Lau, E.P., and Goh, D.H. (2006). In search of query patterns: A case study of a university OPAC. *Information Processing and Management*, 42(5), 1316-1329.

In Search of Query Patterns: A Case Study of a University OPAC

Eng Pwey Lau, Dion Hoe-Lian Goh*

Division of Information Studies, School of Communication and Information

Nanyang Technological University, Singapore 637718

{ps7475787a, ashlgoh}@ntu.edu.sg

* Address correspondence to Dion H. Goh, Division of Information Studies, School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore 637718. Phone: (65) 6790-6290. Fax: (65) 6791-5214. Email: ashlgoh@ntu.edu.sg.

Abstract

A transaction log analysis of the Nanyang Technological University (NTU) OPAC was conducted to identify query and search failure patterns with the goal of identifying areas of improvement for the system. One semester's worth of OPAC transaction logs were obtained and from these, 641 991 queries were extracted and used for this work. Issues investigated included query length, frequency and type of search options and Boolean operators used as well as their relationships with search failure. Among other findings, results indicate that a majority of the queries were simple, with short query lengths and a low usage of Boolean operators. Failure analysis revealed that on average, users had an almost equal chance of obtaining no records or at least one record to a submitted query. We propose enhancements and suggest future areas of work to improve the users' search experience with the NTU OPAC.

Keywords

Online Public Access Catalog; Transaction log analysis; Failure analysis; Information retrieval

1. Introduction

The Internet and more specifically the Web has undoubtedly become the largest available repository of information. The increased reach of Web-based search has made information retrieval (IR) systems a global tool for information access (Wolfram & Xie, 2001). One problem however is that the content of most Web documents retrieved are often of suspect credibility and reliability (Center For Communication Policy, 2003). In contrast, traditional bibliography IR systems such as online catalogs, which are often available in academic and public libraries, are perceived to be more authoritative (Drabenstott & Weller, 1996).

However, online catalogs have been criticized as being hard to use because their designs do not incorporate sufficient understanding of search behaviors (Borgman, 1996). For example, Warren (2000) discovered that the general design of the Urica OPAC system hindered rather than helped users in the searching process. Scrolling through large retrieval sets may be one of the causes of user frustration. While record ranking based on queries has been commonly adopted in IR systems, a successful search in any IR system is still dependent on its composition and query construction (Borgman 1996; Walkerdine & Rodden, 2001). However, the ability of IR systems to analyze query terms and intelligently interpret a user's actual information needs is still far from being perfect. For example, Large and Beheshti (1999) reported that users encounter many problems in choosing suitable search terms to represent their subject interests. There is thus a need to establish a closer working relationship between systems designers and users to develop useful IR systems.

The success of an IR system can be evaluated via a number of quantitative or qualitative methods (Jansen, 2000; Spink, Bateman & Jansen, 1998). Transaction log analysis is an example of the quantitative approach that can lead to plausible conclusions about user search behaviors. The contents of transaction logs are usually formatted in fields to facilitate quantitative analysis (Covey, 2002) and contain information about users' search behaviors. This includes the characters entered by users, timestamps, search command identifiers, the number of hits and the responses from the IR system. The logs are thus able to reveal the queries executed and strategies employed by users. Transaction logs analysis avoids the problem of reliability of user responses or normality of searching behaviors when users know they are the subjects of investigation. These drawbacks are often encountered in interviews, questionnaires and observation research (Bains, 1997).

The present study focuses on Nanyang Technological University (NTU), Singapore, with the goal of determining if its OPAC system can help meet its users' information needs. NTU, with over 1 300 academic staff and 24 000 students, provides a wide-range of tertiary-level programs including engineering, information technology, the humanities, social sciences, and business. The NTU OPAC is the main access point to staff and students for meeting their information needs, and serves as an information gateway to a diverse range of resources maintained by the university. Types of resources range from books, journals, theses, databases and multimedia productions.

This study employs transaction log analysis to examine users' search patterns with the aim of assessing the ability of the NTU OPAC in satisfying the collective (as opposed to individual)

needs of its users (Tague-Sutcliffe, 1996). Transaction logs thus appear to be the most reasonable and non-intrusive means of collecting such information since the probability of obtaining results by chance is minimized (Ke, Kwakkelaar, Tai & Chen, 2002) when the data volume is large. In particular, we analyze the characteristics of the queries issued and the use of various search options. This includes information such as query length, use of Boolean operators and so on. In short, the two objectives of this research are:

- To discover patterns in user search behaviors
- To explore the reasons for search failure

To date, no systematic study has been done on the NTU OPAC and the present work is therefore timely.

The remainder of this paper is divided as follows. A review of related literature in Section 2 highlights the relationship between the present study and previous work done in digital libraries, Web-based IR systems and OPAC systems. We then present our data set and our findings. The final section proposes recommendations in the light of our results and suggests opportunities for future work.

2. Related Work

2.1 *Transaction Log Analysis*

Transaction log analysis has been widely used to study the interactions between IR systems and their users (Jansen & Pooch, 2000; Ke, Kwakkelaar, Tai & Chen, 2002). Bysouth (1990) found that the search terms recorded in the logs were very useful for analysis. For example, the logs reveal information such as which search options were selected, which types of

Boolean operators were used and how many terms were used. When all the information is pieced together, an idea of "how" the search is carried out by the user emerges. On the other hand, transaction log analysis falls short of providing an understanding of user perceptions of their searches (Griffiths, Hartley & Willson, 2002). The understanding of "what the users say they do" is impossible to reveal by just analyzing transaction logs. Although such logs provide highly detailed information about how users interact with the system, they do not reveal users' intentions or whether users are satisfied with the results (Larson, 1991).

Nevertheless, researchers have creatively used the information available in these logs to overcome some of these limitations. For instance, Pu, Chuang and Yang (2002) collected queries from different periods from three search engine logs to remove bias due to the effect of time and locality, and to ensure accuracy. Covey (2002) used IP address and domain names captured in the transaction logs to identify user demographics when user authentication data were unavailable. The interesting observations and conclusions drawn from analyzing transaction logs have therefore led researchers to continue using this approach to examine users' query patterns. The following subsections present related work that examine query characteristics in different IR systems.

