Stochastic Modeling of Usage Patterns in a Web-Based Information System

Hui-Min Chen and Michael D. Cooper*
School of Information Management and Systems, University of California at Berkeley, Berkeley, CA 94720-4600. E-mail: hmchen@sims.berkeley.edu, cooper@socrates.berkeley.edu

Users move from one *state* (or task) to another in an information system's labyrinth as they try to accomplish their work, and the amount of time they spend in each state varies. This article uses continuous-time stochastic models, mainly based on semi-Markov chains, to derive user state transition patterns (both in rates and in probabilities) in a Web-based information system. The methodology was demonstrated with 126,925 search sessions drawn from the transaction logs of the University of California's MELVYL® library catalog system (www.melvyl.ucop.edu). First, user sessions were categorized into six groups based on their similar use of the system. Second, by using a three-layer hierarchical taxonomy of the system Web pages, user sessions in each usage group were transformed into a sequence of states. All the usage groups but one have third-order sequential dependency in state transitions. The sole exception has fourth-order sequential dependency. The transition rates as well as transition probabilities of the semi-Markov model provide a background for interpreting user behavior probabilistically, at various levels of detail. Finally, the differences in derived usage patterns between usage groups were tested statistically. The test results showed that different groups have distinct patterns of system use. Knowledge of the extent of sequential dependency is beneficial because it allows one to predict a user's next move in a search space based on the past moves that have been made. It can also be used to help customize the design of the user interface to the system to facilitate interaction. The group CL6 labeled "knowledgeable and sophisticated usage" and the group CL7 labeled "unsophisticated usage" both had third-order sequential dependency and had the same most-frequently occurring search pattern: screen display, record display, screen display, and record display. The group CL8 called "highly interactive use with good search results" had fourth-order sequential dependency, and its most frequently occurring pattern was the same as CL6 and CL7 with one more screen display action added. The group CL13, called "known-item searching" had third-order sequential dependency, and its most frequently occurring pattern was index access, search with retrievals, screen display, and record display. Group CL14 called "help intensive searching," and CL18 called "relatively unsuccessful" both had third-order sequential dependency, and for both groups the most frequently occurring pattern was index access, search without retrievals, index access, and again, search without retrievals.

Introduction

Users of Web-based library catalogs search for information using a set of functions and features that are generally standard among vendor systems. This article uses probabilistic models to analyze the patterns of use of these functions and features. The models take into account two factors: the probability that a user moves from one state to another (such as from performing a search to viewing the results of the search), and the time spent in each of the states.

The system used for this study is the Web version of the MELVYL® catalog of the University of California. MELVYL is a union catalog of the holdings of the nine campuses of the University of California as well as other institutions. The analysis began with an identification of the activities that can be performed by a user of the catalog and a translation of these activities into *states*—defined activities that take place during a user session. A set of about 127,000 user sessions was analyzed to deduce the patterns of users' movement from one state to another.

This article begins with an analysis of the state structure of the MELVYL system and a presentation of a very simple Web-based search episode. The goal is to show how a session can be represented as a series of states. Next is an introduction to discrete-time Markov chains and semi-Markov models, followed by a review of previous studies of user state-transition behavior that have employed some of these models in library catalog analyses.

Received March 13, 2001; Revised September 14, 2001; accepted January 10, 2002

* To whom all correspondence should be addressed.

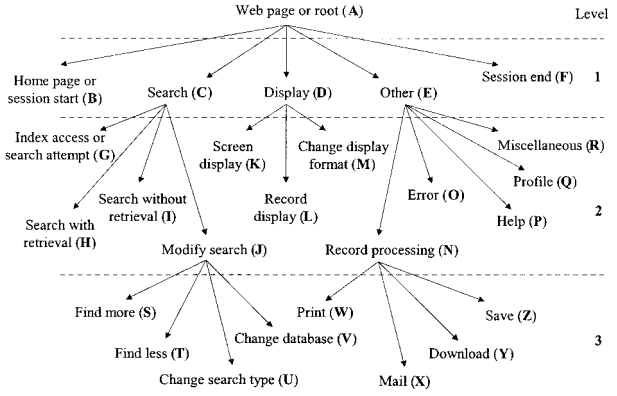

FIG. 1. A hierarchical taxonomy of Web pages in the University of California's MELVYL® on-line catalog system.

Two of the key aspects of an analysis of usage with these semi-Markov models are (1) determining the probability of moving from one state to another, and (2) determining how many previous states likely influenced the user entering the current state, that is, determining the order of the process. Consequently, an analysis of the order of the sample data is presented next. Then the transition rates are calculated and the time a user spent in each state are determined. Finally, the most frequently occurring patterns of use of the system are calculated from the sample data. The article concludes with both qualitative and quantitative analyses of the usage patterns.

Defining a State Space

In a Web-based information system, the user's state at any point in time can be represented by the latest (current) Web page requested by the browser (user) from the Web server. For simplicity, the Web pages in the MELVYL system can be mapped into 26 *states* based on their functionality, as shown in Figure 1. In the figure each state is represented by a one-letter code.

Because there is a trade-off between resolution of states and computational cost, the 12 states on level two constitute the state space for this study. From left to right, they are *index access* or *search attempt activity* (G), *search with retrievals* (H), *search without retrievals* (I), *modify search* (J), *screen display or multiple-record display* (K), *record display or single-record display* (L), *change display format* (M), *record processing* (N), *error* (O), *help (*P), *profile* (Q), and *miscellaneous* actions (R). States B (home page or session start) and F (session end) are omitted because they always appear at the beginning and end of a session. A one-to-one mapping between Web pages in the MELVYL system and states is given in Chen (2000).

The transaction log in Table 1 illustrates the concept of states. In this table there are 12 transaction records representing a user session with the MELVYL catalog. The table shows the date and time the transaction took place, followed by the type of transaction. In the final column of the table is a categorization of the state that the user was in (according to Fig. 1). The transaction records for the user session in Table 1 can be replaced by a sequence of states: B-G-G-G-H-H-K-K-J-P-P-F. That is, the user began in state B, and then entered state G three times in a row, and so on.

Basic Concepts

The basic principle of Markov models is the *Markovian* or *memoryless property*. Given a collection of exhaustive and mutually exclusive states (such as in Fig. 1), the next state in a sequence is independent of the past and solely depends on the current state.

Transition probabilities are used to characterize the movement of users through states, regardless of the time spent in each state. A *zero-order state transition probability* gives the probability of the occurrence of a single state out of all the states a user visits. A *first-order transition probability* gives the probability of moving from one state to another.

TABLE 1. An example of transaction records for a user session.

