Abstract

This paper examines search failures in document retrieval
systems.  Since search failures are closely related to overall
document retrieval system performance, the paper briefly
discusses retrieval effectiveness measures such as precision and
recall.  It examines four methods used to study retrieval
failures: retrieval effectiveness measures, user satisfaction
measures, transaction log analysis, and the critical incident
technique.  It summarizes the findings of major failure analysis
studies and identifies the types of failures that usually occur
in document retrieval systems.


1.0  Introduction

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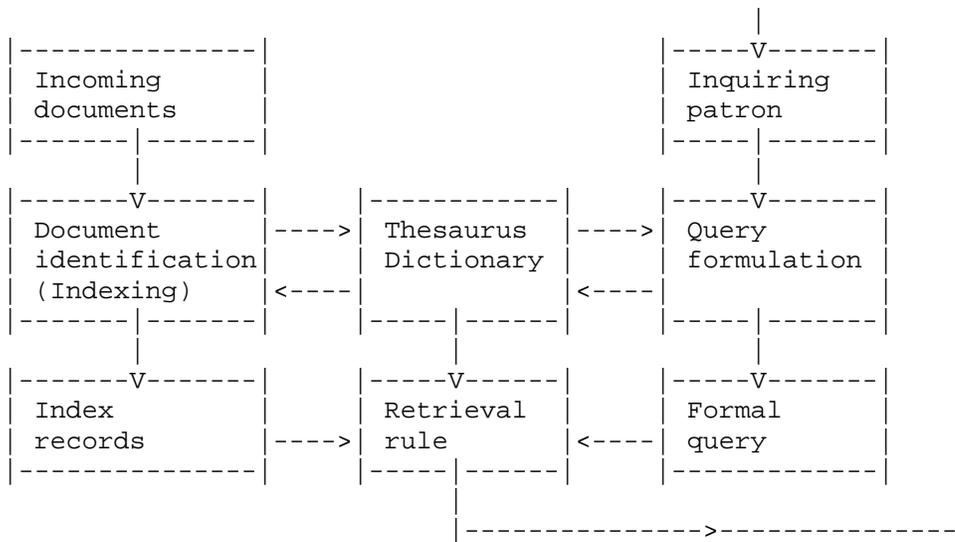interfaces, retrieval rules, and indexing languages.

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.  This paper will
look at various (mostly implied) definitions of "search failure"
and discuss some of the methods used in failure analysis studies.

2.0  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.  Maron [1] provides a
more detailed description of the document retrieval problem and
depicts the logical organization of a document retrieval system
(see Figure 1).

-----------------------------------------------------------------
     Figure 1. Logical Organization of a Conventional Document
             Retrieval System.  Source: Maron [2].

-----------------------------------------------------------------
                                   |------<-----|

```
                                              |           |
   |---------------|                     |-----V-------|  |
   | Incoming      |                     | Inquiring   |  |
   | documents     |                     | patron      |  |
   |-------|-------|                     |-----|-------|  |
           |                                   |          |
   |-------V-------|   |-----------|     |-----V-------|  |
   | Document      |-->| Thesaurus |---->| Query       |  |
   | identification|   | Dictionary|     | formulation |  |
   | (Indexing)    |<--|           |<----|             |  |
   |-------|-------|   |-----|-----|     |-----|-------|  |
           |                |                  |          |
   |-------V-------|   |-----V-----|     |-----V-------|  |
   | Index         |   | Retrieval |     | Formal      |  |
   | records       |-->| rule      |<----| query       |  |
   |---------------|   |-----|-----|     |-------------|  |
                            |                             |
                       |--------------->--------------|

-------------------------------------------------------------------


+ Page 6 +
```

As Figure 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.

Maron 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 [3].

Thus, a document retrieval system consists of (1) a store of
documents (or, representations thereof); (2) a population of
users each of whom makes use of the system to satisfy their
information needs; and (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.  There also should be a user interface to

allow users to interact with the system.

In reality, the ideal document retrieval system discussed in this section 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.

3.0  Search Failure Analysis

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 definitions of "search failure" used by researchers.  After all, we cannot analyze search failures if we do not recognize them.


3.1  Measures of Retrieval Effectiveness

Retrieval effectiveness measures such as "precision" and "recall" are widely used to evaluate the effectiveness of online document retrieval systems.  A few measures, which are discussed below, are also used in the study of search failures.  This paper will not review all the measures of retrieval effectiveness suggested in the literature since they are seldom, if ever, used in the analysis of search failures.

