PAPER *Special Section on Internet Technology V*

Evaluation of Website Usability Using Markov Chains and Latent Semantic Analysis

Muneo KITAJIMA[†], **Noriyuki KARIYA**[††], ***Nonmembers***, **Hideaki TAKAGI**[††], ***Member***, ***and*** **Yongbing ZHANG**[††], ***Nonmember***

SUMMARY

The development of information/communication technology has made it possible to access substantial amounts of data and retrieve information. However, it is often difficult to locate the desired information, and it becomes necessary to spend considerable time determining how to access specific available data. This paper describes a method to quantitatively evaluate the usability of large-scale information-oriented websites and the effects of improvements made to the site design. This is achieved by utilizing the Cognitive Walkthrough for the Web and website modeling using Markov chains. We further demonstrate that we can greatly improve usability through simple modification of the link structure by applying our approach to an actual informational database website with over 40,000 records.

key words: *Latent Semantic Analysis, Website usability evaluation, Markov chains, large-scale information-oriented websites, Cognitive Walkthrough for the Web*

1. Introduction

The development of information and communication technology enables us to access substantial amounts of information on demand, regardless of location or time, and locate specific information. The total amount of available information increases daily, but it is not necessarily organized. People attempting to retrieve targeted information using the Internet tend to use search engines or links within the website.

Survey results indicate that Internet users fail to locate information on a specific site in over 50% of attempts, although it is known to exist somewhere in the site[1]. The desired information is found at a success rate of only 30% using on-site search engines; the success rate is 53% if the information is tracked without using the on-site search engine[2]. These results indicate problems with website usability. The cause lies in the gap between the location and access route to the desired information, and the expectations of the visitors. These flaws in website design cause visitors to either fail at finding the desired information or to spend excessive time locating it.

This problem is not limited to time loss for people searching for information. Websites lose prospective customers because they cannot provide the desired information. This problem with website usability seriously affects not only the visitors, but also the information providers. Websites must be designed so that target information can be obtained quickly and accurately to prevent such problems. In this paper, we describe approaches to usability problems for large-scale information provider websites and provide a method to quantitatively evaluate the improved design of an website that is ideal for both the information searcher and provider.

This paper includes the following sections. The Cognitive Walkthrough for the Web (CWW), which is a usability evaluation method for websites, and the basic model underlying it are briefly explained in Section 2. CWW is applied to a large-scale information provider website in Section 3 to identify usability problems. An actual website with over 40,000 records is used as an example. We describe a model of visitor's link selection process using Markov chains in Section 4 and demonstrate improved usability through a simple web design change. Finally, we present our conclusion in Section 5.

2. Cognitive Walkthrough for the Web (CWW)

Here we briefly explain CWW, the Cognitive Walkthrough for the Web, a website usability evaluation method, and a basic model underlying it. The reader may refer to [3], [4], and [5] for more detail.

2.1 Selecting an Item using the Label-Following Strategy

A visitor wishing to access a page with the desired information from a website with a layered link structure typically selects hyperlinks sequentially. User interfaces with layered link structures are widely employed for applications such as office software, ATMs, and cell phones.

Users encountering such an interface do not typ-

Manuscript received June 24, 2004.
Manuscript revised September 30, 2004.

[†]The author is with National Institute of Advanced Industrial Science and Technology.
[††]The author is with University of Tsukuba.

ically memorize the order of operations and retrieve those actions from memory. They view the contents displayed and select an appropriate item at that point. One study indicated that users do not remember menu items when using menu-based equipment[6], but rather select items using a label-following strategy[7][8]. The primary standard for selection is the degree of semantic matching between the task description and the representation of item. Therefore, the task will not be performed appropriately when the item selection interface design does not match the actual task.

2.2 CWW and its Underlying Models: CoLiDeS and LICAI

CWW is a usability inspection method in which usability is evaluated by simulating the website search process by a user through a cognitive model and thereby usability problems are detected and repairs are suggested. A simulation is carried out based on a cognitive model, the CoLiDeS model (**Co**mprehension-base **Li**nked model of **De**liberate **S**earch) [7]. The CoLiDeS model is a cognitive model of a user's process of manipulating a device while examining an interface, expanded from the LICAI model (**LI**nked model of **C**omprehension-based **A**ction planning and **I**nstruction taking) [9], which simulates the aforementioned label-following strategy.

The psychological processes at the core of those models are three selection processes, i.e., selection of the region, selection of the object of the action, and selection of the action. Each selection is made in two phases. The user first considers the information shown on the interface display in terms of the purpose of the action. The user then selects the item closest to the purpose, based on their understanding of the information. The former is a comprehension process and the latter is a problem-solving process called a means-ends analysis. In CoLiDeS and LICAI models, the process of comprehension is modeled based on the cognitive theory, the Construction-Integration theory[10], that explains how a person comprehends a sentence.

The process of understanding text and graphics, which are objects on a web page, is very important in the CoLiDeS model. A psychological image of the web page object is created, details are added using associated existing knowledge, and the compatibility with the purpose is evaluated. The object most compatible is selected as the object of action. The CoLiDeS model separates the three selection processes described above into two levels, the attention phase and action selection phase, to model a user's web navigation behavior. The user separates the web page into partial regions in the attention phase and assigns an appropriate description to each. The user may use words to describe the heading and page layout. The user then selects the region most compatible with the current purpose (the region selection process). In the action selection phase, the user creates psychological descriptions for all widgets within the selected partial region so as to select the target of action with the greatest compatibility with the current purpose (the object selection process). The user then selects the action (generally by clicking on it) that corresponds with that particular widget (hyperlink, etc.) (the action selection process).