2.2 *Query Patterns*

Jones, Cunningham, McNab and Boddie (2000) analyzed the transaction logs of the Computer Science Technical Report Collections (CSTR) in the New Zealand Digital Library and found that the average number of search terms was 2.43 per query. This result was consistent with a transaction log analysis of ResearchIndex (RI), a digital library of computer science content, with a reported average of 2.32 terms per query (Mahoui & Cunningham

2001). Likewise, Wallace (1993) obtained an average query length of one to two terms per query among users in an OPAC system at the University of Colorado. Similar findings were obtained from the transaction logs analysis of the Excite search engine. From an analysis of 51 473 queries, about 62.0% of queries contained one or two terms, giving an average of 2.21 terms (Jansen, Spink & Saracevic, 2000).

It is often assumed that the correct use of search operators would increase the effectiveness of searches (Eastman & Jansen, 2003). Yet, the usage of operators is relatively low because users tend to use simple queries. Query complexity refers to query commands issued and includes the use of techniques such as stemming, phrase search, and Boolean operators (Jansen & Pooch, 2000). Moukdad and Large (2001) concluded that a high proportion of Web users do not utilize advanced search features and those who do frequently misunderstand them. In their work, only 7.8% of the total queries submitted to WebCrawler used one or more Boolean operators but only 35.8% of the Boolean searches were correctly issued. Another study from Excite revealed that less than 10% of the total queries contained Boolean operators (Jansen, Spink, & Saracevic, 2000). Similar results have been found for digital libraries. For example, Mahoui and Cunningham (2001) found that only 37.6% of queries in RI used Boolean operators while this value was 32.9% for CSTR.

The use of advanced search features is also relatively low in OPACs even though many of these systems are feature rich. For example, a survey conducted on the features provided by OPACs in Tamil Nadu (India) revealed that 86.2% of the systems offered advanced search features in addition to the simple search option (Babu & Tamizhchelvan, 2003). These include options such as author, keyword, subject heading and title searches, plus many others.

Nevertheless, users have been known to encounter difficulties in constructing effective search strategies using these features. For example, Lombardo and Condic (2000) conducted a survey on the new release of the OPAC system at Oakland University and discovered that many of its users were unable to take advantage of more sophisticated search techniques because the features were not intuitive. A separate study of the University of Colorado OPAC also showed that only 8.7% of all queries used advanced search features (Jansen & Pooch, 2000).

2.3 *Failure Analysis*

Search failures have been examined in numerous IR studies. Jones, Cunningham, McNab and Boddie (2000) defined a failed search as one that matches no documents in the collection, and found that 10.5% of queries in CSTR were failures. These failed searches included queries with zero terms, ranked queries and Boolean queries. An analysis of Excite searches found that most failures were due to "incorrect use" – mistakes in the use of Boolean operators, spaces and other operators that were not supported by the search engine (Jansen, Spink & Saracevic, 2000). For example, 37% of "AND NOT" queries were incorrect, 32% of "AND" queries were misused and 26% of "OR" queries contained errors. Another reason for failed searches was found in a study on the Mardigian and Lilly Libraries OPAC (Drabenstott & Weller, 1996). Here, the reason was the lack of user perseverance. This conclusion was obtained from post-search questionnaires which required users to rate whether their searches were successful or not. An analysis of the data showed that some of the supposedly unsuccessful searches were in fact useful ones that could lead users to their topic of interest. Due to a lack of perseverance however, these users chose to terminate their searches rather than look beyond the first screen full of results.

3. Data Set

As discussed, a transaction log analysis was performed on the NTU Library OPAC. These logs provide detailed accounts of user activities which include authentication (login/logout), query operations, renewal requests and so on, with each activity translated into one record in the transaction log. In total, one semester's (4.5 months) worth of records were obtained and these contained 641 991 queries, all of which were used for this work. Among the query records, the data extracted for analysis included search options used, number of terms per query, query terms and the number of records retrieved.

We then proceeded to meet our two objectives by investigating the following issues:

- Average, minimum and maximum query length
- Frequency and type of search options used
- Frequency and types of Boolean operators used
- Types of typographical errors
- Relationship between search failure and search options
- Relationship between search failure and Boolean operators
- Relationship between search failure and query length

4. Findings and Analyses

In this section, we highlight search patterns from the aspects of search options used, length of queries issued, Boolean operators used, and their relationship with search failure.

4.1 *Default Settings in the NTU OPAC*

Since default settings within a system would affect the outcome of searches and how users conduct their search, we first present an overview of the NTU OPAC, its query interface and default search options. Figure 1 shows the NTU OPAC's query interface which allows users to search for both the NTU library's collection as well as those of partner libraries (see Section 1 of Figure 1). Queries may be conducted by author, title, subject or keyword (see Section 3 of Figure 1), with keyword search being the default if the user submitted a query by pressing the "Enter" key on the keyboard. This option searches against all fields of a bibliographic record and is therefore useful for novice searchers or those unsure about their domains of interest. Users may also limit the search by language or material type (see Section 4 of Figure 1) with the default being the most general – all languages and all material types.

[Insert Figure 1 here]

Hildreth (1995b) argued that an online catalog system should never permit a user's search attempt to fail to retrieve at least one bibliographic record for review and action. One way to achieve this is to conduct a broad search, and this is presumably the idea behind the NTU OPAC's default search options, which were keyword search across all languages (e.g. Chinese and English) and all material types (books, journals, theses, computer files, maps, audio-visual and sound recording materials). Nevertheless, a hidden default was found within the system that appeared contradictory to this intended design. That is, when no Boolean operators were explicitly entered, the system assumed a phrase search, requiring terms to be entered in the same order as they appeared in the bibliographic records.

Apart from the occasional use of Boolean operators, there was little evidence shown in the transaction logs that users had changed these default settings to a more restricted search scope. This could be due to the fact that users were unsure about the type of material that would most likely contain the information they were looking for, and/or they would rather to be "conservative", preferring to view more records than receiving zero hits. This observation was consistent with a study on the New Zealand Digital Library (NZDL) in which users rarely changed the default settings for query options (Jones, Cunningham, McNab & Boddie, 2000).

