The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog

Ray R. Larson
School of Library and Information Studies, University of California at Berkeley, Berkeley, CA 94720

Search index usage in a large university online catalog system over a six-year period (representing about 15.3 million searches) was investigated using transaction monitor data. Mathematical models of trends and patterns in the data were developed and tested using regression techniques. The results of the analyses show a consistent decline in the frequency of subject index use by online catalog users, with a corresponding increase in the frequency of title keyword searching. Significant annual patterns in index usage were also identified. Analysis of the transaction data, and related previous studies of online catalog users, suggest a number of factors contributing to the decline in subject search frequency. Chief among these factors are user difficulties in formulating subject queries with Library of Congress Subject Headings, leading to search failure, and the problem of "information overload" as database size increases. This article presents the models and results of the transaction log analysis, discusses the underlying problems with subject searching contributing to the observed decline, and reviews some proposed improvements to online catalog systems to aid in overcoming these problems.

Introduction

The ability to search by subject terms is one of the features that distinguish a library catalog from a simple descriptive inventory of a library's holdings. Subject access, ideally, provides catalog users with a means of locating books of interest to them when they do not know the authors or titles of those books. The provision of subject access changes the catalog from a "data retrieval" tool for locating specific items to an "information retrieval" tool for locating information on a topic when the user has no particular book in mind.

In this study we examine the long-term trends and patterns of index usage in a large online catalog system discovered through transaction log analysis. The results show that there has been a significant decline in the frequency of subject index use over time. The evidence from this analysis, and from previous studies of online catalog users, are examined to identify the probable causes of this decline. The implications of this study are that the subject access mechanisms of the online catalog are failing to satisfy the needs of its users, that the users are consequently turning to alternative means for topical access, and that a redesign of the retrieval capabilities of online catalogs may be required if they are to continue to serve their users.

Previous studies of online catalog use found subject searching to be something of an anomaly: It was the most frequently used form of online catalog search, and, at the same time, users had more difficulties and problems when doing subject searching than in any other type of search. The most widely known and discussed findings on subject searching in online catalogs were those in the nationwide study of online catalog use and users sponsored by the Council on Library Resources (Matthews, Lawrence, & Ferguson, 1983). Among the results reported in that study were the observations:

(1) Subject or Topical searches accounted for up to 59% of searching.
(2) Subject searching was the most likely to prove problematic to users—with 43% reporting difficulty in formulating a subject search.
(3) Enhancements to subject and topical searching were the most desired addition to the capabilities of existing online catalogs with 45% desiring the ability to see words related to their search terms and 42% desiring additional access points (the ability to view the index or table of contents of a book).

In earlier studies of traditional card catalog use, the frequency of subject searching reported varied considerably. Markey (1984) has reviewed the major published studies of card catalog use and found the percentage of subject searches reported to range between 10–62%. However, most of the card catalog use studies con-
ducted in the 1970s, many with an interest in developing specifications for automated catalogs (Palmer, 1972; Lipetz, 1972; Swanson, 1972), found that subject searching accounted for only about 20–40% of catalog searches.

Some designers of online catalogs took these card catalog findings as evidence that subject searching was apparently not of particular interest to the majority of catalog users, and hence, many early online catalogs did not provide it. Indeed, many early online catalogs were designed to mimic the card catalog, in both display and in the provision of a single access point per search. Later, more sophisticated, online catalog systems adopted features previously found in commercial bibliographic retrieval systems (such as Dialog and BRS), including keyword access to author, title and subject fields and the use of Boolean operators to construct queries. Hildreth (1987) has called these types of online catalogs “first generation” and “second generation” respectively.

Studies of the new online catalogs (such as the CLR study) revealed that usage of subject access had increased dramatically over the proportions observed in card catalog studies (or at least those reported in the 1970s). In some online catalogs, subject searches were occurring with twice the frequency expected from studies of card catalogs. Suddenly, it seemed, subject searching had moved from an infrequent mode of access to the most commonly used search access point in online catalogs.

This pattern was shown quite graphically in an interview survey of catalog users at the New York State Library conducted by Lipetz and Paulson (1987). They interviewed three sets of users, one week-long baseline survey was conducted before the introduction of the subject search capability of the online catalog and two follow-up surveys after the change. They found that subject searching, which accounted for 27% of catalog uses during the baseline survey, rose to 41% in the first follow-up survey, and to 49% in the second.

Markey (1985) and Cochrane and Markey (1983) have reviewed several studies of online catalog use that show a majority of users were seeking information on a subject. Some of these studies, and subsequent studies, have shown that a subject approach was frequently employed by users of systems that lacked formal subject access by using title information or shelf browsing (Lipetz, 1972; Lipetz & Paulson, 1987; Hancock, 1987).

This increased reliance by users on the subject access mechanisms of the online catalog led to a renewed interest by researchers in the problems and potentials of subject access in library catalogs. Cochrane (1983) observed that the findings of online catalog studies concerning subject access might be considered as signaling a "paradigm shift" in our knowledge of catalogs and catalog use, pointing out the needs for further research (and for some consensus in the professional community) on the requirements and means for providing subject access in online catalog systems.

Most of the studies of online catalog use, such as the CLR study, relied on questionnaire data and some interpretation of user responses to arrive at their figures for subject search frequency. The limitations of survey data and interpretation are well known. Researchers must assume a great deal about how the survey respondents interpret the questions and choose their responses. Fortunately, in the case of online catalog evaluation, transaction monitoring provides an alternative and more reliable method of observation.

Transaction monitoring, in its simplest form, involves the recording of user interactions with an online system. More complete transaction monitoring will also record the system responses and performance data (such as response time for searches), providing enough information to reconstruct all of the user's interactions with the system. Transaction monitoring alone can provide highly detailed information about how users actually interact with an online system, but it cannot reveal their intentions or whether they are satisfied with the results. Nielsen (1986) and Cochrane and Markey (1983) have discussed the comparative benefits and problems of questionnaires and transaction monitoring in evaluating online catalogs. The transaction monitoring studies carried out in conjunction with the CLR questionnaire study (Larson, 1983; Tolle, 1983; Larson & Graham, 1984) confirmed the high rate of subject access among online catalog users reported in the survey.