Precision is defined as the proportion of retrieved documents which are relevant, whereas recall is defined as the proportion of relevant documents retrieved [4].  These two measures are generally used in tandem in evaluating retrieval effectiveness in document retrieval systems.

Precision can be taken as the ratio of the number of documents that are judged relevant for a particular query over the total number of documents retrieved.  For instance, if, for a particular search query, the system retrieves two documents and the user finds one of them relevant, then the precision ratio for this search would be 50%.

Recall is considerably more difficult to calculate than precision because it requires finding relevant documents that will not be retrieved during users' initial searches [5].  Recall can be taken as the ratio of the number of relevant documents retrieved over the total number of relevant documents in the collection.  Take the above example.  The user judged one of the two retrieved documents to be relevant.  Suppose that later three more relevant documents that the original search query failed to retrieve were found in the collection.  The system retrieved only one out of the four relevant documents from the database.  The recall ratio would then be equal to 25% for this particular search.

"Fallout" is another measure of retrieval effectiveness.  Fallout

can be defined as the ratio of nonrelevant documents retrieved
over all the nonrelevant documents in the collection.  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
nonrelevant documents).  Since the user retrieved one nonrelevant
document out of a total of five nonrelevant ones in the
collection, the fallout ratio would be 20% for this search.


3.2  Methods of Analyzing Search Failures

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


3.2.1  Analysis of Search Failures Utilizing Retrieval
       Effectiveness Measures

If precision and recall are seen as performance measures with the
given definitions, it instantly becomes clear that "performance"
can no longer be defined as a dichotomous concept.  As precision
and recall are defined as percentages, we can think of "degrees"
of search failure or success.  This view would probably best
reflect 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 they are better or worse than one another.

+ Page 9 +


Performance measures such as precision and recall can be used in
the analysis of search failures.

In the precision example in Section 3.1, 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.

It should be noted 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) [6]. (For a comprehensive review of the concept of "relevance," see [7].)

+ Page 10 +

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, the author is not aware of any experiment where fallout ratio has been successfully calculated [8].  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.

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. It should be noted that 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.

Notice that 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 [9] 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. 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 [10]. It would be interesting to study what percentage of searches in online catalogs get abandoned in view of user frustration from fallout failures.

It is also theoretically possible to envision "perverse" document retrieval systems where, for a given query, the system first would retrieve all nonrelevant documents before it would eventually retrieve relevant ones [11]. However, in real life, "perverse" document retrieval systems are unlikely to exist.

Mainly, 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, this paper has examined a few of the measures of retrieval effectiveness and the ways in which they are used in the study of search failures. It was 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. Some would argue that 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.

Users prefer high precision to high recall. They wish to retrieve "some good references without having to examine too many bad ones" [12]. 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.

It also should be noted that not everyone is satisfied with the most commonly used retrieval effectiveness measures (precision and recall). For instance, 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 [13].  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" [14].


### 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.

+ Page 13 +

First, defining user satisfaction is difficult.  Several authors
tried to address this issue.  Tessier, Crouch, and Atherton
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 [15].  Bates 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 [16].  Tessier 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" [17].  Hilchey and
Hurych 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 [18].

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
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"
[19, (original emphasis)].  Tagliacozzo warns us: "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" [20].

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 found that the use of a
two-point (yes-no) scale "appeared to result in inflated success
ratings" [21].  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" [22, (original emphasis)].

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 consider a search as a "failure" even if the system has missed some relevant documents (i.e., recall failure).

User satisfaction measures are influenced by both user group 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 entirely 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, a health professional investigating a medical procedure on "MEDLINE only found records showing it to be safe, missing the reports of fatalities associated with the procedure" [23].

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" [24].  Furthermore, as Cheney notes, we do not "know what end users expect of their search results, because no study has 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" [25].

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 studied, among other things, the relationship between user satisfaction and precision and recall [26].

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, 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 [27].  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 (on a five-
point scale).

+ Page 16 +

They found that "searchers in questions where users indicated
high overall satisfaction with results . . . were 2.49 times more
likely to have higher precision" [28].  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 [29, (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.
Larson 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" [30].  This includes
search queries entered, records displayed, help requests, errors,
and the system responses.  (For a review of online catalog
transaction log studies, see [31].)

+ Page 17 +

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 [32].  A few others employed the same method when they
studied search failures in MEDLINE [33].  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"
[34].  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 [35].  Transaction logs of
such online 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.