2.3 Evaluation of Semantic Similarity Using Latent Semantic Analysis (LSA)

In the CoLiDeS model, overall compatibility of the representation of interface objects and the description of purpose is evaluated to select an appropriate item from the interface. The overall evaluation includes literal matching and semantic matching, and is made by spreading activation within the network with the interface objects and activated knowledge as nodes [11]. However, this procedure is simplified in CWW to evaluate the compatibility using only semantic similarity.

LSA(**L**atent **S**emantic **A**nalysis)[12] is used to quantify semantic similarity. LSA is a mathematical/statistical technique for extracting and representing the similarity of meaning of words and passages by analysis of large bodies of text. It uses singular value decomposition, a general form of factor analysis, to condense a very large matrix of word-by-context data into a much smaller, but still large, typically 100–500 dimensional-representation. The right number of dimensions appears to be crucial; the best values yield up to four times as accurate simulation of human judgments as ordinary co-occurrence measures [13]. An expression synthesized from multiple words is expressed by synthesizing vectors of the component words. Any semantic similarity between two synthesized expressions is defined by the cosine of the angle formed by two corresponding vectors. For example, the semantic similarity between *human computer interaction* and *software engineering* is 0.64. This figure indicates that those two expressions tend to appear in a same context, and thus are similar expressions. In contrast, the similarity of expressions like *parenting* and *human computer interaction* is 0, indicating that they do not appear in the same context. Thus, semantic similarity is objectively quantified by using LSA. There is a web page that interactively evaluates the similarity between words or synthetic expressions based on semantic spaces, constructed using the vocabulary level of American students from several grades (*http://lsa.colorado.edu*).

The user's purpose can be described more realistically using LSA. Goal descriptions tend to be simple and specific in conventional modeling research to locate information, such as "searching for information regarding heart disease." However, users searching the web do not necessarily have such clearly described goals. LSA enables CWW to include in the goal description not

Table 1 Encarta categories and the number of topics and articles they contain

Category	Number of Topics	Number of Articles
Art, Language, & Literature	13	5,309
Geography	13	8,978
History	9	6,087
Life Science	14	5,153
Performing Arts	6	4,845
Physical Science & Technology	16	4,930
Religion & Philosophy	7	2,900
Social Science	12	6,562
Sports, Hobbies, & Pets	4	1,640

only the direct goal to be searched, but also general interest, motivation, and background information.

2.4 Psychological and Practical Validity of CWW

The Cognitive Walkthrough for the Web (CWW) is a partially automated usability evaluation method for identifying and repairing website navigation problems. Relying on LSA produces the same objective answer every time, and laboratory experiments confirm that actual users almost always encounter serious problems whenever CWW predicts that users will have problems doing a particular task [4]. Furthermore, using CWW to guide problem repair yields two-to-one gains in user performance [3].

In a recent study [14], it is reported that CWW has high psychological validity, for both the previously collected data reported in [4] [3] and a new cross-validation experiment, revealed by

1. high hit rates and low false alarms for identifying problems,
2. high rates of correct rejections and low rates of misses for identifying non-problems,
3. accurate measures of problem severity, and
4. high success rates for repairs of identified problems.

In summary, the series of studies concerning CWW [4] [3] [14] confirmed that CWW is a reliable and valid usability inspection method for evaluating the usability of website navigation. This paper uses CWW for detecting a sub-class of usability problems of a large-scale informational website.

3. Usability Evaluation of a Large-Scale Information Provider Website by CWW

In this section, we examine usability problems by applying CWW to an existing large-scale information provider website. We have chosen the Encarta Encyclopedia (*http://encarta.msn.com*) on-line reference site by Microsoft as the website to be analyzed for this paper.

3.1 Overview of the Website to Be Evaluated

The Encarta website is updated daily and its detail changes frequently; our examination is based on the information at the time of analysis, the beginning of 2003. There were 41,952 articles on Encarta, divided in a layered fashion into 9 categories and 94 topics. To access the targeted article, one must first select a category (region selection process) and then select a topic (object selection process). The desired title is then chosen from the list of article titles displayed alphabetically under the specific topic. About 90% of the articles have a unique access route; the remaining 10% (4,454 articles) have multiple access routes. Each category and the number of topics and articles contained within it are listed in Table 1. Encarta is a good target for analysis since CWW only evaluates categories and topics used for navigation.

3.2 Approximation of User Search Goals

The original CWW procedure [4] requires the analyst to enumerate potential user goals with 100-200 words narrative texts, but this procedure is not feasible for evaluating existing large-scale websites consisting of more than 10,000 content pages. To approximate the goal enumeration process, we used the first paragraph of all terminal node pages, so we could automatically harvest goals. The average length of the first paragraph of each article was 93 words, an adequate length for LSA analyses. The first paragraph usually overviews the whole article, so using the first paragraph simulates a person who knows a little about the information covered in the article. The user loads this limited information into working memory while searching for the more detailed information available in the full encyclopedia article by successively selecting the right category and the right topic from the layered menu hierarchy.