Transaction monitoring studies (Larson, 1981, 1983, 1986; Larson & Graham, 1984; Tolle, 1983; Borgman, 1983; Lawrence, Graham, & Presley, 1984; Kaske, 1988a, 1988b) have concentrated on a variety of different aspects of online catalog use and users. Most of the early transaction monitoring studies examined only limited samples of transaction data, and did not (with the exceptions of Borgman (1983) and the recent work by Kaske (1988a, 1988b)) examine the variations in usage patterns over time. Although sampling methods provide fairly reliable pictures of user behavior for a given system and a time period, they do not permit inferences to be made about longer term user behavior.

Recent work by Kaske (1988a, 1988b) has examined the variability of subject searching based on analysis of all transactions collected during one semester from the University of Alabama's VTLS online catalog system. He found that the percentage of subject searches varied from a low of 35% to a high of 52% over the weeks of the semester (with similar wide variations in use for days of the week and hours within the day).

These transaction monitoring studies present rather a confusing picture of subject search frequency, compounded by differences in system provisions for subject access (keyword or exact matching of headings), user population, level of data collection (sample or complete transactions), and the time periods of measurement.
The overall impression is one of random variation. Although the frequency of subject searching is fairly high in these studies, there is no indication whether that frequency is increasing, decreasing, or relatively static within certain boundaries.

In the present research all available transactions from a large online catalog system, collected over a six-year period, are used as a basis for examining some long-term trends and patterns in subject searching in online catalogs. Mathematical models of index usage over time were developed and tested using the transaction monitor data. The analyses show a consistent decline in the use of the subject index, and a corresponding replacement with title keyword searching. The problems with subject access reported in user studies appear to be compounded by database growth, and lead to changes in index usage over time. The principle effect observed is a significant decline in the frequency of subject searching.

The following section briefly describes the online catalog system studied. Then the methods and results of this study are presented. Finally, some implications of the results are discussed and some proposed enhancements to online catalogs that may alleviate some of the problems identified are examined.

Background and Methods

Background: The University of California's MELVYL System

The online catalog system studied in this research was the University of California's Online Union Catalog, MELVYL. This section provides a brief description of the MELVYL system and its features as background for the analyses that follow.

The University of California (UC) libraries include nine main libraries, one on each UC campus. They also include nearly 100 associated branch and specialized libraries (e.g., medical libraries, law libraries, and other separately maintained discipline-oriented libraries). The combined holdings of the UC libraries are estimated to be in excess of 23 million volumes. The goal of the UC Online Union Catalog, MELVYL, is to make these combined holdings accessible to all library users, regardless of campus location.

The user population for the MELVYL system includes the faculty, students, and staff of the nine UC campuses and associated schools and research units (such as Hastings College of Law, the Lawrence Berkeley Laboratories, and the Scripps Institute of Oceanography). The demographic characteristics of the MELVYL system's user population were examined as part of the CLR online catalog user survey (University of California, 1983; Larson & Graham, 1983; Lawrence et al., 1984). The average MELVYL catalog user does not differ significantly from the average online catalog user as described in the CLR survey (Matthews et al., 1983).

A prototype version of the MELVYL system first became available for public use on all nine UC campuses in August 1981. The prototype version used a fixed database of 733,412 bibliographic records representing approximately 1.3 million volumes. No records were added or changed during the prototype period. During the prototype period a new database design was developed for the MELVYL system, providing facilities for batch updating and enhanced authority control. This enhanced or "production version" of the MELVYL system replaced the prototype on April 15, 1984 and since that time the database has grown to over 5 million records, representing approximately 10 million volumes.

Access to MELVYL is provided by dedicated terminals in campus libraries, as well as through campus computing networks, the ARPA Internet and dial-up lines. Users may choose to interact with MELVYL using either a simple set of menus and prompts, called "LOOKUP mode," or using a more powerful command language ("command mode").

The MELVYL system provides access to the database through both keyword and "exact" indexes and provides integrated authority control for personal and corporate name searches. Keyword indexes permit searching on any word from particular field or set of fields in the MARC record. The "exact" indexes provide searches with left-to-right matching of specific fields in the MARC record with optional right truncation. Both keyword and exact indexes ignore case, punctuation, and a small set of stopwords in matching, and indexes may be combined in command mode searches using Boolean operators. Table 1 shows the primary keyword and exact indexes available for searching in the main MELVYL database. (Indexes for the California Academic Libraries List of Serials (CALLS) and the MELVYL/MEDLINE database (Hubble, 1988) are not shown, or included in the analysis).

<table>
<thead>
<tr>
<th>Index Name</th>
<th>Full Name</th>
<th>MARC Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>PA</td>
<td>Personal Author</td>
<td>100, 600, 700</td>
</tr>
<tr>
<td>CA</td>
<td>Corporate Author</td>
<td>11x, 61x, 71x</td>
</tr>
<tr>
<td>SU</td>
<td>Subject (keywords)</td>
<td>6xx</td>
</tr>
<tr>
<td>TW</td>
<td>Title Words</td>
<td>x30, 21x, 24x, 740</td>
</tr>
<tr>
<td>UT</td>
<td>Uniform Title</td>
<td>x30</td>
</tr>
<tr>
<td>SE</td>
<td>Series</td>
<td>4xx, 8xx</td>
</tr>
<tr>
<td>XT</td>
<td>Exact Title</td>
<td>245</td>
</tr>
<tr>
<td>XS</td>
<td>Exact Subject</td>
<td>6xx</td>
</tr>
<tr>
<td>XC</td>
<td>Exact Corporate Author</td>
<td>11x, 61x, 71x</td>
</tr>
<tr>
<td>AU</td>
<td>Combines PA and CA</td>
<td>11x, 61x, 71x</td>
</tr>
<tr>
<td>TI</td>
<td>Chooses TW or XT</td>
<td>11x, 61x, 71x</td>
</tr>
</tbody>
</table>

*Since November 1, 1986 the system attempts to decide if the search is for words or an exact title. The TI index was originally the same as TW.*
A detailed examination of the MELVYL system has been published as a special section in the December 1982 (v. 1, no. 4) and March 1983 (v. 2, no. 1) issues of Information Technology and Libraries.