4.2 *Search Options Used*

Results indicate that the most commonly used search option for the NTU OPAC is the keyword search. The use of keyword searches contributed to 68.9% of all queries while other options such as title, author and subject accounted for 16.5%, 8.2% and 6.4% of all searches respectively (see Table 1 columns (a)-(c)). Three reasons could account for the popularity of keyword search. Firstly, keyword search was the default option if the user submitted a query by pressing the "Enter" key on the keyboard. Secondly, keyword search is helpful when users are not clear about which resources they are looking for. Thirdly, the tip given on the OPAC query page (see Figure 1), suggesting that keyword searches will result in all indexed fields being searched, might have indirectly influenced users to select this option. The tip gives users the impression that keyword search is generally the easiest to perform as all indexed fields of a bibliographic record are searched.

Keyword search exceeded subject search by more than 10 times in the NTU OPAC. There were 442 523 keyword searches as compared to 41 048 subject searches. Unlike keyword search, terms used in subject search need to be matched exactly with subject headings. If matches did not occur, a terse "Item not found" message would be displayed, giving no contextual help to assist in further query refinement. Previous work has shown that there was a consistent decline of subject searching in large OPAC systems. For example, the regression model built by Larson (1991) indicated that the percentage of subject searching in the University of California's OPAC, MELVYL, decreased by about 0.0062% per day over the data collection period. Larson suggested that the major problem that contributed to the decline was due to users' difficulties in formulating subject queries using the Library of Congress Subject Headings (LCSH).

[Insert Table 1 here]

Table 1 also shows that there were more queries on author and title as compared to the subject field. This is interesting because author or title search is more specific and requires a user to know exact author information or the exact first few words of a title before a search can be successfully executed. The finding suggests that certain groups of users knew exactly what they were looking out for and this, to a certain extent, is consistent with Lombardo and Condic's (2000) work which found that most OPAC users could identify whether they required an author or title search, but could not determine when to use subject search.

4.3 *Query Length*

Table 2 (columns (a)-(c)) shows the query lengths (number of terms), their frequencies and percentages in relation to the total number of queries in the transaction logs. The most common length was two terms (34.1%), followed by a single term (23.6%) and three terms (17.3%). Thus, queries comprising not more than three terms constituted approximately 75.0% of all submitted queries. Put differently, only 25.0% of all queries employed more than three terms. Interestingly, 35 queries were submitted without any terms and that became the minimum query length. At the other extreme, the maximum query length was 46.

[Insert Table 2 here]

A further analysis of query lengths was conducted by using the median query length as a cut-off point. Most of the long queries that exceeded the median (queries with greater than or equal to 23 terms) were found to be keyword searches with diverse topics ranging from art, business to science. Two examples are highlighted in Table 3. These queries share a common characteristic in that they have a very narrow focus and provide a large amount of context.

[Insert Table 3 here]

Although long queries were discovered in the NTU OPAC transaction logs, the average query length was found to be only 2.86 terms, with a standard deviation of 2.102. This value is consistent with the short queries found in other OPAC and Web IR systems, and suggest that query lengths have not changed throughout an extended period of time. For instance, Wallace (1993) found that 75.0% of the searches in the University of Colorado OPAC had two or

fewer terms. An analysis of the Alta Vista search engine query logs revealed an average query length of 2.35 (Silverstein, Henzinger, Marais & Moricz, 1998). Our results were also consistent with the investigations conducted on digital libraries (Jones, Cunningham, McNab & Boddie, 2000; Mahoui & Cunningham 2001)..

4.4 *Boolean Operators Used*

Table 4 (columns (a)-(c)) shows limited use of Boolean operators by NTU OPAC users in which only 11.8% of all queries involved the use Boolean operators. Of these, about 10.4% contained only one type of Boolean operator such as "AND", "OR" or "NOT". This is consistent with Matthews (1997) who found that the Boolean feature was infrequently utilized despite the fact that the majority of OPAC systems support Boolean searching. Excite users were also found to hardly ever use Boolean operators in their queries (Spink, Wolfram, Jansen & Saracevic, 2001).

[Insert Table 4 here]

In the NTU OPAC, approximately 88.2% of queries were submitted without the use of Boolean operators (see Table 4). This is despite the fact that a tip on the OPAC query page (see Figure 1) indicated that the use of operators and qualifiers was permitted. However, this brief tip did not provide any usage examples and so may have been ignored by users who either did not understand the meaning of the tip, or who would rather issue simple keyword-only queries similar to Web-based searches. This observation is similar to Yuan (1997), who found out that most OPAC users tended to learn only enough to do simple searches.

Furthermore, most users did not understand how to use Boolean operators in OPAC systems properly (Hildreth, 1997).

In the NTU OPAC, Boolean operators were specified by the operators "AND", "OR", and "NOT" as part of a keyword search. This approach of making Boolean operators optional appears to be consistent with research showing that OPAC users were generally not in favor of using Boolean operators (Dinet, Favart & Passerault, 2004). Nevertheless when Boolean operators were used in the NTU OPAC, the "AND" operator was the most popular and it accounted for approximately 10.3% of all queries. The high frequency of the "AND" operator suggests that a majority of users who used Boolean operators started the search with a narrowing strategy in mind, consistent with the work of Ke, Kwakkelaar, Tai and Chen (2002). The high degree of precision could have possibly been the reason why NTU OPAC users explored more searches using the "AND" operator. The "OR" operator on the other hand, is useful when users are unfamiliar with which terms to employ in a search. However, only a very small percentage (less than 0.1%) used it. Like the prior study on Excite users (Spink, Wolfram, Jansen & Saracevic, 2001), only a minuscule percentage of users used the "NOT" operator (0.01% of all queries).

Boolean operators can also be combined within a single query but our results show that only 1.38% of all queries were submitted as such (see Table 4). A few combinations of Boolean operators are possible:

- OR, NOT
- AND, OR
- AND,NOT
- AND, NOT, OR

Among these, the “AND” and “OR” pair was the most frequently used (about 1.37% of all queries). Formulating useful Boolean queries is indeed an art (Salton, 1988). A large retrieval set may overwhelm users when queries are broadly formulated by the “OR” operator while queries with the “AND” operator may generate too few results. Hence, the combination of “AND” and “OR” operators could balance the number of search results, which perhaps explains why these two operators were the most frequently combined. There was a low percentage of queries using the full combination of “NOT”, “OR” and “AND” operators (0.0039%) as well as the “NOT” and “AND” pair (0.0028%). Only 2 queries (0.0003%) were found using the combination of “NOT” and “OR” operators. These findings suggest that NTU OPAC users seldom negate their query statements.

