A Transaction Log Analysis of a Digital Library

STEVE JONES, SALLY JO CUNNINGHAM, RODGER MCNAB, STEFAN BODDIE

Department of Computer Science

University of Waikato, Hamilton, New Zealand

Telephone: +64 7 838 4021

Fax: +64 7 838 4155

E-mail: {stevej, sallyjo, rjmcnab, sjboddie}@cs.waikato.ac.nz

Abstract

As experimental digital library testbeds gain wider acceptance and develop significant user bases, it becomes important to investigate the ways in which users interact with the systems in practice. Transaction logs are one source of usage information, and the information on user behaviour can be culled from them both automatically (through calculation of summary statistics) and manually (by examining query strings for semantic clues on search motivations and searching strategy). We have conducted a transaction log analysis on user activity in the Computer Science Technical Reports Collection of the New Zealand Digital Library, and report insights gained and identify resulting search interface design issues. Specifically, we present the user demographics available with our library, discuss the use of operators and search options in queries, and examine patterns in query construction and refinement. We also describe common mistakes in searching, and examine the distribution of query terms appearing in the logs.

Keywords

transaction log analysis, search interface, usage analysis.

1 Introduction

There is extensive literature on transaction log analysis of OPACs (see [22] for an overview). However, only recently have these techniques been applied to digital libraries—likely because many digital libraries have only just attained a usage level suitable for log analysis [15, 16, 29]. Since log analysis provides insight into user search behaviour it is useful in the design and evaluation of query interfaces.

Transaction log analysis, as applied to OPACs, has yielded a diversity of results; it appears difficult to generalize about information seeking and search behaviors for all users at all times. Instead, the primary utility of these analysis techniques lies in the production of detailed descriptions of the behavior of a given group of users, on a single retrieval system, for a particular document collection. In this paper we have suggested ways that these fine-grained details can then be used to tailor our system to its target user group. We perform both quantitative and qualitative analyses on transaction logs spanning over a year's use of the largest collection in the New Zealand Digital Library [33]: the 46,000+ collection of computer science technical reports, intended as a resource to support research in computing. Our analysis is significant both for the relatively large span of time studied, and for the relatively focussed nature of the collection; other transaction log analyses have generally captured a much shorter time period (one day, in the case of the Excite studies of Jansen et al [15, 16] and Spink et al [29]), and have not usually been based on a subject-specific set of documents. The computer science research community could be thought of as "best case" users of online search engines, given their familiarity with software and Boolean logic. However, this study indicates that these users experience many of the same difficulties in searching and dealing with query languages that are reported for the general public.

This paper is organised as follows: first we describe the New Zealand Digital library and its collections, focussing on the Computer Science Technical Reports collection. In Section 3 we describe how the data has been collected, and some demographic details of the users are presented. The usage logs are automatically processed by software which extracts specified summary statistics, and it is this data that we analyze in Section 4. A manual analysis of the logs is presented in Section 5. In Section 6 we consider how our observations fit with those of similar studies. Finally, we summarise our observations, and propose directions for future investigations that have been suggested by our study.

2 The New Zealand Digital Library

The New Zealand Digital Library (NZDL) is a publicly-accessible internet-based digital library system that has been active since 1995 (http://www.nzdl.org). At the heart of the library are *collections*—logical groupings of documents. There are currently more than 20 collections, although some of these are private and not accessible to the public. Most collections contain textual documents for which searchable full-text indexes have been built. Examples are *The Oxford Text Archive* (a collection of literary works), *The Humanitarian Development Collection* (information resources for aid workers in developing countries) and a mirror of the *Human-Computer Interaction Bibliography* (over 15,000 bibliographic records of HCI related publications). Other collections deliver multimedia resources. For example, the *Local Oral History Collection* enables searches on the text of interview transcripts and presents electronically stored audio clips as query results. The *Melody Index* enables users to submit recorded music (perhaps a melody that they sing) which is then used to find a match in a database of more than 10,000 tunes. The *Historic New Zealand Newspaper Collection* is bilingual, with parallel Maori and English articles and delivery of scanned images of the newspapers as query results.

One aspect of our research is concerned with evaluating and improving upon the retrieval interfaces that we currently provide. To this end we are undertaking usability studies and analysis of transaction logs, and we report on the latter in this paper.

