Cognitive Walkthrough for the Web

Marilyn Hughes Blackmon[†], Peter G. Polson[†], Muneo Kitajima[‡], Clayton Lewis[†]

[†]Institute of Cognitive Science
University of Colorado at Boulder
Boulder, Colorado 80309-0344 USA
+1 303 492 5063
{blackmon, ppolson}@psych.colorado.edu,

[‡]National Institute of
Advanced Industrial Science and Technology (AIST)
1-1-1, Higashi, Tsukuba, Ibaraki 305-8566 Japan
+81 298 61 6650
kitajima@ni.aist.go.jp, Clayton.Lewis@colorado.edu

ABSTRACT

This paper proposes a transformation of the Cognitive Walkthrough (CW), a theory-based usability inspection method that has proven useful in designing applications that support use by exploration. The new Cognitive Walkthrough for the Web (CWW) is superior for evaluating how well websites support users' navigation and information search tasks. The CWW uses Latent Semantic Analysis to objectively estimate the degree of semantic similarity (information scent) between representative user goal statements (100-200 words) and heading/link texts on each web page. Using an actual website, the paper shows how the CWW identifies three types of problems in web page designs. Three experiments test CWW predictions of users' success rates in accomplishing goals, verifying the value of CWW for identifying these usability problems.

Keywords

Cognitive Walkthrough, web navigation, web usability, cognitive model, information scent, headings, link labels

INTRODUCTION

The goal of this paper is to present a new Cognitive Walkthrough for the Web (CWW) for use in the design and usability evaluation of websites. The CWW, a transformation of the original Cognitive Walkthrough (CW) for applications [19,28,32], is a theoretically-based usability inspection method [24] that can be applied in all stages of the design and development process.

The new CWW, like the CW [15,28], simulates users performing navigation tasks on a website by assuming that users perform goal-driven exploration. However, the CWW is adapted to be a better fit to a realistic website design process by considering the following three features specific to website navigation and design. First, the CWW uses contextually rich descriptions of user goals (100-200 words long) incorporating more information about users' understanding of their tasks and underlying motivation. Second, the CWW assumes that generating an action (e.g., clicking on a link, button, or other widget) is a two step process. The first step involves parsing a new page into subregions and attending to the correct subregion of the page. The second step is selecting a widget from the attended to subregion and acting on it. Third, the CWW evaluation process is organized differently and fits better into website development. A user of CWW works on one web page at a time in relation to a whole set of representative user goals. The CWW evaluation process can start with a detailed description of the home page and a rough outline of its immediate successor pages. The CWW is then applied repeatedly to incrementally design and evaluate each successor page down through the hierarchy.

CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA.

The CW identifies usability problems by simulating step-by-step user behavior for a given task using a prototype interface, and by having the design team answer the following questions at each simulated step: *Q1) Will the correct action be made sufficiently evident to the user? Q2) Will the user connect the correct action's description with what he or she is trying to do? Q3) Will the user interpret the system's response to the chosen action correctly?* The CW is widely used in application development and has been recommended as a useful method for website evaluation (e.g., [9]).

The CWW embodies the same original set of evaluation questions, Q1-3. However, the most critical questions for successful navigation would be these: *Q2a) Will the user connect the correct subregion of the page with the goal using heading information and her understanding of the sites page layout conventions? and Q2b) Will the user connect the goal with the correct widget in the attended to subregion of the page using link labels and other kinds of descriptive information?* The CWW provides the design team with theory-based suggestions concerning the likelihood of users' next heading/link selections.

THEORETICAL FOUNDATION OF THE CWW

"There is nothing so practical as a good theory" [21]

CoLiDeS Simulation Model of Website Navigation

Like the CW [15,28], the CWW is derived from a theory of the cognitive processes that control goal driven exploration, but the model underlying CWW is CoLiDeS [16]. CoLiDeS, an acronym for Comprehension-based Linked model of Deliberate Search, extends a series of earlier models [15] of performing by exploration based on Kintsch's [14] construction-integration theory of text comprehension and

problem solving processes. CoLiDeS is part of a broad consensus among theorists and website usability experts [5,10,16,18,26] that problem solving processes, guided by users' goals and *information scent*, drive users' information-seeking or search behaviors when exploring a new website or carrying out a novel task on a familiar website.

The core process underlying web navigation, in the theory, is comprehension of texts and images. Comprehension processes build, elaborate and compare the mental representations of screen objects on a web page in preparation for selecting and clicking one particular hyperlink or image. Users act on the hyperlink, image, or other screen object that they perceive as being *most semantically similar* to the description of their current goal.

