Predicting the Relevance of a Library Catalog Search

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

Relevance **has been a difficult concept to define, let alone measure. In this paper, a simple operational definition of relevance is proposed for a Web-based library catalog: whether or not during a search session the user saves, prints, mails, or downloads a citation. If one of those actions is performed, the session is considered relevant to the user. An analysis is presented illustrating the advantages and disadvantages of this definition. With this definition and good transaction logging, it is possible to ascertain the relevance of a session. This was done for 905,970 sessions conducted with the University of California's Melvyl online catalog. Next, a methodology was developed to try to predict the relevance of a session. A number of variables were defined that characterize a session, none of which used any demographic information about the user. The values of the variables were computed for the sessions. Principal components analysis was used to extract a new set of variables out of the original set. A stratified random sampling technique was used to form ten strata such that each new strata of 90,570 sessions contained the same proportion of relevant to nonrelevant sessions. Logistic regression was used to ascertain the regression coefficients for nine of the ten strata. Then, the coefficients were used to predict the relevance of the sessions in the missing strata. Overall, 17.85% of the sessions were determined to be relevant. The predicted number of relevant sessions for all ten strata was 11%, a 6.85% difference. The authors believe that the methodology can be further refined and the prediction improved. This methodology could also have significant application in improving user searching and also in predicting electronic commerce buying decisions without the use of personal demographic data.**

1. Introduction

This paper has a number of goals: (1) to propose a simple operational definition of what constitutes a successful and perhaps relevant search of a library catalog by a user, (2) to propose a number of measures that characterize user behavior while searching a Web-based library catalog, (3) to develop a methodology to predict when a search by a user will be perceived as being relevant by the user, and (4) to empirically test the methodology with more than 900,000 user search sessions with the University of California's Melvyl Web-based library catalog (www.melvyl.ucop.edu).

2. An Operational Definition of Relevance for a Web-Based Catalog Search

The usual goal of an information retrieval activity is to find materials, such as bibliographic citations, the text of articles, abstracts, and so on, in a stored database that meets user needs.[1] The goal is to find relevant information, but the exact definition of what constitutes "relevant information" has been the subject of exhaustive discussion. A good definition of "relevant information" would help in the evaluation of information retrieval systems. If a system administrator can ascertain that the system is delivering information that meets user's needs, so much the better. But if this is not the case, something should be done to remedy the situation, for example, improve the the user–system interface, the indexing or classification of the materials in the database, the query language, the help system, user training, or the performance of the computer software and hardware. Thus, a good definition of whether the users of the system are meeting with success in their information-retrieval activities is critical to evaluation of the system and critical to the users' continued willingness to employ the system.

Before the widespread deployment of Web-based library catalogs, the only way to assess user satisfaction was to conduct controlled experiments involving a prescribed document collection, known queries, multiple judges, and a carefully constructed experimental design that took into account a myriad of characteristics of the judge, the document collection, the queries, and the type of judgment made.

Now, consider a user's interaction with a typical Web-based library catalog. The user selects a database, the index to use in searching the database, and keys in the query. The system returns a search history indicating the number of hits

[1] In some cases, the goal may be the opposite, namely not to find anything in the hopes of launching a new investigation into an unexplored area.

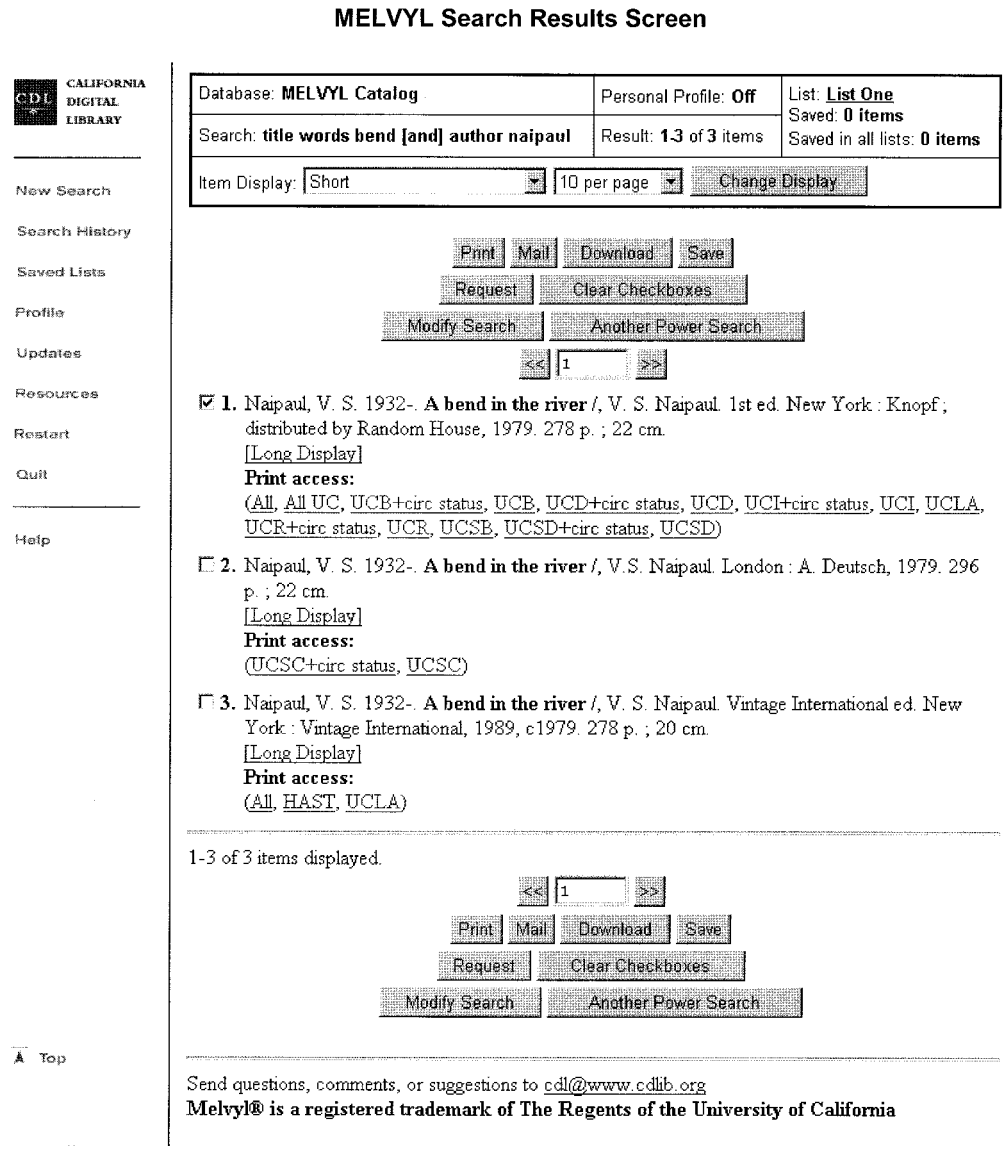

FIG. 1. MELVYL search results screen.

obtained for the search, or sometimes simply a list of items retrieved, depending on the size of the retrieved set. Either way, the user winds up looking at a screen, like the one shown in Figure 1, containing a list of citations. If the user likes one or more items, they can place a check next to the item, as shown for the first item in the list in Figure 1. Then, the system can save the selected items in a list, which later can be sent by e-mail, formatted for printing on the user's local computer, or downloaded to the user's local machine so that it can be incorporated into a personal bibliographic database. Printing, mailing, downloading, or saving (PMDS) are overt acts that the user takes when they perceive that the entry on the screen is worth keeping—for whatever reason. If, during a session, a user performs a search, displays a record from the search, and then takes one of these actions, it is an indicator that the user has found something relevant. We will take this to constitute a simple operational definition of whether the entire session was relevant.[2]

This definition of relevance is binary. If any of the actions are taken, the entire session is marked as relevant. It is possible to expand the measure of relevance to be a quantity, representing the total number of times any of the actions took place, but this was not done.