Figure 1 near here

We have focussed on analysis of the *Computer Science Technical Reports* collection (CSTR). The CSTR contains more than 46,000 publicly available computing-related technical reports harvested from over 300 research institutions from around the world—a substantial collection of "grey literature". Two principles of our digital library architecture are to make a minimum of assumptions about conventions adopted by document repositories, and to avoid manual document processing. Since the CSTR collection is based on a large, diverse set of document repositories, we cannot rely on the presence of bibliographic metadata. The collection is not formally catalogued; however, the full texts of the documents are extracted and indexed. The primary access mechanism for the collection is thus an unfielded keyword search. Both ranked and Boolean querying are supported, and several query options can be specified. Figure 1a shows the *simple query* interface to the NZDL. This is usually the first screen that users see when they visit the library. Figure 1b shows the *advanced query* interface for the CSTR.

The simple query interface presents users with a list of collections that can be searched, and the radio buttons can be used to select a particular collection. An entry box allows the user to type a list of words or phrases to form a query, and a search is activated by clicking the “Quick Search” button. Queries entered at the simple query interface are processed as ranked queries.

An advanced query interface is provided for users who require more control over query options. Again there is an entry area for query terms but the user can additionally select whether the query type is ranked or Boolean. A ranked query contains no operators and returns a list of documents that are ordered according to their similarity to the query as calculated by the retrieval engine. The documents are displayed in order of decreasing similarity. The retrieval system used by the NZDL is MG [34]. A Boolean query can contain operators & (intersection), | (union) and ! (negation) and parentheses can be used to control operator precedence in query evaluation. Query terms can be grouped into phrases using quotation marks.

Figure 2 near here

Other query options control: stemming (whether word stems or whole words must match); case-sensitivity (whether upper/lower case distinctions in query terms must match); and the proximity of query terms within result documents (within the same report, same page or first page of the report). The maximum number of documents to return (options of 50, 100, 200 and 500) and the number of result documents to display on each page (options of 10, 20, 50, 100 and all) can also be specified. A standard results page can be seen in Figure 2. It contains an entry area for a further query in which the current query is displayed, and a description of the options of the current query. An extract of text from the start of each result document is shown—although because they are originally extracted from Postscript the descriptive quality of these can vary. Links associated with each extract enable the user to download the Postscript from the original source, or view a locally held version of the extracted plain text, a facsimile of the front page of the document, Postscript figures extracted from the document, or information describing where it came from and when it was downloaded.

3 Data Collection

All user activity within the NZDL is automatically logged, and although actions can be associated with particular user identifiers, users themselves remain anonymous. The data that we consider here was collected in an 61 week period from April 1996 to July 1997. More than 30000 queries were recorded and analysed for the period in question.

User activities are timestamped and include: query text, query options, documents viewed and the size of result sets. Query options include type (Boolean or ranked), stemming, case sensitivity, term proximity (within the same report, same page or first page), the maximum number of documents to return and the number of returned documents to display on each page of results. The log records the number of resulting documents that the user chooses to view for each query, as well as the location of those documents in the result list. Accesses by local users are not included in this analysis, because many are undertaken for testing purposes rather than real information seeking tasks.

3.1 User demographics

Since users of the CSTR do not register for this database, the only information held on an individual's use is the IP address of the machine through which the collection was accessed. While this prevents us from incorporating detailed user demographics into the transaction log analysis, the design decision has had two practical advantages: users can immediately begin searching without spending time registering or verifying their account (an important consideration, given that this user group appears to prefer brief interactions with search systems); and anonymous access assures users of their privacy, so that user interest profiles specific to given individuals cannot be developed (again, a matter of concern for users of digital libraries [25]).

Table 1 near here

Examination of the search access by domain code (Table 1) indicates that the heaviest use of the collection comes from North America, Europe (particularly Germany and Finland), as well as the local New Zealand community and nearby Australia. As expected for such a collection, a large proportion of users are from educational (.edu) institutions; surprisingly, however, a similar number of queries come from commercial (.com) organizations, perhaps indicating that the documents are seeing use in commercial research and development units.

4 Analysis of summary statistics

The raw data from the transaction logs is automatically processed and collated into tables of summary data. In this section we discuss a selection of this data.