CoLiDeS assumes that selecting a next action is a two-phase process. During the *attention phase*, the user segments/parses the page into a collection of subregions and generates a brief description of each subregion from his/her knowledge of page layout conventions, and titles and headings on the page. The user attends to the subregion whose description is perceived as being most similar to the user's current goal. During the *action selection phase*, the user generates descriptions of all of the widgets in the subregion and acts on the one whose description is most similar to their goal. A CoLiDeS demonstration is available at psych.colorado.edu/~kitajima/ CoLiDeS_Demo.html.

Latent Semantic Analysis, Hill Climbing, Information Scent

A unique feature of CoLiDeS is the use of Latent Semantic Analysis (LSA) [17]. LSA is a mathematical technique that estimates the semantic relatedness of texts, based on a statistical analysis of a large corpus. We use LSA to estimate the semantic similarity between statements of user goals and descriptions of subregions of a page, and between goals and descriptions of possible actions on a web page, e.g., hypertext links, buttons, etc. These estimates enable us to project the likelihood that a user with a given goal will select one or another target for action on a given page.

The use of LSA also enables us to generalize the representation of users' goals. In all previous simulation models of exploratory behavior (e.g., [15,29]), the descriptions of users' goals are brief and very specific. An example would be "Search for information on heart disease." But many authors have made the point that users may not have these explicit goals in mind when they search the Web or other information collections. LSA enables us to represent users' goals as narratives that describe general concerns and motivations, or as specific goals where appropriate, or as specific goals elaborated with background information reflecting the users' understanding of the concepts referred to in the user's goal statement.

Related Models/Studies of User Behavior on the Web

Cognitive Models of User Search Behavior

Information foraging theory [26] incorporates measures of semantic similarity [25] in a model of user search behavior that is closely related to CoLiDeS's use of LSA. Both models take actions that are perceived as being close to a user's description of her goal. However, information foraging theory has a much broader scope, attempting to characterize users' cost/benefit perceptions in making decisions, like terminating search of one website and searching for another site that contains more information relevant to their goals. The critical point is that the two frameworks are complementary, sharing a common model of the search process even though the common models are derived from very different cognitive architectures (ACT-R [2] versus Kintsch's construction-integration model of text comprehension [14]).

Approaches to Understanding User Behavior on the Web

Many research groups are studying user behavior on the Web, since insights are valuable for designing useful websites as well as efficient web servers. One approach focuses on the global behavior of web users. Byrne et al. [4] identified user interaction patterns by analyzing verbal protocols collected during browsing sessions. Tauscher [30] addressed the same issues with usage data. Huberman et al. [12] derived distributions of numbers of user clicks in a site by applying statistical analysis to user log data. Pitkow and Pirolli [27] predicted web pages that users are likely to request by applying data mining technique to user log data. Chi et al. [6] applied techniques used in information retrieval research to estimate the likelihood of selecting each link in a given website for given information needs.

Except for Chi et al. [6], the common characteristic of most of these studies was that user behavior was aggregated over the different user goals. These studies used click stream data to uncover properties of typical sequences of page accesses. In most cases, investigators had no information about the content of users' goals. Thus, these studies did not attempt to show how the content of users' goals controls navigation behavior, despite widespread agreement that goals control search behavior.

Theory-based Design Evaluation Tools

There have been attempts to develop theory-based design evaluation tools and automated tools for the evaluation of websites. Examples of efforts to commercialize automated tools for the evaluation of websites include Accrue Insight [1] and WebCriteria SiteProfile [31]. However, both Nielsen [23] and Chi et al. [5] have been critical of these efforts because they do not capture how the actual meaning of labels and content in a site relate to users' goals.

Chi et al. [5,6] present work that attacks this problem. In [6], Web User Flow by Information Scent (WUFIS), like CoLiDeS, estimates similarity between the information needs or goals and links on the current pages to predict the next pages that users are likely to visit. The similarity measure uses co-occurrence of keywords in goals and link labels, weighted by the discriminative power of different keywords, as reflected in term-document co-occurrence. An evaluation using simulated searches of 19 different websites showed that WUFIS simulations predicted search paths leading to reasonably relevant pages on the sites.

WUFIS's [6] goal-centered analysis is consistent with the consensus in the literature that users' goals are a fundamental determinant of their search behavior (e.g., [7,22]. Byrne et al. [4] and Morrison et al. [20], and others have proposed taxonomies of user goals based on survey data and laboratory studies. CWW shares this foundation.

CWW is intended to extend the WUFIS [6] work in three respects. First, LSA is a more completely developed theory of the acquisition and representation of knowledge [17] and text comprehension [14], and thus can be used to measure the semantic similarity of descriptions of any length. This allows CWW to be applied to richer statements of user goals, longer descriptions of widgets and subregions.

