Order Number 9330763

An analysis of search failures in online library catalogs

Tonta, Yaşar Ahmet, Ph.D.

University of California, Berkeley, 1992

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

An Analysis of Search Failures in Online Library Catalogs

by

Yaşar Ahmet Tonta

B.A. (University of Hacettepe) 1981
M.A. (University of Hacettepe) 1985
M.Lib. (University of Wales) 1986

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Library and Information Studies

in the

GRADUATE DIVISION

of the

UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor Michael D. Cooper, Chair
Professor Ray R. Larson
Professor Lawrence A. Rowe

1992

The dissertation of Yaşar Ahmet Tonta is approved:

Michael D Cooper December 1, 1992

Chair Date

December 1, 1992

Date

Lawrence A. Rowe Dec 3, 1992

Date

University of California at Berkeley

1992

An Analysis of Search Failures in Online Library Catalogs

by

Yaşar Ahmet Tonta

ABSTRACT

AN ANALYSIS OF SEARCH FAILURES IN ONLINE LIBRARY CATALOGS

by

Yaşar Ahmet Tonta

Doctor of Philosophy in Library and Information Studies

University of California at Berkeley

Professor Michael D. Cooper, Chair

This study investigates the causes of search failures that occur in online library catalogs by developing a conceptual model of search failures and examines the retrieval performance of an experimental online catalog by means of transaction logs, questionnaires, and the critical incident technique. It analyzes retrieval effectiveness of 228 queries from 45 users by employing precision and recall measures, identifying user-designated ineffective searches, and comparing them quantitatively and qualitatively with precision and recall ratios for corresponding searches. The dissertation tests the hypothesis that users' assessments of retrieval effectiveness differ from retrieval performance as measured by precision and recall and that increasing the match between the users' vocabulary and that of the system by means of clustering and relevance feedback techniques will improve the performance and help reduce failures in online catalogs.

In the experiment half the records retrieved were judged relevant by the users (precision) before relevance feedback searches. Yet, the system retrieved only about 25% of the relevant documents in the database (recall). As should be expected, precision ratios decreased (18%) while recall ratios increased (45%) as users performed relevance feedback searches. A multiple linear regression model, which was developed to examine the relationship between retrieval effectiveness and users' judgments of the search performance, found that users' assessments of the effectiveness of their searches was the most significant factor in explaining precision and recall ratios. Yet, there was no strong correlation between precision and recall ratios and user characteristics (i.e., frequency of online catalog use and knowledge of online searching) and users' own assessments of search performance (i.e., search effectiveness, finding what is wanted). Thus, user characteristics and users' assessments of retrieval effectiveness are not adequate measures to predict system performance as measured by precision and recall ratios.

The qualitative analysis showed that search failures due to zero retrievals and vocabulary mismatch occurred much less frequently in the online catalog studied. It was concluded that classification clustering and relevance feedback techniques that are available in some probabilistic online catalogs help decrease the number of search failures considerably.

Michael D. Cooper, Chair
December 1, 1992

ACKNOWLEDGEMENTS

I have benefited from the valuable help and encouragement of several people in completing this dissertation. I wish to thank and express my sincere gratitude and appreciation to my adviser Dr. Michael D. Cooper for his guidance and invaluable assistance throughout this project. I owe a special thanks to Dr. Ray R. Larson for allowing me to experiment with the system he designed and for providing computer time and unfailing support in the course of the study. I am thankful to Dr. Lawrence A. Rowe for his meticulous review of this report. I am also thankful to Dr. Michael K. Buckland for his encouragement and continuous interest in my studies. I am grateful to Ms. Judy Baker, Dr. Marcella Genz, and Ms. Elisabeth Rebman for encouraging their students to participate in the experiment and for allowing me to demonstrate the system during their scheduled class times. I am also indebted to MLIS and fellow Ph.D. students of the School of Library and Information Studies at UC Berkeley who performed online searches, filled out questionnaires, and endured my structured interviews.

It is my great pleasure to acknowledge the institutional support I received throughout my studies. I wish to thank the Department of Library Science of the University of Hacettepe for granting me a study-leave to pursue my doctoral studies. My especial thanks go to my teacher Dr. İlhan Kum, the founder and former head of the Department, for his leadership and indefatigable efforts which he so generously expended for the betterment of education of the Turkish library and information professionals. I owe a debt of gratitude to the Fulbright Commission of the United States Information Agency for granting me a doctoral fellowship. I am also indebted in so many ways to all of the faculty and staff of the School of Library and

Information Studies at UC Berkeley for providing fellowships, teaching opportunities, and constant help during my studies.

Above all, I owe heartfelt thanks to my wife and our son for their much needed love and ceaseless support throughout my studies. I am also greatly indebted to our families for their understanding, patience, and endurance throughout the years.

Table of Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

> *No one wants to learn by mistakes, but we cannot learn enough from successes to go beyond the state of the art.*
>
> --Henry Petroski, *To Engineer Is Human: The Role of Failure in Successful Design.* (New York: Vintage Books, 1992), p.62.

1.0 Rationale of the Study

Online catalog users often fail in their attempts to retrieve relevant items from document collections using existing online library catalogs. Most users experience problems especially when they perform subject searching in online catalogs. Confronted with an online catalog that lacks guidance or adequate help features, users tend to abandon their searches without questioning the causes of search failures and the effectiveness of the online catalog.

Although it is users who usually endure online catalogs with ineffective user interfaces and struggle with inflexible indexing and query languages, their involvement in the analysis of search failures is seldom sought. Studies with no user involvement tend to focus on what might have happened during a search, rather than what actually happened. Causes of search failures in online catalogs can be studied best when the users provide invaluable feedback regarding their search queries and retrieval results.

This study is an attempt to investigate the causes of search failures in a third generation experimental online library catalog. It is particularly concerned with the evaluation of retrieval performance in online library catalogs from the users' perspective. The analysis of retrieval effectiveness and search failures was based on transaction log records, questionnaires and critical incident technique. User-designated ineffective searches in an experimental online catalog have been compared with transaction log records in order to identify the possible causes of search failures. The mismatch between the users' vocabulary and the vocabulary used in online library catalogs has been studied so as to find out its role in search failures and retrieval effectiveness. An attempt to develop a conceptual model to categorize search failures in online library catalogs was made.

This study evaluates the retrieval performance of an experimental online catalog by: (1) using precision/recall measures; (2) identifying user-designated ineffective searches; and (3) comparing user-designated ineffective searches with the precision/recall ratios for corresponding searches.

Findings obtained from this study can be used to design better online library catalogs. Designers equipped with information about search failures should be able to develop more robust online catalogs which guide users in their search endeavors. Search failures due to vocabulary problems can be minimized by strengthening existing indexing languages and/or by developing "entry vocabulary systems" to relate users' terms to systems' terms. The results may help improve our understanding of the role of natural query languages and indexing in online catalogs. Furthermore, the findings may provide invaluable insight that can be incorporated in future retrieval effectiveness and relevance feedback studies. The conceptual model developed can be used in other studies of search failures in online catalogs. From the methodological point of view, using critical incident technique may prove to be invaluable in studying search failures and evaluating retrieval performance in online library catalogs.

1.1 Objectives of the Study

The purpose of the present study is to:

1. analyze the search failures in online catalogs so as to identify their probable causes and to improve the retrieval effectiveness;

2. measure the retrieval effectiveness in an experimental online catalog in terms of precision and recall;

3. compare user-designated ineffective searches with the effectiveness results obtained through precision and recall measures;

4. ascertain the relationship between performance of the system as measured by precision and recall and variables that defined user characteristics and users' assessment of retrieval effectiveness;

5. ascertain the extent to which users' natural language-based queries match the titles of the documents and the Library of Congress Subject Headings (LCSH) attached to them;

6. identify the role of relevance feedback in improving the retrieval effectiveness in online catalogs;

7. identify the role of natural query languages in improving the match between users' vocabulary and the system's vocabulary along with their retrieval effectiveness scores in online catalogs;

8. develop a conceptual model to categorize search failures that occur in online library catalogs.

1.2 Hypotheses

Main hypotheses of this study are as follows:

1. Users' assessments of retrieval effectiveness may differ from retrieval performance as measured by precision and recall;

2. Increasing the match between users' vocabulary and system's vocabulary (e.g., titles and subject headings assigned to documents) will help reduce the search failures and improve the retrieval effectiveness in online catalogs;

3. The relevance feedback process will reduce the search failures and enhance the retrieval effectiveness in online catalogs.

1.3 Method

Transaction monitoring and critical incident techniques were used for data gathering in this study. The former method allows one to study the users' search behaviors unobtrusively while the latter helps gather information about user intentions and needs for each query submitted to the system. The critical incident technique, which will be described in Chapter III, is used for the first time, to our knowledge, in this study to examine search failures in online library catalogs.

Users participating in the study were allowed access to an experimental online catalog with more than 30,000 records for a period of one semester (14 weeks). Search queries that the users submitted to the system, the items they retrieved and displayed were recorded in transaction logs along with some other relevant data. These transaction logs were later reviewed to find out the retrieval effectiveness of the online catalog under investigation.

As the logs also included data about the users (e.g., their login id) it was possible to identify the person who submitted each query to the system. Users were later invited to share their experience with regard to the searches they performed on the system. Their comments were audiotaped. A critical incident report was completed for each query based on the user's experience. They also were asked to fill out a questionnaire for each search.

The information furnished by the user for each query regarding its effectiveness was compared with the transaction log records. The searches that the users designated as being 'failures' were identified from the critical incident forms and corroborated with the transaction log records. Users' audiotaped comments were also used to analyze the probable causes of the search failures. Thus, it was possible to determine the performance of the online catalog for each search query using both retrieval effectiveness measures such as precision and recall and the user designated search effectiveness.

The critical incident technique proved useful in the analysis of search failures in online catalogs. Incident reports provided invaluable information about each search query regarding its effectiveness. Furthermore, comparison of critical incident reports with the transaction log records was very helpful in identifying and, consequently, analyzing search failures.

1.4 Organization of the Study

This report consists of eight chapters, a select bibliography, and accompanying appendices. The rationale, objectives, hypotheses, and method of the study are introduced in Chapter I, while Chapters II and III form the theoretical foundations of the present study.

Chapter II examines document retrieval systems in general terms. Retrieval effectiveness measures are defined in Chapter II. Relevance feedback and clustering techniques are also discussed in this chapter.

Chapter III opens with a critical review of methods used in the analysis of search failures in document retrieval systems. A comprehensive review of failure analysis studies is given here.

Chapter IV develops a conceptual model to categorize search failures that occur in online catalogs. Types of search failures are examined by means of a four-step ladder model.

A detailed account of the experiment conducted for this study is presented in Chapter V. It explains the environment in which the experiment has been carried out, provides information about the subjects who participated in the study, and illustrates the tools and methods that were used to gather, analyze and evaluate data.

Findings obtained in this study are presented in Chapter VI and VII. Chapter VI summarizes the descriptive data obtained from the transaction logs, questionnaire forms, and critical incident reports. The results of multiple linear regression analysis are also presented in Chapter VI. The detailed analysis of search queries and search failures is given in Chapter VII.

Chapter VIII gives a brief summary of the findings obtained in this study along with conclusions and recommendations for further research.

CHAPTER II:
DOCUMENT RETRIEVAL SYSTEMS

2.0 Introduction

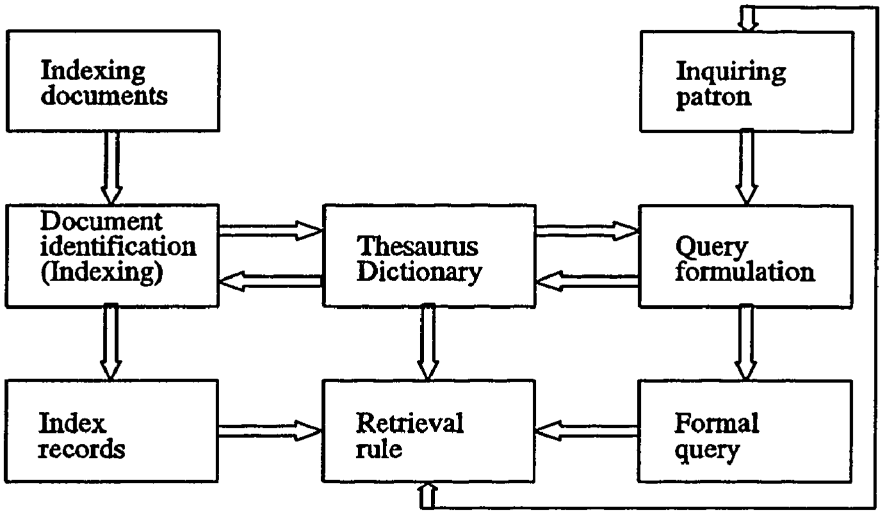
This chapter examines the basic concepts of document retrieval systems and defines major retrieval effectiveness measures such as precision and recall. It also discusses relevance feedback and clustering techniques which are used to enhance the effectiveness of document retrieval systems.

2.1 Overview of a Document Retrieval System

The principal function of a document retrieval system is to retrieve all relevant documents from a store of documents, while rejecting all others. A perfect document retrieval system would retrieve *all* and *only* relevant documents. In reality, the ideal document retrieval system does not exist. Document retrieval systems do not retrieve *all* and *only* relevant documents, and users may be satisfied with systems that rapidly retrieve a few relevant documents.

Maron (1984) provides a more detailed description of the document retrieval problem and depicts the logical organization of a document retrieval system (see Figure 2.1).

FIGURE 2.1 LOGICAL ORGANIZATION OF A CONVENTIONAL DOCUMENT RETRIEVAL SYSTEM (SOURCE: MARON, 1984, P.155)



As Fig. 2.1 suggests, the basic characteristics of each incoming document (e.g., author, title, and subject) are identified during the indexing process. Indexers may consult thesauri or dictionaries (controlled vocabularies) in order to assign acceptable index terms to each document. Consequently, an index record is constructed for each document for subsequent retrieval purposes.

A user can identify proper search terms by consulting these index tools during the query formulation process. After checking the validity of initial terms and identifying new ones, the user determines the most promising query terms (from the retrieval point of view) to submit to the system as the formal query. However, most users do not know about the tools that they can utilize to express their information needs, which results in search failures because of a possible mismatch between the user's vocabulary and the system's vocabulary.

In order for a document retrieval system to retrieve some documents from the database two conditions must be satisfied. First, documents must be assigned appropriate index terms by indexers. Second, users must correctly guess what the assigned index terms are and enter their search queries accordingly. Maron (1984) describes the search process as follows:

> the actual search and retrieval takes place by matching the index records with the formal search query. The matching follows a rule, called 'Retrieval Rule,' which can be described as follows: For any given formal query, retrieve all and only those index records which are in the subset of records that is specified by that search query (p.155).

Thus, a document retrieval system consists of (1) a store of documents (or, representations thereof); (2) a user interface to allow users to interact with the system; (3) a retrieval rule which compares the representation of each user's query with the representations of all the documents in the store so as to identify the relevant documents in the store. It goes without saying that there should be a population of users each of whom makes use of the system to satisfy their information needs.

The major components of an online document retrieval system are reviewed in more detail below.

2.2 Documents Database

The existence of a database of documents or document representations is a prerequisite for any document retrieval system. The term "document" is used here in its broadest sense and can be anything (books, tapes, electronic files, etc.) that carries information. The database can contain the full texts of documents as well as their "surrogates" (i.e., representations).

2.3 Indexing Documents

In order to create a database of documents or document representations, the properties of each document need to be identified and recorded. This process, which is called indexing, can be done either intellectually or automatically. In an environment where intellectual indexing is involved, professional indexers identify descriptive and topical characteristics of the documents and create a record (representation) for each document.

As Fig. 2.1 suggests, indexers can consult the standard tools such as thesauri, dictionaries and controlled vocabulary lists. *Anglo American Cataloging Rules (AACR2)* and the *Library of Congress Subject Headings* List are, among others, used for descriptive and topical analysis of documents, respectively. Indexers then record the document properties and assign subject headings to each document. Recorded descriptive and topical information constitute the representation of the document, which will later be used to provide access points for retrieval purposes.

Automatic indexing, wherein a machine is instructed to recognize and record the properties of documents, has also been used to create index records for retrieval purposes. For topical analysis, automatic indexing relies heavily on terms and keywords used in the full texts (or abstracts) of documents. Words that are useless for retrieval purposes such as "the," "of" and "on" are ignored. Keywords are usually stemmed to their root forms in order to reduce the size of the dictionary of the retrieval-worthy terms. Stemming process also enables the system to retrieve documents bearing variant forms of keywords.

Once the index records are created, the document database will be ready for interrogation by users. The raison d'être of designing a document retrieval system by creating a database of index records is, of course, to serve the information needs of its potential users. We now turn our attention to users' queries and review how the users approach document retrieval systems.

2.4 Query Formulation Process

The query formulation process involves the articulation and formulation of a search query, which by no means is a trivial task. Well-articulated search statements require some knowledge on the user's part. Yet users may not be knowledgeable enough to articulate what they are looking for. Hjerrpe considers this as the fundamental paradox of information retrieval: "The need to describe that which you do not know in order to find it" (Hjerrpe, 1986; cited in Larson, 1991a, p.147).

First time users of document retrieval systems usually act cautiously and tend to enter relatively broad search queries. As the database characteristics (e.g., the number of records and the collection concentration) are not known in the beginning, they try to reconcile their mental models of the system with reality. Sometimes, the reverse may be the case. Users may come up with very specific search queries thinking that the catalog should answer all types of search queries no matter how specific or how broad they happen to be.

As can be seen from Fig. 2.1, dictionaries, thesauri, printed manuals and subject headings lists can be consulted in the course of query formulation process. In addition, some systems offer online help and on-screen instructions to facilitate the query formulation process.

2.5 Formal Query

Once the user's information need is articulated using natural language, a "formal" query statement should be submitted to the system. The syntax of the formal query statement may vary from system to system. In most cases, strict syntactic rules of the command and query languages must be observed in order to enter a formal search statement. Few systems, on the other hand, accept search statements entered in natural language.

Constructing formal query statements is not an easy task. Users must be aware of the existence of a command language and the required commands. In addition, they ought to have some intellectual understanding of how the search query is constructed according to the specifications of the query language. For instance, constructing relatively complex formal query statements using Boolean logic troubles most users.

2.6 The User Interface

Each system is equipped with a user interface which accepts user-entered formal search statements and convert them to a form which will be "understood" by the search and retrieval system. In other words, communication between the system and its users takes place by means of a user interface.

More specifically, the functions of a user interface can be summarized as follows: a) allowing users to enter search queries using either the natural language or the query language provided; b) evaluating the user's query (e.g., parsing, stemming); c) converting it to a form which will be understood by the document retrieval system and submitting the search query to the system; d) displaying the retrieval results; e) gathering feedback from the user as to the relevance of records and reevaluating the original query; and, f) dispensing helpful information (about the system, the usage, the database, and so on).

There are several ways in which users can express their search queries and activate the system (Shneiderman, 1986; Bates, 1989a). The types of user interfaces range from voice input to touch-sensitive screens, from command languages to graphical user interfaces (GUIs), and from menu systems to fill-in-the-blank-type user interfaces. Although the use of voice as input in current document retrieval systems is still in its infancy, other types of user interfaces have been in use for a while. Some are more commonly used than the others. Yet whatever the type of interface used, there is always a "learning curve" involved. To put it differently, users have to master the mechanics of interfaces before they can successfully communicate with the document retrieval systems, submit their search queries and get retrieval results.

Note that an interface is a conduit to the wealth of information that is available in the document database. As far as users are concerned, this conduit should allow every one to tap into the resources regardless of their background and expertise, the amount of information they want, the complexity of the database or the query language, and so on. Mooers' law is also applicable to user interfaces:

> An information retrieval system will tend *not* to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it (Mooers, 1960, p.ii, original emphasis).

It is, perhaps, not too much to suggest that "document retrieval systems will tend not to be used whenever it is more painful and troublesome for patrons to use a poorly designed user interface than not to use it."

2.7 Retrieval Rules

The decisive point in the overall document retrieval process is the interpretation of user's query terms for retrieval purposes. Representation of the formal search requests are matched against that of documents in the database so as to retrieve the record(s) that are likely to satisfy the users' information needs. Thus, the quality of the search outcome hinges very much on the retrieval rule(s) applied in this matching process. Retrieval rules determine which records are to be retrieved and which ones are not.

2.7.1 The Use of Clustering in Document Retrieval Systems

It is important, however, to examine a technique that comes before the application of retrieval rules: *document clustering.*

During earlier document retrieval experiments it was suggested that it would be more effective to cluster/classify documents before retrieval. If it is at all possible to cluster similar documents together, it was thought, then it would be sufficient to compare the query representation with only cluster representations in order to find out all the relevant documents in the collection. In other words, comparison of the query representation with the representations of each and every document in the collection would no longer be necessary. Undoubtedly, faster retrieval to information with less processing seemed attractive.

Van Rijsbergen (1979) emphasizes the underlying assumption behind clustering, which he calls "cluster hypothesis," as follows: "*closely associated documents tend to be relevant to the same requests*" (p.45, original emphasis). The cluster hypothesis has been validated. It was empirically proved that retrieval effectiveness of a document retrieval system can be improved by grouping similar documents together with the aid of document clustering methods (Van Rijsbergen, 1979). In addition to increasing the number of documents retrieved for a given query, document clustering methods proved to be cost-effective as well. Once clustered, documents are no longer dealt with individually but as groups for retrieval purposes, thereby cutting down the processing costs and time. Van Rijsbergen (1979) and Salton (1971b) provide a detailed account of the use of clustering in document retrieval systems.

"Cluster" here means a group of similar documents. The number of documents in a typical cluster depends on the characteristics of the collection in question as well as the clustering algorithm used. Collections consisting of documents in a wide variety of subjects tend to produce many smaller clusters whereas collections in a single field may generate relatively fewer but larger clusters. The clustering algorithm in use can also influence the number and size of the clusters. For instance, some 8,400 clusters have been created for a collection of more than 30,000 documents in Library and Information Studies (Larson, 1989).

Document clustering is based on a measure of similarity between the documents to be clustered. Several clustering algorithms, which are built on different similarity measures such as Cosine, Dice, and Jaccard coefficients, have been developed in the past (Salton & McGill, 1983; Van Rijsbergen, 1979). Keywords in the titles, subject headings, and full texts of the documents are the most commonly used 'objects' to cluster closely associated documents together. In other words, if two documents have the same keywords in their titles and/or they were assigned similar subject heading(s), a clustering algorithm will bring them together.

More recently, Larson (1991a) has successfully used classification numbers to cluster similar documents together. He argues that the use of classification for searching in document retrieval systems has been limited. The class number assigned to a document is generally seen as another keyword. Documents with identical class numbers are treated individually during the searching process. Yet, documents that were assigned the same or similar class numbers will most likely be relevant for the same queries. Like subject headings, "classification provides a topical context and perspective on a work not explicit in term assignments" (Larson, 1991a, p.152; see also Chan, 1986c, 1989; Svenonius, 1983; Shepherd, 1981, 1983). The searching behavior of the users as they search through the book shelves seems to support the above idea and suggests that more clever use of classification information should be implemented in the existing online library catalogs (Hancock-Beaulieu, 1987, 1990).

"Classification clustering method" can improve retrieval effectiveness during the retrieval process. Based on the presence of classification numbers, documents with the same classification number can be brought together along with the most frequently used subject headings in a particular cluster. Thus, these documents will be retrieved as a single group whenever a search query matches the representation of documents in that cluster.

2.7.2 Review of Retrieval Rules

There are several retrieval rules that are used to determine if there is a match between search query terms and index terms. Blair (1990) lists no less than 12 different retrieval rules (called "model") and discusses each in turn in considerable detail.[1] Table 2.1 provides a brief summary of retrieval rules discussed in Blair (1990).

[1] See also Belkin and Croft (1987) for an excellent review of retrieval techniques.

TABLE 2.1
SUMMARY OF RETRIEVAL RULES
SOURCE: COMPILED FROM BLAIR (1990), CHAPTER II.

Model	*Search Request*	*Documents*	*Retrieval Rule*
1	Single query terms	Documents are assigned one or more index terms	If the term in the request is a member of the terms assigned to a document, then the document is retrieved
2	A set of query terms	A set of index terms	Document is retrieved if *all* the terms in the request are in the index record of the document
3	A set of query terms plus a "cut-off" value	A set of one or more index terms	Document is retrieved if it shares a number of terms with the request that *exceeds* the cutoff value
4	Same as 3	Same as 3	Documents showing with the request *more* than the specified number of terms are ranked in order of decreasing overlap
5 Weighted Requests	Set of query terms each of which has a positive number associated with it	Same as 3	Documents are ranked in decreasing order of the sum of the weights of terms common to the request and the index record
6 Weighted Indexing	Set of query terms	Set of index terms each of which has a positive number assigned to it	Documents are ranked in decreasing order of the sum of the weights of terms common to the request and the index record
7 Weighted Requests and Indexing	Same as 5	Same as 6	Documents are ranked by the sum of products each of which results from the multiplication of the weight of the term in the request by the weight of the same term in the index record
8 Cosine Rule	Same as 5	Same as 6	The weights of the terms common to the request and an indexing record are treated as vectors. The value of a retrieved document is the cosine of the angle between the vectors
9 Boolean Requests	Requests are any Boolean combination of query terms with AND, OR, and NOT	A set of one or more index terms	i) AND: Retrieve only documents that match all terms in the request ii) OR: Retrieve only documents that match any term in the request iii) NOT: retrieve all documents that do not match any term in the request
10 Full Text Retrieval	Same as 9	Entire text of the documents is searchable (except stop words)	Same as Model 9 with adjacency operators
11 Simple Thesaurus	Single terms	A set of one or more index terms	The request term is looked up in a thesaurus (online) and semantically related terms are added to the request term
12 Weighted Thesaurus	Single terms	A set of one or more index terms	The request term is looked up in a thesaurus (online) and semantically related terms above a given cut-off value (weight) are added (disjunctively) to the request term. The cut-off value could be given by the inquirer.

Retrieval rules listed in Table 2.1 can be categorized under three broad groups: 1) Exact matches between query term(s) and index terms, along with Boolean retrieval rules (Models 1-4, 9-12); 2) probabilistic retrieval rules (Models 5-7); and 3) vector space model (Model 8).

In group 1, indexing and query terms are binary: i.e., a term is either assigned to a document (or included in a search query) or not. Each term is equally important for retrieval purposes. Cut-off values can be introduced for multi-term search requests (Models 3 and 4). Search terms can be expanded by adding related terms from a thesaurus (Models 11 and 12). Retrieved records can be weakly ordered (retrieved or not) (Models 1-3, 12). Or they can be ranked on the basis of the number of matching terms in the search query and index record (Model 4). Relationships between search terms can be defined using Boolean logic (e.g., retrieve only those documents whose index records contain both search terms *A* and *B*) (Models 9 and 10). The boolean search model is believed to be "the most popular retrieval design for computerized document retrieval systems" (Blair, 1990, p.44).

Retrieval rules under group 2 call for weighted search terms (Model 5), weighted index terms (Model 6), or both weighted search and index terms (Model 7). In other words, the significance of a given term for retrieval purposes can be specified by the user. Retrieved records are ranked on the basis of the strength of the match between search and index terms. Retrieval rules in this category are known as probabilistic retrieval models.

The vector space model (Model 9) in group 3 is, in a way, similar to Model 7 in that both search and index terms are weighted and the retrieved records are ranked.

However, search and index terms in vector space model are treated as vectors in an *n*-dimensional space and the strength of match (e.g., ranking) is determined by calculating the cosine of the angle between search and index vectors. Document retrieval systems utilizing vector space model, notably SMART, have been in use since the early 1960s.

So far the major components of a conventional document retrieval system are reviewed from the following points of view: the document database, query formulation, and retrieval rules. The ultimate objective of a document retrieval system, regardless of which retrieval rule is used, is to retrieve records that best match the user's information needs. Hence, what matters to the user most is the retrieval results (i.e., retrieval effectiveness). The primary measures of retrieval effectiveness are reviewed below.

2.8 Measures of Retrieval Effectiveness

Several different measures are used to evaluate the retrieval effectiveness of document retrieval systems. A few measures that are widely used in the study of search failures such as *precision* and *recall* are discussed below. Other retrieval effectiveness measures suggested in the literature are not reviewed here as they are seldom, if ever, used in the analysis of search failures.

Online document retrieval systems often retrieve some non-relevant documents while missing, at the same time, some relevant ones. Blair (1990) summarizes the retrieval process as follows:

> Because information retrieval is essentially a trial and error process, almost any search for documents on an information retrieval system can

> be expected to retrieve not only useful (or relevant) documents, but also a varying proportion of useless (non-relevant) documents. This uncertainty in the searching process has another consequence: even when useful documents are retrieved from a data base, more useful documents may remain unretrieved despite the inquirer's most persistent efforts. As a result, after any given search the documents in the database can be classified in any four different ways:
>
> Retrieved and relevant (useful)
> Retrieved and not relevant (useless)
> Not retrieved and relevant [missed]
> Not retrieved and not relevant (p.73-74).

He provides a figure representing these four classes of documents:

FIGURE 2.2 A REPRESENTATION OF THE OUTPUT
(SOURCE: BLAIR (1990, P.76).)

	RELEVANT	NOT RELEVANT	
RETRIEVED	x	u	TOTAL NUMBER RETRIEVED=n_1
NOT RETRIEVED	v	y	
	TOTAL NUMBER RELEVANT=n_2		

Based on the above figure, the following retrieval effectiveness measures can be defined:

$$Precision=\frac{x}{n_1}$$

$$Recall=\frac{x}{n_2}$$

$$Fallout=\frac{u}{u+y}$$

where

x = number of relevant documents retrieved,

n_1 = number of documents retrieved ($x+u$ in Fig. 2.2),

n_2 = total number of relevant documents in the collection ($x+v$ in Fig. 2.2),

u = number of non-relevant documents retrieved,

y = number of non-relevant documents not retrieved.

Precision and *recall* are generally used in tandem in evaluating retrieval effectiveness in document retrieval systems. "*Precision* is the ratio of the number of relevant documents retrieved to the total number of documents retrieved" (Van Rijsbergen, 1979, p.10, original emphasis). For instance, if, for a particular search query, the system retrieves two documents (n_1) and the user finds one of them relevant (x), then the precision ratio for this search would be 50% (x/n_1).

Recall is considerably more difficult to calculate than precision because it requires finding relevant documents that will not be retrieved during users' initial searches (Blair & Maron, 1985, p.291). "*Recall* is the ratio of the number of relevant documents retrieved to the total number of relevant documents (both retrieved and not retrieved)" in the collection (Van Rijsbergen, 1979, p.10, original emphasis). Take the above example. The user judged one of the two retrieved documents to be relevant. Suppose that later three more relevant documents (v) that the original search query failed to retrieve were found in the collection. The system retrieved only one (x) out of the four (n_2) relevant documents from the database. The recall ratio would then be equal to 25% for this particular search (x/n_2).

Blair and Maron (1985) point out that "Recall measures how well a system retrieves *all* the relevant documents, and Precision, how well the system retrieves *only* the relevant documents" (p.290).

Fallout is another measure of retrieval effectiveness. *Fallout* can be defined as the ratio of nonrelevant documents retrieved (*u*) over all the nonrelevant documents in the collection (*u*+*y*). Fallout "measures how well a system rejects non-relevant documents" (Blair, 1990, p.116). The earlier example also can be used to illustrate fallout. The user judged one of the two retrieved documents as relevant, and, later, three more relevant documents that the original query missed were identified. Further suppose that there are nine documents in the collection altogether (four relevant plus five non-relevant documents). Since the user retrieved one non-relevant (*u*) document out of a total of five non-relevant ones (*u*+*y*) in the collection, the fallout ratio would be 20% for this search (*u/(u+y)*).

2.9 Relevance Feedback Concepts

It was mentioned earlier (section 2.1) that a document retrieval system should have some kind of user interface which allows users to interact with the system. Furthermore, the functions of a user interface were given (section 2.6) and it was stated that one of the functions of the user interface is to make various forms of feedback possible between the user and the document retrieval system.

As users scarcely find what they want in a single try, the feedback function deserves further explication. Retrieval rules, in and of themselves do not guarantee that retrieved records will be of importance to the user. The user interface may prompt users as to what to do next or suggest alternative strategies by way of system-generated feedback messages (i.e., help screens, status of search, actions to take).

More importantly, the system may allow users to modify their search queries in light of a sample retrieval so that search success can be improved in subsequent retrieval runs (Van Rijsbergen, 1979). Some systems may automatically modify the original search query after the user has made relevance judgments on the documents which were retrieved in the first try. This is known as "relevance feedback" and it is the relevance feedback process that concerns us here.

Swanson (1977) examined some well-known information retrieval experiments and the measures used therein. He suggested that the design of document retrieval systems "should facilitate the trial-and-error process itself, as a means of enhancing the correctability of the request" (p.142).

Van Rijsbergen (1979) shared the same view when he pointed out that: "a user confronted with an automatic retrieval system is unlikely to be able to express his information need in one go. He is more likely to want to indulge in a trial-and-error process in which he formulates his query in the light of what the system can tell him about his query" (p.105).

Van Rijsbergen (1979) also lists the kind of information that could be of help to users when reformulating their queries such as the occurrence of users' search terms in the database, the number of documents likely to be retrieved by a particular query with a small sample, and alternative and related search terms that can be used for more effective search results.

Relevance feedback is one of the tools that facilitates the trial-and-error process by allowing the user to interactively modify his or her query based on search results obtained during the initial run. The following quotation summarizes the relevance feedback process very well:

> It is well known that the original query formulation process is not transparent to most information system users. In particular, without detailed knowledge of the collection make-up, and of the retrieval environment, most users find it difficult to formulate information queries that are well designed for retrieval purposes. This suggests that the first retrieval operation should be conducted with a tentative, initial query formulation, and should be treated as a trial run only, designed to retrieve a few useful items from a given collection. These initially retrieved items could then be examined for relevance, and new improved query formulations could be constructed in the hope of retrieving additional useful items during subsequent search operations (Salton & Buckley, 1990, p.288).

Relevance feedback was first introduced over 20 years ago during the SMART information retrieval experiments (Salton, 1971b). Earlier relevance feedback experiments were performed on small collections (e.g., 200 documents) where the retrieval performance was unusually high (Rocchio, 1971a; Salton, 1971a; Ide, 1971). (For the use of relevance feedback technique in online catalogs, see, for instance, Porter, 1988; Walker, S. & de Gere, 1990; Larson, 1989, 1991a; Walker, S. & Hancock-Beaulieu, 1991.)

It was shown that relevance feedback markedly improved retrieval performance. Recently Salton and Buckley (1990) examined and evaluated twelve different feedback methods "by using six document collections in various subject areas for experimental purposes." The collection sizes they used varied from 1,400 to 12,600 documents. The relevance feedback methods produced improvements in retrieval performance ranging from 47% to 160%.

The relevance feedback process offers the following main advantages:

1. It shields the user from the details of the query formulation process, and permits the construction of useful search statements without intimate knowledge of collection make-up and search environment.

2. It breaks down the search operation into a sequence of small search steps, designed to approach the wanted subject area gradually.

3. It provides a controlled query alteration process designed to emphasize some terms and to deemphasize the others, as required in particular search environments (Salton & Buckley, 1990, p.288).

The relevance feedback process helps in refining the original query and finding more relevant materials in the subsequent runs. The true advantage gained through the relevance feedback process can be measured in two different ways:

1) By changing the ranking of documents and moving the documents that are judged by the user as being relevant up in the ranking. With this method documents that have already been seen (and judged as being relevant) by the user will still be retrieved in the second try, although they are somewhat ranked higher this time. "This occurs because the feedback query has been constructed so as to resemble the previously obtained relevant items" (Salton & Buckley, 1990, p.292). This effect is called "ranking effect" (Ide, 1971) and it is difficult to distinguish artificial ranking effect from the true feedback effect (Salton & Buckley, 1990). Note that users may not want to see the documents a second time because they have already seen them during the initial retrieval.

2) By eliminating the documents that have already been seen by the user in the first retrieval and "freezing" the document collection at this point for the second retrieval. In other words, documents that were judged as being relevant (or nonrelevant) during the initial retrieval will be excluded in the second retrieval, and

the search will be repeated only on the frozen part of the collection (i.e., the rest of the collection from which user has seen no documents yet). This is called "residual collection" method and it ". . . depresses the absolute performance level in terms of recall and precision, but maintains a correct relative difference between initial and feedback runs" (Salton & Buckley, 1990, p.292).

The different relevance feedback formulae are based on the variations of these two methods. More detailed information on relevance feedback formulae can be found in Salton and Buckley (1990). For mathematical explications of relevance feedback process, see Rocchio (1971a), Ide (1971), and, Salton and Buckley (1990).

The relevance feedback process works in practice as follows: a user submits a search query to the system with relevance feedback capabilities and retrieves some documents. When bibliographic records of retrieved documents are displayed one by one to the user, he or she is asked to judge each retrieved document as being relevant or nonrelevant. The user proceeds by making relevance judgments for each displayed record. These relevance judgments will be used to improve the search results should the user decide to perform a relevance feedback search. The system revises and modifies the original query based on the documents judged as being relevant during the first retrieval. In other words, the relevance feedback process enables the system to."understand" the user's query better: the documents that are similar to the query are rewarded by being assigned higher ranks, while dissimilar documents are pushed farther down in the ranking. As a result, the system comes up with potentially more relevant documents.

The relevance feedback search can be iterated as many times as the user desires until the user is satisfied with the search results. However, the relevance feedback technique requires more work for the user who is known to be willing to invest minimal effort only.

2.10 Summary

The major components of a document retrieval system are examined in this chapter. The importance of indexing and query formulation processes are discussed along with the roles of user interfaces and retrieval rules. Some of the more advanced information retrieval techniques such as relevance feedback and clustering are also briefly addressed. A critical review of the major studies related to the present study is given in Chapter III.

CHAPTER III

FAILURE ANALYSIS IN DOCUMENT RETRIEVAL SYSTEMS: A CRITICAL REVIEW OF STUDIES

3.0 Introduction

In Chapter II an overview of a document retrieval system was given along with the definitions of retrieval effectiveness measures such as precision, recall, and fallout. Relevance feedback and classification clustering techniques were briefly explained. The uses of such techniques in enhancing the effectiveness of document retrieval systems were discussed.

This chapter will examine the concepts of failure analysis in document retrieval systems and review the literature on failure analysis studies. A critical review of various methods of analyzing search failures is given in Section 3.2. A brief overview of major studies of search failures in document retrieval systems is given in Section 3.3.

3.1 Analysis of Search Failures

Online document retrieval systems often fail to retrieve some relevant documents. More often than not they also retrieve non-relevant documents. Such search failures may occur due to a variety of reasons, including problems with user-system interfaces, retrieval rules, and indexing languages.

Studying search failures presents extremely complicated problems. For instance, it is not clear exactly what constitutes a "search failure." While some researchers study search failures using retrieval effectiveness measures such as precision and recall, others prefer using "user satisfaction" as a criterion in deciding whether a search has failed or not. Before reviewing major failure analysis studies, it is helpful to examine some approaches used in studying search failures in document retrieval systems and to discuss the various (mostly implied) definitions of "search failure" used by researchers.

3.2 Methods of Analyzing Search Failures

This section discusses the analysis of search failures using retrieval effectiveness methods (e.g., precision and recall), user satisfaction measures, transaction logs, and the critical incident technique.

3.2.1 Analysis of Search Failures Utilizing Retrieval Effectiveness Measures

A detailed discussion of retrieval effectiveness measures such as precision and recall was given in Chapter II. As pointed out earlier, precision is defined as the proportion of retrieved documents which are relevant, whereas recall is defined as the proportion of relevant documents retrieved (Van Rijsbergen, 1979, p.10).

If precision and recall are seen as performance measures with the given definitions, it becomes clear that "performance" can no longer be defined as a dichotomous concept. When precision and recall are defined as percentages, we can think of "degrees" of search failure or success. This view best reflects different performance levels attained by current document retrieval systems. It is impossible to find a perfect document retrieval system. In reality, retrieval systems are imperfect, and one is better or worse than another.

Performance measures such as precision and recall can be used in the analysis of search failures. In the precision example of a calculation of a precision value in Section 2.8 of Chapter II, only 50% of the documents retrieved were relevant, resulting in a precision of 50%. If each nonrelevant document that the system retrieves for a given query represents a search failure, then it is also possible to think of precision as a measure of search failure: failure to retrieve relevant documents *only*. The more nonrelevant documents the system retrieves for a given query, the higher the degree of precision failures. If no retrieved document happens to be relevant, then the precision ratio becomes zero due to severe precision failures.

In the recall example, the recall ratio was 25%, implying that the system missed 75% of the relevant documents in the collection. If each missed relevant document represents a search failure, then it is possible to think of recall as a measure of search failure: failure to retrieve *all* relevant documents in the collection. The more relevant documents the system misses, the higher the degree of recall failure. If the system fails to retrieve any relevant documents from the collection, then the recall ratio becomes zero due to severe recall failures.

Precision and recall are two different quantitative measures of aggregation of search failures. For convenience, search failures analyzed using precision and recall are called precision failures and recall failures.

Precision failures can easily be detected. They occur when the user finds some retrieved documents nonrelevant, even if those documents are assigned the index terms that the user initially asked for in the search query. Users may feel that index terms have been incorrectly assigned to documents that are not really relevant to those subjects.

Note that "relevance" is defined as a relationship ". . . between a document and a person in search of information" and it is a function of a large number of variables concerning both the document (e.g., what it is about, its currency, language, and date) and the person (e.g, person's education and beliefs) (Robertson *et al.*, 1982, p.1).[1]

Recall failures mainly occur because index terms that users would normally utilize to retrieve documents about particular subjects do not get assigned to documents that are relevant to those subjects. As stated earlier, detecting recall failures, especially in large scale document retrieval systems, is much more difficult. Researchers have therefore used somewhat different approximations to calculate recall figures in their experiments.

Although information retrieval textbooks mention "fallout" as a measure of retrieval effectiveness, we are not aware of any experiment where fallout ratio has been successfully calculated.[2] Fallout is the proportion of nonrelevant documents retrieved over all the nonrelevant documents in the collection. Calculating the fallout ratio in large collections is as difficult, if not more difficult, as calculating the recall ratio. To calculate the fallout ratio, all nonrelevant documents retrieved during the search must be identified, all nonrelevant documents in the overall collection must be found, and the size of the collection must be established.

[1] For a comprehensive review of the concept of "relevance," see Saracevic (1975), Schamber *et al.* (1990), and Eisenberg & Schamber (1988).

[2] An attempt has been made in Cranfield II to plot recall/fallout graphs. The size of the collection used in this experiment was relatively small (1,400 documents) and many tests were done with 200 documents. Nevertheless, no analysis has been performed to find out the causes of fallout failures. For details, see: Cleverdon *et al.*, (1966), and Cleverdon & Keen (1966).

It is tempting to say that documents that are not retrieved are probably not relevant; however, since recall failures do occur in document retrieval systems, this is not the case. If all of the unretrieved documents in a collection were scanned, some of them would be relevant. The fallout ratio could then be calculated. This method can only be used for specific queries where the number of relevant documents in the whole collection is known to be small.

"Fallout failures" do occur constantly in document retrieval systems even if it is impractical to quantify them. Whenever the system retrieves too many nonrelevant records, users feel the consequences of fallout failure. Either they must scan long lists of useless records (hence "fallout") or abandon the search.

Fallout failures also can be seen as severe precision failures. Fallout failure has not been adequately studied; however, it is known that users tend to resist scanning through screens of retrieved items. For instance, Larson (1991c, p.188) found that in a large online catalog the average number of records retrieved was 77.5, but users scanned an average of less than 10 records per search. (See also: Wiberley & Dougherty, 1988.) It is not clear why the users stopped scanning after a few records. Some may have been satisfied with the results. Some users might have abandoned their searches due to frustration because the system retrieved too many unpromising, nonrelevant records.[3] It would be interesting to study what percentage of searches in online catalogs get abandoned in view of user frustration from fallout failures.

[3] J. L. Kuhns implied that frustration usually occurs when a user reaches his or her "futility point" in a given search. The futility point is defined as "the number of retrieved documents the inquirer is willing to browse through before giving up his search in frustration" (Kuhns, 1963; cited in Blair, 1980, p.271).

Mainly, then, retrieval effectiveness measures are used to determine and study three types of search failures: (1) retrieving nonrelevant documents (precision failures); (2) missing relevant documents (recall failures); and (3) retrieving too many unpromising, nonrelevant documents (fallout failures). Failure analysis aims to find out the causes of these failures so that existing systems can be improved in a variety of ways.

So far, we have looked at a few of the measures of retrieval effectiveness and the ways in which they are used in the study of search failures. We noted that document retrieval systems are not perfect and that we cannot expect them to achieve, or even approximate, the impossible ideal of retrieving *all* and *only* relevant documents in the collection. Furthermore, users would like to find some relevant documents, but not necessarily *all* of them, unless (as in rare occasions such as patent searching) *all* are wanted. They prefer high precision to high recall. They wish to retrieve "some good references without having to examine too many bad ones" (Wages, 1989, p.80). Consequently, it is more important for a document retrieval system to "distinguish between wanted and unwanted items" quickly than to retrieve all relevant items in the collection.

Not everyone is satisfied with the most commonly used retrieval effectiveness measures (precision and recall), however. For instance, William Cooper has questioned the use of recall as a performance measure because it takes into account not only retrieved documents, but also unretrieved documents. In his view, this is wasted effort since the relevance of unretrieved documents has little bearing on the notion of subjective user satisfaction (Cooper, W., 1973; *cf.* Soergel, 1976). He maintains that "an ideal evaluation methodology must somehow measure the ultimate worth of a retrieval system to its users in terms of an appropriate unit of utility" (Cooper, W., 1973, p.88).

3.2.2 Analysis of Search Failures Utilizing User Satisfaction Measures

Some failure analysis studies are based on user satisfaction measures, rather than on retrieval effectiveness measures. Although it may at first seem straightforward, analyzing search failures utilizing user satisfaction measures is a complex process that provides interesting challenges.

First, defining user satisfaction is difficult. Several authors tried to address this issue. Tessier *et al.* (1977) discussed such factors as the search output, the intermediary, the service policies, and the "library as a whole" as the main determinants of the user satisfaction. Bates (1972, 1977a, 1977b) examined the effects of "subject familiarity" and "catalog familiarity" on search success and found that the former has a slight detrimental effect, while the latter has a very significant beneficial effect on search success. Tessier (1981) used factor analysis and multiple regression techniques to study the influence of various variables on overall search satisfaction. She found that "the strongest predictors of satisfaction were the precision of search, the amount of time saved, and the perceived quality of the database as a source of information" (Tessier, 1981; cited in Kinnucan, 1992, p.73). Hilchey and Hurych (1985, p.455) found "a strong positive relationship between perceived relevance of citations and search value" when they performed a statistical analysis on the online reference questionnaire forms returned by the users in a university library.

Second, user satisfaction relies heavily on users' judgments about search failures or successes; however, users' judgments may be inconsistent for various reasons. For example, Tagliacozzo (1977) found that "MEDLINE was perceived as 'helpful' by respondents who, in other parts of the questionnaire [used in the author's research], showed that they had *not* found it particularly useful" (p. 248, original emphasis). Tagliacozzo warns us that ". . . caution should therefore be used in taking the users' judgments at face value, and in inferring from single responses that their information needs were, or were not, satisfied by the service."

It follows that it is not usually sufficient to obtain a binary "Yes/No" response from the user about being satisfied or not satisfied with the results. Ankeny (1991, p.356) found that the use of a two-point (yes-no) scale ". . . appeared to result in inflated success ratings." When pressed, users are likely to come up with further explanations. For example, a user might say: "Yes, in a way my search was successful even though I couldn't find what I wanted." A second user might say that a given search was not successful because "it did not retrieve anything new."

A researcher getting such answers would have hard time classifying them. The data gathering tools that the researcher employs to elicit information from users should be sensitive enough to handle such answers by asking more detailed questions. After all, a decision has to be made if a search was successful or not. Further conditions have been introduced in some studies to facilitate this decision-making process. In Ankeny's study, for example, a successful search has three characteristics:

> the patron must indicate that s/he found *exactly* what was wanted, that s/he was *fully* satisfied with the search, and that s/he marked none of the 10 listed reasons for dissatisfaction where the reasons for dissatisfaction ranged from 'system problems' to 'too much information,' from 'information not relevant enough' to 'need different viewpoint' (Ankeny, 1991, p.354, original emphasis; see also Auster &

Lawton, 1984).

Nevertheless, it is still possible that a given search may be a failure even if answers given by a user met all three of these conditions. It was noted earlier that users tend to abandon some searches that retrieve too many items. Many users may prefer to retrieve a few relevant documents quickly. They would not be considered a search "failure" even if the system has missed some relevant documents (i.e., recall failure).

User satisfaction measures are influenced by both the type of user and search goal factors. For example, an undergraduate student writing a term paper may be satisfied if a search retrieves a few relevant textbooks. However, the situation is quite different for a health professional. This user may want to know everything about a certain case because the outcome of missing relevant information may have serious consequences. For example, in a search a health professional investigating a medical procedure using the MEDLINE database only found records showing the procedure to be safe, and did not find records that indicated fatalities associated with the procedure (Wilson *et al.*, 1989).[4]

The above examples show that some caution is needed when interpreting users' indication of satisfaction. There are some published studies that show that "in many cases high levels of reported end-user 'satisfaction' . . . may not reflect true success rates" (Ankeny, 1991, p.356). Furthermore, as Cheney (1991, p.155) notes, we do not "know what end users expect of their search results, because no study has

[4] For hypothetical examples as to the importance of unretrieved but relevant documents, see Soergel (1976).

examined end users' expectations of database searching. Neither has any study examined the actual quality of end-user search results measured in terms of precision and recall."

So far, the discussion has concentrated on the analysis of search failures that were based on retrieval effectiveness or "user satisfaction." As part of a carefully designed and conducted experiment under "as real-life a situation as possible," Saracevic and Kantor (1988) studied, among other things, the relationship between user satisfaction and precision and recall.

Their experiment involved 40 users who each submitted a query that reflected a real information need. Thirty-nine professional searchers did online searches on Dialog databases for these queries. Each query was searched by nine different professionals and the results were combined for evaluation purposes. The precision ratio for a given search was estimated as the number of relevant items retrieved by the search divided by the total number of items retrieved by the search. Similarly, the recall ratio was estimated as the number of relevant items retrieved by the search divided by the total number of relevant items in the union of items retrieved by all searchers for that question (Saracevic *et al.*, 1988).[5] Five utility measures were used: (1) whether the user's participation and the resultant information was worth it (on a five-point scale); (2) time spent; (3) perceived (by the users) dollar value of the items; (4) whether the information contributed to the resolution of the research problem (on a five-point scale); and (5) whether the user was satisfied with the results

[5] Note that it is not discussed in this paper how they calculated the precision/recall ratios and what figures (i.e., number of records (a) retrieved, (b) relevant, (c) not relevant) they obtained. As they stressed several times in their report, the recall figures they obtained were not absolute but comparative. For a more detailed account, see Part II of their article (Saracevic & Kantor, 1988).

(on a five-point scale).

They found that "searchers in questions where users indicated high overall satisfaction with results . . . were 2.49 times more likely to have higher precision" (Saracevic & Kantor, 1988, p.193). They interpreted their findings pertaining to the relationship between utility measures and retrieval effectiveness measures as follows:

> In general, retrieved sets with high precision increased the chance that users assessed that the results were 'worth more of their time than it took,' were 'high in dollar value,' contributed 'considerably to their problem resolution,' and 'were highly satisfactory.' On the other hand, high recall did not significantly affect the odds for any of those measures. . . . These are interesting findings in another respect. They indicate that utility of results (or user satisfaction) may be associated with high precision, while recall does not play a role that is even closely as significant. For users, precision seems to be the king and they indicated so in the type of searches desired. In a way this points out to the elusive nature of recall: this measure is based on the assumption that something may be missing. Users cannot tell what is missing any more than searchers or systems can. However, users can certainly tell what is in their hand, and how much is *not* relevant (Saracevic & Kantor, 1988, p.193, original emphasis).

3.2.3 Analysis of Search Failures Utilizing Transaction Logs

The availability of transaction logs, which record users' interaction with the document retrieval systems, provides the opportunity to study and monitor search failures unobtrusively (Tolle, 1983a, 1983b; Borgman, 1983; Simpson, 1989). Larson (1991b, p.198) states: "Transaction monitoring, in its simplest form, involves the recording of user interactions with an online system. More complete transaction monitoring also will 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." This includes search queries entered, records displayed, help requests, errors, and the system responses.[6]

Since transaction logs also contain invaluable information about failed searches, researchers have been interested in scanning transaction logs in order to identify failed searches. Several researchers identified "zero hits" from the transaction logs of selected online catalogs and looked into the reasons for search failures (see, for instance, Dickson, 1984; Peters, 1989; Hunter, 1991; Zink, 1991; Cherry, 1992). A few others employed the same method when they studied search failures in MEDLINE (Kirby & Miller, 1986; Walker, C.J. *et al.*, 1991). These researchers used a rather practical definition of search failure when scanning transaction logs. A search was treated as a failure if it retrieved no records.

Needless to say, the definition of search failure as zero hits is incomplete since it does not include partial search failures. More importantly, there is no reason to believe that all "non-zero hits" searches were successful ones. Such an assumption would mean that no precision failures occurred in the systems under investigation! Furthermore, "not all zero hits represent failures for the patrons . . . It is possible that the patron is satisfied knowing that the information sought is not in the database, in which case the zero-hit search is successful" (Hunter, 1991, p.401). Precedence searching in litigation is an example of a zero-hit search that is successful.

Some newer document retrieval systems such as Okapi and CHESHIRE can accommodate relevance feedback techniques and incorporate users' relevance judgments in order to improve retrieval effectiveness in subsequent iterations (Walker, S. & Hancock-Beaulieu, 1991; Larson, 1991a). Transaction logs of such online

[6] For a review of online catalog transaction log studies, see Simpson (1989).

catalogs also record the user's relevance judgment for each record that is displayed. Using these logs, the researcher is able to determine whether the user found a given record to be relevant or not.

The availability of relevance judgments in transaction logs has opened up new avenues for studying search failures in online library catalogs. Researchers are now able to study not only zero-hit searches, but also failed searches that retrieve nonrevelant records. Obviously, the rendering of relevance judgments makes it easier to identify precision failures, but there still needs to be some kind of mechanism to identify recall failures.

What constitutes a search failure when the relevance judgment for each retrieved document is recorded in the transaction log? Walker, S. and Jones (1987) introduced yet another practical definition of search failure during the analysis and evaluation of an experimental online catalog (Okapi) where they recorded users' relevance judgments in transaction logs. They considered a search query as a failure "if no relevant record appears in the first ten which are displayed" (Walker, S. & Jones, 1987; see also Jones, 1986). This definition of search failure is quite different from one based on precision and recall. It is dichotomous, and it assumes that users will scan at least ten records before quitting. This assumption might be true for some searches and for some users, but not for all searches and users. It also downplays the importance of search failures. Searches retrieving at least one relevant record in ten are considered "successful" even though the precision rate for such searches is quite low (10%).