Transaction record	Activity (state)
1998/02/15 10:08:20.382|1856806655|12|ST|**start of session**	—
1998/02/15 10:08:20.627|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**home.shell**	Session start (B)
1998/02/15 10:10:37.477|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**profile/home.shell**	Presearch (G)
1998/02/15 10:11:18.647|1856806655|08|SH|shell=/meluser/melweb/prd/**shell/search/cat/title.shell**	Database (G)
1998/02/15 10:11:53.493|1856806655|09|SQ|query=TW (**Data mining**) and AT (**UCB**)	Index access (G)
1998/02/15 10:12:14.496|1856806655|10|SR|hits=30	Search (H)
1998/02/15 10:12:14.497|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**history/home.shell**	Search (H)
1998/02/15 10:12:43.817|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**display/Mcit.shell**	Display (K)
1998/02/15 10:13:27.019|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**display/Ocit.shell**"Cdisplay(**1,1Hcit,UCB**)"	Display (K)
1998/02/15 10:14:24.737|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**search/cat/power.edit.shell**	Modify search (J)
1998/02/15 10:14:56.321|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**help/cat/power_guided.shell**	Help (P)
1998/02/15 10:16:09.542|1856806655|08|SH|shell=/meluser/melweb/prd/shell/**resources/home.shell**	Help (P)
1998/02/15 10:36:09.541|1856806655|12|EN|**termination of session**	Session end (F)

Note: Long vertical bars divide fields in the transaction record. In order, the fields are: a date-and-time stamp, a unique identification number for the session, a record code, a record type, and the body of the record. See Cooper (1998) for details on the conceptual design of this log file.

For instance, the zero-order state transition probability for state G in Table 1 is the number of occurrences of G, which is 3, divided by the length of the sequence, which is 12, that is, $3/12 = 0.25$. The first-order transition probability from state G to state H is the number of occurrences of pattern G-H in the sequence divided by the number of occurrences of G, that is, $1/3 = 0.33$. Similarly, the second-order transition probability from state GG to state GH is the number of occurrences of pattern G-G-H in the sequence divided by the number of occurrences of GG, that is, $1/2 = 0.50$. The same rule can be generalized to a higher order model, where the order of the model or the order of sequential dependency is the number of states in a pattern minus one. This is the foundation for discrete-time state transition probability-based studies of usage patterns.

A drawback to these *discrete-time Markov chain* (DTMC) models is that time-dependent variables (such as session length and time between two Web page requests) are never factored into the models. Many researchers (Cooper, 1982, 1983; Fenichel, 1981; Penniman, 1975) have found that session length or length of stay is one of the variables that characterize user behavior. Further, the interval of time the user spends on a page reveals the intensity and evolution of state transition patterns. Two users may behave differently and consume different amounts of time even though they view the same number of Web pages. Besides, the discrete-time models allow duplicate states in succession, such as G-G-G in Table 1. Such a state-transition pattern is too detailed to give useful information.

To avoid the above shortcomings, this article proposes a continuous-time stochastic model based on semi-Markov chains to describe user state-transition behavior in a Web-based information system.

Definition of a Semi-Markov Chain

A semi-Markov chain (SMC) is a continuous-time stochastic (random) process that changes states in accordance with a discrete-time Markov chain but takes a random amount of time between state changes (Ross, 1996). Let X_0 be the initial state and X_n be the state entered on the nth transition, $n < 0$. Then $\{X_n\}$ is a discrete-time Markov chain (embedded) with the following property:

$$\begin{aligned} P_{ij} &= P\{X_{n+1} = j | X_n = i,\ X_{n-1} \\ &\quad = i_{n-1},\ \ldots,\ X_1 = i_1,\ X_0 = i_0\} \\ &\quad = P\{X_{n+1} = j | X_n = i\} \end{aligned}$$

Further,

$$P_{ii} = 0 \quad \text{and} \quad \sum_{j \neq i} P_{ij} = 1,$$

for all states $i_0, i_1, \ldots$ and all $n \geq 0$ (Ross, 1996). That is, the conditional probability distribution of any future state X_{n+1}, given the past states $X_0, X_1, \ldots, X_{n-1}$ and the present state X_n, is independent of the past states and depends solely on the present state.

Let T_n be the amount of time the Web user spent in one of the states in level 2 of Figure 1. Also, let X_n be the state entered on the nth transition, $n \geq 0$. It can be shown that given $\{X_n\}$, $\{T_n\}$ are independent random variables (Wolff, 1989). The distribution of T_n depends only on the states X_n and X_{n+1}. That is,

$$\begin{aligned} H_{ij}(t) &= P\{T_n \leq t | X_{n+1} = j,\ X_n = i, \\ &\qquad X_{n-1} = i_{n-1},\ \ldots,\ X_0 = i_0\} \\ &= P\{T_n \leq t | X_{n+1} = j,\ X_n = i\} \quad \text{for} \quad i, j \in S \end{aligned}$$

As such, the distribution of a duration of stay in state i is

$$H_i(t) = \sum_j P_{ij} H(t)_{ij},\ t > 0$$

with mean

$$u_i = \int_0^\infty t dH(t).$$

Also, the rate at which the semi-Markov process leaves state i at time t is the reciprocal of the mean duration of stay of state i, that is, $-q_{ii}(t) = 1/u_i$. This only applies to a homogeneous model in which the transition rates do not change over time.[1]

The discussion presented in this section is intended to provide a background for the literature review, dataset description, and data transformation discussions that follow. For the moment, we will defer more detailed discussion of the methodology used in the article and the testing procedures.

Previous Research

Several studies have employed discrete-time stochastic models to characterize user search behavior in information retrieval systems. Penniman (1975) defined 11 states and merged them into four categories. He then computed transition probability matrices and determined the number of previous consecutive states upon which the current state depends (the *order* of the discrete-time Markov chain). He found significant differences in both zero- and first-order models when users searched different databases, and between novices and experienced users.

[1] Chen and Cooper (2001) have shown that users of the Web version of the MELVYL system have homogeneous behavior.

Chapman (1981) also studied usage patterns with transition probabilities. The users' commands were recorded and mapped into a set of nine exhaustive and mutually exclusive states. Usage patterns in the form of zero- through fourth-order transition probability matrices were computed for each subject group and compared. To further condense the data and analyze the progression of the users' information-seeking behavior, repetitions of similar states (i.e., two or more of the same or similar states occurring in succession) were grouped together and represented by a string type. Zero- through fourth-order string transition probability matrices were constructed for each subject group, statistically tested for intergroup differences, and compared with those of state transition probability matrices. Chapman has demonstrated that state transition probability matrices can characterize use.

Later, Penniman extended his original work by dividing users by frequency of system use and computing zero- through second-order transition matrices (Penniman, 1982). He found the frequency distribution of usage patterns follows approximately a Zipfian curve across usage groups. Nearly 80% of activities are accounted for by about 20% of the activity types. There are a larger number of infrequent activities as the specificity (levels of generalization) of activities increases. This same observation was made when higher order models were analyzed. Most importantly, Penniman found that significant differences in short sequences of activities do not account for differences in longer sequences of activities.

Qiu (1993) used discrete-time Markov chains to describe user behavior. Eight search states were defined, and empirical testing showed that state transition behavior followed a second-order DTMC model.

Despite their popularity, a drawback to the discrete-time models is that session length, one of the variables that characterize user behavior, is not factored into the models. Previous researchers have probably excluded session length in studying user behavior because data transmission rates and machine speeds were rather slow and unpredictable. Now that data transmission quality has improved, both session length and response time can be measured accurately.