4.5 *Failure Analysis*

In the following, we discuss several aspects of search failure in the NTU OPAC. We adopt a practical definition of search failure as one in which zero hits are returned in response to a query (Hunter & Rhonda, 1991; Jones, Cunningham, McNab & Boddie, 2000).

4.4.1 Search Options

A total of 317 840 queries returned zero hits and this figure accounted for 49.5% of the total number of queries submitted to the NTU OPAC (see Table 1 columns (d)-(e)). This indicates that users have an almost equal chance of submitting a query that returns no hits (a search failure) or one that returns at least one record. On closer examination, the table also indicates that most title searches were successful in that a larger proportion generated non-zero hits.

Table 1 shows that the results returned for author searches were rather balanced for both zero (4.1%) and non-zero hits (4.0%). Among keyword searches, the most popular search option, 34.2% of all queries returned zero hits while 34.7% were non-zero hits. The worst performer was subject search in which users experienced high failure rates (4.5% zero hits versus 1.9% non-zero hits). One possible reason for this high failure rate might be because users were unaware that the NTU OPAC could only retrieve records that started with the search terms. Thus, subject searches were comparatively more difficult because the query must not only start with the correct terms but also conform to LCSH. The lack of knowledge of LCSH on the part of users could possibly account for this high failure rate. Dranbenstott and Weller (1996) highlighted that despite a decade of online catalog research, development and deployment, many of the same problems that the earliest online catalog searchers encountered continued to be experienced by today's users, particularly in the area of subject access. This is consistent with a study on MELVYL that showed that 44% of searches produced no hits with one-third of the subject queries failing to produce retrievals (Lynch, 1988).

4.4.2 Typographical Errors

Typographical errors can be divided into four main categories, namely, substitutions, omissions, insertions and transpositions (Brown, 1988; Hirst & Budanitsky, 2003). Such errors can be major causes for unsuccessful searches because accurate spelling is typically required (Borgman, Hirish, Walter & Gallagher, 1995). An investigation was performed on a random sample of 1 000 queries with zero hits returned. These queries were examined manually to identify patterns behind typographical errors. Next, erroneous terms were grouped into the four error categories.

Of the 1 000 queries, a very small number of only 24 were identified with typographical errors. Table 5 provides some examples of such errors. Of the 24, most of the typographical errors were due to omissions. For instance, the term “Enviromental” was entered without the letter “n” that occurs between the letter “o” and “m”. Substitutions were encountered as frequently as omissions. For example, the term “Fundamentals” was submitted as “Findamentals” instead. Here, the placement of letters on the keyboard is the likely cause, for example, the position of the letter “i” is next to “u”. In addition, there were terms found with letters in the wrong order (transpositional errors). For example, the word “deisngs” was found instead of the correct term “designs”. There were however few cases of errors due to additional characters (insertions).

[Insert Table 5 here]

4.4.3 Query Length

The number of search terms used in a query can affect the returned results. Table 2 (columns (d)-(e)) shows that there were more results with non-zero hits when query lengths were one or two. The percentage of successful queries with one search term was double that of failed queries (15.9% versus 7.7% respectively). For queries with two terms, the success rate was 19.7% against a 14.4% failure rate. In general, when query length increases, there is a higher probability that users would encounter unsuccessful searches. One of the possible explanations might be that the underlying retrieval logic of the NTU OPAC required users to enter search terms in the order in which they appeared in the library’s bibliographic records. Given a longer query length, the probability of matching the query string against these

records would be lowered. This finding is presented in Figure 2 which shows that as query length increases, the ratio of non-zero hit results to zero-hit results decreases. For query lengths greater than 25, there were only zero-hit results. Further support was found via a correlational analysis on query length against ratio (query length of zero was not considered). Here, a negative correlation was obtained [$r=-0.67$, $p<0.001$]. Taken together, our findings suggest that shorter query lengths decrease the likelihood of search failure.

[Insert Figure 2 here]

4.4.4 Boolean Operators

Table 4 (columns (d)-(e)) shows the number of Boolean queries returning non-zero and zero hits. Here, Boolean queries refer to those with single or combined Boolean operators. In total, there were 42 961 zero-hit Boolean queries and 32 517 non-zero hit queries. This outcome was relatively worse than the results obtained from an Excite transaction log analysis (Jansen, Spink & Saracevic, 2000). In that analysis, of the 4 376 Boolean queries, 1 392 queries contained errors leading to search failure. In contrast, more than half of the Boolean queries in the NTU OPAC transaction logs were failed searches.

Our results show that the failure rate of queries with the "AND" operator (6.0%) was higher than those of successful searches (4.3%). Although users could obtain higher precision by using the "AND" operator, they also ran into the risk of not obtaining any results when one or more of the terms were incorrectly used. Put differently, while the introduction of an "AND" clause could potentially reduce the number of records retrieved from the OPAC, it could also return zero hits in the worst case. This, to certain extent, provides a reason why queries with

the “AND” operator had higher search failure rates. In contrast, “OR” Boolean queries returned a higher proportion of non-zero hits relative to those with zero hits. Here, the use of “OR” Boolean operator increased the search scope which in turn reduced the likelihood of obtaining zero hits. Whether these results satisfied users’ information needs is however, not considered in this work. Queries using the “NOT” operator showed an almost equal likelihood of zero or non-zero hit results.

The proportion of queries with combined Boolean operators returning zero hits and non-zero hits varied depending on the types of operators used. Table 4 shows that Boolean queries with the combination of “AND” and “OR” operators had a tendency to produce fewer zero hits while those using “AND” and “NOT” produced an opposite result. The failure rates could be explained by the nature of the operators used. For example, the “NOT” operator is used to exclude a term from the search, which has a narrowing effect on search scope. Hence, when “NOT” is added to the “AND” operator, the search criteria becomes increasingly tightened. Of the 18 queries that used “AND” and “NOT”, 11 returned zero hits while the other seven returned non-zero hits. On other hand, the presence of an “OR” operator in a query with an “AND” operator has the effect of widening the search scope, leading to more non-zero hits being obtained.

