marked.

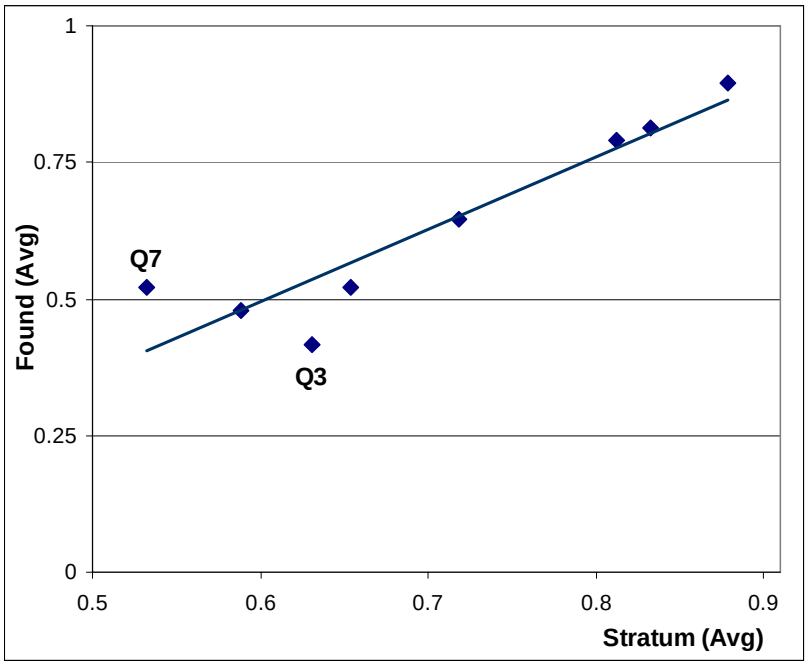


Figure 7. Relationship between task success and stratum (for average values calculated for each question in study TL).

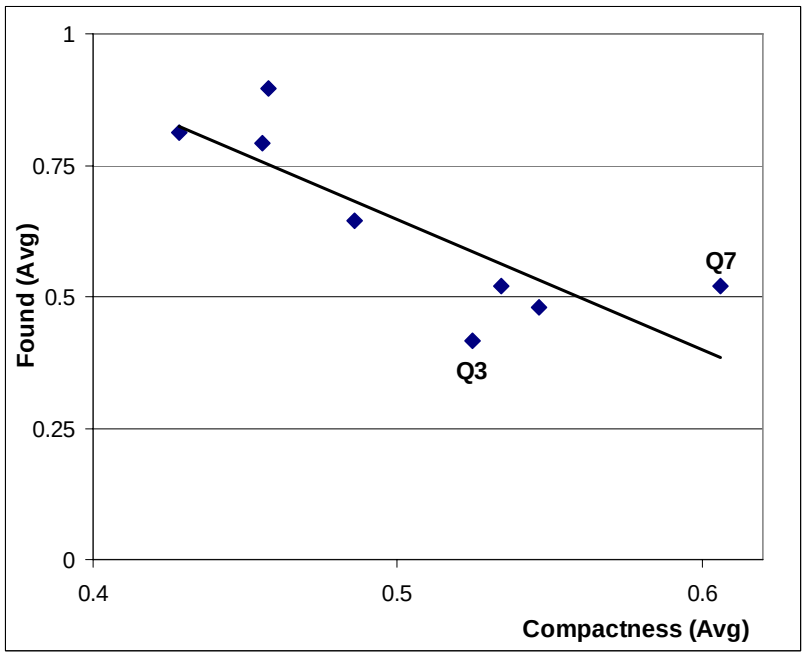


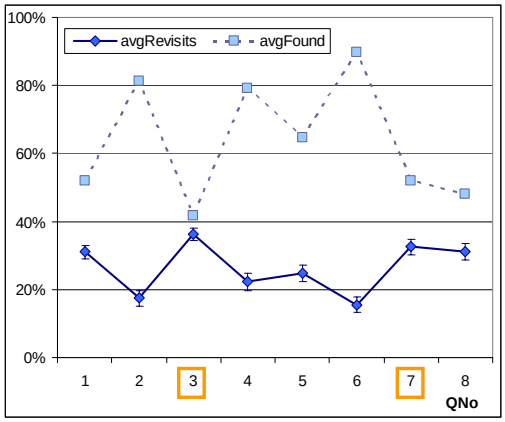
Figure 8. Relationship between task success and compactness (for average values calculated for each question in study TL).