Transaction Monitoring Data

The data used in this study were derived from transaction monitor records of the MELVYL system. The transaction monitor in MELVYL has been described previously (Larson, 1981, 1986). In summary, it was designed to permit detailed analysis of individual user transactions and system performance characteristics. The individual transaction records provide enough information for analysts to reconstruct the events of any user session, including all searches, displays, help requests, and errors, and the system responses. Weekly statistical summaries are generated based on the collected individual transactions for a week, and provide frequency data on index usage in searching, display types, error types, and summarize performance information (such as average response time for search and display transactions.)

The data examined for this research were extracted from weekly summaries of system use from February 1982 through the first week of January 1988, representing over 5.3 million user sessions with over 15.3 million search commands. The analysis concentrates solely on searches issued by “Command Mode” users of the MELVYL system (3.2 million sessions with 10.7 million searches).

The data available for the analysis consisted of two blocks, with a gap during the period of transition from the MELVYL prototype to the production version with an updatable database. The first block, consisting of data from the prototype period, includes transactions between February 1982 and November 1983. The second block, consisting of data from the production version of MELVYL, includes transactions collected from April 1985 through the first week of January 1988. A total of 209 weekly reports were used in the analysis. Some weeks are missing from these blocks due to system downtime, and loss of transaction data. However, for the weeks that data were available, they represent all interactions between the users and the system from that week, not a sample.

Table 2 shows the number of user sessions and searches that were included in the weekly transaction summaries for each year covered in this study. The column labelled “Weeks included” indicates the number of weekly reports available for the given year. The numbers of searches and sessions using either MELVYL LOOKUP mode or command mode are also shown in Table 2.

Analysis of the Data

The weekly summaries from each of the blocks described above were converted to a common format for analysis. The Statistical Analysis System (SAS) was used to process the resulting 209 weekly summary records.

In the following analyses, a “subject search” refers to a command mode search using either the SU or XS indexes as the sole index in the search. A “title keyword search” refers to a command mode search using the TW index or the TI index (when the system treated it as a TW search). Other command mode indexes are referred to by the names in Table 1. This analysis concentrates on command mode searches that used either the subject index, title keyword index, personal author index or corporate author index. All other command mode searches using a single index, those that used Boolean combinations of indexes, and LOOKUP mode searches were excluded from the analysis. Table 3 shows the numbers of searches included in the analysis by the index used for each year.

The figures used below for the frequency of searching in a given index were derived by calculating the percentage of all command mode search activity in the subject, title keyword, personal author, or corporate author indexes for the week. The excluded indexes were not included in these percentages, and therefore, the frequency figures are somewhat inflated when compared to the actual percentage for all indexes. This “normalization” permits comparison between data from

| Table 2. MELVYL transactions during the data collection period, by year. |
|-------------------------|-----------------|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|
| Year | Weeks Included | Total Sessions | Total Searches | LOOKUP Sessions | LOOKUP Searches | Command Sessions | Command Searches |
| 1982 | 45 | 487,233 | 1,531,076 | 309,407 | 871,573 | 177,826 | 659,503 |
| 1983 | 41 | 445,156 | 1,372,542 | 279,004 | 776,735 | 166,152 | 595,807 |
| 1984 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1985 | 23 | 428,215 | 1,197,941 | 153,760 | 360,094 | 274,455 | 837,847 |
| 1986 | 48 | 1,709,768 | 4,916,940 | 590,722 | 1,761,734 | 1,119,046 | 3,655,706 |
| 1987 | 51 | 2,229,493 | 6,220,368 | 743,132 | 1,356,952 | 1,486,361 | 4,863,416 |
| 1988 | 1 | 47,119 | 133,746 | 15,078 | 25,947 | 32,041 | 107,799 |
| Totals | 209 | 5,346,984 | 12,512,613 | 2,091,103 | 4,652,535 | 3,255,881 | 10,720,078 |

* A “session” is the set of transactions conducted between an explicit logon and logoff on the MELVYL system. Usually a session will consist of all transactions for a single user.
the prototype and production versions of MELVYL, by eliminating the variation in the actual percentages caused by the addition of new indexes to the system.

**Frequency of Subject Searching over Time**

The frequency of subject searching, considered as a normalized percentage, was examined over the entire data collection period and separately for data collected from the prototype and production versions of MELVYL. The data were examined mathematically to determine the existence and magnitude of any trends in the frequency of subject searching.

The equation used to describe the relationship between the percentage of subject searches and time took the form of a simple linear model. It should be noted that the intention here is primarily to examine trends in the frequency of subject searching retrospectively, and not to predict future levels of subject searching. Any attempt at prediction based on the models described below should be approached with extreme caution. In the following equations, the percentage of subject searching was termed $Y_s$ and time, as the number of days since the beginning of the data collection period, was termed $X_t$. The first model examined was

$$Y_s = \alpha + \beta_1 X_t + \varepsilon$$

where $\alpha$ and $\beta_1$ are constants and $\varepsilon$ is a random error term.

In this model the sign of the $\beta_1$ parameter indicates whether the frequency of subject searching is increasing (+) or decreasing (−) and its value shows the rate of the change.

All weekly summary data were submitted to least squares regression analysis using this model to estimate the value of the $\alpha$ and $\beta_1$ parameters. The resulting equation was:

$$Y_s = 50.9963 + -0.0062 X_t$$

$$R^2 = 0.3405 \quad F = 106.873 \quad n = 209$$

($t$ statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

Figure 1 shows a scatter plot of the data (subject searching percentage on the $Y$ axis and number of days since the beginning of the data collection, shown as a date, on the $X$ axis) with the regression line superimposed. As the $R^2$ statistic indicates, a little over a third of the variability in the data can be explained by the model. This model indicates that the percentage of subject searching in MELVYL command mode decreases by about 0.0062% per day over the data collection period. This translates to a decline of about 2.26% per year in the frequency of subject searching.