+ Page 18 +

What constitutes a search failure when the relevance judgment for
each retrieved document is recorded in the transaction log?  Some
researchers came up with yet another practical definition of
search failure and analyzed it accordingly.  For example, during
the evaluation of Okapi online catalog, a search was counted as a
failure "if no relevant record appears in the first ten which are

displayed" [36].  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"
[37].

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.

+ Page 19 +

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 remain anonymous in transaction logs,
analysis of these logs "prevents correlation of results with user
characteristics" [38].


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,
Robertson, and Neilson [39] 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 "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 [40] describes the critical incident technique 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 critical incident technique essentially consists of two
steps: (1) collecting and classifying detailed incident reports,
and (2) making inferences that are based on the observed
incidents.

Recently, the critical incident technique has been used to assess
"the effectiveness of the retrieval and use of biomedical
information by health professionals" [41].  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 usually 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 advance permission has to be sought 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.


3.3  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.

+ Page 22 +

4.0  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 [42].  However, few researchers have studied search failures directly [43].  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.

## 4.1 Studies Utilizing Precision and Recall Measures

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

### 4.1.1 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 [44]. They also studied the causes of search failures in document retrieval systems. This paper only reviews findings that pertain to search failures.

In the first study (Cranfield I), Cleverdon 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 [45].

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" [46]. 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]" [47].

+ Page 23 +

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" [48]. 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 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" [49]. 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% [50]; source documents were successfully retrieved for most search queries.

Cleverdon's analysis of search failures was based on 329 documents and queries. The total number of search failures was 495 [51]. 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%).

+ Page 24 +

It appears that well over half of the failures in this study were
caused by the indexing process.  Cleverdon summarizes the results
of the analysis of search failures as follows [52]:

> 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 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 [53].

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" [54].  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.

## 4.1.2  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, studied the retrieval effectiveness of a large biomedical reference retrieval system (MEDLARS) in operation [55].  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 section 3.1.  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 [56, (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%" [57].

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 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 [58].

+ Page 27 +

Lancaster found that recall failures have 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" [59].  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" [60].


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

More recently, Blair and Maron [61] 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" [62].  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" [63].

+ Page 28 +

The results of Blair and Maron's tests showed that the mean
precision ratio was 79% and the mean recall ratio was 20% [64].

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.

Blair and Maron 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" [65].


4.1.4  Markey and Demeyer's Dewey Decimal Classification Online
       Project

Markey and Demeyer studied the Dewey Decimal Classification (DDC)
system "as an online searcher's tool for subject access,
browsing, and display in an online catalog" [66].  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" [67].

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 Section 3.1 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 [68].

+ Page 29 +

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" [69]. 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 [70]. 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" [71]. The
average precision scores ranged from a low of 26% to a high of
65% [72]. 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 [73].

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" [74].

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%) [75]. 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) [76].

Markey and Demeyer apparently did not count "zero retrievals" as search failures.  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.


4.2  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 analyzed 153 online search evaluation forms returned by the users in a university library [77].  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 short, 85% of the respondents felt that search results were at least half relevant.  It should be noted that 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 reviewed the studies reporting user satisfaction in end-user search services such as MEDLINE and BRS/After Dark [78]. Most end-users seemed to be satisfied with the online search services.

Ankeny also reported the results of two studies that 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" [79].

In the second study, Ankeny surveyed some 600 end-users.  He used a stricter measure of search success that had a reliability coefficient of .90.  Search success was not measured on a five-point scale in the second study.  Rather, in order for a search to be qualified as successful, the user had to answer three questions that affirmed that the user was fully satisfied with the search, found exactly what was desired, and was not dissatisfied in any way.  He states: "Of 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" [80].  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 analyzed search failures encountered by MEDLINE
end-users employing the Colleague search software [81].  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" [82, (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" [83].

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).

+ Page 32 +

4.3  Studies Utilizing Transaction Logs

Several researchers have used transaction logs to study search
failures in online catalogs.  Dickson [84] 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 [85] examined transaction logs of the Okapi online catalog
and identified several unsatisfactory areas in the operation of
Okapi 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" [86].

In a follow-up study, it was found that 17 out of 122 sessions (or 13.9%) failed in the Okapi (including 2 sessions that failed due the collection not containing relevant items). (Most sessions contained more than one search.) In 7 sessions, the users' vocabulary did not match that of the catalog (e.g., "sociology of shopping"). Another 4 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") [87].