It is possible that the user finds a session relevant but this information is not captured in the transaction logs. For example, the user may simply use the *print screen* button on his or her computer to print relevant information, or jot down a call number or other data on a piece of paper. Thus,

[2] These actions are recorded for a session as a whole rather than for each search within a session. It would not be difficult to say which search within a session should be considered relevant using this measurement strategy, but this was not done for this paper.

TABLE 1. Reasons why a user would select a citation from a retrieved list.

Reason	Source
I (as the user) consider the item relevant.	
If I say the item is relevant, it is relevant.	Harter (1992)
The item may be relevant, but I cannot be sure. I will examine it further.	
I am selecting this item to give to my colleague.	
I clicked on this item, but in retrospect I think my mouse went crazy (mistake).	
I have selected this item because in selecting it I am making a good trade-off between contextual effort and processing effort.	Harter (1992)
Obsessive concern for completeness is a waste of time.	Wilson (1993)
My internal experiences indicate this citation is going to be ok (subjective about).	Maron (1977)
My perceived utility for this item is high.	Saracevic (1975)
I ran across this item high on the list, and therefore I chose it.	
My experience, background, attitude, and so on, make me believe that this is a good item.	Cuadra and Katter (1967), Rees & Schultz (1967)
I trust the publisher, author, place of publication, the fact that my library holds it, the number of volumes, the length of the back run, the sponsoring agency, and so on.	Cooper (1971), Cooper (1973a), Cooper (1973b), Rees & Schultz (1967)
My current stock of information makes me believe that this is relevant.	Wilson (1968)
My current situation makes me believe that this is relevant (situational relevance).	Wilson (1973)

Many of the paraphrased statements in this table were given by multiple authors. The choice of who to cite as the source does not constitute historical precedence, but rather was an arbitrary decision. Further, the statements are not presented in any particular order of importance.

the results presented here are biased because the study had no way of recording this type of action.

What are the advantages and disadvantages of this definition? On the plus side, it is easy to measure with a reasonably sophisticated transaction logging system. The monitor needs simply to record when the user places a check next to one of the displayed items, and whether the user performs one of the four actions described previously.

The more difficult question is whether this constitutes a valid measure of relevance. Let us consider some possible reasons why a user would take the overt act of selecting one or more items from a list of search results. Table 1 provides a very selective summary. For relatively complete reviews of the literature in this area, see Mizzaro (1997), Schamber et al. (1990), and Saracevic (1975, 1976). Implicit in the entries in this table is the assumption that relevance is a relationship between a user's need, the user's state at any point in time, and the representation of the material displayed on the screen. The user makes the decision whether or not the relationship holds in the particular case. That decision may be judged by others to be incorrect, but it is the user who makes the final determination. Given the entires in Table 1, can we say with certainty that the act of saving a citation is equivalent to stating that the item is relevant for a user, and thus the search was relevant? Probably not, but it is a very good approximation and is very easy to capture.

Now let us examine the opposite side of the question. Why would a user *not* select a citation from the list in Figure 1? Table 2 gives many reasons, principally from the work of Wilson, for not selecting an item, besides the item being perceived as irrelevant. For example, the user may decide that the item is peripheral, is too specialized, is too difficult to comprehend, or is just not appropriate to the information need. Again, for whatever reason, not clicking on a citation in the list in Figure 1 is a pretty good approximation to indicating that, for a variety of possible reasons, the user is not interested in the item. And that is a pretty good approx-

TABLE 2. Reasons why a user would not select a citation from a retrieved list.

Reason	Source
I (as the user) consider the item not relevant.	
I have read the item.	
There is nothing new here.	Wilson (1993)
This has nothing to do with my work.	Wilson (1993)
I don't trust the author or publisher. The title makes me suspicious.	
I have enough material for this project, even though I just did the search.	Wilson (1996)
I fail to recognize this as relevant, even though it really may be relevant.	Wilson (1996)
I am deliberately choosing to ignore this item.	Wilson (1995)
I wish to defer thinking about this item even if it may be relevant.	Wilson (1995)
This item is too specialized for me.	Wilson (1995)
Even though I think this item is relevant, I may safely ignore it.	Wilson (1995)
The document is written in a language I do not understand.	Hjørland (2000)
It will take too much time for me to decide whether this item is relevant, so I am ignoring it.	Hjørland (2000)
The perspective from which the item appears to be written is inappropriate for my purposes.	Hjørland (2000)

imation to saying it is not relevant to the user's information need at that point in time.

3. Measures of User Behavior During a Web-Based Catalog Search

One approach to characterizing user behavior when searching a Web-based library catalog has been presented in Cooper (in press). In that paper, a *session* is defined as beginning when a user makes a connection to the catalog through his or her browser. The session continues until the user overtly disconnects from the library catalog or until the library system's computer detects inactivity for a predetermined *timeout* period. A number of different activities can take place within a session, such as searching or displaying results. That paper also describes a number of macro-level measures charactering a session, such as how users divide their session actions and time between presearch (before the first search takes place), search, display, help, and other activities. It also shows that there are major differences in usage patterns between databases, there are differences in the amount of help requested over a period of time, and that the number of citations displayed per search changes over time.

The current paper takes a micro-level view of the activities that take place during a session. The focus is on developing a set of measures that characterize, in detail, a user's actions during a session. One goal of this activity is to use those variables to predict when a session will be relevant, as defined in the previous section. Other goals are to use the same variables to analyze user behavior using continuous-time stochastic models, to analyze transition probabilities from user search state to search state, and to locate time-invariant sequential usage patterns within a user session. This later work is reported in Chen (2000).

The variables used to characterize a session can be divided into two categories: (1) base variables, and (2) derived variables. Base variables are accumulated from transaction log processing. Derived variables—sums, proportions, and rates that are derived from the base variables, serve, in general, as the independent variables describing the session. There is the potential for considerable multicollinearity in both sets of variables characterizing a session. To reduce it, Principal Components Analysis is used to derive a third and final set of variables that are employed in the logistic regression analysis.

4. Base Variables

The set of base variables is given in Table 3. Note that none of these variables records any demographic information about the user. Of course, it would be relatively straightforward to ask the user personal questions, or to ask the user questions about their perceived relevance of the session once it was completed: we could simply ask if the user would be willing to participate in a survey, and if so, a Web page could appear with appropriate questions (see Cooper, 1998 for the type of information that could be collected). But this was not done.

Because we are all concerned about privacy, especially on the Internet, this study is constructed around the hypothesis that it may be possible to deduce whether a session is relevant to the user without knowing anything personal about the user.