Second, LSA provides a principled representation of the semantic knowledge of a given user population, defined as a semantic space, and estimates of that population's perceptions of semantic similarity. The space is not based on the vocabulary used in a website, but rather on the words contained in texts assumed to have been encountered by members of a given user population.

Third, CWW is based on theory that attempts to account for attention processes in searching the web. The CoLiDeS theory underlying CWW assumes that a user selectively attends to one subregion of a page whose description is most similar to her goal.

APPLYING THE CWW TO A WEBSITE

CWW works by analyzing likely user behavior when the user is pursuing a given goal, and is looking at a particular web page. The technique can be used to critique a design page by page as a site is being designed, or to evaluate the pages in an already completed design, or in an implemented site. We will illustrate the method by applying it to the home page of an actual site, www.AmerLandscape.com. This section sketches a brief overview of the most unique feature of CWW: its ability to identify problems with heading/link texts. A detailed tutorial of the CWW with more examples is available at http://psych.colorado.edu/~blackmon/CWW.html. The tutorial includes images of the AmerLandscape.com web pages that are discussed below.

Step 1: Compiling a Set of Realistic User Goals and Intended Selections

The first task is to compile a diverse, representative set of detailed user goal statements, each 100-200 words long. Here is one of several goals used to evaluate the AmerLandscape.com website:

> Our office park covers over a hundred acres with large areas of lawn and many office buildings with foundation landscape designs of flowering shrubs, evergreens, and some areas of flowers. We also have three garden patio areas for employees to use for eating lunch or snacks and outdoor recreation. Maintaining attractive landscaping is a high priority for most of the companies who rent or own offices here, because many of their clients meet with them on the premises. We want to outsource all the care of our landscape with reliable, well-qualified professionals. This includes frequently mowing the lawns, planting annual flowers, expert pruning of shrubs and trees, raking up the leaves in the fall, and proper fertilizing of the lawn, shrubs, trees, and flowers. It also requires expert diagnosis and treatment of disease and insect problems [136 words].

Cooper [8] has suggested constructing various personas and designing the site to fit the goals of these personas, and Cooper's approach meshes well with the process of compiling the set of user goals for the CWW.

For each goal, the analyst must choose the semantic space that offers the best fit to the presumed reading level of the persona whom the goal represents. For example, the LSA website offers five semantic spaces for American-educated speakers of English: average reading comprehension and background knowledge levels for first-year college (the default space) and for grades 3, 6, 9, and 12.

The analyst must also identify the intended correct selection on the web page for each goal. CWW will identify a problem whenever there is a good likelihood that the user will choose something other than this intended selection. The purpose of Steps 2-4 is to validate that a user with a specific goal will select the correct link or act on the correct widget on the page currently under analysis.

Step 2: Using LSA to Estimate Semantic Similarity of Goals, Headings, and Link Labels

2a) The analyst submits each detailed user goal statement, each heading on the page, and each link label on the page to Latent Semantic Analysis (LSA) for a document-to-document, One-to-Many Comparison on the LSA website (http://lsa.colorado.edu/). LSA returns a table of cosines, which are estimates of the semantic similarity between the goal and each of these elements of the page. The analyst then ranks both the goal-heading and the goal-link cosines from highest to lowest. The CWW predicts that the user will select the headings/links with the top-ranking cosines.

2b) A term-to-term, One-to-Many Comparison is then done on the LSA site to produce term vector lengths for one- and two-word heading/link labels (or for the two most important words selected from heading/link labels with more than two words). Term vector lengths are estimates of how much knowledge about an element is embedded in the designated LSA semantic space (e.g., first-year college).

2c) The analyst also performs Matrix Comparisons, on the LSA site, for the set of all headings and the set of all link labels. This produces cosines for each pair, estimating how similar in meaning one heading or link label is to another.

Step 3: Identifying Problematic Heading/Link Labels

According to CoLiDeS, a user first focuses on a subregion of a page based on its heading (if the page is subdivided). A heading will cause trouble either because it is *unfamili*ar (the user does not know what the heading means), or because it is easily *confusable* with another heading. A heading with an LSA term vector length of less than .8 (for two most meaningful words) is likely to be unfamiliar (use the result of 2b for this decision). Any heading pair yielding a cosine of .6 or more in the LSA analysis is likely to be confusable (use

the result of 2c). Link labels liable to be unfamiliar, or a pair of labels under the same heading liable to be confusable, are identified in the same way.