6.2.3 Prediction of Task Success on Individual Tasks

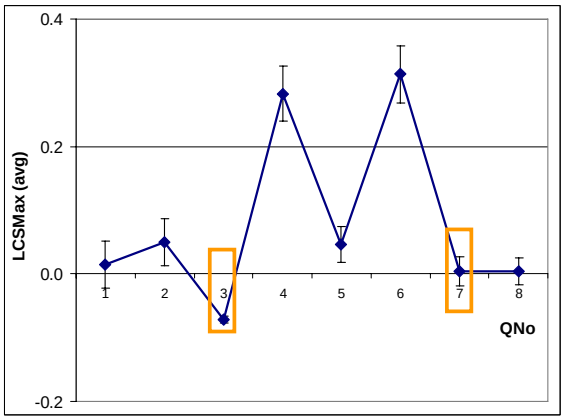
All information-seeking tasks used in this study (i.e. each of the eight questions) were of the fact-finding type. The tasks, however, differed in terms of how difficult it was to find the requested information. To assess the effect of the task, we examined whether LCSMax was also the best predictor of task success for each question.

Significant regression models (n=48) were obtained for five out of eight questions. Three of those models confirmed LCSMax to be the best task success predictor. In two of those five significant models, however, either compactness or stratum turned out to be slightly better predictors[2]. In the first case (question Q3), lower values of compactness (sparsely linked web usage graphs corresponding to fewer returns to previously visited pages) were related to higher task success. In the second case (question Q7), higher values of stratum (more linear user navigation path) were related to higher task success.

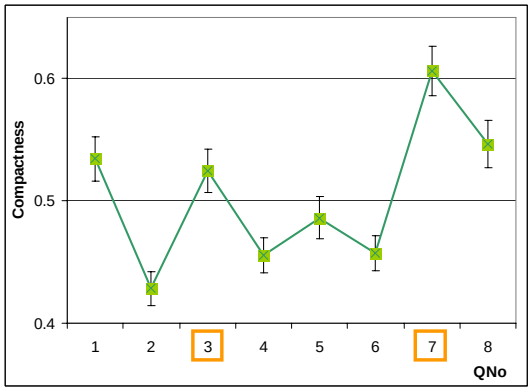
[2] LCSMax was still a significant predictor if fitted alone.



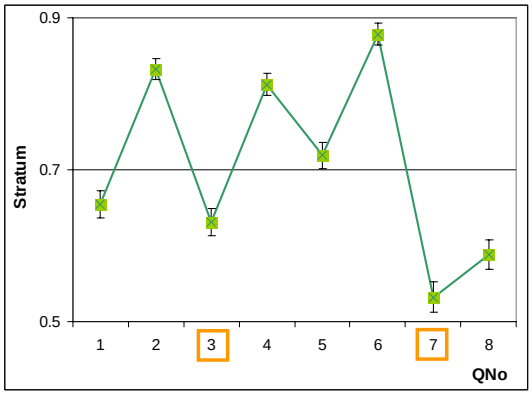
Avg. number of revisits and task success per question



Average similarity to the optimal path per question



Average compactness per question



Average stratum per question

Figure 9. Characteristics of question 3 and 7 (for study TL)

These two questions (Q3 and Q7) were characterized by some of the most extreme values of the web navigation measures – the highest revisit ratios, some of the lowest task success levels, some of the lowest similarity levels to an optimal path, the highest compactness, and the lowest stratum (Figure 9).

6.2.4 Prediction of Task Success by The Second Order Navigation Factors

The two factors (*Inefficiency* and *Laboriousness*) established by principal component analysis (section 6.2.1) were used in another logistic regression model. Both variables were found to be significant predictors of task success. The Wald chi-squared statistic of the regression model was 64.6, $p<.0001$; 89.4% of predicted successes and failures agreed with the observed values and the R^2 of the model was .43. The Wald chi-squared statistic for *Inefficiency* was 61.4, $p<.0001$,

while for *Laboriousness* the Wald chi-squared statistic was 50.1, p<.0001. Lower *Inefficiency* and lower *Laboriousness* predicted a greater chance of task success.