The base variables fall into six categories: relevance indicator variables, session variables, search variables, display variables, error variables, and help variables. The relevance indicator variables count the number of print (LP), mail (LM), download (LD), and save (LS) actions the user performs and serve as a surrogate for a relevance measure and for the dependent variable in the analysis that follows.

One of the session variables (PF) indicates whether the searcher activates a stored profile of their session preferences, such as e-mail address, the language in which the materials they retrieve should be written, or the branch library in which the materials must be located. Users have the option of allowing the system to record these preferences and the option of activating the profile when a search commences.

Other session variables include the number of different databases (TD), the number of different indexes (TI), the session length in seconds (SL), and the number of searches performed in a session (SQ). There are also variables that record the total number of Web pages requested (NP), and the the number of different pages requested (TP).

The search variables characterize the type of searching done during a session. Various approaches to searching, such as author (SA), title (ST), subject (SU), and power (SW), are possible in the Melvyl system. A power search is one in which a number of different indexes can be used at one time in the same search. Once the user has completed a search, the Melvyl system allows the user to find more citations like the one(s) retrieved or to find fewer citations. These actions are recorded in the FR, FH, and FL variables.

The display variables characterize the display of the search results, including the amount of time the user spends displaying citations (RD), the number of records that are displayed one-to-a-screen (OC) vs. multiple records on one screen (MC). In addition, the number of different display formats used in showing results on the screen (DF) is recorded, as is the time spent viewing sing-record screens (VI) and multiple-record screens (VS). Finally, information about the time spent and the number of errors and help requests are recorded.

5. Derived Variables

Although the base variables provide a very comprehensive quantitative picture of a session, various sums, proportions, and rates derived from them can extract meaningful information about a session. The derived variables can, in a quantitative manner, differentiate one searcher from another, relevant from nonrelevant searches, one behavior

TABLE 3. Base variables used to characterize a Web-based catalog search.

Symbol	Description
Relevance variables	
LP	The number of times during a session the user requests citations be formatted for printing on the user's local computer (Print)
LM	The number of times during a session the user requests citations be sent by e-mail (Mail)
LD	The number of times during a session the user requests citations be downloaded to their local computer (Download)
LS	The number of times during a session the user requests that citations be saved to a list that can be later formatted for printing at the local machine, downloaded to the local machine, or sent by e-mail to the user (Save)
Session variables	
PF	An indicator of whether the user activates a stored profile of their preferences (e.g., language of retrieved items, e-mail address for mailing materials, campus on which to locate materials)
TD	The number of different databases used during a session
TI	The number of different indexes used during a session
SL	The length of a session in seconds
SQ	The number of searches performed during a session
NT	The total number of items retrieved in a session
SR	The number of searches in a session for which the user retrieves more than zero records (searches with retrievals)
NP	The number of Web pages requested during a session
TP	The number of different Web pages used during a session
PL	The number of Web pages the user accesses before beginning the first search of a session (presearch)
TS	The time in seconds the user spent 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 in a session
SB	The number of searches in a session that use Boolean operators
CS	The number of times the user changed the search index (e.g., author, title, publisher) used during a session
FL	The number of times a user requests the system to find less citations than the current search retrieves
FR	The number of times a user requests the system to find more citations like the current one
FH	The number of times a user requests the system to find more citations like the current one after the user has performed a power search
Display variables	
RD	The total time (seconds) in a session spent on record display activities
OC	The total number of single-record display screens used during a session
MC	The total number of multiple-record display screens used during a session
SO	The number of searches in a session that are followed by the display of a single record
SP	The number of searches in a session that are followed by the display of multiple records
DF	The number of different types of display formats used during a session
VD	The total viewing time (seconds) for single-record display screens during a session
VS	The total viewing time (seconds) for multiple-record display screens during a session
Error variables	
SE	The number of searches that have system-detected errors in them
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
Help variables	
HP	The number of Web pages containing help text that the system displays during a session (a measure of the help rate)
TH	The number of different types of help requests made during a session
SH	The total time (seconds) spent in help activities during a session

pattern from another, and experienced from inexperienced searchers.

Table 4 summarizes the derived variables and shows the formulas used in their computation. The formulas rely on the base variables from Table 3 for their derivation. In a few cases, indicated in the table, some logic besides a simple formula is necessary for the variable's derivation. The variables are again divided into six categories: an effectiveness variable, session variables, search variables, display variables, error variables, and help variables.

The variable *R* is the session relevance indicator. If a search in a session is followed by a display action, and then followed by a print (LP), mail (LM), download (LD), or save action (LS), then the relevance indicator *R* for the session is set to 1 (indicating relevance); otherwise it is set to 0 (indicating a nonrelevant session).

Session variables include the average number of hits per search (HT), the average search length (SD), average time between Web page requests (VT), and proportion of total time spent in presearch activity (PP).

The number of derived search variables is relatively large. Some capture the relative number of searches with retrieval (RR), author searches (RA), title searches (RT), and so forth. Others quantify the rate at which Boolean

TABLE 4. Derived variables used to characterize a Web-based catalog search.

Symbol	Description	Formula
Effectiveness variable		
R	An indicator of whether the session is relevant	R = If a search is followed by a display action, and then followed by any of the print, mail, download, or save actions, R is set to 1, otherwise it is set to 0
Session variables		
HT	The average number of items retrieved per search when the search retrieves one or more items (average number of hits)	HT = NT/SR
SD	The average search length in seconds (excludes presearch activities)	SD = (SL − TS)/SQ
VT	The average time in seconds between Web page display requests	VT = SL/NP
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 proportion of all searches in a session that are subject searches	RU = SU/SQ
RW	The proportion of all searches in a session that are power searches	RW = SW/SQ
SK	The number of known-item searches in a session	SK = SA + ST
RB	The proportion of all searches in a session that use Boolean operators	RB = SB/SQ
MD	The number of searches that are followed by other searches that modify the original search in any way (search modifications)	MD = CS + FL + FR + FH
RC	The proportion of all search modifications in which the user switches from searching using one index to another (e.g., title to author)	RC = CS/MD
MM	The number of times a user requests the system to find more citations like the current one	MM = FR + FH
PM	The proportion of all search modifications that are related to finding more citations like the one the user already found	PM = MM/MD
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 less citations than the current search retrieved	RL = FL/MD
RM	The average number of search modifications in a session	RM = MD/SQ
Display variables		
AV	The proportion of the total time in a session that the user spent displaying records	AV = (VS + VD)/SL
NR	The average number of times per search the user displays a single record on the screen	NR = OC/SQ
NS	The average number of times per search the user displays multiple records on the screen	NS = MC/SQ
RO	The proportion of all searches that are immediately followed by the display of a single citation	RO = SO/SQ
RP	The proportion of all searches that are immediately followed by the display of multiple citations	RP = SP/SQ
Error variables		
ER	The proportion of searches that have system-detected errors in them	ER = SE/SQ
RE	The proportion of all Web pages displayed during a session that relate to errors (the error rate)	RE = NE/NP
Help variables		
PH	The proportion of the total session time the user spent 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 relate to help activities (the help rate)	RH = HP/NP

operators are implicitly used in a search (RB), the average number of search modifications in a session (RM), and the average number of Web pages requested during a search (SI).

Most derived display variables (NR, NS, RO, RP) try to differentiate between the rate at which single- vs. multiple-record displays of citations are requested. Another (AV) measures the proportion of the total time in a session spent

displaying records. And, finally, error and help rates and times are derived.