The first AmerLandscape.com home page had six headings:

1) Masters in Landscape Management/Maintenance,
2) Masters in Landscape Installation,
3) Masters in Tree Services,
4) Masters in Snow and Ice Services,
5) Customers We Serve, and
6) Member of Better Business Bureau.

Of these, Landscape Installation has an LSA term length of only .53, and so is flagged as *unfamiliar*. There is also a pair of *confusable* headings: Landscape Management/ Maintenance and Tree Services have a cosine of 0.65.

Step 4: Finding Goal-Specific Problems

Unfamiliar or confusable headings or labels are bad whatever the user's goal might be. But some problems emerge only for some goals. For example, two headings may not be very similar to one another, but may both be equally similar to a possible goal.

The cosines from LSA's One-to-Many analysis, created by step 2a, can be compared to identify these problems. If the similarity of a heading to the goal is equal to or greater than the similarity of the correct heading to the goal, the analyst marks the intruder as a *goal-specific competing heading* unless the analyst judges the similarity a false alarm – a heading not likely to attract users' attention for accomplishing that goal.

On the AmerLandscape.com site one goal that fared badly was the example goal quoted above, the office park seeking full-service landscape management. Its strongest similarity was to Landscape Installation (cosine of 0.61), a wrong heading, and the cosine to the Landscape Management, the correct heading, was weaker (0.52).

The CWW standard for a *goal-specific competing link* has three criteria: (1) the competing link label must be under the same heading as the correct link, (2) the competing link label must have a cosine indicating at least 80% of the similarity to the goal that the correct link label has, and (3) not be judged by the analyst as a *false alarm*, a link that real users would probably not select.

This third criterion compensates for a bias in LSA's similarity estimates with respect to actual user judgment. LSA is more likely to overestimate than underestimate the similarity of items, in that a human judge may recognize that phrases that use similar terms are in fact sharply different in meaning. Allowing the analyst to reduce LSA's similarity estimate, and reject a proposed competing link label, is an approximate response to this problem.

On the AmerLandscape page, a goal centered on repairing damage to trees caused by a severe storm illustrates goal-specific competing links. The correct heading is Tree Services, and LSA predicted no competing headings. Under the correct heading, the link label that the developer intended to be correct was *Storm damage specialists* (0.24 cosine), but five competing links all had higher cosines (ranging from 0.81 to 0.27). The analyst might reject *Cabling trees* (0.6) as a *false alarm* (goal did not mention cabling), but the other four competing links identified by LSA were reasonable choices for the storm damage goal, e.g., *Removing trees and dead tree limbs* (0.81).

Method Notes

The analyst must use discretion in deciding what portion of each link label text to submit to LSA for estimating familiarity. For any heading or link label text that uses only one or two words, the analyst should submit the text to LSA as is. But if the link label is longer, the analyst must trim it to one or two words, to avoid false indications from LSA term vector lengths. Term vector lengths increase with the number of words, so longer labels are almost certain to pass the .8 threshold for familiarity, no matter what.

On the other hand, discarding words decreases LSA's ability to assess accurately the meaning of the label. To compensate for this, the analyst should retain the two most meaningful words, the two words that best distinguish that label from other labels on the same page. For example, "snow ice" captures the distinctive features of Snow and Ice Services, but the word "services" belongs as much to Tree Services as it does to Snow and Ice Services.

When analyzing headings or link labels for semantic similarity, the full texts should be used. If headings are very short, the texts submitted to LSA for a heading should include the link labels grouped under the heading, as a way of increasing the content available to LSA. Doing this may distort the meanings of some headings, but less than the distortion involved in applying LSA to very short texts.

On To Other Pages

When one page has been worked over, analysis can proceed to subordinate pages. But here work can be saved, because only a subset of the original sample of goals need be considered for each subordinate page: the subset of goals routed to that specific page from other pages. Thus the CWW speeds up as the analyst goes deeper into the site.

Repairs

The point of finding problems is to fix them, and the CWW output provides quite specific guidance in doing this (see http://psych.colorado.edu/~blackmon/CWW.html). For example, the problems of confusable headings on the AmerLandscape page were solved by rewording the link labels nested under the headings (recall that the link labels under a heading are taken to influence the interpretation of the heading.) Rewording magnified the distinguishing features of each heading and reduced overlap among the words nested under the three problematic headings. The CWW analysis of the revised page found no cosine greater than 0.39 for any heading pair.