3.3 Discovering Usability Problems Using CWW

We explain a method to discover usability problems using CWW in this section and describe the results of an actual Encarta website usability evaluation.

3.3.1 Problem definitions

The following three potential usability problems could have been examined using CWW[4]:

1. Are the categories and topics described in terms understandable to the visitors? (*Unfamiliar categories/topics problem*)
2. Are the descriptions easy to identify? (*Confusing categories/topics problem*)

3. Can the correct category be selected, and will the correct topic under the correct category be selected? (*Goal-specific competing categories/topics problem*)

The primary topic of this paper is a usability evaluation for a website with a given search goal. Therefore, we will focus on the third usability problem above within the structure of our object of analysis, the Encarta website. The followings describe concrete criteria for discovering usability problems used in this paper, consistent with those defined in the original CWW paper [4]:

1. **Problems associated with categories**

   a. **Weak scent problem in category** (called **C-WS** hereafter): When the semantic similarity value between the representation of an article to be searched and that of each category is below a certain threshold value (δ), we designate this article as having a problem with an insufficient scent for category selection. We set $\delta = 0.1$ for this paper. The condition for C-WS is expressed as follows:

   $$\text{If } \max_{i=1,\cdots,9} sim(<A>,<C_i>) \ < \ \delta,$$
   $$\text{then C-WS applies,}$$

   where $sim(<A>,<C_i>)$ denotes the semantic similarity value between the representation of target article A and that of ith category C_i. Its detailed definition will be given in 4.1.2.

   b. **Goal-specific competing category problem** (called **GSCC** hereafter): Even if there is no problem with insufficient scent for the category selection, we determine that there is a problem with correct category selection if the category with the greatest semantic similarity value does not belong to the category containing the target article. The condition for GSCC is expressed as follows:

   $$\text{If C-WS does not apply,}$$
   $$\text{and if}$$
   $$\max_{\substack{i=1,\cdots,9 \\ i\neq i_r}} sim(<A>,<C_i>) > sim(<A>,<C_{i_r}>),$$
   $$\text{then GSCC applies,}$$

   where C_{i_r} is the correct category where the target article is to be found.

2. **Problems associated with topics**

   a. **Weak scent problem in topic** (called **T-WS** hereafter): An article has a problem with insufficient scent for topic selection when both the maximum value of semantic similarity with the correct topic and the maximum value of semantic similarity with incorrect topics are below a certain threshold value (δ') for each topic under the category containing the target article. We set $\delta' = 0.1$ for this paper. The condition for T-WS is expressed as follows:

   $$\text{If } \max_{j=1,\cdots,N_{i_r}} sim(<A>,<T_j^{(C_{i_r})}>) \ < \ \delta',$$
   $$\text{then T-WS applies,}$$

   where C_{i_r} is the correct category where the target article is to be found, and $T_j^{(C_{i_r})}$ denotes the topics nested under the correct category and N_{i_r} is the number of topics of the correct category.

   b. **Goal-specific competing topic problem** (called **GSCT** hereafter): Even if there is no problem with insufficient scent for topic selection, we determine that there is a problem with correct topic selection when there is a topic for which the ratio between the value of semantic similarity for the correct topic and incorrect topics falls below a certain threshold value (γ). We set $\gamma = 0.8$ for this paper. The condition for GSCT is expressed as follows:

   $$\text{If T-WS does not apply,}$$
   $$\text{and if}$$
   $$\exists j; \max_{\substack{j=1,\cdots,N_{i_r} \\ j\neq j_r}} sim(<A>,<T_j^{(C_{i_r})}>) > \gamma \times sim(<A>,<T_{j_r}^{(C_{i_r})}>),$$
   $$\text{then GSCT applies,}$$

   where C_{i_r} and $T_{j_r}^{(C_{i_r})}$ are the correct category and topic, respectively, where the target article is to be found.

3.3.2 Usability evaluation results

Figure 1 summarizes our evaluation result for whether a visitor seeking a particular article on Encarta can reach the target page by following categories and topics. The evaluation was derived using the aforementioned four criteria for a visitor's search process. The number of articles successfully found with no problem for each criterion is designated under *no* in the figure and the number of articles with problems is designated under *yes*. This figure indicates that only 15%(6,648) of all articles could be searched without a problem. While only 19% were categories and topics with insufficient selection scent (C-WS and T-WS) ($6,050+191+1,631 = 7,872$), 27% of all articles had problems with correct category selection (GSCC) (11,200) and 65% of all articles revealed problems with correct topic selection (GSCT)