The remaining variability in the model was examined to determine if significant patterns of subject search frequency could be found. Previous research (Borgman, 1983) had suggested that the percentage of subject searching was lower during Summer term, so model (1) was modified to account for holiday and Summer breaks in the academic year. (This was somewhat complicated by the differing calendars on the nine University of California campuses.) A second model was formulated as

$$Y_s = \alpha + \beta_1 X_t + \beta_2 X_h + \varepsilon $$

where $\alpha$ and the $\beta_1$ are constants and $\varepsilon$ is the error term.

The $X_t$ variable remains the same as defined above, and $X_h$ has the following definition

$$X_h = \begin{cases} 
1 & \text{if the week occurred during a holiday or Summer break} \\
0 & \text{otherwise} 
\end{cases}$$

All weekly summary data were submitted to least squares regression analysis using this model. The resulting equation was

$$Y_s = 55.6756 + -0.0059 X_t + -10.1193 X_h $$

$$R^2 = 0.7655 \quad F = 336.326 \quad n = 209$$

($t$ statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

Figure 2 shows the regression line for this model superimposed on the frequency data. As the $F$ statistic
indicates, this model provides a better fit to the data. The model accounts for about 76% of the variability in the data as shown by the $R^2$ statistic. The $\beta$, parameter estimate, indicating a 0.0059% decline per day (2.15% per year) in the frequency of subject searching, remains quite close to that in the first model. The $\beta_3$ parameter estimate shows that a rather drastic reduction (10.12%) in the frequency of subject searching occurs during holiday and summer break periods. This "seasonal" variation will be discussed further below.

A third model, separating the $X_3$ variable into three separate components (i.e., Summer break, Winter or Christmas break, and Spring break) was also examined. Although this model provided a marginal increase in the
and showed some small differences in the $\beta_i$ parameter estimates (there was less reduction in subject searching frequency during Spring break), the $F$ statistic indicated a better fit for model (2).

Each of the blocks of data was analyzed separately to determine variations from the model (2) estimates in the MELVYL prototype and production systems.

The resulting equation for the prototype period (transactions between February 1982 and November 1983) was

$$Y_s = 52.8771 + 0.0008X_t - 8.6045X_h$$

$$R^2 = 0.6902 \quad F = 92.454 \quad n = 86$$

(the $t$ statistics are given in parentheses below each coefficient. The $t$ statistic for $\beta_i$ was not statistically significant).

Although the model shows a slight increase in the frequency of subject searching during the prototype period, the $F$ value indicates a poor fit to the data and the lack of significance for the $\beta_i$ parameter estimate suggest that time was not a factor in the variability of subject searching during the prototype period.

The prototype data was reexamined using the following equation

$$Y_s = \alpha + \beta_h X_h + \epsilon$$

where $\alpha$ and $\beta_h$ are constants and $\epsilon$ is the error term. In this equation the sign of the $\beta_h$ parameter indicates whether the percentage increases (+) or decreases (−) during a holiday and its value shows the amount of the change.

The data were submitted to least squares regression analysis using this model. The resulting equation was

$$Y_s = 53.1565 - 8.5927X_h$$

$$R^2 = 0.6893 \quad F = 186.315 \quad n = 86$$

($t$ statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

Comparing these models we find that the frequency of subject searching during the MELVYL prototype period was dominated by the seasonal variation observed in model (2), but that the passage of time does not appear to be a major factor in the frequency of subject searching. Figure 3 shows the model (3) regression line superimposed on the prototype period data.

The data from the production version of the MELVYL system (transactions collected from April 1985 through the first week of January 1988) were examined using the model (2) equation. The resulting equation for the production period was

$$Y_s = 55.1646 - 0.0054X_t - 11.1579X_h$$

$$R^2 = 0.6599 \quad F = 116.434 \quad n = 123$$

($t$ statistics are given in parentheses below each coefficient. All statistics are significant at the .0001 level.)

Figure 4 shows the regression line from this equation.
superimposed on the production period data. As the \( \beta_t \) parameter estimate indicates, the decline in the frequency of subject searching during the production period was quite close to that observed for the entire data collection period, with a somewhat stronger effect observed for the \( \beta_h \) "seasonal" component.

From the results of these separate analyses of the prototype and production period we can see that there is a time-like element affecting the frequency of subject searching during the production period that is not a significant factor in the prototype period. This factor appears to be the size of the database, which was static during the prototype period and has, like time, increased monotonically during the production period.

**Replacements for Subject Searching**

The preceding analyses of subject search frequency indicate a gradual decline in the use of the subject index as the size of the database increases. The frequency of subject searching was considered as a percentage, thus the frequency of searching in other indexes must increase to offset the observed decline. This section examines these other indexes, concentrating on the frequency of title keyword searching.

**Title keyword searching frequency on the MELVYL system presents an additional problem in analysis. That is, on November 1, 1986 the MELVYL user interface was modified to interpret a search in the "TI" index as either a title keyword (TW) or exact title (XT) search based on the number of words and usage of words in the search. Prior to this date all exact title searches had to specify the XT index, and all TI searches were title keyword searches. In the following analysis a "title keyword search" is any search that either (a) used the TI index prior to November 1, 1986, (b) specified the TW index after that date, or (c) specified the TI index and was interpreted by the system as a TW search. The following equation was used to model the frequency of title keyword searching

\[
Y_T = \alpha + \beta_r X_r + \beta_s X_s + \beta_d X_d + \epsilon
\]

where \( \alpha \) and the \( \beta_i \) constants and \( \epsilon \) is the error term. \( Y_T \) is the frequency of title keyword searching. The \( X_r \) and \( X_s \) variables remain the same as defined in model (2) above, and \( X_d \) has the following definition

\[
X_d = \begin{cases} 
1 & \text{if date > Nov. 1, 1986} \\
0 & \text{otherwise}
\end{cases}
\]