Table 1. Evaluation of the CWW triage of web page problems using data from three separate experiments

Type of usability problem identified by CWW	Mean first-click success rate and number of affected goals (in parentheses)		
	Experiment 1	Experiment 2	Experiment 3
Unfamiliar link labels	24% (33)	60% (12)	28% (6)
Confusable link labels	53% (45)	75% (15)	55% (1)
Goal-specific competing headings for goals not affected by *unfamiliar* or *confusable* link labels.	No headings on these web pages	No headings on these web pages	40% (6)
Goal-specific competing links for goals not affected by *unfamiliar* or *confusable* link labels.	45% (17)	67% (7)	44% (6)
Summary of all goals with ANY of the above problems	**41% (95)**	**68% (34)**	**38% (19)**
Goals with NO problems	**70% (82)**	**89% (29)**	**62% (13)**

Experiment 1: 177 goals, each attempted by 1-6 experimental participants; data from Larson & Czerwinski [18].
Experiment 2: 63 goals, each attempted by 22-23 experimental participants; this web-based experiment is available at http://psych.colorado.edu/~blackmon/Expt000403Subsite/Expt000403Subsite.html
Experiment 3: 32 goals, each attempted by 20 experimental participants; this web-based experiment is available at http://psych.colorado.edu/~blackmon/PaigeHome.html (select LLLList option).

Additional Steps in a Complete CWW

There are plenty of ways a website can fail besides the navigation problems CWW identifies. The design may use widgets that the user has not seen before and cannot operate, or the cues used to subdivide the page may not be recognized. Nielsen [22] argues forcefully for abiding by standard interface conventions for hyperlinks. The CWW analyst should check these matters with the background of intended users in mind, in the manner of the original CW.

Another problem area is backing out of a failed search. The CWW aims for a site in which users will usually be guided forward successfully, but failures will occur. Instone [13] has developed a Navigation Stress Test that isolates each web page and asks a series of questions that reveal how easy it might be for a typical user to figure out how to move from that page to other locations within the same site. Instone's analysis would be a useful adjunct to the CWW.

EMPIRICAL TEST OF CWW PREDICTIONS

The question is, *How well do CWW's predictions of problems or lack of problems stack up against user behavior?* We present data from three experiments, all designed originally for other purposes, that allow us to test the accuracy of CWW for identifying problems with headings/link texts. The experimental task was simulated search of an online encyclopedia to find an article on a specified topic. On each trial, the participant was presented with a web page containing the target topic and a collection of category links. Clicking on a category link led to a page with a list of target articles. We used CWW to divide the trials into four basic types: unfamiliar heading/link labels, confusable heading/link labels, goal-specific problems, and no problems. Below, we report the percent correct data for these first clicks on the category labels.

Since the experimental participants in all three experiments were college-educated, all LSA comparisons and estimates of term vector length were made using the semantic space for first-year college.

Larson & Czerwinski [18] Data on Two Web Pages

Kevin Larson and Mary Czerwinski generously shared with us the data from an experiment they designed for a related but distinct question about breadth/depth factors that influence information scent perceived by users [18]. Our reanalysis applied the CWW to two web pages from their experiment, which we will call the E16 and E32 pages. E16 presents a single content area with no subregions or headings, and contains a randomly arranged list of 16 links. E32 is very similar to E16 except that it contains 32 links.

The CWW analysis for the E16 and E32 pages was the same as described for the AmerLandscape.com website, except that there were no headings to analyze on the pages, only link labels. The CWW process identified many goals affected by each of the three types of usability problems: *unfamiliar*, *confusable*, and *goal-specific competing* link labels. *Paleontology*, *Anthropology*, and *Theology and Practices* were predicted to be unfamiliar. Examples of confusable pairs of links include *Music* with *Musicians and Composers*, and *Theology and Practices* with *Religions and Religious Groups*. The CWW also found a subset of goals with no usability problems.

For each subset of goals the reported success rate is the mean percentage of experimental participants who selected the correct link on the first click. The "Experiment 1" column of Table 1 displays the performance data for goals affected by each type of usability problem. It then contrasts the 41% success rate for the 95 goals with ANY usability problems identified by the CWW, with the 70% success rate for the 82 goals that had NO identified usability problems. A two-way

ANOVA of their results (mean success rates for the 177 goals) finds a main effect for *unfamiliar* link texts, F (1, 174) = 24.698, p < .0001, and for *goal-specific competing* link texts, F (1, 174) = 5.027, p = .026, and no interaction between the two independent variables.

Replication Using Longer Goal Statements

A replication experiment was run (click link Expt000403 at http://psych.colorado/edu/~blackmon) in which participants were given longer goal statements. Larson and Czerwinski [18] asked participants to find items described by such unelaborated terms as *Pink Floyd* or *Tlingit*. In contrast, the replication also described the term. For example, for *Tlingit* experimental participants saw the following description:

> Tlingit, group of Native American tribes of the Northwest Pacific Coast culture area and the Pacific coast of southeastern Alaska. The economy of the Tlingit is based mainly on fishing, and they are especially noted for their skill in woodcarving. In both appearance and social customs, they closely resemble the neighboring Haida. Today, the largest concentration of Tlingit is in Alaska, where many Tlingit work in the logging and fishing industries.

