A Quantitative and Qualitative Investigation into Failures in the Prediction of Relevance in Catalog Searches

Michael D. Cooper

School of Information
University of California
Berkeley, California 94720-4600

January 16, 2017

Abstract

It appears possible to predict when an online library catalog search will be relevant, without interviewing the person performing the search, with a simple operational criterion--whether during a session the user prints, emails, downloads, or saves to a list any retrieved citation. Previous research performed predictions of the relevance of a session using many observable session characteristics such as session length, number of indexes used, viewing time per page, citations retrieved, number of hits viewed, and so on. This paper refines the prediction methodology using logistic regression residual analysis. One residual variable was found to do an excellent job in partitioning a sample of nearly 600,000 sessions with the University of California's Melvyl® catalog into two sets. One set of sessions consisted of outliers--sessions with variations in the observable variables too great to allow consistent prediction. For the other set, the non-outlier sessions, the number of sessions predicted to be relevant differed from the number known to be relevant by only 1.51%.

The second part of the research was a qualitative investigation into why relevance prediction failed. The transaction logs for 400 randomly selected sessions were analyzed to determine distinguishing characteristics of relevant vs. non−relevant sessions. In addition, patterns emerged in the log files that allowed the identification of searching procedures used by experienced and inexperienced searchers. It was also possible to identify the reasons that searches failed, and causes of the prediction algorithm failure. This paper concludes with suggestions for improvement of the algorithm used to predict the relevance of a session.

Introduction

When a user searches a library catalog, a set of documents is retrieved that may be relevant to the query. With a Web−based catalog it is difficult and expensive to personally interview users to ascertain if the result set was indeed relevant. One solution to this problem is to employ a simple observable measure as a surrogate for user satisfaction. Previously, Cooper and Chen (2001) proposed that if a user saves a citation to a list, prints a citation, downloads the citation to his or her computer, or emails the citation, this constituted a recognition that the result was relevant. In that work, the authors examined more than 905,000 search sessions conducted by users of the University of California's Melvyl® catalog. Using the criteria above, the total number of sessions found to be relevant was determined to be 17.82%. A logistic regression model was used to see if it was possible to predict when a session was relevant using a number of observable variables that characterize a session, including the number of items retrieved, the length of the session, the number of databases used, and a host of others. The result of the logistic regression prediction was an estimate that 10.82% of the sessions were relevant, a difference of 7%.

This paper explores two major issues that arise in the prediction of relevance. First, are there mathematical techniques that can be used to refine the prediction process? Second, will a subjective qualitative analysis of the sessions in which there are prediction failures yield insight into search failures and relevance prediction failures?

The motivation for this study is to refine techniques for predicting relevance and predicting search success and failure. Online catalogs are not perfect, and users do not always succeed in finding materials. If techniques can be developed to determine if a user has succeeded in a search, that information would validate the operation of the catalog. Suppose a system can monitor a user's search and detect that the current search is unlikely to produce meaningful results. Such a detection might be based on a comparison of the patterns of known successful searches with the pattern evolving with the current search. When such a situation is detected, the system, with the user's consent, could intervene to suggest alternative strategies. The result would be improved user satisfaction.

Logistic Regression Analysis of Entire Sample

The methodology used to predict whether a user's session with the library catalog produces relevant results is to form a logistic regression equation in which the dependent variable is a binary variable indicating whether the session was determined to be relevant (R). The independent variables (x_i' s) are a series of measures that characterize the session. The model is given as

$$R = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

where $\alpha$is the intercept; the βs are coefficients; the x's are the independent variables; and ϵ is the error term, which will be the focus of discussion in the next section of the paper.

In several recent papers, Chen and Cooper (2001, 2002) and Cooper and Chen (2001) defined and utilized a series of non−demographic base and derived variables (thex_i 's) that characterize a user's session with a library catalog. These variables are given in Tables 1 and 2. They have proven to be important indicators and have formed the basis for a clustering and a stochastic analysis of user behavior. They have also been used to predict session relevance, the topic that is being explored in this research.

In this paper, a logistic regression analysis was performed on 590,458 sessions from the Melvyl® library catalog. The independent variables used in the logistic regression equation were a principal components transformation of the variables in Tables 1 and 2. They were regressed against the measure of session relevance. The results are given in the column labeled "Entire Database" in Table 3.[1] The table gives a number of performance measures for the logistic regression analysis and shows that the R^2 measure of goodness−of−fit is about 0.48.

The coefficients from the regression analysis were used to form a regression equation and predict whether a session was relevant. Table 4, which gives the results of that analysis, shows that out of 590,458 sessions from the Melvyl® online catalog, 17.82% were known to be relevant and 82.18% were known not to be relevant because the user did not perform a print, mail, download, or save operation during the session. The logistic regression predicted that 89.18% of the sessions would not be relevant and 10.82% would be relevant.

The question that this paper addresses is: Can the prediction that a session is relevant be improved? In other words, can the difference between the number of known relevant and predicted relevant sessions be reduced?

[1] See Cooper and Chen (2001) for details of this transformation.

Residual and Outlier Analysis

A regression equation approximates a set of empirical data points with a mathematical formula that generates a set of expected data points. The degree to which the empirical and expected data points match is measured by the error term. It is a residual value−−the amount left unexplained by approximating the actual data with that derived from the mathematical equation.

When a statistical analysis program computes the coefficients in a regression equation (the β's), it computes the error term and also a number of variables that can be used for residual analysis. Residual analysis is an intrinsic part of regression analysis because it indicates whether, and the extent to which, the equation fits the data. Chapter 5 in Hosmer and Lemeshow's (2000) classic book on logistic regression provides an extensive discussion of methods for assessing and calculating goodness-of-fit and residuals for this type of equation. Pregibon (1981) provides conceptual background and proposes a number of measures for analysis. Allison (1999) and the SAS manual for the LOGISTIC regression procedure (SAS Institute, Inc. 1989) provide additional resources.

The error term is the difference between the observed and predicted data points. It is a measure of the unexplained difference or residual. In linear regression analysis an error term is computed for each pair of observed and predicted values. In logistic regression, the difference is computed, not between each observed and predicted value, but between each covariate pattern.

There are two general categories of diagnostic statistics that will be used here. In the first category are the Pearson residual (CHI), the deviance residual (DEV), and the diagonal element of the hat matrix (HAT). These are so called building block statistics because they are used as a basis for deriving other more complex measures. The mathematical formulation of each of these measures is derived in Appendix 1. Both the Pearson residual and the deviance residual measure the goodness-of-fit between the observed and expected values and identify observations that are poorly accounted for by the model. The diagonal elements of the hat matrix contain values that are useful for identifying sessions that are outliers.

The second category of measures include the one−step difference in the Pearson residual (DIFCHI) and the one−step difference in the deviance (DIFDEV). In essence, these measures do what you would expect: they compute a difference in successive values of the measures. Again, the mathematical formulas to derive these two measures can be found in Appendix 1.

The goal of the residual analysis is to find a residual variable that proves to be good in detecting those sessions that are so far from normal that a statistical procedure could not success− fully predict their session relevance. That is, we want first to separate the outlier sessions from the non−outliers and then to concentrate our analysis on the non−outliers. To accomplish this, a residual analysis was performed as part of the logistic regression analysis of the entire sample of sessions reported earlier. Values of each

residual variable described above were computed for each session in the sample, means calculated, and the results plotted for each session to obtain a visual image of the distribution of the variable. In the graphs the objective is to decide on a value of the residual variable that will provide the separation required between outlier and non−outlier sessions.

Selection of Outliers Based on the Hat Matrix Diagnostic

Table 5 summarizes the results of a subjective imposition of a cutoff value on each of the residual variables. To understand the role of the cutoff value, first consider the hat matrix diagnostic measure (HAT). The goal of this analysis is to establish what constitutes an outlier. Consider the HAT measures and set a criterion that says that an outlier is one in which the value of C or CBAR is greater than or equal to 0.0002.