All weekly summary data were submitted to Least Squares regression analysis using this model. The resulting equation was

\[
Y_T = 18.7223 + 0.0077X_r + 4.4087X_s - 6.7870X_d
\]

\[ (46.280) \quad (21.245) \quad (11.817) \quad (-11.522) \]

\[ R^2 = 0.7683 \quad F = 226.625 \quad n = 209 \]

(\( t \) statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

Figure 5 shows the regression line for this equation superimposed on the title keyword frequency data for the entire data collection period. The \( \beta_i \) parameter esti-
mate indicates that the decline in subject searching seen in model (2), $-0.0059\%$ per day, has been replaced with an increase of $0.0077\%$ per day in the frequency of title keyword searching. The $\beta_h$ parameter estimate suggests, however, that title keyword searching is not the sole replacement index for observed decreases in subject searching during holiday and Summer break periods. The $\beta_a$ parameter estimate shows a $6.79\%$ decrease in title keyword searching when the automatic interpretation of TI searches by the system was installed. It appears, however, that users quickly learned the new feature and began using the TW index in place of the TI index.

Figure 6 shows the frequency of subject searching (solid line) and the frequency of title keyword searching (dashed line) over the data collection period. Figure 6 demonstrates the overall relationship of subject searching and title keyword searching, based on the normalized data. The frequency of title keyword searching exceeds that of subject searching at several points during the production period. The actual (unnormalized) percentages show a similar pattern, though the actual subject searching percentage is lower, particularly following November 1986 due to a large increase in exact title searching. Figure 7 shows the actual percentages for subject searching and title keyword searching over the data collection period.

Searches in the Personal Author (PA) index account for most of the remaining searches on the MELVYL system. PA search frequency was modelled using the following equation

$$Y_p = \alpha + \beta_i X_i + \beta_h X_h + \epsilon \quad (5)$$

where $\alpha$ and the $\beta_i$ are constants and $\epsilon$ is the error term. The $X_i$ variables are the same as defined above for model (2).

All weekly summary data were submitted to Least Squares regression analysis using this model. The resulting equation was

$$Y_p = 22.6642 + 0.0010X_i + 4.3966X_h$$

$$R^2 = 0.5328 \quad F = 117.443 \quad n = 209$$

(the $t$ statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

The low value for the $\beta_i$ parameter estimate indicates that PA search frequency has changed very little (about 2%) over the data collection period. The $\beta_h$ parameter estimate shows that PA searching makes up most of the remaining "seasonal" variation in subject searching.

Corporate author searches make up only about 1.9% of the total searches on the MELVYL system. CA search frequency was modelled using the following equation

$$Y_c = \alpha + \beta_i X_i + \beta_h X_h + \epsilon \quad (6)$$

where $\alpha$ and the $\beta_i$ are constants and $\epsilon$ is the error term. The $X_i$ variables are the same as defined above for model (2).

All weekly summary data were submitted to least squares regression analysis using this model. The re-
The resulting equation was
\[ Y_c = 1.6367 + 0.0002X_i + 0.7482X_h \]

\[ (21.506) \quad (4.458) \quad (10.327) \]

\[ R^2 = 0.3873 \quad F = 65.103 \quad n = 209 \]

(\(t\) statistics are given in parentheses below each coefficient. All statistics were significant at the .0001 level.)

The very low value for the \(\beta_i\) parameter estimate indicates corporate author search frequency has remained virtually the same over the data collection period. The \(\beta_h\) parameter estimate shows that CA searching ac-
counts for less than 1% of "seasonal" variation in observed in subject searching. These results indicate that there have been no significant changes in the frequency of corporate author searching over the data collection period.

**Variability in Index Use during the Year**

Figure 8 summarizes the pattern of search frequency in the various indexes over the year. The Y axis in Figure 8 represents the average frequency (mean normalized percentage) over all data from given month, regardless of the year, and the X axis shows the month. All weekly data from all years contribute to these mean values, so they do not sum to 100%. They do, however, present a general picture of index use patterns during the year. As indicated in the models described above, months with a holiday or Summer break have a lower average frequency of subject searching, with concomitant increase in title keyword and personal author search frequencies.

**Discussion**

The preceding analysis shows a persistent decline in the use of the subject index on the MELVYL system, with the rate of decline at about 2.2% per year over a six-year period. Title keyword access appears to be adopted as a replacement for subject index access by users. However, if we consider the traditional dichotomy between "known item" and "subject" searches, there is a large increase in known item searching.

Figure 9 shows the percentages of known item (defined as the sum of PA, CA, XCA, XTI, and AU search percentages) and "topical" searches (defined as the sum of SU, XSU, and TI percentages), with other indexes and Boolean combinations of indexes excluded. This is a liberal definition of topical searching, because many title keyword searches were for known items, especially during the prototype period when the TI index was the only title index available. In spite of this, Figure 9 shows that over the data collection period the known item search percentage has continually risen, and began to exceed the topical search percentage in late 1986. The following discussion considers some of the likely reasons behind the decline in subject searching and the rise of known item searching.

**Problems Leading to Reduced Subject Searching**

Users experience a number of problems when doing searches in the subject index, whether in an online catalog or a card catalog. These have been discussed by many researchers (see, for example, Markey (1983, 1984, 1985, 1986), Bates (1977, 1986), Borgman (1986), McCarthy (1986), Van Pulis and Ludy (1988), Carson (1985), Kern-Simirenko (1983), and Blazek and Bilal (1988)). The subject index, even after the decline discussed above, is still one of the most commonly used search access points in the online catalog. It is also, unfortunately, the access point most likely to fail, often providing either no items that match the user's search request, or too many items to evaluate conveniently (Tolle, 1983; Larson, 1986; Markey, 1988). In the follow-
ing discussion we will refer to a search that retrieves nothing, sometimes called a "zero hit" search, as a search failure, and a search that retrieves "too much" as "information overload."

