Using Clustering Techniques to Detect 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; E-mail: cooper@socrates.berkeley.edu

Different users of a Web-based information system will have different goals and different ways of performing their work. This article explores the possibility that we can automatically detect usage patterns without demographic information about the individuals. First, a set of 47 variables was defined that can be used to characterize a user session. The values of these variables were computed for approximately 257,000 sessions. Second, principal component analysis was employed to reduce the dimensions of the original data set. Third, a two-stage, hybrid clustering method was proposed to categorize sessions into groups. Finally, an external criteria-based test of cluster validity was performed to verify the validity of the resulting usage groups (clusters). The proposed methodology was demonstrated and tested for validity using two independent samples of user sessions drawn from the transaction logs of the University of California's MELVYL® on-line library catalog system (www.melvyl.ucop.edu). The results indicate that there were six distinct categories of use in the MELVYL system: knowledgeable and sophisticated use, unsophisticated use, highly interactive use with good search performance, known-item searching, help-intensive searching, and relatively unsuccessful use. Their characteristics were interpreted and compared qualitatively. The analysis shows that each group had distinct patterns of use of the system, which justifies the methodology employed in this study.

Introduction

The purpose of this study was to develop an analytical approach to detecting groups with homogeneous usage patterns in a Web-based information system. It is our belief that dividing use into groups with common features aids in the study of usage patterns. The analysis used principal component analysis for data reduction, and cluster analysis for categorizing usage into groups. The methodology was demonstrated and tested for validity using two independent samples of user sessions from the transaction logs of the Web-based University of California's (UC) MELVYL® on-line library catalog system (www.melvyl.ucop.edu). All of the sessions in the samples involve the user conducting one or more searches of a bibliographic database. The samples exclude tourists—users who log onto the system to look around. This article also attempted a qualitative interpretation of the characteristics of the groups found through the clustering methods.

Many people could benefit from a knowledge of usage patterns: managers could see how their systems are used; researchers and designers could take this information and improve the interfaces; and system support personnel could improve not only their answers to users' questions, but also improve automated assistance procedures and documentation.

Usage patterns can be studied in a laboratory (an experimental study) or remotely (an empirical study). An experimental study allows controlled variables and careful monitoring of all user interaction; an empirical study can take place at a distance from the user-system interaction (Peters, 1998). With either type of study, data on user–system interaction can be collected in many ways, such as surveys, interviews, questionnaires, videotaping, verbal protocols, and transaction logs. Subjects in an experimental study are often selected based on social variables (education level, status, experience with the information system, gender, age, socioeconomic level, and other demographic information) and/or psychological variables (intelligence, cognitive style, learning style, and symbolic reasoning skill) (Yuan & Meadow, 1999). Although an experimental study would be relatively easy to arrange, and is therefore appealing, it is unlikely to be conducted for a Web-based information system. It would be very difficult to get user demographic information except through an on-line survey, but asking users to reveal personal information over the Web may be perceived by some as unethical, and by others as a violation

Received November 8, 2000; revised February 1, 2001, accepted February 13, 2001

of privacy laws. Thus, for a Web-based information system, an empirical study seems to be a better approach.

This article begins with a review of previous studies of usage patterns in information systems. It then discusses the variables used to characterize a search session, characteristics of the data used in the study, principal components methodology, clustering methodology, tests of clustering validity, and a qualitative review of the clusters obtained in the research.

Previous Research

Previous studies on usage patterns in an information system classified users into categories based on social variables and/or psychological variables. The behaviors of different user groups were compared and (usually) tested with statistical procedures. For example, Penniman (1975) studied on-line user transition behavior. He analyzed 231 hours of connect time, including 934 interactive sessions with the BASIS on-line retrieval system. Users were divided into two categories, novice and experienced, based upon number of on-line sessions. Penniman defined 11 states (or activities) and then merged them into four categories: search by index, logic formation, document display, and other. He found significant differences in the proportion and pattern of functions used in sessions.

To examine the extent to which the amount of experience and type of training affect the use of an on-line bibliographic retrieval system, Chapman (1981) conducted an experimental project, Individualized Instruction for Data Access (IIDA), using Lockheed's DIALOG system. The experimental subjects were divided into six groups based on individual searching experience and the type of training received (by IIDA or professional searchers).

The results showed that the level of searching experience, the type of training, and their interaction did affect the use of an on-line information system. The subjects who were trained either by IIDA or by professional searchers tended to follow exactly the advice of the trainer. This observation supports the finding that searchers were conservative in using commands; that is, they tended to use "known, taught" commands to search for information on-line. The two groups of inexperienced searchers trained by IIDA had similar patterns of searching. However, searchers who received another kind of training method seemed to have diverse usage patterns. The most significant difference among subject groups was observed between the experienced, professionally trained searchers, and the inexperienced, IIDA-trained searchers. Very little significant difference was observed between the inexperienced, professionally trained searchers and the inexperienced, IIDA-trained searchers. This study suggests that the level of searching experience has a greater impact on user information-seeking behavior than the type of training.

In the same year, Fenichel (1981) did an experimental study with the DIALOG system to find factors that would differentiate users with different types of experiences. Seventy-two subjects were divided into five groups: very experienced with the system and with ERIC (Educational Resources Information Center) database experience, very experienced with the system but without ERIC experience, moderately experienced with ERIC experience, moderately experienced without ERIC experience, and novice. Four kinds of variables (environmental, searcher, search process, and search outcome) were collected for a statistical test of differences. Her research had two objectives: (1) to identify the variables that differentiate among searches by users with different amounts of overall experience with the system, and (2) to identify the variables that differentiate between searches by users who are familiar with the database and searches by users who are unfamiliar with the database.

Statistical tests such as one-way analysis of variance (ANOVA) and the Student–Newman–Keuls test revealed that the searching behavior of users with various levels of experience differed only to a limited extent. The only significant difference was that the novices searched more slowly and made more errors than the experienced subjects did. The novices had a longer total connection time, but they issued fewer commands per minute. It seemed that the novices took more time thinking, learning, and getting familiar with the system. However, the novices were not inferior to the experienced subjects in terms of performance. The very experienced users who also had ERIC database experience tended to have higher values for their search effort variables (e.g., the total number of commands used, the total number of descriptors searched, and connection time) than other experienced users did. There was only slight evidence supporting the assumption that ERIC experience made a difference. The results showed that the subjects with ERIC experience used more thesaurus terms than those without ERIC experience.

Fenichel's major contribution is that she tried to isolate the factors that differentiate users with various levels of searching experience. However, analysis of the search transcripts showed that the differences among various user groups were not large. Rather, the results indicated unexpectedly large individual differences in search behavior. Even for subjects in the same group, the variables tested varied by a factor of 10 or more. This suggests that either the variables chosen for comparison were not appropriate, or classifying users by experience alone (the predominant factor in Fenichel's study) is not optimal. A better way to form user groups, so that users of a group have similar behavior, would be desirable.

Later, Penniman (1982) analyzed transaction logs of the MEDLINE database and compared results to previous studies by Penniman (1975), Chapman (1981), and Fenichel (1981), to determine what usage patterns (sequences of activities) are universal and what the influence of system usage experience is on user behavior. A total of 1,519 user sessions, each having four or more transactions, were collected to gather statistics for each user. Penniman divided

users into three categories by the following criterion: *frequent users*—20–23 hours of use on at least 25 different days; *moderate users*—6–8 hours of use on at least 8 but not more than 20 different days; *infrequent users*—0.5–4 hours of use on at least four different days with at least five but not more than nine sessions.

The results indicated that the three user groups behave differently in many ways. For example, the use of term, advanced term, Boolean, and display commands showed distinguishing patterns among the three classes of users. Infrequent users tended to search more slowly (agrees with Fenichel), but they did not make more errors than the other users (disagrees with Fenichel). Frequent users used more commands, connect time (agrees with Fenichel), and advanced search strategies (agrees with Chapman, but disagrees with Fenichel) than infrequent users. Moreover, experienced users did not frequently employ long sequences of advanced select commands (disagrees with Chapman), and inexperienced users frequently employed long sequences of display commands (agrees with Chapman).

Qiu (1993) is one of the few researchers in recent years to develop mathematical models for user behavior in a hypertext information system. She formed a variety of user groups based on the following factors: gender (male/female), search experiences (low/high), search task (general/specific), and the user's academic background (social science/engineering). Search-state patterns were compared and statistically tested for differences across user groups.

The results revealed that there were significant differences between user groups on all factors. The most significant difference found is between the two task groups (general/specific). For a specific task, users tended to adopt a structured search pattern (i.e., a combination of keyword searching, displaying title, and displaying article); for a general task, users tended to perform random browsing. Highly experienced users were seen to be more confident than novices in using hypertext features such as traversing in a nonlinear manner, while the novices made a lot of linear or structured movements. Male users were more likely than females to adopt a structured search strategy, but female users were more likely than males to continuously revisit previous nodes and perform linear browsing. Differences between users with a technical background and those with no technical background were significant but small.