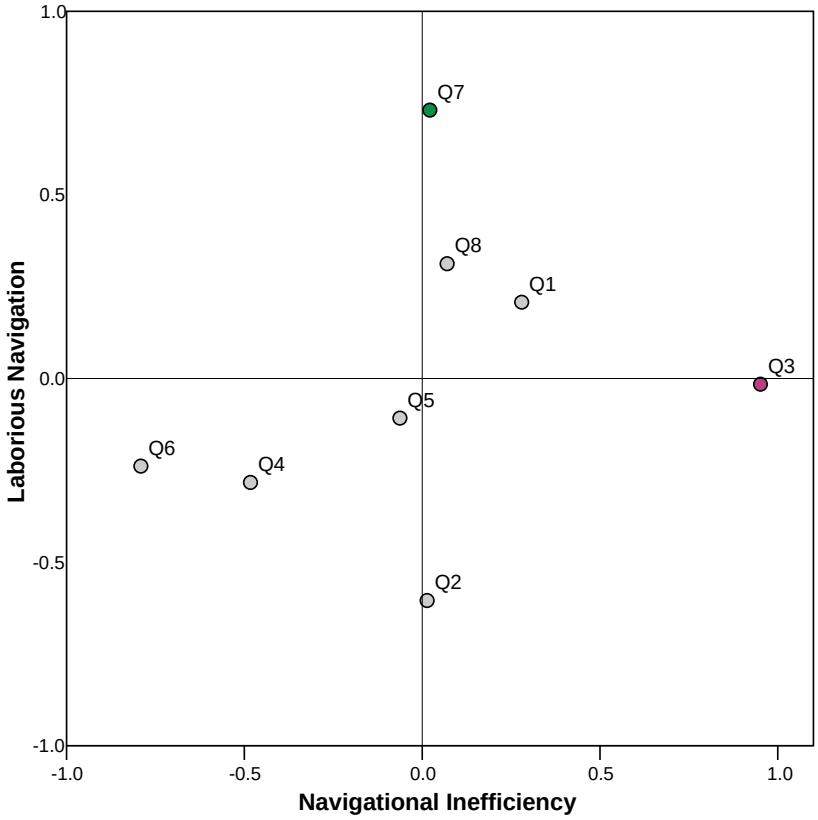


Figure 10. Eight questions from the TL study in 2D space defined by the two extracted factors *Inefficiency* and *Laborious Navigation.*

As can be seen in Figure 10, question Q3 loaded high on *Navigational Inefficiency*, while Q7 loaded high on the *Laborious Navigation* factor. Highly *Laborious* and *Inefficient* surfing paths are "far" from the optimal paths, hence similarity to the optimal path may not be differentiating among them very well. As our results indicate, these navigation paths are indeed better differentiated by the shape of the navigation graph (expressed by stratum and compactness).

Although LCSMax was, in general, the best single predictor of task success, on particular questions it was sometimes the case that measures of graph compactness or linearity proved to be better predictors.

7. DISCUSSION

We first reviewed a variety of structural and temporal measures that may be used to characterize user web navigation (section 3.1). We then examined the measures by applying them in the analysis of two empirical studies (sections 5 and 6). Our evaluation was made in the context of factual information-seeking tasks. We also identified aggregate measures that can be used to characterize user navigation styles and, by experiment, we appraised each of the measures as a predictor of lostness and task success.

7.1 Aggregate Web Navigation Measures.

We identified two navigation styles (Navigational inefficiency and Laborious navigation) that bear some similarities to those suggested in previous studies (Juvina & van Oostendorp, 2004; Herder & Juvina, 2004; Juvina & Herder 2005). The differences between our aggregate measures and those identified in the other studies, probably stem from the fact that those studies employed different sets of first-order measures.

We applied the two identified aggregate measures to the characterization of individual tasks and showed how combining these measures differentiates among the tasks. A combination of measures may more successfully characterize user navigational behavior than any single measure alone. Juvina & Herder (2005) used an aggregate measure (highly similar to our Laborious Navigation factor) to evaluate a new navigational mechanism (link suggestion). Their aggregate measure discriminated between the old and the new navigational mechanisms. Aggregate measures and their combinations have considerable potential in the diagnosis and evaluation of web navigation behavior.