Experiments 1 and 2 used the same web pages (E16 and E32), but the mean success rates in Experiment 2 are consistently about 23 percentage points higher than the corresponding figures for Experiment 1. We can attribute these differences to the presence/absence of a goal description. Experiment 1 participants who were unfamiliar with a term probably resorted to trial-and-error search. Most real-world users know something about what they are searching for, so the Experiment 2 data are more realistic.

The "Experiment 2" column of Table 2 shows that the mean first-click success rate was 89% for the 29 goals with NO problems but only 68% for the 34 goals for which the CWW identified usability problems. A two-way factorial ANOVA for mean success rates of 63 goals showed a main effect for *unfamiliar* links, F (1, 59) = 17.447, p <.0001, a main effect for *goal-specific competing* links, F (1, 59) = 5.058, p = .028, and no significant interaction.

A Web Page with Subregions Labeled by Headings

For the third experiment we used a website Gamble [11] designed for an experiment on the effects of adding additional text to two-word link labels. We applied the CWW to a page (LLLList) with 51 links distributed over eight subregions, each with a two-word heading label, and then tested the predictions against data from 20 experimental participants (10 each in two differently ordered sequences). The pattern of results (see Table 1) substantiates the conclusions drawn from the other two experiments. The first-click success rates were 38% for goals with one or more problems identified by CWW, contrasted with 62% for goals with no problems.

This experiment also provides valuable data on the effect of headings in directing – or misdirecting – the user's attention to a particular subregion of the web page. The page had a very conventional layout of subregions, an eight-cell rectangular matrix, four cells wide by two cells high. The attention-grabbing headings were two-word link labels set off by larger, boldface font in a contrasting color. Nested under each subregion were five to eight link labels, each consisting of a two-word bold title followed by a list of examples of items that could be found under that label on the linked-to web page. Link label texts averaged 10 words in length and ranged from 7 to 15 words.

To examine the effects of headings in directing attention, we computed statistics on the percentage of total links that experimental participants clicked under the correct heading for each goal, competing headings for each goal, and other headings. The data support the claim of CoLiDeS that users quickly narrow attention and ignore links nested under unattended subregions. For goals with no *goal-specific competing* heading, 91% of the clicks fell under the correct heading, compared to 68% for goals with one or more *goal-specific competing* headings, F (1, 638) = 95.574, p<.0001. Total clicks under correct or competing headings did not differ significantly, totaling 91% and 94%, respectively. Thus, almost all clicks occur under correct or competing headings, the particular subregions that the CWW predicted would be the most likely foci of users' attention.

In addition, participants clicked more links on the web page for goals with competing headings (mean 3.67) than for goals with no competing headings (mean 2.15), and the difference was significant, F (1, 638) = 46.92, p <.0001.

Due to the large number of independent variables affecting performance on the LLLList web page, we performed a multiple regression analysis. Three independent variables explained 15% of the variance. Participants' mean clicks accounted for 1.000 clicks (p <.0001), competing headings added 1.417 clicks (p <.0001), and unfamiliar link label texts added .817 clicks (p = .0019). Confusable and goal-specific competing links and goal sequence were not significant and were dropped from the regression analysis.

Experiments 2 and 3 used nearly identical procedures but different web pages, and the success rates for Experiment 3 are consistently about 29 points lower. The data suggest that the difference may be attributable in part to the number of links per page. In the ideal situation the mean links clicked per web page would be 1.0 (first link clicked would always be the correct one). In reality, the mean links clicked was 1.38 for the E16 web page (16 links), 1.77 for the E32 web page (32 links), and 2.67 for the Experiment 3 web page (51 links). Performance deviates farther from the ideal as the number of links per page increases, and headings that misdirect attention cause serious problems.

DISCUSSION

As Table 1 shows, the CWW can identify characteristics of web pages that differentially affect user performance. While overall performance levels vary from example to example, differences are consistently shown that are consonant with the CWW analysis. Further, when CWW flags a problem, it provides a specific diagnosis that can guide the repair or mitigation of the problems it identifies.

The CWW, the underlying theoretical model (CoLiDeS), and the empirical data presented here, call attention to three sets of interrelated attributes of a site and their associated design problems for the website developers. The first is the knowledge needed to interpret the vocabulary used in heading and link labels. The second is potential problems with subregion headings and link labels that are meaningful to users but may still pose difficult decision problems. The third is the repertoire of conventions used in a site to mark subdivisions of pages into subregions, and to represent page elements, such as links, on which users must act.