From the previous studies, we can draw the following conclusions: (1) users behave quite differently; (2) users can be divided into groups based on observable factors; (3) categorizing users by experience, tasks, or demographic information is effective but by no means exhaustive; and (4) transaction logs are effective in capturing user behavior unobtrusively.

But at the same time, a number of problems are present in these studies: (1) research results may be unreliable due to small samples; (2) it is hard to represent a real environment in an experimental study; (3) subjects who participated in an experiment are usually not real users with real information needs; (4) user groups are formed in a subjective manner, not by what users did, but by who they are (e.g., male or female) or what they have (e.g., level of experience, education, etc.).

To attempt to correct these and other problems, a new approach to studying usage patterns in a Web-based information system is introduced. It categorizes them into groups based on operational features of a session.

Although our research focuses on a Web-based catalog, the search and usage results reported here are more indicative of patterns of usage of library catalog rather than of a Web-based search engine. The library catalog has a much more structured database with precisely defined field names and contents. Two articles that reveal the character of the more general Web-searching activities (noncatalog) are Jansen, Spink, and Saracevic (2000), and Spink, Wolfram, Jansen, and Saracevic (2001). However, the methodology proposed in this research can be equally applied to the study of the more general Web-searching activities (noncatalog).

Session Characteristics

The goal of this research was to ascertain whether there were different patterns of system usage. The approach was to develop a set of variables that can be used to characterize a session. Table 1 lists the base variables used to characterize a session, and Table 2 lists the 47 variables derived from the base variables in Table 1. The variables are similar to the sets found in Chen (2000), and Cooper & Chen (in press). Note that none of them contain any demographic information about the user.

In the model that we used to characterize sessions, a session may have presearch, search, display, help, error, and other actions in it. Presearch activities occur before the first search statement is executed and may include invoking or establishing a profile of search preferences (e.g., language of material to be retrieved, library location to search, e-mail address to which to send retrieved material) or looking at screens describing the system. Once the user begins a search, the normal assortment of options is available, such as author, title, subject, power, or Boolean. A power search feature lets the user search a variety of indexes and use Boolean operators at the same time. Users can choose to find more citations like the ones they just retrieved, or fewer. If the user wishes, retrieved citations can be printed, downloaded, mailed, or saved to a list (PMDS). Taking this action is an indicator that the user has found something worth keeping—for whatever reason. The transaction log records these actions along with variables like time spent on tasks, number of items retrieved, and number of Web pages viewed. The analysis program took the transaction log records for a session and computed the values of the variables in Tables 1 and 2. It also computed certain measures of performance, such as estimated precision for searches with retrievals (PR). A measure of relevance (LR) was also

TABLE 1. Base variables used to characterize a Web-based catalog session.

Symbol	Description	Formula
	Session variables	
NT	The total number of items retrieved in a session	
SR	The number of searches in a session for which the user retrieved at least one item (searches with retrievals)	
TS	The time in seconds the user spent before beginning the first search of a session (presearch)	
PL	The number of Web pages the user accessed before beginning the first search of a session (presearch)	
	Search variables	
SA	The number of author searches in a session	
ST	The number of title searches in a session	
SU	The number of subject searches in a session	
SW	The number of power searches (i.e., with all search options) in a session	
SB	The number of searches in a session that use Boolean operators	
MD	The number of searches that are followed by other searches that modify the original search in any way (search modifications)	
CS	The number of times the user changed the search index (e.g., author, title, subject heading) used during a session	
FL	The number of times a user requested the system to find fewer citations than the current search retrieves	
FR	The number of times a user requested the system to find more citations like the current one after the user has performed a title search, an author search, or a subject search	
FH	The number of times a user requested the system to find more citations like the current one after the user has performed a power search	
SK	The number of known-item searches in a session	SK = SA + ST
SM	The number of search modifications in a session	SM = CS + FR + FH + FL
MM	The number of times a user requested the system to find more citations like the current one	MM = FR + FH
	Display variables	
SO	The number of searches in a session that are followed by the display of a single record (a single-record display screen)	
SP	The number of searches in a session that are followed by the display of multiple records (a multiple-record display screen)	
DT	The total time (seconds) in a session spent on record display activities	
SV	The total viewing time (seconds) for single-record display screens during a session	
MV	The total viewing time (seconds) for multiple-record display screens during a session	
	Relevance variables	
NL	The number of searches in a session that are followed by a PMDS activity	
	Error variables	
SE	The number of searches that have system-detected errors in them	
	Help variables	
SH	The total time (seconds) spent in help activities during a session	

computed. The reader is referred to Cooper & Chen (in press) for an extensive discussion of this variable.

The session data was collected from transaction logs of the University of California's Web-based MELVYL® online catalog system. The MELVYL system is part of the California Digital Library (CDL) and provides bibliographic access to the university-wide library collection, and also to a variety of databases, such as BIOSIS, INSPEC, MEDLINE, and so on, for the University of California (UC) community. It can be accessed at www.melvyl.ucop.edu. Each day, thousands of users visit the MELVYL Web site. The broad functionality and the rich user population of the MELVYL system provide an ideal platform for studying user behavior patterns.

Two samples of sessions were selected for analysis: one for initial data analysis, and the other for validation. The first sample, called Sample 1, consists of 126,925 sessions collected from February 15, 1998, to March 14, 1998. The second sample, called Sample 2, comprises 130,902 sessions collected from April 12, 1998, to May 9, 1998. Both samples are from the same Spring semester. Because the MELVYL system serves primarily an academic population, its usage has peaks in the fall and spring and periods of relatively low usage during the summer and academic vacations (Berger, 1994; Cooper, 2001). Despite the varied use of the MELVYL system within an academic year, its user population is unlikely to change dramatically during a semester.

Characteristics of the Samples

A user session with the MELVYL system was represented as a vector of the 47 variables shown in Table 2.

TABLE 2. Derived variables used to characterize a session.

Symbol	Description	Formula
	Session variables	
SL	The length of a session in seconds	
NP	The number of Web pages requested during a session	
TP	The number of different Web pages viewed during a session	
TD	The number of different databases used during a session	
TI	The number of different indexes used during a session	
SQ	The number of searches performed during a session	
PF	An indicator of whether the user activated a stored profile of his or her preferences (e.g., language of retrieved items, e-mail address for mailing materials, campus on which to locate materials)	
VT	The average time in seconds between Web page requests.	VT = SL/NP
SD	The average search length in seconds (excludes presearch activities)	SD = (SL − TS)/SQ
HT	The average number of items retrieved per search when the search retrieved one or more items (average number of hits)	HT = NT/SR
PP	The proportion of the total session time spent before the first search took place (presearch time)	PP = TS/SL
PS	The proportion of all Web pages the user accessed before beginning the first search of the session (presearch)	PS = PL/NP
	Search variables	
SI	The average number of Web pages requested per search	SI = (NP − PL)/SQ
RR	The proportion of all searches that result in the retrieval of one or more citations	RR = SR/SQ
RA	The proportion of all searches in a session that are author searches	RA = SA/SQ
RT	The proportion of all searches in a session that are title searches	RT = ST/SQ
RU	The proportions of all searches in a session that are subject searches	RT = SU/SQ
RW	The proportion of all searches in a session that are power searches	RW = SW/SQ
RB	The proportion of all searches in a session that use Boolean operators	RB = SB/SQ
RD	The proportion of all searches in a session that are followed by a search modification	RD = MD/SQ
RC	The proportion of all search modifications in which the user switched from searching using one index to another (e.g., title to author)	RC = CS/SM
PM	The proportion of all search modifications that are related to finding more citations like the one the user already found	PM = MM/SM
RK	The proportion of searches in a session that are known-item searches	RK = SK/SQ
RL	The proportion of all search modifications that are related to finding fewer citations than the current search retrieved	RL = FL/SM
RM	The average number of search modifications in a session	RM = SM/MD
	Display variables	
MC	The number of multiple-record display screens used during a session	
OC	The number of single-record display screens used during a session	
DF	The number of different types of display formats (e.g., short, long) used during a session	
AV	The proportion of the length of a session that the user spent on displaying records	AV = DT/SL
NS	The average number of times per search that the user displayed multiple records on the screen	NS = MC/SR
NR	The average number of times per search that the user displayed a single record on the screen	NR = OC/SR
RP	The proportion of all searches that are immediately followed by the display of multiple citations	RP = SP/SQ
RO	The proportion of all searches that are immediately followed by the display of a single citation	RO = SO/SQ
VD	The average viewing time (seconds) for single-record display screens during a session	VD = SV/OC
VS	The average viewing time (seconds) for multiple-record display screens during a session	VS = MV/MC
	Relevance variables	
LL	The proportion of searches in a session that are followed by a PMDS activity	LL = NL/SQ
PR	The estimated precision per search with retrievals. PR is calculated as the average of the ratio of the number of citations viewed with single-record display screens to the number of citations browsed with multiple-record display screens per search with retrievals. For example, assume that a session contains two searches with retrievals. The first search has ten single-record display screens and three multiple-record display screens, and the second one has five single-record display screens and two multiple-record display screens. Also assume that 10 citations are shown in a multiple-record display screen. Then the estimated precision for this session is [10/(10 ∗ (3 − 1) + 5) + 5/(10 ∗ (2 − 1) + 5)]/2 = 0.37 because the last screen may not contain exactly 10 citations. Note that this estimation is an upper bound to the actual precision measure because the user may find an item irrelevant during a single-record display.	
LR	The actual precision per search with retrievals. LR is calculated in the same way as PR but only when a search contains PMDS activities.	
	Error variables	
TE	The number of different types of errors made during a session	
NE	The number of Web pages containing text about errors that the system displays during a session	
ER	The proportion of searches that have system-detected errors in them	ER = SE/SQ