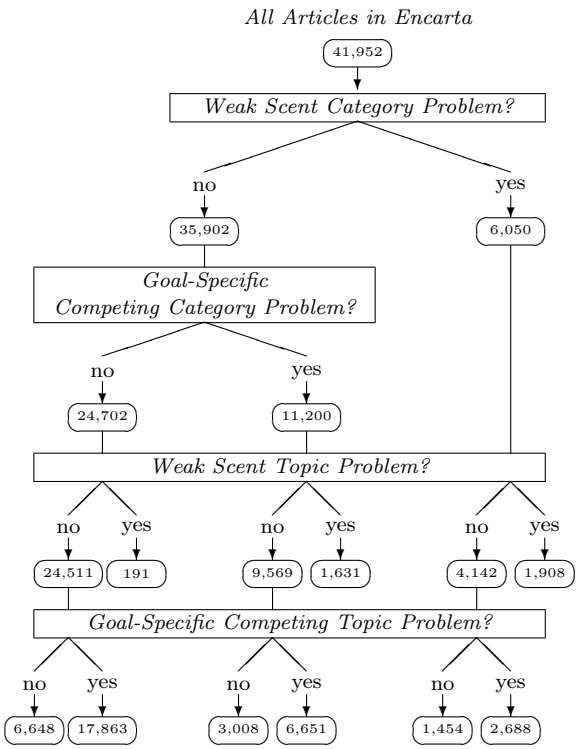


Fig. 1 Result of the Usability Evaluation of the Encarta Website

$(17,863+6,651+2,688=27,202)$. This indicates that while Encarta does not have many problems providing scent in category and topic selection, it is difficult to find the correct access routes.

Note that we used only the first paragraph of the target articles for this usability evaluation since it would not be an efficient use of time to use entire target articles on such a large-scale information provider website, and because Encarta articles tend to summarize their contents in the first paragraph. In addition, it was necessary to simulate the actual user determination process when calculating semantic similarity between the target article and each category/topic using LSA. Encarta allows the user to select a category by viewing and assessing the topics contained within. Therefore, we applied the categories and all topics within them as the representation of category for simplicity in this analysis, assuming this mimics the actual user determination process. In the same vein, we used the topics alone as the representations of topics since no other additional information appeared with the topics. Semantic space representative of the vocabulary level of American university students was used for the analysis of semantic similarity, using LSA for each necessary word.

Part of the problems detected in this usability analysis were examined through user testing with respect to the accuracy measures including hit/false alarm and miss/correct rejection rates with several hundred users and several dozen articles from the Encarta website. The results were promising; the detected problems were real problems that the participants encountered, and CWW missed few problems in the evaluation process[14].

4. Website Modeling and Usability Evaluation Using Markov Chains

Our focus in this section is to express the Encarta website using a Markov chain. We derive the average number of clicks before a visitor reaches his goal to evaluate the usability of an entire website quantitatively. We have identified the usability problems in the previous section. We will now provide a method to improve website usability by repairing the problems according to their characteristics, and evaluate the result of improvement quantitatively by describing the improved website using a Markov chain.

There have been previous attempts to evaluate usability using Markov chains. These attempts included research to evaluate the user interface efficiency and usability, such as modeling the operation success rate and execution efficiency in the interface design for devices such as cellular phones and microwave ovens using Markov chains[15]. Researchers have also evaluated the layered menu for each user through various user models[16]. Usability can be similarly evaluated by applying the Markov chain to sites on which the design space for the analysis target (the whole website for this paper) is vast and it is virtually impossible to evaluate its usability with individual user tests. However, the evaluation is not simply done on existing designs, but rather focuses on an evaluation of the effect of improvement on the usability problems, using CWW as a preliminary stage for detection and suggestion of improvement.

4.1 Website Modeling

4.1.1 Overview

In this section, we describe how to model the Encarta website using a Markov chain. The Encarta website has a three-tiered strata consisting of categories, topics, and articles. The process of a user reaching a target article based on the CoLiDeS model described in 2.2 was simulated as follows.

1. **Category selection:** A category that matches the target is first selected from the nine categories displayed on the home page.
2. **Topic selection:** A topic that matches the target is then selected from the list of topics under the selected category. Alternatively, the user may return to the category selection if there is no apparent appropriate route to the target article when selecting a topic.
3. **Article selection:** Selecting a topic causes a list

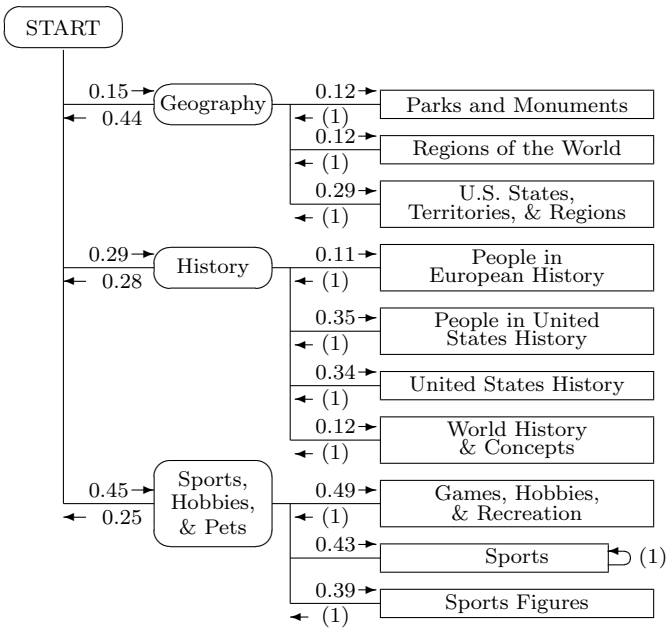


Fig. 2 Markov chain (probability transition example)

of article titles related to that topic to be displayed. The user determines whether there is a title that corresponds to the target article. If the target article is found, the title is selected and the article is located. The user returns to the topic selection if no target article is found.