The MELVYL Transaction Log Dataset and Previous Clustering Analysis

The goal of this research was to analyze user state-transition patterns of a Web-based library catalog using a continuous-time stochastic model based on semi-Markov chains. The data used in the analysis was the same employed by Chen (2000), Cooper (2001a), Cooper (2001b), and Chen and Cooper (2001). The source of the data in this article is usage transactions logs from the University of California's Web-based MELVYL library catalog encompassing 126,925 search sessions conducted between February 15, 1998, and March 14, 1998.

In Chen and Cooper (2001) a methodology was developed to group sessions with similar characteristics using clustering techniques. Six clusters of usage patterns were derived, and it was verified using statistical techniques that the clusters were not found by chance. Replicating the analysis in another independent dataset did this. They were labeled CL6—knowledgeable and sophisticated usage containing 10,455 sessions; CL7—unsophisticated usage, with 47,372 sessions; CL8—highly interactive usage with good search results, with 17,192 sessions; CL13—known-item searching, with 34,363 sessions; CL14—help-intensive searching, with 3,919 sessions; and CL18—relatively unsuccessful usage, with 13,596 sessions. Each of these six clusters was found to be homogeneous using various statistical tests and in addition each group had distinct patterns of use of the system. Now, in this present article, we took those six clusters and analyzed the usage patterns again using different methodologies.

Data Transformation

The first step in the analysis was to take each user session in each of the six clusters and represent it by a sequence of states using the state mapping in level 2 of Figure 1. Thus, the search in Table 1 would be represented by the sequence of letters B-G-G-G-H-H-K-K-J-P-P-F. Multiple occurrences of a state in succession in a session were replaced with a single occurrence of the state. For example, the final state list for the search in Table 1 would be B-G-H-K-J-P-F.

Next, the duration of stay was calculated. The duration of stay in a state (i.e., the time the user stays in a state before entering another state) in the transformed session is simply the sum of the duration of stay of all its components. For instance, if the user spends 30 and 60 seconds in each of the two states H-H above, the duration of stay in state H after the data transformation will be 90 seconds.

The data transformation avoids deriving user state-transition patterns that are too detailed to give useful information. It also eliminates repetition of a state that results not from the user's intention, but rather from the dictates of the system design. For example, it might take a series of steps (Web pages) to modify a search in one system, whereas the same modification can be done in a single step (Web page) in another system. Although both sequences of states perform the same task, the former appears to be a series of successive modifications if repetition of a state is not eliminated.

Testing for the Order of Sequential Dependency

The order test determines how many previous states (including the current state) influence the choice of the next state probabilistically. The test results are able to indicate whether the transitions are zero-, first-, second-order processes (or higher). This allows a cause-and-effect interpretation of sequences of activities. The order test validates

TABLE 2. Results of testing the order of sequential dependency.

Usage group	Average number of states in a session	First-order sequential dependency	Second-order sequential dependency	Third-order sequential dependency	Fourth-order sequential dependency
CL6	15.70	20,825	5,178	3,148*	6,339
		(1,086, df = 980)	(3,264, df = 3079)	(4,334, df = 3905)	(5,497, df = 5256)
CL7	13.35	10,891	3,994	3,625*	9,256
		(1,268, df = 1038)	(3,296, df = 3110)	(5,081, df = 4616)	(7,236, df = 6680)
CL8	43.25	34,850	12,276	12,289	16,860*
		(1,368, df = 1249)	(5,713, df = 5220)	(11,998, df = 11281)	(21,631, df = 20,667)
CL13	9.13	36,879	3,779	2,138*	6,090
		(963, df = 863)	(2,462, df = 2140)	(2,845, df = 2498)	(3,719, df = 3322)
CL14	19.74	15,266	7,400	8,756*	17,309
		(1,469, df = 1221)	(6,114, df = 5604)	(12,534, df = 11,081)	(16,699, df = 15,852)
CL18	3.95	2,692	4,008	3,060*	4,165
		(721, df = 635)	(2,181, df = 2030)	(3,186, df = 3003)	(3,827, df = 3626)

Notes: An entry in the table gives the value of the chi-square statistic. Below each chi-square statistic is a parenthesized pair of numbers giving the critical value of the chi-square statistic and the number of degrees of freedom. Entries labeled with an asterisk indicate the null hypothesis is accepted.

stochastic modeling of user transition behavior using Markov chains.

Consider the user session G-H-K-N-R, which represents index access, search with retrievals, screen display, record processing, and miscellaneous actions. If the user transition behavior has an order of 2, then there are three transitions occurring in the session (GHK, HKN, KNR). That is, the likelihood of being in state K depends on having been in states G and then H previously, that of state N depends on having been in states H and then K previously, and that of state R depends on having been in states K and then N previously.

In this research, a stepwise procedure based on the chi-square goodness-of-fit test was employed to test the order of sequential dependency. It is described in Anderson and Goodman (1963) as follows.

Order Test

(1) Determine an upper bound to the order of the DTMC, say, N. Set $k = 1$.

(2) Formulate a hypothesis test, where
H_0: The DTMC has an order of k. That is, $P_{i_0 i_1 \ldots i_k, i_{k+1}} \neq P_{i_1 \ldots i_k, i_{k+1}}$ for all $i_0, i_1, \ldots, i_{k-1}, i_k, i_{k+1}$.
H_1: The DTMC does not have an order of k. That is, $P_{i_0 i_1 \ldots i_k, i_{k+1}} \neq P_{i_1 \ldots i_k, i_{k+1}}$ for at least one combination of $i_0, i_1, \ldots, i_{k-1}, i_k, i_{k+1}$.

(3) Calculate the chi-square statistic as follows:

$$C = \sum_{i_0} \sum_{i_{k+1}} \frac{(N_{i_0 i_1 \ldots i_k, i_{k+1}} - E[N_{i_0 i_1 \ldots i_k, i_{k+1}}])^2}{E[N_{i_0 i_1 \ldots i_k, i_{k+1}}]},$$

$$\forall\, i_1, i_2, \ldots, i_k$$

where $E[N_{i_0 i_1 \ldots i_k, i_{k+1}}] = N_{i_0 i_1 \ldots i_{k-1}, i_k} * N_{i_1 \ldots i_k, i_{k+1}} / \Sigma_{i_{k+1}} N_{i_1 \ldots i_k, i_{k+1}}$ is the expected number of transitions from state $(i_0, i_1, \ldots i_k)$ to state i_{k+1}, $N_{i_0 \ldots i_{k-1}, i_k}$ is the number of transitions from state $(i_0, i_1, \ldots, i_{k-1})$ to state i_k, $N_{i_1 \ldots i_k, i_{k+1}}$ is the number of transitions from state $(i_1 i_2 \ldots i_k)$ to state i_{k+1}, and $N_{i_1 \ldots i_k, i_{k+1}} / \Sigma_{i_{k+1}} N_{i_1 \ldots i_k, i_{k+1}}$ is the maximum likelihood estimation (MLS) of the transition probability $P_{i_1 \ldots i_k, i_{k+1}}$.