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)" [88]. The authors of the report claim that the improvement is primarily due to: (1) Okapi's "best match" search, and (2) stemming and automatic cross-referencing [89].

+ Page 33 +

Peters [90] 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, it should be noted that 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" [91].

Hunter [92] 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 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. Walker and her colleagues [93] 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 [94]. 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 [95] 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) [96].

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 benefitted from transaction
monitoring to the extent that "zero-hit" searches were identified
from transaction logs [97].  Researchers examined the zero-hit
searches in order to find out why a particular search query
failed to retrieve anything in the database.  Unlike Lancaster
[98], they did not attempt to identify the causes of recall and
precision failures.


4.4  Studies Utilizing the Critical Incident Technique

It was mentioned earlier (Section 3.2.4) that Wilson, Starr-
Schneidkraut, and Cooper studied searching in MEDLINE using the
critical incident technique [99].  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?'" [100].  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" [101].  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 [102].

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.

+ Page 36 +

4.5  Other Search Failure Studies

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

Gouke and Pease [103] 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).

Alzofon and Van Pulis [104] 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.

Janosky, Smith, and Hildreth [105] 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% [106].

+ Page 37 +

It appears that users experienced serious problems with the 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. Failures 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" [107].


4.6  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 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" [108].  This may be 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 [109].  Larson [110] found that almost half of
all subject searches in MELVYL retrieved nothing.  More recently,
Larson [111] 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 and the percentage of
subject searching.  This 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" [112].

+ Page 38 +

Larson [113] 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 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" [114].  A study
conducted by Van Pulis and Ludy [115] showed that 53% of the
users' terms matched subject headings in the online catalog.
Vizine-Goetz and Markey Drabenstott 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"
[116].  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" [117].


5.0  Conclusion

It appears that 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.

+ Page 39 +

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 nothing 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 [118].  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.

+ Page 40 +

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.

+ Page 41 +

Notes

1.  M. E. Maron, "Probabilistic Retrieval Models," in Progress in
Communication Sciences, vol. 5, ed. Brenda Dervin and Melvin J.
Voigt. (Norwood, NJ: Ablex, 1984), 145-176.

2.  Ibid., 155.

3.  Ibid.

4.  C. J. Van Rijsbergen, Information Retrieval, 2nd ed. (London:
Butterworths, 1979), 10.

5.  David C. Blair and M. E. Maron, "An Evaluation of Retrieval
Effectiveness for a Full-Text Document-Retrieval System,"
Communications of the ACM 28 (March 1985): 291.

6.  S. E. Robertson, M. E. Maron, and W .S. Cooper, "Probability
of Relevance: A Unification of Two Competing Models for Document
Retrieval," Information Technology: Research and Development 1
(1982): 1.

7.  Tefko Saracevic, "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 (1975): 321-343.
See also: Michael Eisenberg and Linda Schamber, "Relevance: The
Search for a Definition," in ASIS '88: Proceedings of the 51st
ASIS Annual Meeting, Atlanta, Georgia, October 23-27, 1988, ed.
Christine L. Borgman and Edward Y .H. Pai. (Medford, NJ: Learned
Information, 1988), 164-168.

8.  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: Cyril Cleverdon, Jack Mills, and Michael Keen,
Factors Determining the Performance of Indexing Systems, Volume
1, Design (Cranfield, England: Aslib, 1966); and Cyril Cleverdon
and Michael Keen, Factors Determining the Performance of Indexing
Systems, Volume 2, Test Results (Cranfield, England: Aslib,
1966).

9.  Ray R. Larson, "Between Scylla and Charybdis: Subject
Searching in the Online Catalog," in Advances in Librarianship,
vol. 15, ed. Irene P. Godden (San Diego, CA: Academic Press,
1991), 188.  See also: S. E. Wiberley and R. A. Dougherty,
"Users' Persistence in Scanning Lists of References," College &
Research Libraries 49 (1988): 149-156.

10.  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."  Source: David C. Blair, "Searching
Biases in Large Interactive Document Retrieval Systems," Journal
of the American Society for Information Science 31 (July 1980):
271.

11.  Michael Buckland and Fredric Gey, personal communication,
1991.

12.  Robert Wages, "Can Easy Searching be Good Searching? A Model
for Easy Searching," Online 13 (May 1989): 80.

13.  William S. Cooper, "On Selecting a Measure of Retrieval
Effectiveness," Journal of the American Society for Information
Science 24 (1973): 87-100, 413-424.  Compare this with: Dagobert
Soergel, "Is User Satisfaction a Hobgoblin?," Journal of the
American Society for Information Science 27 (July-August 1976):
256-259.