Actual user behavior is probably not quite this simple. The user may abandon the search or get sidetracked in the midst of it. However, the purpose of this modeling is the quantitative evaluation of usability of an entire website, not an exact simulation of user behavior. We assume that users display the forward search-oriented behavior pattern described here if there is no problem with website usability[9], otherwise they may devise various problem-solving approaches. Therefore, we can conduct a comparative evaluation of website usability (between searches with and without problems) using a Markov chain by describing the search process model.

4.1.2 Forward transition

Figure 2 depicts the Encarta website model when an article on *Baseball* is sought. The right arrow in Figure 2 corresponds to the forward-transition model to reach a target article from each state. The values on the lines are semantic similarity values between this article and each of the categories and topics. Categories and topics with similarity values of less than 0.1 with the target article on the Encarta website are not shown in this figure. The semantic similarity between a category C and a target article A is defined by the cosine value of the angle formed by the two corresponding vectors in the semantic space that represent the concept of the category C, denoted as $<C>$, and the concept of the target article A, denoted as $<A>$, respectively. In sum, the semantic similarity value between a target article A and a category C is defined through respective concept, $<A>$ and $<C>$, denoted as $sim(<A>,<C>)$, and is obtained by calculating the cosine value of the angle formed by the corresponding vectors in a semantic space.

$$sim(<A>,<C>) = \cos(<\vec{A}>,<\vec{C}>)$$

Likewise, the similarity value between a target article A and a topic T is defined as follows:

$$sim(<A>,<T>) = \cos(<\vec{A}>,<\vec{T}>)$$

In this paper, $<C>$ is operationally represented as a collection of the label of the category itself and those of the topics nested under the category, $<T>$ as the label of the topic itself, and $<A>$ as the first paragraph of the target article (see 3.3.2 for rationale). For example, the semantic similarity between the category C = "Sports, Hobbies, and Pets" and the article A = "Baseball" can be expressed as follows.

$$\begin{aligned} & 0.45 \\ = \ & sim(<\text{Baseball}>, \\ & \qquad <\text{Sports, Hobbies, and Pets}>) \\ = \ & sim(<\text{Baseball}>, \text{“Sports Hobbies Pets} \\ & \qquad \text{Game Hobbies Recreation} \\ & \qquad \text{Pets} \\ & \qquad \text{Sports} \\ & \qquad \text{Sports Figures”}) \end{aligned}$$

In this expression, the concept of the category $<$ Sports, Hobbies, and Pets$>$ is expressed as a multiple-word concept that consists of the label of the category "Sports, Hobby, and Pets", and the labels of its topics, "Game, Hobbies and Recreation", "Pets", "Sports", and "Sports Figures." We used only the first paragraph for each article when calculating similarity as described previously. The following expressions are used for *Baseball* in the equation:

Baseball, *competitive game of skill played with a hard ball and bat between two teams of nine players each. Baseball is often called the national pastime of the United States, because of its strong tradition and great popularity. It is played throughout the world by people of all ages.*

4.1.3 Backward transition

The left arrow in Figure 2 corresponds to a backward transition model in which there is a return to the previous state from each state. This is the result of deriving similarity for returning from the category C_i to *Start*,

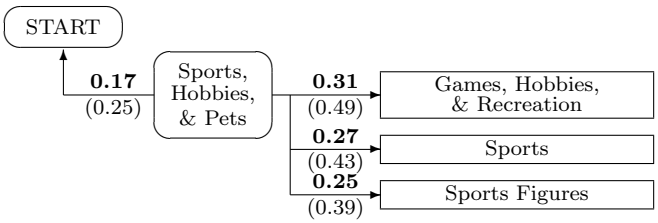


Fig. 3 a Markov chain example

as defined by

$$sim(<A>, \sum_{k=1}^{9} <C_k> - <C_i>)$$

Note that the summation in the above formula expresses the operation of combining multiple concepts that are derived from the categories, which are effectively represented as addition of corresponding vectors in the semantic space. For example, the transitional similarity when returning from the category $C =$ "Sports, Hobbies, and Pets" to *Start* is defined as follows when the target is $A =$ "Baseball" and $Back(A, C)$ is the transitional semantic similarity for returning from the category C to *Start*:

$$\begin{aligned}
& 0.25 \\
= \ & Back(\text{"Baseball"}, \\
& \qquad \text{"Sports, Hobbies, and Pets"}) \\
= \ & sim(<\text{Baseball}>, \\
& \qquad (<\text{Art, Language, and Literature}> + \\
& \qquad <\text{Geography}> + \cdots + \\
& \qquad <\text{Sports, Hobbies, and Pets}>) - \\
& \qquad <\text{Sports, Hobbies, and Pets}>) \\
= \ & sim(\text{Baseball}, \\
& \qquad <\text{Art, Language, and Literature}> + \\
& \qquad <\text{Geography}> + \cdots \ ) \\
= \ & sim(Baseball, \text{"Art Language Literature} \\
& \qquad \text{Architecture Artists} \\
& \qquad\qquad \cdots \text{ Geography } \cdots \text{ "})
\end{aligned}$$

The values inside the parenthesis under the lines between each topic and category in the figure are transition probabilities in a Markov chain for the transition from the state after topic selection, followed by category selection, and back to topic selection. The user must return to topic selection (state of searching for the correct topic) if he or she cannot find the target article after topic selection; therefore, the probability of the transition to that state is 1. The user achieves his goal without having to return to topic selection if the selected topic is the correct route to the target article, making the selected topic the absorption state for the Markov link. The topic *Sports* in Fig. 2 indicates such a state.