TABLE 2. (continued)

Symbol	Description	Formula
	Relevance variables (continued)	
RE	The proportion of all Web pages displayed during a session that are related to errors (the error rate)	RE = NE/NP
	Help variables	
HP	The number of Web pages containing help text that the system displays during a session (a help request)	
TH	The number of different types of help requests made during a session	
PH	The proportion of the length of a session spent in using the help system	PH = SH/SL
VH	The average viewing time (seconds) per page for each help request	VH = SH/HP
RH	The proportion of all Web pages displayed during a session that are related to help activities (the help rate)	RH = HP/NP

Table 3 shows the means of the variables for Sample 1 and Sample 2. The length of a session (SL) was about 14 minutes (827.58 seconds) for Sample 1 and about 15 minutes (897.96 seconds) for Sample 2. Users in Sample 1 viewed an average of 22.62 Web pages (NP); in Sample 2, 20.51. The average time in seconds between Web page requests (VT) was 38.25 and 47.50, respectively, for the two samples. Users in Sample 1 got an average of about 563 hits per session (HT), while those in Sample 2 got 575. Sample 1 users employed an average of 1.38 databases (TD) and used 1.51 indexes (TI) in their searches. Sample 2 users averaged 1.37 databases and 1.81 indexes. Approximately 42% of the session time of users in Sample 1 was devoted to display activities (AV), while it was only 39% in Sample 2.

To verify that the two samples of sessions were from the same population and have the same distributions, the Wilcoxon Rank-Sum test and the Kruskal–Wallis *H*-test (both nonparametric tests) were performed on the 47 variables for 2,002 randomly selected sessions from each sample. The null hypothesis was accepted for 45 out of 47 variables.

The only cases where the null hypothesis was rejected were for variables TI, the number of different types of indexes (e.g., title, author, etc.) employed in a session, and AV, the proportion of time spent in display activities. Searchers in Sample 2 employed more complex searches and view fewer records than those in Sample 1. Although it seems that searchers in Sample 2 knew how to manipulate the system better than those in Sample 1, they did not outperform their counterparts; there is no significant difference between the samples in search performance as measured by estimated precision (PR and LR). It seems very likely that Sample 1 and Sample 2 were from the same user population.

Principal Component Analysis

A user session was represented as a vector of the 47 variables shown in Table 2. Although this set is believed to be comprehensive, it is too large for easy data analysis. Because there is a possibility that some of the 47 variables are correlated, to simplify the analysis, the representations were transformed into lower dimensional vectors using principal component analysis. The lower dimension vectors were then used to represent a session and to cluster sessions. Principal component analysis transforms a set of correlated variables into a set of uncorrelated (orthogonal) variables (components). The new variables that are derived from principal components analysis are linear combinations of the original variables and are derived in descending order of importance according to the degree of variation in the original data. Mathematically, suppose that $X = (X_1, X_2, \ldots, X_P)$ is a P-dimensional random variable with mean $u = (u_1, u_2, \ldots, u_P)$ and covariance matrix Σ. The goal of principal component analysis is to find a new set of variables, $Y = (Y_1, Y_2, \ldots, Y_P)$, that are uncorrelated and whose variances decrease from first to last. Each Y_j is a linear combination of the Xs as follows:

$$Y_j = a_{j1}X_1 + a_{j2}X_2 + \cdots + a_{jP}X_P = Xa_j \quad \text{or} \quad Y = XA$$

where

$$a_j^T = (a_{j1}, a_{j2}, \ldots, a_{jP}),\ A = (a_1, a_2, \ldots, a_P),$$

are called component loadings, i.e., the degree to which each of the variables X_j correlates with each of the components Y_j.

The power of principal component analysis lies in using only the first few new variables, called the *principal components*, to represent the original data if these variables account for a majority of the variation in the original data. Because $\mathrm{Var}(Y_j) = \lambda_j$ is the degree of the variance in the original data explained by component Y_j, it is intuitive to treat the eigenvalues, λ_j, as the importance or value of Y_j. However, there is no absolute way of deciding which eigenvalues are "large" and which are "small." A rule of thumb is to extract the components whose variances are larger than or equal to one, or to choose the components having

$$\lambda_j \Big/ \sum_{j=1}^{P} \lambda_j \geq \frac{1}{P}$$

TABLE 3. Means of the 47 derived variables used to characterize sessions in Sample 1 and Sample 2.

Variable	Description	Sample 1	Sample 2
	Session variables		
SL	Session length (in seconds)	827.58	897.96
NP	Number of Web pages requested	22.62	20.51
TP	Number of different Web pages	7.68	7.10
TD	Number of different databases	1.38	1.37
TI*	Number of different indexes	1.51	1.81
SQ	Number of searches	4.58	4.35
PF	Use of personal profile	0.06	0.05
VT	Average time between Web page displays	38.25	47.50
SD	Average search length	213.89	263.78
HT	Average number of hits	562.68	575.46
PP (100×)	Presearch time proportion	24.98	24.50
PS (100×)	Presearch Web page proportion	22.19	22.86
	Search variables		
SI	Average Web pages requested per search	5.07	4.73
RR (100×)	Proportion of searches with retrievals	69.17	70.53
RA (100×)	Proportion of author searches	20.23	21.14
RT (100×)	Proportion of title searches	24.00	24.35
RU (100×)	Proportion of subject searches	42.52	39.53
RW (100×)	Proportion of power searches	13.24	14.96
RB (100×)	Proportion of Boolean searches	59.60	57.27
RD (100×)	Proportion of searches followed by modifications	22.05	20.60
RC (100×)	Proportion of change indexes	27.01	25.84
PM (100×)	Proportion of find-more searches	10.52	9.53
RK (100×)	Proportion of known-item searches	44.23	45.50
RL (100×)	Proportion of find-fewer searches	7.76	7.19
RM (100×)	Average number of search modifications	53.18	49.16
	Display variables		
MC	Number of multiple-record displays	5.65	5.36
OC	Number of single-record displays	3.37	3.00
DF	Number of different types of display formats	1.77	1.62
AV (100×)*	Display time proportion	42.11	38.59
NS	Average number of multiple-record displays	2.42	2.38
NR	Average number of single-record displays	1.61	1.51
RP (100×)	Proportion of multiple-record displays	51.43	52.94
RO (100×)	Proportion of single-record displays	32.05	32.24
VD	Average viewing time per single-record display	33.68	38.11
VS	Average viewing time per multiple-record display	47.37	46.59
	Relevance variables		
LL (100×)	Proportion of searches followed by PMDS activities	7.98	9.41
PR (100×)	Estimated precision per search with retrievals	27.98	27.20
LR (100×)	Actual precision per search with retrievals	3.17	3.21
	Error variables		
TE	Number of different types of errors	0.30	0.21
NE	Number of Web pages containing error text	0.70	0.58
ER (100×)	Proportion of system errors	8.47	5.55
RE (100×)	Web page error rates	2.79	2.10
	Help variables		
HP	Number of Web pages containing help text	0.05	0.05
TH	Number of different types of help requests	0.04	0.04
PH (100×)	Proportion of time using help	0.30	0.28
VH	Average time per page for help	1.90	1.92
RH (100×)	Web page help rate	0.27	0.25

Notes: An * indicates that the test results (both Wilcoxon Rank-Sum test and Kruskal-Wallis *H* test) are significant ($p < 0.01$); that is, the variable does not have the same distribution in Sample 1 and Sample 2. Because some mean values are small, they are rescaled in the table (100×).

as principal components (Chatfield & Collins, 1980). That is, a principal component should account for at least the average variance in the original data in *P*-dimensional space.