4.1 User Acceptance of Default Settings

Table 2 near here

The logs reveal that users rarely amend default settings for either query or result display options. During the period for which data was collected the default query type was changed. For the first 46 weeks the default query type was Boolean, and for the next 15 weeks the default query type was ranked. For both of these periods virtually identical proportions of queries (two out of three) used the default setting (Table 2). This is consistent regardless of whether the setting was Boolean or ranked. Although the total figures show that there was a higher proportion of Boolean queries (58%) over the full period this is somewhat misleading because of the shorter period in which ranked queries were the default. We expect the total figures to tend towards the pattern of more than two out of three queries using the default query type.

Other query options were also unlikely to be changed. One in five queries (21%) changed the default term proximity setting (terms must appear in the same report). Default settings for case-sensitivity and stemming were changed even less frequently—in only one in twenty queries (5% and 6% of queries respectively). The default result set size (50 documents) was changed in only 10.5% of user queries.

There are two possible interpretations of these observations. First, the default settings may be appropriate to the requirements of the majority of users. However this hypothesis is confounded by the fact that users tend to accept the default query type even though this default varied over the observation period. The second interpretation, that users tend to accept whatever defaults are set is, we believe, more likely. Consequently care must be taken to ensure the efficacy of those settings. Given the reluctance of searchers to use Boolean operators and the relatively small number of terms appearing in most queries (see section 4.2), we have settled on ranked querying as a default. Firstly, ranked queries are simpler to form, and the presence of the occasional extraneous Boolean operator in a ranked query often does not materially affect the result list (we also automatically detect and flag this situation as an error). Additionally, the ranking technique returns documents only partially matching the query, which often provides a richer set of hits than the full-match required by Boolean searching—and thus provide greater return for the short, simple queries preferred by users. Similarly, by setting query term stemming and case insensitivity as the defaults, the system can partially compensate for brief queries through a defacto query expansion.

It is less clear what setting should be used as a default for term proximity. The CSTR interface supports three levels of proximity: query terms must appear within the same document, within the same page, or on the first page. The latter option is used mainly to force an approximation of title and author searching in the collection, as the documents are not formally catalogued, and restricting the search to the first page is likely to pick up this sort of information. Currently, we set the proximity default at the whole document level, again to return as large a set of hits as possible. In practice, it is unclear whether this setting returns too many false drops (irrelevant documents returned by the search engine); an additional user study is needed to confirm this default setting.

4.2 Query Complexity

Table 3 near here

Queries tend to be short and simple—over all queries the average number of search terms in a query is 2.43. There is a small difference between the average number of terms in ranked queries (2.23) and Boolean queries (2.57). Approximately 80% of queries contained one, two or three terms (see Table 3), and 98% of both ranked and Boolean queries contained six or fewer terms.

Table 4 near here

Boolean queries were also simple in terms of the operators used within them. Table 4 shows the frequency of operators in the 19026 of queries where the query type was Boolean. By far the majority of these queries (two out of three) use no Boolean operators at all. Union and negation operators are rarely used—fewer than one in 40 Boolean queries contained them. The most commonly used operator (intersection) occurs in only one quarter of all Boolean queries. When users explicitly specified Boolean query type (the period of default ranked queries) just over half of the queries contained some form of Boolean operator. Again, intersection was by far the most prevalent, but here occurred in just over two out of five of those queries. Fewer than one in ten contained each of union or negation operators, or were compound expressions. Although Boolean query type was explicitly selected over 40 percent of the queries still contained no Boolean operators. When Boolean query type was the default, it is understandable that so many (more than two thirds) of those queries did not contain Boolean operators. Users simply typed query terms into the simple query window.

We might expect the target users of the CSTR (computing researchers) to be conversant with Boolean logic, yet they appear unwilling to apply it when searching. One explanation for this observation is that Boolean logic is ill-suited to specifying queries for information retrieval. Another is that the Boolean query language provided is too complex or restrictive to allow users to effectively specify queries. The literature suggests that difficulties with textual Boolean query languages are common [2, 3, 12]. Users must remember the appropriate symbols or keywords for the Boolean operators. There is a conflict between the inclusive AND of the English language and the exclusive AND of Boolean logic, and also between the exclusivity of OR in English compared to its inclusivity in Boolean logic. Shneiderman [27, p542] offers a good example. The query "List all employees who live in New York and Boston" would normally result in an empty list—"and" is an intersection and only people who lived in both cities would qualify. In English "and" normally expands a set, and so the result would be people who live in either or both of the cities. In English one expects the statement "I'll have Coke or Pepsi" to result in the return of only one type of drink—"or" is exclusive. In Boolean logic OR is inclusive and could expand the result to both types of drink.