14.  Ibid., 88.

15.  Judith A. Tessier, Wayne W. Crouch, and Pauline Atherton,
"New Measures of User Satisfaction with Computer-Based Literature
Searches," Special Libraries 68 (November 1977): 383-389.

16.  Marcia J. Bates, "Factors Affecting Subject Catalog Search
Success," Journal of the American Society for Information Science
28 (May 1977): 161-169.

17.  Mark T. Kinnucan, "The Size of Retrieval Sets," Journal of
the American Society for Information Science 43 (January 1992):
73.

18.  Susan E. Hilchey and Jitka M. Hurych, "User Satisfaction or
User Acceptance?  Statistical Evaluation of an Online Reference

service," RQ 24 (Summer 1985): 455.

19.  Renata Tagliacozzo, "Estimating the Satisfaction of Information Users," Bulletin of the Medical Library Association 65 (April 1977): 248.

20. Ibid.

21.  Melvon L. Ankeny, "Evaluating End-User Services: Success or Satisfaction," Journal of Academic Librarianship 16 (January 1991): 356.

22.  Ibid., 354.  See also: Ethel Auster and Stephen B. Lawton, "Search Interview Techniques and Information Gain as Antecedents of user satisfaction with Online Bibliographic Retrieval," Journal of the American Society for Information Science 35 (March 1984): 90-103.

23.  Sandra R. Wilson, Norma Starr-Schneidkraut, and Michael D. Cooper, Use of the Critical Incident Technique to Evaluate the Impact of MEDLINE. (Palo Alto, CA: American Institutes for Research, 1989), AIR-64600-9/89-FR.  For hypothetical examples as to the importance of unretrieved but relevant documents, see: Soergel, "Is User Satisfaction a Hobgoblin?," 258-259.

24.  Ankeny, "Evaluating End-User Services," 356.

25.  Debora Cheney, "Evaluation-Based Training: Improving the Quality of End-User Searching," Journal of Academic Librarianship 17 (July 1991): 155.

26.  Tefko Saracevic and Paul Kantor, "A Study of Information Seeking and Retrieving.  II.  Users, Questions, and Effectiveness," Journal of the American Society for Information Science 39 (May 1988): 177-196.

27. Tefko Saracevic, Paul Kantor, Alice Y. Chamis, and Donna Trivison, "A Study of Information Seeking and Retrieving.  I.  Background and Methodology," Journal of the American Society for Information Science 39 (May 1988): 161-176.  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.

28.  Saracevic and Kantor, "A Study of Information Seeking and Retrieving.  Part II," 193.

29. Ibid.

30.  Ray R. Larson, "The Decline of Subject Searching: Long Term Trends and Patterns of Index Use in an Online Catalog," Journal of American Society for Information Science 42 (April 1991): 198.

31.   Charles W. Simpson, "OPAC Transaction Log Analysis: The First Decade," in Advances in Library Automation and Networking, vol. 3, ed. Joe A. Hewitt (Greenwich, Conn.: JAI Press, 1989), 35-67.

32.   J. Dickson, "Analysis of User Errors in Searching an Online Catalog," Cataloging & Classification Quarterly 4 (Spring 1984): 19-38; Thomas A. Peters, "When Smart People Fail: An Analysis of the Transaction Log of an Online Public Access Catalog," Journal of Academic Librarianship 15 (November 1989): 267-273; Rhonda N. Hunter, "Successes and Failures of Patrons Searching the Online Catalog at a Large Academic Library: A Transaction Log Analysis," RQ 30 (Spring 1991): 395-402; and Steven D. Zink, "Monitoring User Search Success through Transaction Log Analysis: the WolfPac Example," Reference Services Review 19 (1991): 49-56.

+ Page 45 +

33.   Martha Kirby and Naomi Miller, "MEDLINE Searching on Colleague: Reasons for Failure or Success of Untrained End Users," Medical Reference Services Quarterly 5 (1986): 17-34; and Cynthia J. Walker et al., "Problems Encountered by Clinical End Users of MEDLINE and GRATEFUL MED," Bulletin of the Medical Library Association 79 (January 1991): 67-69.