Although transaction monitoring offers unprecedented opportunities to study search failures in document retrieval systems and provides "highly detailed information about how users actually interact with an online system, . . . it cannot reveal their intentions or whether they are satisfied with the results" (Larson, 1991, p.198).

Some of the shortcomings of transaction monitoring in studying search failures are as follows.

First, it is not clear what constitutes a "search failure" in transaction logs. As mentioned earlier, defining all zero-hit searches as search failures has some serious flaws.

Second, transaction logs have very little to offer when studying recall failures in document retrieval systems. Recall failures can only be determined by using different methods such as analysis of search statements, indexing records, and retrieved documents. In addition, additional relevant documents that were not retrieved in the first place can be found by performing successive searches in the database.

Third, transaction logs can document search failure occurrences, but they cannot explain why a particular failure occurred. Search failures in online catalogs occur for a variety of reasons, including simple typographical errors, mismatches between users' search terms and the vocabulary used in the catalog, collection failures (i.e., requested item is not in the system), user interface problems, and the way search and retrieval algorithms function. Further information is needed about users' needs and intentions in order to find out why a particular search failed.

Finally, since the users usually remain anonymous in transaction logs, analysis of these logs "prevents correlation of results with user characteristics" (Seymour, 1991, p.97).

3.2.4 Analysis of Search Failures Utilizing the Critical Incident Technique

Based on their empirical investigation of tools, techniques, and methods for the evaluation of online catalogs, Hancock-Beaulieu *et al.* (1991, p.532) found that "transaction logs can only be used as an effective evaluative method with the support of other means of eliciting information from users." One of the techniques to elicit information from users about their needs and intentions is known as the "critical incident technique." Data gathered through this technique, which is briefly discussed below, facilitates the study of search failures in document retrieval systems. When it is used in conjunction with the analysis of transaction log data, the critical incident technique permits search failures to be correlated with user characteristics.

The critical incident technique was first used during World War II to analyze the reasons that pilot candidates failed to learn to fly. Since then, this technique has been widely used, not only in aviation, but also in defining the critical requirements of and measuring typical performance in the health professions. Flanagan (1954, p.327) describes it as follows:

> The critical incident technique consists of a set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems and developing broad psychological principles. The critical incident technique outlines procedures for collecting observed incidents having special significance and meeting systematically defined criteria.
>
> By an incident is meant any observable human activity that is sufficiently complete in itself to permit inferences and predictions to be

made about the person performing the act.

The major advantage of this technique is to obtain "a record of specific behaviors from those in the best position to make the necessary observations and evaluations" (Flanagan, 1954, p.355). In other words, it is observed behavior that counts in critical incident technique, not opinions, hunches and estimates.

The critical incident technique consists of two steps: (1) collecting and classifying detailed incident reports, and (2) making inferences that are based on the observed incidents. Wilson *et al.* (1989, p.2) summarize these two steps as follows:

> The collection and careful analysis of a sufficient number of detailed reports of such observations of effective and ineffective behaviors results in comprehensive definition of the behaviors that are required for success in the activity in question under a wide range of conditions. These organized lists of critical requirements (generally termed performance 'taxonomies') can then be used for a variety of practical purposes such as the evaluation of performance, the selection of individuals with the greatest likelihood of success in the activity, or the development of training programs or other aids to increase the effectiveness of individuals.

The critical incident technique can also be used to gather data "on observations previously made which are reported from memory." Flanagan (1954) claims that collecting data about incidents which happened in the recent past is usually satisfactory. However, the accuracy of reporting depends on what the incident reports contain: the more detailed and precise the incident reports are the more accurate, it is assumed, the information contained therein.

Recently, the critical incident technique has been used to assess "the effectiveness of the retrieval and use of biomedical information by health professionals" (Wilson *et al.*, 1989, p.2). In the same study, researchers have used this technique to analyze and evaluate search failures in MEDLINE. Using a structured interview process that included administering a questionnaire, they asked users to comment on the effectiveness of online searches that they performed on the MEDLINE database. Each report obtained through structured interviews was called an "incident report." Researchers matched these incident reports against MEDLINE transaction log records corresponding to each search in order to find out the actual reasons for search success or failure. These incident reports provided much sought after information about user needs and intentions, and they put each transaction log record in context by linking search data to the searcher.

Although the critical incident technique enables the researcher to gather information about user needs and intentions so that he or she can better explain the causes of search failures, it also has some shortcomings. Information gathered through the critical incident technique has to be corroborated with transaction log data. The verification of user satisfaction or dissatisfaction via transaction log data may provide further clues as to why searches succeed or fail. However, the researcher may not be able to confirm each and every user's account of his or her search from the transaction logs. As the users are generally not identified in the transaction logs, it is sometimes difficult to find the search in question in the logs.

There are a variety of reasons for this problem. First, the user's permission has to be sought in advance in order to examine his or her search(es) in the transaction logs. Second, users may not be able to recall the details of their searches after the fact. Third, the logs may not contain enough data about the search: the items displayed and users' relevance judgments are not recorded in most transaction logs.

The lack of enough data in transaction logs also influences the effectiveness of the critical incident technique. The researcher has to rely a great deal on what the user says about the search. For instance, if the items displayed by the user along with relevance judgments are not recorded in the transaction logs, the researcher will not be able to find the precision ratio. Furthermore, the critical incident technique per se does not tell us much about the documents that the user may have missed during the search: we still have to find out about recall failures using other methods.

3.2.5 Summary

This section discussed various methods of analyzing search failures in document retrieval systems. It emphasized that the issue of search failure is complex. It demonstrated that no single method of analysis is self-sufficient to characterize all the causes of search failures. The next section will review the findings of major studies in this area.

3.3 Review of Studies Analyzing Search Failures

Numerous studies have shown that users experience a variety of problems when they search document retrieval systems and they often fail to retrieve relevant documents. The problems users frequently encounter when searching, especially in online catalogs, are well documented in the literature (Alzofon & Van Pulis, 1984; Bates, 1986; Blazek & Bilal, 1988; Borgman, 1986; Cochrane & Markey, 1983; Gouke & Pease, 1982; Hartley, 1988; Henty, 1986; Hildreth, 1982, 1985, 1989; Janosky *et al*, 1986; Kaske, 1983; Kern-Simirenko, 1983; Kinsella & Bryant, 1987; Larson, 1986, 1991c; Lawrence *et al.*, 1984; Markey, 1980, 1984, 1985, 1986; Matthews, 1982; Matthews *et al.*, 1983; Mitev *et al.*, 1985; Nielsen, 1986; Wang, 1985). However, few researchers have studied search failures directly (Cleverdon, 1962; Cleverdon *et* *al.*, 1966; Cleverdon & Keen, 1966; Lancaster, 1968, 1969; Dickson, 1984; Blair & Maron, 1985; Jones, 1986; Markey & Demeyer, 1986; Walker, S. & Jones, 1987; Wilson *et al.*, 1989; Klugman, 1989; Peters, 1989; Ankeny, 1991; Hunter, 1991; Walker, C.J. *et al.*, 1991; Zink, 1991; Cherry, 1992). What follows is a brief overview of major studies of search failures in document retrieval systems. Not surprisingly, the results of these studies are not directly comparable because they use different definitions and methods of analysis.

3.3.1 Studies Utilizing Precision and Recall Measures

Several major studies have employed precision and recall measures to analyze search failures.

3.3.1.1 The Cranfield Studies

Cyril Cleverdon, who was Librarian of the College of Aeronautics at Cranfield, England, and his colleagues conducted a series of studies in late 1950s and early 1960s to investigate the performance of indexing systems (Cleverdon, 1962; Cleverdon *et al.*, 1966, and Cleverdon & Keen, 1966). They also studied the causes of search failures in document retrieval systems. Findings pertaining to search failures are reviewed here.

In the first study (Cranfield I), Cleverdon (1962, p.1) compared the efficiency of retrieval effectiveness of four indexing systems: the Universal Decimal Classification, an alphabetical subject index, a special facet classification, and the uniterm system of co-ordinate indexing. Some 18,000 research reports and periodical articles in the field of aeronautics were indexed using these four indexing systems, and 1,200 queries were used in the tests.

The main purpose of the Cranfield I experiment was to test the ability of each indexing system to retrieve the "source document" upon which each query was based. Researchers knew beforehand that "there was at least one document which would be relevant to each question" (pp.8-9). The recall ratio was calculated based on the retrieval of source documents. However, this recall ratio should be regarded as a type of "constrained" recall since the objective was just to find source documents in the collection. Cranfield I tests have shown that "the general working level of I.R. systems appears to be in the general area of 60%-90% recall and 10%-25% of relevance [i.e., precision]" (pp.8-9).[7]

During the tests, each search was "carried on to the stage where the source document was retrieved or alternatively the searcher was unable to devise any further reasonable search programmes" (p.11). Each query was judged to be a success or failure: a search was a success if the source document was retrieved, a failure if it was not. Swanson (1965, p.5) states: "The decision to measure retrieval success solely in terms of the source document was prompted by an understandable, though unfortunate, desire to determine whether any given document was or was not relevant to the question." Relevant documents other than source documents, which would have been retrieved during the search, were not taken into account.

The success rate for all searches was found as 78%;[8] source documents were successfully retrieved for most search queries.

[7] The design and findings of the Cranfield I experiment have been criticized by many authors. For example, see: Swanson (1965). For a review of the Cranfield tests, see: Sparck Jones (1981a).

[8] This percentage was obtained by averaging the figures given in the fifth column of Table 3.1 of Cleverdon (1962, p.22).

Cleverdon's analysis of search failures was based on 329 documents and queries. The total number of search failures was 495.[9] He classified the causes of search failures under four main headings: (1) question, (2) indexing, (3) searching, and (4) system. Each heading included further subdivisions to specify the exact cause(s) of each search failure. For example, questions could be "too detailed," "too general," "misleading" or just plain "incorrect." Likewise, insufficient, incorrect, or careless indexing; insufficient number of entries; and lack of cross references caused further search failures. Included under searching were "lack of understanding," "failure to use all concepts," "failure to search systematically," and "incorrect" or "insufficient searching." The lack of some features in indexing systems, such as synonymity and inability to combine particular concepts, also caused search failures.

The number of failed searches under each subdivision is given in several tables. The reasons for failures in searches carried out by the project staff are as follows: questions, 17%; indexing process, 60%; searching 17%; and, indexing system, 6%. The percentages of failures in searches performed by the technical staff (i.e., the end-users) were somewhat higher for searching (37%).

It appears that well over half of the failures in this study were caused by the indexing process. Cleverdon (1962, p.88) summarizes the results of the analysis of search failures as follows:

> The analysis of failures . . . shows most decisively that the failures were, for more than all other reasons together, due to mistakes by the indexers or searchers, and that a third of the failures could have been

[9] This summary is based on Chapter 5 of Cleverdon's Cranfield I report (1962). The report also includes the complete summary of the analysis of search failures (Appendix 5A) and "some examples of the complete analysis of the individual documents" (Appendix 5B).

> avoided if the project staff had indexed consistently, as well as they were capable of doing. Put another way, this means that in every hundred documents, the indexers failed to index adequately five documents, the failure usually consisting of the omission of some particular concept.

The second study (Cranfield II) conducted by Cleverdon and his colleagues was an attempt to investigate the performance of indexing systems based on such factors as the exhaustivity of indexing and the level of specificity of the terms in the index language. The test collection consisted of some 1,400 research reports and periodical articles on the subject of aerodynamics and aircraft structures. Some 221 queries (all single theme queries) were obtained from the authors of selected published papers. However, most tests were based on 42 queries and 200 documents (Cleverdon *et al.*, 1966, and Cleverdon & Keen, 1966).

Precision and recall were used to determine the retrieval effectiveness of indexing systems. It is difficult to cite a single performance figure because the Cranfield II experiment involved a number of different index languages with a large number of variables. It was found that there exists an inverse relationship between recall and precision and that "the two factors which appear most likely to affect performance are the level of exhaustivity of indexing and the level of specificity of the terms in the index language" (Cleverdon & Keen, 1966, p.i).[10] As noted in the preface to volume two of the report, a detailed intellectual analysis of the reasons for search failures was not carried out.

[10] For the detailed performance figures along with recall/precision graphs, see volume 2 of the full report (Cleverdon & Keen, 1966).

3.3.1.2 Lancaster's MEDLARS Studies

The Cranfield projects tested retrieval effectiveness in a laboratory setting, and the size of the test collection was small (1,400 documents). By contrast, Lancaster (1968), studied the retrieval effectiveness of a large biomedical reference retrieval system (MEDLARS) in operation. The MEDLARS database (Medical Literature Analysis and Retrieval System) contained some 700,000 records at that time. Some 300 "real life" queries were obtained from researchers and were used in the tests.

The retrieval effectiveness of the MEDLARS search service was measured using precision and recall. The precision ratio was calculated according to the definition given in Chapter 2. However, it would have been extremely difficult to calculate a true recall figure in a file of 700,000 records because this would have meant having the requester examine and judge each and every document in the collection. Lancaster explains how the recall figure was obtained:

> We therefore estimated the MEDLARS recall figure on the basis of retrieval performance in relation to a number of documents, judged relevant by the requester, *but found by means outside MEDLARS.* These documents could be, for example,
>
> 1. documents known to the requester at the time of his request,
>
> 2. documents found by his local librarian in non-NLM [National Library of Medicine] generated tools,
>
> 3. documents found by NLM in non-NLM-generated tools,
>
> 4. documents found by some other information center, or
>
> 5. documents known by authors of papers referred to

by the requester (pp.16, 19, original emphasis).

Relevant documents identified by the requester for each query made up the "recall base" upon which the calculation of the recall figure was based. An example illustrates how recall was calculated. The recall base consists of six documents that are known to the requester to be relevant before the search. Under these circumstances, if "only 4 are retrieved, we can say that the recall ratio for this search is 66%" (pp.19-20).

Based on the results of 299 test searches, Lancaster found that the MEDLARS Search Service was operating with an average performance of 58% recall and 50% precision.

Lancaster also studied the search failures using precision and recall. He investigated recall failures by finding some relevant documents using sources other than MEDLARS and then checking to see if the relevant documents had also been retrieved during the experiment. If some relevant documents were missed, this was considered as a recall failure and measured quantitatively. Precision failures were easier to detect since users were asked to judge the retrieved documents as being relevant or nonrelevant. If the user decided that some documents were nonrelevant, this was considered to be a precision failure and measured accordingly. However, identifying the causes of precision failures proved to be much more difficult because the user might have judged a document to be nonrelevant due to index, search, document, and other characteristics as well as the user's background and previous experience with the document.

To date, Lancaster's study is the most detailed account of the causes of search failures that has been attempted. As Lancaster (1969, p.123) points out:

> The 'hindsight' analysis of a search failure is the most challenging aspect of the evaluation process. It involves, for each 'failure,' an examination of the full text of the document; the indexing record for this document (i.e., the index terms assigned . . .); the request statement; the search formulation upon which the search was conducted; the requester's completed assessment forms, particularly the reasons for articles being judged 'of no value'; and any other information supplied by the requester. On the basis of all these records, a decision is made as to the prime cause or causes of the particular failure under review.

Lancaster found that recall failures occurred in 238 out of 302 searches, while precision failures occurred in 278 out of 302 searches. More specifically, some 797 relevant documents were not retrieved. More than 3,000 documents that were retrieved were judged nonrelevant by the requesters. Lancaster's original research report contains statistics about search failures along with detailed explanations of their causes.

Lancaster discovered that almost all of the failures could be attributed to problems with indexing, searching, the index language, and the user-system interface. For instance, the indexing subsystem in his research "contributed to 37% of the recall failures and . . . 13% of the precision failures" (Lancaster, 1969, p.127). The searching subsystem, on the other hand, was "the greatest contributor to all the MEDLARS failures, being at least partly responsible for 35% of the recall failures and 32% of the precision failures" (Lancaster, 1969, p.131).

3.3.1.3 Blair and Maron's Full-Text Retrieval System Study

More recently, Blair and Maron (1985) conducted a retrieval effectiveness test on a full-text document retrieval system. They utilized a database that "consisted of just under 40,000 documents, representing roughly 350,000 pages of hard-copy text, which were to be used in the defense of a large corporate law suit" (pp.290-291). The tests were based on some 51 queries obtained from two lawyers.

Precision and recall were used as performance measures in the Blair and Maron study. The precision ratio was straightforward to calculate (by dividing the total number of relevant documents retrieved by the total number of documents retrieved). Blair and Maron used a different method to calculate the recall ratio. The way they found unretrieved relevant documents (and thus studied recall failures) was as follows. They developed "sample frames consisting of subsets of the unretrieved database" that they believed to be "rich in relevant documents" and took random samples from these subsets. Taking samples from subsets of the database rather than the entire database was more advantageous from the methodological point of view "because, for most queries, the percentage of relevant documents in the database was less than 2 percent, making it almost impossible to have both manageable sample sizes and a high level of confidence in the resulting Recall estimates" (pp.291-293).

The results of Blair and Maron's tests showed that the mean precision ratio was 79% and the mean recall ratio was 20% (p.293).

Blair and Maron found that recall failures occurred much more frequently than one would expect: the system failed to retrieve, on the average, four out of five relevant documents in the database. They showed quite convincingly that high recall failures can result from free-text queries, where the user's terminology and that of the system do not match. They also observed that users involved in their retrieval effectiveness study believed that "they were retrieving 75 percent of the relevant documents when, in fact, they were only retrieving 20 percent" (p.295).

3.3.1.4 Markey and Demeyer's Dewey Decimal Classification Online Project

Markey and Demeyer (1986) studied the Dewey Decimal Classification (DDC) system "as an online searcher's tool for subject access, browsing, and display in an online catalog" (p.1). Two online catalogs were employed in the study: "(1) DOC, or Dewey Online Catalog, in which the DDC had been implemented as an online searcher's tool for subject access, browsing, and display; and (2) SOC, or Subject Online Catalog, in which the DDC had not been implemented" (p.109).

They also conducted online retrieval performance tests using recall and precision measures to reveal problems with online catalogs and to identify their inadequacies. Precision was defined in their study as the proportion of unique relevant items retrieved and displayed. This definition of precision differs from the one given in Chapter 2 in that it takes into account only retrieved and displayed items (instead of all retrieved items) in the calculation of precision ratio. The researchers made no attempt to have users display and make relevance assessments about all the retrieved items in order to calculate the absolute precision ratio (p.162).

Their estimated recall scores were also based on retrieved and displayed items only, not on all the relevant items in the collection. Understandably, they found it impractical to scan the entire database for every query to find all the relevant items in the collection. They used an estimated recall formula "that combined the relevant items retrieved and displayed in the SOC search for a query and the relevant items retrieved and displayed in the DOC search for the same query" (p.144). In order to find the estimated recall ratio for each search, the number of unique relevant items retrieved and displayed in one catalog was divided by the total number of unique relevant items retrieved and displayed for the same query in both catalogs. No attempt was made to find other potentially relevant items in the database.

The estimated recall scores in the study ranged from a low of 44% to a high of 75%. They found that "searches were likely to retrieve and display a large proportion of relevant items that were unique . . . for the same topic in SOC and DOC" even though DOC's estimated recall was lower than that of SOC (p.146). They also asked users if they were satisfied with the search results, and "the majority of patrons expressed satisfaction with the search in the system yielding higher estimated recall" (p.149). The average precision scores ranged from a low of 26% to a high of 65% (p.165, Table 42). Considering that only a fraction of items retrieved in the searches were actually displayed, the authors noted that precision was affected by the order in which retrieved items were displayed. They found precision to be a less reliable criterion with which to measure the performance of an online catalog (p.162).

They asked users which system gave more satisfactory results for their searches and compared users' responses with the precision scores. They concluded that "there was no relationship between patrons' search satisfaction and the precision of their online searches" (p.166; *cf.* Tessier, 1981).

Markey and Demeyer also analyzed a total of 680 subject searches as part of the DDC Online Project and found that 34 out of 680 subject searches (5%) failed. Two major reasons for subject search failures were identified as follows: (1) the topic was marginal (35%), and (2) the users' vocabulary did not match subject headings (24%) (p.182). Their research report gives a detailed account of the failure analysis of different subject searching options in an online catalog enhanced with a classification system (DDC) (p.182).[11]

Markey and Demeyer apparently did not count "zero retrievals" as search failures.[12] Nor did they include in their analysis partial search failures that retrieved at least some relevant documents. Presumably, that's why the number of search failures they analyzed were relatively low.

3.3.2 Studies Utilizing User Satisfaction Measures

It was noted earlier (Section 3.2.2) that analyzing search failures utilizing user satisfaction measures is extremely complicated. Few researchers have attempted to look at search failures in light of user satisfaction.

Hilchey and Hurych (1985) analyzed 153 online search evaluation forms returned by the users in a university library. Almost half of the respondents (47%) found the search results "most relevant." An additional 32% of the respondents graded the results as "half relevant." Only 6% found all search results relevant. In

[11] Especially, see Chapter 8, pp.173-291.

[12] Tables 50-52 in Markey and Demeyer's (1986) report summarize the reasons for failure and success in subject retrieval in the DDC Online Project. Yet, "zero hit" or "zero retrieval" was not mentioned as a reason for failure in the tables. They acknowledge nonetheless that "there were many searches . . . in which searchers entered one or more access points into the catalog . . . that failed to result in retrieval and display of relevant (or any) items for the same reasons as listed in tables 50-52" (p.189).

short, 85% of the respondents felt that search results were at least half relevant. The return rate in this study was about 10%. Although authors claim that the return rate was "unprejudiced in any way," returned questionnaire forms may have primarily come from satisfied users.

Ankeny (1991, pp.352-354) reviewed the studies reporting user satisfaction in end-user search services such as MEDLINE and BRS/After Dark and also reported the results of two studies he conducted. In the first study, he surveyed 190 end-users and found that 78% of the users located what they wanted in two business databases (DIALOG Business Connection and Dow Jones News/Retrieval). More than 81% of the users rated the services favorably by giving "an overall rating of 4 or 5 on the five-point scale" (p.354).

In the second study, he surveyed some 600 end-users. He used a stricter measure of search success (with a reliability coefficient of .90) in the second study in which a search query was considered as successful when the user: a) was fully satisfied with the search; b) found exactly what was desired; and c) was not dissatisfied in any way. He found that "[o]f the 600 searches in the sample, 233 met all three criteria for complete success and 367 were less than successful, yielding an overall success rate of 38.8 percent" (p.354). Reported reasons for dissatisfaction in 367 "less-than-successful" searches were as follows: system problems; amount, relevancy, or level of the information retrieved; lack of better printed instructions; and lack of more informed and accommodating staff.

Kirby and Miller (1986) analyzed search failures encountered by MEDLINE end-users employing the Colleague search software. In order to find the search successes and failures, end-users compared their search results with the mediated follow-up search results. "Successful" and "incomplete" end-user searches were identified as follows:

> 'Successful' Colleague searches were those for which the follow-up search added nothing important, as indicated by one of two questionnaire responses: 'My search gave satisfactory results, and nothing *essential* was added by the second search' . . . or 'Neither search provided satisfactory results.' Both responses were regarded as 'successful' in that the end user was no less successful in meeting the information need than the trained search analyst. 'Incomplete' Colleague searches were those which had missed important articles, according to end user questionnaire responses after reviewing the follow-up search results (p.20, original emphasis).

However, end-users were not asked to judge each record retrieved by either search. Rather, "the comparison was based on search terms and combinations recorded on the follow-up search form, and on the number of citations printed in the follow-up search" (p.20).

Kirby and Miller examined 52 searches. Of the 52 searches, 31 were "incomplete." The major cause of search failures (67.7%) was the search strategy. The rest of the search failures were due to system mechanics and database selection (22.6% and 9.7%, respectively).

3.3.3 Studies Utilizing Transaction Logs

Several researchers have used transaction logs to study search failures in online catalogs. Dickson (1984, p.26) studied a sample of "zero-hit" author and title searches using the transaction log of Northwestern University Library's online catalog and analyzed why the searches failed. She found out that about 23% of author searches and 37% of title searches retrieved nothing. Misspellings and mistakes in the search formulation were the major causes of zero-hit searches.

Jones (1986) examined transaction logs of the Okapi online catalog and found several unsatisfactory areas in its operation due to, among others, spelling errors, failures in subject searching, and user-system interface problems. He analyzed some 300 subject searches performed on Okapi and found that 25% of them failed: "Using relevance assessments based on a display of the first ten records, the experimenter decided that 62.4% of searches were almost certainly successful, 13% may have been successful, 4.5% were collection failures and 25% failed absolutely" (pp.7-8).

In a follow-up study, it was found that 17 out of 122 sessions (or 13.9%) failed in the Okapi (including two sessions that failed due the collection not containing relevant items). (Most sessions contained more than one search.) In seven sessions, the users' vocabulary did not match that of the catalog (e.g., "sociology of shopping"). Another four sessions failed because the topics expressed by the users were too specific (e.g., "textile industry input-output tables"). Two searches failed because searches did not describe users' needs (e.g., one user entered his query simply as "sterling" although the interviewer found out he was actually looking for "economics--sterling shares and gold") (Walker, S. & Jones, 1987, pp.117-119).

The most recent Okapi report states that "the proportion of (non-aborted) searches which failed to retrieve any records is very low indeed (3.9% overall)" (Walker, S. & Hancock-Beaulieu, 1991, p.30).[13] The authors of the report claim

[13] The authors also surveyed the users to find out if they were satisfied with their search results using a five-point satisfaction scale. Ninety-five out of a total of 120 users (or 80%) indicated that they were satisfied with the search outcome (they marked 4 or 5 on the scale), 19 users (or 16%) "had some reservations" (i.e., they marked 3 on the scale), and 6 users (or 4%) "were negative" (i.e., they marked 1 or 2) (Walker, S. & Hancock-Beaulieu, 1991, pp.24-25).

that the improvement is primarily due to: (1) Okapi's "best match" search, and (2) stemming and automatic cross-referencing (p.31).

Peters (1989) analyzed the transaction logs of a union online catalog (the University of Missouri Information Network) and found that 40% of the searches in that catalog produced zero hits. He classified the causes of search failures under 14 different groups, including typographical and spelling errors (10.9% and 9.9%, respectively) and the search system itself (9.7%). Approximately 40% of the failures were collection failures (i.e., the item sought was not in the database). However, Peters' study was not based on a rigorous analysis of zero-hit searches by re-entering queries to determine the exact causes of failures. Rather, "the analyzers made intelligent guesses . . . of the probable causes" (p.270).

Hunter (1991) analyzed thirteen hours of transaction logs, amounting to some 3,700 searches performed in a large academic library online catalog. She used the same classification schema as Peters (1989) and categorized the causes of search failures under 18 different groups. The overall search failure rate in Hunter's study was found to be 54.2%. The major causes of search failures were identified as the controlled vocabulary in subject searching (29%), the system itself (18%), and the typographical errors (15%). However, it was not explained in detail what sorts of controlled vocabulary failures occurred and what the specific causes were.

C.J. Walker and her colleagues (1991) obtained similar results when they studied the problems encountered by clinical end-users of MEDLINE and GRATEFUL MED. They defined search failure, which they called "unproductive search," as "one that did not retrieve any citations," and they analyzed 172 such searches (p.68). They found that 48% of the search failures occurred because of some flaw in the search strategy. The software in use was responsible for 41% of the search failures. System failures constituted some 11% of all search failures.

Zink (1991) analyzed transaction logs of 6,118 searches that took place on the WolfPAC online catalog at the University of Nevada. He found that:

> more than one of every four (27.81 percent or 1,702) failed to retrieve at least one bibliographical record. Subject searches yielded 667 unsuccessful searches, or 39.19 percent of the total number of unsuccessful searches. Author searches resulted in 250 unsuccessful searches (14.69 percent of the total). Searches by all other criteria accounted for 300 unsuccessful searches (17.63 percent of the total) (p.51).

Collection failures (57.60%), misspellings (18%), and placing first name "improperly" before last name (15.20%) caused most of the author search failures. Similar failure rates were also observed for the title searches (collection failures, 61.86%, and misspellings, 14.23%). In 111 unsuccessful title searches (22.89%), searchers seemed to be attempting to find subject or author information. Sixty-three percent of the subject searches failed because the user-entered subject words were not "legitimate" Library of Congress subject headings. Misspellings and collection failures accounted for 23.24% and 10.64% of all subject search failures.

Most of the studies summarized above benefited from transaction monitoring to the extent that "zero-hit" searches were identified from transaction logs.[14] Researchers examined the zero-hit searches in order to find out why a particular

[14] The following studies should be exempted from this as their analyses were not based on zero-hit searches only: Jones (1986), Walker, S. & Jones (1987), and Walker, S. & Hancock-Beaulieu (1991).

search query failed to retrieve anything in the database. Unlike Lancaster (1968), they did not attempt to identify the causes of recall and precision failures.

3.3.4 Studies Utilizing the Critical Incident Technique

It was mentioned earlier (Section 3.2.4) that Wilson *et al.* (1989) studied searching in MEDLINE using the critical incident technique. The researchers first devised a sampling strategy and developed an interview protocol to elicit the desired information from the subjects. They then developed three "frames of reference" to analyze the interview data: "(1) 'Why was the information needed?,' (2) 'How did the information obtained impact the decision-making of the individual who needed the information?,' and (3) 'How did the information obtained impact the outcome of the clinical or other situation that occasioned the search?'" (p.5). After a qualitative analysis of the critical incident reports, the frames of reference were used to create three similar taxonomies.

In the same study, they asked users to explain what they needed the information for and whether they were satisfied with the search outcome. They used incident forms to record the user's account of why a particular search failed or succeeded and, with permission, they tape-recorded the user's comments. They later tried to match these "incident reports" against MEDLINE transaction log records for each search in order to find out the actual reasons for search failures and successes.

They examined some 26 user-designated ineffective incident reports in order to "characterize the nature of the ineffective searches, analyze the relationship between what the user said and what the transaction log said happened during the search, and ascertain, by performing an analogous MEDLINE search, whether a search could have been performed which would have met the user's objective" (p.81). Most ineffective searches (23 out of 26) were identified as such because the users "could not find what they were looking for and/or could not find relevant materials." An appendix summarizing the analysis of each ineffective search accompanied their research report.

After extensive examination of interview transcripts and transaction logs for ineffective searches, the researchers concluded that users did not appear to comprehend:

> 1. How to do subject searching.
> 2. How MeSH [Medical Subject Headings] works.
> 3. How they can apply that understanding to map their search requests into a vocabulary that is likely to retrieve considerably more relevant materials (pp.83-84).

It appears that critical incident technique can successfully be used in the analysis of search failures in online catalogs as well. Matching incident reports against transaction logs is especially promising. Since the analyst will, through incident reports, gather contextual data for each search query, more informed relevance judgments can be made. Furthermore, this technique also can be utilized to compare user-designated search effectiveness with that obtained through traditional retrieval effectiveness measures.

3.3.5 Other Search Failure Studies

Some experimental studies looked into strict matching failures that occurred when users tried to do catalog searches.

Gouke and Pease (1982) analyzed the success rates of the users in matching titles and found that the success rate in finding "nonproblem" titles was 82%, whereas the rate was 48% for "problem" titles. Almost half of the users failed to match simple titles in the online catalog for various reasons (e.g., titles appearing as subject, hyphenated words, words on stoplist, foreign titles, and abbreviations) (p.139).

Alzofon and Van Pulis (1984) surveyed 430 users of the LCS online catalog of the Ohio State University Libraries to identify the patterns of searching. They also studied the success rates for known-item and subject searches. They replicated the users' searches on the catalog and found that the author-title search had a success rate of 85% compared with 77% for author searches and 68% for subject searches (p.113).

Janosky *et al.* (1986) studied the errors that users made in performing searches in the LCS online catalog of the Ohio State University Libraries. They hired 30 volunteer students who had no prior experience with the online catalog under investigation. Each student searched four queries in the catalog. (Queries were the same for all students.) They performed one subject search and three known-item searches. Authors summarize the procedure and results as follows:

> They [users] were asked to search until they either found the item(s) in question or believed that the item(s) was not present in the library system. They were told that it was possible that the item in question was not contained in the library. While searching, subjects were asked to think aloud. . . . A success rate was computed for each search. Since all search items were actually in the library system (subjects were not told this fact), 'success' is defined as correctly locating the information requested about an item. . . . For the four searches, the success rate ranged from a high of 58% to a low of 0% (p.576).

It appears that users experienced serious problems with mechanical aspects of searching in this catalog, which in turn influenced the success rate considerably. For instance, "HELP-AUTHOR" was the "correct" help command, and users who entered "HELP AUTHOR" failed to get any help about author searches (notice the hyphen between the two words). On-screen and offline instructions in this system that advised users to type in commands "exactly as listed" did not seem to help users much to recover from such search failures. A more forgiving user interface would have easily prevented similar failures from occurring in the first place. The authors concluded: "It is not sufficient to simply tell users that they have made an error. Failure to deal with the causes of an error often snowballed into a whole string of misinterpretations, resulting in complete failures to solve the problem of using LCS" (p.591).

Cherry (1992) studied some 100 search sessions using the University of Toronto Libraries' online catalog (FELIX). She analyzed, among others, a small number (42) of zero-hit subject searches "in an effort to identify conversions that would improve recall" (p.97). Each zero-hit subject search was re-entered as, among others, title, keyword title, and keyword subject search so as to see if it would retrieve any documents. She found that:

> keyword subject, keyword title, or title searches using the original query from the user's zero-hit subject search were as fruitful or more fruitful than new searches constructed from cross-references provided by *LCSH*. Thus, it is suggested that educating users in the use of *LCSH* or providing OPAC [online public access catalog] software to automatically provide LCSH cross-references will not solve the problems with the majority of zero-hit subject searches (p.99).

Seaman (1992) examined the interlibrary loan borrowing requests made by the users for items that were listed in the online catalog of the Ohio State University. Approximately 9% of the requests were for such items. The author reasoned that each interlibrary loan borrowing request for a known item that was already in the online catalog represents "either a failure of the user to search the system correctly or a failure of the catalog to retrieve the required record" (p.113). He took a sample of 226 interlibrary loan borrowing requests and identified user errors (such as spelling errors, incorrect author or title) and catalog errors (such as punctuation or corporate word order). Approximately half of the failures in the sample were due to user errors while catalog failures represented the other half.

3.3.6 Related Studies

A few studies that were not directly concerned with the causes of search failures, but which nevertheless addressed relevant issues are summarized below.

Hildreth (1989) considers the "vocabulary" problem as the major retrieval problem in today's online catalogs and asserts that "no other issue is as central to retrieval performance and user satisfaction" (p.69). It may be so because controlled vocabularies are far more complicated than users can easily grasp in a short period of time. Several researchers have found that the lack of knowledge concerning the Library of Congress Subject Headings (LCSH) is one of the most important reasons why searches fail in online catalogs (Bates, 1986; Borgman, 1986; Byrne & Micco, 1988; Dale, 1989; Frost, 1987a, 1987b, 1989; Frost & Dede, 1988; Gerhan, 1989; Holley, 1989; Kaske, 1988a, 1988b; Kaske & Sanders, 1980; Lawrence, 1985; Lewis, 1987; Markey, 1983, 1984, 1985, 1986, 1988; Mischo, 1981; Svenonius, 1986; Svenonius & Schmierer, 1977; Wang, 1985). Larson (1991c, p.181) found that almost half of all subject searches in the MELVYL® online catalog retrieved nothing.

More recently, Larson (1991b) analyzed the use of MELVYL over a longer period of time (six years) and found that there is a significant positive correlation between the failure rate, which is defined as the proportion of search queries that retrieved nothing, and the percentage of subject searching (p.208). This result confirms the findings of an earlier formal analysis of factors contributing to success and satisfaction: "problems with subject searching were the most important deterrents to user satisfaction" (*University of California Users Look at MELVYL*, 1983, p.97).

Larson (1991a, 1991c) reviewed the literature on subject search failures in online catalogs along with remedies offered to reduce subject search problems. Subject retrieval failures in online catalogs could be reduced in a number of ways, including assigning more subject headings to bibliographic records, providing keyword searching, and enhancing classification retrieval.

Carlyle (1989) studied the match between users' vocabulary and LCSH using transaction logs and found that "single LCSH headings match user expressions exactly about 47% of the time" (p.37). A study conducted by Van Pulis and Ludy (1988) showed that 53% of the users' terms matched subject headings in the online catalog (pp.528-529). Vizine-Goetz and Markey Drabenstott (1991) extracted queries from transaction logs of three online catalogs (SULIRS, ORION, and LS/2000) and analyzed them "both by computer and manually to determine the extent to which they matched subject headings" (p.157). They found that less than half of the subject query terms exactly matched the Library of Congress subject headings. The findings suggest that some search failures can be attributed to controlled vocabularies in online catalogs. However, as the authors note, "such analyses . . . reveal little about whether matching terms satisfactorily represent users' topics of interest" (p.161).

3.4 Conclusion

There is no agreed-upon definition of what constitutes search failure in document retrieval systems. In part, this is due to the multiplicity of data gathering tools and techniques used in the analysis of search failures (e.g., the critical incident technique, controlled experiments, interviews, questionnaires, talk-aloud techniques, and transaction monitoring). Different data gathering methods have different strengths and weaknesses.

Many of the studies reviewed in this paper examined search failures based on zero retrievals in online catalogs. Partial search failures have been studied much less frequently. Experiments that investigate the relationship between search failures and user needs or characteristics are even scarcer. This is not surprising because identifying zero retrievals from transaction logs is relatively easy and inexpensive. By contrast, analyzing search failures using precision and recall measures is more expensive and time-consuming. So is the investigation of user needs and interests, which could help researchers make more informed judgments about search failures identified through other means. No single method or technique is self-sufficient to analyze all search failures in document retrieval systems and to interpret the findings.

As for the causes of search failures, transaction logs of the searches that retrieved no records in online catalogs reveal that users are having numerous mechanical problems, such as improperly keying commands and misspelling words. Such problems can be alleviated to a certain extent by designing more intuitive user interfaces that would not only take into account user expertise and task complexity, but also would give advice and simplify the user's task (Buckland & Florian, 1991). Newer online catalogs are dealing with these problems by incorporating more sophisticated stemming algorithms and Soundex-type techniques to correct misspellings.

Transaction log analysis also reveals that users' lack of knowledge of controlled vocabularies and query languages causes many search failures and, subsequently, results in user frustration. Most users are not aware of the role of controlled vocabularies in document retrieval systems. They do not seem to understand the structure of rigid indexing and query languages. Consequently, their search query terms, which are expressed in their own words, often fail to match the titles and subject headings of the documents, causing search failures. "Brittle" query languages based on Boolean logic tend to exacerbate this situation further, especially for complicated search queries.

Transaction monitoring is the most appropriate technique to study search failures when the cause(s) of search failures are obvious (e.g., zero retrievals due to misspellings or collection failures). However, transaction monitoring seems to be less efficient in dealing with more complicated failures. For example, partial failures can be best studied with the help of the user. After all, the user is the key person in the analysis of search failures. It is the user who can explain what he or she was trying to do and whether it was successful. Such input from the user puts each search into perspective and provides much needed contextual information. However, users do not get identified in most transaction log studies. Without user feedback, researchers are faced with the unenviable task of coming up with a rational explanation as to why a particular search failed.

Notwithstanding the circumstantial evidence gathered through various online catalog studies in the past, studies examining the match between users' vocabulary and that of online document retrieval systems are scarce. Moreover, the probable effects of mismatching on search failures are yet to be fully explored.

Users prefer to be able to express their information needs in natural language, but most contemporary online catalogs cannot accommodate search requests submitted in natural language form. However, it is believed that natural language query interfaces may reduce search failures in document retrieval systems. Natural language search terms will more likely match the titles of the documents in the database. Consequently, the role of natural language interfaces in reducing search failures in document retrieval systems needs to be thoroughly studied.

User input should be sought when analyzing search failures with retrieval effectiveness measures such as precision and recall. The same can be said for failure analysis studies that are based on user satisfaction measures. We should strive for full-scale user involvement as much as possible in every stage of analysis of search failures. Despite user participation in the evaluation process, search failures in document retrieval systems are unlikely to be eliminated altogether. However, only through user participation will we find the real causes of search failures and, consequently, build better document retrieval systems.

CHAPTER IV

SEARCH FAILURES IN ONLINE CATALOGS:
A CONCEPTUAL MODEL

4.0 Introduction

The methods by which search failures are studied in document retrieval systems were discussed in Chapter III along with a critical review of the literature. In this chapter we present a model which enables us to explicate all types of search failures occurring in online catalogs. The concept of "search failure" will be used in its broadest possible sense in this presentation in order to illustrate a wide range of search failures.

4.1 Searching and Retrieval Process

In Chapter II, the major components of an online document retrieval system were given as a store of document representations, a population of users, a retrieval rule and a user interface. The roles of indexing, query formulation, user interface, and retrieval rules are explained in more detail. Moreover, it was pointed out that a search and retrieval process takes place by matching query term(s) entered by users with the document representations on the basis of the retrieval rule.

Searching for and retrieval of information is inherently a complex process. Borgman (1986, p.388) summarizes this process as follows:

> It involves the articulation of an information need, often ambiguous, into precise words and relationships that match the structure of the system (either manual or automated) being searched. In an automated environment, the user must apply two types of knowledge: knowledge

> of the mechanical aspects of searching (syntax and semantics of entering search terms, structuring a search, and negotiating through the system) and knowledge of the conceptual aspects (the 'how' and 'why' of searching --when to use which access point, ways to narrow and broaden search results, alternative search paths, distinguishing between no matches due to a search error and no matches because the item is not in the database, and so on.

Users have to make decisions and relevance judgments during search and retrieval process. Although it is the users who must initiate actions in most cases, they may not have total control over, nor understanding of, all the steps that have to take place in this process. For instance, users are confined with the capabilities of the search and retrieval subsystem. Furthermore, the assumptions the users make and the background knowledge they possess may not always help them in their search endeavor.

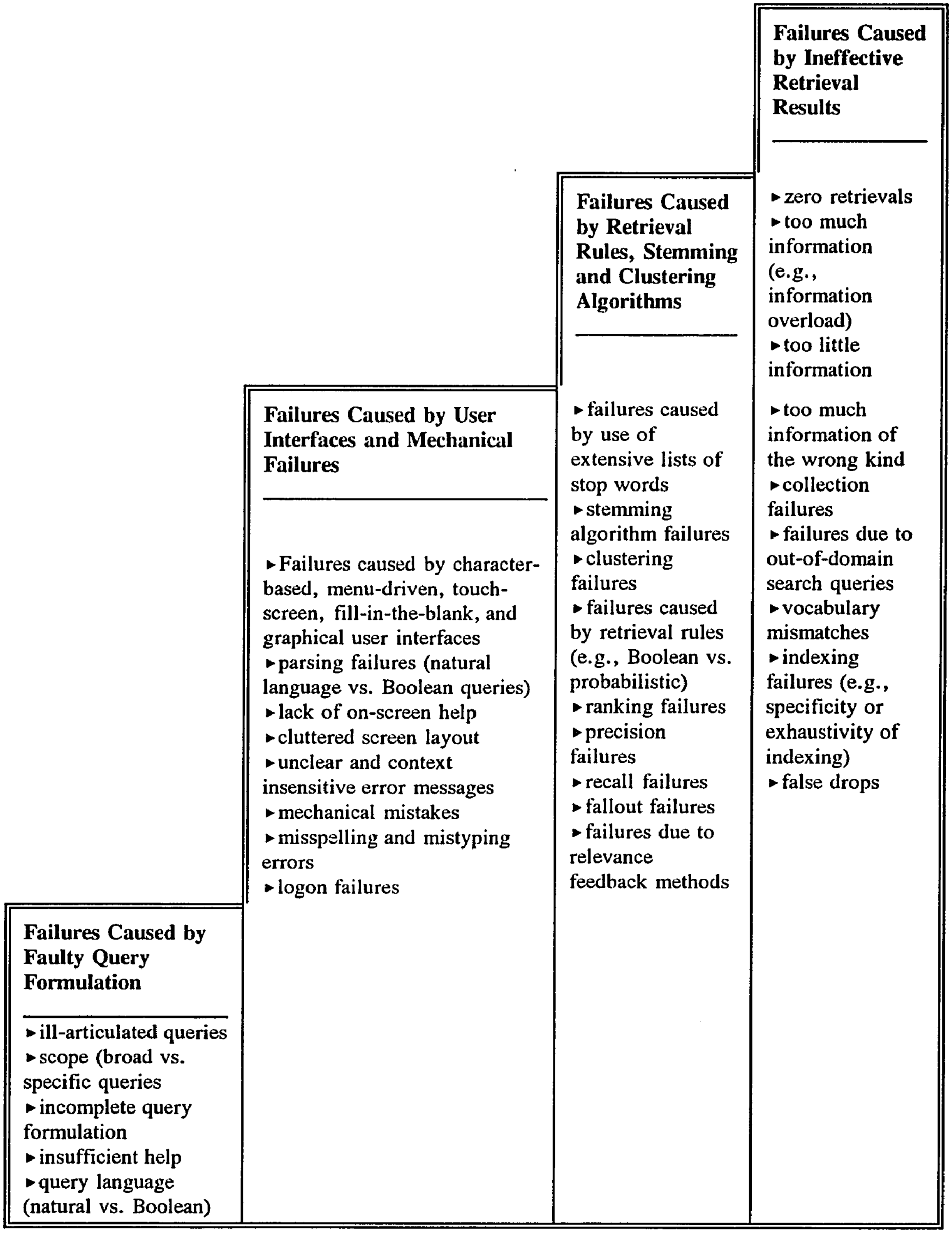
In the following sections we first present a model to describe search failures in online catalogs. We then discuss in detail the types of search failures that occur in online catalogs along with the description of what causes them and why they occur.

4.2 Search Failures in Online Catalogs: A Conceptual Model

It was pointed out in Chapter III that a considerable percentage of online catalog searches fail to satisfy users' information needs. In order to perform successful online catalog searches, users have to engage in an intellectual undertaking. They have to overcome numerous hurdles before they retrieve relevant records. Some hurdles are easier to conquer than others. Some may be invisible to experienced users while users may have no control over some others. Users who accomplish to overcome all hurdles are rewarded with successful search results.

A successful online catalog search process can be likened to climbing a ladder with uneven steps, each step representing the likely places where search failures may occur. To put it differently, each step can be a stumbling block for unprepared users. Figure 4.1 depicts such a ladder with four uneven steps where the size of each step is arbitrary. The degree of difficulty that may be encountered in each step may change from search to search and from one user to the other.

FIGURE 4.1 CATEGORIZATION OF SEARCH FAILURES IN ONLINE CATALOGS



Failures Caused by Faulty Query Formulation

- ill-articulated queries
- scope (broad vs. specific queries
- incomplete query formulation
- insufficient help
- query language (natural vs. Boolean)

Failures Caused by User Interfaces and Mechanical Failures

- Failures caused by character-based, menu-driven, touch-screen, fill-in-the-blank, and graphical user interfaces
- parsing failures (natural language vs. Boolean queries)
- lack of on-screen help
- cluttered screen layout
- unclear and context insensitive error messages
- mechanical mistakes
- misspelling and mistyping errors
- logon failures

Failures Caused by Retrieval Rules, Stemming and Clustering Algorithms

- failures caused by use of extensive lists of stop words
- stemming algorithm failures
- clustering failures
- failures caused by retrieval rules (e.g., Boolean vs. probabilistic)
- ranking failures
- precision failures
- recall failures
- fallout failures
- failures due to relevance feedback methods

Failures Caused by Ineffective Retrieval Results

- zero retrievals
- too much information (e.g., information overload)
- too little information
- too much information of the wrong kind
- collection failures
- failures due to out-of-domain search queries
- vocabulary mismatches
- indexing failures (e.g., specificity or exhaustivity of indexing)
- false drops

As Fig. 4.1 suggests, search failures in online catalogs can be categorized under four broad groups:

1) faulty query formulation;

2) inadequate user interfaces and mechanical failures;

3) retrieval rules; and

4) ineffective retrieval results.

Each category of search failures is discussed below. The discussion follows the logical progression of an hypothetical search query and points out the possible places where events leading to search failures may occur. First, we define major types of search failures.

4.3 Failures Caused by Faulty Query Formulation

The first step in the ladder is the formulation of a formal search query. Several factors play significant roles in formulating successful search statements: the user's background knowledge on the topic for which more information is sought, the document database in use, the query languages available to interrogate the database, and the retrieval rules.

A search query may fail to retrieve records if the search statement contains errors or if it does not describe the user's information need adequately. Typographical errors in search statements or vague, incomplete, too specific or too broad search queries are examples.

Online catalog studies carried out in the past have shown that users experience a wide range of difficulties in formulating their search queries. They have to articulate their search statements and come up with well-thought-out plans so that successful searches can take place. This is not the case for majority of the users, however. Most users formulate their search queries "on-the-fly" and they type in whatever pop into their minds (Markey, 1984, p.70). In some cases they tend to enter incomplete search queries or queries which may not necessarily reflect their real information needs.

The scope of the search queries does not always illustrate whether the user is interested in a broad or specific search on a given topic. Users often issue broad queries and then indicate that they were looking for more specific sources. As they do not know much about collection characteristics, it is understandable to a certain extent that users issue broad queries because they are initially concerned with retrieving "something." Yet very few attempt to revise their original queries.

"Scope failures" also occur when users approach the online catalog unaware of its capabilities. For instance, if the system offers no subject or call number searching and the user attempts to perform this type of searching, the search query will fail. Similarly, if the database contains only monographs and the user expects to retrieve periodical articles, a scope failure will occur.

The availability of printed manuals or on-screen instructions as to how to formulate a successful search query tends to improve things very little. This is not surprising in that very few users are aware of or regularly use such tools. It is open to conjecture whether the poor design of help screens or manual instructions may have something to do with this low level use.

4.4 Failures Caused by User Interfaces and Mechanical Failures

Failures occurring in the course of communicating with the system through user interfaces (e.g., entering and modifying search queries, displaying results) constitute the second category of search failures in online catalogs. Failures due to the user interface of the online catalog stem from the nature of the interface; the nature of the dialog; and the availability of on-screen, context sensitive assistance through the user interface. In other words, interface failures occur when the interface gets in the user's way and prevents the user from finding what he or she is looking for in the online catalog.

4.4.1 Failures Caused by Menu-Driven and Touch-Screen User Interfaces

When interfaces involving touch screens are used, the dialog is extremely rigid because there is no flexibility for the user to enter anything but what is on the screen. Searching becomes tedious as the database size grows and users cannot issue search queries involving more than one concept at a time using touch-sensitive screens. Menu-driven interfaces are the preferred method of search query entry mode of novice users who are either unable or unwilling to invest time to learn the command language of the online catalog. Yet, menu systems offer less capabilities compared with command languages, and users may not enter complex search queries involving the use of more than one indexes (e.g., an author *and* title search).

4.4.2 Failures Caused by Command Language Interfaces

Interfaces based on command languages where users have to formulate their search queries by complying with the strict syntactic rules of the command language, are probably the most common method of interrogating online catalogs. Mastering the use of any character-based command language requires some understanding of the various commands along with their capabilities. The functionality of the online catalog can only be tapped when the user knows which command to use and how to use it.

4.4.2.1 Failures Caused by Parsing Process

The search query entered by the user is "parsed" by the system in order to identify the components of the search statement (e.g., command, index type, search terms, Boolean operators), which enables the system to 'understand' what the user tries to do and thus to take needed actions. Parsing process in character-based interfaces causes several search failures, however. Unless the search terms are entered by observing the rigid syntax rules, parser cannot identify the components of the search query accurately (e.g., command, index to be searched, and search terms). A considerable percentage of search queries submitted to online catalogs through menu-driven interfaces fails because users are often unaware of the fact that the components of a search query must be entered in the exact predetermined order. Otherwise, the interface produces an error message, which may not necessarily be intelligible to the novice user. When this happens, users simply re-enter the same search query without thinking about why the search failed.[1] In fact, such search failures occur frequently in online catalogs with command languages. For instance, almost 10% of all commands that users submitted to a large multi-campus online catalog contained mechanical errors. Of these, almost 50% can be considered parser failures where the system did not recognize the search statements because the components of the search query were not entered in a predetermined order (i.e., first word of command (e.g.,

[1] Siochi and Ehrich (1991) studied the use of repetition indicator ("maximal repeating patterns") as a means of identifying the most troublesome commands in the evaluation of a large image-processing system. It would be interesting to see if the repetition indicator can also be used in the evaluation of command line user interfaces in online catalogs.

FIND) was invalid in 32% of the cases, and an invalid index name was used in 15.3% of the cases.)[2]

4.4.2.1.1 Boolean Searching

When a user must combine search terms to form a query, many errors can be introduced. They include errors related to a complex syntax, errors related to the meaning of Boolean operators (AND, OR, and NOT), and errors stemming from ambiguous parsing of Boolean search requests. More often than not, users do not know how to use Boolean operators correctly. They do not know how Boolean operators may actually affect their search outcome. Most users are unaware of the use of the "implied" Boolean AND operator when the search query contains more than one search terms (or concepts).

Users can be shielded, to a certain extent, from the difficulties with regard to using Boolean operators as part of the query language syntax. Some systems allow users to enter their search queries by filling in the blanks. For example, if a user wishes to find all the titles written by a given author on a given subject, the author name can be entered in the author 'field' and the subject in the subject 'field.' Thus, the system combines these two pieces of information and performs a Boolean search. Note, however, that search failures may still occur in systems with the fill-in-the-blank-type user interfaces due to inherent problems with Boolean logic (see section 4.6).

Fill-in-the-blank-type search query entry mode is more commonly used in CD-

[2] The catalog we refer to is MELVYL®, online catalog of the University of California libraries. Figures reported were taken from the use statistics (September 1992) that are available online on MELVYL.

ROM databases. Most online catalogs have yet to allow their users to manipulate the full screen, rather than a single line, to enter their search queries.

4.4.3 Failures Caused by Natural Language Query Interfaces

More recently, some third generation online catalogs with more advanced retrieval techniques began to allow users to enter their search queries using natural language (Doszkocs, 1983; Mitev *et al.*, 1985, Larson, 1989, 1991a). That is to say, the user is not bound with the syntax of the command language or the use of menus or Boolean operators: he or she simply starts describing the search query using the preferred terms. The lack of a formal query language syntax, it is believed, facilitates the user to express the search query more effectively. It was pointed out earlier that users unaware of the existence of the command language may keep re-entering the same query despite the error message. Such users may benefit from natural language user interfaces where simply typing in the search statement will suffice to retrieve some records out of the database in most cases. Moreover, other users who tend to type in whatever pop into their minds may also benefit from entering their search queries in natural language form.