Search Failure

There are several reasons for failure of subject searches, but the most common are misspelling of search terms and lack of knowledge concerning Library of Congress Subject Headings (LCSH) on the part of the user (Bates, 1977, 1986; Markey, 1983, 1986; Steinberg & Metz, 1984). Larson (1986) found that 48.5% of subject searches on the MELVYL system failed to retrieve any records. Other estimates of subject search failure range from 35% to over 50% (Markey, 1986).

Some researchers have suggested that keyword searching of LCSH alleviates some of the problems experienced by users (Lawrence, 1985; Markey, 1988), but it will often introduce "false drops," due to the appearance of same keywords in headings from divergent topical areas. User difficulties in understanding and applying Boolean operations in keyword-based systems also contribute to search failures (Borgman, 1986).

Because no syndetic structure is currently available in most online catalogs to aid the user by providing cross-references to preferred (LCSH) terms, the users must either know the preferred terms, or are likely to fail in their searches. Frost (1987) and Markey (1984), based on surveys and transaction monitor data, suggest that the majority of users use "whatever popped into the searcher's mind" (Markey, 1984, p. 70) for their search terms. Although bibliographic instruction pro-
grams by libraries can promote use of the LCSH list in constructing subject searches, it is not likely that such programs will reach more than a small percentage of library users.

Even when the machine readable version of the Library of Congress Subject Authority file is made available (Markey, 1988, Van Pulis & Ludy, 1988), it adds only those terms recognized by LC as potential lead-in vocabulary to the established headings, and lacks terms based on the names of persons or organizations and other proper nouns, such as chemical compounds. Analysis of headings used in a large catalog and LCSH (Frost & Dede, 1988) showed that only 43.9% of headings used in the catalog matched LCSH exactly. However, a study by Van Pulis and Ludy (1988) has shown that inclusion of LCSH authorities in an online catalog does provide a significant number (9% of their sample of 203 searches) of matches on cross-references leading to authorized terms.

Figure 10 illustrates the failure rate for all types of searches (i.e., the percentage of searches that retrieved nothing, regardless of the index used) on the MELVYL system over the data collection period. Analysis of the data showed a significant negative correlation between the failure rate and time (Pearson's $r = -0.47, p = .0001, n = 209$), and a significant positive correlation between the failure rate and the percentage of subject searching (Pearson's $r = 0.46, p = .0001, n = 209$). This suggests that the improvement in search success is related to the decrease in the use of the subject index.
**Information Overload**

As an online catalog database grows, increasing numbers of bibliographic records will match a user's subject search. When the system is keyword-based, the rate of increase is more rapid than in those relying on exact matching of LCSH. Use of truncation also increases the number of matching records. This raises the "information overload" problem common to conventional online catalog systems; eventually users are faced with retrievals that provide too much information to evaluate.

There is no precise definition of what constitutes "too much," the dividing line between acceptable retrieval and information overload depends on the individual user's needs and tolerance for scanning through screens of retrieved items. Wiberley and Daugherty (1988) provide a good summary of observations on the information overload problem in both online and manual information systems. Blair (1980) has discussed the concept of information overload in terms of a user's "futility point" in examining the results of retrieval. According to Blair's definitions, it is not enough for the searcher to predict one or several of the terms used to index the desired documents, he must also retrieve these documents in a set small enough for him to browse through. The number of documents in a retrieved set which includes the desired documents should be smaller than the searcher's "futility point." This futility point is the number of retrieved documents the inquirer is willing to browse through before giving up his search in frustration. But there is another kind of futility point—what might be called the "anticipated futility point." This point is the maximum number of retrieved documents that an inquirer would be willing to begin browsing through. It represents the largest size retrieved set of documents he is willing to look at. (Blair, 1980, p. 271)

The futility point for a particular user depends, to a large extent, on the goal or motivation for the search. Examination of transaction logs has indicated that many users rapidly reach their futility point when faced with a large number of retrieved records. Larson (1986) found that the average number of records retrieved by searches (in all indexes) on MELVYL was 77.5, but users only looked at (displayed) an average of about 9.1 records per search.

The MELVYL system handles searches that have large (or potentially large) retrieval sets by using special "long search processing." That is, when any keyword in a search has a large number of associated records, a message is displayed to the user indicating that the search is a "long search" and giving the option of canceling the search. When the user elects to continue such a search, it is processed at a lower priority than normal searches in a number of stages (known as "long search cycles"). Long searches commonly take over a minute to complete (some have taken over 20 minutes), and may still fail due to null Boolean intersections. User cancellation of long searches is quite common, adding to the failure rate for subject searching.

Evidence of the information overload problems may also be found in the transaction log data. In January 1986, the weekly summaries from the transaction monitoring system began recording the average number of

---

**FIG. 10. Combined search failure rate for all indexes.**
items retrieved for each search index, as well as the frequency of use. Figure 11 shows the average numbers of items retrieved by searches in the MELVYL keyword indexes (i.e., the mean number of items retrieved for all searches in the index for a given week). As Figure 11 shows, the average number of items retrieved by subject keyword searching is in the neighborhood of 200 items, and is generally the highest of any keyword index. Obviously, in individual searches the results must often exceed the user's futility points.

We believe that the replacement of subject searching with title keyword searching, as discussed above, indicates that users are attempting to avoid the search failure problems presented by LCSH. Keywords from titles offer a limited form of "natural language" access to online catalog records, and thus avoid the need to know LCSH. However, the lack of vocabulary control for title keywords places the user the burden of finding any synonyms to the search terms chosen. This lack also leads to the retrieval of much irrelevant material. For example, a searcher interested in mirrors would probably not want Barbara Tuchman's book *A Distant Mirror*, yet a searcher interested in Medieval Europe would not want to miss it. But, as Figure 11 shows, title keyword searches usually result in a much more manageable retrieval set size (an average of about 30 items) than do subject searches.