5. Discussion and Conclusion

The ultimate goal of an IR system is to assist users in fulfilling their information needs. Analyzing an IR system’s transaction logs to uncover search and failure patterns can therefore yield ways to further improve the system. Such improvements not apply to the

system being analyzed but may lead to enhancements in the design of other systems as well. However, it is also important to overcome the present limitations for future work.

5.1 *Recommendations*

Our results show that simple searches with short query lengths and low usage of Boolean operators comprised the majority of NTU OPAC queries. The findings are consistent with results from prior work such as Mahoui and Cunningham (2001) and Jansen, Spink and Saracevic (2000). Our failure analyses show that on average, users had an almost equal chance of submitting a query that returned no results and one that returned at least one record. This is surprising given that with the popularity of search engines, one would assume that users would have become more proficient in searching. In sum, our results suggest that that search patterns have not changed very much across the years and across geographical boundaries despite many years of experience in the use, research and development of information retrieval systems.

With the results obtained in our work, we propose some recommendations for enhancing the NTU OPAC, all of which are applicable to other systems as well:

1. **Interactive query reformulation**. This refers to a family of techniques that assist users in query formulation by recommending alternative query terms (Belkin et al., 2001; Efthimiadis, 1996). A technique that we are currently exploring is to employ querying recommending techniques in the NTU OPAC (Fu, Goh & Foo, 2005). Here, submitted queries stored in transaction logs and mined, and related ones are recommended to users during query formulation. Users also have the opportunity to

explore relationships between queries using a graph-based visualization interface. Our approach thus allows users to peruse queries directly related to a submitted query, as well as those that are indirectly related.

2. **Browsing**. Given the poor performance of subject searches, the ability to browse subject indexes would be helpful. Browsing would also be a helpful complement for author searches since different name variants for the same person makes it harder for the users to retrieve all records by the same individual at once (Feitelson, 2004; Smith 2000). In addition, browsing is useful when information needs are vague, such as when a user is a novice in the domain (Hildreth, 1995a). Through browsing, users can identify documents of interest and from these, ideas for formulating/reformulating a query can then be established. Several techniques are available. One approach would be to include browsing options for authors, titles and subjects within the OPAC query page. A potential problem is that the list of entries could be large especially for LCSH subject headings. Hence a second browse-search approach could allow users to browse portions of various indexes that match specified search terms. A final technique involves hyperlinking selected terms (e.g. author names, keywords) in the bibliographic records or search results listings for browsing. This feature is similar to databases such as Web of Science and the ACM Digital Library.

3. **Context-sensitive assistance**. Borgman (1996) argued that searchers require varying levels of help in constructing queries. Further, differences between expert and novice users in articulating their information needs and the search strategies they adopt suggest that such assistance has to be tailored according to individuals (Sutcliffe,

Ennis & Watkinson, 2000). An approach that we are investigating uses concepts in subjective relevance to develop user-centered IR systems (Lee, Theng, Goh & Foo, 2005). In our work, desirable features of IR systems are elicited from novice and expert searchers, and these are mapped to different subject relevance types (Saracevic, 1996) and stages in an adopted online information seeking model (Marchionini, 1995). This mapping is then used to implement a prototype IR system that supports various subject relevance types in each stage of the information seeking process. Examples of features include term and document recommendations, customized help and links to experts and users with similar interests. However, for such user-centered services to be provided in the NTU OPAC, user profiles may have to be created and therefore users would have to authenticate themselves before using the system.

5.2 *Conclusion*

The findings of the present work have played a useful role in informing our ongoing research in the design and implementation of user-centered IR systems, including interactive query formulation using query recommending and the support for subjective relevance in information seeking. There are however limitations to transaction log analysis such as the inability to measure levels of user satisfaction. Similarly, the assumption of zero-hit queries being equated with search failure may not always hold, as a non zero-hit result does not necessarily indicate that a user's information need is met. Thus, one area of work that could be explored is to supplement the present analysis with other techniques including questionnaire surveys and interviews. For example, data obtained from survey responses was used to assess user search patterns in the Excite search engine (Spink, Bateman & Jansen, 1998). In addition, research in user-centered design suggests that it is essential to evaluate the

NTU OPAC interface and determine the extent it has helped users in translating their information needs into explicit query terms. Evaluation could be focused on aspects such as user satisfaction, ease of learning, ease of use and error prevention.

Acknowledgements

The authors would like to thank the NTU Library for providing their OPAC transaction logs for this research.

References

Babu, B.R., & Tamizhchelvan, M. (2003). An investigation into features of OPACs in Tamil Nadu (India). *Library Review*, 52(6), 257-267.

Bains, S. (1997). End-user searching behavior: Considering methodologies. *The Katherine Sharp Review*, 4. Retrieved July 20, 2005, from http://www.lis.uiuc.edu/review/winter1997/bains.pdf

Belkin, N.J., Cool, C., Kelly, D., Lin, S.-J.,Park, S.Y., Perez-Carballo, J., et al. (2001). Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information processing and management*, 37(3), 403-434.

Borgman, C.L. (1996). Why are online catalogs still hard to use? *Journal of the American Society for Information Science*, 47(7), 493-503.

Borgman, C.L., Hirsh, S.G., Walter, V.A., & Gallagher, A.L. (1995). Children's searching behavior on browsing and keyword online catalogs: The Science Library Catalog Project. *Journal of the American Society for Information Science*, 46(9), 663-685.

Brown, A. (1988). A Singaporean corpus of misspellings: Analysis and implications. *Journal of the Simplified Spelling Society*, 3. Retrieved July 2, 2005, from http://www.spellingsociety.org/journals/j9/singapore.php

Bysouth, P.T. (1990). Evaluating the use of several approaches to online literature retrieval by research scientists. In P.T. Bysouth (Ed.), *End-user searching: The effective gateway to published information* (pp. 105-123). London, UK: The Association for Information Management.