When we look at the characteristics of the sessions that are considered outliers versus non−outliers, an interesting pattern emerges. Compare the number of sessions that are known to be relevant versus the number that are predicted to be relevant, broken down by whether the observation has a HAT value less or greater than 0.0002. For sessions below the threshold, the actual percent of non−relevant/relevant is 83%/18% (81% + 2%/10% + 8%). For sessions above the threshold, the percent is 61%/39% (44% + 17%/17% + 22%). Thus sessions below the threshold are much more likely to be known to be relevant. For sessions below the threshold the predicted percentage of non−relevant/relevant is 91%/10% (81% + 10%/8% + 2%); above the threshold it is 61%/39%.[2] Again, sessions below the threshold are much more likely to be predicted to be relevant.

Further, sessions below the threshold are much more likely to be misclassified. For all sessions, the percent of sessions that are known to be relevant but predicted non−relevant is 9.79% (Table 4). For sessions above the threshold using the HAT diagnostic, the value is 17%. Similarly, 2.79% of all sessions are known to be non−relevant and predicted to be relevant (Table 4), but in the outlier class 17% are classified this way.[3] Clearly the use of the HAT diagnostic threshold results in a significant number of misclassified sessions, many more than the normal level.

---

[2] The 61% value is the sum of 44% plus 17%, while the 39% value is the sum of 22% plus 17%. While Table 5 indicates that the percentage of known relevant and predicted not relevant (17%) is the same as the percentage of known not relevant and predicted relevant (17%), this is due to rounding. The percentages are actually different by a few tenths.

[3] See the previous footnote.

Selection of Outliers Based on the One−Step Difference in Deviance

The results of imposing a cutoff value of 2.0 on the one−step difference in deviance, DIFDEV, are very impressive (see Table 5). If a session is considered an outlier when the DIFDEV value is greater than 2.0, two significant things happen. First, the cutoff value succeeds in constructing an outlier set where there is an almost complete coincidence between the known and predicted percentages. Second, the cutoff value allows a non−outlier set to be constructed where the difference between the percentage of predicted relevant (10.23%) sessions and known relevant sessions (11.74%) is only 1.51%. Thus establishing this cutoff substantially improves the prediction on the remaining set (see Table 6). By imposing this cutoff on the DIFDEV value for each session, a system could be considerably more certain of making a correct prediction of relevance.

The DIFDEV measure appears very good at detecting sessions that are outliers. In the following sections, that measure will be used to partition the database, and an analysis of the characteristics of the outliers versus the entire database will be presented.

Table 1

Mean Values for Base Variables Characterizing a Web-based Catalog Session

Symbol	Description	All Sessions			Non-Outlier Sessions			Outlier Sessions		
		N	Mean	St. Dev.	N	Mean	St. Dev.	N	Mean	St. Dev.
	Session variables									
PF	Active profile	905970.00	0.05	0.23	828303	0.05	0.22	77667	0.08	0.27
TD	Databases used	905970.00	1.37	0.90	828303	1.35	0.87	77667	1.57	1.19
TI	Indexes used	905970.00	1.73	1.32	828303	1.69	1.26	77667	2.08	1.73
SL	Session length (seconds)	905970.00	833.77	1023.02	828303	808.88	1014.29	77667	1099.17	1076.7
SQ	Searches performed	905970.00	4.24	5.21	828303	4.1	5.01	77667	5.67	6.84
SR	Searches with retrievals	905970.00	2.66	3.26	828303	2.57	3.17	77667	3.59	4
NP	Web pages requested	905920.00	19.53	21.69	828253	18.77	21.07	77667	27.63	26.08
TP	Unique web pages	905970.00	7.09	3.42	828303	6.92	3.36	77667	8.87	3.59
PL	Web pages in presearch	905970.00	2.36	1.45	828303	2.35	1.45	77667	2.43	1.45
TS	Presearch time (seconds)	897855.00	62.02	55.45	820188	61.4	54.99	77667	68.63	59.75
	Search variables									
SA	Author searches	905970.00	0.83	2.38	828303	0.82	2.33	77667	0.98	2.89
ST	Title searches	905970.00	0.96	2.69	828303	0.96	2.66	77667	0.95	2.99
SU	Subject searches	905970.00	1.60	3.26	828303	1.52	3.12	77667	2.41	4.4
SW	Power searches	905745.00	0.84	2.83	828138	0.8	2.71	77607	1.28	3.85
SB	Boolean searches	905613.00	2.58	3.94	828019	2.49	3.81	77594	3.53	5.03
CS	Unique indexes used	905970.00	0.91	2.60	828303	0.86	2.48	77667	1.43	3.6
FL	Find-less searches	905970.00	0.21	0.84	828303	0.2	0.8	77667	0.34	1.17
FR	Find-more searches	905970.00	0.12	0.54	828303	0.11	0.52	77667	0.18	0.75
FH	Find-more following power search	905970.00	0.10	0.53	828303	0.09	0.5	77667	0.18	0.75
	Display variables									
RD	Display time (seconds)	905970.00	20.38	28.85	828303	19.88	28.77	77667	25.68	29.11
OC	Single-record display screens	905970.00	2.77	5.58	828303	2.66	5.4	77667	3.96	7.09
MC	Multi-record display screens	905970.00	5.06	8.17	828303	4.82	7.94	77667	7.56	9.97
SO	Searches with single-record displays	905970.00	1.13	1.77	828303	1.09	1.73	77667	1.49	2.06
SP	Searches with multiple-record displays	905970.00	1.88	2.30	828303	1.82	2.24	77667	2.55	2.72
DF	Unique display formats	905970.00	1.66	1.03	828303	1.63	1.04	77667	1.98	0.96
VD	Single-record display viewing time	904132.00	33.48	94.29	826465	33.22	96.03	77667	36.28	73.28
VS	Multiple-record display viewing time	692606.00	58.18	107.04	620303	58.43	110.7	72303	56.04	67.89
	Error variables									
SE	Error count	905970.00	0.34	1.19	828303	0.32	1.12	77667	0.52	1.73
TE	Unique errors	905970.00	0.23	0.48	828303	0.22	0.47	77667	0.31	0.54
NE	Web pages of errors	905970.00	0.52	1.87	828303	0.49	1.75	77667	0.81	2.81
	Help variables									
HP	Web pages of help	905970.00	0.05	0.36	828303	0.05	0.35	77667	0.07	0.44
TH	Unique help requests	905970.00	0.04	0.25	828303	0.04	0.24	77667	0.05	0.3