7.2 Predictive Ability of Web Navigation Measures.

Since the findings of the Talk-Aloud (TA) and Time-Limited (TL) studies are similar—and compatible—this discussion draws on the results of both studies.

We first consider the predictive capabilities of stratum and compactness (S&C) and we compare our results with previous studies that also used S&C. We then consider measures of similarity to the optimal path and their predictive capabilities. We conclude with a discussion of lostness.

Our results indicate that lower values of compactness and higher values of stratum tend to be associated with a higher probability of task success. This relation is opposite to the one shown by (McEneaney, 2001). However, there is no necessary contradiction; McEneaney used a different navigational task (learning from a hypertext handbook vs. factual information finding). It seems that the navigational strategies which are successful may be quite different in the two situations. Shih and his colleagues (Shih et al., 2004) found that S&C differentiated between expert and novice navigation paths, and that, for experts, S&C differentiated also among the navigation task phases (exploration, resolution, completion). Results from our study show that, when compared with other measures (e.g., LCSMax), S&C are good predictors of task success for tasks on which user behavior tends to be inefficient and more laborious. Since inefficiency and laboriousness are associated with more difficult tasks, the predictive power of S&C seems to depend on task difficulty. The findings of the three relevant studies are summarized in Table 3.

Table 3. The predictive ability of stratum and compactness in different contexts.

task attribute	context	description	study
task success and task type	n/a	higher stratum was associated with success in other studies; lower stratum was associated with success in our study;	learning task–(McEneaney, 2001; Shih et al., 2004); factual information finding–this study
task phases	expert users	stratum and compactness differed among three task phases	task phases: exploration, resolution, and completion–(Shih et al., 2004)
task success	difficult tasks	lower stratum (or higher compactness) predicted task success	this study

Direct comparisons may be problematic, since each study used different combinations of measures to predict task success or lostness. However, the results are sufficient to suggest a number of plausible reasons for the observed differences among the studies.

1. The effects of different tasks may affect the sensitivity of predictive models; prediction may only be useful, if specific contextual factors are known and their relationships understood. The relationships between stratum/compactness and task outcomes are complex and are likely mediated by contextual factors that vary with the task.

2. Variation in the potency of the same predictors, depending on the particular task (question), further supports the conclusion that the success of a search strategy is dependent on the nature of the information-seeking task. In particular, while they are generally good

predictors, the stratum and compactness measures may not be as effective in predicting task success with easier navigational tasks.

We showed that similarity to the optimal path is a good predictor of both lostness and task success for information-seeking tasks. For other tasks the results may differ. Depending on the task, the notion of a optimal path may not make sense, or an optimal path may not exist and therefore the similarity measures would be ill-defined. Success on other tasks may be better predicted by the shape of the exploration path, using measures like stratum and compactness.

Since we used existing complex websites, we did not control for the differences in website hierarchies, and thus we cannot compare our results with the work of Larson and Czerwinski (1998).

Like Herder (2003), we found that lostness can be predicted by observing user actions. However, the most effective predictors of lostness in our studies (similarity to the optimal path and total time on task), are different from those found by Herder. Also, it is important to be aware of slight differences among the definitions of lostness and how the measures of lostness were operationalized in the two studies.

As Otter and Johnson (2000) have argued, lostness is a complex phenomenon and a diverse set of quantitative measures is likely needed to characterize lostness in different circumstances and on different tasks. Since several studies have demonstrated the merits of different measures in different circumstances, it is critically important to consider the user's goals and the nature of the task required to achieve these goals.

8. CONCLUSIONS & FUTURE WORK

Previous work (Herder, 2003; McEneaney, 2001; Otter & Johnson, 2000; Shih et al., 2004) has shown that the notion of lostness is useful in predicting success in information-seeking tasks. Furthermore, these studies showed that a variety of easily computed measures could be useful in characterizing and predicting lostness, and that lostness, in turn, is strongly associated with task success. While we strongly endorse this approach to a better understanding of web navigation

behavior, we believe that more empirical work is required to refine and select the best measures for a variety of search tasks.