6. Methodology

The major goal of this research is to predict the relevance of a session using the relevance indicator discussed earlier. This prediction process involves using principal components analysis, followed by logistic regression analysis to form the prediction equations. We also use a stratified random sampling experimental design to divide the population of observations into ten test groups, and then evaluate the predictions on the test strata.

Relevance is predicted by forming a regression equation in which the relevance indicator variable is regressed against some of the variables in Tables 3 and 4. In this paper, the relevance indicator is a binary variable, which indicates whether the session is considered relevant by the user. The number of times the user performs a print, mail, download, or save operation is not considered, only the fact that at least one of these operations took place. Because the independent variable is binary, the linear regression model is not appropriate, and logistic regression must be used.

7. Logistic Regression

In multiple linear regression, the model used is

$$R = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

where R is the dependent variable (the indicator of relevance), the β_is are coefficients, the x_is are the independent variables, and ε is the error term. For this regression model to hold, the error term must be independent of the x_is; that is, the error term must be uncorrelated with the xs. In addition, the mean of the distribution of the error term must be 0 and the variance equal to σ^2. Finally, the error terms must be normally distributed.

Consider the case of our binary relevance variable, where a session is marked as 0 if it is not relevant (contains no PMDS actions), and 1 if it is relevant. Because a binary variable can take on only two values, the error term cannot have a normal distribution. Further, the variance of the error term is not equal to σ^2. Thus, the assumptions necessary for the linear regression model do not hold in this case.

An alternative is to use the logistic regression model. The logistic model uses the concept of odds and odds ratios in its calculations. Let p be defined as the probability of an event. Then the odds, O, of an event is defined as

$$O = \frac{p}{1 - p}$$

There is a direct relationship between probability and odds:

$$p = \frac{O}{1 + O}$$

Thus, for example, a probability of 0.5 corresponds to odds of 1, and a probability of 0.2 corresponds to odds of 0.25. In the analysis of logistic regression results, it is much easier to think about the odds of certain occurrences than the probability.

The logistic regression model calculates the estimated odds of an event as follows:

$$\log\left[\frac{p_i}{1 - p_i}\right] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$$

where the αs, βs, and xs are defined as before.

8. Principal Components Analysis

The methodology described so far is to form a logistic regression equation with the probability of a session being relevant regressed against some of the variables in Tables 3 and 4. Because the variables may be correlated, there is the possibility of multicollinearity among them, which could make the results of the regression analysis invalid or misleading. To minimize this possibility, the logistic regression analysis is preceded by a preprocessing step: a principal components analysis is performed on the raw data to transform it into a new set of variables, which reduces multicollinearity.

The new set of variables, called *principal components*, has the desirable property of being a linear combination of the original variables and can be used in place of them in a regression equation. The n-independent variables are transformed into n principal component variables such that the coefficients of the transformed variables are equal to the eigenvectors of the correlation matrix.

Consider the set of variables

$$X = (X_1, X_2, \ldots, X_n)$$

The goal of principal components analysis is to find a new set of variables

$$Y = (Y_1, Y_2, \ldots, Y_n)$$

such that the Ys are uncorrelated, that each Y_i is a linear combination of the Xs, and the variances of the Ys decrease from left to right in the list.

Using the principal components methodology, the new set of Y_j variables are formed from the original Xs using the following equation:

$$Y_j = a_{j1} X_1 + a_{j2} X_2 + \cdots + a_{jn} X_n$$

where the a_js are called *component loadings*. The component loadings express the degree to which each of the X variables correlates with each of the Y components.

The result of the principal component analysis is a new set of variables, Y_j, to use in place of each variable X in the regression equation. Now the logistic regression equation becomes

$$R_j = a_{j1}Y_1 + a_{j2}Y_2 + \cdots + a_{jn}Y_n$$

To summarize our revised methodology so far, take one part of the raw data file, and transform the variables through principal components analysis to derive new principal components variables that are linear combinations of the old variables but that are uncorrelated. Next, use the principal components variables as inputs to the logistic regression model to derive the coefficients of the regression equation that fit the data (relevance indicator versus principal components variables). Finally, use the coefficients derived from the logistic regression on another set of data to see how accurately the equation can predict the relevance of a session.

9. Sampling Methodology

The data for this experiment consists of a smaller subset of the dataset described in Cooper (in press). This experiment uses all 905,970 sessions conducted with the Melvyl Web-based library catalog between February and October 1998. The methodology followed in this paper is one proposed by Breiman et al. (1984). His goal was to determine the predictive ability of a solution. Our goal here is similar—to see how well the logistic regression equation can predict the relevance of a session using the transformed variables. Breiman uses random sampling to divide the population into strata; in our case, the sessions were randomly sampled and divided into ten strata of 90,597 sessions each, using the relevance indicator as the stratification variable. The next step is to run the logistic regression on nine of the ten strata and obtain the coefficients of the regression equation. Finally, we take those coefficients, form a regression equation, and use the data from the remaining stratum to predict the relevance outcome for the cases in that stratum.

To determine whether the predictive results are consistent, this process is performed ten times and the results compared. In the first run, strata 1–9 are used in determining the regression coefficients, then stratum 10 is used for testing. In the second run, strata 2–10 are used in determining the coefficients and stratum 1 is used for testing, and so forth.

The sampling methodology involved using simple random sampling in which a sample of n elements is selected from a population " . . . in such a manner that each *combination* of n elements has the same chance or probability of being selected as every other element" (Hansen, 1953, p. 12). In stratified random sampling, the population is divided into groups (in this case ten), and the selection of sessions to be included in a group is done without replacement. That is, once a session is placed in one group, it cannot be placed in another group. The selection is done such that the proportion of sessions that are defined as relevant in the population is maintained in each stratum. In this case, 82.15% of the sessions in the population of 905,970 sessions are not relevant, and 17.85% are relevant. With the stratified random sampling without replacement methodology, where stratification takes place using the relevance indicator variable, this same proportion of relevant vs. nonrelevant sessions is present in the ten strata containing 90,597 sessions each.

TABLE 5. Frequency distribution of print, mail, download, and save actions.

	Percent of sessions			
Frequency per session	Print	Mail	Download	Save
0	91.97	94.27	96.14	90.47
1	4.39	2.78	2.42	5.87
2	1.83	1.45	0.83	1.52
3	0.79	0.58	0.29	0.69
4	0.42	0.35	0.14	0.44
5	0.23	0.18	0.06	0.29
6	0.13	0.12	0.04	0.20
7	0.08	0.07	0.02	0.13
8	0.05	0.06	0.02	0.10
9	0.03	0.03	0.01	0.06
10	0.02	0.02	0.01	0.06
>10	0.06	0.08	0.02	0.12

N = 905,970 for all actions.

10. Descriptive Data

The independent variable in the logistic regression equation is an indicator of the relevance of the session. As discussed earlier, 17.85% of the sessions are defined as relevant according to the proposed definition. That definition required that the user take one of the print, mail, download, or save search actions during a session. Table 5 shows the percent distribution of these actions in a session. For example, in 90.47% of all sessions, users did not save any citations from their searches, and in 94.27% of all sessions, users did not e-mail citations to themselves.[3] Table 5 shows that the actions that make up the relevance indicator occur very infrequently, and Table 6 shows, surprisingly, that the correlation between the actions is relatively small. Saving takes place most frequently (Table 5), but its correlation is never greater than 0.21 (for *print*) for any action.