34.   Hunter, "Successes and Failures," 401.

35.   Stephen Walker and Micheline Hancock-Beaulieu, Okapi at City: An Evaluation Facility for Interactive Information Retrieval (London: The British Library, 1991), British Library Research Report 6056; and Ray R. Larson, "Classification Clustering, Probabilistic Information Retrieval and the Online Catalog," Library Quarterly 61 (April 1991): 133-173.

36.   Stephen Walker and Richard M. Jones, Improving Subject Retrieval in Online Catalogues, 1: Stemming, Automatic Spelling Correction and Cross-Reference Tables  (London: The British Library, 1987), 139, British Library Research Paper 24.  See also: R. Jones, "Improving Okapi: Transaction Log Analysis of Failed Searches in an Online Catalogue," Vine no. 62 (1986): 3-13.

37.   Larson, "The Decline of Subject Searching," 198.

38.   Sharon Seymour, "Online Public Access Catalog User Studies: A Review of Research Methodologies, March 1986-November 1989," Library and Information Science Research 13 (1991): 97.

39.   Micheline Hancock-Beaulieu, Stephen Robertson and Colin Neilson, "Evaluation of Online Catalogues: Eliciting Information from the User," Information Processing & Management 27 (1991): 532.

40.   John C. Flanagan, "The Critical Incident Technique," Psychological Bulletin 51 (1954): 327.

41.  Wilson, Starr-Schneidkraut and Cooper, Use of the Critical Incident Technique to Evaluate the Impact of MEDLINE, 2.

42.  Sammy R. Alzofon and Noelle Van Pulis, "Patterns of Searching and Success Rates in an Online Public Access Catalog," College & Research Libraries 45 (March 1984): 110-115; Marcia J. Bates, "Subject Access in Online Catalogs: a Design Model," Journal of American Society for Information Science 37 (1986): 357-376; Christine L. Borgman, "Why are Online Catalogs Hard to Use? Lessons Learned from Information-Retrieval Studies," Journal of American Society for Information Science 37 (1986): 387-400; Pauline A. Cochrane and Karen Markey, "Catalog Use Studies Since the Introduction of Online Interactive Catalogs: Impact on Design for Subject Access," Library and Information Science Research 5 (1983): 337-363; Mary Noel Gouke and Sue Pease, "Title Searches in an Online Catalog and a Card Catalog: A Comparative Study of Patron Success in Two Libraries," Journal of Academic Librarianship 8 (July 1982): 137-143; Charles R. Hildreth, Intelligent Interfaces and Retrieval Methods for Subject Searching in Bibliographic Retrieval Systems (Washington, DC: Cataloging Distribution Service, Library of Congress, 1989); Beverly Janosky, Philip J. Smith, and Charles Hildreth, "Online Library Catalog Systems: An Analysis of User Errors," International Journal of Man-Machine Studies 25 (1986): 573-592; Neal N. Kaske, A Comprehensive Study of Online Public Access Catalogs: an Overview and Application of Findings (Dublin, OH: OCLC, 1983), OCLC Research Report # OCLC/OPR/RR-83- 4; Cheryl Kern-Simirenko, "OPAC User Logs: Implications for Bibliographic Instruction," Library Hi Tech 1 (1983): 27-35; Ray R. Larson, "Workload Characteristics and Computer System Utilization in Online Library Catalogs" (Ph.D. diss., University of California at Berkeley, 1986); Gary S. Lawrence, V. Graham, and H. Presley, "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 (1984): 85-208; Karen Markey, Subject Searching in Library Catalogs: Before and After the Introduction of Online Catalogs (Dublin, OH: OCLC, 1984); Karen Markey, "Users and the Online Catalog: Subject Access Problems," in The Impact of Online Catalogs, ed. J.R. Matthews. (New York: Neal-Schuman, 1986), 35-69; Joseph K. Matthews, 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., 1982); Joseph Matthews, Gary S. Lawrence, and Douglas Ferguson, eds., Using Online Catalogs: a Nationwide Survey. (New York: Neal-Schuman, 1983); and Chih Wang, "The Online Catalogue, Subject Access and User Reactions: A Review," Library Review 34 (1985): 143-152.