In both Sample 1 and Sample 2, each of the 47 derived variables was normalized to have zero mean and unit variance (i.e.,

$$\frac{X_i - \bar{X}}{\mathrm{Std}(X_i)}\Bigg),$$

and the covariance matrix (i.e., the correlation matrix of the original variables) was employed to generate the principal components. The purpose of normalization is to eliminate the bias caused by different measuring units. Tables 4 and 5 display the principal components and associated eigenvalues for Sample 1 and Sample 2, respectively.[1]

As mentioned previously, a component whose variance is larger than or equal to one (or the average variance in the

TABLE 4. Principal component analysis eigenvalues of the covariance matrix in Sample 1.

	Eigenvalue	Difference from next eigenvalue	Proportion of variance explained	Cumulative proportion explained
PRIN1	7.65790	3.57923	0.162934	0.16293
PRIN2	4.07867	0.83598	0.086780	0.24971
PRIN3	3.24269	0.25149	0.068993	0.31871
PRIN4	2.99120	0.81481	0.063642	0.38235
PRIN5	2.17638	0.11428	0.046306	0.42866
PRIN6	2.06211	0.28102	0.043875	0.47253
PRIN7	1.78109	0.9879	0.037896	0.51043
PRIN8	1.68230	0.04564	0.035794	0.54622
PRIN9	1.63666	0.21459	0.034822	0.58104
PRIN10	1.42207	0.11711	0.030257	0.61130
PRIN11	1.30496	0.06389	0.027765	0.63906
PRIN12	1.24107	0.03563	0.026406	0.66547
PRIN13	1.20543	0.08935	0.025648	0.69112
PRIN14	1.11609	0.03937	0.023747	0.71486
PRIN15	1.07672	0.07268	0.022909	0.73777
PRIN16	1.00404	0.01484	0.021363	0.75914
PRIN17	0.98920	0.01136	0.021047	0.78018
PRIN18	0.97784	0.01099	0.020805	0.80099
PRIN19	0.96686	0.12536	0.020571	0.82156
PRIN20	0.84150	0.00878	0.017904	0.83946
PRIN21	0.83273	0.12285	0.017718	0.85718
PRIN22	0.70988	0.03526	0.015104	0.87228
PRIN23	0.67462	0.09225	0.014354	0.88664
PRIN24	0.58237	0.04986	0.012391	0.89903
PRIN25	0.53252	0.09254	0.011330	0.91036
PRIN26	0.43998	0.03150	0.009361	0.91972
PRIN27	0.40848	0.03289	0.008691	0.92841
PRIN28	0.37559	0.02848	0.007991	0.93640
PRIN29	0.34710	0.01260	0.007385	0.94379
PRIN30	0.33450	0.05967	0.007117	0.95090
PRIN31	0.27483	0.02015	0.005847	0.95675
PRIN32	0.25468	0.01152	0.005419	0.96217
PRIN33	0.24316	0.03465	0.005174	0.96734
PRIN34	0.20851	0.00688	0.004436	0.97178
PRIN35	0.20163	0.02360	0.004290	0.97607
PRIN36	0.17803	0.01722	0.003788	0.97986
PRIN37	0.16081	0.01667	0.003422	0.98328
PRIN38	0.14414	0.00855	0.003067	0.98635
PRIN39	0.13559	0.01080	0.002885	0.98923
PRIN40	0.12479	0.01088	0.002655	0.99189
PRIN41	0.11391	0.00970	0.002424	0.99431
PRIN42	0.10421	0.02698	0.002217	0.99653
PRIN43	0.07723	0.01180	0.001643	0.99817
PRIN44	0.06543	0.04491	0.001392	0.99956
PRIN45	0.02052	0.02052	0.000437	1.00000
PRIN46	0.00000	0.00000	0.000000	1.00000
PRIN47	0.00000		0.000000	1.00000

Notes: $N = 126{,}925$. Underlined components are considered "important."

TABLE 5. Principal component analysis eigenvalues of the covariance matrix in Sample 2.

	Eigenvalue	Difference from next eigenvalue	Proportion of variance explained	Cumulative proportion explained
PRIN1	7.54491	3.33520	0.160530	0.16053
PRIN2	4.20971	0.82665	0.089568	0.25010
PRIN3	3.38306	0.19369	0.071980	0.32208
PRIN4	3.18937	0.42490	0.067859	0.38994
PRIN5	2.76447	0.50138	0.058818	0.44876
PRIN6	2.26308	0.37697	0.048151	0.49691
PRIN7	1.88612	0.22314	0.040130	0.53704
PRIN8	1.66297	0.13820	0.035382	0.57242
PRIN9	1.52477	0.12890	0.032442	0.60486
PRIN10	1.39587	0.09511	0.029699	0.63456
PRIN11	1.30076	0.09540	0.027676	0.66224
PRIN12	1.20536	0.10173	0.025646	0.68788
PRIN13	1.10363	0.01520	0.023481	0.71136
PRIN14	1.08843	0.02887	0.023158	0.73452
PRIN15	1.05956	0.04312	0.022544	0.75707
PRIN16	1.01644	0.01676	0.021626	0.77869
PRIN17	0.99968	0.02873	0.021270	0.79996
PRIN18	0.97095	0.10511	0.020659	0.82062
PRIN19	0.86584	0.04892	0.018422	0.83904
PRIN20	0.81692	0.02572	0.017381	0.85642
PRIN21	0.79121	0.01518	0.016834	0.87326
PRIN22	0.77603	0.15575	0.016511	0.88977
PRIN23	0.62028	0.07342	0.013197	0.90297
PRIN24	0.54686	0.08363	0.011635	0.91460
PRIN25	0.46323	0.04663	0.009856	0.92446
PRIN26	0.41660	0.05942	0.008864	0.93332
PRIN27	0.35718	0.02290	0.007600	0.94092
PRIN28	0.33428	0.06427	0.007112	0.94803
PRIN29	0.27002	0.01649	0.005745	0.95378
PRIN30	0.25352	0.01577	0.005394	0.95917
PRIN31	0.23776	0.04273	0.005059	0.96423
PRIN32	0.19502	0.00958	0.004149	0.96838
PRIN33	0.18544	0.01938	0.003946	0.97233
PRIN34	0.16606	0.00459	0.003533	0.97586
PRIN35	0.16147	0.00826	0.003436	0.97930
PRIN36	0.15321	0.00705	0.003260	0.98255
PRIN37	0.14616	0.01165	0.003110	0.98566
PRIN38	0.13451	0.01163	0.002862	0.98853
PRIN39	0.12288	0.00891	0.002614	0.99114
PRIN40	0.11397	0.01743	0.002425	0.99357
PRIN41	0.09654	0.01139	0.002054	0.99562
PRIN42	0.08515	0.01533	0.001812	0.99743
PRIN43	0.06982	0.03206	0.001485	0.99892
PRIN44	0.03775	0.02462	0.000803	0.99972
PRIN45	0.01314	0.01314	0.000279	1.00000
PRIN46	0.00000	0.00000	0.000000	1.00000
PRIN47	0.00000		0.000000	1.00000

Notes: $N = 130{,}902$. Underlined components are considered "important."

[1] In the tables that follow, PRIN1, PRIN2, . . . are variables used to describe the first, second, etc., principal components.

TABLE 6. Results of cluster analysis on Sample 1.

Cluster number	Cluster joined (1)	Cluster joined (2)	Frequency of new cluster	Cubic clustering criterion	Pseudo F	Pseudo t^2
CL20	OB27	CL42	8,875	591.29	8,186.59	2,977.05
CL19	CL26	CL56	12,213	548.10	8,382.49	1,777.47
CL18	CL34	OB64	13,596	525.46	8,550.66	5,821.21
CL17	CL35	CL38	15,521	496.08	8,692.07	6,583.36
CL16	CL48	CL19	14,797	459.62	8,803.95	2,588.81
CL15	CL28	CL37	22,936	409.69	8,838.27	9,729.39
CL14	CL27	CL53	3,919	360.85	8,903.50	1,807.72
CL13	CL22	CL15	34,363	314.09	9,012.57	6,731.11
CL12	CL17	CL41	21,550	238.52	9,150.56	6,340.60
CL11	CL23	CL29	7,536	190.48	9,276.56	2,556.95
CL10	CL12	CL24	29,762	145.88	9,484.24	5,644.20
CL9	CL10	CL20	38,637	98.12	9,794.91	5,064.65
CL8	CL16	CL33	17,192	60.65	10,207.31	3,338.10
CL7	CL9	CL25	47,372	−10.19	10,316.55	7,121.68
CL6	CL11	CL21	10,455	−65.25	10,590.22	**3,325.70**
CL5	CL18	CL13	47,959	113.93	10,968.46	13,414.33
CL4	CL6	CL8	27,647	144.63	11,484.95	4,425.40
CL3	CL5	CL7	95,331	159.12	12,357.27	11,488.31
CL2	CL4	CL14	31,566	158.68	13,618.66	5,681.81
CL1	CL2	CL3	126,897	0.00		13,618.66

Note: Underlined rows indicate a possible number of clusters. The row in boldface indicates the optimal number of clusters.

original data) is considered important. In this case, the first sixteen principal components (underlined in Tables 4 and 5) meet the criterion (i.e., eigenvalue ≥ 1 or proportion $\geq 1/20 = 0.05$) and are used in the data transformation (reduction). Both sets of extracted principal components explain as much as 76% of the variance in the original data in approximately one-third of the original dimensions.[2]

Cluster Analysis

Cluster analysis is a statistical analysis technique to create categories that fit observations. In our cluster analysis, the goal is to find groups of sessions that are similar and then ascertain the patterns of usage represented by that group. Our cluster analysis of sessions was divided into three phases and produced three different sets of results. In the first phase, the vectors of principal component variables representing sessions in Sample 1 were formed into clusters in a two-stage methodology that will be described in a moment. In the second, the vectors of principal component variables representing sessions in Sample 2 were clustered using the same methodology. In the third, sessions from Sample 2 were assigned to clusters derived in Sample 1. The purpose of the second and third phases was to validate the results of the clustering to ensure the clusters that were formed in Sample 1 and Sample 2 were not a random phenomenon.