[3] There is no way of knowing from the transaction log whether users click on the print button on their browser, or cut and paste from the browser window. It is possible to instrument users browsers (see Choo et al., 1998) to capture this information, but that is a problematical solution given a geographically distributed user group and concern for user privacy.

TABLE 6. Correlations among relevance indicator variables.

Variable	Description	LP	LM	LD	LS
LP	Print	1.000	0.128	0.038	0.212
LM	Mail	0.128	1.000	0.098	0.158
LD	Download	0.038	0.098	1.000	0.143
LS	Save	0.212	0.158	0.143	1.000

$N = 905{,}970$.

Thus, if a user saves a citation in a session, the most likely event (but with only a 0.21 correlation) is that the user will print a citation.

There are very interesting differences in the mean values of the base variables between relevant and nonrelevant sessions (see Table 7). A relevant session is more than twice as long as a nonrelevant session (1437 vs. 703 sec). Likewise, a user in a relevant session views more than twice as many Web pages. In a relevant session, an average of 1.56 databases are used vs. 1.33 in a nonrelevant session, thus indicating a broader search. Similarly, 2.01 different indexes are used to access the databases vs. 1.67 for a nonrelevant session. Relevant sessions average 27 sec displaying records vs. 19 sec for nonrelevant sessions, and significantly more single- and multirecord displays are requested for relevant sessions.

Although there are a few base variables that are similar for relevant and nonrelevant sessions, the viewing time per record for single- and multiple-record displays is almost the same for each category, and the number of author searches is comparable. In general, the sessions have very different means for the base variables, indicating a true partitioning.

Similar differences exist in the mean values of the derived variables between the session types (see Table 8). Overall, a relevant search retrieved an average of 528 hits per session, whereas a nonrelevant session retrieved 380. A user who has a relevant session spends a very low propor-

TABLE 7. Mean values for base variables.

		All sessions		Nonrelevant sessions		Relevant sessions	
Symbol	Description	Mean	SD	Mean	SD	Mean	SD
	Sessionvariables						
PF	Active profile	0.05	0.23	0.04	0.19	0.12	0.33
TD	Databases used	1.37	0.90	1.33	0.84	1.56	1.15
TI	Indexes used	1.73	1.32	1.67	1.22	2.01	1.65
SL	Session length (seconds)	833.77	1023.02	702.77	921.60	1436.58	1229.57
SQ	Searches performed	4.24	5.21	3.93	4.80	5.64	6.62
SR	Searches with retrievals	2.66	3.26	2.39	2.95	3.92	4.20
NP	Web pages requested	19.53	21.69	16.10	17.32	35.35	30.85
TP	Unique Web pages	7.09	3.42	6.36	2.69	10.44	4.31
PL	Web pages in presearch	2.36	1.45	2.29	1.22	2.67	2.19
TS	Presearch time (seconds)	62.02	55.45	60.04	53.44	71.24	63.14
	Search variables						
SA	Author searches	0.83	2.38	0.82	2.23	0.89	2.97
ST	Title searches	0.96	2.69	1.00	2.69	0.75	2.66
SU	Subject searches	1.60	3.26	1.39	2.95	2.57	4.28
SW	Power searches	0.84	2.83	0.72	2.53	1.39	3.87
SB	Boolean searches	2.58	3.94	2.34	3.63	3.68	4.99
CS	Unique indexes used	0.91	2.60	0.76	2.22	1.60	3.82
FL	Find-less searches	0.21	0.84	0.18	0.76	0.36	1.12
FR	Find-more searches	0.12	0.54	0.11	0.51	0.15	0.67
FH	Find-more following power search	0.10	0.53	0.08	0.48	0.17	0.69
	Display variables						
RD	Display time (seconds)	20.38	28.85	18.92	28.42	27.07	29.82
OC	Single-record display screens	2.77	5.58	2.44	5.00	4.31	7.51
MC	Multiple-record display screens	5.06	8.17	3.83	6.34	10.70	12.26
SO	Searches with single-record displays	1.13	1.77	1.05	1.68	1.50	2.08
SP	Searches with multiple-record displays	1.88	2.30	1.67	2.08	2.84	2.93
DF	Unique display formats	1.66	1.03	1.54	0.97	2.21	1.14
VD	Single-record display viewing time	33.48	94.29	33.48	99.52	33.49	65.07
VS	Multiple-record display viewing time	58.18	107.04	59.06	116.08	55.13	66.83
	Error variables						
SE	Error count	0.34	1.19	0.29	1.11	0.54	1.48
TE	Unique errors	0.23	0.48	0.20	0.45	0.37	0.58
NE	Web pages of errors	0.52	1.87	0.44	1.75	0.86	2.33
	Help variables						
HP	Web pages of help	0.05	0.36	0.04	0.34	0.07	0.43
TH	Unique help requests	0.04	0.25	0.03	0.24	0.05	0.30

TABLE 8. Mean values for derived variables.

Symbol	Description	All sessions Mean	All sessions SD	Nonrelevant sessions Mean	Nonrelevant sessions SD	Relevant sessions Mean	Relevant sessions SD
	Sessionvariables						
HT	Average number of hits	405.93	1921.57	379.51	1859.35	527.50	2181.01
SD	Average search length	202.73	290.34	173.50	270.00	337.86	338.80
VT	Average time between Web page displays	37.54	33.20	36.10	33.44	44.02	31.27
PP	Presearch time proportion	23.24	24.85	26.13	25.99	9.93	11.58
PS	Presearch Web page proportion	22.55	16.28	24.71	16.47	12.57	10.73
	Search variables						
SI	Web pages requested	4.64	4.87	3.89	3.58	8.10	7.71
RR	Searches with retrievals	72.48	33.81	70.59	35.27	81.17	24.26
RA	Proportion of author searches	22.44	39.68	23.60	40.49	17.12	35.25
RT	Proportion of title searches	25.46	41.54	28.04	42.94	13.58	31.74
RU	Proportion of subject searches	37.20	46.21	34.50	45.59	49.66	47.03
RW	Proportion of power searches	14.89	33.49	13.86	32.56	19.64	37.08
SK	Known item searches	1.79	3.53	1.82	3.42	1.64	4.00
RB	Proportion of Boolean searches	56.52	42.57	55.18	43.05	62.66	39.73
MD	Searches followed by modifications	1.19	2.74	1.00	2.35	2.05	3.97
RC	Proportion of change indexes	25.64	41.47	23.69	40.61	34.62	44.15
MM	Number of find-more searches	0.21	0.79	0.19	0.73	0.32	1.01
PM	Proportion of find-more searches	8.98	26.18	8.65	26.00	10.47	26.95
RK	Proportion of known-item searches	47.90	47.99	51.64	48.11	30.70	43.48
RL	Proportion of find-less searches	7.23	23.41	6.49	22.46	10.67	27.08
RM	Average number of search modifications	0.00	0.15	0.00	0.10	0.00	0.32
	Display variables						
AV	Display time proportion	39.81	30.35	38.26	31.52	46.95	22.96
NR	Average number of single-record displays	1.44	2.54	1.31	2.30	2.02	3.36
NS	Average number of multiple-record displays	2.31	3.67	1.83	2.86	4.49	5.65
RO	Proportion of single-record displays	32.76	37.31	32.85	37.98	32.35	34.07
RP	Proportion of multiple-record displays	54.47	37.82	52.31	38.84	64.42	30.87
	Error variables						
ER	Proportion of system errors	6.38	17.98	5.65	17.09	9.78	21.26
RE	Web page error rate	2.29	6.75	2.32	7.08	2.18	4.95
	Help variables						
PH	Proportion of time using help	0.26	2.71	0.26	2.76	0.26	2.42
VH	Average time per page for help	0.93	8.06	0.78	7.26	1.61	10.99
RH	Web page help rate	0.23	1.81	0.24	1.90	0.19	1.28

tion of time before beginning the first search (presearch time) compared to the user with a nonrelevant session.