Insufficient Information Scent and Understanding of Headings and Labels

A given LSA semantic space is an approximate model of a specified user population's text comprehension abilities [17]. Small values of LSA term vector lengths predict that a word or phrase will have little meaning for users modeled by a given semantic space, and hence that these items will offer insufficient scent to those users. Our data clearly indicate the impact of this problem, with link labels flagged as unfamiliar by CWW consistently yielding the lowest user performance. Such problems are also prominent in the analysis of medical websites, where non-medically trained users have a great deal of trouble finding relevant information and comprehending descriptions of medical conditions and treatment recommendations [3]. Note that developers have trouble detecting such problems with link and heading vocabulary, because they nearly always have considerable knowledge of the content of the website.

Confusable and Goal-Specific Competing Labels

Models like CoLiDeS or WUFIS, whose navigation behaviors are controlled by a process analogous to scent following, predict that users will have trouble with confusable link and heading labels. When subregion headings or link labels within a subregion are semantically very similar to each other, problems occur because a goal that is semantically similar to one will probably be semantically similar to the other as well. Goal-specific competing headings and links cause similar problems. Here an incorrect heading or link (even though it is not highly similar to the correct heading or link) will emit a high scent for a specific user goal, confusing the user with multiple high-scent headings or links. Examination of Table 1 shows that both kinds of problems, as flagged by CWW, lead to increases in navigation errors. Preliminary evidence from the third experiment suggests that goals associated with the highest mean-click rates have garden-path headings that actively mislead the user to focus first/most on the wrong subregion. We plan follow-up experiments to test this finding on how headings direct/misdirect users' attention.

Website Conventions

Recall that CWW assumes that selecting a link is a two-stage process: 1) attending to a subregion of a page, and 2) selecting a widget to act on in the attend-to subregion. CWW uses the labels on subregion headings to predict what region users will attend to, and the data from Experiment 3 suggest that these headings do play a key role in navigation. However, how headings direct attention, and hence exactly when or whether a potentially confusable link label comes into play, is shaped by the conventions used to demarcate subregions, whether these involve color, font, boundary markers, or geometry. Developers either have to speculate about users' understanding of these conventions or test representative members of the population of intended users.

One of the major usability problems for users of the web is that any object on a page may be a target for action (e.g., hyperlink text, buttons, graphics, etc.). The evaluation question Q1 in the original CW, shared by CWW, asks if users are able to identify objects on a page as targets for action and to construct a description of them that can be compared to their goals. This is another place where design conventions, and deviations from them, need to be critically appraised, as Nielsen [22] has argued. Novel graphics, objects with short and potentially ambiguous labels, and the like, will cause problems.

CONCLUSIONS

The CWW overcomes a serious limitation of the original Cognitive Walkthrough: instead of relying on subjective estimates of semantic similarity, the CWW uses estimates of semantic similarity from Latent Semantic Analysis. Subjective estimates of semantic similarity are problematic, especially when designing sites for users from underserved populations. Developers have good intuitions about individuals like themselves, but these intuitions tail off as differences in educational level, computer experience, or specialized knowledge separate developer from user.

In principle, LSA allows meaningfulness to be assessed for different user populations, by using semantic spaces derived from different corpora. As mentioned earlier, spaces are available that approximate reading experiences at different grade levels, and these have been used successfully to match content presentations to different audiences [33]. We have not tested the use of these spaces in CWW, nor have we assessed the feasibility of constructing semantic spaces for user populations with special characteristics, like particular technical knowledge or distinctive cultural perspectives. But we think the potential is there to build new semantic spaces in LSA in any language/culture (LSA already has some French language semantic spaces) and to represent diverse users at any level of reading comprehension ability and background knowledge on either side of the digital divide.

ACKNOWLEDGMENTS

We thank Tom Landauer, Walter Kintsch, and Darrell Laham for helpful discussions of LSA and problems of assessing semantic relatedness, and Kevin Larson, Mary Czerwinski, and Paige Gamble for use of stimuli and data.

REFERENCES

1. Accrue Insight. http://www.accrue.com.
2. Anderson, J. R., & Lebiere, C. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, 1998.