Center for Communication Policy (2003). *Surveying the digital future. The UCLA Internet report. Year Three.* Retrieved July 15, 2005, from http://www.forbes.com/fdc/mediaresourcecenter/UCLA03.pdf

Covey, D.T. (2002). *Usage and usability assessment: Library practices and concerns*. Washington, DC: Digital Library Federation and Council on Library and Information Resources.

Dinet, J., Favart, M., & Passerault, J. (2004). Searching for information in an online public access catalog (OPAC): the impacts of information search expertise on the use of Boolean operators [Abstract]. *Journal of Computer Assisted Learning*, 20(5), 338 – 346.

Drabenstott, K.M., & Weller, M.S. (1996). Failure analysis of subject searches in a test of a new design for subject access to online catalogues. *Journal of the American Society for Information Science*, 47(7), 519-37.

Eastman, C.M., & Jansen, B.J. (2003). Coverage, relevance, ranking: The impact of query operators on Web search engine results. *ACM Transactions on Information Systems*, 21(4), 383-411.

Efthimiadis, E.N. (1996). Query expansion. In M.E. Williams (Ed.), *Annual Review of Information Science and Technology*, 31 (pp. 121-187). Medford, NY: Information Today, Inc.

Feitelson, D.G. (2004). On identifying name equivalences in digital libraries. *Information Research*, 9(4). Retrieved February 3, 2006, from http://informationr.net/ir/9-4/paper192.html

Fu, L., Goh, D.H., and Foo, S. (2005). Evaluation of a collaborative querying system. *Proceedings of the 9th European Conference on Digital Libraries*. *Lecture Notes in Computer Science 3652* (pp. 502-503). Berlin, Germany: Springer-Verlag.

Griffiths, J.R., Hartley, R.J., & Willson, J.P. (2002). An improved method of studying user-system interaction by combining transaction log analysis and protocol analysis. *Information Research*, 7(4). Retrieved July 20, 2005, from http://Informationr.net/ir/7-4/paper139.html

Hildreth, C.R (1997). The use and understanding of keyword searching in a university online catalog. *Information Technology and Libraries*, 16(2), 52-62.

Hildreth, C.R. (1995a). *Online catalog design models: Are we moving in the right directions?* Retrieved July 23, 2003, from http://myweb.cwpost.liu.edu/childret/clr-opac.html

Hildreth, C.R. (1995b). *The GUI OPAC: Approach with caution.* Retrieved July 23, 2003, from http://info.lib.uh.edu/pr/v6/n5/hild6n5.html

Hirst, G., & Budanitsky, A. (2003). *Correcting real-word spelling errors by restoring lexical cohesion*. Retrieved July 20, 2005, from http://ftp.cs.toronto.edu/pub/gh/Hirst+Budanitsky-2005ms.pdf

Hunter, R.N., & Rhonda, N. (1991) Successes and failures of patrons searching the online catalog at a large academic library: A transaction log analysis. *Research Quarterly*, 30(3), 395-402.

Jansen, B.J. (2000). The effect of query complexity on Web searching results. *Information Research*, 6(1). Retrieved July 15, 2005, from http://Informationr.net/ir/6-1/paper87.html

Jansen, B.J., & Pooch, U. (2000). A review of Web searching studies and a framework for future research. *Journal of the American Society for Information Science and Technology*, 52(3), 235-246.

Jansen, B.J., Spink, A., & Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*, 36(2), 207-227.

Jones, S., Cunningham S. J., McNab, R., & Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal Digital Libraries*, 3(2), 152-169.

Ke, H., Kwakkelaar R., Tai Y., & Chen, L. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research*, 24(3), 265-291.

Large, A., & Beheshti, J. (1999). OPACs: A research review. *Library & Information Science Research*, 19(2), 111-133.

Larson, R.R. (1991). The decline of subject searching: Long-term trends and patterns of index use in an online catalog. *Journal of the American Society for Information Science*, 42(3), 197-215.

Lee, S.S., Theng, Y.L., Goh, D.H., and Foo, S. (2005). Subjective relevance: Implications on interface design for information retrieval systems. *Proceedings of the 8th Conference*

on Asian Digital Libraries. Lecture Notes in Computer Science 3815 (pp. 424-434). Berlin, Germany: Springer-Verlag.

Lombardo, S.V., & Condic K.S. (2000).Empowering users with a new online catalog. *Library Hi Tech*, 18(2), 130-141.

Lynch, C.A. (1988). Large database and multiple database problems in online catalogs. In M. Dillon (Ed.), *OPACS and Beyond* (pp. 51-56). Dublin, OH: Online Computer Library Center.

Mahoui, M., & Cunningham, S. (2001). Search behavior in a research-oriented digital library. *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries. Lecture Notes in Computer Science 2163* (pp. 13-24). New York: Springer-Verlag.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge, UK: Cambridge University Press.

Matthews, J.R. (1997). Time for new OPAC Initiatives: An overview of landmarks in the literature and introduction to WordFocus. *Library Hi Tech*, 15(1/2), 111-122.

Moukdad, H., & Large, A. (2001).User's perceptions of the Web as revealed by transaction log analysis. *Online Information Review*, 25(6), 349- 358.

Pu, H., Chuang, S., & Yang, C. (2002). Subject categorization of query terms for exploring Web users' search interests. *Journal of the American Society for Information Science and Technology*, 53(8), 617-630.

Salton, G. (1988). *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Boston, MA: Addison-Wesley.

Saracevic, T. (1996). Relevance reconsidered '96. In Ingwersen, P. and Pors, N.O. (Eds.). *Information Science: Integration in Perspective* (pp. 201-218). Copenhagen, Denmark: Royal School of Librarianship.

Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1998). *Analysis of a very large AltaVista query log.* Retrieved July 20, 2005, from http://gatekeeper.research.compaq.com/pub/DEC/SRC/technical-notes/abstracts/src-tn-1998-014.html

Smith, A.G. (2000). Search features of digital libraries. *Information Research, 5(3).* Retrieved December 17, 2005, from http://informationr.net/ir/5-3/paper73.html

Spink, A., Bateman, J., & Jansen, B.J. (1998). Searching heterogeneous collections on the Web: Behavior of Excite Users. *Information Research*, 4(2). Retrieved May 10, 2004, from http://informationr.net/ir/4-2/paper53.html

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Sutcliffe, A.G., Ennis, M., & Watkinson, S.J. (2000). Empirical studies of end-user information searching. *Journal of the American Society for Information Science,* 51(13), 1211-1231.