Table 2

Mean Values for Derived Variables Characterizing a Web-based Catalog Session

Symbol	Description	All Sessions			Non-Outlier Sessions			Outlier Sessions		
		N	Mean	St. Dev.	N	Mean	St. Dev.	N	Mean	St. Dev.
	Session variables									
HT	Average number of hits	904668	405.9	1921.6	827001	401.0	1918.6	77667	459.0	1952.3
SD	Average search length	886888	202.7	290.3	809221	197.9	291.4	77667	253.2	274.6
VT	Average time between Web page displays	878866	37.5	33.2	801199	37.1	33.4	77667	42.6	31.0
PP	Presearch time proportion	905970	23.2	24.9	828303	24.2	25.5	77667	12.5	13.0
PS	Presearch Web page proportion	905970	22.6	16.3	828303	23.2	16.5	77667	15.2	11.7
	Search variables									
SI	Web pages requested	905970	4.6	4.9	828303	4.6	4.9	77667	5.6	4.9
RR	Searches with retrievals	905970	72.5	33.8	828303	72.1	34.4	77667	76.7	26.7
RA	Proportion of author searches	905970	22.4	39.7	828303	22.6	39.8	77667	20.7	38.3
RT	Proportion of title searches	905970	25.5	41.5	828303	26.2	42.0	77667	18.0	35.9
RU	Proportion of subject searches	905970	37.2	46.2	828303	36.6	46.1	77667	44.2	47.0
RW	Proportion of power searches	905970	14.9	33.5	828303	14.7	33.3	77667	17.1	34.9
SK	Known item searches	905970	1.8	3.5	828303	1.8	3.5	77667	1.9	4.2
RB	Proportion of Boolean searches	905970	56.5	42.6	828303	56.3	42.8	77667	59.2	40.4
MD	Searches followed by modifications	905970	1.2	2.7	828303	1.1	2.6	77667	1.9	3.8
RC	Searches that change indexes	905970	25.6	41.5	828303	25.0	41.2	77667	32.8	43.7
MM	Number of find-more searches	905970	0.2	0.8	828303	0.2	0.8	77667	0.4	1.1
PM	Proportion of find-more searches	905970	9.0	26.2	828303	8.7	25.9	77667	12.0	29.0
RK	Proportion of known-item searches	905970	47.9	48.0	828303	48.8	48.1	77667	38.7	46.2
RL	Proportion of find-less searches	905970	7.2	23.4	828303	7.0	23.1	77667	9.9	26.4
RM	Average number of search modifications	526947	0.0	0.2	491677	0.0	0.1	35270	0.0	0.3
	Display variables									
AV	Display time proportion	905970	39.8	30.4	828303	39.2	30.8	77667	46.8	23.9
NR	Average number of single-record displays	905970	1.4	2.5	828303	1.4	2.5	77667	1.8	2.9
NS	Average number of multiple-record displays	905970	2.3	3.7	828303	2.2	3.6	77667	3.2	4.5
RO	Proportion of single-record displays	905970	32.8	37.3	828303	32.7	37.6	77667	33.4	34.7
RP	Proportion of multiple-record displays	905970	54.5	37.8	828303	54.0	38.3	77667	59.1	32.6
	Error variables									
ER	Proportion of system errors	905970	6.4	18.0	828303	6.3	17.9	77667	7.7	18.3
RE	Web page error rate	905970	2.3	6.8	828303	2.3	6.8	77667	2.4	5.9
	Help variables									
PH	Proportion of time using help	905970	0.3	2.7	828303	0.3	2.8	77667	0.2	2.0
VH	Average time per page for help	904819	0.9	8.1	827152	0.9	7.8	77667	1.5	10.3
RH	Web page help rate	905970	0.2	1.8	828303	0.2	1.8	77667	0.2	1.5

Table 3
Comparison of Logistic Regression Runs for Entire Database
Versus Non-Outlier Database

Variable	Entire Database	Non-Outlier Database
Number of observations	561,863	511,339
Maximum rescaled R-square	0.48	0.86
Number of parameter estimates not significant at 0.05 level	5	2
Predicted Probability Information		
Percent concordant	88.9	99.3
Percent discordant	10.9	0.7
Sommers' D	0.779	0.986
Gamma	0.781	0.986
Tau-a	0.229	0.2
c	0.89	0.993
Multicollinearity Test		
Number of variables with Tolerance below 0.40	0	8

Table 4
Number and Percentage of Known Relevant vs. Predicted Relevant Sessions for Entire Population

		Predicted Not Relevant	Predicted Relevant	Total
Known Not Relevant	Frequency	468,752	16,481	485,233
	Percent	79.39	2.79	82.18
Known Relevant	Frequency	57,797	47,428	105,225
	Percent	9.79	8.03	17.82
	Total	526,549	63,909	590,458
		89.18	10.82	100

Table 5
Analysis of Residual Variables for Establishing Cutoff Values

Variable	Name	Cutoff Value	Percent of Obs. In This Category	Percent of Observations That Are				
				Known Relevant		Known Not Relevant		
				Predicted Relevant	Predicted Not Relevant	Predicted Relevant	Predicted Not Relevant	
HAT	Diagonal element of the Hat matrix	>0.0002	4	22	17	17		44
		<=0.0002	96	8	10	2		81
DIFCHI	One-step difference in Pearson chi-square	>2.0	8	0	84	16		0
		<=2.0	92	9	3	2		86
DIFDEV	One-step difference in deviance	>2.0	9	0	83	17		0
		<=2.0	91	9	3	1		87
CHI	Pearson residual	<-25	5	0	17	0		83
		>=-25	95	8	9	3		79
DEV	Deviance residual	<-3.5	5	0	17	0		83
		>=-3.5	95	8	9	3		79

Note: The values for DIFDEV greater than 2.0 are shown in more detail in Table 6.
The derivation of each variable is given in Appendix 1.

Table 6

Number and Percentage of Known Relevant vs. Predicted Relevant Sessions For Non-Outliers Using a DIFDEV Value <= 2.0

		Predicted Not Relevant	Predicted Relevant	Total
Known Not Relevant	Frequency	468,752	7,795	476,547
	Percent	86.82	1.44	88.26
Known Relevant	Frequency	15,959	47,428	63,387
	Percent	2.96	8.78	11.74
	Total	484,711	55,223	539,934
		89.77	10.23	100

Note: See the text for an explanation of the use of the DIFDEV residual variable.

Session Characteristics

As is to be expected, when the set of sessions is partitioned into non−outliers and outliers based on the value of the one−step difference in deviance (DIFDEV), there are important differences in the mean values of the base and derived variables for the sets. Tables 1 and 2 present the results. While the magnitude of the differences between the means of the outliers and non−outliers is small in some cases, a t−test confirmed that there is a significant difference between the means for all variables in Table 1, and for all variables except RM (search modifications) in Table 2. In Table 1, the mean session length (SL) of the outliers is about 4.8 minutes longer than non−outliers. On average a non−outlier session contains about 1.6 more searches and displays 9 fewer Web pages. Non−outlier sessions have about 59 fewer hits (Table 2), more title searches, less subject searches, and more known−item searches than outliers. And the proportion of time spent displaying records is much lower for non−outliers (39%) than for outliers (47%).

In an analysis of the usage patterns of the University of California's Melvyl® catalog over a 479−day period, Cooper (2001) showed how certain session characteristics change over time. A similar analysis was conducted for this investigation, only this time the goal was to determine whether there was a difference in the patterns between outlier sessions and all sessions. Over the sample period, the total number of sessions has a generally upward pattern. However, over the last three months of the sample period, the number of outlier sessions tends to remain flat. This could indicate that users are becoming more familiar with the system, and their observable behavior pattern (as judged by the variables in Tables 1 and 2) is less erratic. Although some days of the week see a heavier use of the Melvyl® catalog, the proportion of outliers to the total population dies not vary appreciably throughout the week. As noted above, the mean outlier session length is shorter than for the entire population, and it remains so throughout the sample period. Figure 1 shows how the number of display actions, search actions, presearch actions, and other actions varied in the outlier set over time. Figure 2 shows the same information for the entire database. It is taken from Cooper 2001, (Figure 4, p. 142). The time series in Figure 1 generally follows the pattern of the entire dataset (Figure 2), except that the number of display actions drops over time in the outlier set, while it remains high and steady in the entire database. Interestingly, although one would expect an outlier session to be error−prone, in the outlier database the number of errors per session is in fact so small that it is not even plotted.

The Medlars database is accessed with the same command language as the Catalog database and displays similar patterns for the number of searches per session. Generally, there are more searches per session in the outlier database for Medlars sessions (in the range of 1.25−−1.75), than for the entire database of Medlars sessions (0.8−−1.5). The amount of time spent displaying records also follows a similar pattern over time between the out− lier and entire set of observations. There is much less differentiation in the patterns in the outlier dataset (as opposed to the entire set of observations) between different databases for variables such as the number of other actions, the number of

records displayed, the display−to−hit ratio, the total display time, and the number of help requests.

Although we can conclude from this analysis that there are differences between outliers and non−outliers, they are relatively difficult to ascertain by simple graphic observation techniques. However, when the logistic regression analysis is repeated for the full dataset versus the non−outlier dataset, the results become much more clear (Table 3). The R^2value increases from 0.48 to 0.86, the number of variables that are not significant in the regression equation drops from five to two, and measures of similarity all improve substantially.