Search failures occur in full-text systems or online catalogs with natural language user interfaces. The main cause of search failures that occur in natural language user interfaces is their lack of natural language understanding capabilities. Retrieval-worthy search terms are often ignored because the parser cannot distinguish them properly. For instance, a search query such as "I'm interested in books on user interfaces but *not* graphical user interfaces" might retrieve books on user interfaces as well as graphical user interfaces. For, parser may not have any understanding of what "not" means in natural language (Krovetz & Croft, 1992, p.128).

Third generation online catalogs with natural language interfaces do not, unlike second generation online catalogs, provide a wide variety of search options such as author/title searching or Boolean searching, although they handle topical search queries fairly effectively.

From the users' vantage point, there appears to be some differences in terms of formulating and submitting search queries to online catalogs using different types of interfaces. As Buchanan (1992) pointed out, it is not "self-evident that users will submit exactly the same set of keywords to an interface that invites 'natural language' input as to an interface that requires Boolean set construction." As we shall see later (section 4.6), retrieval and display rules may also have some impact on the choice of the user interface.

At present, users are constrained by only a single type of user interface on a given online catalog. Even though there may exist more than one type of user interface for the same online catalog (i.e., menu-driven vs. command language), they may not necessarily co-exist on the same system for the user to select. The availability of several different user interfaces on different systems makes the user's task more difficult. Users may be unfamiliar with all types of user interfaces. An experienced user of a command language with Boolean capabilities may have problems in the course of entering his or her query in an online catalog which accepts queries in natural language form. In other words, searching skills acquired using one type of interface may not necessarily be transferable to other types of interfaces, which is likely to cause an increase in the number of search failures. It is expected that future online catalogs will be equipped with "mapping" facilities to convert a query from one user interface mode to another.

4.4.4 Failures Caused by Mechanical Errors

It was pointed out earlier that commands that users enter tend to contain mechanical errors. In fact, a majority of such commands, unless intercepted by the user interface, retrieve nothing. Search failures due to mechanical errors are very common and not limited to the ones committed during the query entry process. Such errors may occur anywhere during the retrieval process: from logon procedures to entering commands, from displaying search results to interpreting system prompts.

In the simplest case, users may not know how to perform a search or how to proceed. Clearly, they may be aware of the fact that "one has to tell the system something in order to get anything out" (Bates, 1989b, p.405). Yet, what it is they should do may not be clear. If this is the case, users tend to make a lot of mechanical errors. Unless the user interface provides some clues as to what the next step is, they may feel helpless. In such instances, needless to say, search queries often fail to retrieve any records from the database.

Despite the availability of on-screen help, users may have difficulty figuring out what advice they are given and what they are supposed to do. To put it differently, help screens provided by user interfaces are often not "context sensitive." They offer "boiler plate" explanations and tend to read like an "essay." More often than not such explanations fill the full screen. They are usually ignored by the users because of the poor display layout.

It goes without saying that clear and understandable explanations offered by the user interface system as to the use of each command enable users to improve their expertise in using more advanced features of the online catalogs. In general, very few commands are used regularly by the users, which means that the full functionality of a given online catalog remains unexplored for most users. If the user interface provides help with the more advanced features available in the system, search success will also increase.

Poorly-worded error messages may also discourage some users because few users would be pleased to be reminded that they made an error. Judgmental error messages that offer no further help are especially damaging in that errors, especially mechanical ones, tend to beget new errors, sometimes compelling users to abandon their searches (Penniman, 1975a, 1975b; Penniman & Dominic, 1980; Cooper, M., 1991).

Misspellings and typographical errors constitute a considerable percentage of search failures occurring in online catalogs. Such search failures can be avoided provided simple programs that check spelling available in the user interfaces.

Several types of search failures occur due to awkward user interfaces. Users may not necessarily know how to formulate search queries or how the retrieval rules work. They may not necessarily be aware of the database characteristics, system limitations, and many other things. Yet they may still be able to perform successful searches provided they get help from the user interface.

Online catalogs are used by people with a wide range of skills. Some know nothing about their search topics while others know nothing about the online catalog in use, or vice versa. Some are first-time users while others are experienced in online searching. A well-designed user interface is the one which accommodates the needs of all types of users and guide them step-by-step if necessary. It reveals the structure of the system as users get more experienced so that they can understand what the outcomes of their actions will be.

Some argue that if the user interface is intuitive and "user-friendly," an inexperienced user should be able to figure out how to use the system and get results in about 10 minutes. There is no doubt that searching online catalogs for information is a complicated task for some users. A context-sensitive user interface takes the level of user expertise (both system and subject expertise) into account when dispensing information or offering help. Online catalog user interfaces claiming to support the needs of all types of users generally offer inferior service to inexperienced users. For instance, novice users who choose to use the menu-based version of a given interface usually cannot use the more advanced features that are available to experienced users. In such cases, designing "easy-to-learn" user interfaces with context sensitive help features becomes an extremely useful asset in online catalogs.

To sum up, then, a well-designed user interface is one of the most notable components of an online catalog. It is no exaggeration to suggest that the quality of the user interface often determines the success or failure of searches users perform in online catalogs.

4.5 Retrieval Rules

Failures caused by retrieval rules constitute the third category of search failures. This type of failure occurs when the user is unfamiliar with the search and retrieval logic of the system or when the search statement entered by the user gets misinterpreted by the system. Stemming algorithm failures, failures caused by Boolean and probabilistic retrieval rules, clustering and ranking failures, precision, recall and fallout failures can be grouped under this category.

What takes place in the course of retrieving information from online catalogs is often unknown to the user. In the eyes of most users, an online catalog is often seen as a "black box." They may have an understanding of the main function of the online catalog in terms of matching their search queries with document representations in the database. Yet they may not know what types of activities take place and how retrieval rules are applied.

Various retrieval rules used in document retrieval systems were briefly explained in Chapter II. Search failures caused by constructing Boolean search queries were addressed in section 4.4. The difficulties that the use of Boolean logic impose on the effective use of document retrieval systems are discussed in the literature (see, for instance, Cooper, W. (1988), and Blair & Maron (1985).) In view of the inherent difficulties in its use, Boolean logic was referred to as "the Curse of Boole" (Bing, 1987). In a recent survey, conducted at Indiana State University Libraries, 40% of the respondents did not answer one of the questions on Boolean searching. Their comments indicated that "they did not know what Boolean operators are, and it is likely that some of the respondents who did answer the question did not know much about Boolean operators" (Ensor, 1992, p.215). Such figures seem to indicate that Boolean logic as a retrieval rule is responsible to a certain extent for search failures occurring in online catalogs.

As discussed in Chapter III, precision, recall and fallout failures occur in document retrieval systems. To reiterate, precision failures occur when the system fails to retrieve relevant documents *only* whereas recall failures occur when the system fails to retrieve *all* relevant documents. Fallout failures occur when the system retrieves a lot of nonrelevant documents (i.e., false drops).

Search failures also occur in document retrieval systems where probabilistic retrieval rules are applied. In fact, one of the major objectives of the present study is to document search failures that occur in a probabilistic document retrieval system. The results of our analysis are presented in detail in the following chapters. Suffice to say here that all types of search failures mentioned in this chapter may also be encountered in probabilistic document retrieval systems. In addition, we mention three types of search failures, caused by clustering, ranking, and relevance feedback techniques, that occur primarily in probabilistic online catalogs.

Document clustering techniques were briefly discussed in Chapter II (section 2.7.1). Some probabilistic online catalogs preprocess search queries by clustering similar records together and presenting the contextual information to the user before they actually retrieve individual bibliographic records. Contextual information in this case can be subject headings and classification numbers attached to the documents. Thus, the user would be able to identify the clusters that matches his or her query best, thereby eliminating the ones that might otherwise retrieve useless bibliographic records.

The success of the clustering process depends on how well the search statement describes the user's information needs and how well the clustering technique works. If the search statement fails to describe correctly what the user is looking for, the retrieved clusters may not be very relevant. On the other hand, if, despite the correct query formulation, the system retrieves broad, unpromising, ambiguous cluster records, this may further confuse the user as to the function of the clustering and the overall retrieval process.

Cluster failures then occur when the user selects none of the retrieved cluster records as relevant as the query expansion is based on the subject headings and classification numbers extracted from selected records.

Ranking failures occur, on the other hand, when less promising records are presented at the top of the list due to imprecise query description or term weighting formulae used in probabilistic online catalogs. Users quickly reach their "futility points" and give up displaying records once they encounter nonrelevant records in the retrieved list. Failures occurring during relevance feedback searches are somewhat similar to ranking failures in that retrieved records that are based on user feedback may not necessarily be what the user wants. In fact, research that was carried out on a probabilistic online catalog with relevance feedback mechanism showed that there was a high proportion of false drops among the records retrieved during relevance feedback searches. "The reason appeared to be connected with the fact that too many irrelevant terms were being used in the feedback" (Walker, S. & Hancock-Beaulieu, 1991, p.62).

The question to ask at this point is how much impact does the users' system knowledge (mechanical aspects as well as the retrieval rules) have on search failures? It is generally believed that the users' search success in online catalogs depends very much on their previous experience. They feel in control when they know what they are doing even though they may still not know much about how it is that the system retrieves what it actually retrieves.

Experienced users tend to question the search results more often. They modify their search queries depending on the search outcome and adapt to the system easily (since most, if not all, current online catalogs cannot adapt, or be adapted, according to the needs of different types of users). This is certainly an important factor in reducing search failures.

Conversely, users seeing the online catalog as a "black box" tend not to question the search results. They trust the system and readily accept the results. For instance, if they do not succeed retrieving anything in their first try due to various reasons (e.g., zero retrieval, misspelling), they come to the conclusion very quickly that the items they seek are not listed in the catalog. Such confidence in the online catalog may sometimes work against their finding the desired records in the database.

However, that does not necessarily mean that all search queries issued by inexperienced users should be seen as potential failures. For instance, an inexperienced user looking for books on domestic animals may be tempted to enter a query such as "FIND SUBJECT DOGS CATS PARROTS" if he or she is unaware of how (implied) Boolean AND may affect the search outcome. If all user wants is a book that talks about all three animals listed in the query, there should be nothing wrong with it. Yet, due to the current indexing practices, the probability of a book being assigned all three index terms is pretty slim, not to mention the possibility of retrieving nothing.

It is also likely that the user may not know the difference between Boolean

AND and OR operators. This user may be surprised, then, to learn that there is no books on either of those animals. Similarly, the use of AND in real life is quite different from its use as part of the Boolean retrieval rule, which confuses many users (e.g., "welfare AND housing of primates in zoos").[3]

Several activities, which are transparent to users, take place between the submission of the search request and retrieval of the results. Stop lists, stemming algorithms, clustering and ranking techniques that are in use in online catalogs constitute the basic components of the retrieval rule. Each component may affect the search outcome either directly or indirectly.

The search terms entered by the users are subject to preprocessing in most, if not all, online catalogs prior to the application of the retrieval rule. Excluding "stop" words from the search queries through stop lists is the most commonly applied preprocessing activity in online catalogs. Function words such as "the," "of" and "to" are eliminated from queries because such words are not retrieval-worthy. Most quantifiers such as "all," "few," "little" and "much" are eliminated because they "are not helpful as indicators of word relation." Similarly, general words such as "above," "again," "always," and "already" are also excluded from the search queries as they "have no technical meaning within the subject domain" (Vickery & Vickery, 1992, pp.261-262).

The use of stop lists usually do not impair the search results. Yet there may be some situations wherein users would have liked to be aware of the existence of such a stop list. For instance, a naïve-looking search query consisting of the title

[3] This example is taken from Vickery and Vickery (1992), p.272.

words "to be or not to be" retrieved almost 30,000 records in a large online catalog which all share the word "be" in their titles.[4] On the other hand, the same query used to produce zero results before it was fixed in the early CD-ROM version of the New Oxford English Dictionary because the text retrieval engine, which used stemming, treated all words including the "be" in this query as stop words.[5]

Similarly, the stemming algorithms or automatic truncation used in online catalogs may cause search failures. This type of search failure occurs when the stemming algorithm fails to recognize some or all of the search terms correctly. Search terms submitted to the system are stemmed to their root form using weak or strong stemming algorithms to improve the recall rate. However, in some cases increasing recall through stemming may retrieve useless records, thereby cluttering the relevant records with nonrelevant ones. Furthermore, "[t]he stemming rules need to include look up of a table of exceptions, listing words that should not be stemmed, for example analysis, gas, chaos, axes, matrices, mechanics, porous, quantify, rabies" (Vickery & Vickery, 1992, p.262). For example, single character search terms or little-known abbreviations are often ignored.

Stop lists are used in both second- and third generation online catalogs to eliminate words that are not retrieval-worthy. After excluding stop words, catalogs with Boolean search capabilities accept the search query terms as is and treat each term equally retrieval-worthy. On the other hand, third generation online catalogs

[4] From the results of a title word search that was carried out on MELVYL®, the nine-campus union catalog of the University of California libraries (November 1992). Note also that this title word search was treated as a Boolean query in the form of *A OR NOT A*, which retrieved all the records on *A*! However, MELVYL has more advanced features to deal with such difficulties.

[5] Fredric Gey, private communication, November 8, 1992: a remark made by one of the researchers associated with the project during the IASSIST Conference (1990).

accepting queries in natural language form treat the search terms rather differently. In addition to stemming words to their root form, search query terms are indexed in order to identify retrieval-worthy terms. Each retrieval-worthy term is weighted on the basis of search query and document database characteristics and the retrieval results are ranked. But because most, if not all, third generation online catalogs lack natural language understanding capabilities, they sometimes attribute undue weight to an otherwise useless term in a natural language query, thereby causing search failures. For instance, a user issuing a search query such as "I'd like to see books about . . ." is unlikely to be interested in "books" per se. Yet this term (books) will be regarded as highly retrieval-worthy in a database concentrating on library and information studies.

Most catalogs accept acronyms as regular search terms rather than attempting to replace them with the spelled out forms. A few catalogs can automatically replace a search query entered as, say, "ALA" with "American Library Association," which would improve search results to a certain degree if the expansion is correctly inferred. However, it may not always be desirable to supply spelled out forms automatically, especially in large collections with documents on several disciplines. For instance, "ALA" may also stand for, say, "American Lutherans Association." It is best to consult the user beforehand to prevent such search failures before they occur.

Synonyms exhibit the same problems as acronyms. Again, the lack of natural language understanding capabilities is the main reason behind search failures caused by synonymity. For instance, a user interested in "The Netherlands" is most likely to be interested in "Holland," too.

So far our discussion on retrieval rule has concentrated on preprocessing activities that may take place before the retrieval of records, such as the use of stop lists, stemming algorithms and clustering techniques. It should be stressed, however, that users play a most significant role in the evaluation of retrieval results. To put it differently, no matter how advanced and sophisticated they may be, retrieval rules in and of themselves cannot affirm the relevance of retrieved records. As Van Rijsbergen (1981, p.40) put it forthrightly, ". . .a retrieval system cannot be all things to all men." We now turn our attention to retrieval results and discuss the types of search failures that take place after the system retrieves some records in response to the user's query.

4.6 Ineffective Retrieval Results

Failures caused by ineffective retrieval results are the most important type for the user, which make up the last category of search failures in online catalogs (Fig. 4.1). Peters (1991) defines them as follows:

> Failed outcomes can include no information, too little information, too much information, and too much information of the wrong kind (too much noise or too many false hits) (p.90).

4.6.1 Zero Retrievals

Zero retrievals ("no information") occur when the system retrieves nothing in response to the user's query. This may happen due to a variety of reasons: collection failures, mismatch between the user's vocabulary and that of the system, misspellings or typographical errors, to name but a few. Searches that fail to retrieve any information are relatively easy to identify through transaction monitoring.

Zero retrievals due to collection failures occur when the requested item(s) is not owned by the library and thus not listed in the database. Collection failures may occur regardless of the search rule or retrieval mechanism used. Such failures can be minimized through collection development efforts only.

Zero retrievals due to vocabulary mismatch occur when the user-entered search terms fail to match the authority records or controlled vocabulary of the system. There are several ways to prevent collection failures that occur due to vocabulary mismatch. The use of cross-references (*see* and *see also* in LCSH) in controlled vocabularies or authority records, and the creation of a *Superthesaurus* (Bates, 1989b) can be given as examples. Authority control of personal names is relatively straightforward compared to the vocabulary control of subject headings.

Zero retrievals due to misspelled or mistyped query terms occur because such terms do not match the system vocabulary. To minimize these failures, some online catalogs implement semi-automatic spell-checkers to scan the query for mistakes. Yet this is the exception rather than the rule. Most online catalogs accept the user-entered query terms without checking for mistakes or typographical errors.

4.6.2 Collection Failures

Collection failures will occur if the database contains no bibliographic records on a given topic and retrieves nothing (zero retrieval) or retrieves no relevant records in response to a search query. "Out-of-domain" search queries are not considered collection failures. For instance, a search query on "classification of materials on gay and lesbian studies" can be categorized as collection failure if it retrieves no relevant records in a library and information studies database, whereas "blood transfusion" will be considered an out-of-domain search query.

4.6.3 Information Overload

Information overload, or too much information, can also cause search failures. In most cases, users are not interested in retrieving *all* the relevant sources (e.g., high recall) on a certain topic but, presumably, only the *good* ones (Wilson, 1983). Yet online catalogs cannot distinguish the good ones from not-so-good ones; they retrieve records that fit the user's query description using the retrieval rule provided. They do not contain evaluative information about the items they list, either. Thus each user has to judge by himself or herself whether the sources retrieved are good or not. As the database size grows, retrieved sets can get very large, thereby causing the user either to scan several records or to abandon his or her search.

Various types of search failures may occur due to information overload. In some cases, the user may simply stop after displaying a given number of records without seeing all the records in the retrieved set. This in itself cannot be seen as a search failure. If the user identifies some acceptable records among the records displayed, then the search, as far as the user is concerned, is a success. The records displayed may not necessarily be the best ones among the retrieved.

Search failure in this case may occur when the user fails to identify at least some good records among those displayed. It could be that the search query entered is too broad so that retrieved records are general. This happens frequently in online catalogs. If this is the case, the search system cannot be seen as the cause of the failure, but rather the query may have been formulated poorly, or the query term truncated too generously.

As the database size grows, search failures caused by information overload become inevitable because of the nature of the probability of assigning terms to documents: a relatively few number of broad index terms are assigned to a large number of documents in the database whereas a majority of index terms get assigned to only a handful of documents. It is not unusual to retrieve thousands of records in large online catalogs for broad subject searches such as "history," "education," or "medicine." Although it may seem unreasonable, such broad queries may describe exactly what a few users want. Yet, it is scarcely conceivable that overwhelming majority of users issuing such broad queries would be interested in seeing, let alone evaluating, all the retrieved records. In fact, some online catalogs began to restrict the use of such broad terms in search queries as they are costly and slow down the system for everyone.

Information overload can cause search failures in known-item searches as well as subject searches. There may be too many postings attached to some titles and author names (including corporate authors) (e.g., "Bible" or "Shakespeare"). Obviously, it is easier to deal with information overload occurring in known-item searches than in subject searches.

Zero retrievals and "information overload" are seen the most compelling types of search failures in online catalogs. Larson (1991c) compares these two types of search failures to the twin monsters Scylla and Charybdis and emphasizes that users must navigate skillfully in online catalogs to avoid "smashing onto Scylla's rock (search failure) and being pulled into Charybdis' whirlpool (information overload)" (p.182). As discussed in Chapter III, existing online catalogs cannot deal successfully with either Scylla or Charybdis. Take, for instance, the information overload problem. In order to solve the information overload problem in Boolean systems, the search query is usually made more specific by adding more terms conjunctively because it reduces the number of retrieved records to a manageable size. But this strategy also excludes many potentially relevant documents from the retrieved set (Blair & Maron, 1985, p.297). In fact, adding search terms liberally (and, thus, using Boolean AND) may quickly deteriorate the search outcome to the point where the query may retrieve nothing (Scylla). Attempting to solve the information overload problem by using Boolean operator AND was likened to "entering into some sort of Boolean lottery, where it is quite incidental whether he [the user] actually wins a relevant document as a prize" (Bing, 1987, p.200).

Third generation online catalogs with ranked retrievals provide a more sophisticated solution to the information overload problem. As the sources retrieved are ranked according to the degree of match between the search query terms and titles and subject headings of the documents in the collection, users, it is assumed, can notice the difference between top-ranked and lower-ranked retrievals in terms of their relevance. That is to say, they can discontinue searching once retrieved records become less and less relevant, safely assuming that records farther down the list contain no more promising ones. This is not the case in Boolean searching where records at the top and bottom all have equal chance of being relevant. Users cannot assume, unlike in probabilistic systems, that rest of the retrieved records contain no relevant ones.

Note, however, that information overload caused by single term broad search queries may not necessarily be handled more successfully by probabilistic online catalogs, either. In other words, a broad, single term search query such as "history" and "education" would match several records equally well and the weakly ordered ranked retrievals may not forcefully separate more relevant records from less relevant ones.

4.6.4 Retrieving Too Little Information

Search failures caused by retrieving "too little information" differ slightly from zero retrievals or information overload in that it is difficult for users to determine if they retrieved too little information. From the user's vantage point it can be more detrimental in some cases to retrieve "too little information." They may not necessarily know that they missed some relevant documents. Yet the consequences of retrieving too little information as a type of search failure cannot be overlooked, especially in medicine, legal research and patent searching.

Although it takes longer for users to scan retrieved records, "too much information" (high recall) does not hurt as much as "too little information" (low recall) does. Users can always stop scanning records once they find some that are relevant to their needs. Furthermore, online catalogs that rank retrieved records on the basis of their similarity to the search query facilitate the scanning process. It was pointed out in Chapter III that most users do not consider recall failures as search failures because they cannot tell from the retrieved records that they are missing some other relevant ones. In most cases, unretrieved but relevant records would not hurt them if they are satisfied with what they have already seen.

4.6.5 False Drops

False drops will occur when the system cannot distinguish the subject domain in which a given search term is being used and thus retrieve all the records indexed under this term. Incorrect term relationships may also cause false drops. Such failures mainly stem from the lack of natural language understanding capabilities in present document retrieval systems.

False drops may clutter the retrieved set of records and cause search failures, especially in broad or vague search queries. The main cause of false drops in online catalogs is that the system has no way of distinguishing records with the same titles or subject headings unless further information is provided by the user. As the retrieval rule is based on simple keyword matching, the system cannot differentiate terms that are semantically unrelated. A user interested in natural disasters like "fires" would be hard-pressed to make the connection when presented with Lakoff's monograph titled *Women, Fire, and Dangerous Things* (unless the user is either knowledgeable about categorization and psycholinguistics, or he or she speaks Dyrbal, an aboriginal language of Australia). Similarly, a simple search on the subject "rubbish" may retrieve several unrelated records ranging from rubbish theory to intellectual rubbish to rubbish filtering.

4.6.6 Failures Caused by Indexing Practices and Vocabulary Mismatch

Indexing failures occur when documents are assigned incorrect index terms. They also occur when indexers fail to assign no index terms at all. Indexing failures may also explain some of the search failures caused by ineffective retrieval results (zero retrievals, too much information, too little information, false drops). Furthermore, precision, recall, and fallout failures are mainly caused by indexing practices. For instance, precision failures occur in part when documents are assigned broad index terms. Recall failures occur when relevant documents are not assigned appropriate index terms.

Indexing practices thus play a significant role in search success. Assigning index terms exhaustively to reduce recall failures, for instance, may cause other types of search failures such as information overload. As discussed earlier, second generation online catalogs with Boolean searching capabilities cannot deal with information overload in large retrieved sets successfully.

Vocabulary failures occur when the user-entered search terms fail to match the system's vocabulary (i.e., titles and subject headings of the bibliographic records in the database). In other words, presence or absence of certain index terms and title words may determine whether the user will success in his or her search endeavor.

At this point, it is essential to indicate the role of clustering techniques not only in decreasing false drops but also vocabulary mismatches. False drops occur when the system cannot determine the context in which the search terms are used and therefore retrieves nonrelevant records. To prevent false drops, some third generation online catalogs display the context in which the search terms are used before retrieving the bibliographic records. Vocabulary mismatches, on the other hand, occur when the user-selected terms do not match that of the controlled vocabulary of the system or the titles of documents in the database. Clustering techniques help users match their terminology with that of the system by checking the occurrence of search terms both in title and subject indexes. This is advantageous compared to Boolean searching in that the user is automatically given an extra chance to be able to match his or her terms rather than facing zero retrievals due to vocabulary mismatch. For instance, a user performing a subject search under "sarcophagi" in a second generation online catalog will not be well-served because overwhelming majority of such items were cataloged under "sepulchral monuments." Yet, this search will in third generation online catalogs retrieve records that have "sarcophagi" in either their titles or subject headings (or both). In other words, the user will be automatically directed to the titles which were not cataloged under "sarcophagi" but nevertheless are likely to be relevant (e.g., cataloged under "sepulchral monuments").

That clustering techniques help reduce search failures due to vocabulary mismatches should be seen as a considerable achievement. For, controlled vocabularies such as *LCSH* have for long been criticized for being, among other things, outdated, obscure, biased, and scarcely applied (Chan, 1986a, 1986b). Users experience the most persisting problems with subject searching because of the constraints of controlled vocabularies.

4.7 Summary

Categories of search failures are analyzed by means of a four-step ladder model in this chapter. Each step represents a category of search failures. It was suggested that in order to perform successful searches users have to climb all four steps. The roles of query formulation process, user interfaces, retrieval rules, and ineffective retrieval results in search failures are thus addressed using the ladder model. The types of search failures discussed under each category are as follows: 1) Failures caused by faulty query formulation: specific or broad search queries; vague search statements; scope failures; lack of on-screen help in query formulation process; 2) Failures caused by user interfaces and mechanical failures: failures caused by menu-driven, touch-screen, command language, and natural language user interfaces; parsing failures; users' familiarity (or lack thereof) with Boolean searching; poor screen layout or error messages; 3) Failures caused by retrieval rules: Boolean searching failures; precision, recall, fallout failures; failures caused by clustering, ranking and relevance feedback techniques; failures caused by stemming algorithms; failures caused by the use of acronyms or synonyms; 4) Failures caused by ineffective retrieval results: collection failures; information overload failures; zero retrievals; retrieving too little information; false drops; failures caused by indexing and vocabulary mismatch.

CHAPTER V
THE EXPERIMENT

5.0 Introduction

The theoretical foundations of the present study were presented in the last three chapters. An overview of document retrieval systems is given in Chapter II. Chapter III examined the methods used in the study of search failures and reviewed the major works in the field. A conceptual model of search failures in online catalogs was presented in Chapter IV. A detailed description of the experiment conducted for the present study is given in this chapter.

5.1 The Experiment

The purposes of this study are, among others, to analyze search failures in an experimental online catalog with advanced information retrieval capabilities; to measure the retrieval performance in terms of precision and recall ratios and user designated retrieval effectiveness; and to develop a conceptual model to categorize search failures that occur in online catalogs. An experiment was conducted in order to test the hypotheses and to address the research questions presented in Chapter I. The hypotheses were as follows:

1. Users' assessments of retrieval effectiveness may differ from retrieval performance as measured by precision and recall;

2. Increasing the match between users' vocabulary and system's vocabulary (e.g., titles and subject headings assigned to documents) will help reduce the search failures and improve the retrieval effectiveness in online catalogs;

3. The relevance feedback process will reduce the search failures and enhance the retrieval effectiveness in online catalogs.

Data was gathered on the use of the catalog from September to December 1991. Data concerning users' actual search queries submitted to the catalog, the records retrieved and displayed to the users, users' relevance judgments for each record displayed, records retrieved and displayed after relevance feedback process represents the kinds of data collected during this experiment. Further data was collected, by means of the critical incident technique, from the users about their information needs and intentions when they performed their searches in the online catalog. This data was then analyzed in order to find out the retrieval effectiveness attained in the experimental online catalog. The search failures are documented and their causes investigated in detail.

5.2 The Experimental Environment

5.2.1 The System

The research was carried out at the School of Library and Information Studies, University of California at Berkeley on the CHESHIRE system. CHESHIRE (California Hybrid Extended SMART for Hypertext and Information Retrieval Experimentation) is an experimental online library catalog system "designed to accommodate information retrieval techniques that go beyond simple keyword matching and Boolean retrieval to incorporate methods derived from information retrieval research and hypertext experiments" (Larson, 1989, p.130). It uses a modified version of Salton's SMART system for indexing and retrieval purposes (Salton, 1971b; Buckley, 1987) and currently runs on a DECStation 5000/240 with about one gigabyte of disk space and 64 megabyte of memory.[1] Larson (1989) provides a more detailed information about CHESHIRE.[2]

[1] CHESHIRE ran on a Sun 3/50 workstation under UNIX and the SunTools windowing system, with 320 megabytes of disk storage, during the time period when most of our research was conducted.

[2] For the theoretical basis of, and the probabilistic retrieval models used in, CHESHIRE online catalog, see Larson (1992).

CHESHIRE accommodates queries in natural language form. It currently supports subject searching only. The user describes his or her information need or interest(s) using words taken from natural language and submits this statement to the system. This statement is then "parsed" and analyzed to create a vector representation of the search query. The query is submitted to the system for the retrieval of relevant classification clusters from the collection that best match the user's query. Each cluster record contains the most common title keywords, subject headings and the normalized classification number for the records represented in that cluster. Upon the user's selection of one or more clusters, the query gets further enriched with the terms that appeared in relevant clusters before it is submitted to the system for the retrieval of individual documents from the database.

The classification clustering technique which Larson developed and implemented in CHESHIRE is used for query expansion in CHESHIRE (Larson, 1989, 1991a). The technique is briefly mentioned in Chapter II (section 2.7.1). The method used to retrieve and rank classification clusters is based on probabilistic retrieval models (Larson, 1992). What follows is a more detailed overview of the use of "classification clustering method" in CHESHIRE.

Fig. 5.1 illustrates the classification clustering procedure diagrammatically.

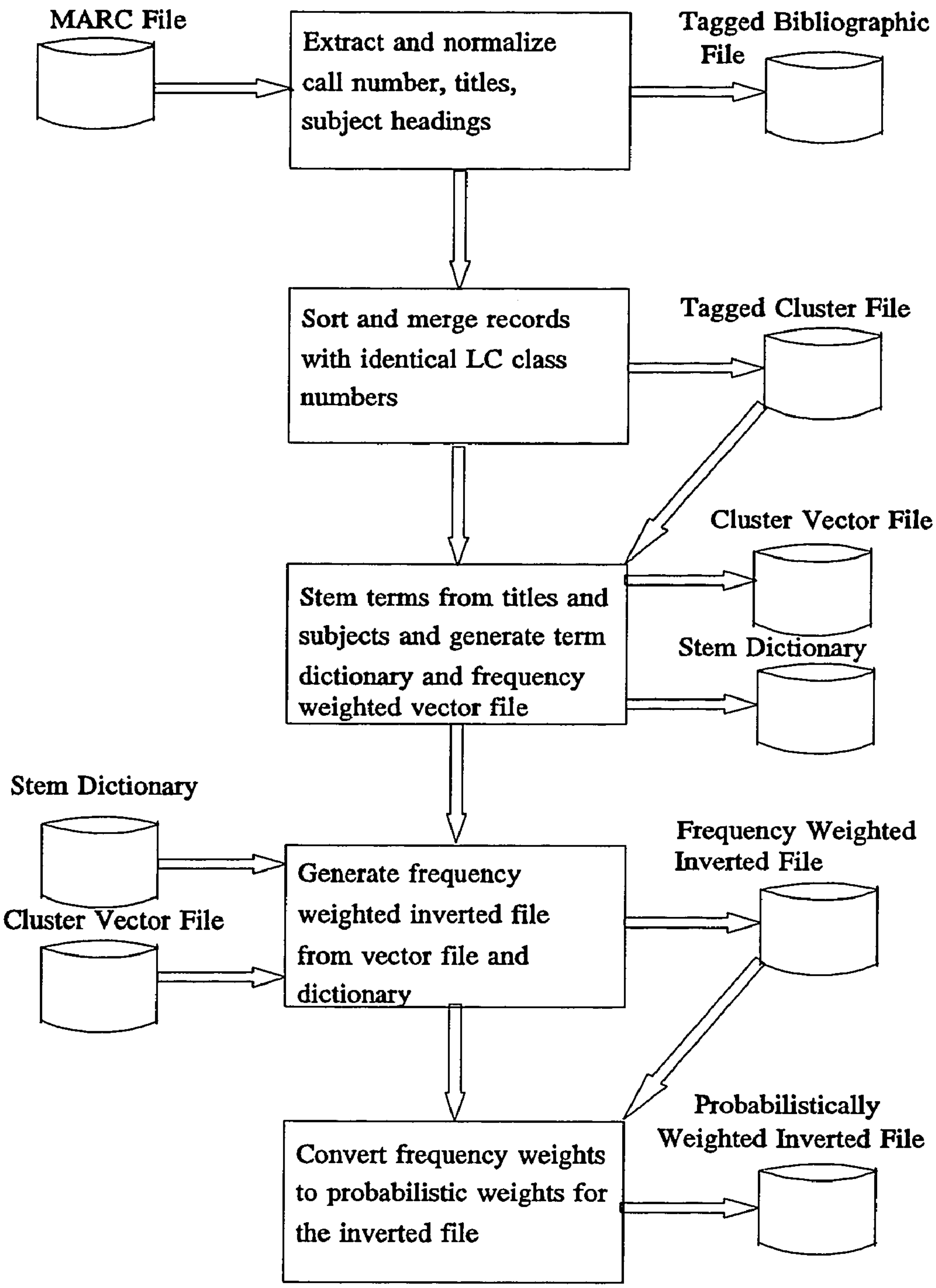

MARC File
Extract and normalize call number, titles, subject headings
Tagged Bibliographic File
Sort and merge records with identical LC class numbers
Tagged Cluster File
Cluster Vector File
Stem terms from titles and subjects and generate term dictionary and frequency weighted vector file
Stem Dictionary
Stem Dictionary
Cluster Vector File
Generate frequency weighted inverted file from vector file and dictionary
Frequency Weighted Inverted File
Probabilistically Weighted Inverted File
Convert frequency weights to probabilistic weights for the inverted file

FIGURE 5.1 CLASSIFICATION CLUSTERING PROCEDURE (SOURCE: LARSON, 1991A, P.157)

Larson (1991a, 1992) provides a more formal presentation of the classification clustering method he developed. He (1992) states:

> [the method] involves merging the topical descriptive elements (title keywords and subject headings) for all MARC records in a given Library of Congress classification. The individual records are *clustered* based on a normalized version of their class number, and each such *classification cluster* is treated as a single 'document' with the combined access points of all the individual documents in the cluster.... The clusters can be characterized as an automatically generated pseudothesaurus, where the terms from titles and subject headings provide a lead-in vocabulary to the concept, or topic, represented by the classification number (p.39).

The classification clustering method improves retrieval effectiveness during document retrieval process as follows:

Suppose that a collection of documents has already been clustered using a particular classification clustering algorithm. Let's further suppose that a user has come to the document retrieval system and issued a specific search query (e.g., "intelligence in dolphins"). First, a retrieval function within the system analyzes the query, eliminates the useless words (using a stop list), processes the query using the stemming and indexing routines and weights the terms in the query to produce a vector representation of the query. Second, the system compares the query representation with each and every document cluster representation in order "to retrieve and rank the cluster records by their probabilistic "score" based on the term weights stored in the inverted file. . . .The ranked clusters are then displayed to the user in the form of a textual description of the classification area (derived from the LC classification summary schedule) along with several of the most frequently assigned subject headings within the cluster." (Larson, 1991a, p.158).

Once the system finds the potentially relevant clusters, the user can judge some of the clusters as being relevant by simply identifying the relevant clusters on the screen and pushing a function key. "After one or more clusters have been selected, the system reformulates the user's query to include class numbers for the selected clusters and retrieves and ranks the individual MARC records based on this expanded query" (Larson, 1991a, p.159).

Larson (1991a) describes how it is that this tentative relevance information for the selected clusters can be utilized for ranking the individual records:

> In the second stage of retrieval . . . , we still have no information about the relevance of individual documents, only the tentative relevance information provided by cluster selection. In this search, the class numbers assigned to the selected clusters are added to the other terms used in the first-stage query. The individual documents are ranked in decreasing order of document relevance weight calculated, using both the original query terms and the selected class numbers, and their associated MARC records are retrieved, formatted, and displayed in this rank order . . . In general, documents from the selected classes will tend to be promoted over all others in the ranking. However, a document with very high index term weights that is not from one of the selected classes can appear in the rankings ahead of documents from that class that have fewer terms in common with the query (pp.159-60).

Although the identification of relevant clusters can properly be considered a type of relevance feedback, we prefer to regard it as some sort of system help before the user's query is run on the entire database.

After all of the above re-weighting and ranking processes, which are based on the user's original query and the selection of relevant clusters, are done, individual records are displayed to the user. This time the user is able to judge each individual record (rather than the cluster) that is retrieved as being relevant or nonrelevant, again by simply pushing the appropriate function key. The user can examine several records by making relevance judgments along the way for each record until he or she thinks that there is no use to continue displaying records as the probability of relevance gets smaller and smaller.

To sum up, classification clustering method brings similar documents together by checking the class number assigned to each document. It also allows users to improve their search queries by displaying some retrieved clusters for the original query. At this point users are given a chance to judge retrieved clusters as being relevant or nonrelevant to their queries. Users' relevance judgments then get to be incorporated into the original search queries, thereby making the original queries more precise and shifting them in the "right direction" to increase retrieval effectiveness.

CHESHIRE has a set of both vector space (e.g., cosine matching, term frequency - inverse document frequency matching (TFIDF)) and probabilistic retrieval models available for experimental purposes. Formal presentations of these models can be found elsewhere (e.g., Larson, 1992). In essence, cosine matching measures the similarity between document and query vectors and "ranks the documents in the collection in decreasing order of their similarity to the query." TFIDF matching is similar to cosine matching. However, TFIDF takes the term frequencies into account and attaches "the highest weights to terms that occur frequently in a given document, but relatively infrequently in the collection as a whole, and low weights to terms that occur infrequently in a given document, but are very common throughout the collection" (Larson, 1992, p.37). Probabilistic models (Model 1, Model 2, Model 3), on the other hand, approach the "document retrieval problem" probabilistically and assume that probability of relevance is a relationship between the searcher and the document, not between the terms used in indexing documents and the terms used in expressing search queries (Maron, 1984).

CHESHIRE also has relevance feedback capabilities to improve retrieval effectiveness.[3] Upon retrieval of documents from the database, the user is asked to judge if the retrieved document is relevant or not. Based on users' relevance judgments on retrieved documents, the original search queries are modified and a new set of, presumably more relevant, documents is retrieved for the same query. Users can repeat the relevance feedback process in CHESHIRE as many times as they want.

Probabilistic retrieval techniques, along with classification clustering and relevance feedback capabilities, have been used for evaluation purposes in this experiment. The feedback weight for an individual query term *i* was computed according to the following probabilistic relevance feedback formula:

$$\log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

where

$$p_i = \frac{relret + \frac{freq}{numdoc}}{numrel + 1.0}$$

$$q_i = \frac{freq - relret + \frac{freq}{numdoc}}{numdoc - numrel + 1.0}$$

where

[3] Relevance feedback concepts are explained in Chapter II (section 2.9).

freq is the frequency of term *i* in the entire collection;

relret is the number of relevant documents term *i* is in;

numrel is the number of relevant documents that are retrieved;

numdoc is the number of documents.

This formula takes into account only the "feedback effect," not the artificial "ranking effect" (i.e., documents retrieved in the first run are not included in the second run) (see Chapter II, section 2.9).

5.2.2 Test Collection

The test collection used for this experiment was that of the bibliographic records of the Library of the School of Library and Information Studies (LSL) of the University of California at Berkeley. LSL has a specialized collection concentrated in library and information sciences, publishing and the book arts, management of libraries and information services, bibliographic organization, censorship and copyright, children's literature, printing and publishing, information policy, information retrieval, systems analysis and automation of libraries, archives and records management, office information systems, and the use of computers in libraries and information services.

The test database for the CHESHIRE system consists of 30,471 MARC records representing the machine-readable holdings of the LSL up to February 1989. Using the test database, Larson (1989, 1991a) created a bibliographic file containing the titles, subject headings and classification numbers from the MARC records. He then generated a cluster file from the bibliographic file using the classification clustering technique. Due to the nature of the LSL's highly specialized collection, more than 80% of the records in the test database fall into LC main class Z. MARC records in the database had some 57,000 Library of Congress subject headings (LCSH) assigned to them, which amounts to about two subject headings per record (Larson, 1991a, p.162). Table 5.1 provides some collection statistics for the test database.

TABLE 5.1
MARC TEST COLLECTION STATISTICS (SOURCE: LARSON 1992, P.40)

	Cluster File	*Bibliographic File*
No. of document vectors	8,435	33,371
No. of distinct terms	33,883	33,891
Total term occurrences	221,042	397,790
Avg. terms per document	26.21	11.92
Avg. term freq. in vectors	2.03	1.14
Avg. documents per term	6.52	11.74
Max. documents per term	2,754	11,999
Avg. documents per cluster	3.95	-

Larson (1992) interprets the data in the table as follows:

> the bibliographic database generated only 8345 clusters, giving an average of just under four bibliographic records per cluster (the standard deviation was 19.50). The majority of records in the database fall into LC main class Z, and in that class the average is about 4.8 records per cluster with a standard deviation of about 23.29 records. As the large standard deviations would suggest, the distribution of bibliographic records to classification clusters is very uneven, with many clusters (67%) consisting of a single record, and some (1.1%) with more than 40 records. The large number of single record clusters is primarily due to the enumerative nature of the LC classification, where Cutter numbers are used to order items alphabetically within broad classes (e.g., the 'By Name A-Z' direction in the schedules). It should be noted also that these single record clusters represent only about 16.9% of the input records. The number of document vectors generated for the test database is larger (33,371) than the number of input MARC records [30,471] due to variant record forms generated

> for MARC records having more than one class number.
>
> The searchable terms in both of the files consist of keyword stems extracted from titles and subject headings, and the normalized class number. Most common words (e.g., 'and,' 'of,' 'the,' 'a,' etc.) are included in a *stop list* and ignored during indexing and retrieval. All other keywords are reduced to word stems using a stemming algorithm. . . . Terms from subject headings and titles are treated separately and considered to be different terms, even if they are based on the same word.
>
> . . .there were an average of about 12 terms per document (standard deviation 5.92) in the bibliographic file and about 26 terms per cluster (standard deviation 65.21) in the clustered file. . . . These statistics indicate that the classification clustering process is having the desired effect of grouping similar bibliographic records together, but the enumerative nature of the classification scheme prevents some records from clustering (pp.40-41).

5.2.3 Subjects

Forty-five entering master's and continuing doctoral students (hereafter "users") in the School of Library and Information Studies of the University of California at Berkeley voluntarily participated in the experiment. They were not compensated for their participation in the study.

5.2.4 Queries

Users performed a total of 228 catalog searches on CHESHIRE during the fall semester of 1991. The topics of search queries were determined by the users, not by the researcher. Most, if not all, search queries originated from users' real information needs.

The number of queries users searched on CHESHIRE is thought to be appropriate for evaluation purposes as most information retrieval experiments in the past had been conducted with either comparable or much fewer number of queries. For instance, some 221 search queries were used in Cranfield II tests, one of the earliest information retrieval experiments. The search queries were "obtained by asking the authors of selected published papers ('base documents') to reconstruct the questions which originally gave rise to these papers" (Robertson, 1981, p.20). Similarly, 302 genuine search queries were used in the MEDLARS study. Search queries used in MEDLARS tests originated from the real information needs of the system's users (Lancaster, 1968). More recently, Blair and Maron (1985) used some fifty-one real search queries, obtained from two lawyers, to test the retrieval effectiveness of the STAIRS system. Tague (1981) observes that "the number of queries in information retrieval tests seems to vary from 15 to 300, with values in the range 50 to 100 being most common" (p.82).

5.3 Preparation for the Experiment

The experiment was carried out on CHESHIRE online catalog. The users' complete interaction with the online catalog was captured on transaction logs. A self-administered questionnaire for each search was filled out by the users. In addition, a post-search structured interview was conducted with the users.

5.3.1 Preparation of Instructions for Users

A two-page handout and a booklet were prepared for instructional purposes. The handout contained background information about CHESHIRE as well as guidelines for CHESHIRE searches (see Appendix A). The booklet demonstrated, with step-by-step instructions, how to get access to CHESHIRE, how to log on, enter a search query, display clusters and bibliographic records, make relevance judgments, and how to perform relevance feedback searches (see Appendix B). Both the handout and booklet were pilot-tested on two users who were unfamiliar with the system.

5.3.2 Preparation of the Data Gathering Tools

A comprehensive analysis of search failures in an online catalog requires the use of a number of data gathering tools, the most important ones being transaction logs, questionnaires and critical incident forms used during the structured interviews to collect critical incident reports about search failures.

Transaction logs were used to record relevant data about the entire session for each search conducted on CHESHIRE. Transaction record for each search consists of the user's password, logon and logoff times and dates (to the nearest second), the full search statement entered by the user, the stemmed roots of search terms and their weights, cluster and bibliographic records retrieved along with their id numbers, ranks, and the user's relevance judgments on displayed records. Relevance feedback data, if applicable, was also captured in the transaction record. The types of data recorded for each search in the transaction logs are illustrated in Appendix C.

A questionnaire and critical incident report forms were designed to record user's experience for each search carried out on CHESHIRE. Both the questionnaire and critical incident report forms were pretested, and suggestions obtained from users (e.g., slight changes in wording of some questions) were incorporated into the final versions of questionnaire and critical incident report forms.

The questionnaire aims to measure, in more precise terms, users' perceived search success for each query submitted to CHESHIRE. It included such questions as: the type of the user; how long ago the search was performed and whether the user was successful in the first try; if not, what was the reason for the search failure; what percentage of the sources the user found especially useful (precision); whether relevance feedback was performed or not; if yes, what was its impact on the search results; and, most helpful and most confusing features of CHESHIRE (see Appendix D for a copy of the questionnaire form).

Two types of critical incident report forms were devised (modified from Wilson *et al.* (1989)): one for reporting "effective searches" and the other for "ineffective searches" (see Appendices E and F for effective and ineffective incident report forms, respectively). The critical incident report form aims to gather, for each search query submitted to CHESHIRE, data on the effectiveness or ineffectiveness of the search query, the user's information needs that triggered the search, the types of sources retrieved and whether they were helpful or not, relevance feedback process, whether CHESHIRE retrieved most of the useful sources or not (recall), and whether sources retrieved were useful or not (precision). Incident reports also include users' own assessments of the effectiveness of their searches. Note that critical incident report form was intended to be used as a structured interview form during the interview with the user. (Interviews were audiotaped (with permission) for further analysis.)

The critical incident report form and the questionnaire form consist of similar questions. The questionnaire form was designed to complement the critical incident reports and to corroborate the findings to be obtained from the critical incident reports.

5.3.3 Recruitment of Users to Participate in the Experiment

Potential participants (all entering master's and continuing doctoral students) were invited to take part in the experiment (see Appendices G and H). The guidelines and detailed instructions were sent to doctoral students (see Appendices A and B) along with the invitation letter. A live demo introducing logon and the search procedures in CHESHIRE was offered to the interested doctoral students. Permission to review their transactions was obtained from participating doctoral students.

Entering master's students were handed out the invitation letter during their scheduled class times for a course offered in the School of Library and Information Studies called LIS 210: Organization of Information (Fall 1991).[4] This was followed by a 20-minute presentation in which an example search session on CHESHIRE was demonstrated. Students were told that, should they decide to participate, the system would be open to their use throughout the semester. They were encouraged to use the system as often as they desired. The written consents of participating master's

[4] LIS 210 is about the organization of information. It covers such issues as subject access and classification. It was emphasized during the presentation that students may want to try an online catalog with more sophisticated search capabilities such as relevance feedback and classification clustering techniques. This can, in a way, be seen as a positive incentive to participate in the experiment.

students to review their transactions were obtained after the presentation.

After the presentations, the transaction log file was monitored daily to see if any searches had been performed. "Thank you" messages were sent to first time users. Later in the semester, students were reminded periodically that they could continue to perform searches on CHESHIRE.

5.4 Data Gathering

As was indicated earlier, users' full interaction with CHESHIRE (user names, search statements, records displayed, relevance judgments, and so on) was recorded in the transaction log file. Users carried out a total of 228 search queries on CHESHIRE. By the time the data collection period ended (mid-December 1991), more than 200,000 lines of data were gathered through transaction monitoring.

The transaction log file was scanned, using data reduction techniques, to extract information about users, search statements, and the outcome of their searches. Such information proved to be the foundation of data gathering process through questionnaires and structured interviews.

Participating users were interviewed throughout December 1991 and spring semester of 1992. Users filled out a questionnaire form for each search they performed (see Appendix D). Afterwards, a structured interview, which was audiotaped, was carried out with users for each search query. If a given search was judged as being "effective" by the user, questions in the Effective Incident Report Form (see Appendix E) were asked. If not, questions in the Ineffective Incident Report Form (see Appendix F) were asked. In addition to audiotaping, users' answers were also recorded on critical incident report forms. More than sixteen hours worth of user comments were audiotaped. These tapes were later transcribed in order to facilitate the analysis process.

All searches submitted to CHESHIRE during the data gathering period were repeated on MELVYL®, the nine-campus University of California online catalog, using its title and subject keyword options. The results were recorded in script files. In addition, searches that retrieved nothing (zero retrievals) on CHESHIRE were redone just to make sure that that was the case. The results of both MELVYL and CHESHIRE searches were later used to calculate the recall ratio for each query.

The limited resources available for this study prevented an experimental design in which the participants would be divided into a control and experimental group so as to compare the results obtained from each of the two groups. In addition, it was not possible to have more than one evaluator to examine the search results or critical incident report forms.

To sum up, then, four methods were used to gather data for each search query performed on CHESHIRE: 1) The outcome of the full search process (query statement, clusters and records retrieved, relevance judgments, etc.) was recorded in the transaction log file; 2) A questionnaire form were filled out for each search query; 3) A structured interview, which was both audiotaped and recorded on critical incident report forms; and 4) Search queries submitted to CHESHIRE were repeated on MELVYL and the results were recorded in script files. Table 5.2 summarizes the data types, and methods of data collection and analysis.

TABLE 5.2
SUMMARY OF DATA TYPES, METHODS OF DATA GATHERING AND ANALYSIS

Data collection	*Data collection methods*	*Data analysis methods*
1. Quantitative data from transaction logs, questionnaires, and critical incident report forms 2. Qualitative data from transaction logs, structured interviews, and searches on MELVYL	1. Transaction logs 2. Questionnaire forms 3. Critical incident report forms and audiotaped structured interviews 4. Repetition of searches on MELVYL	1. Statistical analysis of quantitative data 2. Qualitative analysis of search sessions recorded on transaction logs, audiotapes, and MELVYL script files 3. Comparison of results from both analyses

5.5 Data Analysis and Evaluation Methodology

A comprehensive quantitative and qualitative analysis and evaluation was carried out on the raw data gathered through by means of transaction logs, questionnaires, and critical incident reports.

5.5.1 Quantitative Analysis and Evaluation

5.5.1.1 Analysis of Transaction Logs

The quantitative analysis of transaction logs revealed a wealth of data about the use of the CHESHIRE catalog during the experimental period. For instance, such statistical data as the number of searches conducted, number of searches that retrieved no records, number of different users participated in the experiment, number of records displayed and judged relevant (i.e., precision), and average number of terms in search statements were easily computed. Figures obtained from the quantitative analysis of transaction logs were entered into a spreadsheet package for further evaluation.

Searches that retrieved nothing (zero retrievals) as well as searches wherein users selected no clusters as being relevant were identified from the transaction logs. As discussed in Chapter IV, zero retrievals may occur due to, among others, collection failures, misspellings, and vocabulary mismatch. A search on CHESHIRE may also fail to retrieve any bibliographic records even if the search query terms match the terms in titles and subject headings of the items in the database. The way CHESHIRE works at present is such that it first retrieves some classification clusters if there exists a match between the query term(s) and titles and subject headings. If there is a match, CHESHIRE displays up to twenty clusters for user's relevance judgment. The user has to select at least one cluster as relevant in order for CHESHIRE to continue the search and retrieve individual bibliographic records from the database. The implicit assumption here is that if the user finds no cluster as being relevant, then it is highly unlikely that the document collection may have any relevant records to offer to the user.

5.5.1.2 Calculating Precision and Recall Ratios

In addition to a comprehensive analysis of search failures and zero retrievals, retrieval effectiveness of the CHESHIRE experimental online catalog was studied using precision and recall measures.

As the user's relevance judgment for each record displayed was recorded in the transaction log file, it was possible to calculate the precision ratio for each search that retrieved some records. If the record scanned was relevant, the user was simply asked to press the "relevant" key. If it was nonrelevant, hitting the carriage return key would display the next record. Thus for each and every record displayed, there was a piece of relevance judgment data attached to it in the transaction log file (see Appendix C).

Note that relevance assessments were based on retrieved references with full bibliographic information including subject headings, not the full text of documents. Relevance judgments were done by the users themselves who submitted search queries to satisfy their real information needs.

The precision value for a given search query was taken as the ratio of the number of documents judged relevant by the user over the total number of records scanned before the user either decided to quit or do a relevance feedback search. There is a slight difference between the original definition of the precision formula (given in Chapter II) and that which was used in this experiment: instead of taking the total number of retrieved records in response to a particular query, we took the total number of records scanned by the user no matter how many records the system retrieved for a particular query. For instance, if the user stopped after scanning two records and judged one of them being relevant, then the precision ratio was 50%. Precision ratios for retrievals during the relevance feedback process were calculated in the same way.

Precision ratios were calculated from the transaction logs without much difficulty. Calculating recall ratio for each search query proved to be the most challenging task as it required finding relevant documents that were not retrieved in the course of user's initial search (Blair & Maron, 1985). The procedure went as follows:

The approximate 'recall base' for each search query performed on CHESHIRE was found by repeating all search queries on both CHESHIRE and on the UC online catalog MELVYL. (The database used in CHESHIRE is a subset of the MELVYL database.) This was done for a variety of reasons. First, it is believed that repeating the same searches on CHESHIRE would somewhat facilitate the task as the researcher is familiar with both the database (i.e., records mainly about Library and Information Science) and the search system (CHESHIRE). Second, CHESHIRE and MELVYL have completely different retrieval rules. The CHESHIRE experimental online catalog utilizes probabilistic retrieval techniques along with classification clustering mechanism whereas MELVYL uses the Boolean operators AND, OR, and NOT to retrieve records from the database. It was thought that searching on two different systems for the same queries would expand the recall base by retrieving different records.