Also, textual Boolean query languages use a wide and inconsistent variety of representations for the operators. All of these issues lead users to produce erroneous queries or avoid Boolean expression if at all possible.

However, the use of Boolean expressions can support expressive and powerful querying. There is evidence to suggest that other presentations of the Boolean query model can be effective [8, 13]. For this reason we are investigating alternative interface metaphors for Boolean querying [17, 18].

4.3 Query Terms

Table 5 near here

There were 4,993 unique terms in the query logs, excluding terms appearing in quoted phrases. In Table 5, we present the 50 most commonly occurring search terms. Note that seven of these are stopwords (and, of, or, for, a, in, the). Examination of the query strings gives the impression that in many cases the searcher intended to perform a phrase search but omitted the quotation marks (for example, the searches *overview & of & Fiber & distributed & data & interface*; *A Study of Integrated Prefetching and Caching Strategies*).   The presence of the terms "and" and "or" appears to be either the mistaken use of those words for the MG Boolean operators (& and |, respectively), or to stem from the omission of quotation marks around a phrase.

The 50 most common query terms account for, in total, 5585 of the 32800 terms appearing in the transaction logs—approximately 17%. Other studies of web search engines [15, 16, 29] or digital libraries [5] have noted similar clusters of common search terms, so it is unlikely that this is an artefact of our relatively focussed target user group.

If we omit the stopwords from this list, then the remaining 43 content-bearing terms occur 4534 times (approximately 13.8%). Many of the individual terms appeared as portions of a phrase in the queries (for example, neural network/networks, information retrieval, case based reasoning). Since these search terms also appear relatively frequently in isolation, it is likely that using them without specifying that a phrase search be conducted is leading to a large number of false drops. This is additional evidence that phrase searching is an important tool for obtaining search precision for our library users; we must improve the ability of our search engine to efficiently process phrases. Additionally, as noted above, we must support the users in constructing proper phrases (in our system, delimited by quotes) rather than simply and-ing terms together. A prototype interface based on the Kea keyphrase extraction system [11] may prove useful in that task [21].

The proportion of common terms is significant; it appears that we may achieve a substantial performance improvement if we simply cache the results for those terms. The cached results would not take the form of full documents, rather the document surrogates (brief summaries) that are displayed in result pages as shown in Figure 2. Given that most viewed documents appear within the first 25 items in a result list (see Section 4.8) we might choose to cache 25 surrogates for each of the 50 most common query terms, excluding any stopwords. Even if each list of 25 surrogates was unique, storage requirements for 1250 surrogates would not be prohibitive. We estimate 373Kb for simple surrogates, and 1.8Mb for simple surrogates plus abstracts as maximum required cache sizes for 25 documents per term.

This approach would be useful in handling single-term queries, but less so for queries containing multiple terms. When a query contains more than one term, the ranking of results is dependent upon all the terms that it contains and is normally determined by the retrieval engine. In the case of Boolean queries, a small set of cached results can not produce an exhaustive result list for multiple-term queries.

Knowledge of the most frequent query terms can further be exploited in structuring term indexes for the collection. Indexes can be reordered to provide more rapid response for common terms.

Thematic coding of the query terms into categories corresponding to subject classifications has not yet been attempted, but would be an interesting exercise. This type of analysis would allow us to track specific research sub-fields that are high/low users of the digital library, so that we could identify user groups that are/are not having their needs met by the collection.

4.4 Term Specificity

Table 6 near here

We considered the question of whether our users were likely to be using overly general or overly precise query terms. We matched each individual query term to the document collection, discarding phrases and stripping the terms from the context of their query. Table 6 shows these results. Note, however, that the size of the CSTR collection has grown steadily over the trial period; we have therefore matched these query terms to the present contents of the collection, as we were unable to calculate exact query/document match information for intermediate states of the collection. The data in Table 6, then, should be considered as relative numbers of matches against the collection, rather than as absolute figures.