(4) Set a significance level α and degrees of freedom,

$$d.f. = \sum_{|(i_1 i_2 \ldots i_k)|} (S - I_{(i_1 i_2 \ldots i_k)} - 1)(S - O_{(i_1 i_2 \ldots i_k)} - 1)$$

where S is the number of different states in the state space, $I_{(i_1 i_2 \ldots i_k)}$ is the number of transitions toward state $(i_1 i_2 \ldots i_k)$, and $O_{(i_1 i_2 \ldots i_k)}$ is the number of transitions from state $(i_1 i_2 \ldots i_k)$ that cannot occur. $I_{(i_1 i_2 \ldots i_k)}$ and $O_{(i_1 i_2 \ldots i_k)}$ are termed *structural* or *sampling zeros* (i.e., impossible events) in the analysis of contingency tables (Everitt, 1992), and should be subtracted from the numbers of degrees of freedom. If $C \leq C_\alpha$, stop. The order of the DTMC is k. Otherwise, go to step 5.

(5) Let $k = k + 1$. If $k > N$, then stop the computation because the order of the DTMC is more than N. Otherwise, go to step 2.

The order of sequential dependency for each usage group was iteratively tested to see whether it was first through fourth order.[2] For example, the first-order test is to deter-

[2] Testing for higher order dependency beyond fourth-order is costly and was not pursued. The number of computational steps to test a particular order is given by

$$M*(M-1)^{N-1}*S*(L-N) + M^3*(M-1)^{3(N-1)},$$

where M is the number of states in the zero-order state space, N is the hypothesized order of sequential dependency, S is the number of sessions processed, and L is the number of states in a session. For instance, in searcher group CL7, M = 12, S = 47,372, and $L \approx 14$ (see Table 3). The number of computational steps required to test first-order, second-order, third-order, and fourth-order sequential dependency are 7,391,760; 77,337,216; 3,817,882,992; and 4,082,099,865,888, respectively.

mine whether or not the probability of being in state i_2, given that the process has been in states i_0 and then i_1 previously ($P_{i_0 i_1 i_2}$), is equal to the transition probability from state i_1 to state i_2 ($P_{i_1 i_2}$). The test results are displayed in Table 2. In each cell of columns three and beyond in Table 2, the value in the first row is the chi-square statistic, and the first value in the parentheses is the critical value of the chi-square statistic at a 0.01 level of significance given the corresponding degrees of freedom.[3]

All usage groups except cluster CL8 were found to follow third-order sequential dependency. In cluster CL8, user transition behavior has fourth-order sequential dependency. The average number of states in a session in cluster CL8 is larger (43.25) than the other usage groups. It seems that users in cluster CL8 are more likely to reflect back on what has been done in the past than those in the other usage groups because when they move to the next state, a larger number of previous states (including the current state) will influence their decisions. As reported in Chen and Cooper (2001), cluster CL8 is characterized as highly interactive usage with good search results. Users in this group are characterized by having the lengthiest sessions, the highest number, and most types of Web pages requested, the longest viewing time per Web page, and above all, the heaviest search activities. It seems that they made a series of interrelated searches during their visits, and thus, required more past information to make a decision. This hypothesis could be verified by comparing the query terms employed by a user over time in a longitudinal study. Although it is a research issue worth investigating, it is beyond the scope of this study.

Transition Rates Analysis

As mentioned previously, a semi-Markov chain has first-order sequential dependency; that is, it depends on the previous state. However, none of the six usage groups in the MELVYL system meet this criterion (see Table 2). Therefore, the state space had to be modified by combining zero-order states (so-called functions of Markov chains) to obtain first-order sequential dependency. In this case, the theoretical number of states that are possible is given by $M^*(M-1)^{N-1}$, where M is the number of states (i.e., 12—see level 2 in Fig. 1), and N is the order of the process. For example, there will be 132 states for the second-order cases, 1,452 states for the third-order cases, and 15,972 states for the fourth-order cases (see Table 3).

TABLE 3. The actual number of states in the refined state space for each usage group.

Usage group	Order of sequential dependency	Theoretical number of states for the given order of sequential dependency	Actual number of states that occur in the sample data
CL6	Third	1,452	803
CL7	Third	1,452	949
CL8	Fourth	15,972	5,126
CL13	Third	1,452	795
CL14	Third	1,452	924
CL18	Third	1,452	342

Recall that in a semi-Markov chain, the duration of stay T_i of state i has a general probability distribution (not necessarily an exponential distribution) with mean u_i. The transition rate at which the semi-Markov process leaves state i at time t is the reciprocal of the mean duration of stay in state i; that is, $-q_{ii}(t) = 1/u_i$. Because the refined state space is organized by combining zero-order states, the mean duration of stay in a state in the refined state space is simply the sum of the mean duration of stay of its components, namely the zero-order states.

Assume that the duration of stay of zero-order states {G, H, I, J, K, L, M, N, O, P, Q, R} in a usage group has a general probability distribution $\{f_G, f_H, f_I, f_J, f_K, f_L, f_M, f_N, f_O, f_P, f_Q, f_R\}$ with mean $\{v_G, v_H, v_I, v_J, v_K, v_L, v_M, v_N, v_O, v_P, v_Q, v_R\}$. For an Nth-order state i $= (s_1, s_2, \ldots S_N)$ in the refined state space, where $s_j \neq s_{j+1}$, , $j = 1, 2, \ldots, N$, its mean duration of stay is determined by

$$u_i = \sum_{j=1}^{N} E(T_{sj}) = \sum_{j=1}^{N} v_{sj}.$$

For example, the third-order state GHK (index access, search with retrievals, and screen display) has a mean duration of stay $v_G + v_H + v_K$. In this way, the cost of calculating the transition rate of each state for every usage group can be reduced to a minimum because only a limited number of calculations (the number of zero-order states) are required for every usage group. Table 4 shows the mean duration of stay of zero-order states for each of the six usage groups.

The mean duration of stay of a zero-order state is not the same as the average viewing time of Web pages associated with (mapped into) that state. Instead, it is the average aggregated viewing time of Web pages that are associated with that state and are requested by the user in succession. In other words, it represents the mean time the user stays in that state before entering another state (mean-time-to-leave or MTTL). Consider the sequence of user actions G-G-H-K-K that are reduced by our data transformation to G-H-K (index access, search with retrievals, and screen display).

[3] For the large values of degrees of freedom (>100), the critical values are approximated by the following equation (Law & Kelton, 1991):

$$\chi^2_{(K,1-\alpha)} \approx (K)\left\{1 - \frac{2}{9(K)} + Z_{1-\alpha}\sqrt{2/[9(K)]}\right\}^3$$

where $\chi^2_{(K,1-\alpha)}$ is the upper $1-\alpha$ critical point of the chi-square distribution with K degrees of freedom, and $Z_{1-\alpha}$ is the upper $1-\alpha$ critical point of the standard normal distribution.

TABLE 4. Mean duration of stay of zero-order states for searcher groups.