We can establish the transition probability from one state to each of the other states based on the semantic similarity thus derived and by applying weight to each transition. However, selecting a topic under an incorrect category produces an insufficient scent for the topic when the semantic similarity for all the topics is below a certain threshold value (here, δ'). Those topics are not selected and the user returns to *Start* with a probability of 1. In addition, the transition probability is assumed to be 0 for those with negative semantic similarity values. Figure 3 illustrates an example of the state after conversion to the transfer probability from the semantic similarity. This indicates the transition probability to the connected node when selecting the category "Sports, Hobbies, and Pets", based on the semantic similarity indicated in Figure 2. The values above the lines are transition probabilities and the values inside the parenthesis below them are the semantic similarity values given in Figure 2.

4.2 Formulation Using a Markov Chain

The modeling described above enables us to assess a user's category/topic selection behavior at a layered website like Encarta using a Markov chain. The web page transition process can be seen as a finite Markov chain with a discrete state space and a discrete time parameter if we consider each page as one state and one click as a time parameter and consider it sufficient to have viewed the previous page before arriving at the current page.

Consequently, the following equation holds true if X_n is the state of the page appearing at the nth click.

$$\begin{aligned}
& P\{X_{n+1} = j \ \mid \ X_0 = i_0, X_1 = i_1, ..., X_n = i_n\} \\
& \qquad = P\{X_{n+1} = j | X_n = i_n\} \qquad (1)
\end{aligned}$$

The term "click" used here represents a mouse click necessary to navigate a page within a website by selecting hyperlinks (categories, topics, or "Back" button), not clicks associated with operations other than selections. The only absorption state for this Markov chain is the page following the correct topic selection where the title of the target article appears. Thus, modeling enables us to analyze a user's behavior starting from the home page at *Start* until absorbed at the absorption state, discovery of the target article (or its title).

4.3 Average Number of Clicks

If the average steps before a Markov chain starting with non-recursive state i becomes absorbed by some recursive equivalence class, the average absorption steps would be $\mu^{(i)}$, and a set in a non-recursive state of $\mathcal{T}$ satisfies the simultaneous equations:

$$\mu^{(i)} = 1 + \sum_{k \in \mathcal{T}} P_{ik} \mu^{(k)} \qquad (2)$$

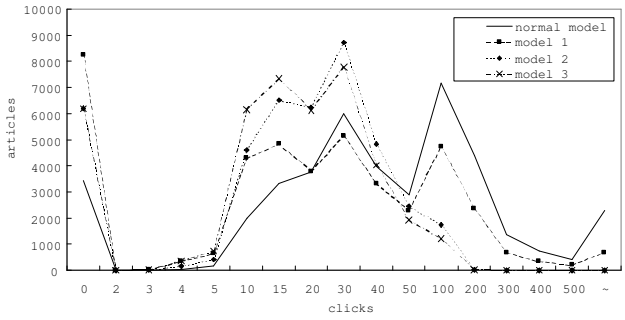


Fig. 4 Comparison of the number of clicks between before and after switching links

Interpreting Eq. (2) semantically indicates that it takes one step to transfer from state i to another state (the first argument on the right side), and it takes an average of $\mu^{(k)}$ steps to move from the transferred state k to an absorption state. All cases of states k where it is possible to transfer from state i are added together (the second argument on the right side). This equation enables us to derive the average number of clicks necessary for a user visiting a website to reach the absorption state, discovery of the target article, counting from *Start*.

4.4 Website Usability Improvement and Evaluation Using a Markov Chain

The result of the usability evaluation in 3.3 suggests that, while there are a few articles with category or topic weak scent problems at the Encarta website, there are several articles with problems relating to correct category or topic selection. This creates significant potential for a user to select incorrect categories and topics. Therefore, repairing this problem is important for improving the usability of the entire Encarta site. We will repair the usability problems through simple link changes and describe a method to evaluate its effect using a Markov chain.

4.4.1 Website improvement: Link alteration

We can make it easier for a visitor to reach the target article by repairing usability problems through the following three types of link alteration.

Link Alteration 1: We attempt to improve usability by altering the links to topics to make it easier to select a correct topic when the user completes the correct category selection. The most prominent GSCT problems in Encarta website usability are expected to be repaired through this link alteration.

Method: If there are neither C-WS nor T-WS problems, we alter the topic with the maximum semantic similarity with the target article out of all incorrect topics under the correct category to create a correct topic.

Link Alteration 2: A user searching for the target article selects a category first and then selects a topic under that category. Therefore, we attempt to improve usability to make a visitor's selection behavior most efficient by altering the category with the maximum semantic similarity with the target article as well as correcting the topic with the maximum semantic similarity under that category. GSCC is eliminated with this alteration and T-WS and GSCT should be improved.