3. Berland, G.K. at al. Health information on the Internet: Accessibility, quality, and readability in English and Spanish. *Journal of the American Medical Association*, 285, 2612-2621, 2001
4. Byrne, M. D., John, B. E., Wehrle, N. S., & Crow, D. C. The tangled Web we wove: A taskonomy of WWW use. In *Proceedings of CH'99,* ACM Press, 544-551, 1999.
5. Chi, E., Pirolli, P., & Pitkow, J. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of CHI 2000,* 161-168, ACM Press, 2000.
6. Chi, E., Pirolli, P., Chen, K., & Pitkow, J. Using information scent to model user information needs and actions and the Web. In *Proceedings of CHI 2000*, ACM Press, 490-497, 2001.
7. Cooper, M. Evaluating accessibility and usability of web pages. In *Proceedings of 2nd Int. Conference on Computer-Aided Design of User Interfaces*, Kluwer Academics, 33-42, 1999.
8. Cooper, A. *The Inmates are Running the Asylum: Why High Tech Products Drive Us Crazy and How To Restore the Sanity*. Macmillan Publishing Co., 2000.
9. E-marketing.Com: http://www.e-marketing.com.au/ documents/useabilityintro.htm
10. Furnas, G. W. Effective view navigation. In *Proceedings of CHI'97*, ACM Press, 367-374, 1997.
11. Gamble, P. E. *Does Link Label Length and Content Affect Web User Performance?* Unpublished Senior Honor's Thesis, Department of Psychology, University of Colorado at Boulder, 2001.
12. Huberman, B. A., Pirolli, P., Pitkow, J., & Lukose, R. Strong regularities in World Wide Web surfing. *Science*, **280**, 95-97, 1998.
13. Instone, K. STRESS-TEST your favorite, or least favorite website. *Michigan Ohio Computer-Human Interaction Meetings (March 14, 2001)*; find Navigation Stress Test also at http://keith.instone.org/navstress/.
14. Kintsch, W. *Comprehension: A Paradigm for Cognition*, Cambridge University Press, 1998.
15. Kitajima, M. & Polson, P. G. A comprehension-based model of exploration, *Human-Computer Interaction*, **12**, 345-389, 1997.
16. Kitajima, M., Blackmon, M. H., & Polson, P. G. A Comprehension-based model of Web navigation and its application to Web usability analysis. In *People and Computers XIV*, Springer, 357-373, 2000.
17. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211-240, 1997.
18. Larson, K., & Czerwinski, M. Web page design: Implications of memory, structure and scent for information retrieval. In *Proceedings of CHI'98*, ACM Press, 25-32, 1998.
19. Lewis, C., Polson, P., Wharton, C., & Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of CHI'90*, ACM Press, 235-241, 1990.
20. Morrison, J. B., Pirolli, P., & Card, S. K. A taxonomic analysis of what World Wide Web activates significantly impact people's decisions and actions. *UIR Technical Report UIR-R-2000-17*, Xerox PARC, 2000.
21. Newell, A. & Card, S. K. The prospects for psychological science in human-computer interaction. *Human-Computer Interaction*, **1**, 209-242, 1985.
22. Nielsen, J. *Designing Web Usability,* Indianapolis: New Riders Publishing, 2000.
23. Nielsen, J. http://www.useit.com/alertbox/991212.html, 1999.
24. Nielsen, J. & Mack, R. L. *Usability Inspection Methods*, New York: John Wiley & Sons, Inc., 1994.
25. Pirolli, P. Computational models of information scent-following in a very large browsable text collection. In *Proceedings of CHI'97*, ACM Press, 3-10, 1997.
26. Pirolli, P. & Card, S. Information foraging. *Psychological Review*, **106**, 643-675, 1999.
27. Pitkow, J. & Pirolli, P. Mining longest repeated subsequences to predict World Wide Web surfing. In *Proceedings of the USENIX Conference on Internet*, 1999.
28. Polson, P. G., Lewis, C., Rieman, J., & Wharton, C. Cognitive walkthroughs: A method for theory-based evaluation of user interfaces. *International Journal of Man-Machine Studies*, **36**, 741-773, 1992.
29. Rieman, J., Young, R. M., & Howes, A. A dual-space model of iteratively deepening exploratory learning. *International Journal of Human-Computer Studies*, **44**, 743-775, 1996.
30. Tauscher, L. & Greenberg, S. How people revisit web pages: Empirical findings and implication for the design of history systems. *International Journal of Human-Computer Studies*, **47**, 97-137, 1997.
31. WebCriteria SiteProfile. http://www.webcriteria.com.
32. Wharton, C., Rieman, J., Lewis, C., & Polson, P. The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods*, New York: John Wiley & Sons Inc., 105-140, 1994.
33. Wolfe, M. B., Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Learning from text: Matching readers and text by Latent Semantic Analysis. *Discourse Processes,* **25**, 309-336.