Zero-order state		CL6	CL7	CL8	CL13	CL14	CL18
G (index access)	Proportion of transitions into this state	11%	19%	14%	23%	22%	41%
	Proportion of time in this state	6%	17%	11%	20%	19%	58%
	Mean duration of stay (in seconds) in this state	41.79	38.66	38.31	34.94	43.94	42.63
H (search with retrievals)	Proportion of transitions into this state	10%	18%	16%	23%	12%	11%
	Proportion of time in this state	2%	7%	4%	8%	4%	6%
	Mean duration of stay (in seconds) in this state	20.20	16.47	13.76	13.47	18.58	19.67
I (search without retrievals)	Proportion of transitions into this state	2%	11%	10%	9%	12%	28%
	Proportion of time in this state	<1%	5%	4%	7%	6%	<1%
	Mean duration of stay (in seconds) in this state	32.61	23.90	22.63	32.08	26.32	34.73
J (modify search)	Proportion of transitions into this state	3%	12%	13%	7%	12%	3%
	Proportion of time in this state	1%	9%	7%	5%	8%	20%
	Mean duration of stay (in seconds) in this state	40.38	32.48	27.37	29.61	35.49	25.47
K (screen display)	Proportion of transitions into this state	31%	19%	22%	20%	14%	7%
	Proportion of time in this state	47%	34%	37%	33%	25%	2%
	Mean duration of stay (in seconds) in this state	117.23	83.23	81.06	75.52	88.89	55.65
L (record display)	Proportion of transitions into this state	20%	11%	15%	13%	8%	1%
	Proportion of time in this state	25%	19%	23%	20%	12%	4%
	Mean duration of stay (in seconds) in this state	99.04	86.69	75.91	87.61	79.18	50.29
M (change display format)	Proportion of transitions into this state	4%	1%	2%	1%	1%	<1%
	Proportion of time in this state	1%	<1%	<1%	<1%	<1%	<1%
	Mean duration of stay (in seconds) in this state	23.23	25.89	22.02	26.95	29.95	26.88
N (record processing)	Proportion of transitions into this state	12%	2%	4%	1%	4%	<1%
	Proportion of time in this state	12%	3%	7%	1%	7%	<1%
	Mean duration of stay (in seconds) in this state	87.30	70.58	83.51	61.58	97.37	78.50
O (error)	Proportion of transitions into this state	4%	4%	3%	1%	5%	2%
	Proportion of time in this state	<1%	<1%	<1%	<1%	<1%	<1%
	Mean duration of stay (in seconds) in this state	0.74	1.83	1.37	1.17	2.13	0.61
P (help)	Proportion of transitions into this state	<1%	1%	<1%	1%	8%	2%
	Proportion of time in this state	<1%	<1%	<1%	<1%	10%	3%
	Mean duration of stay (in seconds) in this state	49.92	50.51	47.74	47.68	76.53	60.01
Q (profile)	Proportion of transitions into this state	1%	<1%	<1%	<1%	<1%	1%
	Proportion of time in this state	<1%	<1%	<1%	<1%	<1%	<1%
	Mean duration of stay (in seconds) in this state	25.57	22.91	24.54	20.83	29.95	21.85
R (miscellaneous)	Proportion of transitions into this state	2%	1%	1%	2%	2%	3%
	Proportion of time in this state	1%	1%	2%	2%	2%	1%
	Mean duration of stay (in seconds) in this state	112.08	73.98	91.60	93.87	71.13	36.51

Now consider another raw session sequence G-G-G-H-K-N-K-K that is transformed to G-H-K-N-K. If these are the only two sessions, and we are calculating the duration of stay in state G, the time must be weighted by the original count of the number of states visited.

Because Table 4 delineates usage patterns on a rate basis, it can answer questions such as "How often does the user enter this particular state?" or "At what rate does the user leave this state?" It also indicates how the system has been used in terms of functionality. For instance, cluster CL6 represents knowledgeable and sophisticated use of the MELVYL system. The majority of transitions (73%[4]) and time (86%[5]) for this usage group are in states "search with retrievals (H)," "screen display (K)," "record display (L)," and "record processing (N)." Users in this group spend the most time in displaying retrieved items (117.23 seconds for screen display and 99.04 seconds for record display) among all usage groups. They seldom have a search failure (I) (i.e., a search without retrievals) and seldom make errors (O). This all demonstrates that they are skillful in searching and experienced with the system.

Clusters CL7 and CL8 represent unsophisticated usage, and highly interactive usage with good search results with the MELVYL system, respectively. Their patterns of use in all states are similar. Nearly 56% of the transitions for CL7 and 52% for CL8 occur in "index access activities (G)," "search with retrievals (H)," and "screen display (K)." They spend 19 and 21% of their time, respectively, in the state "screen display (K)" and 19 and 14%, respectively, in state "index access activities (G)." They are more likely to have a search failure: nearly one-tenth of both clusters' transitions are into "search without retrievals (I)." This explains why their search performance is inferior to that of searchers in cluster CL6. However, these two groups seldom request on-line help (P), or create or activate a personal profile (Q).

Clusters CL13 and CL14 represent other categories of usage patterns in the MELVYL system. Searchers in cluster

[4] For cluster CL6, the proportion of transitions into states H, K, L, and N are 10, 31, 20, and 12%, respectively. They sum to 73%.

[5] For cluster CL6, the proportion of time in states H, K, L, and N are 2, 47, 25, and 12%, respectively. They sum to 86%.

TABLE 5. The top 20 usage patterns for each group.

Usage group	CL6	CL7	CL8	CL13	CL14	CL18
Sequential dependency	Third-order	Third-order	Fourth-order	Third-order	Third-order	Third-order
Number of patterns	2,906	4,378	15,807	2,882	3,754	854
1	KLKL (15.40%)	KLKL (3.65%)	KLKLK (4.89%)	GHKL (6.13%)	GIGI (2.53%)	GIGI (19.11%)
2	LKLK (14.29%)	GHKL (3.57%)	LKLKL (3.99%)	GHKG (3.30%)	KLKL (2.00%)	IGIG (10.20%)
3	GHKL (3.09%)	HKLK (2.58%)	JIJIJ (1.60%)	HKGH (2.79%)	IGIG (1.96%)	GIGH (4.27%)
4	GHKN (2.81%)	LKLK (2.53%)	HKLKL (1.59%)	KGHK (2.57%)	GHKL (1.71%)	GOGI (4.02%)
5	HKNK (2.15%)	GIGI (2.52%)	IJIJI (1.33%)	GIGH (2.23%)	IJIJ (1.54%)	GIJI (3.42%)
6	HKLK (2.10%)	IGIG (2.12%)	GIGIG (1.18%)	HKLK (2.05%)	LKLK (1.51%)	IGHK (3.12%)
7	KNKN (1.75%)	JHKL (1.80%)	HKJHK (1.09%)	GHLG (2.04%)	JIJI (1.25%)	GIHK (2.34%)
8	KNMN (1.69%)	IJIJ (1.72%)	JHKLK (1.08%)	LGHK (1.93%)	GIJG (1.15%)	OGIG (1.84%)
9	NMNM (1.17%)	JIJI (1.56%)	GHKLK (1.07%)	KLKL (1.89%)	HKLK (1.14%)	GOGH (1.83%)
10	MNMN (1.14%)	GHKG (1.43%)	HKGHK (1.00%)	IGHK (1.86%)	GHKG (1.10%)	GHKG (1.77%)
11	NKLK (1.13%)	GIGH (1.38%)	GHKGH (0.99%)	HKLG (1.75%)	GIGH (1.04%)	GIJG (1.53%)
12	KLKN (1.13%)	KGHK (1.31%)	JHKJH (0.91%)	HLGH (1.60%)	IJGP (1.02%)	IGIJ (1.32%)
13	KGHK (0.94%)	HJHK (1.28%)	IGIGI (0.79%)	GIGI (1.57%)	GIJI (0.91%)	OGHK (1.31%)
14	KLNK (0.92%)	HKGH (1.25%)	GIGHK (0.78%)	HKHK (1.50%)	IGHK (0.89%)	GIGR (1.28%)
15	HKLN (0.87%)	KJHK (1.24%)	JIJHK (0.64%)	HKJH (1.43%)	JHKL (0.87%)	HKGI (1.20%)
16	NKNM (0.86%)	GHJH (1.23%)	KJHKJ (0.62%)	KLGH (1.41%)	KGHK (0.82%)	IJIG (1.04%)
17	KOKN (0.81%)	HKJH (1.17%)	KLJHK (0.60%)	KJHK (1.35%)	HKGH (0.81%)	IJIJ (1.04%)
18	OKOK (0.73%)	IGHK (1.13%)	KGHKG (0.60%)	GHKJ (1.24%)	PGHK (0.74%)	IGIR (0.93%)
19	GHKG (0.70%)	HKLJ (1.04%)	GHJHK (0.56%)	IGIG (1.21%)	GHKJ (0.71%)	PGIG (0.89%)
20	KOKO (0.67%)	GHKJ (1.00%)	KJHKL (0.55%)	IJIJ (1.17%)	PGPG (0.70%)	GHKL (0.86%)
Cumulative percentage of transitions explained	54.45%	35.61%	25.97%	41.13%	24.48%	63.41%