Tague-Sutcliffe, J.M. (1996). Some perspective on the evaluation of information retrieval systems. *Journal of the American Society for Information Science*, 47(1), 1-3.

Walkerdine, J., & Rodden, T. (2001). Sharing searches: Developing open support for collaborative searching. *Proceedings of INTERACT 2001: 8th TC13 IFIP International Conference on Human-Computer Interaction*. Retrieved July 2, 2005, from http://www.comp.lancs.ac.uk/computing/users/walkerdi/Papers/I2001PDF.pdf

Wallace, P. (1993). How do patrons search the online catalog when no one's looking? The transaction log analysis and implications for bibliographic instruction and system design. *Reference & User Services Quarterly*, 33(2), 239-252.

Warren, P. (2000). Why they still cannot use their catalogs. *Proceedings of the 2001 Information Science Conference*. Retrieved July 1, 2005, from http://www.upe.ac.za/citte2000/docs/pwarren.pdf

Wolfram, D., & Xie, H. (2001). Traditional IR for Web users: A context for general audience digital libraries. *Information Processing and Management*, 38(5), 627-648.

Yuan, W. J. (1997). End-user searching behavior in information retrieval: A longitudinal study. *Journal of the American Society for Information Science*, 48(3), 218-234.

Vitae

Dion Goh has a PhD in Computer Science from Texas A&M University. He is an Associate Professor of the School of Communication and Information, Nanyang Technological University, Singapore. Besides teaching IT-related courses for the Master's program in Information Studies, he is also the Director of the Master's program in Information Systems. His research interests are in information retrieval, digital library applications, the use of information technology in education, and the application of Internet technologies to support these fields.

Eng Pwey Lau is a graduate of the Master's in Knowledge Management program from Nanyang Technological University, Singapore. She is an IT professional in Singapore Airlines where she manages the company's information systems.

Tables

Table 1. Type of search options used, their frequencies and failure rates.

(a) Search Option	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
Author	52 326	8.2000	26 619	4.1463	25 707	4.0043
Keyword	442 523	68.9000	219 634	34.2114	222 889	34.7184
Subject	41 048	6.4000	28 632	4.4599	12 416	1.9340
Title	106 094	16.5000	42 955	6.6909	63 139	9.8349
Total	**641 991**	**100**	**317 840**	**49.5085**	**324 151**	**50.4915**

Table 2. Query lengths, their frequencies and failure rates.

(a) Query Length	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
0	35	0.0055	35	0.0055	0	0
1	151 462	23.5925	49 410	7.6964	102 052	15.8961
2	218 609	34.0517	92 273	14.3729	126 336	19.6788
3	110 863	17.2686	67 299	10.4829	43 564	6.7857
4	60 112	9.3634	39 605	6.1691	20 507	3.1943
5	40 714	6.3418	27 190	4.2353	13 524	2.1065
6	21 691	3.3787	15 212	2.3695	6 479	1.0092
7	13 121	2.0438	8 929	1.3908	4 192	0.6530
8	9 072	1.4131	5 368	0.8362	3 704	0.5770
9	4 426	0.6894	3 283	0.5114	1 143	0.1780
10	4 742	0.7386	3 040	0.4735	1 702	0.2651
11	3 003	0.4678	2 455	0.3824	548	0.0854
12	1 650	0.2570	1 483	0.2310	167	0.0260
13	911	0.1419	803	0.1251	108	0.0168
14	481	0.0749	435	0.0678	46	0.0071
15	301	0.0469	276	0.0430	25	0.0039

Table 2 continues.

Table 2 continued.

(a) Query Length	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
16	203	0.0316	182	0.0284	21	0.0032
17	102	0.0159	91	0.0142	11	0.0017
18	91	0.0142	85	0.0132	6	0.0010
19	72	0.0112	71	0.0111	1	0.0002
20	50	0.0078	49	0.0076	1	0.0002
21	48	0.0075	48	0.0075	0	0
22	28	0.0044	26	0.0041	2	0.0005
23	33	0.0051	22	0.0034	11	0.0017
24	28	0.0044	28	0.0044	0	0
25	28	0.0044	27	0.0042	1	0.0002
26	22	0.0034	22	0.0034	0	0
27	12	0.0019	12	0.0019	0	0
28	15	0.0023	15	0.0023	0	0
29	15	0.0023	15	0.0023	0	0
30	12	0.0019	12	0.0019	0	0
31	6	0.0009	6	0.0009	0	0

Table 2 continues.

Table 2 continued.

(a) Query Length	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
32	6	0.0009	6	0.0009	0	0
33	11	0.0017	11	0.0017	0	0
34	5	0.0008	5	0.0008	0	0
35	3	0.0001	3	0.0005	0	0
36	1	0.0002	1	0.0002	0	0
37	1	0.0002	1	0.0002	0	0
39	1	0.0002	1	0.0002	0	0
43	1	0.0002	1	0.0002	0	0
46	4	0.0006	4	0.0006	0	0
Total	**641 991**	**100.0**	**317 840**	**49.5085**	**324 151**	**50.4915**

Table 3. Examples of long queries.

Query	Query Length
The 8088 and 8086 microprocessors : $ b programming, interfacing, software, hardware, and applications : including the 80286, 80386, 80486, and Pentium processors	23
Risk evaluation and management of build - operate - transfer (BOT) projects / B.K. Lee and R.L.K. Tiong -- p. 711-718 [in] 3rd National Undergraduate Research Programme Congress 97.	29

Table 4. Frequency with which Boolean operators are used and their failure rates.