Method: When there are no C-WS problems, we repair the topic with the maximum semantic similarity under the category with the maximum semantic similarity with the target article. However, all of the semantic similarity values for all topics under the category should be above the threshold value δ.

Link Alteration 3: Performing link alterations 1 and 2 simultaneously will greatly improve the usability of the entire Encarta website.

Method: Link alterations 1 and 2 are performed simultaneously. We perform all alterations that are possible under the existing conditions.

Note that we regard the proposed method of link alterations to alleviate navigational problems as a quick fix, in which alternative access routes to target articles with navigational problems are added while preserving most of the original link structure which must reflect the design policy of the site developer. If it is suggested a substantial amount of link alterations should be applied, it would not be wise to apply them immediately because it might signal the discrepancy between the conceptualization of the design policy by the site developer and the perceived design policy of the site by the target users from the current implementation of the site. Thus, the site developer would have to step back and reconsider an alternative representation of information structure that should fit with the target users better than the current implementation.

4.4.2 Evaluation of the effect of improvement and the plausibility of the evaluation

We will now evaluate the effect of the usability improvements described in the previous section through analysis with a Markov chain. Figure 4 compares the average number of clicks required to arrive at the target article before and after link alteration. The horizontal axis indicates the average number of clicks, and the vertical axis indicates the number of articles. The usability evaluation result in 3.3 indicates that $34,080$ articles are candidates for alteration 1, $35,902$ articles are candidates for alteration 2, and $34,080$ articles are candidates for alteration 3. Of these candidates, the numbers of articles that do not meet the alteration conditions are $7,872$, $6,048$, and $6,070$, respectively, and the number of articles that meet the link alteration conditions but

unreachable are 390, 123, and 123, respectively. From these, of the candidates for the link alterations, 8,262, 6,171, and 6,193 articles in the respective alteration conditions are not considered for the further calculations. These numbers along with the number of unreachable articles before the link alterations, 3,422, are plotted at 0 on the horizontal axis of the figure.

The number of articles requiring several dozens to hundreds of clicks is reduced by any of the alterations, replaced by articles reachable after 20 to 30 clicks. This becomes evident, particularly with link alterations two and three, indicating that those approaches are effective. The graphs illustrate that more than half of the articles can be reached within 100 clicks. The problems are substantially reduced.

Thus, modeling through a Markov chain enables us to easily confirm the effects of improvement. The least number of clicks to arrive at a goal in actual operations at the website is two; therefore, the appropriate number of clicks calculated using a Markov chain model should be a few clicks to a few dozen clicks. The number of clicks is greater using the model because user behavior in a situation involving problems with category/topic selections is not modeled precisely using a Markov chain. The excessive number of clicks indicates that there *is* a problem with the Encarta website, not with the method of discovering usability problems or performing comparative evaluations.

Multiple access paths to a target article can be established using the method of improving web usability through link alteration based on the CWW usability evaluation result described in this paper. The effectiveness of this improvement method has been confirmed through user tests executed for hard-to-find articles discovered through this evaluation [14]. The average efficiency actually doubled in terms of the number of clicks.

This demonstrates that discovery and repair of usability problems using CWW as described in this paper is a certain and plausible method. We can discover usability problems using CWW by referring to the semantic similarities among targets, categories, and topics when the correct categories and topics are not significantly more similar than the others, making correct selection difficult. The number of clicks would thus increase in a model of visitor's selection behavior using a Markov chain. However, the number of clicks decreases if the similarity values are sufficiently high. This tendency should be maintained even when the approximation is crude.

We are now researching how accurately the Markov chain model simulates a visitor's selection behavior, as well as how we can improve the accuracy. We have not yet strictly correlated the results of user tests and the prediction using a Markov chain model. However, we anticipate no problems with improvement, provided the characteristics described above are maintained.

5. Conclusion

We evaluated the usability of an existing large-scale information provider website in this paper using the Cognitive Walkthrough for the Web, a usability inspection method. We modeled the website using a Markov chain and demonstrated that the average number of clicks before a visitor reaches a goal can be analyzed simply and that the effect of the usability improvement method suggested by the Cognitive Walkthrough for the Web can be evaluated quantitatively. As a result, we discovered that simple link alteration can greatly improve usability. We believe we can evaluate the degree of potential usability improvement by applying this method to similar websites.

References

[1] J.M., Spool, T. Scanlon, W. Schroeder, C. Snyder, and T. DeAngelo, *Web Site Usability: A Designer's Guide*, Morgan Kaufmann, San Francisco, 1999.

[2] E. Ojakaar and J.M. Spool, Getting Them to What They Want, *UIE Reports: Best Practices Series*, 2001.

[3] M.H. Blackmon, M. Kitajima, and P.G. Polson, Repairing Usability Problems Identified by the Cognitive Walkthrough for the Web, *Proceedings of the conference on human factors in computing systems (CHI'2003)*, 497–504, 2003.

[4] M.H. Blackmon, P.G. Polson, M. Kitajima, M. and C. Lewis, Cognitive Walkthrough for the Web, *Proceedings of the conference on human factors in computing systems (CHI'2002)*, 463–470, 2002.

[5] M. Kitajima, Cognitive Walkthrough for the Web, *Journal of Japan Society for Fuzzy Theory and Systems*, **14**, 446–460, 2002.