Note: Values in parentheses indicate the percentage of the entire transitions for that cluster.

CL13 employ known-item searches exclusively, whereas those in cluster CL14 are new to the MELVYL system and request considerable on-line help (P). Approximately 23% of their transitions and 20% of their time are in state "index access activities (G)." They seldom create or activate a personal profile (Q) and seldom change display format (M). However, their patterns of use in states related to search performance are different. For searchers in CL13, nearly 56%[6] of their transitions are into "search with retrievals (H)," "screen display (K)," and "record display (L)," which are positively related to search performance. Another 9% of their transitions are into "search without retrievals (I)" and 7% into "modify search (J)," which are negatively related to search performance. On the other hand, searchers in cluster CL14 have less of their transitions (29% altogether) into "search without retrievals (I)," "modify search (J)," and "errors (O)" than into "search with retrievals (H)," "screen display (K)," and "record display (L)," which totals 34%. Nevertheless, the use of "record processing (N)" indicates that searchers in cluster CL14 are more likely to retrieve relevant items than those in cluster CL13. This implies that either most searchers in cluster CL13 are unaware of the PMDS (print, mail, download, and save retrieved items) functions or they simply do not want to use them.

Finally, searchers in cluster CL18 are relatively unsuccessful in their use of the MELVYL system. They spend the majority of their time (58%) in state "index access activities (G)," which accounts for 41% of their total transitions in the system. Despite heavy index access activities, they are far more likely to have a search failure (I) than a search with retrievals (H). As such, they seldom display retrieved items (K and L), or try to improve retrievals by modifying searches (J). Interestingly, they make errors at one of the lowest rates of all clusters. They spend 3% of their time in help activities (P) and stay there for about 1 minute before entering another state.

State-Transition Patterns Analysis

Every usage group in the MELVYL system has at least a third-order sequential dependency. Because of limited space, it would be impossible to show all the transition rates as well as transition probability matrices. Therefore, we will consider the top twenty state-transition patterns based on frequency of occurrence in sessions in a usage group instead (see Table 5). In each cell, the usage pattern of *N*th-order sequential dependency should be interpreted as a transition from an *N*th-order state into another *N*th-order state, and the value in parentheses is the proportion of *N*-step transitions that are from the former into the latter. For instance, the most frequent usage pattern for users in cluster CL6 (knowledgeable and sophisticated usage is KLKL (screen display, record display, screen display, and record display), which accounts for 15.40% of their transitions (see Table 5). It denotes the transition from third-order state KLK (screen

[6] For cluster CL13, the proportion of transitions into states H, K, and L are 23, 20, and 13%, respectively. They sum to 56%.

display, record display, and screen display) to third-order state LKL (record display, screen display, and record display). For simplicity, it can be interpreted as a transition into "record display (L)," given that the process or the user has been in state KLK (screen display, record display, and screen display) previously. These two display-related patterns account for 30% of the transitions in this group.

Moreover, 12 out of the top 20 usage patterns for this group contain "record processing (N)." None of the top 20 usage patterns contain "search without retrievals (I)" or "modify search activities (J)." The most frequent search patterns (i.e., patterns that contain search activities) are GHKL (index access, search with retrievals, screen display, and record display) and GHKN (index access, search with retrievals, screen display, and record processing). It seems that most of the queries submitted retrieve relevant items so that one-third (at least) of their transitions are into display activities (K and L) and another one-sixth (at least) of their transitions are into record processing (N). Although they are prone to making errors (O) when they employ screen displays (K) (patterns OKOK and KOKO), they rarely resort to on-line help (P): none of the top 20 usage patterns contain "help activities (P)."

Cluster CL7 represents relatively unsophisticated usage of the MELVYL system. The top four usage patterns for this group are mainly related to display activities (K and L). However, the most frequent search patterns for this group are the same as for CL6: GHKL. Due to a probable lack of experience and skills, users in this group are more likely to encounter a search failure (I). Although some of the queries are modified (J) to improve retrievals (9 out of the 20 patterns listed in Table 5), they usually do not work: in this group "search modification (J)" is most likely followed by "search without retrievals (I)," as patterns IJIJ (search without retrievals, modify search, search without retrievals, and modify search) and JIJI (modify search, search without retrievals, modify search, and search without retrievals) suggest. Even so, they rarely request on-line help (P).

Cluster CL8 consists of highly interactive usage of the MELVYL system with good search results. A usage pattern in this group represents the transition from a fourth-order state into another fourth-order state. It has the greatest number of different usage patterns (15,807) among all searcher groups. Again, the most frequent pattern is KLKLK (screen display, record display, screen display, record display, and screen display), which denotes a series of display activities, followed by pattern LKLKL.

Compared to searchers in cluster CL7, those in cluster CL8 are less likely to encounter a search failure (I) when they perform index access activities (G) (patterns GIGIG and IGIGI), but are more likely to have an invalid search modification (J), which can be deduced from the fact that "search modification (J)" is most frequently followed by "search without retrievals (I)," as patterns IJIJI and JIJIJ suggest. In fact, half of the top 20 usage patterns for cluster CL8 contain "search modifications (J)." In addition, users in this group seldom make errors (O) or request on-line help (P).

Cluster CL13 consists of searchers who employ known-item searches exclusively. The most frequent search pattern in this group is GHKL (index access, search with retrievals, screen display, and record display). However, it only accounts for 6.13% of the transitions in this group. Unlike the preceding searcher groups, in which the top four usage patterns are largely related to display activities (K and L), cluster CL13 has more patterns on the top of the list that reveal successive search activities. For example, the second through fifth most frequent usage patterns are GHKG (index access, search with retrievals, screen display, and index access), HKGH (search with retrievals, screen display, index access, and search with retrievals), KGHK (screen display, index access, search with retrievals, and screen display), and GIGH (index access, search without retrievals, index access, and search with retrievals), respectively. Because searchers in this group have a poor search performance, these patterns suggest that most of the queries submitted in this group retrieve a few relevant items or nothing, and thus another search follows immediately. Nevertheless, searchers in this group seldom consider search modification (J) (in 4 out of the 20 usage patterns listed in Table 5) as a way to improve retrievals. In addition, they seldom make errors (O) or request on-line help (P).