There are two general types of clustering methods: hierarchical and nonhierarchical.[3] A drawback to nonhierarchical methods is that the number of clusters is *a priori* information. Without knowledge of the underlying structure, it is difficult to determine the number of clusters in the data set objectively.

Clustering Phase 1—Cluster the Sessions in Sample 1

To avoid subjectivity and reduce computational cost, a hybrid approach with two stages was employed to discover potential user groups in Sample 1. In the first stage, 125,625 sessions were divided into 100 clusters using a nonhierarchical method with the SAS procedure FASTCLUS. Sixteen clusters each with five or fewer observations (amounting to a total of only 28 sessions) were treated as outliers and removed from further analysis to prevent them from distorting the results. The remaining 84 clusters served as

[2] A pair-wise comparison between principal components extracted in Sample 1 and in Sample 2 is inappropriate because there is no underlying statistical model (e.g., normal distribution in linear regression analysis) for variance components due to sampling errors (Chatfield & Collins, 1980). A feasible approach to evaluating the similarity between two sets of principal components is to compare the roles of principal components qualitatively. In such an approach, subjectivity is unavoidable. Because principal component analysis is not the ultimate goal of this study but only an intermediary step, such a qualitative comparison will not be pursued here.

[3] The computational cost of hierarchical methods ranges from $O(N^2)$ to $O(N^5)$, but the computational cost of nonhierarchical methods is usually proportional to the size of data, $O(N)$. Thus, a nonhierarchical method is attractive when the size of the data set is very large, i.e., greater than 10,000 records (SAS, 1990).

input to the second stage; a hierarchical method using the SAS procedure CLUSTER with Ward's algorithm.

Table 6 displays the truncated results (the last 20 iterations) of the cluster analysis on the sessions in Sample 1. It is easier to understand the clustering in Table 6 by interpreting it backwards. For instance, in the last row of Table 6, cluster CL1 (126,897 sessions) was formed by joining clusters CL2 (31,566 sessions) and CL3 (95,331 sessions); in the second-to-last row, cluster CL2 (31,566 sessions) was formed by joining clusters CL4 (27,647 sessions) and CL14 (3,919 sessions); and so on. In the first row of Table 6, cluster CL20 was formed by cluster number 27 (OB27), which was formed using the nonhierarchical method in the first stage, and newly aggregated sessions in cluster CL42. Figure 1 illustrates the truncated dendrogram (tree structure) of the hierarchical clustering of the sessions of Sample 1. The horizontal lines that intersect the lightning bolt constitute the nonoverlapping clusters (usage groups). The value in the parentheses associated with a cluster is the number of sessions in that cluster.

Note that the number of nonoverlapping clusters at any point in the hierarchical structure is equal to the cluster number in Table 6. For example, in the last row of Table 6, there is only one cluster, CL1, in the data set; in the second-to-last row, there are two clusters, CL2 (formed by joining clusters CL4 and CL14) and CL3, in the data set; in the third-to-last row, there are three clusters (CL3, CL4, and CL14) in the data set; and so on. See Figure 1 for a graphical presentation.

A major goal at this stage of the analysis was to decide the correct number of clusters or groups. If there were too few groups, we cannot be as precise as possible about the differences between searchers, and if there were too many groups, we lose generality. Several measures are available to help in this decision. One looks for local peaks in the value of the cubic clustering criterion (CCC), the pseudo F statistic, and/or a local minimum value of the pseudo t^2 statistic. These indicate that the optimal number of clusters in the data set is found (SAS, 1990). Using the t^2 criteria, the possible number of clusters (underlined in Table VI) is 2, 4, 6, 8, 11, 14, 16, or 19 (see Fig. 2).

The optimal number of clusters in Sample 1 was set at six (boldface in Table 6) because it had the maximum increase in the pseudo t^2 statistic (13,414.33 − 3,325.70 = 10,088.63) toward the next iteration (i.e., from CL6 to CL5) in the hierarchical clustering. In Figure 1, from top to bottom, the clusters are CL18 (13,596 sessions), CL13 (34,363 sessions), CL7 (47,372 sessions), CL14 (3,919 sessions), CL8 (17,192 sessions), and CL6 (10,455 sessions). The pseudo t^2 statistic is mainly used to test the hypothesis that the two clusters being merged are actually from the same cluster. A sudden increase in the pseudo t^2 statistic leads to the rejection of that hypothesis; in other words, the two clusters being considered are dissimilar to each other. Therefore, the maximum increase in the pseudo t^2 statistic strongly suggests that the optimal number of clusters has been found.

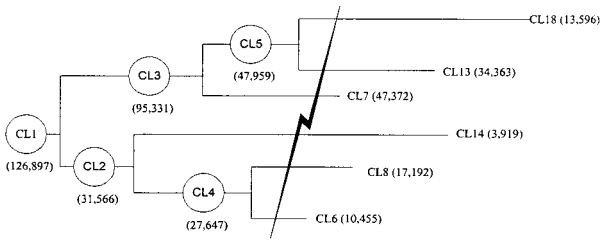

FIG. 1. Dendrogram of the hierarchical clustering on Sample 1.

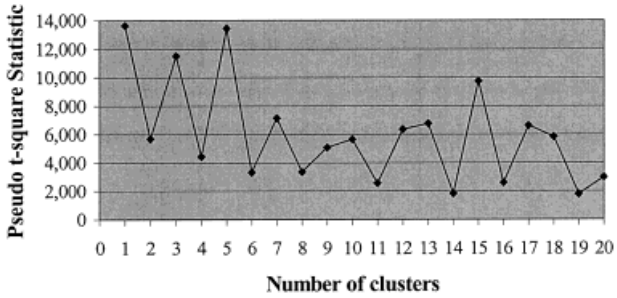

FIG. 2. Plot of number of clusters vs. pseudo t^2 statistic for Sample 1.

Clustering Phase 2—Cluster the Sessions in Sample 2

The second phase in the clustering process was to perform the same procedure on Sample 2 as was performed on Sample 1. It was performed independently, using the same variable selection, data normalization, clustering method, and beginning number of clusters as in Sample 1. The results are given in Table 7 for the last 20 iterations. The truncated dendrogram of the hierarchical clustering is illustrated in Figure 3.[4]

The possible number of clusters in Sample 2 (underlined in Table 7) is 3, 6, 8, 11, 13, 16, or 18 because of a local minimum of the pseudo t^2 statistic. The maximum increase in this statistic occurs when the number of clusters is six (i.e., from CL6 to CL5, with 12,781.31 − 3,447.99 = 9,333.32). As can be seen from Tables 6 and 7, both cluster analyses suggested that there were six distinct usage groups in the MELVYL system. From top to bottom (see Fig. 3), the clusters are CL21 (6,746 sessions), CL19 (10,796 sessions), CL18 (3,756 sessions), CL14 (21,478 sessions), CL7 (15,068 sessions), and CL6 (72,952 sessions). Comparing Figures 1 and 3, one finds that Sample 1 and Sample 2 have similar clustering structures, but the distribution of sessions over the six clusters in each sample is slightly different.

[4] Twenty-two clusters were considered outliers in this phase. They contained a total of 106 sessions. Thus, the second stage of phase 2 began with 78 clusters.

TABLE 7. Results of the cluster analysis on Sample 2.