Conclusions from the Quantitative Analysis

A number of conclusions become apparent from the preceding analysis. First, it is possible to use a set of quantitative variables to characterize a catalog session, and these variables appear to be reasonable substitutes for personal interviews about the session. A logistic regression analysis can be used to explain and then predict when a session will produce relevant results. The accuracy of the prediction can be substantially improved by regression residual analysis by measuring the difference between the observed and fitted values of the independent variables. If an appropriate cutoff value of 2.0 is applied to the one−step difference in deviance (DIFDEV), the sessions can be successfully partitioned into two sets. This partitioning allows a significant increase in the accuracy of the prediction and lowers the difference between the percentage of predicted relevant (10.23%) sessions and known relevant sessions (11.74%) to only 1.51%.

The next part of the paper, focusing on a subjective analysis of failures, has two goals: to determine qualitatively the characteristics of sessions that were misclassified by the regression analysis and to discover how the selection algorithm might be improved. But first we will review previous quantitative and qualitative work in the area of failure analysis.

Figure 1
Outlier Session Characteristics

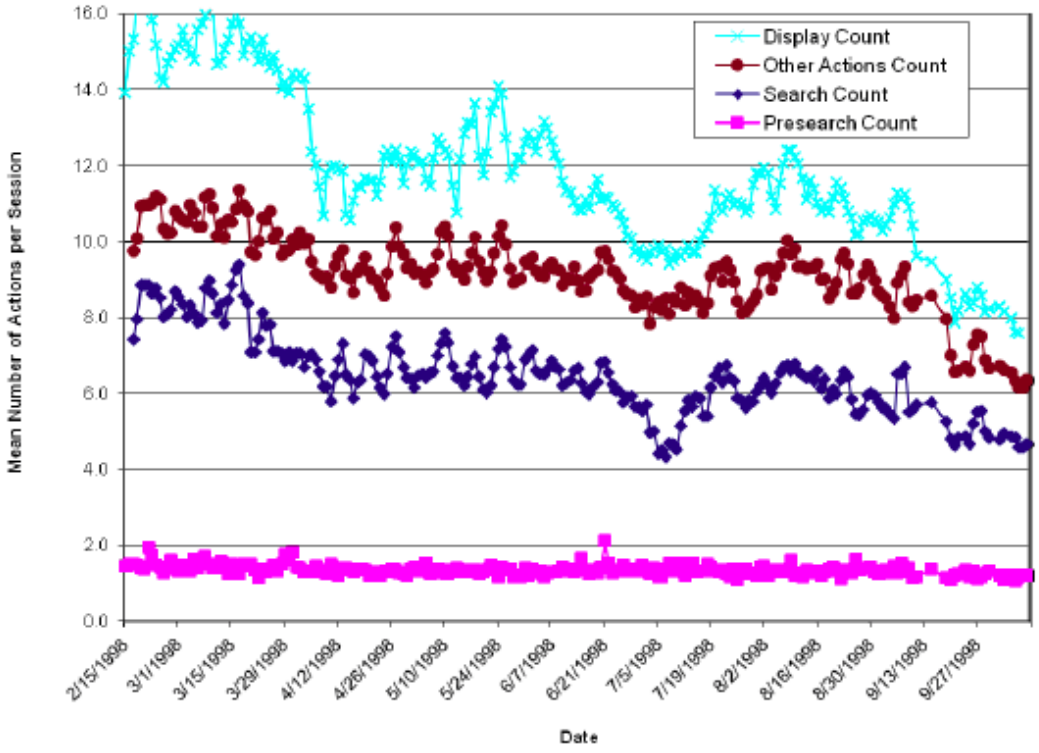
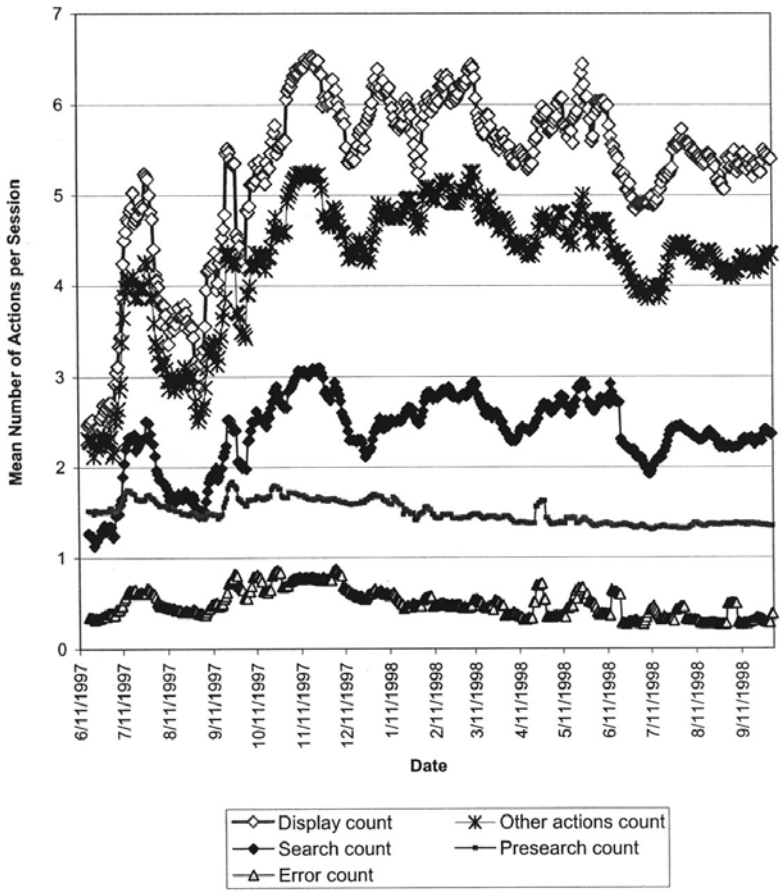


Figure 2
Session Characteristics



Source: Cooper (2001) Figure 4, p. 142

Previous Failure Analysis Research

The analysis of system failures plays an important role in many fields: the development of physical objects, political processes, medical research and development, and software and communications systems, to name but a few. Failure analysis is especially prevalent in engineering as illustrated by the many texts on this topic such as the Failure Analysis of Engineering Materials (Brooks and Choudhury, 2002).

Some of the most detailed failure investigations are undertaken by the U.S. National Transportation Safety Board in their analysis of aircraft accidents. Typical of these are the reports on the USAir accident in Pennsylvania in 1994, the TWA accident off New York City in 1996, and the EgyptAir accident off Massachusetts in 1999 (U.S. National Transportation Safety Board, 1999, 2000, 2002). In these studies, every conceivable aspect that could have contributed to the accident is explored meticulously.

In medical and biological research, a technique that is often employed is failure−time analysis. This statistical methodology is used to examine when a failure occurs in a temporal sequence of events (Kalbfleisch and Prentice, 1980). For example, in a controlled experiment, a failure event might occur when an animal dies during a drug research experiment. This type of failure analysis has also been applied to the study of office information systems to predict when a study participant would adopt or stop using a system's feature (Cooper, 1991).

The commercial world can also suffer failures−−witness the collapse of a number of major U.S. corporations in the early 2000s. Scherer and Ross (1990) provide the classic text on the analysis of economic market failures. They discuss what can be done to evaluate and regulate markets where there is the potential for a lack of fair competition.

The library and information science literature contains a number of examples of the use of failure analysis. One of the earliest−−and now considered a classic investigation of information retrieval failures was conducted by F. W. Lancaster in his evaluation of the National Library of Medicine's MEDLARS system (Lancaster, 1968). Among other things, Lancaster tried to determine why documents were not being retrieved in response to well−constructed queries. His investigation of this problem focused on the indexing of the documents in the collection, and he suggested improvements in indexing strategy.