Although the database used in CHESHIRE is a subset of MELVYL, calculating the recall base for each search query proved to be a formidable task. On

MELVYL, it was not possible to restrict the retrievals to the holdings of the Library of the School of Library and Information Studies (LSL) only. Each and every record retrieved by MELVYL was checked to identify the ones located at LSL. The publication date of each retrieved record was also checked as the CHESHIRE database contains records up to the beginning of 1989. Records with publication dates 1989 or later were deleted from the MELVYL retrievals.[5] Unique records retrieved by each system (CHESHIRE and MELVYL) were identified. The total number of unique records retrieved by both systems constituted the 'recall base' for a given search.

Next, clusters and bibliographic records retrieved by the user and judged as being relevant were reviewed. User's search statement, the questionnaire form and the script of the structured interview belonging to this query were examined in order to determine the user needs and intentions that generated the query. Given the fact that the user judged the retrieved documents as being relevant in the way he or she did, given the user needs and intentions that generated the search query, the question asked was: "In addition to the records the user selected as being relevant in the original CHESHIRE search, which records would he or she have selected as being relevant had he or she seen all the records retrieved by CHESHIRE, MELVYL, or both?"

In order to answer this question, each record in the recall base was reviewed and judged as being relevant or not relevant. Thus, the total number of relevant

[5] Some records with nontraditional classification numbers (i.e., records with no topical LC class numbers) were also deleted from the MELVYL retrievals. Such records were excluded from the CHESHIRE database because classification clustering process cannot cluster records with no topical LC class numbers (e.g., records whose class numbers start with "MICROFILM").

records in the recall base for a given query was identified. Relevant records retrieved by the user were then compared with all the relevant records in the recall base. The number of relevant records retrieved by the user was divided by the total number of relevant records in the recall base to find the recall ratio for a given search query. The following example illustrates how the recall base was determined for a·given query and how recall ratio was calculated.

The search query "human-computer interaction" (query # 211) was submitted to the CHESHIRE system and two records were displayed. One was marked as being relevant. The precision ratio was computed to be 50% (1/2) from the transaction log. During the interview the user said she was looking for "anything on the topic of human-computer interaction."

Next, we searched the UC online catalog MELVYL under title keyword and subject keyword indexes (with truncations) using several synonyms. A title keyword search under "human-computer interaction" retrieved two unique items (i.e., retrieved by MELVYL, but not retrieved by CHESHIRE). Similarly, a title keyword search under "user interface" retrieved seven more unique[6] items. A subject search "human computer interaction" retrieved two more unique items whereas a subject search under "user interfaces" retrieved no unique items. Title words and subject headings in the retrieved items were examined. It was found that "man-machine interaction" has also been used in relevant records. A title keyword search under "man-machine interaction" retrieved two more unique items. One of these items was cataloged

[6] What is meant by "unique record" is that a given record has not been retrieved in earlier searches. Some records may be indexed under more than one term. Needless to say, the more searches done using synonyms, the less likely it is to come across unique records. The number given for unique records should not be confused with the total number of postings.

under a general LC subject heading **Information storage and retrieval systems.** A subject search under "man-machine systems" retrieved seven more unique items. A subject search under "human engineering" retrieved one more unique item. A subject and title keyword search under "interactive computer systems" retrieved three more unique items. All in all, 24 unique items were retrieved using MELVYL which were located at the LSL collection. So the recall base for this search query was 24. (Searches under "man-machine communication" and "system engineering" retrieved no unique items.) Table 5.3 gives the search query terms used and the type of searches conducted in order to retrieve those 24 unique items:

TABLE 5.3
SEARCHES CONDUCTED TO FIND THE RECORDS
CONSTITUTING THE RECALL BASE FOR QUERY #211

Search query	*Type of Search*	*N*
human-computer interaction	title keyword	2
user interface	title keyword	7
human-computer interaction	subject keyword	2
user interfaces	subject keyword	0
man-machine interaction	title keyword	2
man-machine systems	subject keyword	7
human engineering	subject keyword	1
interactive computer system	title/subject	3
	TOTAL	24

Note: N represents the number of "unique" records retrieved at each step.

Some of the terms used in the titles of books about human-computer interaction are as follows: "human-computer interaction," "user interface," "user/computer interface," "computer interfaces for user access," "interactive computer systems," "human-computer environment," "man-machine interaction," "machine-human interface," "interactive title computer environment," "person-computer interaction," "man-computer dialogue," "human-machine interaction," "patron interface," etc. These items were indexed under the following LC subject headings: **Human-computer interaction, User interfaces (Computer systems), Computer interfaces, Man-machine systems, Interactive computer systems, Human engineering,** and **Information storage and retrieval systems.**

After finding the number of records that made up the recall base (24), the percentage of records retrieved by CHESHIRE was calculated. CHESHIRE retrieved 13 out of 24 records that were in the recall base. As indicated earlier, CHESHIRE ranks the retrieved records in the order of their similarity to the search query and presents the top 20 records in the output list to the user. Assuming that all 20 records CHESHIRE retrieved could have been relevant yet only 13 of them actually were, the recall ratio was calculated as 65% (13/20) for this query.[7] The rest of the records retrieved by CHESHIRE were about online communities, computer output microfilm, development and testing of computer-assisted instruction, and so on.

It is worth repeating that the relevance judgments when calculating recall were made by the researcher, not by the user. Relevance judgments to calculate the recall ratios were based on the analysis of user's query statements, records retrieved and judged relevant by the user, analysis of the user's needs and intentions from the structured interviews and questionnaire forms. Contextual feedback gained from users for each query and the review of retrieved records facilitated, to a certain extent, making relevance judgments for recall calculation purposes. It was assumed

[7] Note that the denominator is 20, not 24, as CHESHIRE displays the top 20 records to the user. (The limit can be changed, however. It was set as 20 throughout the experiment.)

that, with all this feedback, objective relevance judgments reflecting actual users' decision-making processes could be made by the researcher. Nevertheless, recall ratios obtained in this study should be taken as approximate, not absolute, figures.

Once precision and recall ratios for queries retrieving some records were calculated, recall/precision graphs were plotted. Precision/recall graphs illustrate the retrieval effectiveness that users attained in CHESHIRE.

Precision and recall values were averaged over all search queries in order to find the average precision/recall ratio for CHESHIRE. "Macro evaluation" method was used to calculate average precision and recall values. This method provides both adequate comparisons for test purposes and meets the need of indicating a user-oriented view of the results (Rocchio, 1971b). It uses the average of ratios, not the average of numbers. (The latter is called "micro evaluation.") For instance, suppose that we have two search queries. The user displays 25 documents and finds 10 of them relevant in the first case. In the second case, the user displays 10 documents and finds only one relevant document. The average precision value for these two queries will be equal to 0.25 using the macro evaluation method ((10/25)+(1/10)=0.25). (Micro evaluation method, on the other hand, will give the result of 0.31 for the same queries ((25+1)/(25+35)=0.31).) As Rocchio (1971) points out, macro evaluation method is query-oriented while micro evaluation method is document-oriented. The former "represents an estimate of the worth of the system to the average user" while the latter tends to give undue weight to search queries that have many relevant documents (i.e., document-oriented) (Rocchio, 1971b; *cf.* Tague, 1981).

"Normalized" precision and recall values would have been easier to calculate, as was done in some studies (Salton, 1971). However, normalized recall does not take into account of all relevant documents in the database. Whenever the user stops scanning records, the recall value at that point is assumed as 100% even though there might be more relevant documents in the database for the same query which the user has not yet seen. The recall figures to be obtained through normalized recall may not reflect the actual performance levels. It is believed that more reliable recall values were obtained in this study. For, the comprehensive analysis of transaction logs and other records retrieved through exhaustive searches on CHESHIRE and MELVYL established the basis for the calculation of recall ratios. In addition, review of questionnaire forms and critical incidence reports provided much helpful information about users' information needs and intentions.

The users tend to be more concerned with precision values. They seem to value highly systems that could retrieve some relevant documents from the database which are not too diluted with nonrelevant ones. As long as they are able to find some relevant documents among the retrieved ones, they may not necessarily think of the fact that the system might be missing some more relevant documents. Recall values, on the other hand, are of greater concern to system designers, indexers and collection developers than users. Recall failures tend to generate much needed feedback to improve retrieval effectiveness in present document retrieval systems, although they are more difficult and time-consuming to detect and analyze.

5.5.1.3 Analysis of Questionnaire Forms and Critical Incident Report Forms

Questionnaire forms were analyzed to identify the effective and ineffective searches and to tabulate the user-designated reasons for search failures. Most useful and confusing features of the CHESHIRE experimental online catalog were also noted.

The questionnaire form included a question about the search success in terms of precision (Question #5: ". . . what percent of the sources you found were especially useful?"). This was an attempt to quantify users' perception of search success in terms of precision and to compare it with that obtained from the transaction logs.

As indicated earlier, questionnaire form and the critical incident report forms used during the structured interviews contain similar questions. Some answers from questionnaire forms were compared with the answers given in the critical incident report forms.[8] For instance, both the questionnaire and incident report forms included some questions so as to determine what the users thought of the effect of relevance feedback technique on the overall retrieval effectiveness in CHESHIRE. It is difficult to determine the exact role of relevance feedback in improving the retrieval effectiveness in CHESHIRE. Larson (1989, p.133) points out that "experience with the CHESHIRE system has indicated that the ranking mechanism is working quite well, and the top ranked clusters provide the largest numbers of relevant items."

Scripts of structured interviews were also analyzed and compared with results that were obtained from both the questionnaire forms and the transaction logs. The relationship between user-designated retrieval effectiveness and precision/recall measures was studied. The results were compared with the precision/recall ratios found for corresponding search queries recorded in transaction logs. This three-way

[8] Users' answers to some questions in the critical incident form (question #5 through #7) were marked on the form (e.g., parts that were answerable by a simple "yes" or "no"). In addition, the full structured interview was audiotaped.

comparison for some questions (e.g., search effectiveness) enabled us to investigate the causes of search failures more carefully.

5.5.2 Qualitative Analysis and Evaluation

The main objective of this study is to find out the causes of search failures in an experimental online catalog with sophisticated information retrieval capabilities. Therefore a comprehensive qualitative analysis and evaluation of the available data from transaction logs, questionnaires, and structured interview scripts was essential.

A wide variety of strategies were used to identify search failures that occurred in CHESHIRE. First, searches that retrieved no records were easily identified from the transaction logs. Analysis of the causes of zero retrieval searches showed that some searches retrieved nothing due to collection failures and misspellings whereas some others retrieved nothing because they were personal author or known-item searches, which are not supported by CHESHIRE. Yet some others failed to retrieve any records because they were out of domain search queries.

Second, search queries that retrieved some clusters but nevertheless were not pursued by the users to the end were identified. As indicated earlier, the user had to select at least one cluster record as relevant in order for the search query to retrieve bibliographic records from the database. If no cluster records were selected, then the search ended there with failure. All such failures were not necessarily due to collection failures. Some occurred because cluster records did not seem relevant while others were abandoned because of the user interface problems. False drops and stemming algorithm were also responsible for some of the cluster failures.

Analysis of search failures that occurred because no clusters were chosen as relevant required some additional work. The cluster records for such searches were not recorded in the transaction log file. These searches were redone on CHESHIRE just to record the cluster records so that the reason why the user selected no clusters could be understood.

Third, ineffective search queries were identified from the critical incident forms. Ineffective search queries were those for which users retrieved some bibliographic records but they nevertheless thought that retrieved records were not satisfactory. Precision and recall ratios for such searches were identified. Search statements, clusters, and bibliographic records were examined from transaction logs to determine what caused the search query to fail in the user's eyes.[9]

Once search failures were identified, analysis then concentrated on the causes of search failures. Again, a wide variety of methods were used: analysis of search queries (broad vs. specific), users' information needs, cluster records, bibliographic records and the subject headings attached, false drops, collection failures, precision and recall ratios, are to name but a few. Out-of-domain search queries where the user entered a search query that could not be answered using the CHESHIRE database were examined. So were personal author, known-item or call number searches.

Questionnaire forms were examined to determine what the users thought of the system's effectiveness along with user-designated reasons for failures and users' perception of search success.

[9] In fact, this was done for each and every search query that was conducted on CHESHIRE, no matter what the user said in terms of search effectiveness.

Finally, the scripts of structured interviews (i.e., incident reports) were studied. The detailed examination of critical incident reports proved useful to understand users' information needs and intentions better, which facilitated the evaluation of retrieval effectiveness performance in CHESHIRE. Other observable data about the characteristics of users and search queries were also noted.

Based on the comprehensive analysis presented above, types of failures were recorded and classified along with the cause(s) of each search failure.

5.6 Summary

The overall experiment was summarized in this chapter. Features of the system and the document collection database were explained in detail. Data gathering tools were introduced along with instructional materials that were used to recruit users to participate in the experiment. Finally, quantitative and qualitative data analysis and evaluation methodologies were explicated.

CHAPTER VI

FINDINGS

6.0 Introduction

This chapter presents the findings obtained from the experiment described in Chapter V. The first part provides descriptive statistics about users, searches and search statements captured through transaction logs, questionnaires and critical incident report forms. Qualitative analysis and evaluation of successful and unsuccessful searches is presented in the second part.

As discussed in Chapter V, an experiment was carried out in the School of Library and Information Studies of the University of California at Berkeley involving master's and doctoral students. They were given access to an experimental online catalog (CHESHIRE) for one semester (Fall 1991) and their complete interactions with the catalog were recorded in transaction logs. The purpose of the experiment was to identify search failures occurring in this experimental online catalog with a view to explicate the causes of search failures. The data analyzed below came from a variety of sources including transaction logs, questionnaire forms, and critical incident report forms.

6.1 Users

Users who agreed to participate in the experiment were asked to fill out a pre-search questionnaire and signed consent forms (see Appendix G and Appendix H). A total of 45 users participated in the experiment, 30 entering Masters-level (MLIS) students

(69.8%) and 13 Ph.D. students (30.2%) (Table 6.1).[1] Fifty-eight percent of the participating users indicated through the questionnaire that they search online catalogs daily whereas 37% used them weekly (Table 6.2). Two users (4.7%) indicated that they used online catalogs four times a year.

TABLE 6.1
USERS PARTICIPATING IN THE EXPERIMENT *(N=43)*

User Type	*N*	%
MLIS	30	69.8
Ph.D.	13	30.2
TOTAL	43	100.0

TABLE 6.2
ONLINE CATALOG USE BY PARTICIPANTS *(N=43)*

Catalog Use	*N*	%
Daily	25	58.1
Weekly	16	37.2
Four times a year	2	4.7
TOTAL	43	100.0

A large majority of participating users stated that they know how to use several application software packages such as word-processing, database management systems (DBMSs) and spreadsheets (Table 6.3). More than 80% of the users knew how to perform online searching. Almost 63% could use at least one computer programming language (e.g., BASIC, C, Pascal). Similarly, 65% of the users were familiar with electronic mail and bulletin board systems (BBSs).

[1] Two users logged onto CHESHIRE anonymously. One of the users issued the sample password ("MARY SMITH") that was provided in the instructional handout when he or she performed searches. Thus, the identities of those two users could not be verified. They were excluded from the data analysis whenever warranted.

TABLE 6.3
USERS' KNOWLEDGE OF COMPUTER SOFTWARE APPLICATIONS *(N=43)*

Knowledge of application	*N*	*%*
Word-processing	43	100.0
Online searching	35	81.4
Database Management Systems	32	74.1
Spreadsheets	31	72.1
Electronic mail & bulletin board systems	28	65.1
Programming languages	27	62.8
Other (e.g., SPSS)	3	7.0

These users performed a total of 228 search queries on CHESHIRE online catalog. Of the 228 search queries conducted throughout the experiment, 175 were (76.8%) carried out by MLIS students and 53 (23.2%) by Ph.D. students (Table 6.4). A more detailed description and analysis of searches, which is based on transaction logs, is presented in the next section.

TABLE 6.4
THE NUMBER OF CHESHIRE SEARCH QUERIES
CONDUCTED BY USER TYPE *(N=228)*

User Type	*N*	*%*
MLIS	175	76.8
Ph.D.	53	23.2
TOTAL	228	100.0

6.2 Description and Analysis of Data Obtained From Transaction Logs

6.2.1 Description and Analysis of Searches and Sessions

The average number of search queries performed by users was 5.3. Number of searches performed by each user varied a great deal with a mode of 4 searches and a minimum of one and a maximum of 21 searches. Almost 80% of the users issued between 1 and 6 search queries, which represented almost half the total. Two users issued a total of 41 search queries, 18% of all search queries submitted to CHESHIRE during the experiment. Table 6.5 gives the distribution of search queries issued by all participating users.

TABLE 6.5
DISTRIBUTION OF SEARCH QUERIES BY USERS

No. of search queries issued by users	*No. of users performing searches*	*Total number of queries (col. 1 x col. 2)*	*% distribution of total searches*
1	4	4	1.8
2	8	16	7.0
3	6	18	7.9
4	11	44	19.3
5	3	15	6.6
6	2	12	5.3
7	2	14	6.1
8	4	32	14.0
10	2	20	8.8
12	1	12	5.3
20	1	20	8.8
21	1	21	9.2
TOTAL	45[2]	228	100.1

Note: Percentage totals do not always equal 100% due to rounding

Using the definition that "[a] session is defined as one continuous period of time during which one user with a single user logon performs a search" or a number of searches (Tremain & Cooper, 1983, p.67), it was found that 228 search queries were issued in 106 search sessions, which represents just over 2 searches per session.

[2] The discrepancy in the total number of participating users is due to two anonymous users mentioned earlier (see footnote 1 in this Chapter).

Almost half the search sessions (52) consisted of a single search query. Twenty-two sessions consisted of two search queries. Table 6.6 provides the session information along with the number of search queries performed. The fourth column gives percentages of search queries performed in those sessions within the total.

TABLE 6.6
DISTRIBUTION OF SEARCH QUERIES BY SESSION

No. of search queries issued by users	*No. of sessions*	*Total queries (col.1 x col. 2)*	*% distribution of total searches*
1	52	52	22.8
2	22	44	19.3
3	11	33	14.5
4	10	40	17.5
5	5	25	11.0
6	2	12	5.3
7	2	14	6.1
8	1	8	3.5
TOTAL	106	228	100.0

Two-thirds of participating users (29) performed only one or two sessions. Number of search queries (99) carried out in those sessions constituted 43% of all search queries. Nine users performed three sessions each, which made up of almost 20% of all searches (or 44 searches). The highest number of sessions performed by any user was 10. This single user performed 9% of all searches. Table 6.7 gives the distribution of search sessions by number of users.

TABLE 6.7
DISTRIBUTION OF SEARCH SESSIONS BY USERS

No. of sessions	*No. of users performing that many sessions*	*Total no. of queries issued*	*% distribution of total searches*
1	17	49	21.5
2	12	50	21.9
3	9	44	19.3
4	4	43	18.9
5	1	10	4.4
6	1	12	5.3
10	1	20	8.8
TOTAL	45[3]	228	100.1

Note: Percentage totals do not always equal 100% due to rounding

Users spent almost 23 hours searching on CHESHIRE. The average search query took just under 6 minutes to complete. However, the time it took to complete a search query varied a great deal. More than one-third of search queries (36%) took less than one minute to complete (Table 6.8). Those search queries appear to be primarily the ones which retrieved nothing or which were discontinued by users. Ninety search queries (39.5%) took between one and eight minutes to complete. Forty (17.5%) took between nine and 16 minutes to complete. Few searches (7%) were completed in more than 17 minutes. The longest search took 35 minutes to complete.

[3] See footnote 1 in this chapter.

TABLE 6.8
DISTRIBUTION OF SEARCH QUERIES BY COMPLETION TIME

Time it took to complete a search (in minutes)	*No. of search queries*	*% distribution of total searches*
Less than 1	82	36.0
1-5[4]	40	17.6
5-9	50	21.9
9-13	26	11.4
13-17	14	6.1
17-21	8	3.5
More than 21	8	3.5
TOTAL	228	100.0

6.2.2 Description and Analysis of Search Statements

The full list of all search queries submitted to CHESHIRE during the experiment is given in Appendix I. The total number of search terms (excluding stop words) contained in 228 search queries was 802, which represents an average of 3.5 search terms per query (mode 2, median 3). The average number of stop words per search query was 1.3.

One- and two-term search queries represented 40% of all search queries (15% and 25%, respectively) (Table 6.9). Twenty-two percent of all search queries consisted of three terms. Four- and five-term search queries represented more than a quarter of all search queries (17.5% and 8.8%, respectively). Queries with six or more search terms constituted 11.4% of all search queries. The highest number of search terms in a single query was 24 (two instances), which was followed by a query

[4] Includes all search questions that took between 1:00 and 4:59 minutes to complete. The other rows should be read in the same way (i.e., "between 5-9" means all search queries that took between 5 (exactly) and 8:59 minutes).

with 19 search terms.

TABLE 6.9
THE NUMBER OF SEARCH TERMS (EXCLUDING STOP WORDS)
INCLUDED IN SEARCH QUERIES *(N=802)*

Number of	*Search queries*		*Search terms*	
search terms	*N*	*%*	*N*	*%*
0[5]	1	.4	0	0.0
1	34	14.9	34	4.2
2	57	25.0	114	14.2
3	50	21.9	150	18.7
4	40	17.5	160	19.9
5	20	8.8	100	12.5
6	11	4.8	66	8.2
7 or more	15	6.6	178	22.2
TOTAL	228	99.9	802	99.9

Note: Percentage totals do not always equal 100% due to rounding

There were a total of 85 search terms that were taken into account during the retrieval process even if they were not retrieval-worthy. That is to say, some of the search terms users entered should not have been evaluated as part of the search query. For instance, search queries "I want *information* on . . ." or "I want *books* on . . ." contain terms such as "information" and "books" that had nothing to do with the user's query. However, CHESHIRE cannot identify such terms and exclude them from the query. This requires natural language understanding capabilities in the system's part.

Whether such terms are retrieval-worthy or not depend on the context. For instance, "information" and "books" in the previous example are not retrieval worthy.

[5] One of the search queries was "the", which was a stop word.

Yet in a query like "find books on *information* policy," the term "information" is crucial for retrieval purposes whereas "books" is not.

The most frequently used unretrieval-worthy search terms (in context) were "books" (13 times), "find" (7 times), "information" and "want" (5 times each), "subject" and "search" (4 times each), "materials" (3 times), "library" and "studies" (2 times each). Inclusion of these terms (except "find" and "want") during the retrieval process may be especially undesirable for some search queries in a library and information studies database like that of CHESHIRE. For there are many sources in the CHESHIRE database with these terms in their titles and subject headings which may cause false drops when they match the users' search terms.

The CHESHIRE system only allows subject searching and does not support qualification or Boolean operators. Nevertheless, users entered queries where they asked the system to limit their searches by period (10 times), title (8 times), author (5 times), subject and language (4 times each), and form (3 times) qualifiers. Similarly, one search query contained a negation ("not dissertations"), in which the term "not" was treated simply as a stop word. Two search queries contained as part of the query truncated search terms ("librar" and "bibliograph#").

A search concept can be described by one or more terms or phrases. The analysis of search statements shows that search queries contained a total of 384 search concepts, an average of 1.7 concepts per search query. Although CHESHIRE does not support Boolean searching, users utilized (sometimes implicitly) Boolean operators to describe their search queries. Boolean AND was used in 133 search queries whereas Boolean OR was used in 41 search queries. There was only one search query with a Boolean NOT operator.

There was a total of 20 misspelled or mistyped terms in all search queries. In other words, a mere 2.5% of all search terms (20/802) entered by the users contained spelling or typographical errors. Table 6.10 lists the misspelled and miskeyed search terms.

TABLE 6.10
SPELLING AND TYPOGRAPHICAL ERRORS *(N=20)*

Search term	*Search term*	*Search term*	*Search term*
Abut	englich	marchant	seamenship
Acookery	fin	managment	suess (2)
alred	Finalnd	profesiions	systenm
basball	hitchock	policyand	vctorian
childrens	infor	salors	

There were 295 search terms that were treated as stop words (e.g., "in," "of," "on") and thus not retrieval-worthy. Some of the most frequently used stop words in search queries were: "of" (47 times), "and" (43 times), "the" (38 times), "in" (28 times), "on" (23 times) and "or" and "for" (12 times each). Table 6.11 gives the ranked list of stop words that were used five or more times in all search queries.

TABLE 6.11
RANKED LIST OF STOP WORDS USED IN SEARCH QUERIES *(N=295)*

Stop word	*N*	*Stop word*	*N*	*Stop word*	*N*
of	47	or	12	to	7
and	43	for	12	all	5
the	38	about	8	like	5
in	28	I'm	8	(all others)	51
on	23	I'd	7		

6.2.3 Analysis of Search Outcomes

The analysis of transaction logs showed that 18 out of 228 search queries (7.9%) retrieved nothing. Users selected no cluster record as being relevant in 61 (26.8%) search queries. Users proceeded to view bibliographic records in 149 search queries (65.3%). That is to say, they displayed records and selected some records as being relevant. Users performed relevance feedback searches in 91 search queries out of 149 (61.1%). In other words, almost two-thirds of searches were followed up by relevance feedback searches.

Relevance feedback searches were performed more than once for some search queries. For instance, users repeated relevance feedback searches once for 91 search queries, twice for 28 search queries, three times for 6 queries, and four times for two search queries.

The number of records users displayed and selected as relevant was recorded in transaction logs. Table 6.12 provides descriptive data about precision ratios obtained in the original search and relevance feedback (RF) iterations. Search queries that retrieved nothing due to collection failures were also included in the precision ratio calculations.

TABLE 6.12
DESCRIPTIVE STATISTICS ON NUMBER OF RECORDS SEEN AND SELECTED, AND PRECISION RATIOS[6] (SOURCE: TRANSACTION LOGS) "SELECTED" MEANS "SELECTED AS BEING RELEVANT."

Retrieval	*Total no. of records*		*Average no. of records*		*Precision*
stage	*Seen*	*Selected*	*Seen*	*Selected*	*ratio (%)*
Original	1928	369	12.6	2.4	21.49
RF (1)	1173	156	12.4	1.6	13.21
RF (2)	352	58	11.3	1.2	8.44
RF (3)	82	0	11.7	0.0	0.0
RF (4)	26	1	8.6	0.3	2.5
TOTAL	3561	584	AVERAGE PREC.		15.84

Notes: Macro evaluation method was used in the calculation of precision ratios. "RF" refers to relevance feedback cycles.

As the table shows, users displayed a total of 3,561 records and selected 584 of them as being relevant. The precision ratio was just over 20% during the original retrieval. In other words, users selected, on the average, one in five records as being relevant. It is interesting to note that as users continue their searches with relevance feedback iterations, precision ratios went down sharply, from 21.49% during the original retrieval (e.g., before relevance feedback cycles) to 13.21% in the first relevance feedback cycle to 8.44% in the second cycle. It became 0% by the time users reached the third relevance feedback cycle.

It is not possible at this stage to explain why precision ratios went down just

[6] Precision ratios given here also represent 35 "out-of-domain" search queries. An "out-of-domain" search query is such that it cannot be answered through the CHESHIRE database which concentrates on, in general, library and information studies. Thus, search queries such as "syrian asceticism," "blood transfusion," and "drugs, sex and rock and roll" are defined as "out-of-domain" search queries. Therefore precision ratios given above are somewhat lower than they actually are because of the inclusion of the out-of-domain search queries.

by looking at the precision figures obtained through transaction log data. It can only be conjectured that users who performed relevance feedback searches might have been more demanding. Or, retrieved records might have become less and less promising as the user proceeded. It has also been suggested that records retrieved during relevance feedback searches often contain a high proportion of false drops because too many nonrelevant terms are being used in the feedback process (Walker, S. & Hancock-Beaulieu, 1991, p.62).

Table 6.12 records the total and average number of records displayed and selected as being relevant in each step (i.e., original retrieval and relevance feedback iterations). However, the number of records displayed and selected vary a great deal from search to search. Table 6.13 and 6.14 provide the distribution of the number of records displayed and selected as being relevant during the original retrieval and first two relevance feedback cycles.

TABLE 6.13
THE NUMBER OF RECORDS DISPLAYED IN SEARCH QUERIES

No. of records	*Original retrieval*		*Relevance feedback cycle 1*		*Relevance feedback cycle 2*	
displayed	*N*	*%*	*N*	*%*	*N*	*%*
1-5	29	25.0	17	23.0	5	20.8
6-10	10	8.6	8	10.9	4	16.7
11-15	13	11.2	11	14.9	3	12.5
16-20	64	55.2	38	51.3	12	50.0
TOTAL	116	100.0	74	100.1	24	100.0

Note: Percentage totals do not always equal 100% due to rounding

In about a quarter of search queries, users displayed between one and five records, which seems to indicate that they were looking for a few relevant records.

More importantly, users displayed between 16 and 20 records in more than half of the search queries. In such search queries user were evidently either performing exhaustive searches or they did not find what they wanted and therefore continued to display subsequent records.

Table 6.14 shows that users selected no records as being relevant in more than a quarter of searches during original retrieval. Number of searches in which users selected no records went up considerably during the relevance feedback cycles (about 50%). An overwhelming majority of users (56.9%) selected between one and six records as being relevant. Very few users selected more than seven records as relevant.

TABLE 6.14
THE NUMBER OF RECORDS SELECTED AS RELEVANT

No. of records	*Original retrieval*		*Relevance feedback cycle 1*		*Relevance feedback cycle 2*	
selected	*N*	*%*	*N*	*%*	*N*	*%*
0	33	28.4	33	44.6	13	54.2
1-2	40	34.5	23	31.1	7	29.2
3-4	14	12.1	6	7.9	0	0.0
5-6	12	10.3	9	12.2	2	8.3
7 or more	17	14.6	3	4.1	2	8.3
TOTAL	116	99.9	74	99.9	24	100.0

Note: Percentage totals do not always equal 100% due to rounding.

6.3 Description and Analysis of Data Obtained From Questionnaires

In addition to recording users' complete interaction with CHESHIRE in transaction logs, users were asked to fill out a post-search questionnaire form for the searches they conducted. This section summarizes the data obtained through the questionnaire.

Questionnaire forms were completed for only those searches which retrieved some records. No questionnaire forms were filled out for out-of-domain search queries, either.

The self-administered questionnaire contained 10 questions (Appendix D). In addition to factual questions, it also elicited data about the users' experience with CHESHIRE. For instance, they were asked whether they found what they wanted along with the user-perceived search success rates (precision). Questions about the CHESHIRE system were also included.

Altogether 92 questionnaire forms were filled out by the users, 62 (67.4%) by MLIS students and 30 (32.6%) by Ph.D. students.

As the post-search questionnaire was applied at the end of the data collection period, there was a time lag of one week and 16 weeks between the time the users performed their searches and they answered questionnaire questions. For instance, 74% of questionnaire forms were filled out at least one month after the searches were conducted.

Users were asked (question #3) whether they found what they wanted in their first try when they performed their search queries (Table 6.15). Close to 34% said they did while the remainder were not as positive. When answers to negative categories ("no" and "not quite what I wanted") are collapsed, in 58 searches (64%) users did not find what they wanted.

Table 6.15
Answers to Question #3:
"Did you find what you wanted in your first try?" *(N=92)*

Answer	*N*	%
Yes	31	33.7
No	33	35.9
Not quite what I wanted	25	27.2
Don't remember	4	3.3
TOTAL	92	100.1

Note: Percentage totals do not always equal 100% due to rounding.

The major reasons users did not find what they wanted were that they were looking for something more specific (41.1%) and that sources retrieved did not look as helpful (30.4%) (Table 6.16).

Table 6.16
Answers to Question #4:
Why Users Did Not Find What They Wanted *(N=56)*

Reasons	*N*	%
Sources didn't look helpful	17	30.4
Looking for more specific sources	23	41.1
Looking for more general sources	1	1.8
Had to wade through a lot of useless sources	11	19.6
Had problems with CHESHIRE	3	5.4
Other	1	1.8
TOTAL	56	100.1

Note: Percentage totals do not always equal 100% due to rounding.

Users were asked their perception of search success in terms of precision (Table 6.17). In close to 14% of the cases, users found none of the sources useful, and about 27% of the cases they found less than 10% of the retrieved sources useful. In almost two-thirds of the cases (56 or 63.6%) the percentage of useful sources was found to be less than 50%. In 20 cases only (22.8%) did users found more than 50% of the retrieved sources useful.

TABLE 6.17
PERCENTAGE OF RETRIEVED SOURCES USERS FOUND USEFUL *(N=88)*

Percent Useful	*N*	%
0	12	13.6
Less than 10	24	27.3
Less than 25	17	19.3
Less than 50	15	17.0
More than 50	13	14.8
More than 75	4	4.5
More than 90	2	2.3
100	1	1.1
TOTAL	88	99.9

Note: Percentage totals do not always equal 100% due to rounding.

Figures in Table 6.17 suggest that the precision ratios as perceived by the users were quite low. Their perception of low precision ratios somewhat correspond to how they actually judged the retrieved sources (relevant or nonrelevant). As we shall see later, the average precision ratio was calculated as less than 20%, which was based on users' relevance judgments as recorded in transaction logs.

Users said they performed relevance feedback searches for close to 54% of the queries, and no relevance feedback searches for about 14% of the queries. Of those who performed relevance feedback searches, more than 50% said relevance feedback search improved the search results. More than 20% said the sources retrieved during the relevance feedback search were similar to the original retrievals. In almost 20% of the cases relevance feedback results were either less helpful or not helpful at all (Table 6.18).

TABLE 6.18
ANSWERS TO QUESTION #7:
"DID RELEVANCE FEEDBACK IMPROVE THE SEARCH RESULTS?" *(N=48)*

Relevance Feedback Results	*N*	*%*
More useful	16	33.3
Better	9	18.8
Similar	11	22.9
Less helpful	5	10.4
Not helpful at all	4	8.3
Missing	3	6.3
TOTAL	48	100.0

Users who performed relevance feedback search were asked what percent of the retrieved sources, including the ones retrieved during relevance feedback searches, they found especially useful (Table 6.19). Twelve-and-one-half percent of the users said none of the retrieved sources were useful. Almost 17% said they found less than 10% of the retrieved sources useful. Approximately one-third of the users thought that retrieved sources contained less than 25% useful sources. A further 19% indicated that less than 50% of the retrieved sources were useful. More than 20% said retrieved sources contained more than 50% useful sources.

It is interesting to note that although users thought that relevance feedback searches improved the results and retrieved additional relevant sources in more than 50% of the cases (Table 6.18), their perceptions of precision ratios for retrievals obtained after the relevance feedback searches were quite low (Table 6.19). In other words, they thought that retrievals after the relevance feedback searches contained too many nonrelevant documents, which, in fact, directly corresponds to how the users judged the records retrieved after the relevance feedback searches, as recorded in transaction logs (see Table 6.12). As we discussed earlier, the transaction logs data show that precision ratios for queries for which relevance feedback searches were performed deteriorated quickly and become zero after the third relevance feedback iteration.

TABLE 6.19
PERCENTAGE OF RETRIEVED SOURCES USERS FOUND USEFUL
AFTER RELEVANCE FEEDBACK SEARCHES *(N=48)*

Percent Useful	*N*	*%*
0	6	12.5
Less than 10	8	16.7
Less than 25	15	31.2
Less than 50	9	18.7
More than 50	8	16.7
More than 75	1	2.1
More than 90	1	2.1
TOTAL	48	100.0

The last two questions in the questionnaire form were about users' experience with the CHESHIRE experimental online catalog. They were asked to indicate what was it that they found most useful and most confusing in CHESHIRE.

6.4 Description and Analysis of Data Obtained From Critical Incident Reports

Critical incident report forms were used to gather both qualitative and quantitative information (see Appendix E and Appendix F). Users were asked to evaluate their searches from a number of different perspectives: the overall effectiveness of the search, their information needs, types of sources they were looking for, whether they carried out relevance feedback search, and so on. No critical incident report forms were filled out for searches which retrieved no clusters (zero retrievals and out-of-domain search queries). Similarly, search queries for which users selected no clusters as relevant were also excluded. The quantitative data obtained through critical incident forms are presented below.

A total of 114 critical incident report forms were filled out. Users judged their search queries as being effective in almost 70% of the cases (Table 6.20). The search outcome was found ineffective in the remainder of the cases (31.7%).

TABLE 6.20
USER-DESIGNATED SEARCH SUCCESS *(N=114)*
(SOURCE: CRITICAL INCIDENT REPORT FORMS)

Search Outcome	*N*	*%*
Effective	79	69.3
Ineffective	35	30.7
TOTAL	114	100.0

Of those users who judged their searches as being effective, about 42% said the system retrieved most of the useful sources that they needed for the search (i.e., perceived recall ratio was greater than at least 50%) whereas 15% thought otherwise. Similarly, about 37% of the respondents said more than half the sources they found using the system were useful whereas about 18% thought otherwise.

It is interesting to compare the data obtained through the transaction logs, the questionnaire and the critical incident report forms at this point. As we pointed out in the close of the previous section, users' perceptions of low precision ratios for retrieval performance were confirmed from the transaction logs. Yet, as Table 6.20 indicates, users we interviewed found the search results effective for the majority of search queries, which suggests that there is very little correspondence between the retrieval performance as measured by precision and the ways in which users evaluate the outcome of search queries as a whole. To put it differently, a user may find the search results effective even if the precision ratio for a given search query, as judged by the same user, happens to be low.

Of those users who judged the search results as being ineffective, about 83% said the system failed to retrieve most of the useful sources (i.e., perceived recall ratio was less than 50%) whereas, despite their judging the search outcome as being ineffective, about 14% did not think that the system failed. All the respondents who judged their search results as being ineffective indicated that more than half the sources they found using the system were useless.

This finding suggests that some users judged the search outcome as being ineffective when the majority of the useful records were not retrieved. That is to say, they were more concerned about retrieving most, if not all, relevant records in the database (i.e., high recall) and they attributed a considerable weight to this fact when they judged the overall outcome of the search query.

6.5 Descriptive and Comparative Analysis of Data Gathered Through All Three Data Collection Methods

In section 6.2 above, the results of search queries users performed on CHESHIRE were given. Descriptive data about searches, search sessions, and search statements are delineated and precision ratios recorded in transaction logs for 149 search queries presented in tables (section 6.2.3). The precision ratio is in itself not sufficient to determine retrieval effectiveness in document retrieval systems. In the following analysis, recall ratios for each search query are also calculated.[7] (For detailed explanation of the calculation of recall ratios, see Chapter V, section 5.5.1.2.) Precision and recall ratios obtained before relevance feedback searches are given in Tables 6.21 and 6.22, respectively. Figure 6.1 plots the precision and recall ratios for each search query on the same graph.

TABLE 6.21
PRECISION RATIOS BEFORE RELEVANCE FEEDBACK SEARCHES *(N=118)*

Ranges of precision ratio	*Number of searches having this precision value*	*%*
0 - 10%	24	20.3
11 - 20	2	1.7
21 - 30	15	12.7
31 - 40	8	6.8
41 - 50	12	10.2
51 - 60	6	5.1
61 - 70	9	7.6
71 - 80	7	5.9
81 - 90	15	12.7
91 - 100	20	16.9

Average Precision Ratio Before Relevance Feedback Searches = 50.1%

[7] The out-of-domain search queries were excluded from precision/recall calculations. The definition of the out-of-domain search query was given in footnote 6 in this Chapter.

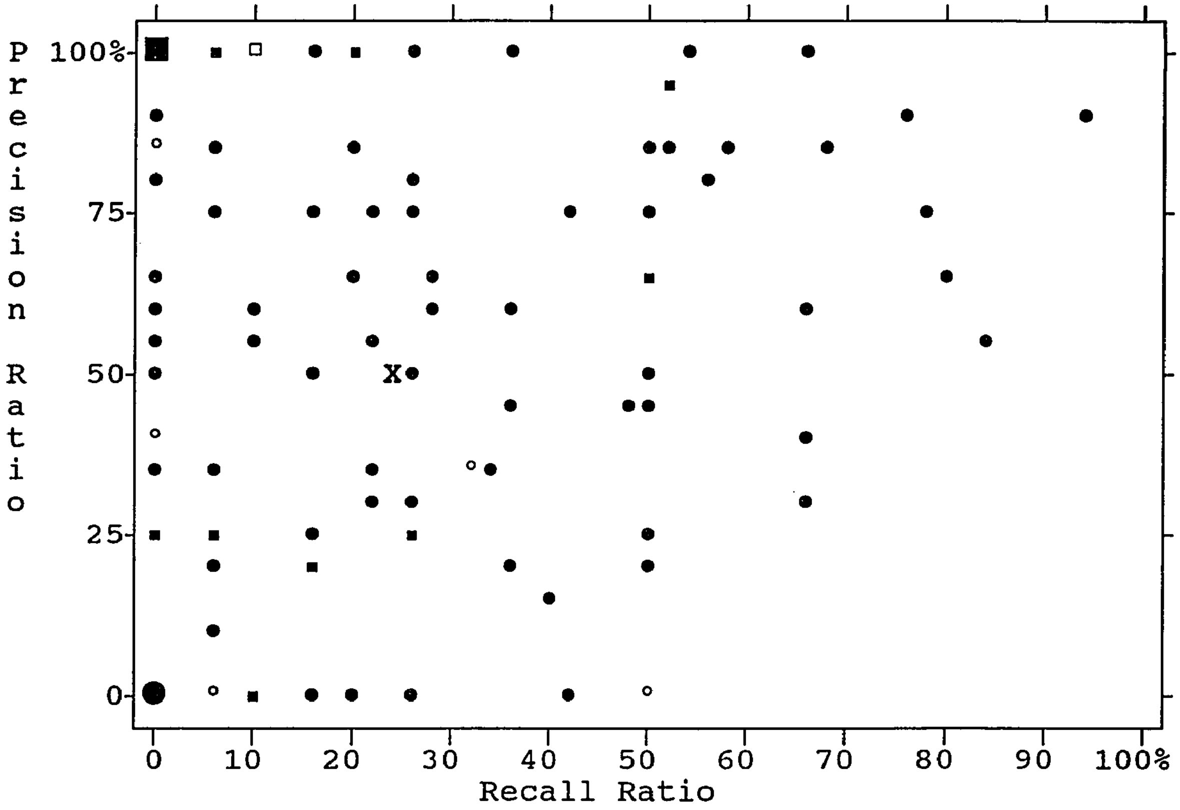
TABLE 6.22

RECA!.L RATIOS BEFORE RELEVANCE FEEDBACK SEARCHES *(N=118)*

Ranges of recall ratio	*Number of searches having this recall value*	*%*
0 - 10%	45	38.1
11 - 20	15	12.7
21 - 30	19	16.1
31 - 40	8	6.8
41 - 50	15	12.7
51 - 60	6	5.1
61 - 70	5	4.2
71 - 80	2	1.7
81 - 90	2	1.7
91 - 100	1	.8

Average Recall Ratio Before Relevance Feedback Searches = 23.6%

FIGURE 6.1
RETRIEVAL PERFORMANCE IN CHESHIRE
BEFORE RELEVANCE FEEDBACK SEARCHES *(N=118)*



Multiple occurrences: "▪"=Twice; "○"=3 times; "□"=4 times; "■"=5 times; "●"=12 times. "X" = Average precision (50.1%) and average recall (23.6%) ratios.

Precision ratios (Table 6.21) obtained before relevance feedback searches show a great deal of variation. The average precision ratio for 118 queries was 50.1%. In other words, half of the retrieved sources were judged as being relevant by the users.[8] Recall ratios, on the other hand, are concentrated in the lower end of the spectrum, indicating that majority of the searches retrieved less than half the relevant sources in the database (Table 6.22). In fact, the recall ratio was about 25% or less for almost 80% of the search queries. The average recall ratio was 23.6%. (Precision and recall ratios for all search queries are given in Appendix J.) The

[8] Users judged none of the retrieved sources as being relevant in several search queries.

figure shows that there is no strong correlation between precision and recall ratios obtained before the relevance feedback searches, and a correlation analysis confirms this (Pearson's $r=.20$, $p=.033$).

The precision and recall ratios presented in Table 6.21, Table 6.22, and Figure 6.1 exhibit some interesting findings. Several studies in the past reported that there is an inverse relationship between precision and recall measures whereas no clear pattern has emerged in this study as to the relationship between precision and recall ratios that were obtained before relevance feedback searches. The discrepancy may be due to two factors: 1) the number of observations in this study was relatively small and the findings regarding precision and recall ratios may not be definitive; and, more importantly, 2) the method of calculation of retrieval performance measures in this study differs from other studies. For instance, precision ratios reported in the past were usually based on all the retrieved records for a given query whereas in this study they were based not on all the retrieved records but only on the retrieved *and* displayed records. The precision ratio was calculated as the proportion of displayed records that were judged as being relevant to all the displayed records, which disregards the fact that there may have been more relevant records among the retrieved ones that the user chose not to display. In fact, this is one of the reasons why precision ratios for individual search queries varied a great deal in this study. Some users displayed only a few records while others displayed several.

It is also conceivable that some users may have been browsing and thus did not necessarily wish to make relevance judgments on the retrieved records, which may have suppressed the precision ratios to a certain extent.

Table 6.21, Table 6.22, and the scatter diagram (Fig. 6.1) presented above represent precision and recall ratios accomplished before relevance feedback retrieval process. As mentioned before (section 6.2.3), users continued their searches with relevance feedback iterations in 91 search queries. Tables 6.23 and 6.24 provide the precision and recall ratios obtained after relevance feedback searches along with the scatter diagram (Fig. 6.2). The average precision and recall ratios given in these figures represent the averages of ratios obtained both before and after relevance feedback searches. That is to say, if the user continued his or her search after the original retrievals and performed a relevance feedback search, the average of both results is taken. For instance, if, for a given search query, the precision ratio is 40% before the relevance feedback search and it increases to 60% after the relevance feedback search, the average precision ratio for the full search will be the average of both ratios (i.e., 50%).

TABLE 6.23
PRECISION RATIOS AFTER RELEVANCE FEEDBACK SEARCHES *(N=116)*

Ranges of precision ratio	*Number of searches having this precision value*	*%*
0 - 10%	47	40.5
11 - 20	24	20.7
21 - 30	23	19.8
31 - 40	5	4.3
41 - 50	10	8.6
51 - 60	1	.9
61 - 70	2	1.7
71 - 80	1	.9
81 - 90	3	2.6
91 - 100	0	.0

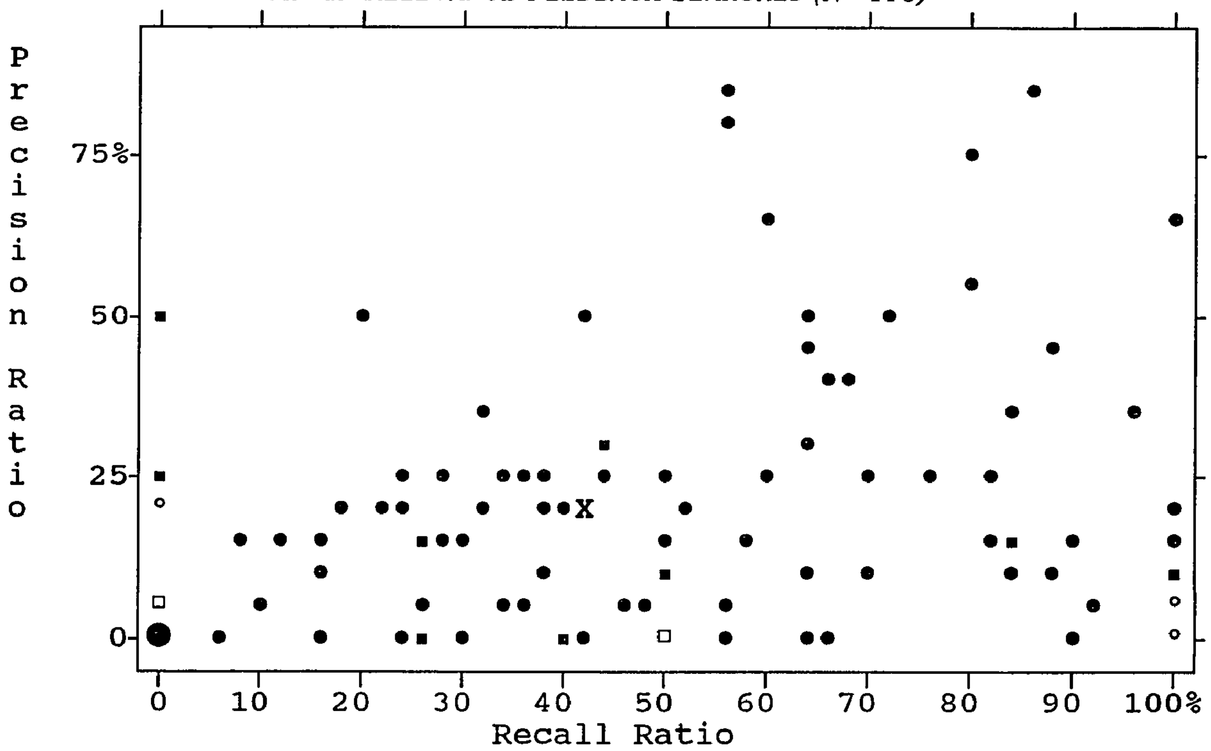
Average Precision Ratio After Relevance Feedback Searches = 18.3%

TABLE 6.24
RECALL RATIOS AFTER RELEVANCE FEEDBACK SEARCHES *(N=116)*

Ranges of recall ratio	*Number of searches having this recall value*	%
0 - 10%	23	19.8
11 - 20	5	4.3
21 - 30	14	12.1
31 - 40	9	7.8
41 - 50	19	16.4
51 - 60	7	6.0
61 - 70	11	9.5
71 - 80	2	1.7
81 - 90	12	10.3
91 - 100	14	12.1

Average Recall Ratio After Relevance Feedback Searches = 45.4%

FIGURE 6.2
RETRIEVAL PERFORMANCE IN CHESHIRE
AFTER RELEVANCE FEEDBACK SEARCHES *(N=116)*



Multiple occurrences: "▪"=Twice; "○"=3 times; "□"=4 times; "●"=10 times.
"X" = Average precision (18.3%) and average recall (45.4%) ratios.

As should be expected, as users proceeded with their searches with relevance feedback iterations, precision ratios decreased whereas recall ratios increased. That is to say, the CHESHIRE managed to retrieve additional relevant records during the relevance feedback searches that were not retrieved in the original searches. On the other hand, as the number of retrieved records increased with relevance feedback searches, so did the ratio of nonrelevant records among the retrieved ones. The average recall ratio went up almost twice from 23.6% to 45.4% whereas the average precision ratio went down from 50% to less than 20%. (See Appendix J for complete precision and recall ratios for all search queries.) Again, there is no strong correlation between precision and recall ratios (Pearson's r=-.13, p=.165) obtained after relevance feedback searches.

The above figures show that relevance feedback technique used in CHESHIRE improved the search results by retrieving additional relevant records from the database. However, there is no strong correlation between the precision ratios obtained before the relevance feedback searches and precision ratios obtained after the relevance feedback searches (Pearson's r=-.09, p=.327). Similarly, there is no strong correlation between the recall ratios obtained before the relevance feedback searches and recall ratios obtained after the relevance feedback searches (Pearson's r=.17, p=.072). However, there was a fairly high correlation between precision ratios obtained before relevance feedback searches and recall ratios obtained after the relevance feedback searches (Pearson's r=.86, p=.0005).

Notice that no observations were recorded in the upper left-hand corner of the scatter diagram, which represents the search queries with higher precision (i.e., greater than 50%) and lower recall ratios (i.e., less than 50%). This was due to two factors. First, precision ratios reported in Fig. 6.2 are the average of precision ratios obtained both before and after relevance feedback searches. For example, if the precision ratio for a given query is 60% before relevance feedback search and 20% after the relevance feedback search, the average precision ratio will be equal to 40% ((60+20)/2).

Second, the upper left-hand corner of the scatter diagram clearly indicates that it is difficult, if not impossible, to score consistently high precision and high recall ratios in online catalogs. More often than not, users have to make compromises (i.e., high recall or high precision, but not both). This finding is consistent with the probabilistic nature of the document retrieval process, too.

T-tests were performed to determine if there was any difference in average precision and recall ratios between the MLIS and Ph.D. students. MLIS students obtained slightly higher precision and recall ratios (before relevance feedback searches) than Ph.D. students did (51% vs. 47% for precision, and 25% vs. 19%, respectively). However, the difference between the two groups is not statistically significant (for precision, t=.50, p=.65; for recall, t=1.41, p=.16). Similarly, there appears to be no difference between precision and recall ratios (after relevance feedback searches) obtained by MLIS and Ph.D. students (20% vs. 14% for precision and 45% vs. 46% for recall, respectively) and the results of t-tests were not statistically significant (t=.50, p=.65 for precision; t=1.41, p=.16 for recall).

T-tests also were carried out to determine if there was any difference in average precision and recall ratios for effective and ineffective searches. As should be expected, precision and recall ratios for effective searches were different. Average precision and recall ratios for effective searches were sometimes as much as two times higher than that for ineffective ones (Table 6.25).

Table 6.25
Descriptive Statistics For Effective and Ineffective Searches

Precision (P) & recall (R) ratios before & after relevance feedback (RF) searches	*Effective searches*			*Ineffective searches*			*Total*		
	N	*Avg*	*SD*	*N*	*Avg*	*SD*	*N*	*Avg*	*SD*
P before RF	71	.64	.29	47	.29	.35	118	.51	.36
R before RF	71	.28	.25	47	.17	.20	118	.24	.24
P after RF	71	.21	.21	45	.14	.17	116	.18	.20
R after RF	71	.56	.27	45	.28	.33	116	.45	.33

The results of *t*-tests indicate that the differences in precision and recall ratios for effective and ineffective searches are all statistically significant. Average precision ratio for effective searches (before relevance feedback) was 64% as opposed to 29% for ineffective searches (t=5.93, p=.0005) whereas the average recall ratio was 28% for effective searches compared with 17% for ineffective ones (t=2.47, p=.015). Similarly, average precision ratio for effective searches (after relevance feedback) was 21% as opposed to 14% for ineffective searches (t=2.01, p=.047) while the average recall ratio was 56% for effective searches compared with 28% for ineffective ones (t=4.84, p=.0005).

The results of χ^2 test show that there was a strong relationship between the user type (MLIS vs. Ph.D.) and that of users' finding what they wanted (χ^2=6.82, df=1, p=0.009), indicating that Ph.D. students are more likely to find what they wanted in their online catalog searches than MLIS students are. MLIS students found what they wanted only in quarter of the searches they performed whereas Ph.D. students found what they wanted in more than half the searches they performed.

6.6 Multiple Linear Regression Analysis Results

A model was developed to examine the relationship between the performance of the system as measured by precision and recall and variables that defined user characteristics and users' assessment of search performance. The models were of the form

$$Y = a + b_1 x\ UTYPE + b_2 x\ CATUSE + b_3 x\ ONSRCH + b_4 x\ PLANG + $$
$$+ b_5 x\ EI + b_6 x\ FINDIT + b_7 x\ RFPERF$$

where *Y* is the dependent variable, and *UTYPE, CATUSE, ONSRCH, PLANG, EI, FINDIT* and *RFPERF* are the independent variables. Four dependent variables were used. They are:

1) ORPREC: Precision ratio obtained before relevance feedback searches

2) ORRCLL: Recall ratio obtained before relevance feedback searches

3) AVPREC: Precision ratio obtained after relevance feedback searches, and

4) AVRCLL: Recall ratio obtained after relevance feedback searches.