(a) Boolean Operators Used	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
AND	66 053	10.2888	38 608	6.0138	27 445	4.2750
OR	487	0.0759	161	0.0251	326	0.0508
NOT	84	0.0131	40	0.0062	44	0.0069
Total Single Boolean Operator (1)	66 624	10.3777	38 809	6.0451	27 815	4.3326
OR, NOT	2	0.0003	2	0.0003	0	0
AND, OR	8 809	1.3721	4 136	0.6442	4 673	0.7279
AND, NOT	18	0.0028	11	0.0017	7	0.0011
AND, NOT, OR	25	0.0039	3	0.0005	22	0.0034
Total Combined Boolean Operators (2)	8 854	1.3791	4 152	0.6467	4 702	0.7324

Table 4 continues.

Table 4 continued.

(a) Boolean Operators Used	(b) No. of Queries	(c) Percent (%)	(d) Zero Hits		(e) Non-zero Hits	
			No. of Queries	Percent (%)	No. of Queries	Percent (%)
Total using Boolean Operators (1) + (2)	75 478	11.7569	42 961	6.691838	32 517	5.0650
No operators (3)	566 513	88.2431	274 879	42.8188	291 634	45.4265
Grand Total (1) + (2) + (3)	**641 996**	**100**	**317 840**	**49.5085**	**324 151**	**50.4915**

Table 5. Examples of typographical errors.

Type of Error	Actual Query	Corrected Query
Substitutions (Misspelled term has an incorrect character)	data minimg Findamentals of engineering materials jack welsh	data mining Fundamentals of engineering materials jack welch
Omissions (Misspelled term has left one character out)	Enviromental Science gold min american standards for tesing of materials	Environmental Science gold mine american standards for testing of materials
Insertions (Misspelled term has an extra character)	Singapore developement policies and trends online obooking readinjg people	Singapore development policies and trends online booking reading people
Transpositions (Misspelled term with transposition of two adjacent characters)	Rice growing industry in Asai digital logic applications and deisngs	Rice growing industry in Asia digital logic applications and designs

Note. Misspelled and corrected terms are underlined.

Figures

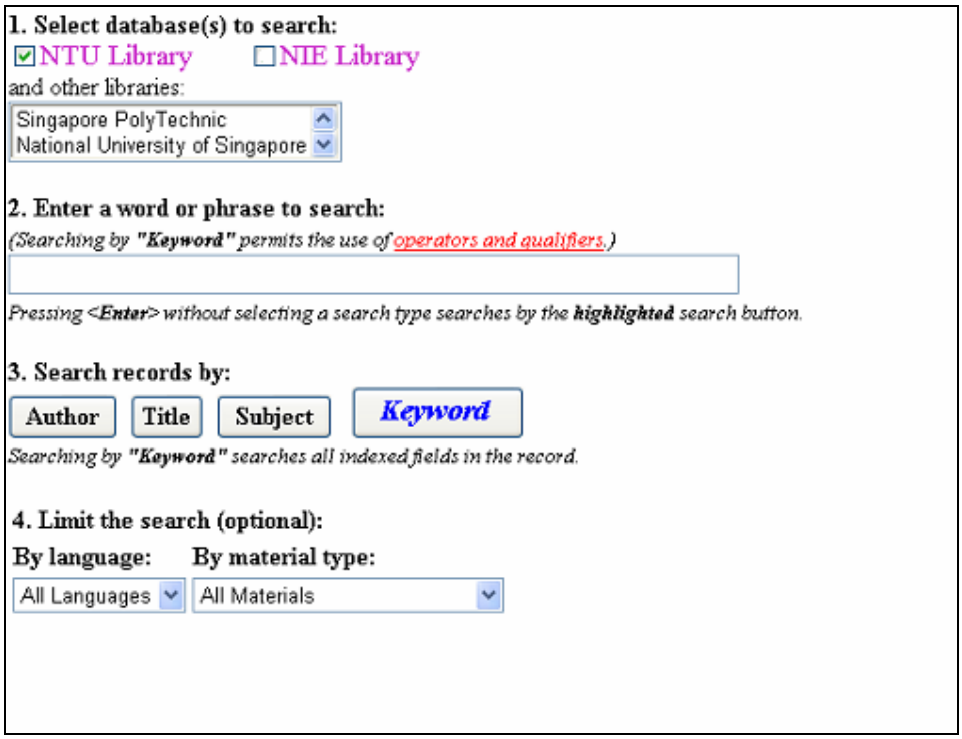

Figure 1. The NTU OPAC query interface.

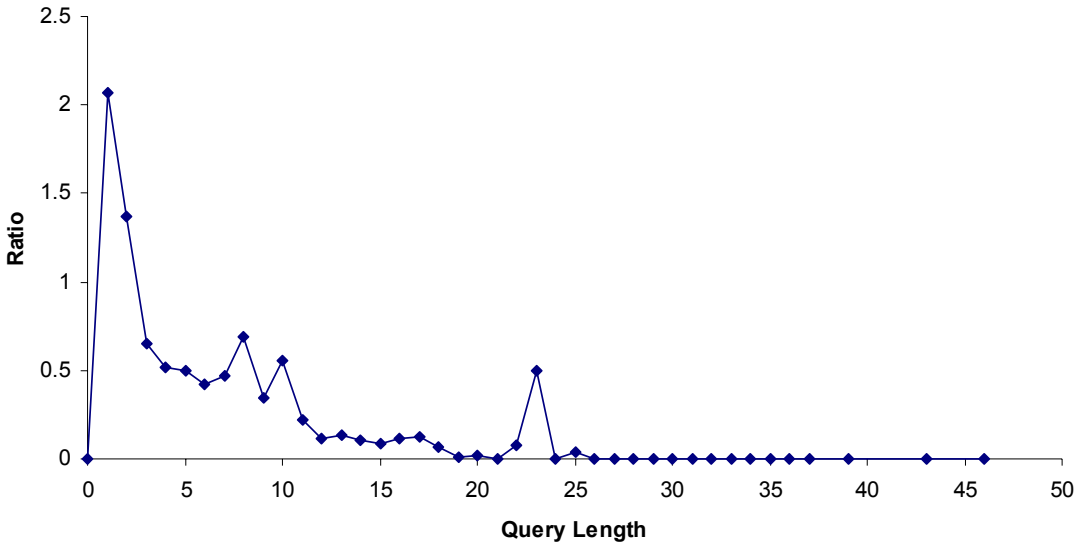

Figure 2. Increasing search failure as query length increases.