In the information retrieval process, a user's query is matched against stored representations of documents, and any documents that match the query are retrieved from the database and returned to the user for review. The matching process is facilitated by stemming−−removing the endings from the roots of words. Harman (1991) conducted a comprehensive analysis of the effects of various stemming algorithms on the success or failure of an information retrieval system to match the query against a document. She concluded that search failures cannot be attributed to the algorithm used for stemming.

Kambil and Bodoff (1998) provide an excellent analysis of information retrieval indexing failures using part of the TREC database. They ran a set of queries against the database

and looked for false hits (documents retrieved incorrectly) and false misses (documents that should have been retrieved but were not). They discovered three causes of false hits: (1) an inappropriate term was used in the indexing, (2) an appropriate term was used in indexing but it was not appropriate for the context of the query, and (3) the ranking system used to present results forced inappropriate documents to the top of the list and forced others off the list. They also found seven causes of false misses: (1) lack of exhaustive indexing, (2) lack of depth of indexing, (3) lack of synonyms in indexing, (4) a typographic error, (5) presence of multiple terms that hid the appropriate meaning of the document in a particular query context, (6) use of a synonym in indexing the document that prevented it from matching a query term, and (7) word−stemming problems.

Another categorization of failures in searching was established by Drabenstott and Weller (1996). They submitted 528 usable queries to two library catalogs and analyzed the results. They categorized search failures into "seemingly unsuccessful", and "unsuccessful" searches. Failures within the first category could be attributed to a lack of user perseverance in the search process, too great a specificity in the query terms used in the search, and conflicting judgments on what constituted relevant material retrieved. Factors that contributed to unsuccessful searches included very large sets of retrieved citations, inappropriate use of search vocabulary, spelling errors, problems navigating the command structure of the catalog, subject−searching problems potentially attributable to the way the search system used its thesaurus, the retrieval of titles that were more specific than desired, and the retrieval of zero or too few titles.

Berger (1994) studied user interaction patterns in 144 sessions with the University of California's Melvyl® catalog. Among other things, he tried to analyze what constituted both an unsuccessful search and also a suboptimal search−−a search in which the user ultimately finds the desired material, but does so in a circuitous route. He categorized searches into known−item, personal author, periodical (i.e., journal title), or subject (topical) searches. For known−item unsuccessful searches, errors included "... use of incorrect title, misspellings, faulty syntax, large results, use of the wrong index, and searching the wrong database." (Berger, 1994, p. 45). For suboptimal personal author searches, errors again included large result sets and a failure to understand how to search "... for works by an author versus works about an author." (Berger, 1994, p. 45). Unsuccessful personal author searches stemmed from misspellings and a lack of perseverance by the searcher to refine searches that returned large result sets. Suboptimal periodical searches were mainly the result of the user's lack of a clear understanding of which index in the Melvyl® system to use. Most unsuccessful periodical searches stemmed from the user's lack of a clear idea of the definition of a periodical−−a source of arguments even for librarians. A lack of precision and completeness in specifying search terms contributed to suboptimal and unsuccessful topical searches.

A comprehensive analysis of MEDLINE searching conducted in 1989 (Wilson, Starr−Schneidkraut, and Cooper, 1989) included an analysis of so−called "ineffective" searches, as defined by the users themselves. That is, the search did not meet the user's self−imposed criteria. When transaction logs were analyzed, it appeared that searches were rendered "ineffective" because the user did not use the correct database for searching, did not consult all appropriate backfiles when doing the search, did not examine retrieved

sets carefully, did not understand the concepts of word truncation, did not understand how to do subject searching, or did not understand how the Medical Subject Headings (MeSH) system works and how to map their search request into MeSH.

Web search engines, such as those provided by Google, provide only a limited capability to construct complex queries, and consequently, the analysis of search failures is also limited. Jansen, Spink, and Saracevic (1998) conducted an investigation of search failures in a sample of more than 51,000 user queries submitted to the Excite search engine. Less than 10% of all queries used Boolean operators, and of these, the authors determined that about 26% were used incorrectly. In the Excite query language it is possible to specify that a term must be present (or absent) in the retrieved documents. About 10% of all queries used one of these operators, but in this case the authors determined that 75% of the uses were incorrect.

In contrast to Web search engines, the DIALOG system offers a rich set of commands and options and is available for hundreds of well−structured databases. Siegfried, Bates, and Wilde (1993) analyzed the searching tactics of 27 art history researchers over a two−year period. The researchers were given some search training and were taught that a good search strategy was to develop individual simple search statements, retrieve the results and then combine them to make complex queries. When the authors analyzed the search transaction logs, they found reasonably successful implementation of this idea. However, only one−third of all searches used any form of Boolean logic.

As part of a much wider investigation into the design of interfaces to library catalogs, Bates (2002) gives a number of examples of search problems and failures attributable to the limited capabilities of the query language. In one case, the query language would not allow users to express simple and short phrases. In another case, the query language could not successfully match complex subject headings when the heading included multiple words (like the Library of Congress Subject headings) and Boolean operators were used in the query. In still another, the query language would not allow users to choose multiple terms for a query from different sections of a faceted thesaurus.

A Qualitative Assessment of Prediction Failures

In the first part of this paper, we developed mathematical criteria that optimize the partitioning of sessions into outliers (whose relevance cannot easily be predicted) and non−outliers (whose behavior patterns are more constrained). Now we will take a subjective look at why the predictions of relevance failed for sessions that were considered outliers. This part of the investigation involved a manual examination of every action taken by users in 400 randomly selected outlier catalog sessions. Each session was reviewed and a subjective rating of the relevance of the session was made by the author.[4] Further, an analysis of why the author thought there was a prediction failure was made, and how the prediction algorithm could be modified to improve its success.

Sampling Methodology

Table 4 divided the original population of 590,458 sessions into four groups: (1) those known not to be relevant and predicted not relevant, (2) those known not to be relevant and predicted relevant, (3) those known to be relevant and predicted not relevant, and (4) those known to be relevant and predicted relevant. For this failure analysis, we will only examine the outlier sessions falling in the groups were there was a mismatch: group 2 (consisting of 8,686 sessions) and group 3 (consisting of 41,836 sessions). Recall that a session is known to be relevant because a print, mail, download, or save action was recorded in the transaction log file during the session. A session is ,I predicted to be relevant when a set of values for a session are substituted into the derived regression equation, and the dependent variable indicates the session is relevant. [5]

The one−step difference in deviance (DIFDEV) was computed for each outlier session in the two groups. The sessions were classified into 10 categories based on the value of DIFDEV (which is always greater than or equal to 0). Sessions with a DIFDEV value of less than 2 were placed in category 1, those with a value greater than or equal to 2 and less than 3 in category 2, and so on. Those with a DIFDEV value greater than 10 were assigned to category 10

Two independent samples (one for each group) were drawn using stratified random sampling without replacement. The goal was to derive a sample for each group consisting of an equal number of sessions in each of the 10 strata or DIFDEV categories. Two independent samples were drawn for each of the two groups. Each sample consisted of 200 observations.

---

[4] The criteria used for this rating will be discussed later in the paper.

[5] These counts are different from those in Table 4 because the analysis is dealing only with outliers, and because of missing observations of some of the variables that prevented either classification or prediction.

Subjective Relevance Rating of Sessions

The transaction log of each session was examined and a numerical relevance ranking was assigned solely by the author. The rating scale extended from 1 through 7, where a session assigned a rating of 1 was considered relevant and 7 was considered not relevant. The relevance of a session, as determined by the author, was based purely on subjective criteria and was evaluated as part of the failure analysis to determine if there was any relationship between the mathematically predicted values of relevance of the session and the author's own assessment.