The seven independent variables are defined below:

1) UTYPE: User type (MLIS vs. Ph.D. students)

2) CATUSE: The frequency of online catalog use (i.e., daily, weekly)

3) ONSRCH: Knowledge of online searching

4) PLANG: Knowledge of programming languages

5) EI: Search effectiveness (i.e., whether the user found his or her search as being effective or not)

6) FINDIT: Finding what is wanted (i.e., whether the user found what he or she was looking for), and

7) RFPERF: Relevance feedback search (i.e., whether the user performed relevance feedback search).

Descriptive statistics about the independent variables are summarized in Table 6.26.

TABLE 6.26
DESCRIPTIVE STATISTICS ABOUT INDEPENDENT VARIABLES

	Frequency distribution	
Independent variable name	*1*	2
User type (1: MLIS 2: Ph.D.)	88	30
Frequency of catalog use (1: Daily 2: Weekly)	69	39
Knowledge of online searching (1: Yes 2: No)	93	25
Knowledge of programming (1: Yes 2: No)	77	41
Search effectiveness (1: Effective 2: Ineffective)	71	47
User finding what he or she wanted (1: Yes 2: No)	29	47
Performed relevance feedback search (1: Yes 2: No)	40	9

Multiple linear regression analysis was used to evaluate relationships between precision and recall ratios and seven independent variables. Table 6.27 shows the correlation between precision ratios obtained before relevance feedback searches and seven independent variables. As can be seen from the correlation coefficients, there was no strong correlation between precision and any of the independent variables. However, two independent variables had some slight correlation with the dependent variable. They were the users' perception of search effectiveness (*r*=-.41) and whether they found in the online catalog what they were looking for (*r*=-.22).

Table 6.27 Relationships of Measures That Are Correlated With ORPREC (Precision Ratio Before Relevance Feedback Searches) *(N=73)*

	UTYPE	CATUSE	ONSRCH	PLANG	EI	FINDIT	RFPERF
ORPREC	-.01	.09	-.02	-.10	-.41*	-.22*	.08
UTYPE		.15	.38*	.46*	.05	-.19	-.19*
CATUSE			.39*	-.04	.07	.02	.16
ONSRCH				.29*	-.24*	-.17	-.19
PLANG					-.07	-.20*	-.21*
EI						.52*	.11*
FINDIT							-.03
RFPERF							

*Statistically significant at or below the .05 level.

Similarly, there was no strong correlation between recall ratios obtained before relevance feedback searches and any of the independent variables (Table 6.28). However, two independent variables had some slight correlation with the dependent variable. They were the frequency of catalog use (r=.26) and the search effectiveness (r=-.22).

TABLE 6.28
RELATIONSHIPS OF MEASURES THAT ARE CORRELATED WITH
ORRCLL (RECALL RATIO BEFORE RELEVANCE FEEDBACK SEARCHES) *(N=73)*

	UTYPE	CATUSE	ONSRCH	PLANG	EI	FINDIT	RFPERF
ORRCLL	-.09	.26*	.06	.19	-.22*	-.11	.02
UTYPE		.15	.38*	.46*	.05	-.19	-.19*
CATUSE			.39*	-.04	.07	.02	.16
ONSRCH				.29*	-.24*	-.17	-.19*
PLANG					-.07	-.20*	-.21*
EI						.52*	.11
FINDIT							-.03
RFPERF							

*Statistically significant at or below the .05 level.

There was no strong correlation between precision and recall ratios obtained after the relevance feedback searches and any of the independent variables (Tables 6.29 and 6.30, respectively). Knowledge of programming was slightly correlated (r=.26) with the dependent variable AVPREC, precision ratios obtained after the relevance feedback searches (Table 6.29). Search effectiveness had some slight correlation (r=-.30) with the dependent variable ORRCLL, recall ratios obtained after the relevance feedback searches (Table 6.30).

TABLE 6.29
RELATIONSHIPS OF MEASURES THAT ARE CORRELATED WITH
AVPREC (PRECISION RATIO AFTER RELEVANCE FEEDBACK SEARCHES) *(N=71)*

	UTYPE	CATUSE	ONSRCH	PLANG	EI	FINDIT	RFPERF
AVPREC	-.10	.14	-.01	.26*	-.19	-.10	-.17
UTYPE		.19	.40*	.43*	.01	-.22*	-.17
CATUSE			.39*	-.01	.02	.01	.15
ONSRCH				.32*	-.24*	.17	-.20*
PLANG					-.14	-.25*	-.18
EI						.51*	.15
FINDIT							-.01
RFPERF							

*Statistically significant at or below the .05 level.

TABLE 6.30
RELATIONSHIPS OF MEASURES THAT ARE CORRELATED WITH
AVRCLL (RECALL RATIO AFTER RELEVANCE FEEDBACK SEARCHES) *(N=71)*

	UTYPE	CATUSE	ONSRCH	PLANG	EI	FINDIT	RFPERF
AVRCLL	.08	.03	-.02	.00	-.30*	-.11	-.02
UTYPE		.19	.40*	.43*	.01	-.22*	-.17
CATUSE			.39*	-.01	.10	.04	.15
ONSRCH				.32*	-.24*	-.17	-.20*
PLANG					-.14	-.25*	-.18
EI						.51*	.15
FINDIT							-.01
RFPERF							

*Statistically significant at or below the .05 level.

No strong intercorrelations were observed amongst the independent variables, either. However, search effectiveness was moderately intercorrelated in all cases with whether the user found what he or she wanted in the online catalog search, indicating that users who found what they wanted are more likely to judge their searches as being effective.

These findings suggest that users' judgment of the effectiveness of their searches turned out to be the most significant factor in predicting precision and recall ratios. Search effectiveness was negatively correlated, although not strongly, with all but one (precision obtained after relevance feedback searches) dependent variables, indicating that those who judged their searches as being ineffective are less likely to have higher precision and recall values.

Nonetheless, it should be emphasized that correlations between the dependent and independent variables were not strong. As can be seen from the multiple linear regression analysis results (Table 6.31), all seven independent variables combined explain only about 25% of the observed variability in precision and recall ratios. Almost 75% of the observed variability in precision and recall ratios remain unexplained.

TABLE 6.31
SUMMARY OF MULTIPLE LINEAR REGRESSION ANALYSIS

Dependent variable name	*N*	r^2	*F*	*Significance of F*
Precision ratio before rel. fdbck. search (ORPREC)	73	.25	3.03	.008
Recall ratio before rel. fdbck. search (ORRCLL)	73	.24	2.91	.010
Precision ratio after rel. fdbck. search (AVPREC)	71	.26	3.21	.006
Recall ratio after rel. fdbck. search (AVRCLL)	71	.14	1.46	.196

The results of the multiple linear regression analysis may not be definitive as the sample size was small. Nonetheless, the results indicate that user characteristics (i.e., frequency of online catalog use, knowledge of online searching and programming languages) and users' own assessment of search performance (i.e., search effectiveness, finding what is wanted) are not adequate measures to predict the system performance as measured by precision and recall ratios. To put it somewhat differently, as a considerable percentages of observed variabilities in precision and recall ratios remain unexplained, the regression model developed earlier cannot reliably explain the correlation between precision and recall ratios and the measures studied here. Therefore, it is difficult to use this model to examine the relationship between the system performance as measured by precision and recall ratios and variables defining user characteristics and users' judgments of search effectiveness.

6.7 Summary

Quantitative data collected by means of transaction logs, questionnaires and critical incident report forms were summarized in this chapter. Descriptive statistics on participating users, search queries, search outcomes in terms of number of records seen by the users and selected as being relevant, users' assessments of search results were given. Retrieval performance of CHESHIRE as measured by precision and recall was also discussed along with the results of a regression model.

The quantitative analysis of retrieval performance of the system was based on a total of 228 queries submitted by the MLIS and doctoral students of the School of Library and Information Studies at the University of California at Berkeley. An average search query took just under six minutes to complete, although one-third of the queries submitted took less than one minute due to search failures (i.e., zero retrievals). On average, a search statement contained 3.5 terms, which is relatively higher than that submitted to second generation online catalogs. This suggests that users may have felt less constrained to describe their requests to an online catalog with a natural language user interface. Misspelling and typographical errors were relatively few; only 2.5% of all search terms contained such errors. Some queries also contained terms that were useless from the retrieval point of view ("I want information on . . .", "please find some books on. . .").

Although users displayed between 16 and 20 records in more than half the searches, they selected only between 0 and 4 records as relevant in more than 75% of all search queries. In users' view, two-thirds of the searches contained less than 25% of the useful sources. The main reason for this was that they were looking for more specific sources and the retrieved sources did not look helpful. The number of records selected further declined as users performed relevance feedback searches. Yet they felt that they retrieved additional useful sources during relevance feedback searches in more than 50% of the cases. Although precision ratios obtained from transaction logs were low, users who were interviewed judged two-thirds of the search queries as being effective. This finding suggests that precision was not the only criterion in their assessments of search effectiveness.

The average precision ratio before relevance feedback searches was about 50% whereas the average recall ratio was about 24%. In other words, one out of every two records retrieved was judged as being relevant by the users. Yet the system retrieved only one out of every four relevant documents in the database. The average precision ratio after relevance feedback searches went down to less than 20% whereas the average recall ratio rose to 45%. In other words, an almost two-fold increase was observed in the recall ratios after relevance feedback searches whereas the average precision ratio declined from 50% to 18%. Although relevance feedback technique helped retrieve additional relevant documents from the database after each iteration, thereby increasing the average recall ratio up to 45%, the average precision ratio went down drastically after each relevance feedback cycle.

T-tests showed that MLIS students obtained slightly higher precision and recall ratios than Ph.D. students, although the difference was statistically insignificant. Yet a χ^2 test indicates that Ph.D. students were more likely to find what they wanted in the online catalog than MLIS students and the difference was statistically significant. MLIS students found what they wanted in less than a quarter of the searches whereas Ph.D. students found what they wanted in more than half the searches. As should be expected, precision and recall ratios for effective searches were significantly higher than for ineffective ones.

Finally, a multiple linear regression analysis, which aimed to examine the relationship between CHESHIRE's retrieval performance as measured by precision and recall ratios and the users' judgment of the system's search performance, found that users' assessments of the effectiveness of their searches was the most significant factor in explaining precision and recall ratios. However, there was no strong correlation between precision and recall measures and user characteristics, and users' assessment of retrieval performance. It was concluded that the regression model developed cannot be used to examine the relationship between these measures as all seven independent variables combined explained only a quarter of the observed variability in precision and recall ratios.

It must be stated that these results were obtained without an experimental design with a control and experimental group (see Chapter V) and thus the results may be biased. Nevertheless, findings we obtained and the conclusion we reached regarding the relationship between performance measures and users' assessments of search effectiveness are commensurate with findings obtained in other studies. For instance, although she did not study recall, Su (1992) found that precision is not correlated with search success. However, more research is needed to validate the findings obtained in this study over larger populations of search queries.

CHAPTER VII

ANALYSIS OF RETRIEVAL PERFORMANCE IN CHESHIRE

7.0 Introduction

Quantitative findings regarding CHESHIRE's retrieval performance as determined by precision and recall measures were discussed in Chapter VI. One of the primary objectives of the present study is to examine the retrieval performance of CHESHIRE more comprehensively. The analysis of CHESHIRE's retrieval effectiveness to be presented in this chapter is based on the results obtained from transaction logs, questionnaire forms and structured interviews with the participating users.

7.1 Determining Retrieval Performance

It was noted in earlier chapters that no single measure of retrieval performance is sufficient to determine the retrieval performance of an online catalog. The results of multiple linear regression analysis that we discussed in Chapter VI showed that there was no strong correlation between traditional retrieval performance measures (precision and recall) and user characteristics and users' assessment of search performance. Furthermore, the performance of the retrieval system for each query as measured by precision and recall ratios varied a great deal. This suggests that a qualitative analysis of search effectiveness for each query will be helpful to explain the variations in the retrieval performance of the system.

The qualitative analysis of retrieval performance in CHESHIRE presented below makes use of several pieces of data that we gathered by means of transaction logs, questionnaires, structured interviews, and comprehensive searches.

As discussed in detail in Chapter V, a pre-search questionnaire was filled out by the participating users which included questions on user type, frequency of catalog use, and the knowledge of computer software packages and online searching. This data was presented in the previous chapter. Then, users performed catalog searches throughout the data collection period. All the searches they performed were recorded in transaction logs along with users' relevance judgments on retrieved records. Precision ratios were calculated from the data recorded in transaction logs. Next, we performed successive searches on CHESHIRE and MELVYL® in order to determine the recall base for each search query submitted to the system. Recall ratios were calculated from this data.

Once the data collection period was over, participating users were asked to fill out a questionnaire form for search queries that they performed on the system. Data on whether users found what they wanted in the catalog along with their perceived search success in terms of precision and recall, both before and after relevance feedback searches, came from the questionnaire. Some of the quantitative findings obtained through the questionnaire were presented in Chapter VI.

After the users filled out the questionnaire, we interviewed them so as to find out about their views of the retrieval performance of the system and audiotaped their comments. Thus, the users' assessments of retrieval effectiveness came from questionnaires and structured interview results.

The overall retrieval performance of the system for a given search query was then determined on the basis of three pieces of information. A search query was considered as being effective if: a) the user found what he or she was looking for (as recorded in the questionnaire form); b) the user judged the search results as being effective (as recorded in the critical incident form and the script of the structured interview); and c) precision and recall ratios were commensurate with, to a certain extent, the user's judgment. However, it is difficult to come up with a formula that would indicate to what extent each piece of information has contributed to the final decision as to the effectiveness or ineffectiveness of a given search query, although, it should be emphasized, users' own assessments of their search queries were weighted more heavily. To put it somewhat differently, the retrieval performance of CHESHIRE for a given search query was judged as being effective unless there was a considerable discrepancy that was unaccounted for between the answers supplied by the user and the precision and recall ratios.

In the qualitative analysis, "out-of-domain" search queries and search queries that retrieved nothing (i.e., zero retrievals) were identified from the transaction logs. Searches that retrieved nothing were later analyzed to determine the causes of failure by examining the search statement and collection make-up. Similarly, queries in which users selected no clusters as being relevant were identified from the logs and all such queries were repeated on CHESHIRE in order to determine the causes of such incidents. Search queries for which precision and recall ratios were available were also analyzed to determine the retrieval effectiveness and to corroborate the findings obtained from questionnaires, critical incident report forms, and structured interview scripts.

7.2 Retrieval Performance in CHESHIRE

The retrieval performance of CHESHIRE as measured by traditional precision and recall ratios was given in Chapter VI. On the average, half the records CHESHIRE retrieved were judged as being relevant by the users (precision) before relevance feedback searches. On the other hand, CHESHIRE retrieved only about 25% of all the relevant documents in the database (recall). As should be expected, precision ratios went down (18%) while recall ratios increased (45%) as users performed relevance feedback searches. To put it differently, successive relevance feedback searches improved the recall ratios to a point where almost half the relevant records in the database were retrieved.

What follows is a comprehensive analysis of retrieval performance in CHESHIRE, which incorporates not only precision and recall measures but also feedback gathered from the users on their assessments of search effectiveness. The analysis consists of two parts: 1) analysis of search failures that occurred in CHESHIRE; and 2) examination of search effectiveness in CHESHIRE. The first part concentrates on the analysis of the causes of search failures in CHESHIRE, the main theme of this dissertation. In the second part, we will emphasize CHESHIRE's strengths as a third generation online catalog and compare it with other catalogs.

7.2.1 Analysis of Causes of Search Failures in CHESHIRE

Altogether users performed 228 search queries on CHESHIRE. A total of 107 search queries (46.9%) failed due to a wide variety of reasons including collection failures. Table 7.1 summarizes the causes of search failures for those 107 search queries.

TABLE 7.1 CAUSES OF SEARCH FAILURES *(N=107)*

Causes of search failures	*N*	*%*
Collection failure	42	39.3
User interface problem	13	12.1
Search statement	11	10.3
Known-item search	11	10.3
Cluster failures	8	7.5
Library of Congress Subject Headings	5	4.7
Stemming algorithm	4	3.7
No apparent reason	3	2.8
Specific query	2	1.9
Cluster selection	2	1.9
Communication problem	2	1.9
Scope	2	1.9
False drops	1	0.9
Call number search	1	0.9
TOTAL	107	100.1

Notes: (1) Percentage totals do not all equal to 100% due to rounding.
(2) Definitions of the categories of search failures can be found in Chapter IV.

As can be seen from Table 7.1, collection failure was the primary cause of almost 40% of all unsuccessful search queries. This was followed by the problems that users experienced with CHESHIRE's user interface (12.1%). Flaws in the search statements caused failures in more than 10% of search queries. Another 10% of the queries failed because some users tried to perform known-item searches, which is not supported by CHESHIRE. (The online catalog supports subject searching only.) Users found the retrieved clusters nonrelevant for eight (7.5%) search queries. They discontinued their searches upon seeing the retrieved, but not-so-promising, cluster records. The Library of Congress Subject Headings (LCSH) was the primary cause of almost 5% of all search failures. CHESHIRE's stemming algorithm was the cause of four (3.7%) search failures. A total of thirteen (12.1%) search queries failed due to, among others, telecommunication (telnet) problems, cluster selection, and false drops.

The detailed findings with regards to each type of search failure are presented below.

7.2.1.1 Analysis of Collection Failures

Some 42 search queries (39.3%) failed as there were no relevant sources in the database. (See Chapter V for a detailed description of how relevant sources were found.) This type of failure is commonly called "collection failure" and it constitutes, generally speaking, a considerable percentage of all search failures in online catalogs.

More than two-thirds (30 out of 42) of all collection failures in this study were coupled with specific search queries. For instance, the Library School Library (LSL) collection simply lacked sources that could have satisfied specific search queries such as "virtual reality cyberspace" (#65)[1], "classification of materials on gay and lesbian studies" (#222), "hypermedia" (#20), "hypertext" (#21), "indexes for information resources on or in networks like Internet and Bitnet," (#15 and #16) "minitel" (#108), "novell" (#125), "cheshire" (#191), "xerox windows" (#80), and "project mercury" (#186).

Some of the collection failures cited above occurred because the search topics were relatively new. Monographic literature on, say, "virtual reality cyberspace,"

[1] Numbers given in parentheses in this Chapter refer to the search query numbers. The text of each search query submitted to the system can be found in Appendix I using the query number (i.e., #65). The retrieval performance of the system for each query can also be found in Appendix J using the query number.

"indexes for information resources on or in networks like Internet and Bitnet," and "project mercury" came into being very recently and the database contains records only up to 1989. Therefore search queries for those relatively new topics failed without retrieving any promising bibliographic records. Similarly, despite the fact that the system retrieved promising items, search queries for the most recent publications (i.e., published since 1989) were judged as being ineffective due to collection failures (i.e., #66 and #73). Some of the search queries were very specific in nature and the database lacked specific sources to satisfy such queries (i.e., #191, #62, #175, #56). The literature simply did not exist in published form for some other search queries (#141, #193). Four search queries in this group retrieved nothing at all (zero retrievals) due to collection failures (#13, #20, #21, #108).

Out-of-domain search queries are not treated as collection failures in this study. Several users were apparently unaware of the domain of the CHESHIRE database and issued out-of-domain search queries on, say, ancient Chinese poetry, romance novels, and Alfred Hitchcock films.

7.2.1.2 Analysis of the Causes of User Interface Problems

CHESHIRE is an experimental online catalog that has been made accessible to the users who participated in this study. It was developed by Larson (1989, 1991a, 1992) for his theoretical research on advanced information retrieval techniques. It is fair to suggest that more emphasis has been given to its functionality than its user interface during the design and implementation stages. Yet the user interface was the primary cause of only 13 (12.1%) unsuccessful search queries. In eight cases the users indicated that they simply did not know how to use the system or how to proceed once they entered their search queries (#27, #29, #40, #41, #53, #72, #189, #190).

Some "got lost" and "couldn't tell from the interface how to select an item." The user interface was "just too foggy" for some others and it "didn't give enough user clues." One user was desperately seeking help (#43, #44) while two others could not figure out how to quit the system (#79, #94). Their help and quit requests were treated as legitimate search queries by the natural language user interface. Another user experienced problems when editing her search statement and could not backspace to previous lines (#98).

Of 13 search queries which failed due to user interface problems, seven occurred when users attempted to search CHESHIRE for the first time. This would seem to suggest that some first time users were not well-served by the user interface. It should also be mentioned that CHESHIRE has no help screens of any significance to guide the novice users.

7.2.1.3 Analysis of Failures Caused by Search Statements

A total of 11 (10.3%) search queries failed due to major flaws in the users' search statements. Vocabulary problems (#28, #32, #38, #178, #184, #194, #223, #225), incomplete search queries (#45), misspellings (#227), truncated search terms (#184), and indecipherable query statements (#64) are classified under this group.

Several factors caused vocabulary problems: some search queries contained abbreviated or truncated terms while others were broad, did not describe the user's information need adequately, and contained search terms that were not retrieval worthy. For instance, the abbreviated search term "cip" retrieved a few records but missed many that were listed under the spelled out form ("Cataloging-in-Publication"). Similarly, some queries contained truncated search terms ("librar"), which was not supported by the system. The stemming algorithm failed to recognize such terms because they were not listed in the system's dictionary in that form and thus ignored during the retrieval.

Some search queries did not describe users' real information needs. For instance, one user entered "freedom of information" (#28) as her search query even though she was looking for information on "national security issues and classification of documents."

Some users qualified their search queries and entered phrases such as "subject search" or "title search" as a part of their complete search statements. However, such terms were treated as legitimate search terms and treated as such, thereby causing some false drops. Some others simply described what they wanted (i.e., "alternatives to traditional subject headings" (#223)) and expected the system to handle the rest. However, the system cannot handle such queries successfully as it has no natural language understanding capabilities.

7.2.1.4 Analysis of the Causes of Known-item Search Failures

It appears that a few users were unaware of the fact that CHESHIRE only allows subject searching. Four users entered a total of 11 known-item search queries: five personal author (#90, #91, #92, #93, #97), and six title searches (three for book titles (#200, #201, #204), and three for periodical titles (#78, #202, #203)). All eleven known-item search queries failed one way or the other.

The stemming algorithm did not recognize personal author names (i.e., "marcia tuttle," "katz," and "patrick wilson") as legitimate query terms because author names are not taken into account during the retrieval. One of the personal author searches was in fact a factual query: "how many books by patrick wilson does the library have?" As the system performs no semantic analysis on the search statement, this query could not be satisfied.

The rest of the known-item searches were for periodical and monographic titles (three search queries each). Needless to say, the system treated all six searches as subject searches and retrieved some items accordingly, although not necessarily the ones sought by the users.

7.2.1.5 Analysis of the Causes of Cluster Failures

As pointed out earlier, CHESHIRE expands the users' original queries on the basis of classification clustering process where users are asked to indicate whether retrieved cluster records seem relevant or not. The query expansion is largely based on the title words, LC subject headings, and classification numbers present in clusters judged as relevant by the users. However, if, for some reason, the user happens to select no cluster as relevant, the search would end without retrieving bibliographic records.

As briefly explained in Chapter V, when the user selects no cluster as relevant, cluster records do not get recorded in the transaction files. In order to see which clusters the users did not like, the search queries were re-created just to record the clusters in the transaction log file. Search queries were re-entered exactly as they were and then displayed one by one. In order to record the clusters in the transaction file, we selected the first cluster in each search as relevant and then quit. We repeated this process for all queries that retrieved some clusters but that none of them was chosen by the user as relevant.

The idea was to see the clusters which the user judged nonrelevant, thereby ascertaining how efficiently the classification clustering process in CHESHIRE brings the relevant clusters (i.e., LCSH and class numbers) together. This process also allowed us to record the bibliographic records as if the user had selected the very first cluster as relevant, which reflects the fact that those would be the kinds of records the user would have retrieved. One of the shortcomings in this process was that we did not know how many clusters the user had seen and decided that it was not worth pursuing his or her search further. In other words, the user may have abandoned the search after seeing only one cluster or all 20 clusters. We simply do not have this information recorded in transaction logs. In fact, some of the search queries suggest that the user, for instance, saw his or her spelling mistake, or wanted to broaden or narrow the query and quickly abandoned the search and re-issued a similar one. We did not classify such queries as cluster failures.

In this section cluster failures that were primarily caused by retrieval of nonrelevant (judged by the user) cluster records were analyzed.

There were eight (7.5%) search queries that were abandoned by the users because the system failed to retrieve relevant clusters (#3, #18, #68, #70, #151, #166, #172, #173). One user issued a search query on "a general history of the Library of Congress" (#151) but did not like the clusters retrieved by the system. In fact, none of the 20 cluster records included the specific LC subject heading **Library of Congress** in it; they were all general. It appears that CHESHIRE's weighting formula underweighted the most important words ("Library" and "Congress") in the search query. The user re-issued her query as "library of congress" (#152) and retrieved relevant clusters and bibliographic records.

Some users found the retrieved clusters not specific enough and thus selected none as relevant. For instance, one user was looking for collection development in law libraries. He repeated his search query twice ("law libraries -- collection development from 1935" (#68), "collection development law libraries only" (#70)). The most promising two clusters he retrieved in his first search were **Collection development -- Libraries** and **Law libraries**. As the user was not satisfied with these somewhat general clusters, he re-issued his query by adding the word "only" after "collection development law libraries." CHESHIRE retrieved, among others, the same two clusters again. Eventually, the user gave up, thinking that there was nothing in the collection that could answer his query.

A similar situation occurred when another user was looking for reference sources in art. Again, the user repeated his query twice ("library reference material on art" (#172), "library resource materials on art" (#173). The former retrieved a few clusters on reference services and reference books. Yet none of them was specific enough to be selected as relevant by the user. The latter was worse: it retrieved nothing whatsoever on either reference sources or reference services. The choice of the term "resource" in the query may have affected the retrieval results negatively because it is not interchangeable with "reference."

Queries on "information policy" (#166) and "cost-effectiveness of library services" (#3) also failed to retrieve relevant clusters. The best cluster CHESHIRE retrieved for the former query was **Information services**. There was no specific cluster on "information policy." The second one was more specific. None of the clusters retrieved had anything to do with the user's query.

Cluster failures summarized above include those search queries for which CHESHIRE failed to retrieve any relevant clusters. We did not consider such cases as cluster failures where users entered out-of-domain search queries and then, upon seeing unpromising clusters or failing to retrieve anything at all, did not want to continue their searches. For instance, users abandoned 17 out-of-domain search queries without selecting any cluster as relevant.

Similarly, search queries that were abandoned before selecting any clusters as relevant because the users simply wanted to revise their queries and resubmit them were not considered as cluster failures, either. There were 11 such search queries.

CHESHIRE's classification clustering mechanism usually helps users get close by to their subject areas by way of displaying promising subject headings which users are likely to find relevant. However, there appears to be some cases where the classification clustering mechanism did not help users to identify their subject areas.

This is simply what happened in majority of the search queries summarized above. Basically, the user found none of the clusters promising or specific enough and did not select any. However, it is highly likely that had the user selected at least one cluster as relevant CHESHIRE would have retrieved some relevant records. A similar case occurred for a query on "library tours" (#113) for which CHESHIRE picked up some general clusters on libraries and some others on "tours, France"! Again, as the user selected no clusters as relevant the search failed. (Selecting some general clusters as relevant just because there are no specific ones available also may cause failures, especially for very specific queries. For instance, both queries on "indexes for information resources on or in networks like Internet and Bitnet" (#15 and #16) failed to retrieve any relevant bibliographic records in spite of the fact that the user judged some clusters as relevant.)

7.2.1.6 Analysis of Search Failures Caused by the Library of Congress Subject Headings

Subject headings assigned to bibliographic records in the CHESHIRE database were taken from the Library of Congress Subject Headings (LCSH) vocabulary. The terminology used in headings and specificity or exhaustivity of assigned subject headings were determined by LCSH.

LC subject headings assigned to documents caused a total of five (4.7%) search failures in our study (#127, #128, #174, #181, #192). Retrieved clusters (hence assigned subject headings) were fairly broad in all but one cases. LC subject headings presented in those clusters were not specific enough to describe users' search topics. Yet users felt that they were compelled to select broad LC subject headings as relevant in order to retrieve bibliographic records.

Three search queries on censorship of children's literature (#127, #128, #181) failed because LC subject headings provided were not specific enough. CHESHIRE retrieved some general clusters on censorship for the first search query ("censorship of children's books"). Yet the most specific cluster on censorship of children's literature was not displayed. CHESHIRE successfully retrieved two titles relevant to the user's query. Yet both titles were cataloged under general LC subject headings **Censorship** and **Censorship -- United States.**

The second query was worded slightly differently ("censorship of children's literature"). The match between the query terms and that of LC subject headings were better for this query. CHESHIRE retrieved three relevant sources that were cataloged under the specific LC subject heading **Children's literature -- Censorship.** On the other hand, this query missed two relevant records cataloged under the broader LC subject headings given above.

The third search query retrieved only one relevant source on censorship of children's literature. Retrieved sources were mostly on either children's literature or censorship, but not necessarily the combination of the two. The user's selection of broad clusters on censorship as relevant helped very little in terms of CHESHIRE's ability to pinpoint more specific items in the database.

Another search query on "children's book reviewing" (#174) also failed because there was no specific LC subject heading provided. The user was looking for theoretical works on children's book reviewing. The majority of the sources CHESHIRE retrieved were on the history of book reviewing and book reviews and children's literature in general.

The last search query that failed because of the lack of specific LC subject headings was on "relevance" (#192). The user was trying to find sources on relevance feedback in information retrieval systems. None of the sources CHESHIRE retrieved was assigned "relevance" as a specific LC subject heading. Rather, broader LC subject headings of **Information storage and retrieval systems -- Testing** and **Information storage and retrieval systems -- Evaluation** were assigned to relevant titles.

The lack of specific LC subject headings appears to have affected the outcome of some other search queries in an unfavorable way, although such searches failed for other reasons. For instance, one user was looking for sources on "letterpress printing" (#29) and there was no specific LC subject heading, which caused the system to retrieve some general clusters on private presses and little presses. One other user was interested in "greek typefaces in Paris in fifteenth and sixteenth centuries" (#180). None of the assigned LC subject headings was that specific.

7.2.1.7 Analysis of Search Failures Caused by CHESHIRE's Stemming Algorithm

The function of a stemming algorithm is to reduce the search terms in the user's query to their root forms so that search terms would match more records in the database, thereby increasing the recall rate. Reducing the search terms to their roots also means that less storage space will be needed to accommodate the dictionary of all the terms occurring in the document database.

Stemming algorithm used to parse the query terms in CHESHIRE caused four (3.7%) search queries to fail completely (#74, #75, #118, #209).

The search query on "C" (#74) retrieved nothing because the stemming algorithm disregarded the term completely. The user revised his query and entered "programming C" (#75). However, revision of the query did not improve the search query very much because the algorithm recognized only the first term and disregarded "C" again. This caused CHESHIRE to retrieve several clusters on programming, but not necessarily C programming. The user abandoned the search upon not finding any relevant clusters.

The other two stemming failures were also similar. The search queries on "r&d" (#118) and "e-journal" (#209) retrieved nothing because the algorithm failed to recognize the abbreviated terms "r&d" and "e". ("r&d" for "research and development", and "e-journal" for "electronic journal".)

There were a few more search queries which the stemming algorithm failed to evaluate properly, although those search queries failed due to some other reasons (e.g., collection failures, user interface failures). A personal author search query for "marcia tuttle" (#90) was reduced to "marc" by the stemming algorithm, which caused the system to pick up several sources on Machine Readable Cataloging (MARC). Similarly, "novell" (#125) (a local area network brand name) was reduced to "novel", which resulted in the retrieval of such clusters as Santa Maria *Novella* Dominican Monastery, American fiction, and the like. The system would have retrieved bibliographic records on, among others, Victorian *novelists* if the search was not abandoned.

7.2.1.8 Analysis of Search Failures Caused by No Apparent Reason

Three (2.8%) search queries failed due to no apparent reason (#157, #176, #219). Retrieved records for a search query on "storytelling" (#176) were all relevant. Relevance feedback search results were also relevant. Yet the user judged this search as ineffective. She said she was asked a reference question in her job about storytelling, but she could not remember the details very well.

A search query on the "history of Library of Congress Subject Headings" (#219) was abandoned by the user although retrieved clusters were relevant. During the interview the user did not recall performing this search. Similarly, a search query on the "history of printing" (#157) was also abandoned by the user even though the system retrieved some excellent clusters.

It is difficult to classify these three searches under a certain category of search failure. Clearly, users had some difficulty recalling their search queries. However, none of the queries was judged ineffective or abandoned because of system problems.

7.2.1.9 Analysis of Search Failures Caused by Specific Queries

Although users submitted several specific search queries to the system, only two (1.9%) search queries failed primarily due to the specificity of search queries. In fact, some search queries submitted to the system were formulated as "research questions" rather than online catalog search requests. For instance, one user was trying to find some sources to support his thesis that "law librarianship is a product of the 1929 stock market crash." He was also interested in if "the federal depository legislation of the early '30s. . . had a major impact on law libraries." His search query was relatively broad ("history of law libraries, history of federal depositories, personal narratives of law librarians, law libraries") (#7). Yet when the system retrieved some general sources on law libraries, he selected none of them as relevant. He was after specific sources that could prove his thesis. Such sources simply did not exist in the database. It is likely that user's query can be answered only after an extensive study of the literature. Yet he was expecting to find specific titles referring directly to his research question.

Another user was looking for information "on the public image of librarians through history" (#162). The system retrieved, among others, two general titles on her topic. Yet none of the items were as specific as the user would have liked.

Hence she selected none as relevant.

7.2.1.10 Analysis of Search Failures Caused by Imprecise Cluster Selection

Two (1.9%) search queries failed due to somewhat imprecise cluster selection by users (#11, #195). A search query on "electronic mail" (#11) retrieved a very promising cluster on "electronic mail systems." Yet the search was abandoned by the user. In the second case the user was interested in reference sources on art and she entered her query simply as "art" (195). Yet she selected some general clusters on reference services in libraries as relevant. Based on the user's cluster selection, the system expanded the original search query in that direction and retrieved some sources on reference services rather than reference sources on art. The user said she did not remember finding anything useful.

7.2.1.11 Search Failures Caused by Telecommunication Problems

One of the users experienced telecommunication problems when she got access to CHESHIRE, which caused her to abandon two search queries (1.9%) in the middle of the sessions (#17, #111). In both cases she managed to establish connection immediately afterwards and carried out her searches. The exact cause of why she was disconnected is not known. Yet several users got access to the system and experienced no telecommunication problems.

7.2.1.12 Analysis of Failures Caused by Users' Unfamiliarity with the Scope of the CHESHIRE Database

One of the users was unaware of the scope of the database and looking for periodical literature on collection development and acquisition practices in law libraries. He carried out two searches (1.9%) and found both of them unsuccessful. He thought the database contained bibliographic records of articles published in law library journals. Yet the database contains no references to periodical literature; bibliographic records in it represent the monographic holdings of the LSL collection. The user suggested that our presentation of the system as a "third generation" online catalog during the classroom demonstrations led him to believe that the database also indexed periodical literature. He maintained that he was "relying entirely too much on CHESHIRE to come up with the definitive answer."

7.2.1.13 Analysis of Search Failure Caused by False Drops

The primary cause of one of the search failures (0.9%) was false drops that occurred during the retrieval. The user was looking for sources on "library tours" (#113) that are given to users as a part of the bibliographic instruction or library orientation program. The top cluster the system retrieved included the following LC subject headings: **Plentin, Cristophe -- ca. 1520-1589, Printing -- France -- History, Printing -- France -- Touraine -- History**. The rest of the clusters retrieved were general.

Apparently, the retrieval algorithm attached more weight to the term "tours" in the user's query than the term "library" (for "library" is the most frequently occurring term in the database). Also, sources on library tours generally bear different title words (e.g., library orientation, bibliographic instruction). Furthermore, the user's query matched none of the LC subject headings in the database completely.

False drops occurred in a few other search queries as well. Yet they affected the outcome very little as they were presented through the end of the retrieval list. For instance, a search query on "cd-rom databases" (#10) retrieved several bibliographic records on CD-ROMs. Yet it also retrieved six records that had nothing to do with CD-ROMs such as *The early editions of the Roman de la Rose* and *Operai tipografi a Roma, 1870-1970.* Fortunately, all six titles ranked lower than the relevant titles on CD-ROMs and were displayed at the end of the list. The reason for why the system picked up such titles was that "CD-ROM" was treated as two separate words, "CD" and "ROM." Thus, after all the records in the database on CD-ROMs were exhausted, the system retrieved the next best matching records. (Apparently, the stemming algorithm reduced "Roma" to "rom.")

Similarly, the search term "quit" (#94), which was intended to be a quit command but entered in the query description screen, retrieved two clusters on the history of printing in Ecuador because the title of one of the books happened to include the word *Quito*! A search query on "Dr. Seuss" (#57) retrieved several items with "Dr." in their titles! One other query on CHESHIRE system (#191) retrieved a cluster with LC subject heading **Libraries -- England -- Cheshire -- Directories.**

7.2.1.14 Analysis of Search Failure Caused by Call Number Search

After displaying cluster records, one of the users thought the call number as another access point and entered "1. Call Number Z00699," apparently trying to retrieve all the items in that call number range. The search failed because the system has no call number searching facility.

7.2.2 Analysis of Zero Retrievals

In the previous section we analyzed the causes of search failures and referred to zero retrievals from time to time in the context of collection failures, misspellings, and so on. In this section we will briefly look at zero retrievals that occurred in CHESHIRE separately. Note that we do not categorize zero retrievals as a separate factor causing search failures, for we have already analyzed some search queries that retrieved no records in the previous section.

A total of 18 search queries retrieved nothing in CHESHIRE.[2] The causes of these zero retrievals were presented in Table 7.2.

TABLE 7.2 CAUSES OF ZERO RETRIEVALS *(N=18)*

Causes of zero retrievals	*N*	%
Collection failure	4	22.2
Out-of-domain search query	4	22.2
Stemming algorithm	2	11.1
Personal author search	2	11.1
Call number search	1	5.6
Misspelling	1	5.6
Help request	1	5.6
Quit	1	5.6
Incomplete search query	1	5.6
Gibberish	1	5.6
TOTAL	18	100.2

Note: Percentage totals do not always equal to 100% due to rounding.

As can be seen from table 7.2, of those 18 search queries, four retrieved no clusters due to collection failures ("z39.50," "hypermedia," "hypertext," and "minitel").[3] Four search queries retrieved nothing because they were out-of-domain ("nanotechnology," "syrian asceticism," "asceticism in syria," and "blood

[2] These were: 13, 20, 21, 37, 43, 45, 64, 71, 74, 79, 91, 92, 108, 118, 138, 139, 155, 226.

[3] For detailed analyses of failures mentioned here, see the previous section.

transfusion"). The stemming algorithm was the cause of two search failures ("C" and "r&d") which retrieved no records. Two search queries failed to retrieve any records because the user attempted to perform a personal author search ("tuttle" and "katz"). One search query retrieved nothing because the user attempted to perform a call number search on CHESHIRE ("1. Call Number Z00699"). Another query failed because it was incomplete ("the"). One search query failed to retrieve any records due to misspelling ("vctorian"). An indecipherable query ("ljkdsf g") retrieved nothing, either. In one case the user needed help ("how do I use this systenm *[sic]*"); one other user entered "Bquit *[sic]*," both in the query description screen. Both queries retrieved nothing due to misspellings.

We examined if the users who got zero retrieval results pursued their searches further by issuing new searches. The user who was looking for sources on "Z39.50" issued a broader search query on "user interface studies." The user who tried "hypermedia" and "hypertext" issued a new search query on a completely different topic. The user who issued search queries "nanotechnology" and "C" was browsing to see if there was anything on these topics in the database. He also was testing CHESHIRE's user interface. When his search query on "C" failed, he renewed his query as "programming C." The user who misspelled her query as "vctorian" renewed her query with the correct spelling. The user who attempted to perform a call number search understood the limitations of the system and issued a topical search next.

The user who performed out-of-domain searches on "asceticism in syria" abandoned his search after two attempts. He seemed to have been unaware of the database limitations. The user who performed personal author searches ("tuttle" and "katz") decided to quit after one more attempt ("katz reference"). The users who were looking for sources on "minitel" and "r&d" stopped using the system afterwards. So did the user who was searching for sources on "blood transfusion." The user who requested help ("how do I use this systenm") stopped searching after entering an incomplete query ("the").

7.2.3 Discussion on Search Failures

Our analysis shows that almost 40% (or 42 search queries) of search failures that occurred in CHESHIRE were mainly due to collection failures. An additional 10% (or 11 queries) of the search queries failed because users attempted to perform known-item searches (author or title searches). Two search queries failed due to user's unawareness of the scope of the CHESHIRE database (i.e., periodical articles are not indexed in the database). Two more search queries failed due to telecommunication problems. One user attempted to perform a call number search which is not supported by the system. These figures suggest that more than half the search failures (58 out of 107) were caused by factors that were outside the control of the CHESHIRE system. As the number and variety of search queries increase collection failures become inevitable no matter how large the size of the database. Furthermore, some 14 search queries failed because users were not well-informed about the limitations of the CHESHIRE system (e.g., lack of known-item and call number search features), despite our efforts of providing demonstrations and documentation.

The rest of the search failures were primarily due to user interface problems (13 queries), search statement (11 queries), cluster failures (8 queries), LCSH (5 queries) and stemming algorithm (4 queries). Specific queries, imprecise cluster selection and false drops also caused a total of 5 search failures.

Several users complained that they had experienced a multitude of difficulties with the user interface. Interviews with users indicate that CHESHIRE's user interface might have affected the outcome of several search queries indirectly even though only 13 search queries failed primarily due to interface problems.

When one of the users observed that that the user interface "looked very much like something invented for an experimental catalog," she was obviously referring to the limited help features available in CHESHIRE. Another user described the interface as "inattentive" when help was not available. Some users thought the interface was hard to understand intuitively. The experience was simply frustrating for some others.

Others compared CHESHIRE's user interface with that of second generation online catalogs. They said they feel more comfortable with, and in control of, the process of searching in traditional online catalogs where Boolean operators AND, OR, and NOT are available. Some thought the user interface was inflexible because it was menu-driven and they had to "plough through [records] screen by screen."

More often than not users issued detailed, descriptive, and yet specific, search queries, which sometimes resulted in failures. It appears that the expectations of users from a system which accepts natural language queries were high. Several users seem to have assumed that CHESHIRE is able to "understand" their search queries completely and retrieve the relevant records.

This assumption has led to poor retrieval in CHESHIRE in some cases because it has no natural language understanding capabilities. As explained earlier, all the system does is it "parses" the search statement and determines the retrieval-worthy search terms in the query. It then matches the query terms with those in the database and brings back the results using probabilistic retrieval algorithms. In fact, there were some queries where the system attributed undue weight to some search terms that should not have been taken into account at all. For instance, the term "books" in the search statement "some books on history of libraries and classification" (#38) is useless for retrieval purposes. Yet it was taken into account by the retrieval algorithm, which cluttered the search results. Similarly, the search request "find all library literature concerning the history and publication of the Federal Register" (#95) contains two words ("library" and "literature") that were useless for retrieval purposes. There were other such examples ("want to find a small set of *books* on historical treatment of mathematics" (#147), "I want *information* on the public image of librarians through history" (#162), and "*subject search* japanese novelists"(#182)).

In some cases users added qualifiers to their search queries, presumably thinking that the system would be able to figure out from their search statements what they exactly wanted. For instance, period qualifiers were introduced in the following examples: 1) "I want books about letterpress printing published after 1950" (#29); 2) "law libraries -- collection development from 1935" (#68); and 3) "banned books after 1980" (#83). Language and publication form qualifiers were also used in some queries ("I'd like to see recent books, in english, about library automation" (#99), "periodical literature on the development of law library collections" (#69)). These queries can be handled with the Boolean operator AND in second generation online catalogs (e.g., FIND SUBJECT LIBRARY AUTOMATION AND LAN ENGLISH).

One user was looking for sources on collection development in law libraries only (#70), which requires a Boolean NOT operator in second generation online catalogs.

None of the above conditions can be satisfied by CHESHIRE since, as pointed out earlier, it has no natural language understanding capabilities. The system cannot distinguish records by date, language and form. Nor can it deal with Boolean operators. It is interesting to note that users carried over some of their previous search experience from other online catalogs to CHESHIRE.

Users issued more complicated search queries which neither second- nor third generation online catalogs can satisfy. Examples are as follows: 1) "alternatives to traditional subject headings" (#223); 2) "how many books by patrick wilson does the library have?" (#97); 3) "projected salaries for special and academic librarians on the west coast" (#190); and 4) "looking for a humorous book about librarianship with cartoons" (#105).

These examples illustrate some very interesting points. Clearly, those search statements were difficult to parse and they all require natural language understanding capabilities. The first two examples were already discussed earlier. The third user was expecting the system not only to interpret her query as "projected salaries for special and academic librarians on the West Coast *of the United States* but also to determine which states constitute the West Coast (e.g., California, Washington) and thus to expand her query by adding the state (or even city) names automatically. Parsing this query requires not only some sound natural language understanding capabilities but also an extensive system vocabulary to convert (or expand) the user-supplied query terms to system's vocabulary.[4]

The fourth example also exhibits similar difficulties. In addition, the question of how one would describe a humorous book and whether such a book would be labeled in its title as a "humorous book" remains to be answered. Without such labels (or "handles") it is difficult to imagine how online catalogs could possibly retrieve records. As far as LC subject headings are concerned, some of the relevant titles (e.g., *Bibliologia comica, Bizarre books*) were cataloged under such headings as **Library science -- Humor, Literary curiosia, Bibliography -- Miscellanea, Bibliography -- Anecdotes, facetiae, satire, etc.**

Examples given above also show us how specific the users' queries can get when they are not bound with Boolean operators. It should be stressed that such specific queries would most likely fail in Boolean online catalogs. Whether the ability to submit search queries to CHESHIRE in natural language form encouraged users to be more specific is open to conjecture. Subject search statements in online catalogs that require Boolean set construction tend to be shorter whereas several search queries submitted to CHESHIRE contained more than five searchable terms (maximum was 24). (In this study the average number of searchable terms in search queries was 3.5 (see Chapter VI).)

Classification clustering process caused some false drops, examples of which were given earlier. CHESHIRE retrieves and ranks the records, generally speaking, on the basis of how closely they match users' search terms. If there are some items

[4] Note that such a system vocabulary can also be used, after negotiations with the user, to spell out user's abbreviated search requests (e.g., "rlg oclc utlas" (#189) and "r&d" (#118)).

in the database that fully match the user's query terms, then such items are listed at the top. If, however, there are not that many items that either fully or partially match the user's query terms, the system lists the best matches at the top and then lists the partial matches. It is those partial matches that confused the users most.

It was confusing when CHESHIRE's classification clustering mechanism failed to retrieve the most promising clusters, and bibliographic records, at the top of the list. When the system came up with nonrelevant records users got curious why CHESHIRE retrieved what it retrieved. For instance, one of the users was interested in library book boycott against South Africa and she entered her query as "cultural boycott of south africa" (#62). There were no relevant items in the database on this topic (e.g., collection failure). The system typically evaluated the search terms and retrieved some items. But because there were no records that fully matched user's query terms (i.e., cultural, boycott, south, africa), it came up with the next best matches such as *Morphotaxonomic studies of the South African representatives of the genus Codicum (Chlorophycophyta)* and *Research materials in South Carolina.* It is not too difficult to see that CHESHIRE retrieved those two records, among others, because they happened to contain some of the terms in the user's query ("south africa" and "south," respectively). The user said she "could not figure out what the system was doing."

In addition to the ones summarized earlier (e.g., "dr. seuss" (#57), "cd-rom databases" (#10)), several examples of such partial matches can be given. For instance, a query on "libraries in mexico" (#47) also retrieved items on libraries in *New* Mexico. A query on "berkeley library school history" (#140) came up with titles on the history of the University of California Berkeley Library. (Note the incorrect term relationships in the retrieved items for this query.) A query on "computer conferencing" (#34) retrieved general sources on computers because the term "conferencing" was not recognized. Similarly, the query "programming C" (#75) brought back general items on programming but not necessarily C programming. One user was surprised to see that her search query on "history of printing in Paris" (#81) included titles "not really connected with printing in Paris." Another user was playing with the system and he wanted to see what the system would do with a query like "please find books on children, basball *[sic]*, and animals" (#30). He said that he should not have retrieved anything. Yet he indicated that he would prefer retrieving some records, even if they do not make much sense, rather than retrieving nothing.

Several users found relevance feedback search hard to understand and confusing. Some did not know what to do with relevance feedback search while others indicated that "there is no indication of the point at which you should stop performing the relevance feedback." A few users found the relevance feedback feature in CHESHIRE not very helpful in some circumstances. For instance, one of the users indicated that "a system which will *always* attempt to give the user something is a system with a problem. The system has to be smart enough to know and inform the user that there is no good information." He added that CHESHIRE doesn't.

As summarized above, some retrieval results puzzled the users. They became curious and wanted to know "how CHESHIRE retrieves what it retrieves." The following excerpts from the interview scripts illustrate, to some extent, their uneasiness:

> I would like to learn more about CHESHIRE and how it does what it does.
>
> I couldn't figure out what it [CHESHIRE] was doing or why; it seems strange...
>
> I don't quite get CHESHIRE. I don't quite get what it's doing and I don't quite know what to do with it.
>
> Personally I would have thought it helpful to understand it a little bit better why it was retrieving what it was retrieving. It wasn't always clear to me why something had come up. . . Personally I find I can use a system better if I have some sense of why it does what it does.
>
> People don't want to interface with optimized retrieval algorithms and data structures. They want something they can work with.

It appears that some users feel less confident with their searching skills when they cannot figure out how the system interprets their commands. Furthermore, the outcome of such ineffective search results may cultivate "distrust" between the system and its users. As Buchanan (1992) pointed out, users may have very little patience when the system presents bibliographic records that should not have been retrieved in the first place.

The number of users who experienced such problems when they performed searches on CHESHIRE was relatively low, however. Several users found the system very helpful and effective. The next section concentrates on retrieval performance in CHESHIRE in terms of success. It examines the search effectiveness in CHESHIRE and summarizes the strengths of the system based on the retrieval results and users' assessments of the system.

7.2.4 Search Effectiveness in CHESHIRE

Many users participated in the experiment expressed their opinions of CHESHIRE with the following words:

> I enjoyed searching CHESHIRE. It was fun.
>
> I think this kind of system is a great idea.
>
> I think it's a marvelous idea.
>
> . . .refreshingly useful. . .intuitively easy to learn to use.
>
> I guess making things a little bit more user-friendly in terms of what was happening in the program would have made things easier to use. But even as it was I found it [CHESHIRE] more effective than other online systems on campus.
>
> It was just great.
>
> I enjoyed using it, actually.
>
> It is really intriguing stuff. . .This type of activity is I think clearly what patrons are looking for. . .People are going to get used to searching in this particular way.

Although an overwhelming majority of participating users was not familiar with probabilistic online catalogs, they quickly became proficient, as the quotes from the interview scripts show, in searching CHESHIRE once they figured out how the system works. In fact, several users compared CHESHIRE to second generation online catalogs with Boolean searching capabilities and said they would prefer CHESHIRE-like online catalogs.

The analysis of retrieval results shows that CHESHIRE's performance was well above the average. If the search failures caused by collection failures and the user interface are to be excluded from the analysis, it becomes clear that search effectiveness in CHESHIRE was much higher than many second generation online catalogs. Take, for instance, the zero retrieval rate in CHESHIRE. Eighteen queries failed to retrieve any records in CHESHIRE, which constitutes 7.9% (18/228) of all search queries submitted. Compared with much higher zero retrieval rates in first- and second generation online catalogs, this low percentage represents a remarkable achievement for CHESHIRE. For instance, Markey (1984) found that percentages of zero retrievals in subject searching range from a low of 35% to a high of 57.5%. Similar findings have been reported in several other online catalog studies (e.g., Larson, 1986; Peters, 1989; Hunter, 1991).

The enormous difference between the zero retrieval rates in CHESHIRE and other online catalogs may be due to a number of factors. First, classification clustering mechanism in CHESHIRE seems to decrease the number of zero retrievals tremendously, for CHESHIRE automatically checks both titles and subject headings of the documents in the database for possible matches during the classification clustering process. If a match is found either in titles or subject headings (or both), CHESHIRE retrieves the clusters and, subsequently, bibliographic records. In other words, a search query in CHESHIRE only fails when neither the title words nor subject headings match the user's query term(s).

Second, stemming algorithm used in CHESHIRE might have helped decrease the number of zero retrievals. In second generation online catalogs the same effect can be achieved by truncating search terms. Yet, unlike in CHESHIRE where search terms are reduced to their roots automatically, the user has to initiate the truncation action. As we have seen earlier, stemming algorithm in CHESHIRE caused false drops in rare occasions (e.g., "novell," "cheshire," "marcia tuttle"). Yet such false drops occur in second generation online catalogs more frequently.

The zero retrieval rate in CHESHIRE could be even lower with the availability of a spell-checker. Scanning search queries for misspelled or mistyped words and informing the users about potential errors before the retrieval would have prevented some zero retrievals before they occurred (see, for instance, "vctorian," "Bquit," "systenm," "ljkdsf q").

In addition to relatively low zero retrieval rate, the number of search failures that were caused by vocabulary mismatch in CHESHIRE were also fewer. That's to say, users were able to match their search queries with the system's vocabulary (i.e., titles and Library of Congress subject headings assigned to bibliographic records). Only five search queries (out of 228) failed due to mismatch between the user's vocabulary and that of CHESHIRE (2.2%) and lack of specific LC subject headings. However, it is not appropriate to compare this figure with those obtained in second generation online catalogs, which consistently showed that users' search terms exactly match the subject headings only about half the time (Carlyle, 1989, p.37; Van Pulis & Ludy, pp.528-529; Vizine-Goetz & Markey Drabenstott, 1991, p.157).

The reason why users were able to match their search statements with CHESHIRE's vocabulary, which also is one of the reasons why the figures cited above are not comparable, is the availability of classification clustering process in CHESHIRE. As mentioned earlier, classification clustering method used in CHESHIRE is the first step in the retrieval process. The user's query is processed first to determine if the query terms match titles or subject headings of the items in the database. CHESHIRE then retrieves and ranks cluster records on the basis of the degree of match between the query terms and the titles and subject headings and displays them to the user. Each cluster record display has the classification number under which most or all bibliographic records are listed, the broad topic (description of which is taken from the LC classification scheme) of the books in the cluster, and the most often assigned three LC subject headings for the books in that particular cluster. The user can then select one or more clusters as relevant, primarily by checking the most frequently assigned LC subject headings. This information will then be used to expand the user's original search query.