The search variables show that the categories have different search approaches. A nonrelevant session uses more author and title searches, whereas a relevant session mostly employs subject searches. Relevant sessions also have more Boolean searching, more search modification, and more time in display activities. Error and help rates are roughly comparable. Again, the differences between the categories are striking and reveal very different patterns of behavior, which reinforces the idea that the simple relevance variable may be a good categorization tool.

One other interesting variable is whether the user activates a stored profile of session preferences (indicating a preferred language for retrieved citations, a campus to which to limit the retrieved set of citations, or a preferred e-mail address for mailing saved citations). Of the user sessions with an activated profile (only 5% of sessions), 41% of the sessions were relevant, and 59% nonrelevant. When no profile was activated, 17% of the sessions were relevant, and 83% nonrelevant. Users who activate a profile are much more likely to have a relevant session.

11. Principal Components Results

Principal components analysis was performed, not to reduce the number of variables used to predict relevance, but to reduce the likelihood of multicollinearity in the variables.[4] The 44 variables used as inputs were:

AV DF ER HP HT MC MM NE NP NR NS OC PF PH

PL PM PP PS RA RB RC RE RH RK RL RO RP RR RT

RU RW SD SI SL SQ TD TE TH TI TP TS VD VH VT

[4] The principal components analysis was performed by the SAS PRINCOMP procedure (SAS Institute Inc., 1990).

Of these, the base variables (Table 3) are DF (display formats), HP (Web pages containing help text), MC (number of multirecord displays), NE (Web pages about errors), NP (Web pages requested during a session), OC (number of single-record displays), PF (profile indicator), PL (Web pages in presearch), SL (session length), SQ (searches performed), TD (different databases), TE (different errors), TH (different help requests), TI (different indexes), TP (different Web pages in session), TS (time in presearch), and VD (viewing time for single-record displays). The other variables in the list are the derived variables from Table 4.

The result of the principal components analysis was 44 new variables, called PRIN1, PRIN2, . . . , PRIN44, that from now on take the place of the 44 original variables listed previously.

The ordering of the principal component values resulting from the analysis (i.e., PRIN1, PRIN2, . . .) has some significance: by using only the first 23 of the 44 principal component variables, 90.92% of the variance can be explained. As one moves toward the end of the list of variables, like PRIN43 and PRIN44, they contribute nothing to explaining variance, and were dropped.

12. Logistic Regression Results

Once the principal components analysis was completed, the original variables were no longer used. Instead, the principal component variables took their place in the logistic regression equation:

$$\log\left[\frac{p_i}{1-p_i}\right] = \alpha + \beta_1 PRIN1 + \beta_2 PRIN2 + \cdots + \beta_{42} PRIN42$$

where the αs, βs, and the PRINs are defined as before.

Ten logistic regression analysis runs were made.[5] In each run, nine different strata of data were used in the regression. Thus, run number 5—3 is the logistic regression analysis on strata 5, 6, 7, 8, 9, 10, 1, 2, and 3 (815,373 sessions out of 905,970). This particular equation has the form:

$$\log\left[\frac{p_i}{1-p_i}\right] = -2.3917 + 0.5276^{*}\, PRIN1 - 0.1912^{*}\, PRIN2 + \cdots - 0.2675^{*}\, PRIN42$$

Recall that the 905,970 observations were randomly assigned to the strata such that the proportion of relevant sessions was constant in each stratum. Because of the overlap in the observations that form each set for the logistic analysis, the coefficients are similar for each run.

Because the regression variables are principal component values, we cannot easily interpret the meaning of the coefficients. That is, one cannot say that a particular principal component variable represents a certain variable because it is a linear combination of other variables. Thus, we cannot say that PRIN37 has a high positive value, and this means that, say, display viewing time is strongly related to relevance. The variables serve only as a guide to the next stage of the analysis.

13. Stepwise Logistic Regression Results for One Run

Each of the ten logistic regression runs was completed using the same methodology, stepwise logistic regression. The regression program picked the order in which the principal component variables were entered into the equation. Rather than present the results of all ten runs, we have selected one, the regression for strata 3 through 1 with stratum 2 excluded, for detailed analysis.

Table 9 gives the parameter estimates, the standard error, the χ^2 statistic, and the probability for the null hypothesis that each coefficient is equal to zero. All the parameters are significant except PRIN35, which accounts for the least variance. The last column in Table 9 is the odds ratio. This is computed as e^{β}. For example, the odds ratio of PRIN1 is $e^{0.5259}$, or 1.692. Recall that odds of 1.00 can be translated into a probability of 0.5. The odds are greater than 1.00 for 15 of the 42 variables in the equation, indicating that some of the variables could have been omitted without loss of precision.

As a results of the analysis, the null hypothesis that all the explanatory variables in the equation have coefficients of zero can be rejected at the 0.0001 level (42 degrees of freedom).[6] Thus, at least one of the coefficients in the equation is not zero.

14. Multicollinearity Validation

As a check on whether multicollinearity was present in the logistic regression, a further step was performed. For each variable, the tolerance was computed. Tolerance is defined as $1 - R^2$ where R^2 is obtained by regressing each variable on all other regressors in the model (see Allison, 1999, pp. 50–51 for details).

[5] The SAS procedure LOGISTIC was used for the analysis (SAS Institute Inc., 1990). Each of the runs took from 6 to 7 hr to complete on a SUN Enterprise 5000 Server.

[6] All three computed tests have the same significance. The likelihood ratio χ^2 value is 269,097, the score χ^2 value is 265,370, and the Wald χ^2 value is 133,668. The values of two other statistics are worth noting. The Akaike's information criterion has a computed value of 728,863 for intercept only and 459,849 for intercept and covariates. And the Schwartz criterion has values of 728,874 and 460,346. These values are used when the fit of the present model is compared to other models, and by themselves are not useful.

TABLE 9. Logistic regression maximum likelihood parameter estimates and odds ratios for strata 3–1.