The relevance of a session was assessed by looking at whether there were any hits; any viewing of the results of the searches; any logical progression in the session; any time spent looking at the results; and any print, mail, download, or save actions. The goal was not to infer the motive of the user, but rather to see whether the user extracted something from the process. Although it is not always possible to know from the transaction logs whether the user printed something using the browser's print button, it is sometimes possible to infer this action from the time spent viewing the search results. Further, if it appeared that the user spent some time viewing search results, that was another indicator of a positive outcome.

As amorphous as the above description may sound, there are actually very few sessions in which the transaction log file does not reveal what the user is doing. Search activity, by its very nature, is not overly complex. Although it may seem counterintuitive, it is possible to "see through" the searcher to understand what is going on. For example, it is possible to "see" when a person is using the system to check the citations in a paper that is to be published, when a user is simply a tourist at the site,[6] when a user is progressing nicely from the general to the specific, or when the user is not very successful. Further, is possible to diagnose from the transaction logs the temporal nature of a session−a very important aspect. There is an overall logical structure of a session that appears to result in success.[7] Finally, it appears to be easier to interpret whether the user has obtained results for searches that employ a more "scientific" database like Medline, rather than a "general" database like Melvyl.

Table 7 summarizes the results of the subjective rating of session relevance. The outcome, although only one individual's personal assessment still is very interesting. For those sessions that are known to be relevant and predicted not relevant, the author's ranking also confirmed that the operational criterion for establishing relevance (presence of a PMDS action （print, mail, download, or save action) was valid. That is, the author validated that a PMDS action was a valid indicator of relevance because the preponderance of the sessions were assigned a relevance rating of 2 (45.5%) or 3 (34.5%). Further, the author agreed with the prediction made using the logistic

[6] See Cooper (1991), where the concept of a search 'tourist' is introduced.

[7] This temporal aspect was investigated in previous research using mathematical clustering and stochastic modeling of sessions. (Chen and Cooper, 2001; Chen and Cooper, 2002).

regression equation that those same sessions were not relevant (note that few sessions were assigned a relevance rating of 5, 6, or 7).

When it comes to the other set of sessions in Table 7, the results are less clearcut. For those sessions known not relevant and predicted relevant, there is a significant split in judgment. About 48.5% of the sessions were given a relevance rating of 2 or 3, while 42% were given a rating of 5 or 6. Thus it was much harder for the author to say whether those sessions were relevant and whether the prediction was correct. Obviously this procedure should be repeated with a panel of judges to verify its accuracy. This analysis does show, however, that such an experiment would be easy to conduct and could lead to additional useful results.

Table 7

Subjective Relevance Ratings for Sample of 400 Melvyl Sessions

Session Relevance Group	Relevance Rating Scale						
	Session judged relevant 1	2	3	4	5	6	Session judged not relevant 7
Known not relevant and predicted relevant	0.00%	11.50%	37.00%	7.00%	21.50%	20.50%	2.50%
Known relevant and predicted not relevant	2.00%	45.50%	34.50%	4.50%	4.50%	9.50%	3.50%

Notes: Each relevance group includes 200 sampled sessions. Entries in the table are the percent of sessions that were assigned the associated relevance rating by the author.

Characteristics of Relevant and Non−Relevant Sessions

Aside from assigning relevance rankings to sessions, an attempt was made to extract from the transaction log file the characteristics that distinguish relevant and non−relevant sessions. Again, although these observations are subjective, they should ring true to the experienced searcher. Sessions that were relevant appeared to be more carefully constructed and precise than their non−relevant counterparts. Terms used in a search were gradually made more specific as the session progressed. Multiple databases were used, and the same search terms were applied to them. A relatively long amount of time was spent looking at result sets. Often the user picked specific citations from a browse list for further investigation, rather than merely scrolling through a list from beginning to end. When many searches were conducted during a session, the user often deleted non-meaningful ones from the search history. When the user decided to save citations to a list, that action would take place frequently throughout the session followed by a return to the search process.

Nonrelevant sessions also had distinguishable characteristics. Many had no apparent focus. During a session a search would take place on one topic, then shift to a second and a third topic. While there are rational explanations for this strategy, other factors seem to confirm that the user was floundering. The search process often resulted in either zero hits or an extremely high number of hits. And sometimes the user repeated the exact same search. with (one suspects) the expectation that perhaps, this time, something different would happen! Often the same search was repeated periodically during the session. Thus the user might enter a particular query at the start of the session, then repeat it 5, 14, and 32 minutes later. This could be attributed to a faulty memory, but again, the evidence seems to indicate the user was optimistic that the results might be different the next time. Often minute changes were made to the query. Many times the user would either not look at the result sets from a session, spend very little time looking at the retrieved set, or perform no (or very few) record display actions at all. Sometimes the user would decide to search a different database, but after changing databases would not even bother searching the new database. When there were errors in the session, they were often repeated over and over, indicating the user did not know what to do to correct the error. It was also possible to detect that many of these non−relevant sessions ended on a downnote. Instead of forming a search that resulted in a small result set and looking at that set, the session would have a less happy outcome: The user would perform a series of searches, each of which returned either large result sets or zero result sets, and they would not bother to look at any results. Then, unable to come up with any new strategies, they would end the session.

Characteristics of Experienced and Inexperienced Searchers

The transaction logs reveal many interesting aspects of searching behavior, and the difference between inexperienced and experienced searchers is quite apparent. Inexperienced searchers often repeat the exact same query multiple times in a row, getting the exact same hit count each time. They formulate incredibly general searches (e.g., SU Jesus), get very large result sets (greater than 300), and then proceed to look through each citation one at a time. They often capitalize search terms in the belief that is necessary (shouting to the system?). They enter multiple search terms in one field, each separated by a semicolon (e.g., SU (receptor; retina)). They randomly insert commas in the middle of a single search term. They use Boolean operators in the middle of a search term (e.g., SU (gender and money)). Or they omit truncation symbols on a search term and expect the system to understand the term.

Experienced searchers perform a complex search in parts, building the pieces of the search and checking the hit count of the pieces as they go. The queries of the session are not unnecessarily narrow to begin with−−few restrictions on the search are made when starting. But as the search continues, it is refined. Individual searches are constructed of a number of terms. If author names are entered, they are formatted as the searcher thinks they would be stored in an index (e.g., last name followed by a comma followed by the first name). If subjects are used, exact subject headings are often employed. Subject headings are often extracted from the terms associated with relevant and retrieved items. Many times truncation symbols are associated with terms. A number of Boolean operators are used in a search. It is not uncommon to find the "and - not" operator in a query. Databases are selected for their specificity, and databases with the most general coverage are avoided. Features associated with each database are used to advantage, such as BIOSIS concept codes or supertaxa, or Medlars MESH headings. As a search progresses, some queries may result in no hits, in which case terms or operators are systematically removed to solve the problem.

A process of constant refinement is present in these sessions. When the searcher gets a reasonable result set, then and only then are citations displayed. Often the display format is changed to reveal more or less of the characteristics of the records. Once displayed, record sets are augmented or reduced by search modification. After an acceptable set of searches has been formulated, the searcher switches to another database using the same or similar queries. Finally, at the end of the session, the searcher begins to save, download, print, or email citations.

Reasons for Search Failures

The analysis of the transaction logs of sessions included an attempt to diagnose subjectively why sessions did not achieve their goals. Again, although it may seem counterintuitive, it appears to be possible to ascertain this information from a transaction log.

At an operational level, there are a number of observable events that can lead to search failure. One is searching for a particular topic in the wrong database. Another is the use of an inappropriate index in searching a selected database. For example, the search may be formulated to find an exact match between the query and the index entry, when it may be more appropriate to search for any occurrence of the query terms in the index. Further, a search qualifier that narrows the search too far (such as limiting results to those held at a specific library) may jeopardize search effectiveness. Finally, spelling errors often lead to search failures: for example, the user enters "microbioloby" when the correct term "microbiology" should have been entered. Spelling errors are sometimes hard for the user to spot, and so they are often repeated again and again on subsequent search statements.