Cluster number	Cluster joined (1)	Cluster joined (2)	Frequency of new cluster	Cubic clustering criterion	Pseudo F	Pseudo t^2
CL20	OB9	OB23	12,053	755.14	9,183.55	1,030.95
CL19	OB13	CL28	10,796	691.98	9,300.21	7,901.00
CL18	CL31	CL39	3,756	662.61	9,446.63	6,055.90
CL17	CL24	CL36	17,828	601.95	9,601.12	6,595.41
CL16	CL30	OB51	5,544	570.33	9,766.14	2,295.25
CL15	CL42	CL26	8,828	506.12	9,917.77	4,815.17
CL14	CL23	CL20	21,478	473.07	10,116.03	2,641.31
CL13	OB18	OB41	16,684	432.46	10,285.19	497.19
CL12	OB5	CL17	23,738	366.11	10,463.57	1,686.93
CL11	CL22	CL27	20,719	316.93	10,614.76	1,076.91
CL10	CL12	CL16	29,282	269.74	10,840.94	6,382.98
CL9	CL10	CL37	35,549	206.39	11,118.85	11,726.49
CL8	CL9	CL11	56,268	137.88	11,215.06	5,970.09
CL7	CL15	CL25	15,068	68.33	11,448.88	12,574.50
CL6	CL8	CL13	72,952	−0.79	11,696.19	**3,447.99**
CL5	CL7	CL6	88,020	−61.04	11,980.75	12,781.31
CL4	CL21	CL19	17,542	102.92	12,471.13	9,007.43
CL3	CL18	CL4	21,298	123.97	13,190.07	7,385.87
CL2	CL14	CL5	109,498	127.02	14,373.26	7,789.57
CL1	CL3	CL2	130,796	0.00		14,373.26

Note: Underlined rows indicate a possible number of clusters. The row in boldface indicates the optimal number of clusters.

Clustering Phase 3—Test of Cluster Validity

To demonstrate that the clustering structure of six clusters found in Sample 1 was valid and did not occur by chance, and was replicable in Sample 2, an external criteria-based test of cluster validity was conducted. External criteria measure the degree to which the cluster structure matches *a priori* information (Jain & Dubes, 1988). As such, they are widely used in replication analysis, which assesses the degree to which two partitions of samples of the same population agree.

The methodology used to perform the test involved taking observations (sessions) from Sample 2 and assigning them to the nearest cluster centroid (a vector of principal component values representing the location in the 16-dimensional space where this cluster center can be found) of Phase 1. This strategy is similar to building a subject classification scheme (Phase 1) and then assigning documents to a subject (Phase 3). To perform the assignment, the Euclidian distance between each session vector in Sample 2 and each centroid in Sample 1 was computed, and the 130,796 observations in Sample 2 were assigned to the closest of the six Sample 1 centroids to form a third clustering.[5] Then, a number of external indices were computed to obtain quantitative measurements of agreement between the characteristics of the clusters found in Phase 2 and Phase 3. The test procedure is an extension of the work of McIntyre and Blashfield (1980) and Morey, Blashfield, and Skinner (1983).

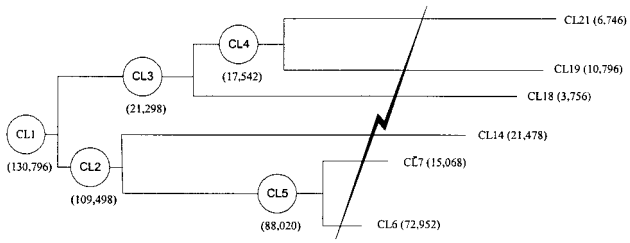

FIG. 3. The truncated dendrogram of the hierarchical clustering in Sample 2.

The cluster centroids in Phase 1 were labeled CL6, CL7, CL8, CL13, CL14, and CL18 (see Fig. 1). Because the contents of the clusters had changed in Phase 3, we retained the cluster number so as to make the association with Phase 1, but replace the label; thus CL6 became F6, CL7 became F7, and so forth. For simplicity, we refer to the clusters from Phase 2 in Sample 2 as E-clusters (i.e., CL6 in Phase 2 becomes E6, CL7 in Phase 2 becomes E7, CL14 becomes E14, CL18 becomes E18, and so on).

Sample 2 has now been divided in two ways. In Phase 2 it was clustered using a two-stage methodology. In Phase 3 it was assigned to centroids derived from Phase 1. Table 8 shows the similarity between the two different clustering methods for the same set of observations, i.e., Sample 2. Each cell in the table gives the number of sessions in Sample 2 that were in the corresponding clusters. Thus, the

[5] A total of 130,796 sessions went into Phase 3 because 106 were eliminated as outliers.

TABLE 8. Contingency table for E-clusters and F-clusters in Sample 2.

Cluster	F6	F7	F8	F13	F14	F18	Total
E6	10	5	0	321	4,289	847	72,952
E7	2,595	3,534	20,300	581	2	2	15,068
E14	280	629	0	807	2,808	5,087	21,478
E18	60,826	848	485	0	796	0	3,756
E19	509	0	297	2,045	2,374	810	10,796
E21	8,732	10,052	396	2	527	0	6,746
Total	5,472	27,014	9,611	62,955	6,035	19,709	130,796

value in cell (E18, F6) is 60,826, the number of sessions in Sample 2 that were in those clusters. The underlined numbers in Table 8 show which clusters were most similar. For example, E-cluster CL6 is most similar to F-cluster CL14, E7 similar to F8, E14 to F18, E18 to F6, E19 to F14, and E21 to F7.

Table 8 can be further simplified into the 2×2 contingency table shown in Table 9. In Table 9, entry (a) is the number of pairs of sessions in Sample 2 that were not in the same cluster in partitions E and F. Entry (b) is the number of pairs of sessions that were not in the same cluster in partition E but were in the same cluster in partition F. Entry (c) is the reverse of entry (b), and entry (d) is the reverse of entry (a).[6] Obviously, larger values for entities (a) and (d) indicate a higher degree of agreement between partitions E and F. This is because if the clustering structure in partition E is valid and replicable in partition F, a pair of sessions that are close to each other in the former should be close to each other in the latter, and a pair of sessions that are distant from each other in the former should be distant from each other in the latter. This analysis suggests close agreement between partitions E and F for Sample 2 because entries (a) and (d) are larger than entries (b) and (c) in Table 9.

[6] The values in the table are computed as follows (see Jain & Dubes, 1988):

$$a = N(N+1)/2 - \left[\sum_{i=1}^{K} n_{i.}^2 + \sum_{j=1}^{R} n_{.j}^2\right]/2$$

$$b = \sum_{i=1}^{K}\sum n_{i.}^2/2 - \sum_{i=1}^{K}\sum_{j=1}^{R} n_{ij}^2/2$$

$$c = \sum_{j=1}^{R} n_{.j}^2/2 - \sum_{i=1}^{K}\sum_{j=1}^{R} n_{ij}^2/2$$

$$d = \sum_{i=1}^{K}\sum_{j=1}^{R} n_{ij}^2/2 - N/2$$

$$N = \sum_{i=1}^{K} n_{i.} = \sum_{j=1}^{R} n_{.j}$$

where K is the number of clusters in partition E and R is the number of clusters in partition F.

Finally, a number of external indices, shown in Table 10, were used as quantitative measurements of agreement between the clustering assignments in Phase 2 and Phase 3. The indexes all have values between 0 and 1. Larger values indicate a higher degree of agreement between the two partitions, E and F. However, how large is "large"? And how can one tell that two highly close partitions have not been developed by chance? For an absolute measurement of agreement, the indices are best tested against their baseline distribution, i.e., pure randomness. Then a clustering can be asserted to be "valid" if it has unusually high values of corrected (normalized) external indices. Let n_{ij} be the value in cell (i, j) in Table 8. Then $n_{i.}$ is the ith row sum and $n_{.j}$ is the jth column sum. Hubert and Arabie (1985) suggested a baseline distribution based on these row and column sums in Table 8 being fixed but the partitions (i.e., K and R) being chosen at random. In that case, an index S is corrected according to the following formula:

$$S' = S - E(S)/[\text{Max}(S) - E(S)]$$

To compute the corrected Rand index, define

$$G = ([2(a+d)/N(N-1)] - E[2(a+d)/N(N-1)])/(1 - E[2(a+d)/N(N-1)])$$

$$H = \sum_{i=1}^{K}\sum_{j=1}^{R} n_{ij}(n_{ij}-1)/2 - [2/N(N-1)] \times \sum_{i=1}^{K} n_{i.}(n_{i.}-1)/2 \sum_{j=1}^{R} n_{.j}(n_{.j}-1)/2$$

TABLE 9. Contingency table for E-clusters and F-clusters in Sample 2.

	Pairs of sessions are not in the same F-Cluster	Pairs of sessions are in the same F-Cluster
Pairs of sessions are not in the same E-Cluster	4,375,691,158 (a)	432,288,537 (b)
Pairs of sessions are in the same E-Cluster	905,410,961 (c)	2,840,340,754 (d)

TABLE 10. Measures of cluster similarity for the sessions.