[6] T.J. Mayes, S.W. Draper, A.M. McGregor, and K. Oatley, Information flow in a user interface: The effect of experience and context on the recall of MacWrite screens, *Proc. HCI'88 Conference on People and Computers IV*, Cambridge, 275–289, Cambridge University Press, 1988.

[7] M. Kitajima and P. Polson, A comprehension-based model of exploration, *Human-Computer Interaction*, **12**, 4, 345–389, 1997.

[8] C. Wharton, J. Rieman, C. Lewis, and P. Polson, The cognitive walkthrough method: A practitioner's guide, *Usability inspection methods*, J. Nielsen and R.L. Mack (Eds.), New York, 105–140, John Wiley and Sons, 1994.

[9] M. Kitajima, M.H. Blackmon, and P.G. Polson, A Comprehension-based Model of Web Navigation and Its Application to Web Usability Analysis. In S. McDonald, Y. Waern & G. Cockton (eds.), *People and Computers XIV - Usability or Else! (Proceedings of HCI 2000)*, Springer, 357–373, 2000.

[10] W. Kintsch, *Comprehension: A Paradigm for Cognition*, Cambridge University Press, Cambridge, 1998.

[11] M. Kitajima and P.G. Polson, A comprehension-based model of correct performance and errors in skilled display-based, human-computer interaction, *International Journal of Human-Computer Studies*, **43**, 65–99, 1995.

[12] T.K. Landauer and S.T. Dumais, A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, **104**, 211–240, 1997.

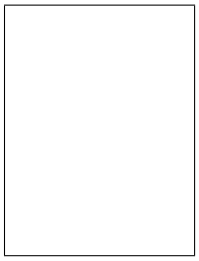
[13] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer,

and R. Harshman, Indexing by Latent Semantic Analysis, *Journal of the American Society For Information Science*, **41**, 391–407, 1990.

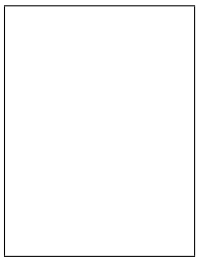
[14] M.H. Blackmon, M. Kitajima, D. Mandalia, and P.G. Polson, Automating Usability Evaluation: Cognitive Walkthrough for the Web Puts LSA to Work on Real-World HCI Design Problems, *LSA Workshop for "LSA: A Road to Meaning"*, 2004.

[15] H. Thimbleby, P. Cairns, and M. Jones, Usability analysis with Markov models, *ACM Transactions on Computer-Human Interactions*, Vol.8, No.2, pp.99–132, June 2001.

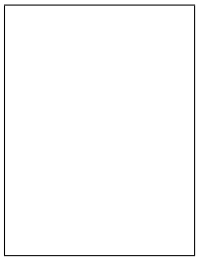
[16] M. Kitajima, H. Takagi, T. Yamamoto, and Y. Zhang, Search process evaluation for a hierarchical menu system with Markov chains constructed by using latent semantic analysis, *Journal of Information Processing Society of Japan*, **43**, 3722–3732, 2002.



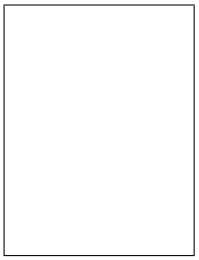
Muneo Kitajima is Leader of Human–Computer Interaction Group at the Institute for Human Science and Biomedical Engineering, National Institute of Advanced Industrial Science and Technology, Japan. He received B. S and M. S in Physics from Tokyo Institute of Technology in 1978 and 1980, respectively, and Dr. Eng. in Applied Physics from the University of Waseda in 1986. Prior to the current position, he was a researcher at the Industrial Products Research Institute, a senior researcher at the National Institute of Bioscience and Human-Technology. His research interest includes cognitive modeling in human–computer interaction.



Noriyuki Kariya received the B. S. and M. S. degrees in Policy and Planning Sciences from University of Tsukuba, Japan, in 2002 and in 2004, respectively. He is now with Future System Consulting Corporation. He is interested in the performance and usability evaluation of web sites.



Hideaki Takagi is Professor at the Graduate School of Systems and Information Engineering of the University of Tsukuba, Japan. He received B. S. and M. S. in Physics from the University of Tokyo, and Ph. D. in Computer Science from the University of California, Los Angeles. Prior to the current position, he was Consultant Researcher at IBM Research, Tokyo Research Laboratory, Professor and Chair at the Institute of Policy and Planning Sciences of the University of Tsukuba, and Vice President of the University of Tsukuba. His research interest includes enumerative combinatorics, probability models, and performance evaluation of computer and communication systems. He is IEEE Fellow and IFIP Silver Core holder.



Yongbing Zhang is an Associate Professor at the Graduate School of Systems and Information Engineering of the University of Tsukuba. He received his M. S. and Dr. Eng. in Computer Science in 1989 and 1992, respectively, both from the University of Electro-Communications, Tokyo, Japan. From 1992 to 1996, he was a Research Associate with the Department of Computer Science at the University of Electro-Communications. In 1996, he joined the Institute of Policy and Planning Sciences, University of Tsukuba. His research interests include wireless communications, optical networks, and performance evaluation.