Appropriate measures can provide useful characterizations of user web navigation behavior and can help to diagnose a variety of problems (such as getting lost) that users encounter when navigating hypertext documents. Such measures can also help to identify the local web structures that are conducive to successful navigation. Thus the basic goal in our research is to establish measures that provide an objective basis for diagnosing and evaluating the information architecture and information design of websites.

Our results showed that three first-order measures (similarity to the optimal path, navigation graph linearity and compactness) and two second-order measures can be useful diagnostics of user web navigation behavior. Further research is needed to determine whether user lostness and success can be identified on tasks other than those used in our studies.

We have not considered individual differences (such as level of web familiarity, domain knowledge, gender, verbal ability, spatial ability). While it is reasonable to expect that individual differences would play a role in the development of feelings of lostness, this aspect was beyond the scope of the present investigation. Future studies should examine such effects.

One of our next goals is to investigate the possibility of real-time automatic detection of when users are becoming lost. The ability to predict lostness and task success would be extremely useful in real-time. The prediction of could be based on behavior of one user or on aggregate behavior of many visitors to a website. An effective diagnostic capability could be used to help to build adaptive web structures. For example, user-help could be created dynamically based on a real-time recognition of increasing lostness.

APPENDIX A. Web Navigation Tasks from Study TL and TA

TA	TL	Task Goal
x		Find a listing of addresses for passport offices in Ontario.
x		What is the history of the West Nile virus?
x	x	Find a listing of documents on the topic of Dealing With Abuse.
x		Why are foods irradiated? Find the page that describes this process.
x	x (Q3)	Find a short description of Ottawa that lists population and area covered, among other information.
x		Find the page that graphs energy consumption in Canada.
x		Find a listing of "Travel Health Advisories" listed by date.
x		Find a page that describes how to deal with stress for women.
x	x	Find a brief (two sentences) listing of Canadian health expenditures for 2000-2001.
x	x	Find the page that discusses "Maternity and Newborn Care". This page includes a chapter listings for a book.
	x	Find the page that defines and describes Smog.
	x	Find the official web-page for Saskatchewan that lists that province's population.
	x	Find the page that describes precautions for "Minimizing your risk" of contracting Hepatitis C.
	x (Q7)	Find the page that lists the key health care issues.

References

Ahuja, J. S. & Webster, J. (2001). Perceived disorientation: an examination of a new measure to assess web design effectiveness. Interacting with computers, 14, 15-29.

Berendt, B. (2002). Using Site Semantics to Analyze, Visualize, and Support Navigation. Data Min. Knowl. Discov. 6(1): 37-59.

Boechler, P. M. (2001). How spatial is hyperspace? Interacting with hypertext documents: cognitive processes and concepts. CyberPsychology and Behavior, 4, 23-46.

Botafogo, R. A., Rivlin, E., & Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. ACM Transactions on Information Systems, 10, 142-180.

Chen, S. Y. & Macredie, R. D. (2002). Cognitive style and hypermedia navigation: development of a learning model. Journal of the American Society for Information Science and Technology, 53, 3-15.

Chi, E. H., Cousins, S., Rosien, A., Supattanasiri, G., Williams, A., Royer, C. et al. (2003). The Bloodhound project: automating discovery of Web usability issues using the InfoScent™ simulator. In Proceedings of the ACM Conference on Human Factors in Computing Systems CHI'2003.

Clark, L., Ting, I., Kimble, C., Wright, P. & Kudenko, D. (2006) "Combining ethnographic and clickstream data to identify user Web browsing strategies" Information Research, 11 (2) paper 249 [Available at http://InformationR.net/ir/11-2/paper249.html]

Conklin, J. (1987). Hypertext: An introduction and survey. Computer, 20, 17-41.

Czerwinski, M., Horvitz, E. and Cutrell, E. (2001). Subjective Duration Assessment: An Implicit Probe for Software Usability. Proceedings of IHM-HCI 2001, Lille, France, September, 2001, pp. 167-170.

Dieberger, A. (1995). Providing spatial navigation for the world wide web. Spatial Information Theory. In Spatial Information Theory - Proceedings of COSIT'95 (pp. 93-106). Semmering, Austria: Springer.

Dieberger, A. (1997). A city metaphor to support navigation in complex information spaces. In Spatial Information Theory - Proceedings of COSIT'95 (pp. 53-67). Springer.