Name	Formula	Value
Fowlkes and Mallows	$d/\sqrt{(d + c)(d + b)}$	0.8683
Jaccard	$d/(b + c + d)$	0.7658
Rand (R)	$2(a + d)/N(N - 1)$	0.8436
Corrected Rand (R′)	$R - E(R)/[\text{Max}(R) - E(R)]$	0.6496

$$D = \left[\sum_{j=1}^{R} n_{.j}(n_{.j} - 1)/2 + \sum_{i=1}^{K} n_{i.}(n_{i.} - 1)/2 \right]/2$$

$$- [2/N(N - 1)] \sum_{i=1}^{K} n_{i.}(n_{i.} - 1)/2 \sum_{j=1}^{R} n_{.j}(n_{.j} - 1)/2$$

Then

$$R' = G = H/D$$

The indexes in Table 10 represent probabilities. Thus, the Jaccard value of 0.76 suggests that there is a 24% chance of disagreement between partitions E and F. The values of all three uncorrected indices indicate that the clustering structure in Sample 2 is replicable to that in Sample 1 for a value over 0.7. Even after normalization, the corrected Rand index is 0.65 for the sessions. It appears that the clustering structure is very unlikely to have been discovered by chance. The 13% (1–0.87) to 35% (1–0.65) chance of disagreement between partitions E and F was largely caused by sampling errors and by the loss of information due to data reduction using principal component analysis. Thus, the test results demonstrated the validity of the cluster analyses on the sessions in the MELVYL system.

An Interpretation and Comparison of the Usage Groups

As discussed previously, six clusters were found in Sample 1. Table 11 displays the means of the 47 derived variables for each usage group, and Table 12 gives a qualitative description of each group.

From Table 11 it can be seen that cluster CL6 had the longest viewing time (around 2 to 3 minutes) per Web page (VT, VS, and VD). The average session length was about 22 minutes (SL), of which nearly 64% is in display activities (AV). Few queries were submitted (SQ), and almost all of them were effective in terms of retrievals (RR). This also explains why sessions in this cluster had the largest proportion of searches immediately followed by a display activity (RP and RO). A subject search (RU) was the most frequently used type of search, followed by a known-item search (RT, RA, and RK). About 59% of the searches employed Boolean functions. Nearly 12% of the queries were modified (RD); "change display format" (RC) was the most frequently used type of search modification.

More than half of the queries that fall in this cluster retrieved relevant items (LL), and the average number of retrievals per effective search (HT) (one that retrieves some items) was manageable. Thus, this group is characterized by good performance in terms of retrieving items that are relevant (PR and LR). In addition, sessions that fall in this group were the longest (around 4 to 5 minutes) in viewing retrieved items (VD and VS). They seemed to retrieve a good set of items, made few errors (NE, RE, and TE), and requested little on-line help (HP, RH, and TH). This all implies that the cluster CL6 probably represents knowledgeable and sophisticated use of the system and its databases being searched, because to have good retrieval results a subject search usually requires knowledge of the indexing structure of the database.

Cluster CL7 had the largest number of sessions (37.3%) in it. The session length was about 11 minutes (SL), of which 20 and 45% was in presearch activities (PP) and in display activities (AV), respectively. The average viewing time per Web page (VT) was a bit more than half a minute, which is a reasonable amount of time to skim the information on one screen. Nearly 69% of the queries submitted by this group retrieved some items (RR); however, the average number of retrievals per effective search (HT) was very large, bordering on unmanageable (950 items). On average, only 21 out of the 950 retrieved items were viewed using a screen display (NS).[7] Few retrieved items were displayed (MC, OC, NS, and NR), and nearly one-third of the queries submitted were modified (RD). Finally, sessions falling in this group barely employed on-line help (HP, RH, TH, VH, and PH). We consider the sessions that fall in this cluster relatively unsophisticated.

The usage characterized by cluster CL8 is quite different. In the cluster were the lengthiest sessions (SL is about 36 minutes), those requesting the highest number and most types of Web pages (NP and TP), the most queries (SQ), and the fewest presearch activities (PS and PP) among all groups. As variable TP indicates, there were a lot of exploratory activities among sessions that fall into this cluster. Despite heavy search activities (SQ, SD, and SI), only 14% of the queries submitted retrieve relevant items (LL). Due to the manageable size of their retrieved sets (HT), the average search performance (PR and LR) remains attractive. In fact, queries that resulted in retrievals performed best in terms of search effectiveness (LR). Sessions that fall in this cluster allocated 55% of their time displaying retrieved items (AV), and the average viewing time per screen display was around

[7] The variable NS is the number of screen displays employed per effective search. By default, 10 items are displayed on one screen in the MELVYL system. Thus, the number of items viewed using a screen display per effective search is approximately $10 * (NS - 1) + 5$ because the last screen may not contain exactly 10 items. In this case, it is $10 * (2.12 - 1) + 5 \approx 16$.

TABLE 11. Means of the 47 derived variables for Sample 1 for the six identified clusters.

Variables	CL6	CL7	CL8	CL13	CL14	CL18
Number of sessions	10,455	47,372	17,192	34,363	3,919	13,596
SL	1,317.60	640.06	2,138.56	401.02	1,040.65	146.75
NP	30.07	17.84	62.78	12.75	27.63	5.67
TP	9.00	7.15	13.60	6.04	11.61	4.08
TD	1.11	1.28	2.25	1.24	1.62	1.12
TI	1.16	1.46	2.39	1.37	1.82	1.14
SQ	2.01	4.20	12.10	3.16	5.23	1.81
PF	0.09	0.02	0.20	0.04	0.10	0.02
VT	105.36	35.08	36.59	28.40	38.63	24.67
SD	996.34	162.70	237.71	105.14	221.59	31.13
HT	483.08	950.42	596.82	160.91	548.72	247.24
PP (100×)	7.55	20.01	6.50	30.27	22.13	66.64
PS (100×)	16.66	19.14	7.18	25.97	19.30	47.35
SI	15.45	4.31	6.29	3.53	5.57	1.83
RR (100×)	90.04	68.77	67.39	81.16	52.39	31.34
RA (100×)	18.91	7.40	12.11	45.57	16.70	13.28
RT (100×)	13.94	9.46	14.39	52.09	22.32	24.10
RU (100×)	55.21	59.76	57.98	1.65	47.12	55.07
RW (100×)	11.94	23.39	15.51	0.70	13.87	7.56
RB (100×)	58.56	64.65	68.33	48.55	60.74	59.39
RD (100×)	11.33	30.98	39.36	11.95	27.99	1.19
RC (100×)	10.88	28.19	59.98	23.37	29.75	2.00
PM (100×)	4.97	19.47	14.49	0.86	20.90	0.07
RK (100×)	32.85	16.85	26.50	97.65	39.01	37.37
RL (100×)	4.65	14.43	10.23	1.16	9.23	0.06
RM (100×)	25.02	73.51	97.81	27.29	75.22	2.14
MC	11.25	4.21	16.38	2.79	4.54	0.34
OC	5.60	2.39	10.34	1.90	2.48	0.06
DF	2.20	1.75	2.96	1.68	1.69	0.32
AV (100×)	63.76	44.95	54.67	42.10	27.07	4.08
NS	7.80	2.12	3.67	1.47	1.94	0.32
NR	4.05	1.40	2.90	1.18	1.24	0.06
RP (100×)	81.25	50.95	46.76	59.47	36.54	20.11
RO (100×)	42.96	31.98	33.39	40.96	22.18	2.59
VD	120.39	28.20	43.68	23.86	25.77	0.67
VS	158.11	43.86	51.08	34.49	40.57	4.28
LL (100×)	51.24	3.59	13.48	1.21	8.42	0.08
PR (100×)	17.63	24.71	46.74	36.52	23.61	3.30
LR (100×)	6.66	0.75	15.98	0.24	3.51	0.01
TE	0.35	0.38	0.66	0.07	0.51	0.10
NE	0.74	0.84	1.66	0.09	1.25	0.11
ER (100×)	19.57	12.28	10.65	1.16	15.04	0.54
RE (100×)	2.31	4.83	2.70	0.48	4.41	1.53
HP	0.08	0.08	0.17	0.06	2.25	0.09
TH	0.07	0.07	0.14	0.05	1.78	0.08
PH (100×)	1.78	0.58	0.47	0.54	16.06	1.51
VH	4.54	2.46	4.93	1.75	49.46	3.08
RH (100×)	0.45	0.63	0.38	0.61	14.08	1.63

1 minute (VS and VD). There were proportionally fewer errors in this cluster (ER and RE), but more on-line help requests (HP, TH, and VH), than in cluster CL7. Thus, the usage characterized by this cluster can be termed "highly interactive with good search performance." In addition, the high use of profiles (PF) revealed that most of the usage came from individuals who were frequent users of the MELVYL system.

Cluster CL13 is substantially different from the preceding groups. As can be seen from Table 11, the usage characterized by cluster CL13 represents known-item searches almost exclusively (RT, RA, and RK). In this group, the average session length (SL) was around 7 minutes, with an average of 13 Web pages requested (NP) per session. Sessions lasted for a shorter period of time because the usage seems to be characterized by a clear information need. Once the item or the information had been found, the session ended.

Generally speaking, sessions falling in cluster CL13 had the worst search performance (LL and LR) than those in the preceding groups. However, variables RR (proportion of searches with retrievals), HT (average number of retrievals

TABLE 12. Summary of cluster characteristics in Sample 1.