Larson (1991a, p.158) suggests that most often "[t]he information in the cluster display usually provides a good indication of the general topics of books under a particular classification number." Furthermore, the utilization of both title words and subject headings during the classification clustering process also increases the users' chances of matching their terminology with that of the system. The display of LC subject headings in the cluster record seems to facilitate the matching process as users are better at recognizing relevant search terms than remembering them.

Classification clustering technique used in CHESHIRE evidently helped decrease both the number of zero retrievals and number of search failures caused by vocabulary mismatch. CHESHIRE's classification clustering process worked remarkably well for especially specific search queries. Despite the fact that there were several specific queries and that the database did not contain many records that could answer such queries, CHESHIRE usually managed to retrieve the relevant ones. It successfully retrieved clusters from different parts of the classification scheme, thereby providing the user an opportunity to view his or her query in different contexts. For instance, one of the users was "interested in works that either were directly in the interdisciplinary area of knowledge utilization or that were tangential to the area of knowledge utilization". He submitted his query as "knowledge utilization"

(#119). CHESHIRE's classification clustering mechanism did an excellent job of pulling together several clusters from different parts of the LC classification schedule: theory of knowledge (BD161), communication (P91), sociology of knowledge (BD175), social science research (H62), and classification of sciences (BD241). Subsequently, the system retrieved several sources on knowledge creation, production, and utilization, which the user was "satisfied to see that they came up." Another user was trying to find out classification sections pertaining to "graphic display of thesauri in electronic format" and CHESHIRE's classification clustering process successfully pulled out records from different areas of the LC classification schedule (Z695, Z699, TK7882). He found out that "there is a section in TK. . .that deals specifically with visual display on computers."

Classification clustering technique helped provide more specific LC subject headings as part of the cluster records for specific search queries. It brings together several records from different parts of the LC classification schedules, which enables the user to retrieve relevant records that are cataloged under slightly different but nonetheless related LC subject headings. For instance, one of the users was interested in "library services for ethnic minorities" (#217). We performed several searches in order to determine the recall base for this search query, and found that the most commonly assigned LC subject heading (**Library services to minorities**) to such books retrieved less than half of the relevant records in the database. There were several unique relevant records that were indexed under 16 different LC subject headings! CHESHIRE successfully collocated most of those records cataloged under different LC subject headings by expanding the user's query on the basis of cluster selection and relevance judgments and retrieved them.

This is one of the reasons why search failures due to vocabulary mismatch occurred much less frequently in CHESHIRE than in second generation online catalogs. For, one cannot expect an ordinary end-user to come up with all the possible LC subject headings under which sources on library services to ethnic minorities are indexed. The user would have missed all the records cataloged under, *inter alia*, **Minorities -- Information services, Libraries -- Services to Hispanic Americans, Mexican Americans and libraries, Library services to Chicanos.** Furthermore, a user "looking for a humorous book on librarianship with cartoons" (#211) would be hard-pressed to remember the LC subject heading **Libraries -- Anecdotes, facetiae, satire, etc.**, under which many such books were cataloged.

It is no exaggeration, then, to suggest that many specific queries submitted to CHESHIRE would have produced zero results in second generation online catalogs with Boolean search capabilities. The availability of automatic query expansion in CHESHIRE, which is based on feedback from the user by means of classification clustering and relevance feedback techniques, helps alleviate the search failures that might have otherwise occurred.

One of the features that is available in CHESHIRE that some users found especially useful was the relevance feedback search capability. As explicated in Chapter II, relevance feedback process enables users to refine their search queries by making relevance judgments on the retrieved records. The system then incorporates this relevance information and retrieves more records that are similar to the ones that the user already judged as being relevant. Users tried relevance feedback option of CHESHIRE for 91 search queries in this study. Relevance feedback usually improved the results by retrieving more relevant records from the database (see Chapter VI).

Users, in general, seemed to have liked CHESHIRE's relevance feedback search capability, although some users admitted that they were "overwhelmed by it." One of the users commented that relevance feedback search "seemed to get her what she wanted." Another user shared the same view when he said relevance feedback search results "get more specific into exactly what he wanted." Yet another user remembered relevance feedback as "being a very nice feature." However, several users found the concept of relevance feedback search hard to understand and confusing.

One of the features of CHESHIRE that users especially liked is being able to describe their search queries in natural language. They thought that entering search statements in natural language without worrying about syntactic rules and Boolean operators was most helpful. The availability of natural language interface seem to have improved users' search statements and made the queries more descriptive.

To conclude, then, that some of the advanced information retrieval techniques that are available in CHESHIRE help decrease the search failures in online catalogs while at the same time increase the search success. Classification clustering and relevance feedback techniques tremendously improve retrieval results. Users can enter very specific search queries using the natural language and yet still retrieve some relevant records because of the availability of classification clustering and relevance feedback techniques. Furthermore, zero retrieval rates and failures caused by vocabulary mismatch are much lower than second generation online catalogs.

7.3 Summary

In this chapter the causes of search failures that occurred in CHESHIRE were analyzed qualitatively. Types of search failures (e.g., collection failures, failures due to user interface problems, cluster failures) were classified and several examples were given in each category. The likely causes of search failures were examined from the analyses of transaction logs, questionnaires, and structured interview scripts. Then, search effectiveness was examined. The strengths of CHESHIRE such as the availability of classification clustering and relevance feedback techniques and its success in decreasing search failures were discussed and the findings were recapitulated.

We found that collection and user interface failures constituted more than half of all the search failures that occurred during the experiment. This was followed by failures that occurred due to, among others, faulty search statements and known-item search queries (which were not supported by the system). To put it differently, well over half the search failures were caused by factors that were outside the control of the retrieval system. On the other hand, failures due to zero retrievals and vocabulary mismatch occurred much less frequently in CHESHIRE than in second generation online catalogs. Similarly, despite the fact that users submitted detailed yet specific search queries in many cases, the system still managed to retrieve some relevant records. This is due, in part, to the fact that probabilistic systems attempt to match the user's search terms both with titles and subject headings of the items in the database. In addition, we also found that users tend to submit longer search statements (with more search terms) to probabilistic online catalogs with natural language interfaces than they would submit to second generation online catalogs with command language user interfaces. For they are not constrained with the syntax rules of the command languages and can describe their information needs with more words.

Nonetheless, parsing natural language queries proved to be difficult because some search terms were useless for retrieval purposes but nevertheless matched records in the database. In addition, some search queries contained Boolean operators as well as language, date and form qualifiers, which suggests that users carried over some of the expertise that they gained using second generation online catalogs. Although the number of such cases was not high, it is likely that such mismatches will persist as the size of the database grows and the collection make-up becomes multi-disciplinary. That's to say, lack of natural language understanding capabilities in user interfaces will continue to cause search failures in online catalogs.

The classification clustering and relevance feedback techniques that are available in CHESHIRE appear to have played significant roles in decreasing search failures. The classification clustering technique provides users an opportunity to expand their search queries by selecting some cluster records thereby increasing their chances of retrieving relevant documents. Similarly, users' relevance judgments on retrieved records are used to automatically expand the original search query so that documents that are "similar" to the ones that were already judged as being relevant can be retrieved from the database.

However, the way the classification clustering technique has been implemented in the system prevented a few users from continuing their searches. In order to continue their searches, users have to select at least one cluster record as relevant. Yet some users were unaware of this and their searches ended prematurely due to not selecting any clusters.

Users should be able to continue their searches even if they select no clusters as being relevant. One can think of two solutions to this problem. The user can simply be asked to select at least one cluster as relevant, presumably the most promising one. This is a rather crude and simplistic solution. Besides, there may be some cases where none of the clusters would seem relevant. The second, and more elegant, solution would be to execute the query without the classification clustering mechanism, rather than forcing the user to choose at least one cluster as relevant (when there is none) against his or her will. If the user does not like any cluster, the system would go ahead and execute the query based on simple frequency distributions of query term(s) that are contained in titles and subject headings.

This may require some changes in the way the system works. At present, the classification clustering mechanism as implemented in the system gets its input from the user: whenever the user chooses one or more clusters as relevant, the system goes back and promotes those records which were listed under the selected clusters. If no clusters are chosen, however, the search ends there. The implicit assumption here is that if the system is unable to bring back possibly relevant clusters, it is highly unlikely that the collection has anything useful for that particular user and search query. This assumption may well hold for most, if not all, users and search queries. Nonetheless there would still be a merit not to end the search there, in spite of the fact that the user chose no clusters. The system could go ahead and execute the query by "bypassing" the classification clustering step.

Such an improvement would benefit some users. First, it could be that some users may find the individual records relevant even if they did not like the clusters. This would cut down the number of searches that abruptly end due to not selecting any clusters. Second, some users may not be aware of the fact that they must choose at least one cluster as being relevant in order to be able to retrieve some individual records. There is some evidence that some users did expect to get to individual records without selecting any clusters. In fact, CHESHIRE's user interface gives no clues to the users that they "have to" select clusters. Third, some users come to the system just to test it and see how it works (so called "tourists"). They do not necessarily want to follow the instructions. Rather they want to explore the system. When reminded during the interviews that they probably did not like the clusters they had seen, several users stated that they were "just exploring." Those students who want to explore the system without selecting any clusters would never get to individual records.

It can be argued that providing access to individual records by bypassing the classification clustering step would mean that the users would not be able to use CHESHIRE to its full strength. This is certainly true. Larson's research indicates that classification clustering mechanism helps users match their vocabulary with that of the system. This, in turn, improves the quality of the searches. Nevertheless it should still be possible to retrieve records based on simple frequency distribution counts. At present, the classification clustering algorithm is not closely tied with the retrieval process. That is to say, the system checks the cluster "centroids" only after the user selects some clusters so that the records in promising clusters be considered more important for retrieval than the others. Bypassing classification clustering would mean that the system need only evaluate the bibliographic records containing the query terms. This, according to Larson, will actually decrease the overall processing needed for each query as there will be fewer terms to consider for retrieval purposes.

The relevance feedback technique helped retrieve more relevant sources from the database, yet the search results tended to deteriorate quickly after the second relevance feedback iteration. It appears that the user-entered search query deviates from the original form with the addition of too many nonrelevant terms during the relevance feedback cycles. Similar findings have also been reported in other probabilistic online catalogs (e.g., Okapi) with relevance feedback search techniques.

Some of the advanced retrieval techniques that constitute the strengths of the CHESHIRE system confused some users. For instance, most users liked the natural language interface and found the relevance feedback feature useful. Yet some users were bewildered with the availability of the very same techniques as they apparently never used a probabilistic online catalog with classification clustering and relevance feedback techniques. A few users indicated that they would prefer to use a Boolean command language to interact with the system rather than a natural language user interface.

CHAPTER VIII
CONCLUSION

8.0 Summary

The hypothesis of this dissertation was the assertion that online catalog users often fail in their attempts to retrieve relevant items from document collections using existing online library catalogs. A conceptual mode[1] was developed to examine and categorize search failures that occur in these catalogs. To test the model, an experiment was designed in which we recorded in transaction logs complete interactions of 45 users performing 228 queries. A questionnaire was administered and participating users were interviewed after the completion of their searches. One of the main objectives of this dissertation has been to analyze search failures comprehensively by employing not only precision and recall measures but also by identifying user-designated ineffective searches and comparing them with the precision and recall measures for corresponding queries.

Using a regression model, we tested the hypothesis that users' assessments of retrieval effectiveness differ from retrieval performance as measured by precision and recall ratios and that increasing the match between the users' vocabulary and that of the system by means of clustering and relevance feedback techniques will improve retrieval effectiveness and help reduce search failures in online catalogs.

8.1 Conclusions

Retrieval performance of the system as measured by precision and recall ratios was such that users judged half the retrieved records as being relevant before relevance feedback searches while the system retrieved less than a quarter of the relevant sources in the database. As users proceeded with relevance feedback searches, they found the retrieved sources less and less helpful although the system retrieved additional relevant sources from the database. In other words, as should be expected, precision ratios dropped sharply after relevance feedback searches while recall ratios almost doubled.

In spite of the fact that users selected less than four records as being relevant in more than 75% of the search queries and, in their view, two-thirds of the searches contained less than 25% of the useful sources, they judged two-thirds of the search queries as being effective. In other words, low precision rates do not necessarily mean that users found their search results ineffective. Furthermore, they indicated the relevance feedback mechanism was helpful and that they retrieved additional relevant sources during the relevance feedback searches, although precision ratios were much lower in those cases.

These seemingly conflicting findings obtained from transaction logs, questionnaires and critical incident reports were confirmed by the results of a multiple linear regression analysis. No strong correlation was found between retrieval performance as measured by precision and recall ratios and users' assessments of search effectiveness (i.e., whether they judged their search as being effective or not, or whether they found what they wanted). Furthermore, there was no strong correlation, either, between precision and recall ratios and the user characteristics such as the frequency of online catalog use and knowledge of online searching. These findings also proved the main hypothesis of this dissertation, which was that retrieval performance as measured by precision and recall ratios differs from users' assessments of retrieval effectiveness and that variables that define users characteristics do not explain the variability in performance measures.

The relationship (or lack thereof) between traditional performance measures such as precision and recall and that of user characteristics and users' assessments of retrieval effectiveness shows, once again, that measuring retrieval performance is a complex task. It also shows that it is difficult to explain the retrieval effectiveness in online catalogs on the basis of variables that define user characteristics and traditional performance measures. No meaningful pattern has emerged as to how the user judges the retrieval results for a given query based on retrieval performance and other variables. Although not directly examined in this dissertation, the findings also indicate that it is extremely difficult to study the search behavior of users when searching online catalogs.

Quantitative findings also suggest that measuring retrieval performance solely on the basis of precision and recall ratios may not satisfactorily explain the causes of all types of search failures that occur in online catalogs. Each search query is unique in the sense that success or failure depends very much on individual circumstances. This observation was confirmed by the qualitative analysis of search failures that occurred during the experiment.

Search queries failed predominantly due to collection and user interface failures in the experiment. More than half the search failures were caused by collection and user interface failures. In addition, search statements, users' unawareness of the capabilities of an experimental online catalog, lack of specific subject headings, cluster failures, among others, also caused search failures. Users experienced some difficulties in adapting to an experimental online catalog with advanced retrieval techniques such as classification clustering and relevance feedback and sometimes they could not figure out how to continue their searches. Some users also experienced problems with the natural language user interface as they expected more from it than a natural language interface can deliver. For instance, they expected such interfaces to be capable of not only interpreting Boolean operators and qualifiers but also, in some cases, providing in-depth or factual answers to research questions. To put it in somewhat different terms, users seemed to have transferred some of the search tactics they developed on Boolean systems over to a probabilistic system and, at the same time, wished to benefit from whatever the probabilistic retrieval systems may have to offer (i.e., "best match" techniques, natural language interfaces). This suggests that if probabilistic retrieval systems are to be alternatives to existing online catalogs, they should have the capabilities of existing online catalogs in addition to more advanced search features such as clustering and relevance feedback techniques. For example, the functionality of probabilistic online catalogs can be further increased by utilizing some of the information that is already in place in a MARC record (i.e., author, title, language, publication date).

Users tend to issue longer and sometimes rather specific search queries in probabilistic online catalogs presumably because they are not constrained with the limitations of the command language and Boolean logic. This, however, complicates the query parsing process as longer search requests are more likely to contain useless words from the retrieval point of view. Presumably in the future, online catalogs will be equipped with a multitude of user interfaces where users will have a choice to select their most favorite user interface type, be it the command language or the natural language user interface. Co-existence of several user interfaces in an online catalog will facilitate the use of the system by all types of users. Thus, it will be possible to perform a search in a probabilistic online catalog using a command language.

One hypothesis tested in the dissertation was that certain types of search failures will occur less frequently in probabilistic online catalogs than in second generation online catalogs. This hypothesis was confirmed in that zero retrievals and failures due to vocabulary mismatch occurred much less frequently during the experiment. Despite the fact that users submitted several very specific search requests to the system, failures due to zero retrievals constituted less than 8% of all the queries, a far better rate than that in second generation online catalogs. It appears that probabilistic online catalogs are less "brittle" than online catalogs with Boolean searching capabilities regarding zero retrievals. Similarly, very few queries completely failed as a direct consequence of users' not matching their search terms with the system's vocabulary (i.e., titles and subject headings assigned to the documents). Thus, the classification clustering and relevance feedback techniques that are available in the experimental online catalog helped decrease these types of search failures because search terms are matched against both titles and subject headings, thereby increasing the chances of a potential match.

The qualitative analysis of search failures showed that the conceptual model to examine and categorize search failures was comprehensive enough to encompass most, if not all, the types of search failures that occurred during the experiment.

8.2 Further Research

As mentioned earlier, we found that there was no strong correlation between traditional retrieval performance measures and variables that defined users' characteristics and users' assessment of search effectiveness. Similar findings also have been reported elsewhere. However, more research is needed to validate the findings obtained in this study over larger populations of search queries.

It would also be useful to see if the conceptual model developed in this study can be employed to examine and categorize search failures in other studies. Moreover, the model can be refined and used as the starting point to create an even more detailed taxonomy of search failures in online catalogs.

Although retrieval performance in probabilistic online catalogs has been studied using precision and recall measures, search failures that occur in such catalogs have not been fully examined. The present study is the first attempt and should be replicated on other probabilistic online catalogs and catalogs with natural language interfaces.

BIBLIOGRAPHY

Alzofon, Sammy R. and Noelle Van Pulis. (1984). "Patterns of Searching and Success Rates in an Online Public Access Catalog," *College & Research Libraries* **45**, 110-115.

Ankeny, Melvon L. (1991). "Evaluating End-User Services: Success or Satisfaction," *Journal of Academic Librarianship* **16**, 352-356.

Auster, Ethel and Stephen B. Lawton. (1984). "Search Interview Techniques and Information Gain as Antecedents of User Satisfaction with Online Bibliographic Retrieval," *Journal of the American Society for Information Science* **35**, 90-103.

Bates, Marcia J. (1972). "Factors Affecting Subject Catalog Search Success," Unpublished Ph.D. Dissertation, The University of California, Berkeley.

_______. (1977a). "Factors Affecting Subject Catalog Search Success,"*Journal of the American Society for Information Science* **28**, 161-169.

_______. (1977b)."System Meets User: Problems in Matching Subject Search Terms," *Information Processing & Management* **13**, 367-375.

_______. (1986). "Subject Access in Online Catalogs: a Design Model," *Journal of the American Society for Information Science* **37**, 357-376.

_______. (1989a). "The Design of Browsing and Berrypicking Techniques for the Online Search Interface," *Online Review* **13**, 407-424.

_______. (1989b). "Rethinking Subject Cataloging in the Online Environment," *Library Resources & Technical Services* **33**, 400-412.

Belkin, Nicholas J. and W. Bruce Croft. (1987). "Retrieval Techniques," *Annual Review of Information Science and Technology* **22**, 109-145.

Bing, Jon. (1987). "Performance of Legal Text Retrieval Systems: The Curse of Boole," *Law Library Journal* **79**, 187-202.

Blair, David C. (1980). "Searching Biases in Large Interactive Document Retrieval Systems," *Journal of the American Society for Information Science* **31**, 271-277.

Blair, David C. (1990). *Language and Representation in Information Retrieval.* Amsterdam: Elsevier.

Blair, David C. and M.E. Maron. (1985). "An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System," *Communications of the ACM* **28**, 289-299.

Blazek, Ron and Dania Bilal. (1988). "Problems with OPAC: a Case Study of an Academic Research Library," *RQ* **28**, 169-178.

Borgman, Christine L. (1986). "Why are Online Catalogs Hard to Use? Lessons Learned from Information-Retrieval Studies," *Journal of the American Society for Information Science* **37**, 387-400.

_______. (1983). "End User Behavior on an Online Information Retrieval System: A Computer Monitoring Study," in Kuehn, Jennifer J. (ed.) *International Conference on Research and Development in Information Retrieval* (pp.162-176). 6th Annual International ACM SIGIR Conference. New York: ACM.

Buchanan, Paul (1992, August 7). E-mail message to PACS-L@UHUPVM1.BITNET.

Buckland, Michael K. and Doris Florian. (1991). "Expertise, Task Complexity, and Artificial Intelligence: A Conceptual Framework," *Journal of the American Society for Information Science* **42**, 635-643.

Buckley, Chris. (1987). *Implementation of the SMART Information Retrieval System.* Ithaca, NY: Cornell University, Department of Computer Science.

Byrne, Alex and Mary Micco. (1988). "Improving OPAC Subject Access: The ADFA Experiment," *College & Research Libraries* **49**, 432-441.

Carlyle, Allyson. (1989). "Matching *LCSH* and User Vocabulary in the Library Catalog," *Cataloging & Classification Quarterly* **10**, 37-63.

Chan, Lois Mai. (1986a). *Library of Congress Subject Headings: Principles and Application.* 2nd edition. Littleton, CO: Libraries Unlimited, Inc.

_______. (1986b). *Improving LCSH for Use in Online Catalogs.* Littleton, CO: Libraries Unlimited, Inc.

_______. (1986c). "Library of Congress Classification as an Online Retrieval Tool: Potentials and Limitations," *Information Technology and Libraries* **5**, 181-192.

Chan, Lois Mai. (1989). "Library of Congress Class Numbers in Online Catalog Searching," *RQ* **28**, 530-536.

Cheney, Debora. (1991). "Evaluation-Based Training: Improving the Quality of End-User Searching," *Journal of Academic Librarianship* **17**, 152-155.

Cherry, J. M. (1992) "Improving Subject Access in OPACs: An Exploratory Study of Conversion of Users' Queries," *Journal of Academic Librarianship* **18**, 95-99.

Cleverdon, Cyril W. (1962). *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems*. Cranfield, England: Aslib.

Cleverdon, Cyril, Jack Mills, and Michael Keen. (1966). *Factors Determining the Performance of Indexing Systems, Volume 1, Design*. Cranfield, England: Aslib.

Cleverdon, Cyril and Michael Keen. (1966). *Factors Determining the Performance of Indexing Systems, Volume 2, Test Results*. Cranfield, England: Aslib.

Cochrane, Pauline A. and Karen Markey. (1983). "Catalog Use Studies -Since the Introduction of Online Interactive Catalogs: Impact on Design for Subject Access," *Library and Information Science Research* **5**, 337-363.

Cooper, Michael D. (1991). "Failure Time Analysis of Office System Use," *Journal of the American Society for Information Science* **42**, 644-656.

Cooper, William S. (1973). "On Selecting a Measure of Retrieval Effectiveness," *Journal of the American Society for Information Science* **24**, 87-100, and 413-424.

______. (1988). "Getting beyond Boole," *Information Processing & Management* **23**, 243-248.

Dale, Doris Cruger. (1989). "Subject Access in Online Catalogs: An Overview Bibliography," *Cataloging & Classification Quarterly* **10**, 225-251.

Dickson, J. (1984). "Analysis of User Errors in Searching an Online Catalog," *Cataloging & Classification Quarterly* **4**, 19-38.

Doszkocs, T.E. (1983). "CITE NLM: Natural Language Searching in an Online Catalog," *Information Technology and Libraries* **2**, 364-380.

Eisenberg, Michael and Linda Schamber. (1988). "Relevance: The Search for a

Definition," in *ASIS '88: Proceedings of the 51st ASIS Annual Meeting, Atlanta, Georgia, October 23-27, 1988* (pp.164-168). Ed. by Christine L. Borgman and Edward Y .H. Pai. Medford, NJ: Learned Information.

Ensor, Pat. (1992). "User Practices in Keyword and Boolean Searching on an Online Public Access Catalog," *Information Technology and Libraries* **11**, 210-219.

Flanagan, John C. (1954). "The Critical Incident Technique," *Psychological Bulletin* **51**, 327-358.

Frost, Carolyn O. (1987a). "Faculty Use of Subject Searching in Card and Online Catalogs," *Journal of Academic Librarianship* **13**, 86-92.

______. (1987b). "Subject Searching in an Online Catalog," *Information Technology and Libraries* **6**, 61-63.

______. (1989). "Title Words as Entry Vocabulary to LCSH: Correlation between Assigned LCSH Terms and Derived Terms From Titles in Bibliographic Records with Implications for Subject Access in Online Catalogs," *Cataloging & Classification Quarterly* **10**, 165-179.

Frost, Carolyn O. and Bonnie A. Dede. (1988). "Subject Heading Compatibility between LCSH and Catalog Files of a Large Research Library: a Suggested Model for Analysis," *Information Technology and Libraries* **7**, 292-299.

Gerhan, David R. (1989). "LCSH *in vivo*: Subject Searching Performance and Strategy in the OPAC Era," *Journal of Academic Librarianship* **15**, 83-89.

Gouke, Mary Noel and Sue Pease. (1982). "Title Searches in an Online Catalog and a Card Catalog: A Comparative Study of Patron Success in Two Libraries," *Journal of Academic Librarianship* **8**, 137-143.

Hancock-Beaulieu, Micheline. (1987). "Subject Searching Behaviour at the Library Catalogue and at the Shelves: Implications for Online Interactive Catalogues," *Journal of Documentation* **43**, 303-321.

______. (1990). "Evaluating the Impact of an Online Library Catalogue on Subject Searching Behaviour at the Catalogue and at the Shelves," *Journal of Documentation* **46**, 318-338.

Hancock-Beaulieu, Micheline, Stephen Robertson and Colin Neilson. (1991). "Evaluation of Online Catalogues: Eliciting Information from the User," *Information Processing & Management* **27**, 523-532.

Hartley, R.J. (1988). "Research in Subject Access: Anticipating the User," *Catalogue and Index* no. 88, 1,3-7.

Henty, M. (1986). "The User at the Online Catalogue: a Record of Unsuccessful Keyword Searches," *LASIE* **17**(2): 47-52.

Hilchey, Susan E. and Jitka M. Hurych. (1985). "User Satisfaction or User Acceptance? Statistical Evaluation of an Online Reference Service," *RQ* **24**, 452-459.

Hildreth, Charles R. (1982). *Online Public Access Catalogs: The User Interface.* Dublin, OH: OCLC, Inc.

_______. (1985). "Online Public Access Catalogs," in *Annual Review of Information Science and Technology* vol. 20, (pp.233-285). Ed. by Martha E. Williams. White Plains, NY: Knowledge Industry Publications.

_______. (1989). *Intelligent Interfaces and Retrieval Methods for Subject Searching in Bibliographic Retrieval Systems.* Washington, DC: Cataloging Distribution Service, Library of Congress.

Hjerrpe, Roland. (1986). "Project HYPERCATalog: Visions and Preliminary Conceptions of an Extended and Enhanced Catalog," in *Intelligent Information Systems for the Information Society.* Ed. by B.C. Brookes. New York: North-Holland Publishing Co.

Holley, Robert P. (1989). "Subject Access in the Online Catalog," *Cataloging & Classification Quarterly* **10**, 3-8.

Hunter, Rhonda N. (1991). "Successes and Failures of Patrons Searching the Online Catalog at a Large Academic Library: a Transaction Log Analysis," *RQ* **30**, 395-402.

Ide, E. (1971). "New Experiments in Relevance Feedback." in Salton, Gerard, (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp.337-354). Englewood Cliffs, NJ: Prentice-Hall.

Janosky, Beverly, Philip J. Smith and Charles Hildreth. (1986). "Online Library Catalog Systems: An Analysis of User Errors," *International Journal of Man-Machine Studies* **25**, 573-592.

Jones, R. (1986). "Improving Okapi: Transaction Log Analysis of Failed Searches in an Online Catalogue," *Vine* no. 62, 3-13.

Kaske, Neal N. (1983). *A Comprehensive Study of Online Public Access Catalogs: an Overview and Application of Findings*. (OCLC Research Report # OCLC/OPR/RR-83-4) Dublin, OH: OCLC.

Kaske, Neal N. (1988a). "A Comparative Study of Subject Searching in an OPAC Among Branch Libraries of a University Library System," *Information Technology and Libraries* **7**, 359-372.

_______. (1988b). "The Variability and Intensity over Time of Subject Searching in an Online Public Access Catalog," *Information Technology and Libraries* **7**, 273-287.

Kaske, Neal K. and Nancy P. Sanders. (1980). "Online Subject Access: the Human Side of the Problem," *RQ* **20**, 52-58.

Kern-Simirenko, Cheryl. (1983). "OPAC User Logs: Implications for Bibliographic Instruction," *Library Hi Tech* **1**, 27-35.

Kinnucan, Mark T. (1992). "The Size of Retrieval Sets," *Journal of the American Society for Information Science* **43**, 72-79.

Kinsella, Janet and Philip Bryant. (1987). "Online Public Access Catalog Research in the United Kingdom: An Overview," *Library Trends* **35**, 619-629.

Kirby, Martha and Naomi Miller. (1986). "MEDLINE Searching on Colleague: Reasons for Failure or Success of Untrained End Users," *Medical Reference Services Quarterly* **5**(3): 17-34.

Klugman, Simone. (1989). "Failures in Subject Retrieval," *Cataloging & Classification Quarterly* **10**, 9-35.

Krovetz, Robert and W. Bruce Croft. (1992). "Lexical Ambiguity and Information Retrieval," *ACM Transactions on Information Systems* **10**, 115-141.

Kuhns, J.L. (1963). Section 7 of *Research on an Advanced NASA Information System.* Los Angeles: Bunker Ramo Corp.

Lancaster, F.W. (1968). *Evaluation of the MEDLARS Demand Search Service.* Washington, DC: US Department of Health, Education and Welfare.

_______. (1969). "MEDLARS: Report on the Evaluation of Its Operating Efficiency," *American Documentation* **20**, 119-142.

Larson, Ray R. (1986). "Workload Characteristics and Computer System Utilization in Online Library Catalogs." Unpublished Ph.D. Dissertation, The University of California at Berkeley. (University Microfilms No. 8624828)

Larson, Ray R. (1989). "Managing Information Overload in Online Catalog Subject Searching,"in *ASIS '89 Proceedings of the 52nd ASIS Annual Meeting Washington, DC, October 30-November 2, 1989* (pp.129-135). Ed. by Jeffrey Katzer et al. Medford, NJ: Learned Information.

_______. (1991a). "Classification Clustering, Probabilistic Information Retrieval and the Online Catalog," *Library Quarterly* **61**, 133-173.

_______. (1991b). "The Decline of Subject Searching: Long Term Trends and Patterns of Index Use in an Online Catalog," *Journal of the American Society for Information Science* **42**, 197-215.

_______. (1991c). "Between Scylla and Charybdis: Subject Searching in the Online Catalog," in *Advances in Librarianship*, vol. 15, (pp.175-236). Ed. by Irene P. Godden. San Diego, CA: Academic Press.

_______. (1992). "Evaluation of Advanced Information Retrieval Techniques in an Experimental Online Catalog," *Journal of the American Society for Information Science* **43**, 34-53.

Lawrence, Gary S. (1985). "System Features for Subject Access in the Online Catalog," *Library Resources & Technical Services* **29**, 16-33.

Lawrence, Gary S., V. Graham and H. Presley. (1984). "University of California Users Look at MELVYL: Results of a Survey of Users of the University of California Prototype Online Union Catalog," *Advances in Library Administration* **3**, 85-208.

Lewis, David. (1987). "Research on the Use of Online Catalogs and Its Implications for Library Practice," *Journal of Academic Librarianship* **13**, 152-157.

Markey, Karen. (1980). *Analytical Review of Catalog Use Studies.* (OCLC Research Report # OCLC/OPR/RR-80/2.) Dublin, OH: OCLC.

_______. (1983). *The Process of Subject Searching in the Library Catalog: Final Report of the Subject Access Research Project.* (OCLC Research Report # OCLC/OPR/RR/83-3) Dublin, OH: OCLC.

_______. (1984). *Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs.* Dublin, OH: OCLC.

Markey, Karen. (1985). "Subject Searching Experiences and Needs of Online Catalog Users: Implications for Library Classification," *Library Resources & Technical Services* **29**, 34-51.

_______. (1986). "Users and the Online Catalog: Subject Access Problems," in Matthews, J.R. (ed.) *The Impact of Online Catalogs* (pp.35-69). New York: Neal-Schuman.

_______. (1988). "Integrating the Machine-Readable LCSH into Online Catalogs," *Information Technology and Libraries* **7**, 299-312.

Markey, Karen and Anh N. Demeyer. (1986). *Dewey Decimal Classification Online Project: Evaluation of a Library Schedule and Index Integrated into the Subject Searching Capabilities of an Online Catalog.* (Report Number: OCLC/OPR/RR-86-1) Dublin, OH: OCLC.

Maron, M.E. (1984). "Probabilistic Retrieval Models," in Dervin, Brenda and Melvin J. Voigt, (eds.) *Progress in Communication Sciences*, vol. 5 (pp.145-176). Norwood, NJ: Ablex.

Matthews, Joseph K. (1982). *A Study of Six Public Access Catalogs: a Final Report Submitted to the Council on Library Resources, Inc.* Grass Valley, CA: J. Matthews and Assoc., Inc.

Matthews, Joseph, Gary S. Lawrence and Douglas Ferguson. (eds.) (1983). *Using Online Catalogs: a Nationwide Survey.* New York, NY: Neal-Schuman.

Mischo, William H. (1981). *A Subject Retrieval Function for the Online Union Catalog.* (Technical Report Number: OCLC/DD/TR-81/4) Dublin, OH: OCLC.

Mitev, Nathalie Nadia, Gillian M. Venner and Stephen Walker. (1985). *Designing an Online Public Access Catalogue: Okapi, a Catalogue on a Local Area Network.* (Library and Information Research Report 39) London: British Library.

Mooers, Calvin N. (1960 July). "Mooers Law Or, Why Some Retrieval Systems Are Used and Others Are Not," (Editorial) *American Documentation* **11**, ii.

Nielsen, Brian. (1986). "What They Say They Do and What They Do: Assessing Online Catalog Use Instruction Through Transaction Monitoring," *Information Technology and Libraries* **5**, 28-34.

Penniman, W. David. (1975a). "A Stochastic Process Analysis of On-line User Behavior," in *Information Revolution: Proceedings of the 38th ASIS Annual Meeting, Boston, Massachusetts, October 26-30, 1975*. Volume 12 (pp.147-148). Washington, DC: ASIS.

Penniman, W. David. (1975b). "Rhythms of Dialogue in Human-Computer Conversation." Unpublished Ph.D. Dissertation. The Ohio State University.

Penniman, W.D. and W.D. Dominic. (1980). "Monitoring and Evaluation of On-line Information System Usage," *Information Processing & Management* **16**, 17-35.

Peters, Thomas A. (1989). "When Smart People Fail: An Analysis of the Transaction Log of an Online Public Access Catalog," *Journal of Academic Librarianship* **15**, 267-273.

_______. (1991). *The Online Catalog: A Critical Examination of Public Use*. Jefferson, NC: McFarland & Co.

Porter, Martin and Valerie Galpin. (1988). "Relevance Feedback in a Public Access Catalogue for a Research Library: Muscat at the Scott Polar Research Institute," *Program* **22**, 1-20.

Robertson, S.E., M.E.Maron and W.S. Cooper. (1982). "Probability of Relevance: A Unification of Two Competing Models for Document Retrieval," *Information Technology: Research and Development* **1**, 1-21.

Rocchio, Jr., J.J. (1971a). "Relevance Feedback in Information Retrieval." in Salton, Gerard, (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp.313-323). Englewood Cliffs, NJ: Prentice-Hall.

_______. (1971b). "Evaluation Viewpoints in Document Retrieval," in Salton, Gerard, (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp. 68-73) Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1971a). "Relevance Feedback and the Optimization of Retrieval Effectiveness." in Salton, Gerard, (ed.) *The SMART Retrieval System: Experiments in Automatic Document Processing* (pp.324-336). Englewood Cliffs, NJ: Prentice-Hall.

_______, (ed.) (1971b). *The SMART Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, Gerard and M.J. McGill. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Salton, Gerard and Chris Buckley. (1990). "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science* **41**, 288-297.

Saracevic, Tefko. (1975). "Relevance: A Review of and a Framework for the Thinking on the Notion in Information Science," *Journal of the American Society for Information Science* **26**, 321-343.

Saracevic, Tefko and Paul Kantor. (1988). "A Study of Information Seeking and Retrieving. II. Users, Questions, and Effectiveness," *Journal of the American Society for Information Science* **39**, 177-196.

Saracevic, Tefko, Paul Kantor, Alice Y. Chamis, and Donna Trivison. (1988). "A Study of Information Seeking and Retrieving. I. Background and Methodology," *Journal of the American Society for Information Science* **39**, 161-176.

Schamber, Linda, Michael B. Eisenberg and M.S. Nilan. (1990). "A Re-examination of Relevance: Toward a Dynamic, Situational Definition," *Information Processing & Management* **26**, 755-776.

Seaman, Scott. (1992). "Online Catalog Failure as Reflected through Interlibrary Loan Error Requests," *College & Research Libraries* **53**, 113-120.

Seymour, Sharon. (1991). "Online Public Access Catalog User Studies: A Review of Research Methodologies, March 1986-November 1989," *Library and Information Science Research* **13**, 89-102.

Shepherd, Michael A. (1981). "Text Passage Retrieval Based on Colon Classification: Retrieval Performance," *Journal of Documentation* **37**, 25-35.

_______. (1983). "Text Retrieval Based on Colon Classification: Failure Analysis," *Canadian Journal of Information Science* **8**, 75-82.

Shneiderman, B. (1986). *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Reading, MA: Addison-Wesley.

Simpson, Charles W. (1989). "OPAC Transaction Log Analysis: The First Decade," in *Advances in Library Automation and Networking*, vol. 3 (pp.35-67). Ed. by Joe A. Hewitt. Greenwich, CT: JAI Press.

Siochi, Antonio C. and Roger W. Ehrich. (1991). "Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions," *ACM Transactions on Information Systems* **9**, 309-335.

Soergel, Dagobert. (1976). "Is User Satisfaction a Hobgoblin?," *Journal of the American Society for Information Science* **27**, 256-259.

Sparck Jones, Karen. (1981a). "Retrieval System Tests 1958-1978," in Sparck Jones, Karen, (ed.) *Information Retrieval Experiment* (pp. 213-255). London: Butterworths.

Sparck Jones, Karen, (ed.) (1981b). *Information Retrieval Experiment.* London: Butterworths.

Su, Louise T. (1992). "Evaluation Measures for Interactive Information Retrieval," *Information Processing & Management* **28**, 503-516.

Svenonius, Elaine. (1983). "Use of Classification in Online Retrieval," *Library Resources & Technical Services* **27**, 76-80.

_______. (1986). "Unanswered Questions in the Design of Controlled Vocabularies," *Journal of the American Society for Information Science* **37**, 331-340.

Svenonius, Elaine and H. P. Schmierer. (1977). "Current Issues in the Subject Control of Information," *Library Quarterly* **47**, 326-346.

Swanson, Don R. (1965). "The Evidence Underlying the Cranfield Results," *Library Quarterly* **35**, 1-20.

Swanson, Don R. (1977). "Information Retrieval as a Trial-and-Error Process," *Library Quarterly* **47**, 128-148.

Tagliacozzo, Renata. (1977). "Estimating the Satisfaction of Information Users," *Bulletin of the Medical Library Association* **65**, 243-249.

Tague, Jean M. (1981). "The Pragmatics of Information Retrieval Experimentation," in Sparck Jones, Karen. (ed.) *Information Retrieval Experiment* (pp. 59-102). London: Butterworths.

Tessier, Judith A. (1981). "Toward the Understanding of User Satisfaction: A Multiattribute Study of User Evaluations of Computer-Based Literatures in Medical Libraries. Unpublished Ph.D. Dissertation, Syracuse University.

Tessier, Judith A., Wayne W. Crouch, and Pauline Atherton. (1977). "New Measures of User Satisfaction with Computer-Based Literature Searches," *Special Libraries* **68**, 383-389.

Tolle, John E. (1983a). *Current Utilization of Online Catalogs: Transaction Log Analysis. Final Report to the Council on Library Resources, Vol. 1.* (OCLC Research Report # OCLC/OPR/RR/83-2) Dublin, OH: OCLC.

______. (1983b). "Transaction Log Analysis: Online Catalogs," in Kuehn, Jennifer J. (ed.) *International Conference on Research and Development in Information Retrieval.* (pp.147-160). 6th Annual International ACM SIGIR Conference. New York, NY: Association for Computing Machinery.

Tremain, Russ and Michael D. Cooper. (1983). "A Parser for On-line Search System Evaluation," *Information Processing & Management* **19**, 65-75.

University of California Users Look at MELVYL: Results of a Survey of Users of the University of California Prototype Online Union Catalog. (1983). Berkeley, CA: The University of California.

Van Pulis, N. and L.E. Ludy. (1988). "Subject Searching in an Online Catalog with Authority Control," *College & Research Libraries* **49**, 523-533.

Van Rijsbergen, C.J. (1979). *Information Retrieval.* 2nd ed. London: Butterworths.

______. (1981). "Retrieval Effectiveness," in Sparck Jones, Karen, (ed.) *Information Retrieval Experiment* (pp.32-43). London: Butterworths.

Vickery, Brian and Alina Vickery. (1992). "An Application of Language Processing for a Search Interface," *Journal of Documentation* **48**, 255-275.

Vizine-Goetz, Diane and Karen Markey Drabenstott. (1991). "Computer and Manual Analysis of Subject Terms entered by Online Catalog Users," in *ASIS '91: Proceedings of the 54th ASIS Annual Meeting. Washington, DC, October 27-31, 1991* (pp.156-161). Ed. by Jose-Marie Griffiths. Medford, NJ: Learned Information.

Wages, Robert. (1989). "Can Easy Searching be Good Searching? A Model for Easy Searching," *Online* **13** (May), 78-85.

Walker, Cynthia J., K. Ann McKibbon, R. Brian Haynes and Michael F. Ramsden. (1991). "Problems Encountered by Clinical End Users of MEDLINE and GRATEFUL MED," *Bulletin of the Medical Library Association* **79**, 67-69.

Walker, Stephen. (1988). "Improving Subject Access Painlessly: Recent Work on the Okapi Online Catalogue Projects," *Program* **22**, 21-31.

Walker, Stephen and Richard M. Jones. (1987). *Improving Subject Retrieval in Online Catalogues. 1: Stemming, Automatic Spelling Correction and Cross-Reference Tables.* (British Library Research Paper 24) London: The British Library.

Walker, Stephen and R. de Vere. (1990). *Improving Subject Retrieval in Online Catalogues. 2: Relevance Feedback and Query Expansion.* (British Library Research Paper, no. 72) London: British Library.

Walker, Stephen and Micheline Hancock-Beaulieu. (1991). *Okapi at City: An Evaluation Facility for Interactive Information Retrieval.* (British Library Research Report 6056). London: The British Library.

Wang, Chih. (1985). "The Online Catalogue, Subject Access and User Reactions: A Review," *Library Review* **34**, 143-152.

Wiberley, S. E. and R. A. Dougherty. (1988). "Users' Persistence in Scanning Lists of References," *College & Research Libraries* **49**, 149-156.

Wilson, Patrick. (1983). "The Catalog as Access Mechanism: Background and Concepts," *Library Resources & Technical Services* **27**, 4-17.

Wilson, Sandra R., Norma Starr-Schneidkraut and Michael D. Cooper. (1989). *Use of the Critical Incident Technique to Evaluate the Impact of MEDLINE.* (Final Report, September 30, 1989. Contract No. N01-LM-8-3529; AIR-64600-9/89-FR) Palo Alto, CA: American Institutes for Research.

Zink, Steven D. (1991). "Monitoring User Search Success through Transaction Log Analysis: the WolfPac Example," *Reference Services Review* **19**, 49-56.

APPENDICES

APPENDIX A
BACKGROUND INFORMATION ABOUT CHESHIRE AND GUIDELINES FOR CHESHIRE SEARCHES

This appendix reproduces a handout distributed to potential participants before the experiment began. It introduces the experiment to participating users and explains what they are asked to do for the experiment.

BACKGROUND INFORMATION ABOUT CHESHIRE AND GUIDELINES FOR CHESHIRE SEARCHES

Before I explain what I would like you to do for this research, let me briefly summarize what CHESHIRE is all about and why your participation is important.

CHESHIRE is one of the next generation online catalogs that is designed to accommodate sophisticated information retrieval (IR) techniques based on sound theoretical backing. The database for the CHESHIRE system consists of some 30,000 records representing the holdings of the Library School Library here at UC Berkeley. The size of the database makes CHESHIRE one of the largest systems that has ever been used for IR research and experimentation.

As it is well known, existing online catalogs in use are based on Boolean logic and simple keyword matching techniques, which are hard to use, brittle, and unforgiving: more than one third of the searches retrieve nothing! CHESHIRE, on the other hand, offers further improvements: it accommodates search queries in natural language form. The user describes his/her information need using words that are taken from the natural language and submits this statement to CHESHIRE. CHESHIRE "evaluates" the query, identifies the records that are most similar to the user's query and comes up with a ranked list of "would-be" relevant records. Furthermore, CHESHIRE is able to incorporate users' relevance judgments through what is called "**relevance feedback process**," which increases the chance of retrieving more relevant documents. That is to say, no matter how poorly the information need is explained, CHESHIRE always retrieves some relevant documents and it helps users to clarify their intentions by way of relevance feedback mechanism. Such features are lacking in traditional online catalog systems.

As for my research, I am trying to find out the causes of search failures in online catalogs. Catalog searches may fail due to a variety of reasons such as a clunky user interface, indexing and vocabulary problems, rigid command languages and retrieval rules. Findings to be obtained from this research can be used in designing better online library catalogs. Designers equipped with information about search failures should be able to develop more robust and "fail-proof" online catalogs. The size of the CHESHIRE database offers a remarkable opportunity to obtain more reliable research results since most IR experiments in the past have been conducted on small test collections, findings of which do not necessarily "scale-up" to large bibliographic databases.

You are to play a very important role in this research. I am sure you are familiar with some other online catalogs. Yet most, if not all, of you presumably never used CHESHIRE before. You are kindly requested to try CHESHIRE and do some searches on it. We will record your entire search (i.e., query entered, records displayed, relevance judgments) in a transaction file so that we can later analyze these records and determine the retrieval effectiveness of CHESHIRE.

Here is what I would like you to do:

1. Go to the Computer Laboratory in the second floor of South Hall and log on to the SLIS Local Area Network using your regular login and password.

2. Once you are in the Main menu (of the SLIS network), follow the instructions in the document entitled "**Access to CHESHIRE: An Experimental Online Catalog**" in order to connect to CHESHIRE. (This document will be handed out in one of your classes.)

3. When you get to the disappearing cat screen (smiling CHESHIRE cat), type a natural language query of your choice (an example of the search process is provided in the above document). Please note that since the CHESHIRE database is restricted to the holdings of the Library School Library in South Hall, questions that could be answered from this collection will get the best results.

   Examples of **subfields** which are supported by the Library School Library are as follows: librarianship and information science in general, publishing and the book arts, management of libraries and information services, bibliographic organization, censorship and copyright, children's literature, printing and publishing, information policy, information retrieval, systems analysis and automation of libraries, archives and records management, office information systems, use of computers in libraries and information services.

4. Mark relevant clusters and records by pressing "s" and skim through records. (You might want to write down or download relevant ones so that you can obtain them from the stacks of the Library School Library.)

5. Perform a "relevance feedback search" and see if it improves the search results.

6. Repeat the above process whenever you have a query that can be answered from the collection of the Library School Library.

7. You might wish to try the same searches on GLADIS or MELVYL and compare the results with that which you obtained through CHESHIRE.

8. After a couple of searches on CHESHIRE, you will be able to compare and contrast the following features of online catalogs:

   a) natural language-based queries vs. command languages (e.g., "subject access in online library catalogs" (in CHESHIRE) vs. "FIND SU SUBJECT ACCESS AND ONLINE CATALOGS" (in MELVYL)).

   b) relevance feedback mechanism and its use in CHESHIRE.

   c) use of LC classification system for subject access in CHESHIRE.

   d) "information overload" and "zero retrieval" in traditional online catalogs. (CHESHIRE solves "overload" problem by presenting records in ranked order so that the most promising records will be displayed first. CHESHIRE almost always retrieves something from the database, unless there is no record in the database for a given query.)

yt 9/91

APPENDIX B

ACCESS TO CHESHIRE: AN EXPERIMENTAL ONLINE CATALOG

(Instructions)

This appendix reproduces a handout distributed to MLIS and Ph.D. students who agreed to participate in the study. It includes step-by-step instructions as to how to get access to CHESHIRE through the local area network of the School of Library and Information Studies at the University of California at Berkeley. It also explains how to perform an online search on CHESHIRE.

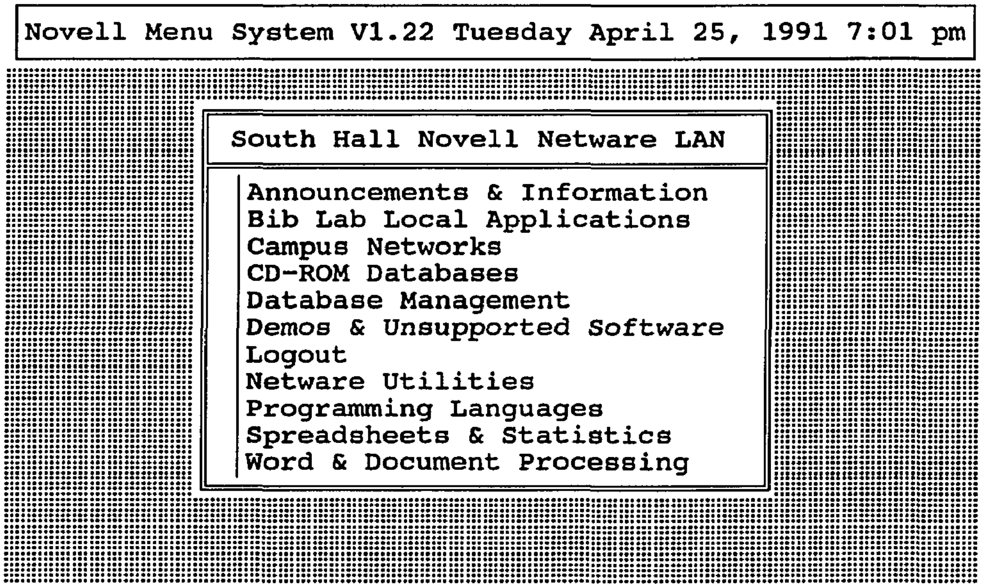
ACCESS TO CHESHIRE:
AN EXPERIMENTAL ONLINE CATALOG
(INSTRUCTIONS)

by

Yasar Tonta

Berkeley
September 1991

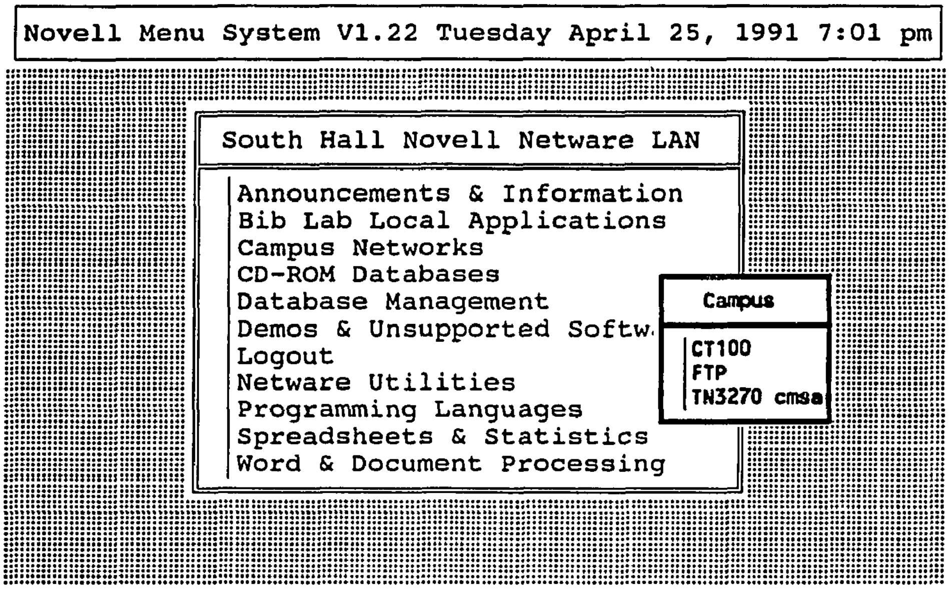
SOUTH HALL LAN MENU: 6-25-91



```
Novell Menu System V1.22 Tuesday April 25, 1991 7:01 pm

South Hall Novell Netware LAN

Announcements & Information
Bib Lab Local Applications
Campus Networks
CD-ROM Databases
Database Management
Demos & Unsupported Software
Logout
Netware Utilities
Programming Languages
Spreadsheets & Statistics
Word & Document Processing
```

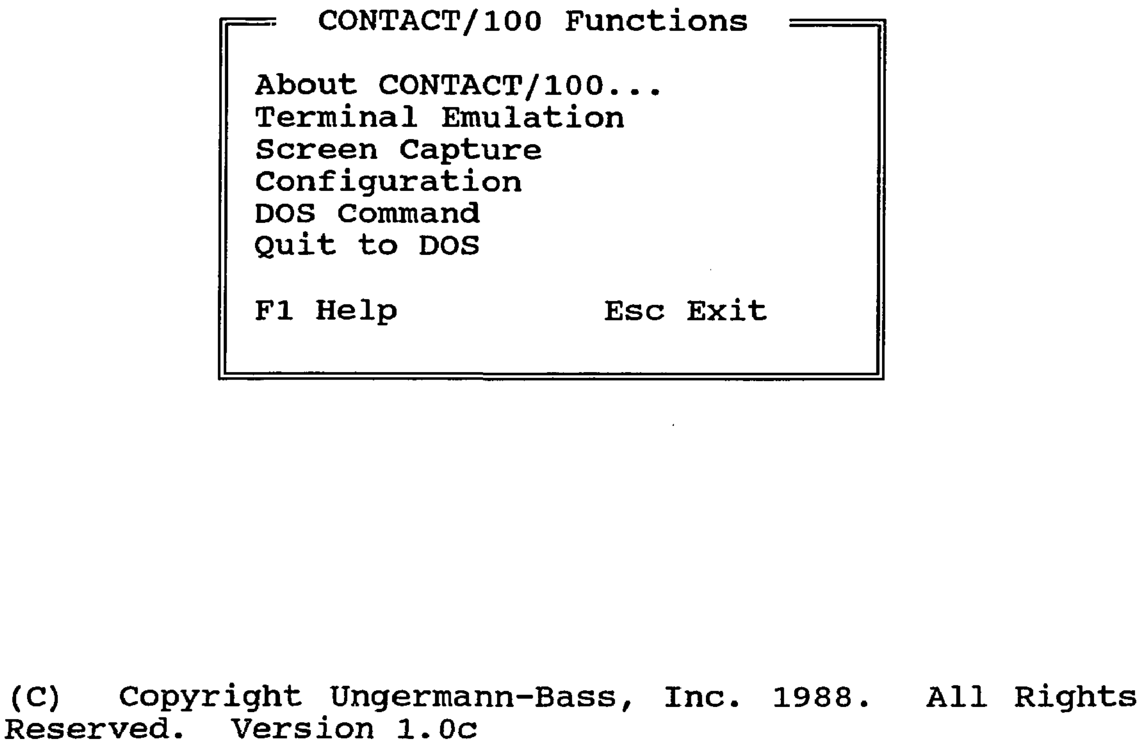
Select **Campus Networks** from the Main Menu and hit <Enter>.