Principal component variable	DF	Parameter estimate	Standard error	Chi-square value	Pr > Chi^2	Odds ratio
Intercept	1	−2.3874	0.0060	160975.7	<.0001	
PRIN1	1	0.5259	0.0022	59713.5	<.0001	1.692
PRIN2	1	−0.1914	0.0022	7290.9	<.0001	0.826
PRIN3	1	0.0936	0.0025	1381.4	<.0001	1.098
PRIN4	1	0.1804	0.0025	5437.8	<.0001	1.198
PRIN5	1	−0.1234	0.0028	1916.6	<.0001	0.884
PRIN6	1	0.1233	0.0026	2184.8	<.0001	1.131
PRIN7	1	−0.0130	0.0033	15.7	<.0001	0.987
PRIN8	1	−0.3694	0.0031	14073.4	<.0001	0.691
PRIN9	1	0.3409	0.0034	9796.3	<.0001	1.046
PRIN10	1	0.0809	0.0036	511.1	<.0001	1.084
PRIN11	1	0.0877	0.0034	655.2	<.0001	1.092
PRIN12	1	−0.4919	0.0040	14822.5	<.0001	0.611
PRIN14	1	−0.4584	0.0045	10212.9	<.0001	0.632
PRIN15	1	0.0592	0.0040	221.8	<.0001	1.061
PRIN16	1	−0.0838	0.0038	479.8	<.0001	0.920
PRIN17	1	0.1563	0.0043	1341.3	<.0001	1.169
PRIN18	1	−0.3654	0.0046	6228.3	<.0001	0.694
PRIN19	1	−0.2767	0.0045	3786.7	<.0001	0.758
PRIN20	1	0.0453	0.0052	76.2	<.0001	1.046
PRIN21	1	0.2237	0.0061	1357.3	<.0001	1.251
PRIN22	1	−0.3345	0.0054	3830.7	<.0001	0.716
PRIN23	1	−0.3735	0.0054	4776.9	<.0001	0.688
PRIN24	1	−0.4038	0.0062	4252.0	<.0001	0.668
PRIN25	1	−0.3830	0.0059	4268.8	<.0001	0.682
PRIN26	1	−0.3972	0.0072	3010.9	<.0001	0.672
PRIN27	1	−0.2531	0.0068	1371.5	<.0001	0.776
PRIN28	1	−0.6563	0.0067	9676.6	<.0001	0.519
PRIN29	1	−0.1710	0.0078	481.9	<.0001	0.843
PRIN30	1	−0.3067	0.0083	1355.6	<.0001	0.736
PRIN31	1	−0.9064	0.0082	12276.2	<.0001	0.404
PRIN32	1	−0.8864	0.0085	10766.7	<.0001	0.412
PRIN33	1	−0.0836	0.0093	81.6	<.0001	0.920
PRIN34	1	−0.1277	0.0087	215.7	<.0001	0.880
PRIN35	1	0.0266	0.0117	5.1	0.0233	1.027
PRIN36	1	−0.0604	0.0130	21.6	<.0001	0.941
PRIN37	1	0.8888	0.0109	6695.9	<.0001	2.432
PRIN38	1	0.4689	0.0111	1800.6	<.0001	1.598
PRIN39	1	0.3674	0.0136	727.7	<.0001	1.444
PRIN40	1	−0.3496	0.0131	714.6	<.0001	0.705
PRIN41	1	−0.4342	0.0178	597.3	<.0001	0.648
PRIN42	1	−0.2456	0.0264	86.4	<.0001	0.782

The results are given in Table 10 for the same test stratum used here. The table lists the principal component variables along with the tolerance value and its reciprocal, the variance inflation factor. To quote Allison (1999, p. 50), the variance inflation factor " . . . tells you how 'inflated' the variance of the coefficient is, compared to what it would be if the variable were uncorrelated with any other variable in the model." As a rule of thumb, if the tolerance is below 0.40, it may indicate multicollinearity. None of the variables meet this criterion, thus indicating that the principal components analysis has done its job.

15. Prediction Results

The final stage in the analysis is to predict the relevance of sessions. Each of the ten logistic regression runs omitted one stratum in computing the coefficients. The run that computed coefficients for strata 1 to 9 omitted stratum 10; the run for strata 8 through 6 omitted stratum 7. Relevance was predicted by using the coefficients for the included strata and applying them against the sessions of the excluded stratum. For example, the coefficients in column 3 of Table 9 were used to predict the relevance of the 90,597 sessions in stratum 2.

The criterion used to determine whether a session was relevant was based on the computed probability, P, from the prediction equation:

$$P = \frac{1}{1 + e^{-R_p}}$$

TABLE 10. Tolerance values and variance inflation factors for principal component values of strata 3–1.

Principal component variable	Tolerance	Variance inflation
PRIN1	0.4975	2.0101
PRIN2	0.6746	1.4823
PRIN3	0.7202	1.3885
PRIN4	0.7839	1.2757
PRIN5	0.8484	1.1787
PRIN6	0.8722	1.1465
PRIN7	0.7029	1.4227
PRIN8	0.8596	1.1633
PRIN9	0.8302	1.2045
PRIN10	0.7646	1.3079
PRIN11	0.9215	1.0852
PRIN12	0.8263	1.2103
PRIN13	0.9444	1.0589
PRIN14	0.7225	1.3840
PRIN15	0.9528	1.0495
PRIN16	0.9543	1.0479
PRIN17	0.9211	1.0857
PRIN18	0.8697	1.1499
PRIN19	0.8695	1.1500
PRIN20	0.8474	1.1801
PRIN21	0.7211	1.3868
PRIN22	0.9070	1.1026
PRIN23	0.9225	1.0840
PRIN24	0.8529	1.1725
PRIN25	0.8944	1.1181
PRIN26	0.7906	1.2649
PRIN27	0.9415	1.0622
PRIN28	0.8863	1.1283
PRIN29	0.8298	1.2051
PRIN30	0.8811	1.1350
PRIN31	0.8959	1.1162
PRIN32	0.8882	1.1259
PRIN33	0.8225	1.2159
PRIN34	0.8835	1.1318
PRIN35	0.7226	1.3839
PRIN36	0.8782	1.1387
PRIN37	0.8816	1.1343
PRIN38	0.8610	1.1615
PRIN39	0.9367	1.0676
PRIN40	0.9165	1.0912
PRIN41	0.8381	1.1932
PRIN42	0.9143	1.0937
PRIN43	0.9977	1.0023
PRIN44	0.9978	1.0022

The prediction equation for R_p is given by

$$R_p = \alpha + \beta_1 PRIN1 + \beta_2 PRIN2 + \cdots + \beta_{42} PRIN42$$

where R_p is the predicted relevance value, the α and the β_is are given for one case in Table 9, and the values of PRIN1—PRIN42 are associated with each session in the omitted stratum.

If the probability value $P < 0.5$, the session is deemed not to be relevant, and if $P >= 0.5$, it is predicted to be relevant. The latter is equivalent to saying that if the odds are greater than or equal to 1, the session is considered relevant.

Table 11 summarizes the ten prediction runs. The total number of nonrelevant sessions in the entire population of 905,970 sessions is 82.15%, and the number of relevant sessions is 17.85%. These proportions exactly hold in each stratum. The equations predict about 11% of the sessions will be relevant across the ten runs, a 6.85% difference. Thus knowing only observable characteristics of a session, we can predict within 7% whether the session will be relevant!

Does the criterion established for determining relevance of a session substantially influence the outcome? The predicted value is a continuous variable, whereas the criterion is discrete. Figure 2 plots the predicted probability values against the cumulative percent distribution for stratum 2. Note that the curve does not experience a significant jump near the 0.50 value, but rather is smooth. Thus the cutoff does not substantially influence the outcome.