43.  Examples of such studies are (in chronological order): Cyril W. Cleverdon, Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems (Cranfield, England: Aslib, 1962); Cleverdon, Mills and Keen, Factors Determining the Performance of Indexing Systems,

Volume 1, Design; Cleverdon and Keen, Factors Determining the
Performance of Indexing Systems, Volume 2, Test Results; F. W.
Lancaster, Evaluation of the MEDLARS Demand Search Service.
(Washington, DC: US Department of Health, Education and Welfare,
1968); F. W. Lancaster, "MEDLARS: Report on the Evaluation of Its
Operating Efficiency," American Documentation 20 (1969): 119-142;
Dickson, "Analysis of User Errors in Searching an Online
Catalog"; Blair and Maron, "An Evaluation of Retrieval
Effectiveness for a Full-Text Document-Retrieval System"; Jones,
"Improving Okapi: Transaction Log Analysis of Failed Searches in
an Online Catalogue"; Karen Markey and Anh N. Demeyer, Dewey
Decimal Classification Online Project: Evaluation of a Library
Schedule and Index Integrated into the Subject Searching
Capabilities of an Online Catalog (Dublin, OH: OCLC, 1986),
Report Number: OCLC/OPR/RR-86-1; Kirby and Miller, "MEDLINE
Searching on Colleague"; S. Walker and Jones, Improving Subject
Retrieval in Online Catalogues; Wilson, Starr-Schneidkraut, and
Cooper, Use of the Critical Incident Technique to Evaluate the
Impact of MEDLINE; Simone Klugman, "Failures in Subject
Retrieval," Cataloging & Classification Quarterly 10 (1989): 9-
35; Peters, "When Smart People Fail"; Ankeny, "Evaluating End-
User Services: Success or Satisfaction"; Hunter, "Successes and
Failures"; C. Walker et al., "Problems Encountered by Clinical
End Users of MEDLINE and GRATEFUL MED"; and Zink, "Monitoring
User Search Success through Transaction Log Analysis: the WolfPac
Example."

44.  Cleverdon, Report on the Testing and Analysis; Cleverdon,
Mills and Keen, Factors Determining the Performance of Indexing
Systems, Volume 1, Design; and Cleverdon and Keen, Factors
Determining the Performance of Indexing Systems, Volume 2, Test
Results.

45.  Cleverdon, Report on the Testing and Analysis, 1.

46.  Ibid., 8-9.

47.  Ibid., 89.  The design and findings of the Cranfield I
experiment have been criticized by many authors.  For example,
see: Don R. Swanson, "The Evidence Underlying the Cranfield
Results," Library Quarterly 35 (1965): 1-20.  For a review of the
Cranfield tests, see: Karen Sparck Jones, "The Cranfield Tests,"
in Information Retrieval Experiment, ed. Karen Sparck Jones
(London: Butterworths, 1981), 256-284.

48.  Ibid., 11.

49.  Swanson, "The Evidence Underlying the Cranfield Results,"
5.

50.  This percentage was obtained by averaging the figures given
in the fifth column of Table 3.1 of Cleverdon, Report on the
Testing and Analysis, 22.

51.  This summary is based on Cleverdon, Report on the Testing

and Analysis, Chapter 5.  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).

52.  Ibid., 88.

53.  Cleverdon, Mills, and Keen, Factors Determining the
Performance of Indexing Systems, Volume 1, Design; and Cleverdon
and Keen, Factors Determining the Performance of Indexing
Systems, Volume 2, Test Results.

54.  Cleverdon and Keen, Factors Determining the Performance of
Indexing Systems, Volume 2, Test Results, i ("Summary").  For the
detailed performance figures along with recall/precision graphs,
see volume 2 of the full report.

55.  Lancaster, Evaluation of the MEDLARS Demand Search Service.

56.  Ibid., 16, 19.

57.  Ibid., 19-20.

58.  Lancaster, "MEDLARS: Report on the Evaluation of Its
Operating Efficiency," 123.

+ Page 49 +

59.  Ibid., 127.

60.  Ibid., 131.

61.  Blair and Maron, "An Evaluation of Retrieval Effectiveness
for a Full-Text Document-Retrieval System."

62.  Ibid., 290-291.

63.  Ibid., 291-293.

64.  Ibid., 293.

65.  Ibid., 295.

66.  Markey and Demeyer, Dewey Decimal Classification Online
Project, 1.

67.  Ibid., 109.

68.  Ibid., 162.

69.  Ibid., 144.

70.  Ibid., 146.

71.  Ibid., 149.

72.  Ibid., 165, Table 42.

73.  Ibid., 162.

74.  Ibid., 166.

75.  Ibid., 182.

76.  Ibid.; especially, see Chapter 8, 173-291.

77.  Hilchey and Hurych, "User Satisfaction or User Acceptance?"

78.  Ankeny, "Evaluating End-User Services," 352-354.

79.  Ibid., 354.

+ Page 50 +

80.  Ibid.

81.  Kirby and Miller, "MEDLINE Searching on Colleague."

82.  Ibid., 20.

83.  Ibid.

84.  Dickson, "Analysis of User Errors in Searching an Online Catalog," 26.

85.  Jones, "Improving Okapi: Transaction Log Analysis of Failed Searches in an Online Catalogue."

86.  Ibid., 7-8.

87.  S. Walker and Jones, Improving Subject Retrieval in Online Catalogues, 117-119.

88.  S. Walker and Hancock-Beaulieu, Okapi at City, 30.  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).  Ibid., 24-25.