SOUTH HALL LAN MENU: 6-25-91



```
Novell Menu System V1.22 Tuesday April 25, 1991 7:01 pm

South Hall Novell Netware LAN

Announcements & Information
Bib Lab Local Applications
Campus Networks
CD-ROM Databases
Database Management
Demos & Unsupported Softw
Logout
Netware Utilities
Programming Languages
Spreadsheets & Statistics
Word & Document Processing

Campus

CT100
FTP
TN3270 cmsa
```

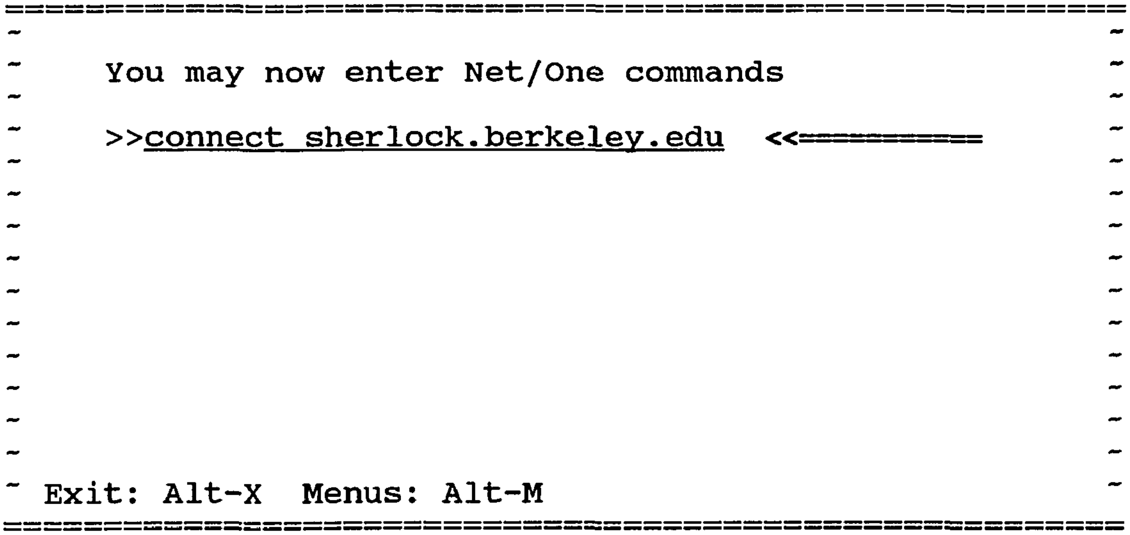
Select **CT100** from the Campus Networks menu and hit <Enter>.



(C) Copyright Ungermann-Bass, Inc. 1988. All Rights Reserved. Version 1.0c

Select **Terminal Emulation** from CT100 menu and hit <Enter>.

Hit <Enter> to get the ">>" prompt and type in the following line (your input is underlined) to connect to the Sherlock computer where the CHESHIRE system resides, and hit <Enter>.

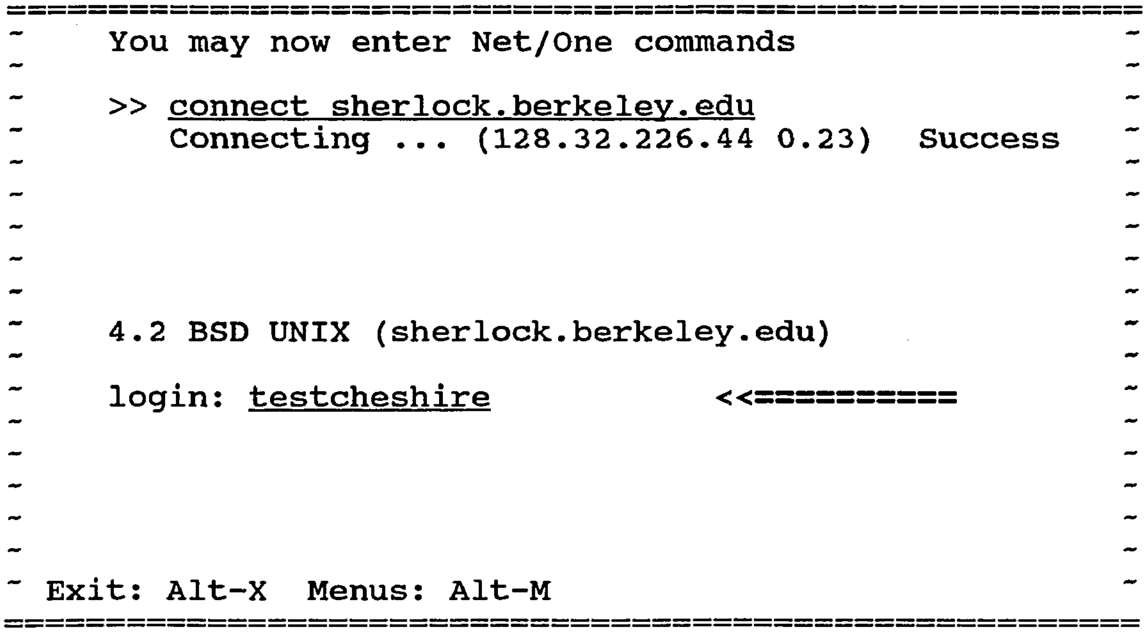


```
=====================================================================
~                                                                   ~
~    You may now enter Net/One commands                             ~
~                                                                   ~
~    >>connect sherlock.berkeley.edu  <<==========                  ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~                                                                   ~
~ Exit: Alt-X  Menus: Alt-M                                         ~
=====================================================================
```

After a few seconds the following lines will appear on your screen with the cursor blinking next to "**login:**". Computer is waiting for **login id** to be entered.

```
=======================================================
~    You may now enter Net/One commands               ~
~                                                     ~
~    >> connect sherlock.berkeley.edu                 ~
~       Connecting ... (128.32.226.44 0.23)  Success  ~
~                                                     ~
~                                                     ~
~                                                     ~
~    4.2 BSD UNIX (sherlock.berkeley.edu)             ~
~                                                     ~
~    login:              <<============               ~
~                                                     ~
~                                                     ~
~                                                     ~
~                                                     ~
~                                                     ~
~ Exit: Alt-X  Menus: Alt-M                           ~
=======================================================
```

Here, type in **testcheshire** and hit <Enter>.

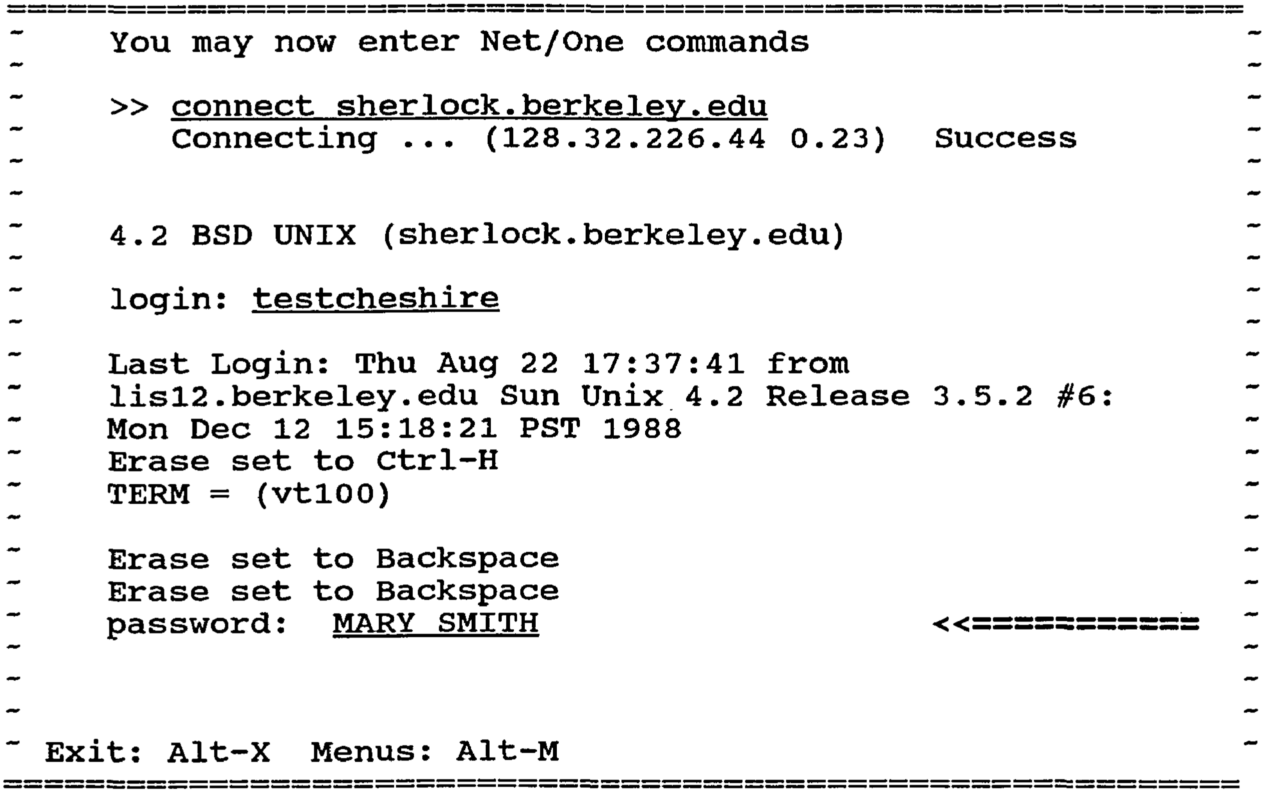


```
=========================================================
~    You may now enter Net/One commands                ~
~                                                      ~
~    >> connect sherlock.berkeley.edu                  ~
~       Connecting ... (128.32.226.44 0.23)  Success   ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~    4.2 BSD UNIX (sherlock.berkeley.edu)              ~
~                                                      ~
~    login: testcheshire           <<==========        ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~ Exit: Alt-X  Menus: Alt-M                            ~
=========================================================
```

After a couple of lines of text (usual Unix greeting information), the cursor will be next to "TERM = (vt100)". Simply hit <Enter>.

```
========================================================
~    You may now enter Net/One commands                ~
~                                                      ~
~    >> connect sherlock.berkeley.edu                  ~
~       Connecting ... (128.32.226.44 0.23)  Success   ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~                                                      ~
~    4.2 BSD UNIX (sherlock.berkeley.edu)              ~
~                                                      ~
~    login: testcheshire                               ~
~                                                      ~
~    Last Login: Thu Aug 22 17:37:41 from              ~
~    lis12.berkeley.edu Sun Unix 4.2 Release 3.5.2 #6: ~
~    Mon Dec 12 15:18:21 PST 1988                      ~
~    Erase set to Ctrl-H                               ~
~                                                      ~
~    TERM = (vt100)              <<=============       ~
~                                                      ~
~                                                      ~
~ Exit: Alt-X  Menus: Alt-M                            ~
========================================================
```

It is now time to enter the "**password**" (ignore a few lines of information such as "Erase set to Backspace," etc.). Type in your **Firstname** and **Lastname** (space in between) and hit <Enter>.

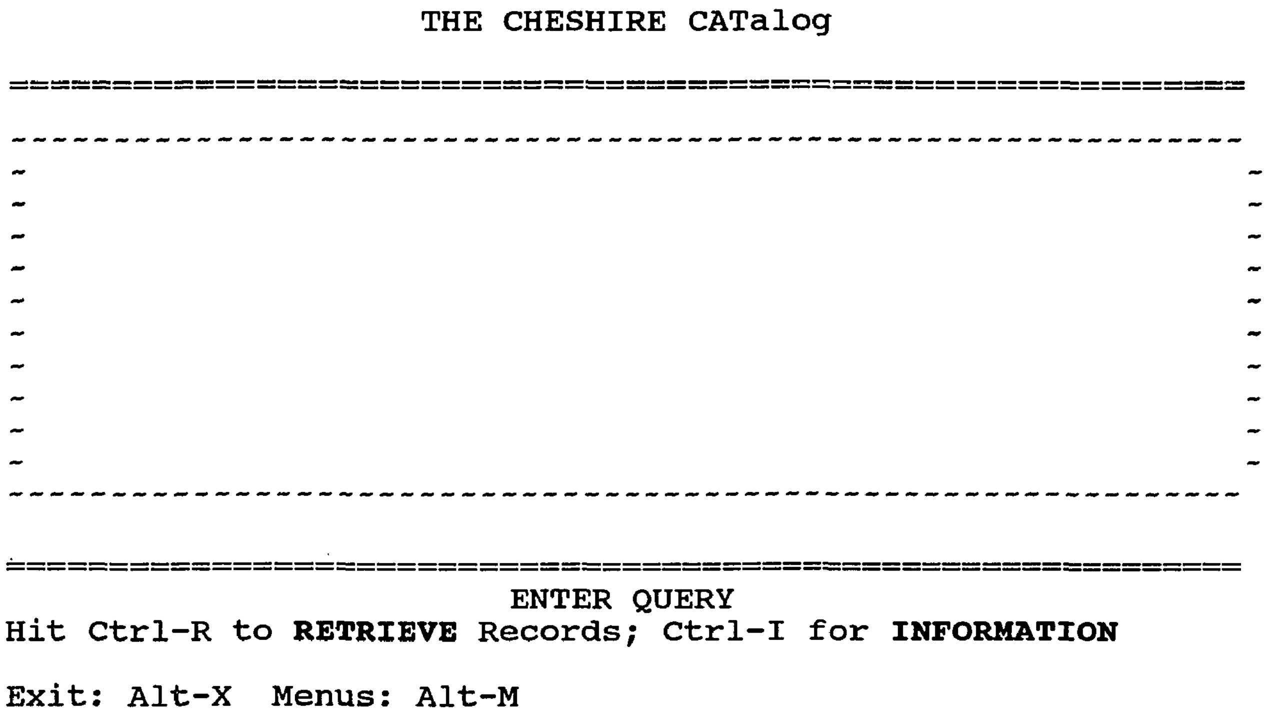


```
    You may now enter Net/One commands

    >> connect sherlock.berkeley.edu
       Connecting ... (128.32.226.44 0.23)  Success


    4.2 BSD UNIX (sherlock.berkeley.edu)

    login: testcheshire

    Last Login: Thu Aug 22 17:37:41 from
    lis12.berkeley.edu Sun Unix 4.2 Release 3.5.2 #6:
    Mon Dec 12 15:18:21 PST 1988
    Erase set to Ctrl-H
    TERM = (vt100)

    Erase set to Backspace
    Erase set to Backspace
    password:  MARY SMITH                      <<===========



 Exit: Alt-X  Menus: Alt-M
```

At this point the CHESHIRE system will start up and the following screen will appear (the disappearing cat).

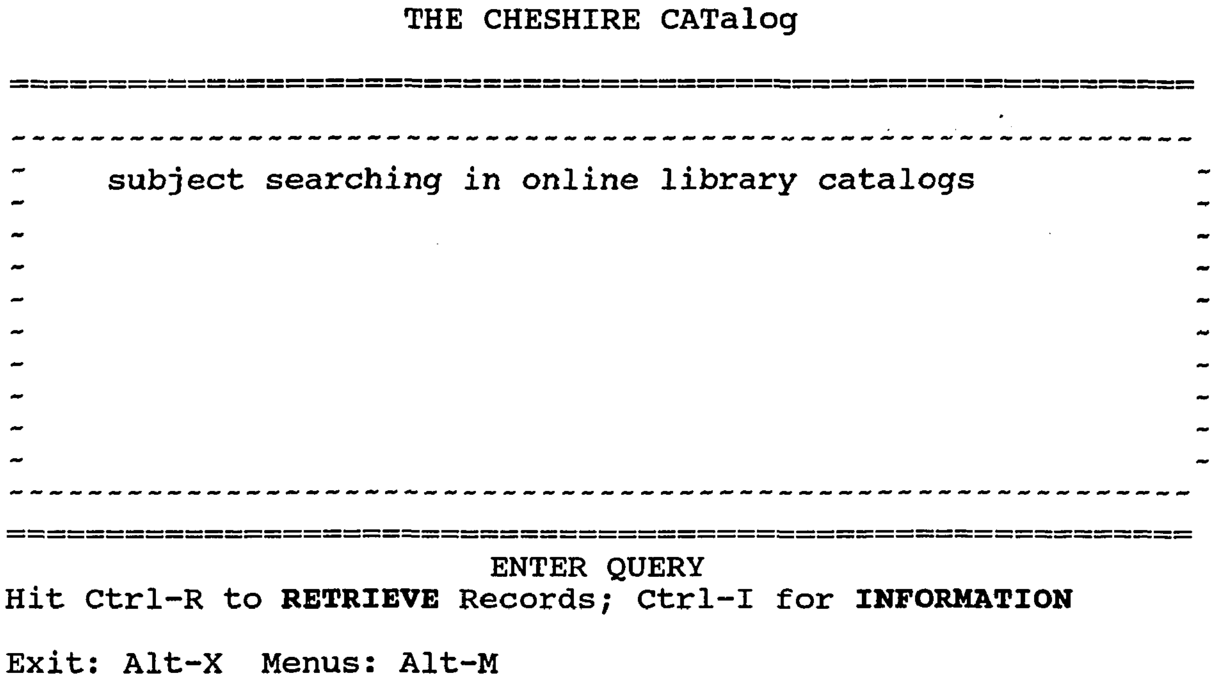
```
                         THE CHESHIRE CATalog
=====================================================================



                              /\__/\
                             { o__o }
                            >(-======-)<



=====================================================================
                        PRESS ANY KEY TO START

Exit: Alt-X     Menus: Alt-M
```

Once you are connected to CHESHIRE, press any key to begin searching.



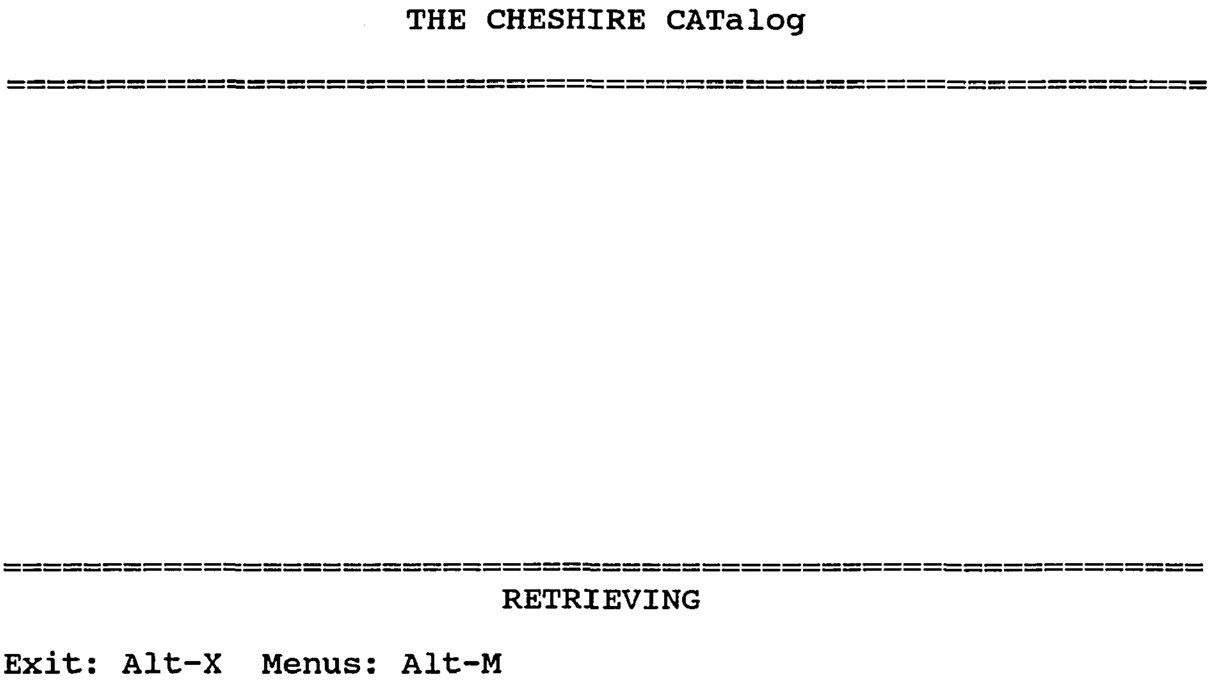
```
                          THE CHESHIRE CATalog
=================================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~                                                               ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

=================================================================
                              ENTER QUERY
Hit Ctrl-R to RETRIEVE Records; Ctrl-I for INFORMATION

Exit: Alt-X  Menus: Alt-M
```

The CHESHIRE search entry screen is a simple text editor. There is no query syntax or "command language."



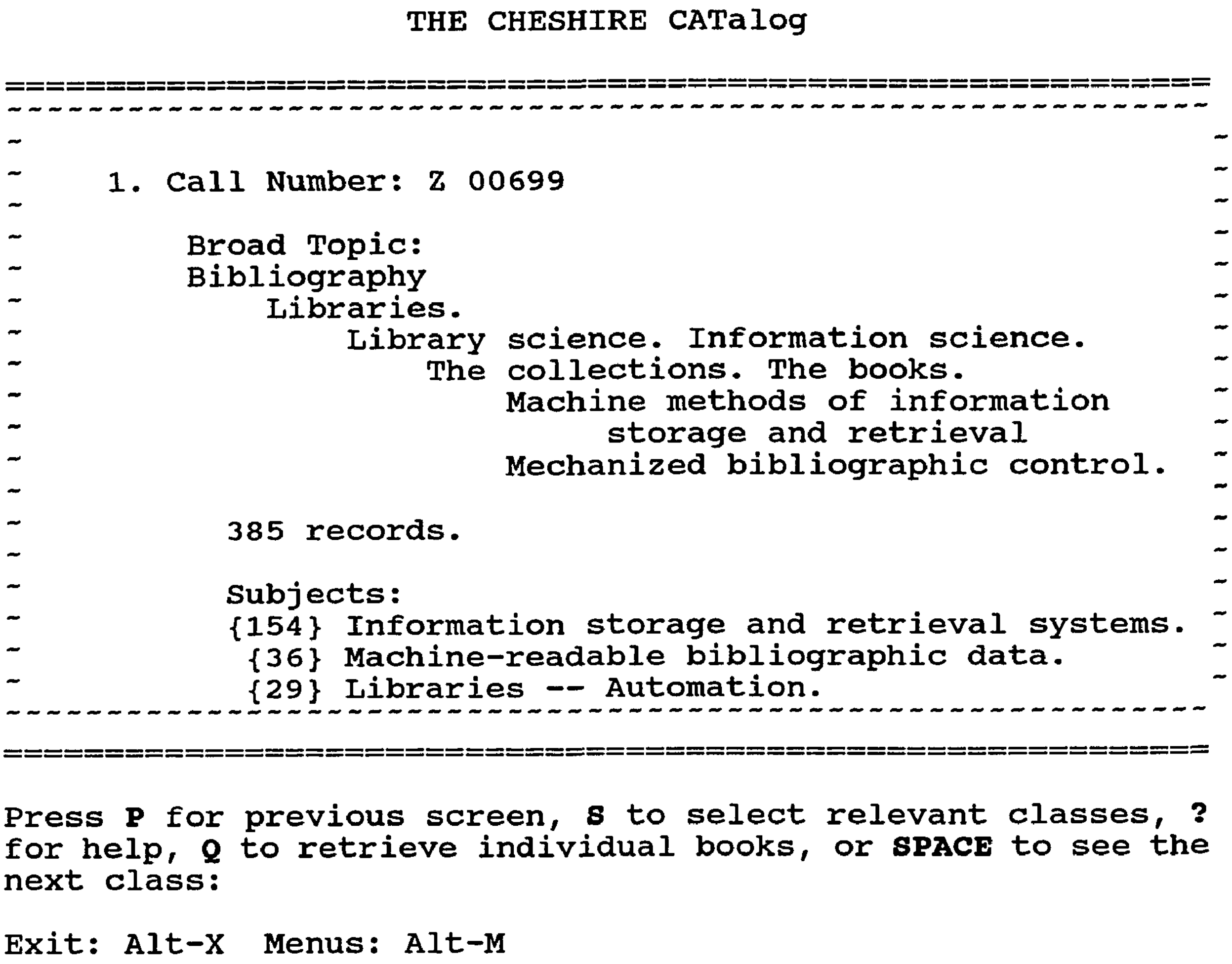
```
                      THE CHESHIRE CATalog
=============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    subject searching in online library catalogs           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~                                                           ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
=============================================================
                          ENTER QUERY
Hit Ctrl-R to RETRIEVE Records; Ctrl-I for INFORMATION

Exit: Alt-X  Menus: Alt-M
```

Just type in your query using as many words as desired (the more accurately you can specify your topic, the better the system will work). (Since the database is restricted to the holdings of the Library School Library, questions that could be answered from that library will get the best results.)



```
                    THE CHESHIRE CATalog
==============================================================




==============================================================
                          RETRIEVING
Exit: Alt-X  Menus: Alt-M
```

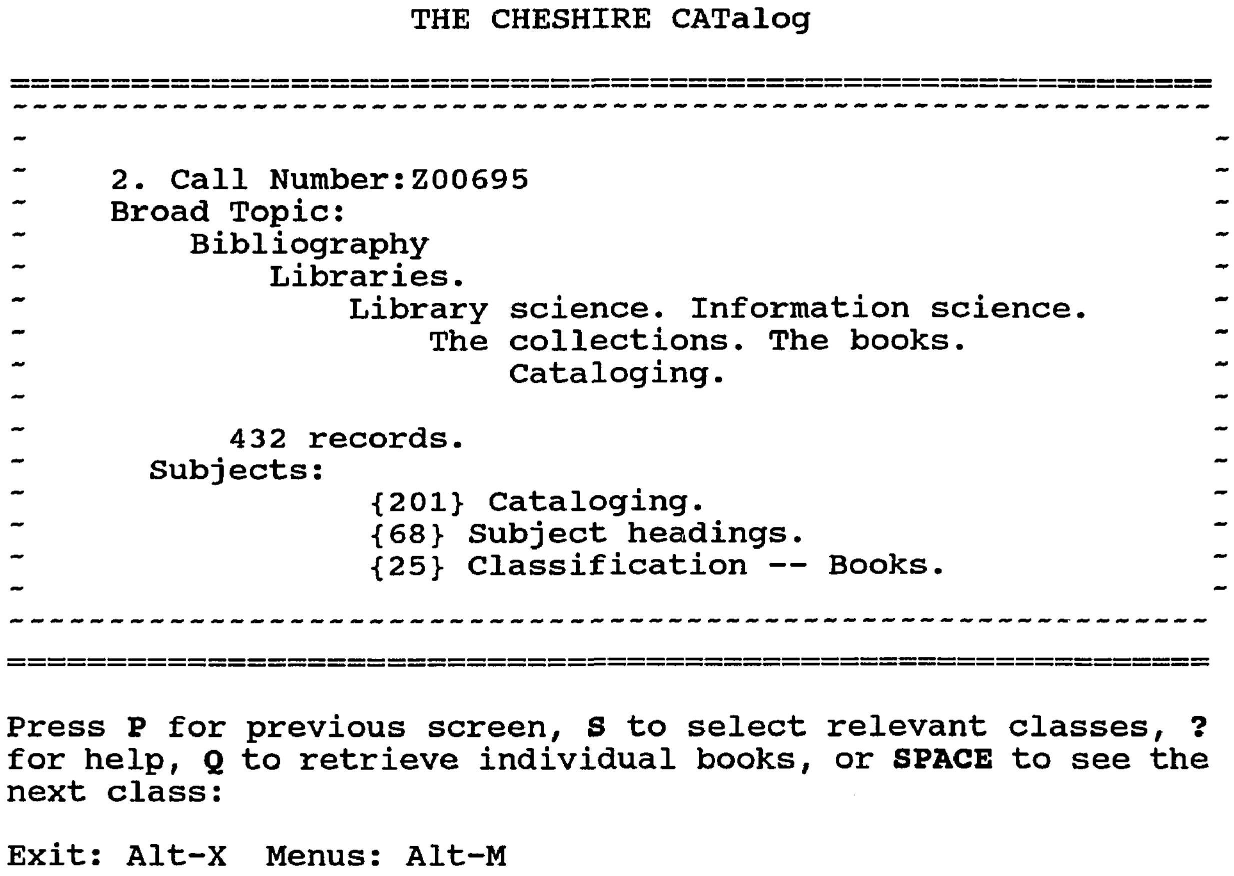
Press **Ctrl-R** to start the search (by holding down the Ctrl key and pressing R simultaneously.



```
                    THE CHESHIRE CATalog

==============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~                                                            ~
~    1. Call Number: Z 00699                                 ~
~                                                            ~
~        Broad Topic:                                        ~
~        Bibliography                                        ~
~            Libraries.                                      ~
~                Library science. Information science.       ~
~                    The collections. The books.             ~
~                        Machine methods of information      ~
~                             storage and retrieval          ~
~                        Mechanized bibliographic control.   ~
~                                                            ~
~          385 records.                                      ~
~                                                            ~
~          Subjects:                                         ~
~          {154} Information storage and retrieval systems.  ~
~           {36} Machine-readable bibliographic data.        ~
~           {29} Libraries -- Automation.                    ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
==============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

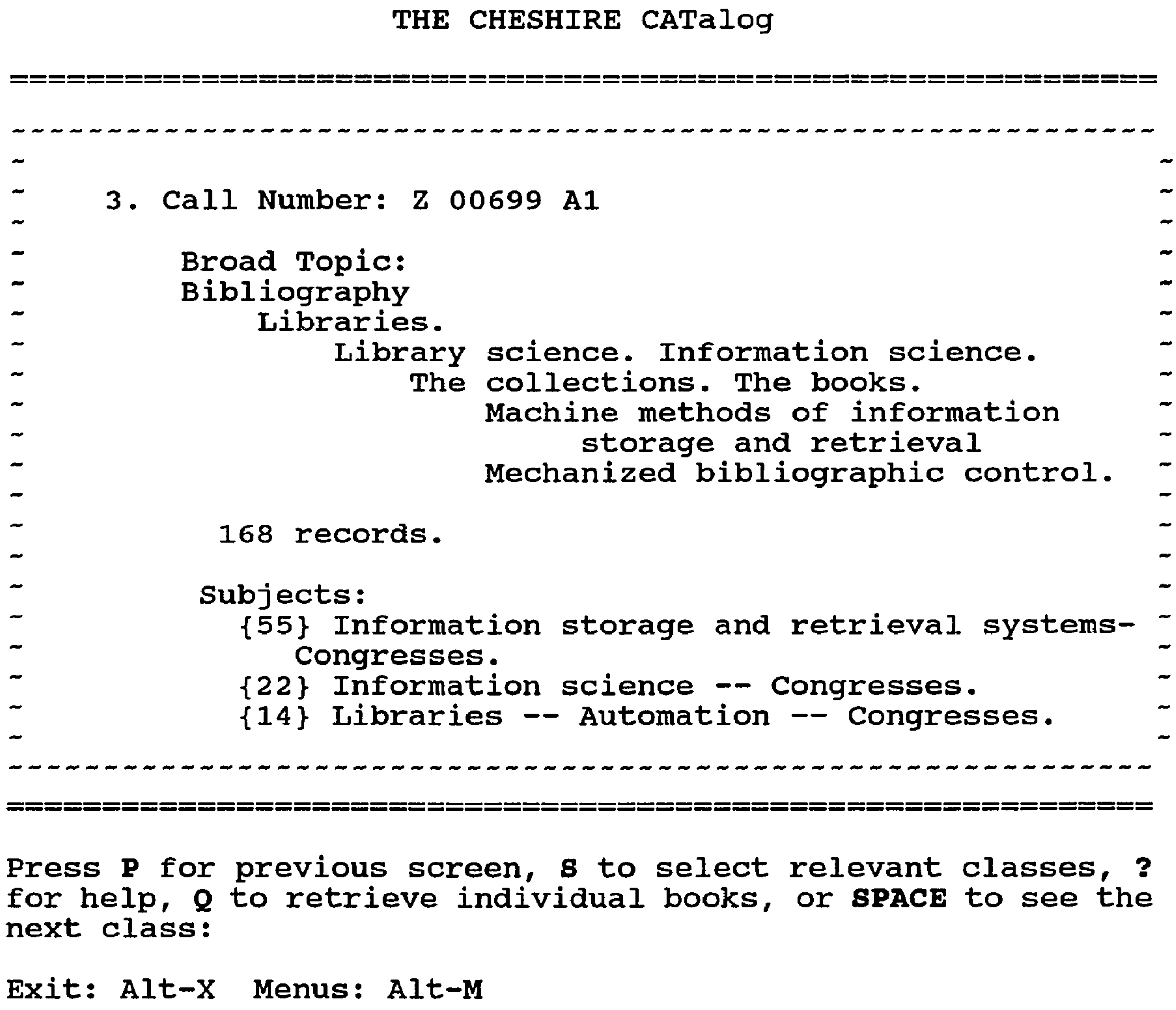
After evaluating the query, the system retrieves and displays the best matching subject headings and classification clusters.

If the cluster seems to be on a relevant topic, press "s" (for select). If not, press space bar to continue.



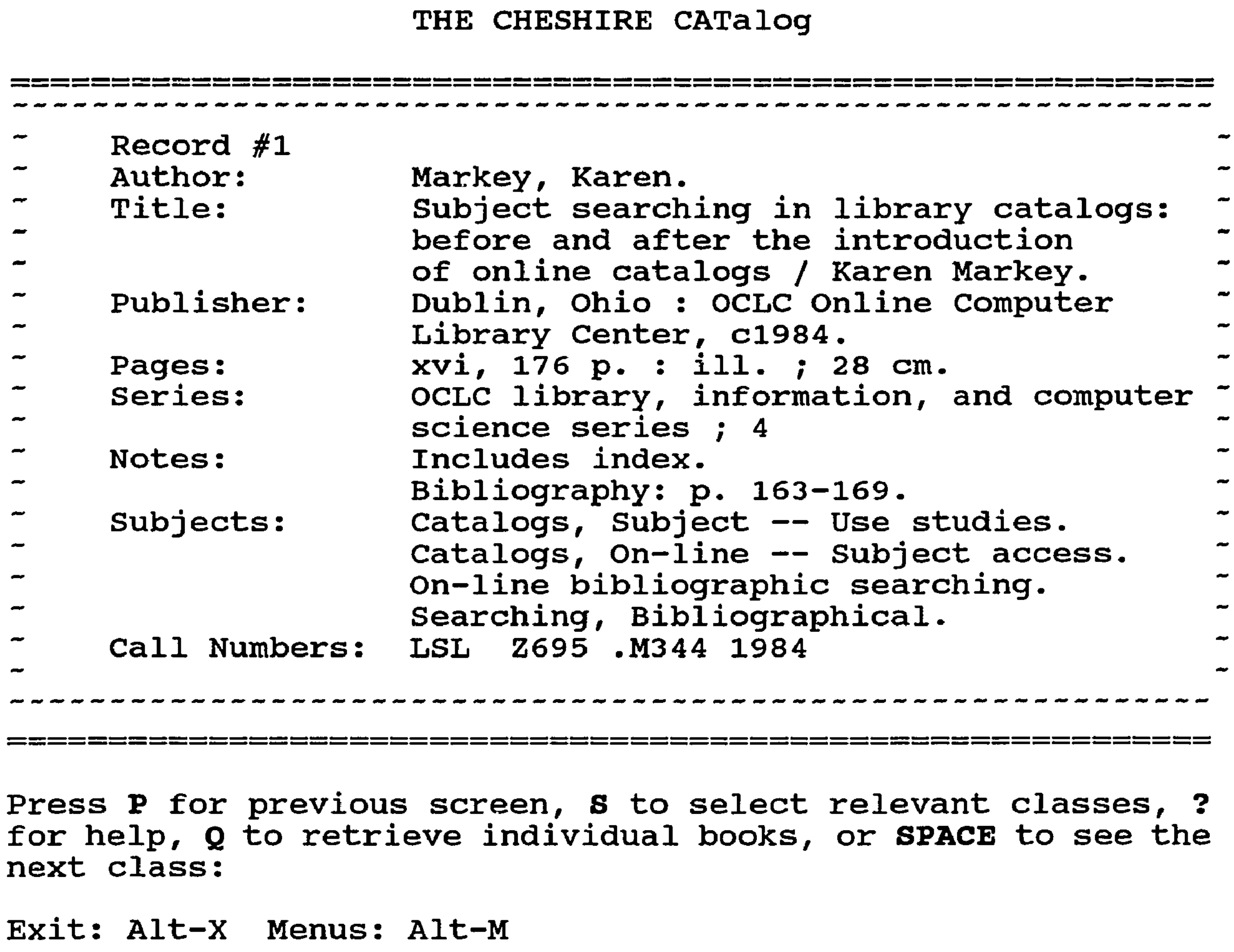
```
                         THE CHESHIRE CATalog

==============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~                                                            ~
~    2. Call Number:Z00695                                   ~
~    Broad Topic:                                            ~
~        Bibliography                                        ~
~            Libraries.                                      ~
~                Library science. Information science.      ~
~                    The collections. The books.            ~
~                        Cataloging.                        ~
~                                                            ~
~          432 records.                                      ~
~      Subjects:                                             ~
~                 {201} Cataloging.                          ~
~                 {68} Subject headings.                     ~
~                 {25} Classification -- Books.              ~
~                                                            ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
==============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

Selected clusters (i.e., subject headings and classification numbers) get added to the original query, thereby increasing the chance of finding more relevant books.



```
                    THE CHESHIRE CATalog
============================================================
------------------------------------------------------------
~                                                          ~
~    3. Call Number: Z 00699 A1                            ~
~                                                          ~
~        Broad Topic:                                      ~
~        Bibliography                                      ~
~            Libraries.                                    ~
~                Library science. Information science.     ~
~                    The collections. The books.           ~
~                        Machine methods of information    ~
~                             storage and retrieval        ~
~                        Mechanized bibliographic control. ~
~                                                          ~
~          168 records.                                    ~
~                                                          ~
~         Subjects:                                        ~
~           {55} Information storage and retrieval systems- ~
~              Congresses.                                 ~
~           {22} Information science -- Congresses.        ~
~           {14} Libraries -- Automation -- Congresses.    ~
~                                                          ~
------------------------------------------------------------
============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

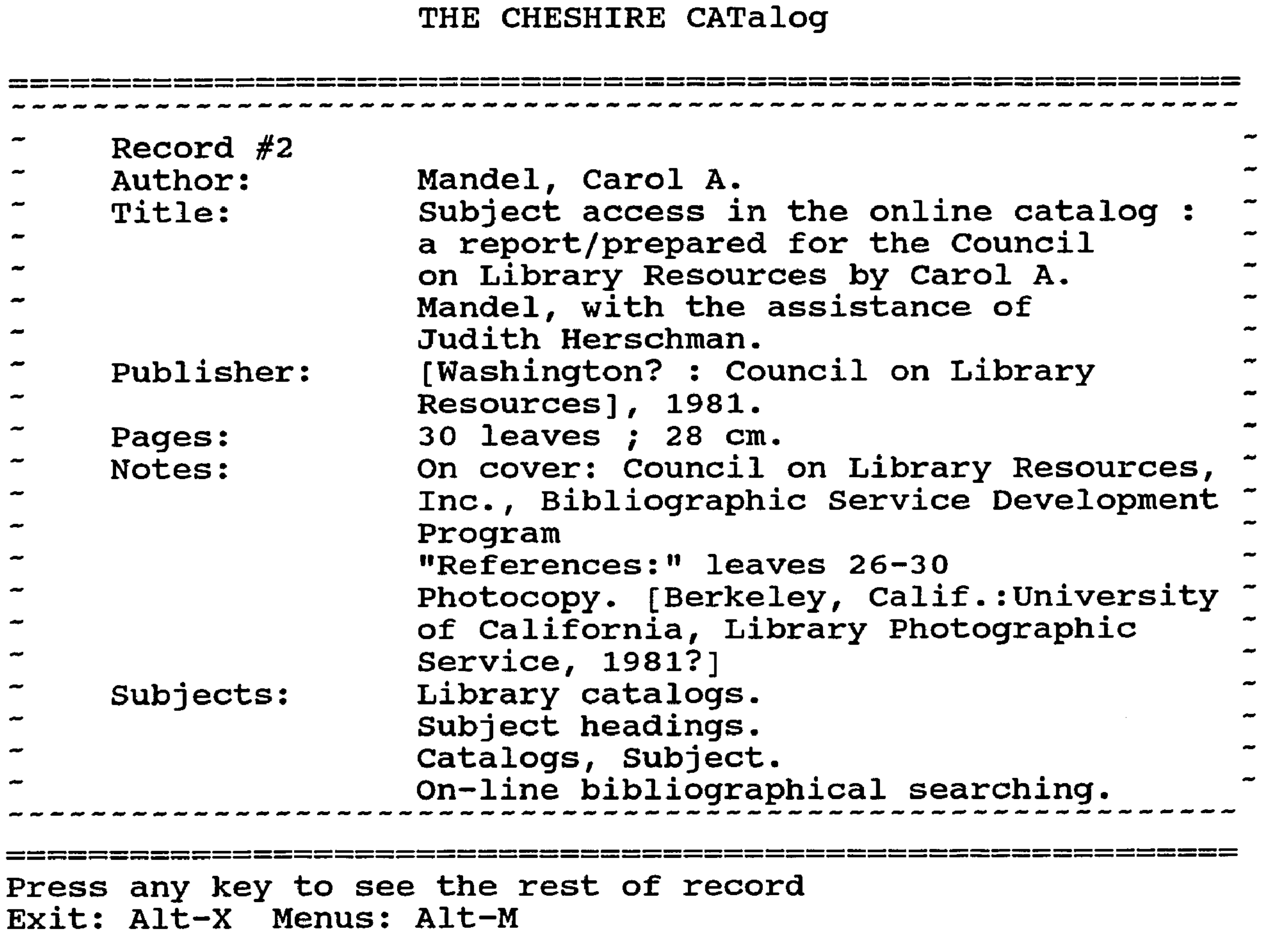
After selecting a few clusters (at least one, at most 20), press "q" to retrieve individual books.



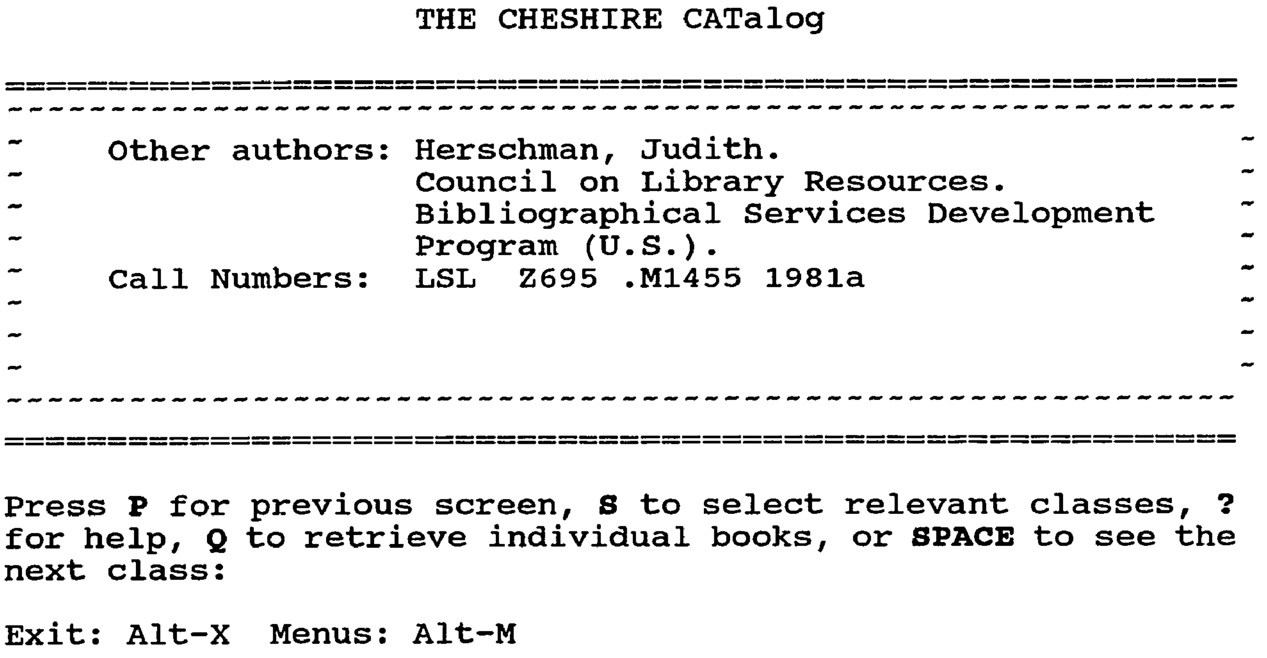
```
                         THE CHESHIRE CATalog

================================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Record #1                                                 ~
~    Author:         Markey, Karen.                            ~
~    Title:          Subject searching in library catalogs:    ~
~                    before and after the introduction         ~
~                    of online catalogs / Karen Markey.        ~
~    Publisher:      Dublin, Ohio : OCLC Online Computer       ~
~                    Library Center, c1984.                    ~
~    Pages:          xvi, 176 p. : ill. ; 28 cm.               ~
~    Series:         OCLC library, information, and computer   ~
~                    science series ; 4                        ~
~    Notes:          Includes index.                           ~
~                    Bibliography: p. 163-169.                 ~
~    Subjects:       Catalogs, Subject -- Use studies.         ~
~                    Catalogs, On-line -- Subject access.      ~
~                    On-line bibliographic searching.          ~
~                    Searching, Bibliographical.               ~
~    Call Numbers:   LSL  Z695 .M344 1984                      ~
~                                                              ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
================================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

Based on the query and selected clusters, CHESHIRE now retrieves the best matching bibliographic records. Press "s" if the item seems relevant to the query. Press **space bar** if not.

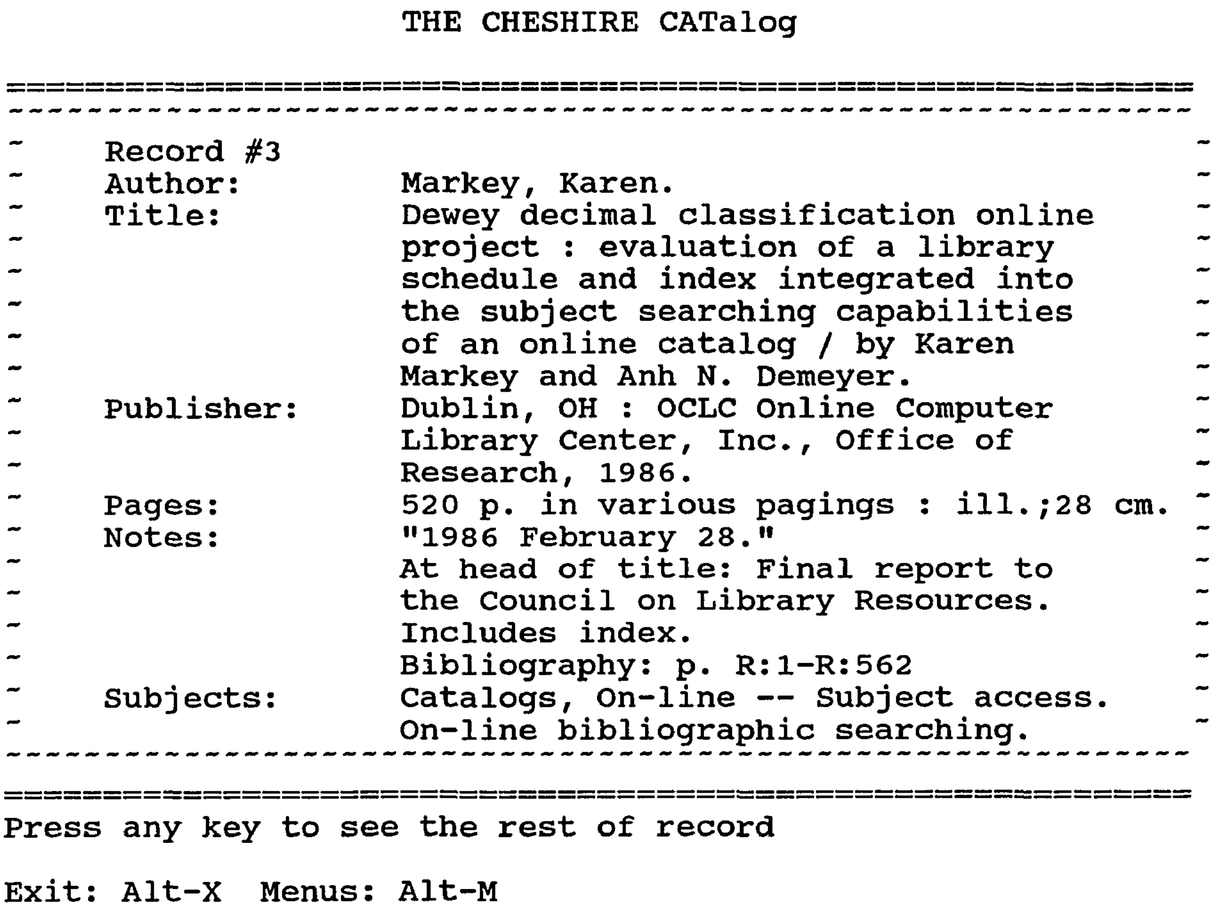
(All items are available in the stacks of the Library School Library.)