Finally, clusters CL14 and CL18 represent help-intensive searching and relatively unsuccessful usage, respectively. The usage patterns are different from those of the preceding search groups. The most frequent usage pattern in these two groups is GIGI (index access, search without retrievals, index access, and search without retrievals). Due to a lack of experience and skills, they are most likely to encounter a search failure (I). Searchers in cluster CL14 perform a little better than those in cluster CL18: two of the top five usage patterns in cluster CL14 reveal an effective search pattern, that is, KLKL (screen display, record display, screen display, and record display) and GHKL (index access, search with retrievals, screen display, and record display). In cluster CL18, all the top five usage patterns suggest a poor search pattern, such as GIGH (index access, search without retrievals, index access, and search with retrievals), GOGI (index access, error, index access, and search without retrievals), and GIJI (index access, search without retrievals, modify search, and search without retrievals). In fact, nearly all (15) of the top 20 usage patterns for cluster CL18 contain "search without retrievals (I)."

Searchers in both groups (clusters CL14 and CL18) also show their intention to improve retrievals by modifying searches (J). Most of their efforts are in vain: "search modification (J)" is most likely followed by "search without retrievals (I)," as patterns IJIJ, JIJI, GIJI, and IJIG suggest. Searchers in cluster CL14 tend to seek help (P) from the system when they are performing index access activities (G), as patterns IJGP (search without retrievals, modify search, index access, and help request) and PGPG (help request, index access, help request, and index access) indicate. Searchers in cluster CL18 seldom request on-line help (P), although they are prone to making errors (O) when they

TABLE 6. Results of testing differences in one-step transition probability matrices within usage groups.

Usage group	CL7	CL8	CL13	CL14	CL18
CL6	606	287	1,186	688	7,209
	(146, df = 109)	(146, df = 109)	(146, df = 109)	(146, df = 109)	(146, df = 109)
CL7		156	314	304	7,266
		(150, df = 112)	(150, df = 112)	(150, df = 112)	(150, df = 112)
CL8			537	229	7,530
			(156, df = 117)	(156, df = 117)	(156, df = 117)
CL13				633	4,274
				(151, df = 113)	(151, df = 113)
CL14					7,109
					(147, df = 110)

are performing index access activities (G), as patterns GOGI (index access, error, index access, and search without retrievals) and GOGH (index access, error, index access, and search with retrievals) indicate.

Testing the Differences in Usage Patterns Between Groups

Due to limited space, it would be impractical to compare all usage patterns in a qualitative manner. Consequently, a quantitative evaluation of the similarity or dissimilarity of usage patterns between usage groups was conducted. A chi-square test was performed to test statistically the significance of differences in usage patterns between groups. Because most of the groups comply with third-order sequential dependency, the chi-square test was chiefly performed on the three-step transition probability matrices. In this study a three-step transition probability matrix is equal to a one-step transition probability matrix of third-order sequential dependency. The chi-square test was performed on one-step and two-step transition probability matrices as well to see whether or not there are differences in transitions of varying levels between usage groups.

The test proceeds as follows: Let A and B be two samples to be compared. Then define f_{ij}^{A} (i,j = 1,2,. . .,K) as a transition frequency matrix for sample A, and p_{ij}^{A} as a transition probability matrix for sample A. Also define $f_{ij}^{A(B)}$ as the number of transitions from state i into state j. Then $p_{ij}^{A(B)}$ is the transition probability from state i into state j, and K is the number of (assumed) states in the state space:

$$p_{ij}^{A(B)} = \frac{f_{ij}^{A(B)}}{\sum_{i=1}^{K} f_{ij}^{A(B)}}$$

If A and B are similar to each other, then f_{ij}^{B} should be close to the expected number of transition from state i into state j in B, that is,

$$E(f_{ij}^{B}) = \sum_{l=1}^{K} f_{il}^{B} * \frac{f_{ij}^{A}}{\sum_{l=1}^{K} f_{il}^{A}},$$

for

$$p_{ij}^{B} = \frac{f_{ij}^{B}}{\sum_{l=1}^{K} f_{il}^{B}}$$

being close to

$$p_{ij}^{A} = \frac{f_{ij}^{A}}{\sum_{l=1}^{K} f_{il}^{A}}.$$

In that case,

$$C = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{[f_{ij}^{B} - E(f_{ij}^{B})]^2}{E(f_{ij}^{B})}$$

will approximate a chi-square distribution with $K^2 - N_1 - N_2$ degrees of freedom, where N_1 is the number of f_{ij}^{B} that can be determined by others (or N_1 is the number of actual states in B) and N_2 is the number of impossible transitions in B.[7] Therefore, the null hypothesis that there is no difference between transition probability matrices A and B is accepted if C is less than the critical value $C_{a}^{K^2-N_1-N_2}$ at α (= 0.01) level of significance. Tables 6-8 display the test results of the chi-square test on one-step through three-step transition probability matrices. In each cell, the value in the first row is the test value C, and the first value in parentheses is the critical value given the degrees of freedom. Again, for large values of degrees of freedom (>100), the critical values are approximated by the equation in Law and Kelton (1991).[8]

As can be seen, the six usage groups have distinct transition behaviors. The null hypothesis that there is no difference in state transition probability matrices within usage groups was rejected in almost all cases.

[7] They are called structure zeros by Everitt (1992).

[8] See Footnote 3.

TABLE 7. Results of testing differences in two-step transition probability matrices within usage groups.

Usage group	CL7	CL8	CL13	CL14	CL18
CL6	7,908 (791, df = 701)	7,456 (791, df = 701)	9,381 (791, df = 701)	7,544 (791, df = 701)	7,651 (791, df = 701)
CL7		2,717 (968, df = 868)	3,088 (968, df = 868)	3,969 (968, df = 868)	6,134 (968, df = 868)
CL8			3,931 (1,167, df = 947)	3,715 (1,167, df = 947)	6,507 (1,167, df = 947)
CL13				11,908 (796, df = 706)	4,803 (796, df = 706)
CL14					7,090 (914, df = 817)

TABLE 8. Results of testing differences in three-step transition probability matrices within usage groups.

Usage group	CL7	CL8	CL13	CL14	CL18
CL6	62,118 (2,422,df=2103)	67,380 (2,422, df = 2103)	37,960 (2,422, df = 2103)	39,745 (2,422, df = 2103)	18,360 (2,422, df = 2103)
CL7		31,347 (3,832, df = 3429)	16,325 (3,832, df = 3429)	30,535 (3,832, df = 3429)	14,460 (3,832, df = 3429)
CL8			18,172 (4,652, df = 4207)	30,372 (4,652, df = 4207)	14,725 (4,652, df = 4207)
CL13				46,966 (2,405, df = 2087)	12,571 (2,405, df = 2087)
CL14					13,969 (3,198, df = 2830)

Third- and fourth-order sequential dependencies were found in this analysis. There is no guarantee that other Web-based information system users will exhibit the same order, perhaps because of the definition of the state space (Fig. 1 in our case), or because of a different purpose of the information system (a library catalog in our study). What is important is the methodology used to derive the dependencies. This methodology can be applied to other information systems.