89.  Ibid., 31.

90.  Peters, "When Smart People Fail."

91.  Ibid., 270.

92.  Hunter, "Successes and Failures."

93.  C. Walker, et al., "Problems Encountered by Clinical End Users of MEDLINE and GRATEFUL MED."

94.  Ibid., 68.

+ Page 51 +

95.  Zink, "Monitoring User Search Success."

96.  Ibid., 51

97.  The following studies should be exempted from this as their
analyses were not based on zero-hit searches only:  Jones,
"Improving Okapi: Transaction Log Analysis of Failed Searches in
an Online Catalogue";  S. Walker and Jones, Improving Subject
Retrieval in Online Catalogues;  and S. Walker and
Hancock-Beaulieu, Okapi at City.

98.  Lancaster, Evaluation of the MEDLARS Demand Search Service.

99.  Wilson, Starr-Schneidkraut and Cooper, Use of the Critical
Incident Technique to Evaluate the Impact of MEDLINE.

100.  Ibid., 5.

101.  Ibid., 81.

102.  Ibid., 83-84.

103.  Gouke and Pease, "Title Searches in an Online Catalog and a
Card Catalog," 139.

104.  Alzofon and Van Pulis, "Patterns of Searching and Success
Rates in an Online Public Access Catalog," 113.

105.  Janosky, Smith and Hildreth, "Online Library Catalog
Systems: An Analysis of User Errors."

106.  Ibid., 576.

107.  Ibid., 591.

108.  Hildreth, Intelligent Interfaces and Retrieval Methods for
Subject Searching in Bibliographic Retrieval Systems, 69.

+ Page 52 +

109.  Bates, "Subject Access in Online Catalogs: a Design Model";
Borgman, "Why are Online Catalogs Hard to Use? Lessons Learned
from Information-Retrieval Studies"; David R. Gerhan, "LCSH in
vivo: Subject Searching Performance and Strategy in the OPAC
Era," Journal of Academic Librarianship 15 (1989): 83-89;
Klugman, "Failures in Subject Retrieval"; David Lewis, "Research
on the Use of Online Catalogs and Its Implications for Library
Practice," Journal of Academic Librarianship 13 (1987): 152-157;
Karen Markey, "Users and the Online Catalog: Subject Access
Problems," in The Impact of Online Catalogs, ed. J.R. Matthews.
(New York: Neal-Schuman, 1986), 35-69; Wang, "The Online
Catalogue, Subject Access and User Reactions: A Review."

110.  Larson, "Between Scylla and Charybdis: Subject Searching in

the Online Catalog," 181.

111.   Larson, "The Decline of Subject Searching," 208.

112.   University of California Users Look at MELVYL: Results of a Survey of Users of the University of California Prototype Online Union Catalog. (Berkeley, CA: The University of California, 1983), 97.

113.   Larson, "Classification Clustering, Probabilistic Information Retrieval and the Online Catalog," 136-144

114.   Allyson Carlyle, "Matching LCSH and User Vocabulary in the Library Catalog," Cataloging & Classification Quarterly 10 (1989): 37.

115.   Noelle Van Pulis, and L.E. Ludy, "Subject Searching in an Online Catalog with Authority Control," College & Research Libraries 49 (1988): 528-529.

116.   Diane Vizine-Goetz and Karen Markey Drabenstott, "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, ed. Jose-Marie Griffiths (Medford, NJ: Learned Information, 1991), 157.

+ Page 53 +

117. Ibid., 161.

118.   Michael K. Buckland and Doris Florian, "Expertise, Task Complexity, and Artificial Intelligence: A Conceptual Framework," Journal of American Society for Information Science 42 (October 1991): 635-643.

Acknowledgements

About the Author

Yasar Tonta, Ph.D. candidate, School of Library and Information Studies, University of California, Berkeley, CA 94720.