In a number of cases the logs show that searches that start with a very narrow focus or that are very specific end in a lack of results. In a "good" session, the sequence of search statements gradually narrows the size of the retrieved set, but in sessions perceived as failures there is no such progression.

Causes of Prediction Algorithm Failures

The statistical methodology used to determine whether a session was a failure is obviously prone to error. Prediction algorithm failures can be classified into two types. The first type of prediction failure is one where the session is known to be relevant because a print, mail, download, or save (PMDS) action took place, but is predicted not to be relevant based on the logistic regression equations. The second type of failure is one where the session is predicted to be relevant but is known not to be relevant because no PMDS action took place. A subjective analysis revealed that the following characteristics seem to cause prediction failures:

(1) A short session in which the user makes errors that are diagnosed by the system.

(2) A short session with many requests for help.

(3) A very long session.

(4) A very short session.

(5) A very short session that included a save action. If a save action took place, the session would be known to be relevant, but the system predicted it not relevant, presumably based on its length.

(6) A session with a large number of system−diagnosed errors but that included a PMDS action.

(7) A session consisting of multiple searches each of which yields zero results and no display actions. The guess here is that because the number of searches performed versus the number of displays performed is not "normal", the algorithm thinks it is an error.

(8) A session consisting of multiple searches and a large number of display actions, but no PMDS action. Here the prediction is that the session is relevant because of the frequency of display actions.

Suggestions for Revision of the Algorithm

It appears that it is possible to refine the prediction algorithm to produce better results by giving less weight to the number of errors that occur in a session and by having less certainty about predictions for very long or very short sessions.

The PMDS criterion seems to be quite good. A PMDS action generally occurs in conjunction with a good search, usually after a search topic has been narrowed by successive queries and successive display actions. A PMDS action occurring in a cycle of search then display actions is good evidence that the criterion is valid. And a PMDS action occurring toward the end of a session, especially after a narrowing of the search and displays of intermediate results, is a strong indicator of success and a good validation of the relevance criteria. All these patterns were found in the sessions examined.

But it is possible that the PMDS criterion may be too broad. In a number of sessions, the user appears to save a citation to a list but then does nothing further with the list, like mailing it to someone. There are even cases where this happens at the very end of a session. One solution is to count a session as relevant only if a print, mail, or download action took place, and not if a save action occurs alone without one of the other three.

Some users have the option of checking the circulation status of an item at their location to determine if it is currently on the shelf. If it has been checked out by another borrower, they can see when it is due back in the library. This action could be incorporated into the criteria as an additional indicator of the relevance of the session.

Finally, a case can be made that a session is too broad a unit for analysis, and that the session should be considered to be composed of logical entities made up of topical units. These units would consist of all searches on a general topic and all display actions related to those searches. Each unit would be analyzed separately for the occurrence of a PMDS action. This type of analysis, however, requires the ability to diagnose when there is a change of search topic in a session, which is not a trivial problem for a machine algorithm.

Summary and Conclusions

This paper has taken two quite different approaches to improving the prediction of relevance of a user's session with an online library catalog. In previous papers we have shown that it is possible to use a set of quantitative measures to characterize a session (Tables 1 and 2). We have also shown that a simple metric can serve as a surrogate to indicate whether the session results were relevant to the user's needs. The measure of relevance is whether, during a session, the user printed, emailed, downloaded, or saved a citation from a search result. The results of a logistic regression analysis of the entire sample of sessions from the Melvyl® catalog showed a difference of about 6.85% between the actual number of sessions determined to be relevant and the predicted number of sessions determined to be relevant. In the quantitative analysis of failures, an extensive review of regression residual measures was undertaken. It was determined that the one−step difference in deviance measure performed well in partitioning the sample into outlier sessions (which would be hard to apply prediction tools upon), and non−outlier sessions. After establishing a cutoff criteria, the one−step difference in deviance measure was used to partition the sample and means of the quantitative measures were calculated for the partitions. With very few exceptions, there were significant differences between the means for the partitions. When the logistic regression was rerun again on the non−outlier partition, and the prediction process repeated, there was a significant improvement. The number of known relevant sessions was 11.74% for this sample, and the number of predicted relevant sessions was 10.23%, a difference of 1.51%. Thus the cutoff variable was successful in isolating homogeneous sessions, and the prediction process was excellent.

With further work, it seems likely that the Logistics regression methodology could be employed together with the outlier analysis methodology to make real-time assessments to predict the relevance outcome of a session.

The second part of this paper involved a qualitative assessment by the author of the nature of the sessions. Two stratified random samples each consisting of 200 sessions (a total of 400 sessions), were extracted from the set of 590,458 sessions for manual analysis.

This qualitative analysis had a number of objectives. One was to determine if the user's goals in performing a search could be ascertained simply from studying the transaction log files. This appears to be quite feasible, given enough experience looking at the logs.

A second was to determine the observable characteristics of a session that make the outcome of that session successful or unsuccessful. Again this was feasible. Successful searches are carefully constructed, employ terms that move from general to specific as the session progresses, selectively examine citations from a browse result list, and spend a relatively long amount of time looking at result sets. Unsuccessful searches lack focus, change topics frequently, or result in large result sets or in zero−hit searches.

A third objective was to diagnose why searches resulted in failure. The log files revealed a number of causes including the use of the wrong database, an incorrect index in searching, requiring an exact match between query and search terms, and spelling errors.

An attempt was made to ascertain why the prediction algorithm failed. A number of session characteristics seemed to be associated with prediction failure, including very short and very long sessions; sessions with many searches and very few display actions; and sessions with many searches and many display actions, but no PMDS action.

Finally, there appears to be ways to revise the prediction algorithm to improve its performance. The principal idea is to refine the operational definition of relevance to include print, mail, and download actions, but to exclude save actions. A number of sessions were found in the log where the user saved a citation to a list but then did nothing further with it. Confining the relevance test to print, mail, and download actions seems a very reasonable solution.

References

Allison, Paul D. (1999). *Logistic Regression Using the SAS System: Theory and Application.* Cary, NC: SAS Institute Inc., 304 pp.

Bates, Marcia J (2002). The cascade of interactions in the digital library interface. *Information Processing and Management*, 38, 381 - 400.

Berger, Michael (1996). *The User Meets the Melvyl System: An Analysis of User Transactions. Technical Report No. 7*. Division of Library Automation, University of California, Office of the President, Oakland, CA.

Brooks, Charlie R., & Choudhury, Ashok (2002). *Failure Analysis of Engineering Materials.* New York: McGraw−Hill.

Chen, Hui−Min, & Cooper, Michael D. (2002). Stochastic modeling of usage patterns in a Web−based information system. Journal of the American Society for Information Science and Technology, 53:7, 536 - 548.

Chen, Hui−Min, & Cooper, Michael D. (2001). Using clustering techniques to detect usage patterns in a Web−based information system. *Journal of the American Society for Information Science and Technology,* 52:11, 888 - 904.

Cooper, Michael D. (1991). Failure time analysis of office system use. *Journal of the American Society for Information Science,* 42, 644 - 656.

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

Cooper, Michael D. (1991). User skill acquisition in office information systems. *Journal of the American Society for Information Science,* 42, 735 - 746.

Cooper, Michael D., & Chen, Hui−Min (2001). Predicting the relevance of a library catalog search. *Journal of the American Society for Information Science and Technology,* 52:10, 813 - 827.

Drabenstott, Karen M., & Weller, Marjorie S. (1996). Failure analysis of subject searches in a test of a new design for subject access to online catalogs. *Journal of the American Society for Information Science,* 47, 519 - 37.

Harman, Donna (1991). How effective is suffixing? *Journal of the American Society for Information Science,* 42, 7 - 15.

Hosmer, David W., & Lemeshow, Stanley (2000). *Applied Logistic Regression*. Second edition. New York: John Wiley & Sons., 375 pp.