Knowing the order of sequential dependencies is obviously beneficial. For example, we can predict a user's next move based on his or her past three steps if we already know that the behavior exhibits a third-order dependency. In some commercial customer relationship management software, the knowledge of such user behavior aids the design of a customized Web page or help screen.

The order of sequential dependency also reveals how users interact with the system. The most common search pattern for users in cluster CL18 (relatively unsuccessful searchers) is to perform a search (state G from Fig. 1), obtain no results (state I), perform another search (state G), and again obtain no results (state I). Are these two search failures related? Are they a search on the same topic? Is poor system design the cause of the search failure (if this is a failure)? Knowing the correct order of sequential dependency gives us considerably more power in studing user interaction.

Conclusion

In this article, a continuous-time stochastic process (a semi-Markov chain) was applied to modeling user state-transition behavior in a Web-based information system. The stochastic model was demonstrated using a sample of approximately 127,000 user sessions, each of which contains one or more search queries with the University of California's Web version of the MELVYL on-line catalog system. Using clustering techniques, user sessions were divided into six usage groups of homogeneous behavior: knowledgeable and sophisticated usage, unsophisticated usage, highly interactive usage with good search results, known-item searching, help-intensive searching, and relatively unsuccessful usage. A three-layer hierarchical taxonomy of system Web pages was proposed to provide a mapping between system Web pages and states that are employed in stochastic modeling. Later, user sessions for every usage group were transformed into a sequence of states. To improve the quality of the derived user state-transition patterns, multiple occurrences of a state in succession were replaced with a single occurrence of the state. A chi-square test was performed to determine the order of sequential dependency in state-transition activities for each usage group. The test results indicated that all of the usage groups but one have a third-order sequential dependency. The sole exception has a fourth-order sequential dependency.

TABLE 9. Summary of state-transition patterns for usage groups.

Usage group	Type of usage	Order of sequential dependency found in group	Most frequent pattern	Pattern features
CL6	Knowledgeable and sophisticated	Third	screen display record display screen display record display	More than one-third of the transitions are into display activities. At least one-sixth of the transitions are into record processing. Tend to make errors while in screen display. Seldom resort to online help.
CL7	Unsophisticated	Third	screen display record display screen display record display	More likely to encounter a search failure while in index access activities. Search modification is most likely followed by a search failure. Rarely request on-line help.
CL8	Highly interactive with good search results	Fourth	screen display record display screen display record display screen display	More likely to encounter a search failure while in search modification. Half of the top twenty patterns contain search modification. Willing to follow the "advice" of the system in modifying searches. Seldom make errors or request online help.
CL13	Known-item searching	Third	index access search with retrievals screen display record display	Tend to search successively. Seldom employ search modification. Rarely make errors or request online help.
CL14	Help-intensive searching	Third	index access search without retrievals index access search without retrievals	Tend to encounter a search failure successively while in index access activities. Most search modifications are ineffective. Tend to seek on-line help while in index access activities.
CL18	Relatively unsuccessful	Third	index access search without retrievals index access search without retrievals	Tend to encounter a search failure successively. Nearly all of the top 20 patterns contain a search failure. Most search modifications are ineffective. Seldom request on-line help. Tend to make errors while in index access activities.

Usage patterns in the form of transition rates and transition probability matrices were derived and interpreted qualitatively for each usage group found in the system. Because the duration of stay in a state was factored into transition rates in the semi-Markov models, the derived usage patterns can answer questions such as "How often does the user enter this particular state?" or "At what rate does the user leave this state?" Factoring in the duration of stay also gives indications of how the system has been used, which is not possible with discrete-time models. Further a chi-square test revealed that the differences in transition patterns between usage groups were statistically significant. The results are summarized in Table 9.

The methodology presented in this article has many applications. For example, it can help in the design of an advanced Web site that provides real-time personalization and/or interaction for its users—a common feature in customer relationship management. Without any demographic information, the system can classify a user into a usage group and thereby predict his or her next activity. In other words, users are identified not by who they are but by how they use the system. Because a user may behave differently in different periods and contexts in a session, the above strategy seems to be more flexible and accurate in capturing user behavior than a system that simply maintains a single profile of a user.

References

Anderson, T., & Goodman, L. (1963). Statistical inference about Markov chains. Readings in Mathematical Psychology. (Vol. 2, pp. 241–262). New York: Wiley.

Chapman, J. (1981). A state transition analysis of online information-seeking behavior. Journal of the American Society for Information Science, 32, 325–333.

Chen, H.M. (2000). An analytical approach to deriving usage patterns in a Web-based information system. Ph.D. dissertation. School of Information Management and Systems, University of California, Berkeley.

Chen, H.M., & Cooper, M.D. (2001). Using clustering techniques to detect usage patterns in a Web-based information system. Journal of the American Society for Information Science and Technology, 52, 888–904.

Cooper, M.D. (1982). Usage patterns of an online search system. Journal of the American Society for Information Science, 34, 343–349.

Cooper, M.D. (1983). Response time variations in an online search system. Journal of the American Society for Information Science, 34, 374–380.

Cooper, M.D. (1998). Design considerations in instrumenting and monitoring Web-based information retrieval systems. Journal of the American Society for Information Science, 49, 903–919.

Cooper, M.D. (2001a). Predicting the relevance of a library catalog search. Journal of the American Society for Information Science and Technology, 52, 813–827.

Cooper, M.D. (2001b). Usage patterns of a Web-based library catalog. Journal of the American Society for Information Science and Technology, 52, 137–148.

Everitt, B. (1992). The analysis of contingency tables, 2nd ed. London: Chapman and Hall.

Fenichel, C. (1981). Online searching: Measures that discriminate among users with different types of experiences. Journal of the American Society for Information Science, January, 23–32.

Law, A., & Kelton, W. (1991). Simulation modeling & analysis, 2nd ed. New York: McGraw-Hill.

Penniman, W. (1975). A stochastic process analysis of on-line user behavior. Proceedings of the 38th Annual Meeting of the American Society for Information Science, Boston, MA; C.W. Husbands, R.L. Tighe (Eds.). Washington, DC: American Society for Information Science, pp. 147–148.

Penniman, W. (1982). Modeling and evaluation of on-line user behavior. Proceedings of the 45th ASIS Annual Meeting, Columbus, OH; A.E. Petrarca, C.I. Taylor, & RS. Kohn (Eds.). White Plains, NY: Knowledge Industry, pp. 231–235.

Qiu, L. (1993). Markov models of search state patterns in a hypertext information retrieval system. Journal of the American Society for Information Science, 44, 413–427.

Ross, S. (1996). Stochastic processes, 2nd ed. New York: John Wiley & Sons.

Wolff, R. (1989). Stochastic modeling and the theory of queues. Englewood Cliffs, NJ: Prentice-Hall.