The difference between the expected and predicted number of relevant sessions may be partially explained by the transformation performed on the odds ratio. Recall that the transformation was nonlinear because it used a logarithm to convert the odds ratio to a binary variable. It is possible that better transformations are available that could minimize the loss of information and thus reduce the difference.

TABLE 11. Classification of predicted probability values by sampling strata.

Stratum predicted	Predicted nonrelevant sessions	Number of relevant sessions	Predicted nonrelevant sessions	Percent of relevant sessions
1	80,729	9,868	89.11	10.89
2	80,735	9,862	89.11	10.89
3	80,841	9,756	89.23	10.77
4	80,800	9,797	89.19	10.81
5	80,798	9,799	89.18	10.82
6	80,801	9,796	89.19	10.81
7	80,787	9,810	89.17	10.83
8	80,839	9,758	89.23	10.77
9	80,665	9,932	89.04	10.96
10	80,806	9,791	89.19	10.81

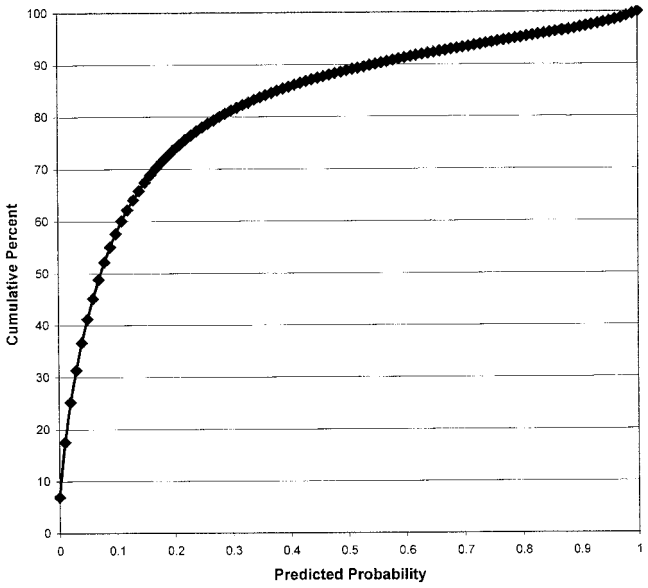

FIG. 2. Cumulative percent distribution of predicted probability values.

16. Conclusions and Implications

This research has shown that it is possible to develop a good operational definition of the concept of "relevance" of a session to a user, and to predict the relevance of the session without any demographic variables about the user. The prediction is based on the way in which the user conducts the session: the time spent performing tasks during the session; and the counts, relative frequencies, and proportions of actions taken during the session. These measures are surrogates for user behavior. For a population of 905,970 sessions, of which 17.85% of the sessions were relevant, the methodology was able to predict that about 11% of the sessions were relevant.

Along what avenues can this approach lead? One is in the direction of best practices and training. If we know the characteristics of a relevant session, we can guide users toward a successful conclusion to their session. If we enhance these techniques, we can use an existing database of best practices and sequential decision-making methods to track a users session, and (if the user is willing) make suggestions for improvement.

The research reported here analyzed server-maintained transaction logs. There was no server-log knowledge of user behavior at the client computer, such as clicking on the print button in a browser window or using cut-and-paste techniques to extract information from the screen of the client. If enhanced logging were available, and the user was willing to be monitored in this way, the predictions could be improved further.

The implications of this research approach extend beyond the prediction of the relevance of a library catalog session. Electronic commerce purchasing screens bear a striking resemblance to the search results screen in Figure 1. The user selects one or more items from the screen to purchase—the equivalent of conducting a relevant session. The methodology given here could be applied to the prediction of that type of behavior. The ability for an electronic commerce organization to predict when a user will buy, or not buy, is extremely important. The organization could use these models to understand each category of behavior, offer assistance (if desired) to those users predicted not to buy, revamp its Web site based on behavior patterns deduced from the models, and, in general, increase the probability of more purchases by its users.

References

Allison, P.D. (1999). Logistic regression using the SAS system: Theory and application. Cary, NC: SAS Institute.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. Belmont, CA: Wadsworth International Group.

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

Choo, C., Detlor, B., & Turnbull, D. (1998). A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and IT specialists use the Web. ASIS '98, Proceedings of the 61st ASIS Annual Meeting, (Vol. 35), Information Today, Medford, NJ, pp. 290–302.

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. (in press). Usage patterns of a Web-based library catalog. Journal of the American Society for Information Science.

Cooper, W.S. (1971). A definition of relevance for information retrieval. Information Storage and Retrieval, 7, 19–37.

Cooper, W.S. (1973a). On selecting a measure of retrieval effectiveness, part 1: The "subjective" philosophy of evaluation. Journal of the American Society for Information Science, 24, 87–100.

Cooper, W.S. (1973b). On selecting a measure of retrieval effectiveness, part 2: Implementation of the philosophy. Journal of the American Society for Information Science, 24, 413–424.

Cuadra, C.A., & Katter, R.V. (1967). Experimental studies of relevance judgments: Final report, 3 Vols., TM-3520/001/00, TM-3520/002/00, and TM-3520/003/00, System Development Corporation, Santa Monica, CA.

Hansen, M., Hurwitz, W.N., & Madow, W.G. (1953). Sample survey methods and theory: Volume 1, Methods and applications. New York: John Wiley & Sons.

Harter, S.P. (1992). Psychological relevance and information science. Journal of the American Society for Information Science, 43, 602–615.

Hjørland, B. (2000). Relevance research: The missing perspective(s): "nonrelevance" and "epistemological relevance" (Letter to the Editor). Journal of the American Society for Information Science, 51, 209–211.

Maron, M.E. (1977). On indexing, retrieval and the meaning of about. Journal of the American Society for Information Science, 28, 38–43.

Mizzaro, S. (1997). Relevance: The whole history. Journal of the American Society for Information Science, 48, 810–832.

Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching (Vol. I). Final Report to the National Science Foundation, NSF Contract Number C-423, Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, OH, PB 176080.

Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. Journal of the American Society for Information Science, 26, 321–343.

Saracevic, T. (1976). Relevance: A review of the literature and a framework for thinking on the notion in information science. In M.J. Voigt & M.H. Harris (Eds.), Advances in librarianship (Vol. 6) (pp. 79–138). New York: Academic Press.

SAS Institute, Inc. (1990). SAS/STAT user's guide (Version 6, 4th ed., Vol. 2). Cary, NC: SAS Institute.

Schamber, L., Eisenberg, M.B., & Nilan, M.S. (1990). A reexamination of relevance: Toward a dynamic, situational definition. Information Processing & Management, 26, 755–776.

Wilson, P.G. (1968). Two kinds of power: An essay on bibliographic control. University of California Publications in Librarianship (Vol. 5). Berkeley, CA: University of California Press.

Wilson, P. (1973). Situational relevance. Information Storage and Retrieval, 9, 457–471.

Wilson, P.G. (1993). Communication efficiency in research and development. Journal of the American Society for Information Science, 44, 376–382.

Wilson, P.G. (1995). Unused relevant information in research and development. Journal of the American Society for Information Science, 46, 45–51.

Wilson, P.G. (1996). Some consequences of information overload and rapid conceptual change. In J. Olaisen, E. Munch-Petersen, & P. Wilson (Eds.), Information science: From the development of the discipline to social interaction. Oslo: Scandinavian University Press.