The switch to title keyword searching seems to indicate that the desire to do topical searching has not diminished, but that the penalties incurred by the user in the process of using the subject index have led to the decline in its use. These penalties might be described as a Scylla of search failure (due to difficulties with LCSH, misspellings, etc.), and a Charybdis of information overload with hundreds or thousands of matching items, with relatively few searches managing to steer a "middle course" between these extremes. On the MELVYL system this middle course is quite narrow, with only about 12% of subject searches retrieving between 1 and 20 items (Larson, 1986, p. 251).

We might consider the decline in subject searching as a case of "aversive operant conditioning." If it is likely that some unpleasant result (search failure or information overload) will occur each time a user does a search in the subject index, he or she probably will not want to repeat the experience.

*User Experience and Subject Searching*

Studies of card catalog use (see Markey (1980) for a good summary) showed that more experienced users of the card catalog, including library staff, faculty, and advanced students, were more likely to perform known-item searches than subject searches. There is also evidence that user experience with the online catalog is also associated with less subject searching. Matthews and Lawrence, in a detailed analysis of data from the CLR survey, found that "topical searching is more prevalent among those who are less experienced with the library and its catalogs" (Matthews & Lawrence, 1984, p. 362). They also confirmed, based on analysis of the CLR data from academic libraries, the "common
wisdom" from card catalog studies that undergraduates are more likely to do subject searches than graduate students (Matthews & Lawrence, 1984, p. 362). Other research has supported these findings, though often indirectly.

Blazek and Bilal (1988) conducted a questionnaire survey of users who approached the reference desk with problems in searching the online catalog. They found about 75% of these problems were related to subject searching, but that users with greater experience had far fewer problems relating to subject access. They did not, however, examine whether experienced users had fewer subject search problems because they did less subject searching.

Frost (1987), however, found that subject searching by one university faculty, in both the card and online catalogs, was more frequent than expected. Although the levels of subject searching reported by Frost are in the low ranges of subject search frequency reported above, her findings also go against the "common wisdom" that faculty use known item searches almost exclusively.

If we make the assumption that users of the online catalog during holiday and Summer break periods are more likely to be library staff, faculty and graduate students than undergraduates (i.e., those with research projects longer than a single term). And further assume that library staff, faculty, and graduate students are likely to have more experience with the online catalog (due to more frequent use of the library) than undergraduates. Then a possible explanation for the dramatic decrease in subject search frequency during the holiday and Summer breaks described above may be simply that users during those times are more experienced in using the system. Lipetz (1972) and Lipetz and Paulson (1987) have observed that catalog users have a tendency to "sublimate" subject searches in a catalog that is inhospitable by using alternative methods of access. Experience in catalog use may not necessarily imply that users have been "conditioned" to avoid subject searches, although such conditioning appears to be a likely result of gaining experience in catalog use, whether card or online catalog. We would suggest, as a hypothesis for further study, that individual users' experiences of subject search failure and information overload lead them to reduce their use of the subject index and to increase their use of alternate means of subject access, such as title keyword searching and shelf browsing following a known item search (Hancock, 1987). An alternative hypothesis is that experienced users are more effective searchers, and need to perform fewer subject searches to achieve satisfactory retrieval.

There is some indirect evidence from the transaction logs that the general level of user experience on the MELVYL system is increasing. This is found in the relative percentages of LOOKUP mode and command mode use. LOOKUP mode is specifically directed towards the new (inexperienced) user, and command mode towards the more advanced user. Command mode use is fostered by bibliographic instruction programs on the U.C. campuses. Figure 12 shows the relative percentages of command mode and lookup mode sessions over the data collection period. As the figure shows, command mode has become the dominant form of interaction with MELVYL.

However, there is a need for further research on the relationship between users' experience with the online catalog and their searching behavior to test the hypotheses offered above. In particular, a longitudinal study of a set of users from their first introduction to the online catalog through one or two years of use, using transaction monitoring to record search behavior, and questionnaires and interviews to trace changing attitudes and needs, would answer many questions that can only be approached indirectly using transaction data alone.

Remedies to Subject Searching Problems

The preceding discussion identified two major problems with subject searching in online catalogs, search failure and information overload. The evidence indicates that these problems are the likely causes for the observed decline in subject search frequency. In this section we will focus on some suggested improvements to online catalog systems as potential remedies for these problems.

There are three primary facets of online catalog systems that can be improved to enhance subject search capabilities. These are:

(1) The database.
(2) The search processing and retrieval algorithms.
(3) The user interface.

Many of the proposed solutions to the problems of subject searching in online catalogs concentrate on a single problem and facet of the system, and tend to neglect the opposing problem. In effect, they suggest that to avoid Scylla, one must steer towards Charybdis, or vice versa. Various proposals and experiments that have dealt with one or more of these facets are discussed briefly below.

Enhancing the Database

Most proposals for improving subject searching in online catalogs tend to focus on alleviating search failure by providing additional access points to the catalog record. Several methods of enriching the topical contents of bibliographic records have been discussed by Mandel (1985). Record enhancement methods include Atherton's (1978) Subject Access Project, where records were expanded by adding words from the book's index and table of contents, and Markey's (1985, 1987) project...
for enhancing records with terms from the Dewey Decimal Classification (DDC) schedules and relative index.

A recent report on the application of Atherton's method at the Australian Defense Force Academy (Byrne & Micco, 1988), indicated that about 15 minutes per item were required to provide an average of 20.7 additional headings (53.6 keywords) derived from tables of contents and indexes. The difficulties of applying this method retrospectively to a large collection are obvious. The primary effect of the method is to increase the number of items retrieved in a keyword search. Byrne and Micco estimate that retrieval increased by 300% with the additional terms. They note that "the number of records retrieved even in this small database of 160,000 titles is unmanageable" (Byrne & Micco 1988, p. 440), although only 6,000 items had been enhanced by additional terms.