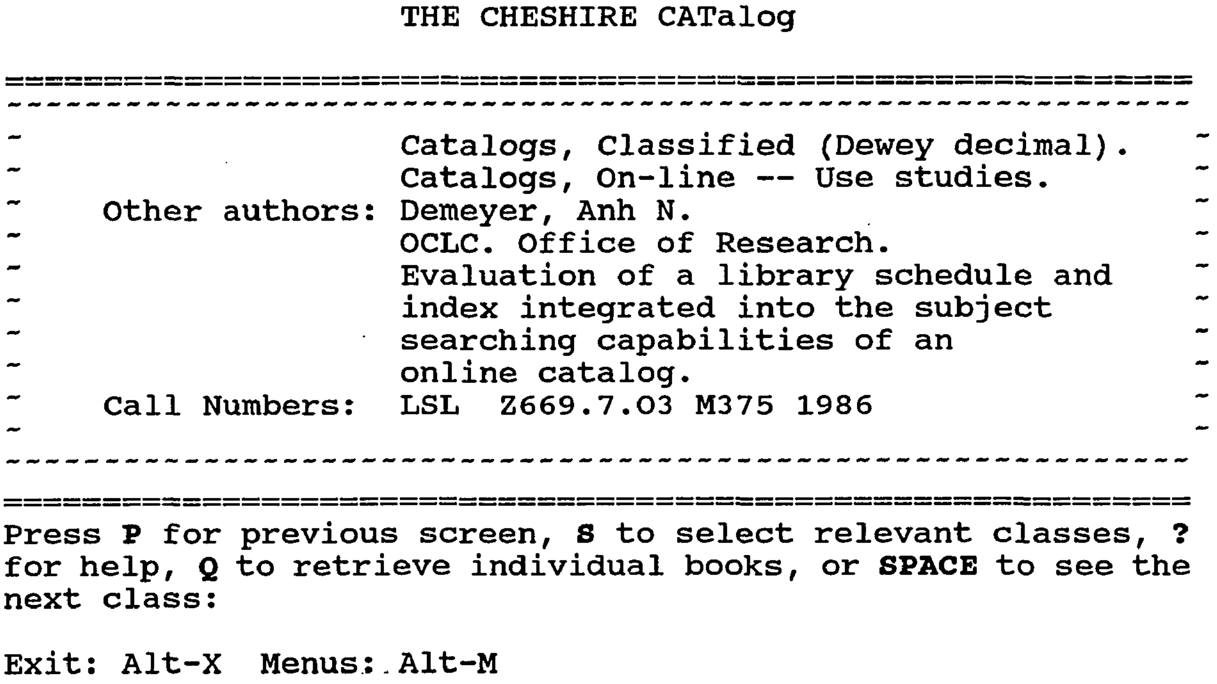
```
                          THE CHESHIRE CATalog

============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Record #2                                              ~
~    Author:         Mandel, Carol A.                       ~
~    Title:          Subject access in the online catalog : ~
~                    a report/prepared for the Council      ~
~                    on Library Resources by Carol A.       ~
~                    Mandel, with the assistance of         ~
~                    Judith Herschman.                      ~
~    Publisher:      [Washington? : Council on Library      ~
~                    Resources], 1981.                      ~
~    Pages:          30 leaves ; 28 cm.                     ~
~    Notes:          On cover: Council on Library Resources, ~
~                    Inc., Bibliographic Service Development ~
~                    Program                                ~
~                    "References:" leaves 26-30             ~
~                    Photocopy. [Berkeley, Calif.:University ~
~                    of California, Library Photographic    ~
~                    Service, 1981?]                        ~
~    Subjects:       Library catalogs.                      ~
~                    Subject headings.                      ~
~                    Catalogs, Subject.                     ~
~                    On-line bibliographical searching.     ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

============================================================
Press any key to see the rest of record
Exit: Alt-X  Menus: Alt-M
```



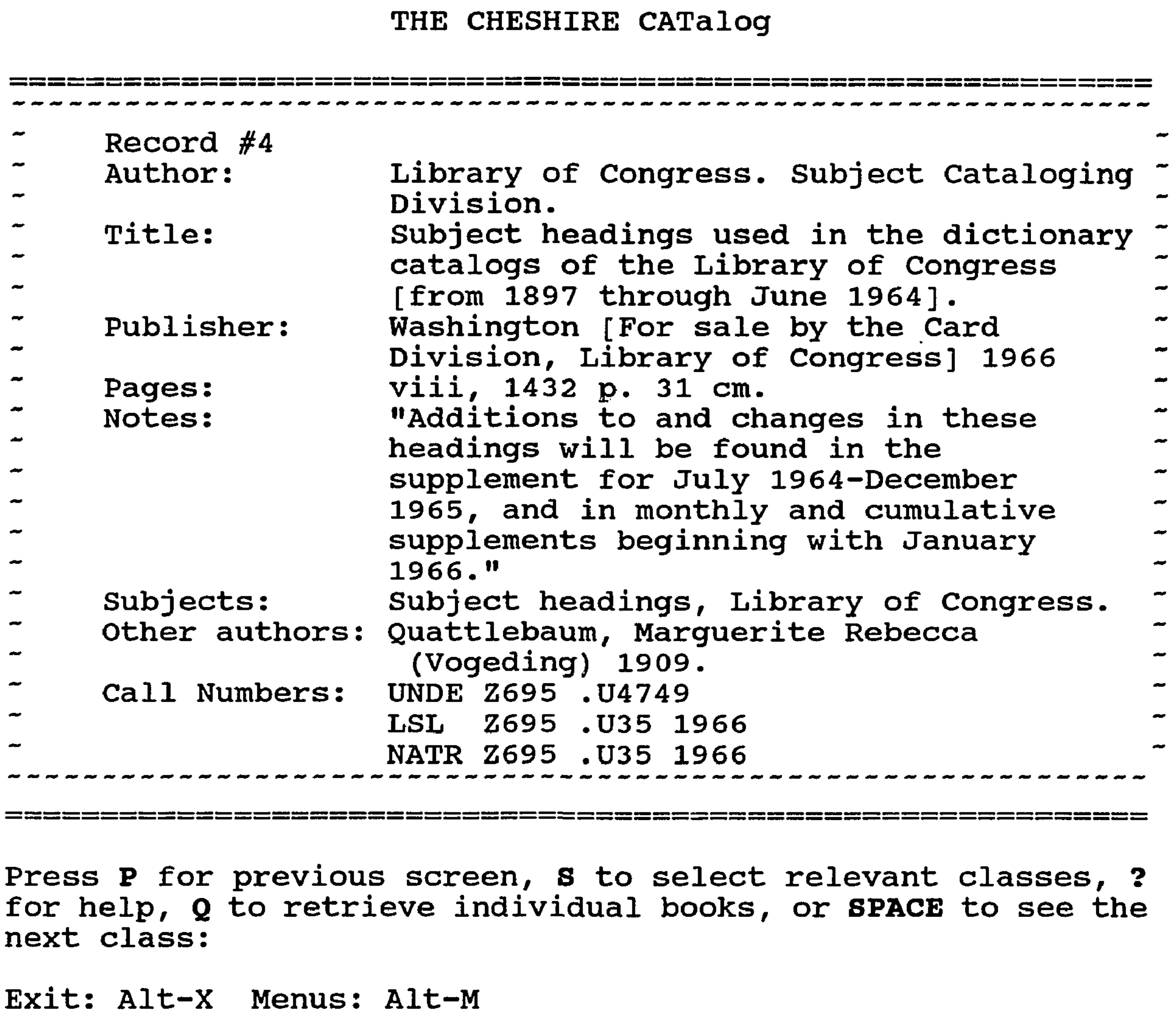
```
                          THE CHESHIRE CATalog

============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Other authors: Herschman, Judith.                      ~
~                   Council on Library Resources.           ~
~                   Bibliographical Services Development    ~
~                   Program (U.S.).                         ~
~    Call Numbers:  LSL  Z695 .M1455 1981a                  ~
~                                                           ~
~                                                           ~
~                                                           ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```



```
                    THE CHESHIRE CATalog

============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Record #3                                             ~
~    Author:        Markey, Karen.                         ~
~    Title:         Dewey decimal classification online    ~
~                   project : evaluation of a library      ~
~                   schedule and index integrated into     ~
~                   the subject searching capabilities     ~
~                   of an online catalog / by Karen        ~
~                   Markey and Anh N. Demeyer.             ~
~    Publisher:     Dublin, OH : OCLC Online Computer      ~
~                   Library Center, Inc., Office of        ~
~                   Research, 1986.                        ~
~    Pages:         520 p. in various pagings : ill.;28 cm. ~
~    Notes:         "1986 February 28."                    ~
~                   At head of title: Final report to      ~
~                   the Council on Library Resources.      ~
~                   Includes index.                        ~
~                   Bibliography: p. R:1-R:562             ~
~    Subjects:      Catalogs, On-line -- Subject access.   ~
~                   On-line bibliographic searching.       ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
============================================================
Press any key to see the rest of record

Exit: Alt-X  Menus: Alt-M
```

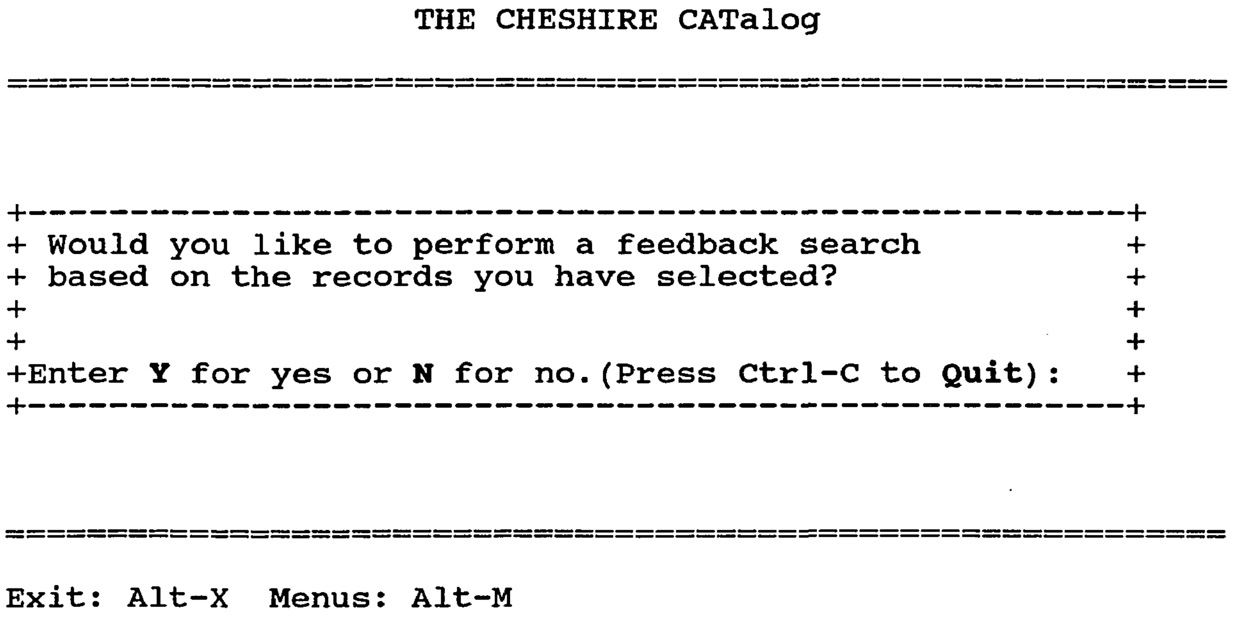


```
                    THE CHESHIRE CATalog

============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~                   Catalogs, Classified (Dewey decimal).  ~
~                   Catalogs, On-line -- Use studies.      ~
~    Other authors: Demeyer, Anh N.                        ~
~                   OCLC. Office of Research.              ~
~                   Evaluation of a library schedule and   ~
~                   index integrated into the subject      ~
~                   searching capabilities of an           ~
~                   online catalog.                        ~
~    Call Numbers:  LSL  Z669.7.03 M375 1986               ~
~                                                          ~
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
============================================================
Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```



```
                      THE CHESHIRE CATalog

============================================================
~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
~    Record #4                                                ~
~    Author:        Library of Congress. Subject Cataloging ~
~                   Division.                                 ~
~    Title:         Subject headings used in the dictionary ~
~                   catalogs of the Library of Congress       ~
~                   [from 1897 through June 1964].            ~
~    Publisher:     Washington [For sale by the Card          ~
~                   Division, Library of Congress] 1966       ~
~    Pages:         viii, 1432 p. 31 cm.                      ~
~    Notes:         "Additions to and changes in these        ~
~                   headings will be found in the             ~
~                   supplement for July 1964-December         ~
~                   1965, and in monthly and cumulative       ~
~                   supplements beginning with January        ~
~                   1966."                                    ~
~    Subjects:      Subject headings, Library of Congress.  ~
~    Other authors: Quattlebaum, Marguerite Rebecca           ~
~                    (Vogeding) 1909.                         ~
~    Call Numbers:  UNDE Z695 .U4749                          ~
~                   LSL  Z695 .U35 1966                       ~
~                   NATR Z695 .U35 1966                       ~
~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~ ~
============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

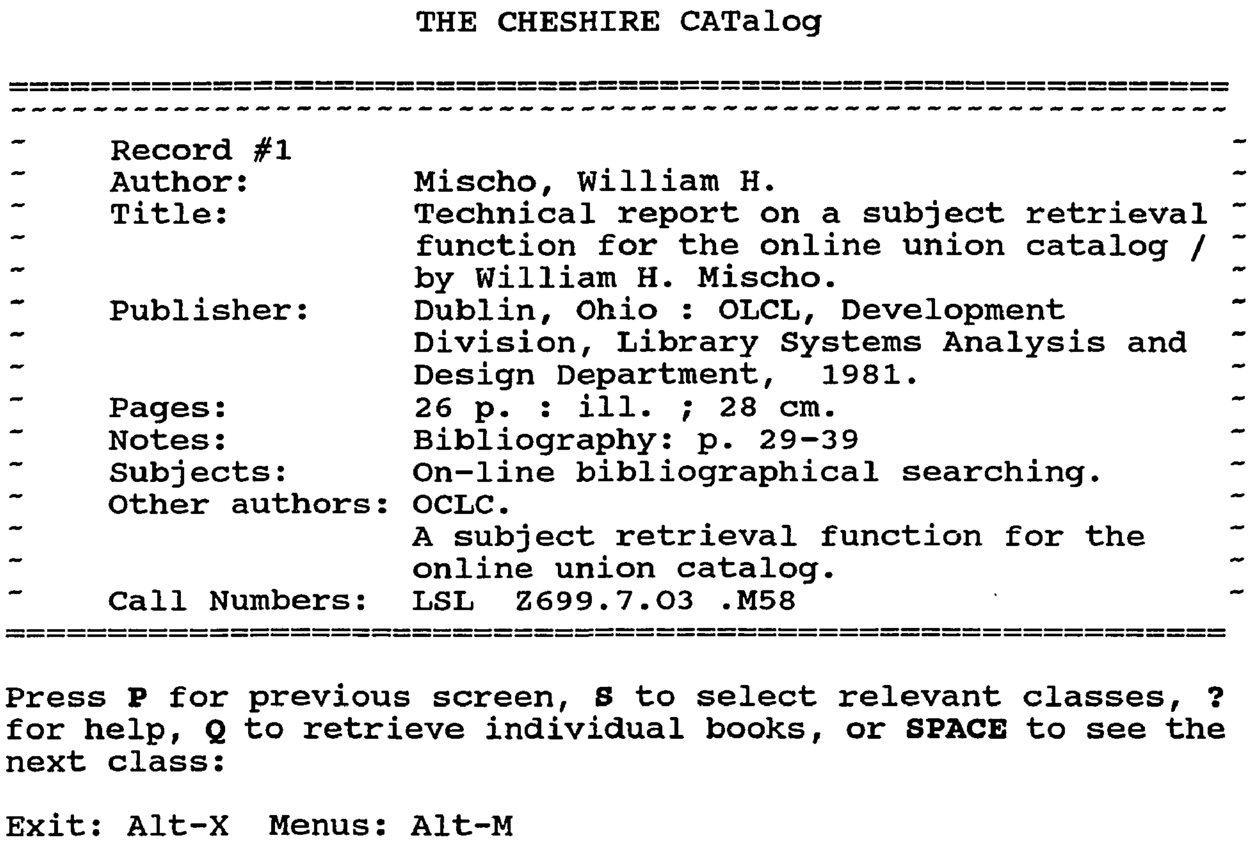
Press "q" to quit.



```
                    THE CHESHIRE CATalog
==========================================================


+------------------------------------------------------+
+ Would you like to perform a feedback search          +
+ based on the records you have selected?              +
+                                                      +
+                                                      +
+Enter Y for yes or N for no.(Press Ctrl-C to Quit):   +
+------------------------------------------------------+


==========================================================

Exit: Alt-X  Menus: Alt-M
```

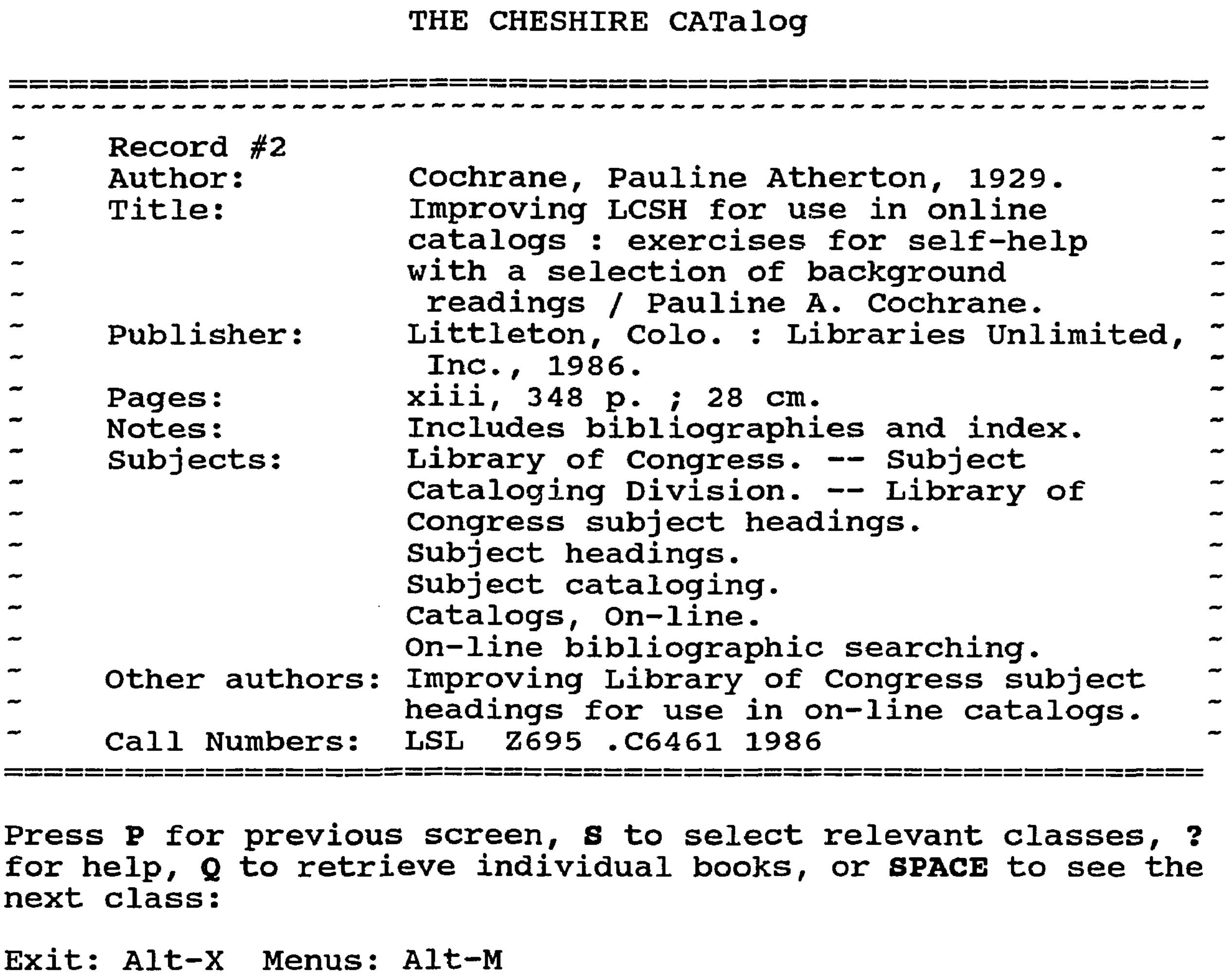
CHESHIRE asks if the user would like to perform a **relevance feedback** search based on the items selected as relevant. If the answer is "**Yes**", the system will try to find additional items that are similar to those that have been selected as relevant. If the answer is "**No**", the opening screen (disappearing cat) will be displayed to start a new search.

Press **Ctrl-C** to leave the system.



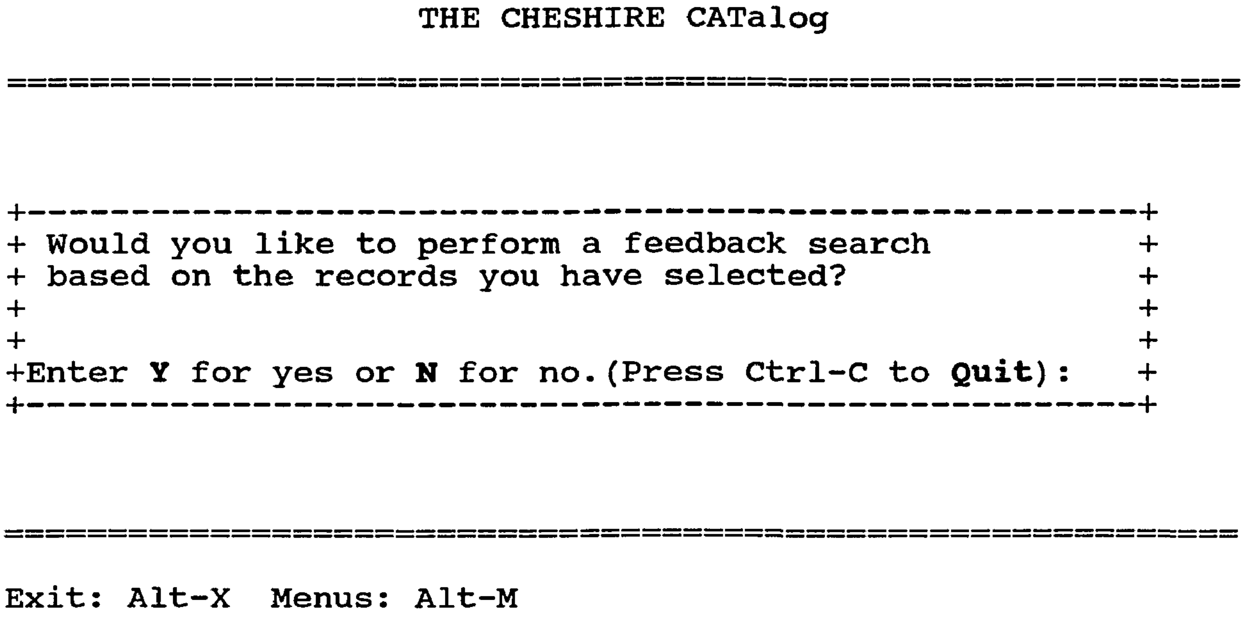
```
                    THE CHESHIRE CATalog

==============================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Record #1                                               ~
~    Author:         Mischo, William H.                      ~
~    Title:          Technical report on a subject retrieval ~
~                    function for the online union catalog / ~
~                    by William H. Mischo.                   ~
~    Publisher:      Dublin, Ohio : OLCL, Development        ~
~                    Division, Library Systems Analysis and  ~
~                    Design Department,  1981.               ~
~    Pages:          26 p. : ill. ; 28 cm.                   ~
~    Notes:          Bibliography: p. 29-39                  ~
~    Subjects:       On-line bibliographical searching.      ~
~    Other authors:  OCLC.                                   ~
~                    A subject retrieval function for the    ~
~                    online union catalog.                   ~
~    Call Numbers:   LSL  Z699.7.O3 .M58                     ~
==============================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

If the item seems to be relevant, press "s" (for select). If not, press the **space** bar to continue.



```
                         THE CHESHIRE CATalog

=================================================================
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
~    Record #2                                                  ~
~    Author:          Cochrane, Pauline Atherton, 1929.         ~
~    Title:           Improving LCSH for use in online          ~
~                     catalogs : exercises for self-help        ~
~                     with a selection of background            ~
~                      readings / Pauline A. Cochrane.          ~
~    Publisher:       Littleton, Colo. : Libraries Unlimited,   ~
~                      Inc., 1986.                              ~
~    Pages:           xiii, 348 p. ; 28 cm.                     ~
~    Notes:           Includes bibliographies and index.        ~
~    Subjects:        Library of Congress. -- Subject           ~
~                     Cataloging Division. -- Library of        ~
~                     Congress subject headings.                ~
~                     Subject headings.                         ~
~                     Subject cataloging.                       ~
~                     Catalogs, On-line.                        ~
~                     On-line bibliographic searching.          ~
~    Other authors:   Improving Library of Congress subject     ~
~                     headings for use in on-line catalogs.     ~
~    Call Numbers:    LSL  Z695 .C6461 1986                     ~
=================================================================

Press P for previous screen, S to select relevant classes, ?
for help, Q to retrieve individual books, or SPACE to see the
next class:

Exit: Alt-X  Menus: Alt-M
```

Page through the records by either "selecting" or pressing the **space bar**. Press "q" to stop the search.



```
                     THE CHESHIRE CATalog
==============================================================


+-------------------------------------------------------+
+ Would you like to perform a feedback search           +
+ based on the records you have selected?               +
+                                                       +
+                                                       +
+Enter Y for yes or N for no.(Press Ctrl-C to Quit):    +
+-------------------------------------------------------+


==============================================================

Exit: Alt-X  Menus: Alt-M
```

Relevance feedback process can be repeated as many times as necessary by pressing "Yes." To start a new search, press "**No.**" Press **Ctrl-C** to exit CHESHIRE.

APPENDIX C

TRANSACTION LOG RECORD FORMAT

This appendix describes the types of data captured for each search query in the transaction log record.

TRANSACTION LOG RECORD FORMAT

Field Name	*Description*
User's password	Each user's password is captured during logon process
Logon date and time	Logon date and time (to the nearest second) is supplied by the system
Full search query	The full search statement entered by the user
Parsed search query.	The full search statement is preparsed. Three pieces of data are recorded: 1) stemmed query terms; 2) term id's; and 3) collection weight of each search term
Normalized call numbers	Normalized call numbers in user-selected clusters are recorded in the log file
Top Ranked Clusters (Up to 20 cluster records can be displayed)	
Query id number	Query id number assigned by the system
Cluster id number	Each cluster record that best represents all the records in a given cluster is assigned a unique number in advance.
Cluster rank	The rank of each retrieved cluster record (1st, 2d, etc.) among all the retrieved cluster records
Cluster weight	Cluster weight determines how closely each cluster record matches the query entered by the user
Action taken by the user	Code indicating whether the user displayed the cluster record and whether he or she selected it as relevant or not
Normalized call number	Classification number that best represents all the bibliographic records in a given cluster
Topical information	Broad topical information taken from the Library of Congress Classification (LCC) scheme (e.g., Bibliography Libraries . . .)
Number of records	Number of bibliographic records represented by a given cluster
Subject headings	The most frequently encountered three Library of Congress subject headings attached to the bibliographic records under a given cluster (along with their frequencies)

(continued)

TRANSACTION LOG RECORD FORMAT *(Continued)*

Field Name	*Description*
	Top Ranked Bibliographic Records (Up to 20 bibliographic records can be displayed)
Query id number	Query id number assigned by the system
Record id number	ID number of record displayed
Record rank	The rank of each retrieved bibliographic record (1st, 2d, etc.) among all the retrieved bibliographic records
Record weight	The weight of bibliographic record that determines how closely the record matches the query entered by the user
Feedback iteration number	Whether the record is retrieved during original search or relevance feedback search (users can perform relevance feedback searches more than once for the same query)
Action taken by the user	Code indicating whether the user displayed the bibliographic record and whether he or she selected it as relevant or not
Bibliographic data	The full MARC data (author, title, imprint, subject headings, etc.) for each retrieved bibliographic record including local call numbers
Completion time	Date and time supplied by the system once the user completes the search

Records Retrieved During Relevance Feedback Searches
(Up to 20 records can be displayed)

Should the user opt for a relevance feedback search after the first retrieval results, CHESHIRE revises the original query based on the user's relevance judgments and retrieves more bibliographic records. The same type of information as given above (under Top Ranked Bibliographic Records) is displayed for each bibliographic record retrieved during relevance feedback searches. The user can continue relevance feedback searches as many times as desired. Starting and ending times are supplied for each relevance feedback cycle.

APPENDIX D

QUESTIONNAIRE

Appendix D is the questionnaire form used after participants performed online catalog searches on CHESHIRE. It contains questions on users' search experience on CHESHIRE and aims to collect data on CHESHIRE's retrieval performance. A questionnaire form was filled out for each search query.

(1-3) Incident ID # Login ID #:

(4) 1. o Effective 2. o Ineffective o Transaction log #

Please do not write above the line

QUESTIONNAIRE

Please answer, by checking appropriate box(es), the following questions about the search you performed on CHESHIRE.

1. You are a: *(5)*
□ MLIS student □ Certificate student □ Doctoral student

2. How long ago did you perform this search? *(6)*
□ 0 - 7 days ago □ 8 to 30 days ago □ 31 to 90 days ago

3. Did you find what you wanted in your first try? *(7)*
□ Yes *[Skip to Q# 5]* □ No □ Not quite what I wanted

4. Why was this? *(8)*
□ The sources did not look helpful
□ I was looking for more specific sources
□ I was looking for more general sources
□ I had to wade through a lot of useless sources
□ I had experienced problems in using CHESHIRE

5. Considering your first try using CHESHIRE for this question, approximately what percent of the sources you found were especially useful? *(9)*
□ 0%
□ Less than 10%
□ Less than 25%
□ Less than 50%
□ More than 50%
□ More than 75%
□ More than 90%
□ 100%

6. Did you perform relevance feedback search? *(10)*
□ Yes □ No *[Skip to Q# 9]* □ I don't remember *[Skip to Q# 9]*

Please turn over

7. Did relevance feedback improve the search results? *(11)*

□ Yes, I found more useful sources

□ Yes, results were better than the first try

□ No, the sources I found were similar

□ No, I found less helpful sources

□ No, not at all

8. Considering your overall experience **(including relevance feedback)** with CHESHIRE*(12)* approximately what percent of the sources you found were especially useful?

□ 0%

□ Less than 10%

□ Less than 25%

□ Less than 50%

□ More than 50%

□ More than 75%

□ More than 90%

□ 100%

9. What was it that you found most helpful in CHESHIRE? [You may mark more than one.] *(13-18)*

□ Ability to enter search queries in natural language

□ Ease of use

□ Relevance feedback

□ Fast retrieval

□ Retrieval results

□ Other (Please explain.)

10. What was it that you found most confusing in CHESHIRE? [You may mark more than one.] *(19-23)*

□ Difficult to use

□ Lack of Boolean searching capability

□ Relevance feedback

□ Time consuming

□ Other (Please explain.)

Thank you.

APPENDIX E

CRITICAL INCIDENT REPORT FORM FOR EFFECTIVE SEARCHES

Appendix E contains the critical incident report form used for effective searches. It was used during the structured interviews that we conducted with participating users. It aims to collect qualitative data about users' search experience with CHESHIRE.

Incident ID #: User login #:

EFFECTIVE INCIDENT REPORT FORM

1. Can you think of a recent instance in which the bibliographic records/sources you obtained through a CHESHIRE search you conducted was especially helpful with your work? Do you have a specific search in mind?

2. What specific information were you seeking? [What was the subject of the search? What was the question in your mind?]

3. How did you carry out this search to get the information you needed? How did you formulate your original search question? What exactly did you type?

4. What information did you obtain as a result of your initial search? What kinds of sources did you find? Were they helpful? In what ways?

5. Did you carry out a relevance feedback search? **Y / N** **[If the answer is "no," skip this question. If "yes," continue.]** How did relevance feedback search help? [Would you say relevance feedback further improved/worsened your research? How?]

6. Considering your search experience with CHESHIRE for this question and the types of sources you found, would you say your search was, in general, successful/effective? **Y / N** Why was this?

7. How would you assess the following sentences?

   a) CHESHIRE retrieved most of the useful sources. **Y / N** Can you be more specific?

   b) More than half of the sources I found using CHESHIRE were useful. **Y / N** Can you be more specific?

APPENDIX F

CRITICAL INCIDENT REPORT FORM FOR INEFFECTIVE SEARCHES

Appendix F contains the critical incident report form used for ineffective searches. It was used during the structured interviews that we conducted with participating users. It aims to collect qualitative data about users' search experience with CHESHIRE. The data was used in the analysis of search failures that occurred in CHESHIRE.

Incident ID #: User login ID #:

INEFFECTIVE INCIDENT REPORT FORM

1. Have you had any recent experience in which you performed a CHESHIRE search that was **unsatisfactory** or **NOT** helpful in finding useful sources that you needed for your work? Do you have a specific search in mind?

2. What specific information were you seeking? [What was the subject of the search? What was the question in your mind?]

3. How did you carry out this search to get the information you needed? How did you formulate your original search question? What exactly did you type?

4. What information did you obtain as a result of your initial search? What kinds of sources did you find? In what ways was the search or its results unsatisfactory?

5. Did you carry out a relevance feedback search? **Y / N** **[If the answer is "no," skip this question. If "yes," continue.]** How did relevance feedback search help? [Would you say relevance feedback further improved/worsened your research? How?]

6. Considering your search experience with CHESHIRE for this question and the types of sources you found, would you say your search was, in general, ineffective/unsatisfactory? **Y /N** Why was this?

7. How would you assess the following sentences?

   a) CHESHIRE failed to retrieve most of the useful sources. **Y / N** Can you be more specific?

   b) More than half of the sources I found using CHESHIRE were useless. **Y / N** Can you be more specific?

APPENDIX G

INVITATION LETTER SENT TO MLIS STUDENTS

Appendix G contains the text of the invitation letter that was handed out to entering MLIS students of the School of Library and Information Studies at the University of California at Berkeley who took LIS 210: Organization of Information in the fall semester of 1991. The letter was also used to get participating users' permission to record their interaction with CHESHIRE in transaction logs.

DATE: September 1991
TO: MLIS students
FROM: Yasar Tonta, Doctoral Student, SLIS

Dear MLIS Student:

I am a doctoral student here at SLIS and currently working on my dissertation under the supervision of Professor Michael Cooper. My research topic has to do with the design of online library catalogs which, more often than not, fail to retrieve relevant materials. More specifically, I am trying to find out the causes of search failures so that we can improve the design of online library catalogs.

I am in the process of gathering data in order to test my hypotheses and would like to have your help in this respect. If agreed, you will be given access to the CHESHIRE system, an experimental online library catalog which provides access to the collection of the SLIS Library. Your interaction with the online catalog will be recorded for further analysis. Please be assured that under no circumstances will the individuals be identified, nor will the data be used for purposes other than research. A follow-up interview and questionnaire will be administered through the end of the fall semester.

Further information about this research will be provided in one of your classes. A demo of CHESHIRE will follow up my introduction. Detailed written instructions for remote access to CHESHIRE will also be handed out in the class.

Please provide the following information; it will be kept strictly confidential.

Name:

Address:

Phone #: **Locker #:**

I also need to know about your prior online catalog and computer experiences. Please kindly answer the following questions.

1. I am familiar with online catalogs such as MELVYL and GLADIS and I use them on a regular basis.

□ Daily □ Weekly
□ Monthly □ About four times a year
□ Once a year □ Never

Please turn over

2. I am familiar with the following application programs (you may mark more than one):

□ Word processing
□ Spreadsheets
□ Database management systems
□ Online searching
□ Electronic mail and bulletin boards
□ programming language(s) (Please list the name(s).)

____

____

□ Other Please explain ______________________

I hereby give my consent that doctoral student Yasar Tonta may record my interaction with the CHESHIRE system, an experimental online catalog. I understand that he shall not use the information I furnish for purposes other than his research. Under no circumstances will my identity be revealed to third parties.

Signed__________________________________ Date: ________________

APPENDIX H

INVITATION LETTER SENT TO DOCTORAL STUDENTS

Appendix H contains the text of the invitation letter that was sent to doctoral students of the School of Library and Information Studies at the University of California at Berkeley. The letter was also used to get participating users' permission to record their interaction with CHESHIRE in transaction logs.

DATE: September 1991
TO: Doctoral Students
FROM: Yasar Tonta, Doctoral Student, SLIS

Dear Doctoral Student:

I am a doctoral student here at SLIS and currently working on my dissertation under the supervision of Professor Michael Cooper. My research topic has to do with the design of online library catalogs which, more often than not, fail to retrieve relevant materials. More specifically, I am trying to find out the causes of search failures so that we can improve the design of online library catalogs.

I am in the process of gathering data in order to test my hypotheses and would like to have your help in this respect. If agreed, you will be given access to the CHESHIRE system, an experimental online library catalog which provides access to the collection of the SLIS Library. Your interaction with the online catalog will be recorded for further analysis. Please be assured that under no circumstances will the individuals be identified, nor will the data be used for purposes other than research. A follow-up interview and questionnaire will be administered through the end of the fall semester.

Further information about this research will be provided should you agree to participate. A demo of CHESHIRE will be given along with the detailed written instructions for remote access to CHESHIRE.

Please provide the following information; it will be kept strictly confidential.

Name:

Address:

Phone #: (Please include your e-mail address)

I also need to know about your prior online catalog and computer experiences. Please kindly answer the following questions.

1. I am familiar with online catalogs such as MELVYL and GLADIS and I use them on a regular basis.

□ Daily □ Weekly
□ Monthly □ About four times a year
□ About once a year □ Never

Please turn over

2. I am familiar with the following application programs (you may mark more than one):

□ Word processing
□ Spreadsheets
□ Database management systems
□ Online searching
□ Electronic mail and bulletin boards
□ programming language(s) (Please list the name(s).)

____
____

□ Other Please explain ______________________

I hereby give my consent that doctoral student Yasar Tonta may record my interaction with the CHESHIRE system, an experimental online catalog. I understand that he shall not use the information I furnish for purposes other than his research. Under no circumstances will my identity be revealed to third parties.

Signed________________________ Date: ____________________

APPENDIX I:

QUERIES SUBMITTED TO CHESHIRE (N=228)

There were 228 queries analyzed in this study. They are listed below and include all spelling errors made by the users. See Appendix J for the outcome of search queries submitted to CHESHIRE.

1. computer applications in library operations
2. market analysis of library services
3. cost-effectiveness of library services
4. performance measures of library services
5. performance measures for library services
6. all material regarding the development of law libraries in the united states between 1840-1970, paying particular attention to the years 1925 - 1965. also everything regarding the history of federal depositories the architechture of law libraries between the years 1925-65. include personal narratives of law librarians, lawyers, and legal support staff. provide similar information for the development of law libraries in england and canada.
7. history of law libraries, history of federal depositories, personal narratives of law librarians, law libraries,
8. information poverty in the united states
9. cd-rom
10. cd-rom databases
11. electronic mail
12. local area networks
13. z39.50
14. user interface studies
15. Indexes for information resources on or in networks like Internet and Bitnet

16. indexes for information resources on or in networks like internet or bitnet

17. medical libraries in medical schools

18. medical school libraries

19. medical school libraries

20. hypermedia

21. hypertext

22. online authority control

23. computer generated type fonts

24. computer typesetting

25. information society privacy

26. manuscript cataloging

27. subject freedom of information and national security

28. freedom of information

29. I want books about letterpress printing published after 1950.

30. please find books on children, basball, and animals

31. what about obscenity and literature

32. computer mediated communication

33. stasz communications

34. computer conferencing

35. vocational education

36. fin me systems in libraries

37. 1. Call Number Z 00699

38. some books on history of libraries and classification

39. organizational effectiveness in libraries

40. I want books about slavic incunabula

41. the children's book market in the united states

42. the children's book market in the united states

43. how do I use this systenm

44. help i am sooooo confused

45. the

46. all materials on the history of law libraries, all materials on the history of law librarianship

47. libraries in mexico

48. libraries in berlin, germany

49. career in library automation

50. information poverty united states

51. information ownership -- united states

52. information resource policy of the united states information poverty knowledge gap

53. public library systems in Finalnd and Sweden public library systems in Finland and Sweden

54. books censorship united states

55. john dewey biography

56. cataloging methods France

57. dr. seuss

58. suess englich language

59. dr suess english

60. dr seuss english

61. serials acquisition

62. cultural boycott of south africa

63. ethics and policy

64. ljkdsf q

65. virtual reality cyberspace

66. user interface

67. i want holdings on the history of law librarianship, collection development in law libraries from 1935-1970, trends in acquisition practices for law libraries during the same period and i want see the material on these headings that are in the periodicals collection

68. law libraries - collection development from 1935

69. periodical literature on the development of law library collections

70. collection development law libraries only

71. nanotechnology

72. interface

73. interface

74. C

75. programming C

76. library history methodology

77. library history methodology comparison international

78. information economics and policy journal

79. Bquit

80. xerox windows

81. history of printing in Paris

82. management budget

83. banned books after 1980

84. romance novels

85. squid Acookery american

86. rlg oclc utlas

87. hygiene

88. natural language computers

89. serials automation acquisition

90. marcia tuttle

91. tuttle

92. katz

93. katz reference

94. quit

95. find all library literature concerning the history and publication of the Federal Register

96. graphic display of thesauri in electronic format

97. how many books by patrick wilson does the library have?

98. I'd like to see books about library automation Cq

99. I'd like to see recent books, in english, about library automation

100. alred hitchock films

101. film librarianship

102. university presses and book reviews

103. university presses

104. scholarly publishing

105. looking for a humorous book about librarianship with cartoons

106. anything about su tung po and ancient chinese poetry

107. anything about online fulltext or the retrieval of information over networks like the Internet or Bitnet or Janet or EARN or Minitel

108. minitel

109. access points in catalogs

110. katz on reference

111. teaching library users

112. british serials collections

113. library tours

114. variations in catalog use among library user populations

115. Morrill Act

116. armed forces overseas libraries

117. r&d communication

118. r&d

119. knowledge utilization

120. history of manichaeans and their persian origins

121. manichaeism and persia 200-400 ad

122. collection development in american literature

123. access in online catalogs

124. local area network

125. novell

126. local area network

127. censorship of children's books

128. censorship of children's literature

129. recommended books ofor gifted children

130. books for mentally handicapped children

131. books for the mentally handicapped children

132. sex education books for adolescents

133. television's effect on literacy

134. bibliographies of occult books

135. manichaeism and augustine

136. assam, india

137. drugs, sex and rock and roll

138. syrian asceticism

139. asceticism in syria

140. berkeley library school history

141. seismic berkeley

142. application of artificial intelligence in information systems

143. expert systems and databases

144. prolog and databases

145. expert systems

146. electronic serials

147. want to find a small set of books on historical treatment of mathematics

148. optical disk technology

149. subject cataloging

150. alumni

151. a general history of the library of congress

152. library of congress

153. book-binding -- bibles

154. bible

155. blood transfusion

156. rare book collection

157. history of printing

158. annotated bibliography on anthropology

159. anthropology

160. anthropology

161. reference sources on film or movies

162. I want information on the public image of librarians through history

163. need information on information scientists and information profesiions professionals and current trends in the field information professionals or information scientists--not dissertation Abut want information on the careers

164. librarians and stereotypes and attitudes toward librarians

165. librarians and professional recognition and paraprofessionals find information about librarians and professional recognition and image of paraprofessional librarians

166. information policy

167. United States information policyand freedom of information United States infor

168. United States information policy and freedom of information

169. federal information policy and freedom of information

170. art library resources

171. how to research women artists

172. library reference material on art

173. library resource materials on art

174. children's book reviewing

175. impact of book reviewing on the publishing industry

176. storytelling

177. asian american illustrators

178. cip

179. medieval manuscript illumination and development of the book

180. greek typefaces in paris in fifteenth and sixteenth century

181. subject searching childrens literature and censorship

182. subject search japanese novelists

183. subject search japanese book reviews

184. title search japanese librar

185. integrated academic information managment system

186. project mercury

187. express

188. israel

189. rlg conspectus issues

190. projected salaries for special and academic librarians on the west coast

191. cheshire

192. relevance

193. remodel

194. guide to literature

195. art

196. interior design

197. bibliography art design interior decorating wood woodworking

198. online systems in middle eastern libraries

199. libraries in egypt

200. The Printed Press

201. the private presses

202. \library resources and technical services

203. booklist

204. the story of language

205. the wood engraving of gwen raverat

206. illumination and calligraphy in manuscripts shown at exhibitions

207. the history of writing

208. electronic books

209. e-journal

210. folklore for children

211. human computer interaction

212. find bibliographies on archaeology

213. bibliograph# and qumran

214. find bibliograph# and Middle East

215. find a bibliography of archaeological books

216. archaeology

217. library services for ethnic minorities

218. bibliographies or indexes covering merchant marine, salors, seamanship, marine cargo transportation, maritime law

219. history of library of congress subject headings

220. bibliography of transportation, mercantile aspects, sailors, seamenship, marchant marine

221. history of libraries in italy

222. classification of materials on gay and lesbian studies

223. alternatives to traditional subject headings

224. art

225. alphabet

226. vctorian

227. victorian

228. relieurs

APPENDIX J

RETRIEVAL PERFORMANCE IN CHESHIRE

Appendix J contains precision and recall ratios for all search queries submitted to CHESHIRE throughout the experiment. The causes of search failures, if applicable, are also given. The figures in Column 1 refer to search query number (*Q. no.*); the full text of the query can be found in Appendix I.

Precision and recall ratios obtained before the relevance feedback searches are given in Columns 2 and 3, respectively. No precision and recall ratios are available for discontinued searches. Columns 4 through 9 give precision and recall ratios obtained after relevance feedback searches. No figures are available if a relevance feedback search was not performed for a given search query. Average precision and recall ratios are given in Columns 10 and 11, respectively.

Search effectiveness for each query is given in Column 12. No data is provided in this column for out-of-domain search queries or queries that was discontinued for some reason.

The cause of search failure, if applicable, is briefly explained in Column 13.

Q.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	
			First iteration		*Second iteration*		*Third iteration*					
no.	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		*Causes of search failure*
001	.000	.850	.166	.812					.083	.831	E	
002	.150	.750	.055	1.00					.105	.875	I	Collection failure
003												No clusters selected; user did not like the clusters
004												No clusters selected; user wanted to revise the query
005	.210	.850	.105	.818					.158	.834	E	
006	.5	.231	.00	.00					.167	.077	E	
007	.00	.231							.000	.231	I	Specific query
008	.263	.312	.200	.273	.00	.250			.154	.278	E	
009												No clusters selected; user wanted to revise the query
010	.214	.750	.667	1.00					.440	.875	E	
011												Faulty cluster selection
012	.00	.833	.00	.00					.00	.417	E	
013												Collection failure; zero retrieval
014	.00	.400	.461	.500					.230	.450	E	
015	.167	.00							.167	.00	I	Collection failure
016	.105	.00	.133	.00	.00	.00			.079	.00	I	Collection failure
017												No clusters selected; telecommunication problems
018												Cluster failure; no clusters selected as being relevant
019	.053	.033							.052	.033	I	Collection failure
020												Collection failure; zero retrieval
021												Collection failure; zero retrieval
022	.222	.333							.222	.333	E	
023	.316	.333	.200	.333					.258	.333	E	*(Continued)*

Q.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	
			First iteration		*Second iteration*		*Third iteration*					
no.	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		*Causes of search failure*
024	.316	.727	.312	.555					.314	.641	E	
025	.00	.100	.00	.00					.00	.500	E	
026	.368	.467	.00	.375					.184	.421	E	
027												No clusters selected; user interface problems
028	.067	.850	.00	1.00					.033	.925	I	Search statement
029	.400	.133	.00	.385					.200	.259	I	User interface problems
030	.00	.00							.00	.00	I	Out of domain search query
031	.316	.350	.150	.461	.30	1.00			.255	.604	E	
032	.667	.417	.100	.286	.00	.00			.255	.234	I	Vocabulary problem
033												No clusters selected; out of domain search query
034	.00	.00							.00	.00	I	Collection failure
035												No clusters selected; out of domain search query
036	.667	.600							.666	.600	E	
037												Zero retrieval; call number search
038	.00	.00							.00	.00	I	Broad search query
039	.00	.400	.00	.417					.00	.408	E	
040												No clusters selected as being relevant
041												No clusters selected; user interface problems
042	.053	.250	.055	.333	.50	1.00			.202	.528	E	
043												Zero retrieval; help request
044											I	Help request
045												Zero retrieval; "the" is a stop word
046	.00	.00	.00	.500					.00	.250	E	*(Continued)*

Q. no.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	*Causes of search failure*
			First iteration		*Second iteration*		*Third iteration*					
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		
047	.053	.555							.053	.555	E	
048	.273	.643	.00	1.00					.136	.821	E	
049	.75	.9	.00	1.00					.375	.95	E	
050												No clusters selected; user wanted to revise the query
051	.053	.25	.091	.667					.072	.458	I	Collection failure
052	.267	.25	.00	.00					.133	.125	E	
053												No clusters selected; user interface problems
054	.5	.85	.00	.778					.25	.814	E	
055												No clusters selected; out of domain search query
056	.5	.5	.00	.00	.00	.00			.167	.167	I	Collection failure
057	.067	1.00							.067	1.00	E	
058												No clusters selected; user wanted to revise the query
059												No clusters selected; user wanted to revise the query
060	.263	1.00	.053	.00					.158	.5	E	
061	.474	.45	.00	.15	.44	.7			.306	.433	E	
062	.00	.00	.00	.00	.00	.00			.00	.00	I	Collection failure; search statement
063	.1	1.00							.1	1.00	E	
064												Zero retrieval; user entered gibberish characters
065												Collection failure; no clusters selected
066	.00	.25							.00	.25	I	Collection failure; most recent items needed
067												Scope failure; periodical literature search
068												Cluster failure; user did not like the clusters
069											I	Scope failure; periodical literature search *(Continued)*

Q. no.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches* — *First iteration*		*Second iteration*		*Third iteration*		*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	*Causes of search failure*
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		
070											I	User did not like clusters; no clusters selected as relevant
071												Out of domain search query; zero retrieval
072											I	User interface problems; no clusters selected as being relevant
073	.00	.00							.00	.00	I	User interface problems; collection failure
074											I	Stemming algorithm; did not recognize "C"; zero retrieval
075											I	Stemming algorithm failure; no clusters selected as relevant
076												No clusters selected; user wanted to revise her query
077	.00	.850	.00	.650	.40	1.00			.133	.833	E	
078	.091	1.00	.00	.00					.045	.500	I	Known-item search
079												User typed "quit" in the query description screen
080	.667	1.00							.667	1.00	I	Collection failure
081	.500	.200							.500	.200	E	
082	.500	.650							.500	.650	E	
083	.579	.850	.250	.450	.11	.100	.00	.00	.235	.350	E	
084												Out of domain search query
085												No clusters selected; out of domain search query
086	.555	.800							.555	.800	E	
087												Out of domain search query
088	.00	.555							.00	.555	E	
089	.800	.650	.833	.454					.817	.552	E	
090												Author search; not supported in CHESHIRE; false drops
091												Author search; not supported in CHESHIRE; zero retrieval
092												Author search; not supported in CHESHIRE; zero retrieval

Q. no.	Precision (P) & recall (R) ratios before relevance feedback searches		Precision & recall ratios after relevance feedback searches						Average precision & recall ratios		Search performance: effective (E)/ ineffective (I)	Causes of search failure
			First iteration		Second iteration		Third iteration					
	P	R	P	R	P	R	P	R	P	R		
093												Author/title search; not supported in CHESHIRE
094												Quit entered in the query description screen
095	.00	.00							.00	.00	I	Collection failure
096	.368	1.00	.100	.00					.234	.500	E	
097												Author search; not supported in CHESHIRE
098												No clusters selected as being relevant
099	.500	.650	.500	.800					.500	.725	E	
100												Out of domain search query
101	.667	.300	.00	.350					.333	.325	E	
102	.500	.00							.500	.00	I	Collection failure
103	.210	1.00							.210	1.00	E	
104	.105	1.00	.00	1.00					.053	1.00	E	
105	.263	.500	.00	.00					.131	.250	E	
106												Out of domain search query
107											I	Collection failure; no clusters selected as being
108											I	Collection failure; zero retrieval
109	.222	.300	.300	.300	.00	.150			.173	.250	E	
110											E	
111											I	Telecommunication problem; no cluster selected
112	.053	.00	.050	.00	.00	.00			.051	.00	I	Collection failure
113											I	False drops; no clusters selected as being relevant
114*	.158	.210	.450	.600	.00	.00	.00	.00	.122	.162	E	
115	.500	.00	.00	.00	.05	.00			.183	.00	E	*(Continued)*

Q.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	
			First iteration		*Second iteration*		*Third iteration*					
no.	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		*Cause of search failure*
116	.158	.200	.150	.875	.05	1.00			.119	.692	E	
117	.263	.263	.105	.143					.184	.203	E	
118											I	Stemming algorithm; "r&d" not recognized; zero retrieval
119	.684	.850	.250	.450					.467	.650	E	
120												Out of domain search query
121												Out of domain search query; no clusters selected as relevant
122	.333	.333	.500	1.00					.117	.667	I	Out of domain search query
123	.263	.750	.100	.500	.10	.50			.154	.583	E	
124	.00	.800	.00	1.00					.00	.900	E	
125											I	Collection failure; no clusters selected as being relevant
126	.00	1.00							.00	1.00	E	
127	.00	.400							.00	.400	I	Library of Congress Subject Headings
128	.272	.600	.00	.00					.136	.300	I	Library of Congress Subject Headings
129	.00	1.00							.00	1.00	E	
130												Out of domain search query; user wanted to revise the query
131	.00	.500							.00	.500	I	Collection failure
132	.00	1.00							.00	1.00	E	
133												Out of domain search query
134												Out of domain search query
135												Out of domain search query; no clusters selected as relevant
136												Out of domain search query; no clusters selected as relevant
137												Out of domain search query; no clusters selected as relevant
138												Out of domain search query; zero retrieval *(Continued)*

Q. no.	Precision (P) & recall (R) ratios before relevance feedback searches		Precision & recall ratios after relevance feedback searches						Average precision & recall ratios		Search performance: effective (E)/ ineffective (I)	Cause of search failure
			First iteration		Second iteration		Third iteration					
	P	R	P	R	P	R	P	R	P	R		
139												Out of domain search query; zero retrieval
140	.105	1.00	.050	.00	.00	.00	.00	.00	.039	.250	I	Collection failure
141											I	Collection failure
142	.368	.615	.077	.200					.223	.408	E	
143												No clusters selected; user wanted to revise her query
144												No clusters selected; user wanted to revise her query
145	.00	.615	.00	.00					.00	.308	E	
146	.500	.00							.500	.00	I	Collection failure
147	.316	.545	.250	1.00	.00	.00			.205	.515	E	
148	.500	.769	.00	.00					.250	.385	E	
149	.00	.500							.00	.500	E	
150	.158	1.00							.158	1.00	I	Collection failure
151											I	Cluster failure; no clusters selected as being relevant
152	.053	.769	.125	.333	.00	.00			.059	.367	E	
153											I	Collection failure
154											I	Collection failure; no clusters selected as being relevant
155												Out of domain search query; zero retrieval
156	.00	.900	.00	.400					.00	.650	E	
157											I	No apparent reason; relevant clusters, but not selected
158	.00	1.00							.00	1.00		Out of domain search query
159												Out of domain search query
160	.00	1.00	.00	.00					.00	.500		Out of domain search query
161												Out of domain search query *(Continued)*

Q. no.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	*Cause of search failure*
			First iteration		*Second iteration*		*Third iteration*					
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		
162	.00	.333	.00	.00					.00	.167	I	Specific query
163	.053	.100							.053	.100	I	Collection failure
164	.00	.667							.00	.666	E	
165	.00	.00									I	Collection failure
166											I	Cluster failure
167												No clusters selected; user wanted to revise his query
168												No clusters selected; user wanted to revise his query
169	.526	.950	.00	.437					.263	.694	E	
170	.526	.850	.300	.500					.413	.675	E	
171	.143	.500	.050	1.00					.096	.750	E	
172												No clusters selected as being relevant
173												No clusters selected as being relevant
174*	.167	.500	.100	.500	.17	.500	.00	1.00	.097	.500	I	Library of Congress Subject Headings
175	.263	.00							.263	.00	I	Collection failure
176	.538	1.00	.267	1.00	.10	1.00	.00	.00	.226	.750	I	No apparent reason given
177												Out of domain search query; no clusters selected as relevant
178	.500	.429							.500	.429	I	Search statement; abbreviation used
179	.053	1.00							.053	1.00	E	
180	.421	.00	.00	.00					.210	.00	I	Collection failure
181	.368	.214	.250	.364	.25	.714			.289	.431	I	Library of Congress Subject Headings
182												Out of domain search query
183												Out of domain search query
184	.158	.250	.250	.210					.204	.230	I	Search statement; truncated word not recognized *(Cont'd)*

Q. no.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	*Cause of search failure*
			First iteration		*Second iteration*		*Third iteration*					
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		
185	.200	1.00	.00	.00					.100	.500	E	
186	.00	.00	.00	.00					.00	.00	I	Collection failure
187												No clusters selected as being relevant
188	.00	1.00	.00	.00					.00	.500	E	
189	.00	.00							.00	.00	I	User interface problems
190	.00	.00	.500	.00					.250	.00	I	User interface problems
191												Collection failure; no clusters selected as being relevant
192	.214	.571	.062	.333	.00	1.00			.092	.635	I	Library of Congress Subject Headings
193												Collection failure
194	.053	.200	.00	.00	.00	.00			.017	.067	I	Search statement
195	.250	.300	.100	.300					.170	.300	I	Out of domain search query
196												Out of domain query; no clusters selected as being relevant
197												Out of domain search query
198	.00	.00							.00	.00	I	Collection failure
199	.263	.818	.050	1.00					.156	.910	E	
200												Title search; not supported in CHESHIRE
201	.100	1.00							.100	1.00		Title search; not supported in CHESHIRE
202												Title search; not supported in CHESHIRE
203												Title search; not supported in CHESHIRE
204												Title search; not supported in CHESHIRE
205												Title search; not supported in CHESHIRE
206												Title search; not supported in CHESHIRE
207												Title search; not supported in CHESHIRE

Continued

Q. no.	*Precision (P) & recall (R) ratios before relevance feedback searches*		*Precision & recall ratios after relevance feedback searches*						*Average precision & recall ratios*		*Search performance: effective (E)/ ineffective (I)*	*Causes of search failure*
			First iteration		*Second iteration*		*Third iteration*					
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*		
208	.00	.00							.00	.00	I	Collection failure
209												Stemming algorithm failure; abbreviation ("e") not recognized
210	.789	.750	.700	.867					.744	.808	E	
211	.200	.650	.00	.100					.100	.370	E	
212												Out of domain search query
213												No clusters selected as being relevant
214												No clusters selected as being relevant
215	.00	.00							.00	.00		Out of domain search query
216												Out of domain search query; no clusters selected as relevant
217	.947	.900	.800	.800					.874	.850	E	
218												Out of domain search query; no clusters selected as relevant
219											I	No reason given; relevant clusters retrieved but not selected
220												Out of domain search query; no clusters selected as relevant
221	.316	.350	.222	.210					.269	.280	E	
222											I	Collection failure; no clusters selected as being relevant
223	.053	.00							.053	.00	I	Search statement
224	.053	.400	.100	.400	.06	.200	.00	.250	.054	.312	I	Faulty cluster selection
225	.526	.950	.454	.700	.00	.850			.327	.833	I	Search statement
226												Misspelling; zero retrieval
227	.105	.625	.00	.333					.053	.479	E	
228	.846	.55							.846	.550	E	