Cluster number	Type of usage	Operational features
CL6	Knowledgeable and sophisticated usage	• The longest viewing time per Web page • Two-thirds of session length in index access activities • Almost all queries were effective in terms of retrievals • The largest proportion of searches followed by a display activity • The most time in viewing retrieved items • A subject search was the most frequently used type of search • Search performance was good, the average number of retrievals was manageable, the largest proportion of searches that retrieved relevant items • Made few errors, and requested little online help.
CL7	Unsophisticated usage	• The largest number of sessions • One-fifth of session length and nearly half of Web pages requested in presearch activities • Nearly half of session length in display activities • A high use of subject searches and power searches, two-thirds of queries retrieved some items, the average number of retrievals was too big to be manageable, few retrieved items displayed per search, nearly one-third of queries were modified, poor search performance due to lack of skills and index knowledge of the database being searched • Made more errors due to information overload; barely requested online help.
CL8	Highly interactive usage with good search results	• The lengthiest sessions • The highest number and most types of Web pages requested, the most queries submitted, as low as one-seventh of queries submitted retrieved some items, one-sixth of effective searches retrieved relevant items, the average number of retrievals was manageable, the best search performance in terms of search effectiveness, displayed more retrieved items per search • Made fewer errors proportionally but requested more on-line help, and tended to activate an existing profile.
CL13	Known-item searching	• The second largest searcher group, shorter sessions • Employed known-item searches exclusively, four-fifths of queries retrieved some items but poor search performance • Rarely requested online help.
CL14	Help-intensive searching	• The smallest searcher group • Characterized by the extraordinary use of online help, one-third of queries were modified, longer viewing time per Web page • The longest viewing time per help page, better search performance than that of cluster CL7 but proportionally fewer searches with retrievals.
CL18	Relatively unsuccessful usage	• The least active searchers, the shortest sessions, the smallest number and fewest types of Web pages requested, the shortest viewing time per Web page • Characterized by extraordinary presearch activities, rarely performed display activities, more than two-thirds of queries retrieved nothing, the smallest proportion of searches that retrieved relevant items, the worst search performance, seldom modified search queries • Rarely requested online help.

per effective search), and PR (estimated precision per effective search) indicate that about 81% of the queries submitted retrieved some items, of which 37% were actually viewed (PR). This implies that either most users whose sessions fall in this group were unaware of the PMDS (print, mail, download, and save) functions or they simply did not want to use them. In addition, the sessions consist of a simple search strategy—less than half of the searches employed Boolean functions (RB). Moreover, the infrequent use of subject searches (RU) or power searches (RW) might indicate a lack of user skills or a lack of knowledge of system features or the indexing scheme of the database being searched. Sessions in this cluster rarely requested on-line help (HP, RH, VH, and TH). The amount of time spent in presearch activities (PP and PS) suggests a level of inexperience. We think this cluster represents known-item searching.

Cluster CL14 had the fewest sessions in it (3%) among all the groups. The average session length was about 17 minutes (SL), with 28 Web pages requested (NP). Sessions in this group were characterized by the extraordinary use of on-line help (HP, RH, TH, VH, and PH), and we label it as a "help intensive cluster." In addition, more than one-fifth of total session time was spent in presearch activities (PP and PS). Compared to the sessions in clusters CL7 and CL13, these sessions showed much better searching performance (LL and LR), but had fewer searches with retrievals (RR). A subject search (RU) was the most favored type of search, followed by a known-item search (RT, TA, and RK). Nearly one-third of the queries submitted were modified (RD); "change search type" (RC) was the most frequent search modification, followed by "find more" (PM). These sessions were also characterized by spending more time reading each Web page (VT, VH, VS, and VD).

Finally, CL18 is a cluster of sessions with the shortest session length (SL is about 2.5 minutes per session), the lowest number (NP) and fewest types (TP) of Web pages (about 6 and 4, respectively), and the shortest viewing time

per Web page (VT is about 25 seconds). There was barely any display activity (MC, OC, NS, NR, VS, VD, and AV), because 69% of the queries submitted retrieved nothing (RR). Although the average number of retrievals (HT) was normal, very few searches retrieved relevant items (LL). This explains why the sessions had the worst search performance in terms of search effectiveness (PR and LR) among all groups. There were seldom any modifications of search queries (RD) and seldom any requests for help from the system (HP, RH, TH, and VH). Further, the high use of subject searches (RU) along with the lowest search performance suggests that the use was relatively unsuccessful, and we label it as such. The extraordinarily heavy presearch activities (PS and PP) also suggest that the users whose sessions fall into this cluster were new to the system, and spent most of their time learning about the system before doing a search.

Conclusions

The goal of this article was to bring a new concept to studying usage patterns: segmenting user not by who the users are or their demographics, but rather by what patterns of usage they had in common. The methodology was illustrated and tested for validity using two independent samples of user sessions drawn from the transaction logs of the University of California's MELVYL® on-line library catalog system. A session was represented as a vector of 47 variables. Though this set of variables was believed to be comprehensive, it was too large for easy data analysis. Therefore, sessions were transformed into lower dimensional vectors using the principal component analysis technique. Sixteen principal components, which altogether account for 76% of the overall variance, were extracted during the data transformation. Despite the loss of information, the reduction was beneficial because the dimensions of the data after transformation were one-third of the original.

In cluster analysis, an efficient hybrid approach with two stages, combining a hierarchical method and a nonhierarchical method, was employed to discover potential usage groups in Sample 1. In the first stage, user sessions were divided into 100 clusters using a nonhierarchical method. In the second stage, a hierarchical method was applied. Through a series of statistical tests it was ascertained that six clusters could be defined as the optimal partitioning of the samples.

Next, an independent cluster analysis was performed on Sample 2 to test the validity of the approach. This analysis used the same variable selection, data normalization, clustering method, and beginning number of clusters as in Sample 1. Again, the results suggested that there were six clusters in Sample 2 as well. To see how close these two clustering structures were, a number of indices were computed to gauge the degree of agreement. The test results demonstrated that the clustering structures found in Sample 1 were valid, and did not occur by chance, and were replicable to other samples drawn from the same user population.

Finally, the characteristics of each usage group were discussed and compared qualitatively. Groups can be labeled as knowledgeable and sophisticated use, unsophisticated use, highly interactive use with good search performance, known-item searching, help-intensive searching, and relatively unsuccessful use. Quantitative differences between groups in the values of the variables were not statistically tested because cluster analysis attempts to maximize the separation between them; therefore, the assumptions of the significance tests, parametric or nonparametric, are violated.

The methodology presented in this article could have many applications. It could aid in the design of an advanced on-line system that provides customized or situational help, depending on the category of usage. It could also be employed in the design of Web sites that adapt their content to the preferences and behavior of users. Or it could be applied to customer relationship management in electronic commerce, enabling a company to provide an adaptive system or service that meets its customer needs. This study also lays the groundwork for further analysis of user behavior on the Web. For example, the authors also analyzed user transition behavior using continuous-time stochastic models; these results will be reported in a later article.

References

Berger, M.G. (1994). Information-seeking in the online bibliographic system: An exploratory study. Ph.D. dissertation. School of Library and Information Studies, University of California, Berkeley.

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

Chatfield, C., & Collins, A. (1980). Introduction to multivariate analysis. New York: Chapman and Hall.

Chen, Hui-Min. (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.

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

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

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

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193–218.

Jain, A., & Dubes, R. (1988). Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall.

Jansen, B., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. Information Processing and Management, 36, 207–227.

McIntyre, R., & Blashfield, R. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. Multivariate Behavioral Research, 15, 225–238.

Morey, L., Blashfield, R., & Skinner, H. (1983). A comparison of cluster analysis algorithms—A comparison of hierarchical and nonhierarchical methods. Multivariate Behavioral Research, 20, 283–304.

Penniman, W.D. (1975). A stochastic process analysis of on-line user behavior. In C.W. Husbands & R.L. Tighe Eds.), Proceedings of the annual meeting of the American Society for Information Science. (pp. 147–148). Washington, DC: American Society for Information Science.

Penniman, W.D. (1982). Modeling and evaluation of on-line user behavior. Proceedings of the ASIS Annual Meeting, Columbus, OH: American Society for Information Science, 19, 231–235.

Peters, T.A. (1998). Remotely familiar: Using computerized monitoring to study remote use. Library Trends, 47, 7–20.

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.

SAS. (1990). SAS/STAT user's guide vol. 1, ANOVA-FREQ; Version 6, 4th ed. Cary, NC: SAS Institute.

Spink, A., Wolfram, D., Jansen, B., & Saracevic, T. (2001). Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology, 52.

Yuan, W., & Meadow, C.T. (1999). A study of the use of variables in information retrieval user studies. Journal of the American Society for Information Science, 50, 140–150.