The method employed by Markey in the DDC project does not need manual reindexing of records, but it does require that machine readable versions of the classification schedules be available. This effectively limits the application to the DDC, because no complete machine-readable schedules are available for the LC classification scheme. The hierarchical structure of the DDC was exploited to permit browsing of broader and narrower topics. Markey (1987) presents the results of comparative testing between the DDC-based system and a similar subject term-based system. Although no clear preference on the part of the users was found, Markey suggests that this may be due to problems with the particular implementation of the system, and user difficulties with DDC terminology.

Bates (1986) has offered a convincing argument that the search failure problem would be best handled not by increasing the number of headings assigned to individual records, but by providing a large "end-user thesaurus" with a rich syntactic structure. The end-user thesaurus envisioned by Bates could incorporate a number of lesser thesauri, classification structures, and subject heading lists. Moreover, it would include a very large entry vocabulary that could be used to map from users' natural language terms to the controlled vocabulary applied to the records. Such a thesaurus would constitute a separate database, albeit linked to the bibliographic records. Integration of the machine-readable version of LCSH into an online catalog is one step in the direction of an end-user thesaurus, but requires much enhancement to deal with the inconsistencies and limited entry vocabulary of LCSH (Markey, 1988).

Although enhancing the database seems to offer at least a partial solution to the problem of search failure, when used alone it runs the risk of exacerbating the information overload problem.

Enhancing Search Processing and Retrieval Algorithms

Many suggestions for dealing with information overload take a "Procrustean Bed" approach to limiting search results. That is, Boolean intersections using dates or additional terms are suggested as a way to re-
duce the size of search results. This rather arbitrary "chopping" of the retrieval set doesn’t consider the potential relevance of older material or that the scarcity of terms in MARC records (when not enhanced) leads quickly to empty Boolean intersections.

A more promising line of research for improving online catalog subject access appears to be the application of methods and algorithms derived from information retrieval (IR) research. These methods include partial matching and stemming of keywords (i.e., removal of suffixes and prefixes from keywords), ranking of retrieval output according to probability of relevance or coordination level matching (that is, the number of terms in common between the query and the record (Losee, 1987)) in place of Boolean logic, automatic mapping from input search terms to controlled vocabulary terms through thesaurus look-up, and relevance feedback. Salton (1989) and Cooper (1988) provide good introductions to these IR techniques.

Some of these enhancements aid in alleviating search failure by dealing with the search itself, such as the use of stemming, expansion of the search by thesaurus look-up or relevance feedback, and automatic spelling correction or phonetic matching of terms. Others, such as ranked output, require modifications to the existing retrieval algorithms of the online catalog.

The IR techniques address both sides of the subject search problem: stemming and thesaurus lookup help avoid search failure, and ranked output reduces information overload. Hildreth (1987) has suggested that systems using these methods will be the “next generation” of online catalogs.

Some “next generation” online catalog systems already exist, and have implemented some of these IR methods. They include PaperChase at Beth Israel Hospital (Horowitz & Bleich, 1981), CITTE at the National Library of Medicine (Doszkocs, 1983), OKAPI at the Polytechnic of Central London (Mitev, Venner, & Walker, 1985). Other experimental online catalog systems that use IR techniques and hypertext features include HYPERCATalog under development at LILAB in Sweden (Hjerppe, 1986), the BOOK HOUSE system in Denmark (Pejtersen, 1989), and the CHESHIRE system under development at Berkeley (Larson, in press).

As observed above, when a subject search succeeds in a large online catalog like MELVYL, the user is usually presented with a large number of records, arranged by author and title, with no attempt to differentiate between variant uses of a term. If the results were presented, in summarized form, as clusters of strongly related records, organized according to classification areas and ordered by the degree of match between the query and the cluster, it is believed that users would be able to select relevant items more quickly and largely avoid the information overload problem. Similarly, if stemmed keywords from both titles and LCSH are used to provide an entry vocabulary to these “classification clusters,” many search failures can be avoided. These methods are being explored using the CHESHIRE system.

Enhancing the User Interface

Other methods of enhancing subject access have approached the problem from a different point of view. Rather than adding additional access points or changing the basic retrieval techniques of the system, these methods propose improving the online catalog’s user interface to increase the “browsability” of existing subject headings and classification assignments (Hildreth, 1982). There is increasing interest in the design of the user interface, including the means of specifying searches and the formats and organization of retrieved information.

Some of the proposed changes are intended to be easily applicable to existing online catalog systems. For example, Massicotte (1988) has proposed a method of compressing lengthy LCSH “browse” displays by merging form and geographic subdivisions in subject headings and permitting the user to selectively expand those of interest. Similar display methods have been discussed by Markey (1986). These proposals would help relieve the problem of information overload by simplifying the display of search results, but they do little to aid the searcher in coming up with a correct heading to begin browsing.

Other approaches utilize the graphical display capabilities of workstations or personal computers to simplify interactions and facilitate browsing. These range from front-end gateways to existing online catalogs (most commonly using HyperCard on Macintosh PCs) to complete artificial environments combining graphical displays with powerful underlying search techniques (see, for example, Pejtersen (1989)).

We would suggest that no single method will provide a complete solution to the problems of subject searching, but that each of the facets of the system (database, retrieval algorithms, and user interface) need to be enhanced to contribute to a solution. We should begin to consider the topical elements of online catalog databases (subject headings, keywords, and classification assignments) not as separate inviolate entities, but as related clues to the content of books that can be combined with synergistic effects.

Conclusions

This study has shown that a constant decline has taken place in the use of the subject index in a large online catalog. Title keyword searching, which provides a limited form of natural-language access to the topics of books, was found to be the primary replacement for subject index use. The causes and implications were explored and evidence provided suggesting that user frustration in using the subject index effectively has led to
this decline. The two major problems contributing to user frustration are search failure, due to difficulties in dealing with LCSH, and information overload, due to the ever-increasing size of the database.

As online catalogs grow and expand their range of services beyond the traditional book catalog to include journal indexes and other information sources (Potter, 1989), it is time to reconsider the nature and structure of subject access in these systems. We should learn from the failures of the past, and not perpetuate them in a new generation of online information systems.

References