Jansen, Bernard J., Spink, Amanda, & Saracevic, Tefko (1998). Failure analysis in query construction: Data and analysis from a large sample of Web queries. In: *Proceedings of the Third ACM Conference on Digital Libraries,* Pittsburgh, PA. pp. 289 - 290.

Kalbfleisch, J.D., & Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.

Kambil, Ajit, & Bodoff, David (1998). Partial coordination. II. A preliminary evaluation and failure analysis. *Journal of the American Society for Information Science*, 49, 1270 - 1282.

Lancaster, F. Wilfrid (1968). *Evaluation of the MEDLARS Demand Search Service*. Washington: U.S. Dept. of Health, Education, and Welfare, Public Health Service.

Pregibon, Daryl (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9, 705 - 724.

SAS Institute Inc. (1990). *SAS/STAT User's Guide, Version 6. Fourth edition, Volume 2*. Cary, NC: SAS Institute Inc., 846pp.

Scherer, F. M., & Ross, David R. (1990). I*ndustrial Market Structure and Economic Performance.* Third edition. Boston: Houghton Mifflin.

Siegfried, Susan, Marcia J. Bates, and Deborah N. Wilde (1993). A profile of end−user searching behavior by humanities scholars: The Getty online searching project report no. 2. *Journal of the American Society for Information Science,* 44, 273 - 291.

United States National Transportation Safety Board (1999). *Air craft Accident Report: Uncontrolled Descent and Collision with Terrain USAir flight 427 Boeing 737−300, N513AU near Aliquippa, Pennsylvania September 8, 1994*. Adopted March 24, 1999. Washington, D.C.: National Transportation Safety Board, 366 pp. NTIS Stock Number PB99−910401. Report Number NTSB/AAR−99/01. Available online at http://www.ntsb.org

United States National Transportation Safety Board (2000). *Aircraft Accident Report: In−flight Breakup over the Atlantic Ocean Trans World Airlines Flight 800 Boeing 747−131 N93119 near East Moriches, New York July 17, 1996*. Adopted August 23, 2000. Washington, D.C.: National Transportation Safety Board, 441 pp. NTIS Stock Number PB2000−910403. Report Number NTSB/AAR/−00/03. Available online at http://www.ntsb.org.

United States National Transportation Safety Board (2002). *Aircraft Accident Brief: EgyptAir Flight 990 Boeing 767−366ER SU−GAP 60 Miles South of Nantucket, Massachusetts October 31, 1999.* Adopted March 13, 2002. Washington, D.C.:

National Transportation Safety Board, 160 pp. NTIS Stock Number PB2002−910401. Report Number NTSB/AAB/−02/01. Available online at http://www.ntsb.org.

Wilson, Sandra R., Starr−Schneidkraut, Norma, & Cooper, Michael D. (1989). *Use of the Critical Incident Technique to Evaluate the Impact of MEDLINE. Final Report.* September 30, 1989. Contract No. N01−LN−8−3529. Palo Alto, CA: American Institutes for Research.

Appendix 1
Logistic Regression Diagnostic Statistics

This appendix provides the mathematical derivation of the diagnostic statistics that are used to partition sessions into the outlier and non−outlier sets. The notation and mathematical formulations used here are from Hosmer and Lemeshow (2000). The material is based on Chapters 1 and 5 in their text.

The basic logistic regression equation for n independent variables is given by

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x+\ldots+\beta_n x}}{1+e^{\beta_0+\beta_1 x+\ldots+\beta_n x}}.$$

This equation can be fitted by standard linear regression techniques if the logistic transformation is made:

$$g(x) = \ln[\pi\frac{(x)}{1-\pi(x)}]$$

$$=\beta_0 + \beta_1 x + \ldots + \beta_n x + \epsilon.$$

Here g(x) becomes a linear equation.

The error term ϵ is the difference between the observed (y) and predicted ($\hat{y}$) data points. The value of the error term for a particular pair of data points j is $\hat{y}_j$. The error term is a measure of the unexplained difference or residual. For linear regression analysis, an error term is computed for each pair of observed and predicted values. In Logistics regression, however, the difference is computed, not between each observed and predicted value, but between each covariate pattern. Theoretically, the number of covariate patterns can greatly exceed the number of observed vs. predicted pairs, but because most of the independent variables are continuous in this model, the number of covariate pairs should be similar.

The difference is given as

$$\hat{y}_j = m_j\hat{\pi}_j$$

$$= m_j \frac{e^{\hat{g}(x_j)}}{1+e^{\hat{g}(x_j)}},$$

where $m_j$is the number of sessions with the same covariate pattern.

The Pearson Residual

The Pearson residual, r, for a particular covariate pattern, j is given as

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}.$$

In the text, the Pearson residual will be referred to as CHI.

The Pearson residual is an example of one category of goodness-of-fit measures, which measures the individual differences between observed and expected values. A second category summarizes each individual measure into one aggregate statistic. The summary statistic for the Pearson residual is the Pearson chi-square statistic. It simply sums individual r_j values:

$$X^2 = \sum_{(j=1)}^{J} r(y_j, \hat{\pi}_j)^2,$$

where J is the number of distinct covariate patterns in the sessions being examined.

The Deviance Residual

The deviance residual$d_j$is defined as

$$d(y_j, \pi_j) =$$

$$\pm\sqrt{2(y_j \ln(\frac{y_j}{m_j\pi_j}) + (m_j - y_j)\ln(\frac{m_j - y_j}{m_j(1-\pi_j)}))}$$

where the sign associated with the computed value is the same as the sign of the quantity

$$y_j - m_j\hat{\pi}_j.$$

When the value of $y_j = 0$, the deviance residual is given by

$d(y_j, \pi_j) =$

$\sqrt{2m_j|\ln(1-\pi_j)|}$

When $y_j = m_j$, the deviance residual is given by

$d(y_j, \pi_j) =$

$\sqrt{2m_j|\ln(\pi_j)|}$

The deviance residual d_j will be termed DEV in the text.

The Diagonal Element of the Hat Matrix

The design matrix X contains all the J covariate patterns found in the sessions being analyzed. There is a row in the design matrix for each session. The first column is set to one to indicate there is an intercept in the regression model. Succeeding columns record each of the p covariate values$x_j$for a specific session.

Hosmer and Lemeshow (2000, pp. 168-169) and Pregibon (1981) show that the hat matrix H is given by

$H = X(X'X)^{-1}X'V^{1/2}$

where $V$is a JxJdiagonal matrix with the general element

$v_j = m_j\hat{\pi}(x_j)[1 - \hat{\pi}(x_j)]$

The hat matrix diagonal elements h_j have the desirable property that they can flag sessions that are outliers. They are referred to as HAT in the text.

The value of each element is defined as

$$h_j = m_j \hat{\pi}(x_j)[1 - \hat{\pi}(x_j)](X'VX)^{-1}x_j' = v_j x b_j$$

Here

$$b_j = x_j'(X'VX)^{-1}x_j'$$

The value x_j' is given by

$$x_j' = (1, x_{1j}, x_{2j}, x_{3j}, \ldots x_{pj})$$

and is a vector of covariate values defining the j^{th} covariate pattern across all sessions.

The Pearson residual for the j^{th} covariate pattern, namely r_j, is given by Hosmer and Lemeshow (2000, p. 173) as

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \pi_j)}}$$

The standardized Pearson residual *x sub j* has a variance equal to 1. It is given by

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_j}}$$

Chi-Square Difference and Deviance Difference Measures

There are two measures that compute the change in value of statistics that have already been derived. One is the change in value of the Pearson chi-square (CHI) statistic, called DIFCHI. The other is the change in the deviance (DEV) statistic, termed DIFDEV. Each records the decrease in the value of the base statistic from deleting outlier observations. See Hosmer and Lemeshow (2000, p. 174).

The change in the chi-square statistic, ΔX_j^2, is given as

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)}$$

$$= r_{sj}^2$$

The change in the deviance, ΔD_j, is

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)}$$