Ericsson, K. & Simon, H. (1980). Verbal Reports as Data. Psychological Review. 87, 215-251.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. ACM Transactions on Information Systems, 23, 147-168.

Gansner, E. R. & North, S. (1999). An open graph visualization system and its applications to software engineering. Software: Practice and Experience, 30, 1203-1233.

Herder, E. (2003). Revisitation Patterns and Disorientation. In Proceedings of the German Workshop on Adaptivity and User Modeling in Interactive Systems ABIS 2003 (pp. 291-294).

Herder, E. & Juvina, I. (2004). Discovery of Individual Navigation Styles. In Proceedings of Workshop on Individual Differences in Adaptive Hypermedia at Adaptive Hypermedia 2004 (AH2004).

Juvina, I. & van Oostendorp, H. (2004). Individual Differences and Behavioral Aspects Involved in Modeling Web Navigation. LECTURE NOTES IN COMPUTER SCIENCE, 3196, 77-95.

Juvina, I. & Herder, E., (2005). The Impact of Link Suggestions on User Navigation and User Perception. UM2005 User Modeling: Proceedings of the Tenth International Conference.

Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Modeling information content using observable behavior. In Proceedings of the ASIST Annual Meeting.

Kelly, D. & Belkin, N. J. (2001). Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback. SIGIR Forum (ACM Special Interest Group on Information Retrieval), 408-409.

Kelly, D. & Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. SIGIR Forum, 37, 18-28.

Kim, H. & Hirtle, S. C. (1995). Spatial metaphors and disorientation in hypertext browsing. Behaviour & Information Technology, 14, 239-250.

Krahmer, E.& Ummelen, N. (2004). Thinking about Thinking Aloud: A Comparison of Two Verbal Protocols for Usability Testing, IEEE Transactions on Professional Communication, 47, 105-117.

Larson, K. & Czerwinski, M. (1998). Web page design: Implications of memory, structure and scent for information retrieval. In Proceedings Form CHI 98. 25-32.

Masahiro, M. & Yoichi, S. (1994). Information filtering based on user behavior analysis and best match text retrieval. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 272-281). Dublin, Ireland: Springer-Verlag New York, Inc.

McEneaney, J. E. (2001). Graphic and numerical methods to assess navigation in hypertext. International Journal of Human Computer Studies, 55, 761-786.

Morrison, J., Pirolli, P., & Card, S. K. (2001). A taxonomic analysis of what world wide web activities significantly impact people's decisions and actions. In Proceedings of CHI' 2001. Extended abstracts (pp. 161-162).

Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology, 48, 443-453.

Newell, A. & Simon, H. (1972). Human Problem Solving, Englewood Cliffs, NJ: Prentice Hall,

North, S. C. & Koutsofios, E. (1994). Applications of graph visualization. In (pp. 235-246).

Oard, D. & Kim, J. (2001). Modeling information content using observable behavior. In Proceedings of the ASIST Annual Meeting (pp. 481-488).

Open Source (2005). Pathalizer [Computer software].

Otter, M. & Johnson, H. (2000). Lost in hyperspace: metrics and mental models. Interacting with computers, 13, 1-40.

Pitkow, J. E. & Pirolli, P. (1999). Mining Longest Repeating Subsequences to Predict World Wide Web Surfing. In USENIX Symposium on Internet Technologies and Systems The USENIX Association.

Russo, J.E., Johnson, E.J. & Stephens, D.L. (1989). The Validity of Verbal Methods. Memory and Cognition, 17, 759-769

Shih, P.-C., Mate, R., Sanchez, F., & Munoz, D. (2004). Quantifying user-navigation patterns: a methodology proposal. In Poster presented at the 28th International Congress of Psychology in Bejing Bejing 2004.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. Interacting with computers, 8, 365-381.

Spink, A. & Losee, R. M. (1996). Feedback in Information Retrieval. Annual Review of Information Science and Technology, 31, 33-78.

Tauscher, L. & Greenberg, S. (1997). How people revisit web pages: Empirical findings and implications for the design of history systems. International Journal of Human Computer Studies, 47, 97-137.

Wang, W. & Zaïane, O. R. (2002). Clustering Web Sessions by Sequence Alignment. In Proceedings of DEXA Workshops 2002 (pp. 394-398).