Approximately one in twenty query terms did not match any of the documents in the expanded collection. Inspection of these terms reveals that the majority consist of mis-spellings (repesentation, sofhware, algorythm), with a smattering of what appear to be personal or product names (gillingham, geobase) and a very few legitimate-appearing technical terms that simply do not appear in the collection (jsort, octrie, echocardiology). At this end of the spectrum, the problem appears to be spelling terms incorrectly, rather than that users have difficulty in matching their information needs to the collection terminology.

Instead, the difficulty for searchers may lie in selecting terms with the appropriate specificity. Over 40% of the query terms were matched by 1000 or more documents—and approximately 3% of the query terms were matched by over half of the collection! It seems likely that low search precision may be more problematic than low recall, particularly in light of the fact that users tend to submit queries containing relatively few terms (and hence have less opportunity to winnow down a large result set).

4.5 "Failed" and "erroneous" queries

Search failures have been variously defined in the retrieval literature—with the success or failure of a search being examined through retrieval effectiveness (precision/recall) measures, polls of user satisfaction, critical incident analysis, and transaction log analysis. The first three methods can give a finer-grained understanding of the causes of retrieval and searching failures, at the cost of also requiring far more contextual and demographic information from users [30]. While transaction log analysis can give no insight into user motivations and the semantic side of searching, its application requires no additional input or effort on the part of the users being studied.

We follow the practical definition of a failed search as one that matches no documents in the collection ("zero hits"), a definition which has been proposed in other studies [9, 14, 23, 34]. This definition has obvious drawbacks—for example, the user may be looking to confirm a hypothesis that no one else has studied a certain problem before, in which case zero hits would be regarded as a success. Such insight into user goals and criteria for determining whether or not searches have been successful can not be gleaned from transaction logs, and is more effectively gathered by alternative methods of analysis. Our definition of failure does not necessarily regard all non zero-hit searches as successes, but rather identifies the zero hits searches as extreme (and therefore potentially interesting) examples of unsuccessful searches.

For the 61 weeks studied in this transaction log analysis, a total of 1020 queries returned zero hits (10.53% of all queries). Of these failed queries, 105 had no query terms; 411 were ranked queries; and 504 were Boolean. These Boolean queries averaged 2.4 terms apiece—close to the overall average of 2.5. The ranked queries, however, averaged only 1.2 terms. Not surprisingly, the fewer the terms in a ranked query, the more likely that the query will fail. Moreover, 300 of the ranked queries contain a quoted phrase (approximately 73%); again, this confirms the intuition that a phrase is less likely to be matched than an unquoted set of terms in a ranked query.

Table 7 near here

Syntactically incorrect queries were rejected before being entered into the transaction logs, so we are unable to calculate the percentage of mal-formed queries. Earlier analyses of library online catalogs [23] and the Excite Web search engine [15, 16, 29] revealed significant difficulties in conforming to query syntax, particularly for the more demanding Boolean syntax. We have, however, automatically detected four types of probable errors in both ranked and Boolean query formulation (Table 7). Users receive a warning error when these potential problems are detected. A significant number of the processed queries—over a quarter—contain potential problems of one sort or another. Given that the obviously malformed queries had been rejected by the query preprocessor, it seems that users require strong support in creating correctly-formed queries, and that the instructions on syntax and information on search defaults should be more prominently displayed. It appears likely, for example, that users creating queries containing more than one phrase had not read through the help pages to note that this is not permitted by our search engine, or that users inserting capitalized terms into casefolded queries may not have realized that the casefolding default was in force.

4.6 User Sessions

Table 8 near here

Approximately a fifth (21.51%) of all user sessions were visits to the NZDL WWW pages which did not entail the submission of a query. Just over a half (51.68%) of all sessions included submission of only one or two queries. Slightly more than a fifth (21.69%) included submission of three, four, five or six queries, and 5.12% included seven or more. These figures are shown in Table 8. The average number of queries issued per session was 2.04. From these figures it appears that many users are prepared to expend little effort in the development of sequences of queries to focus in on their topic of interest. Given that few, short queries resulting in no documents being viewed are most common (see Table 14), we assume that a substantial portion of users end a session without having met their information seeking needs. We will need to carry out further investigation, such as observational analysis, to determine just why so many users seem to abandon their query process prematurely.

Figure 3 near here

The length of user sessions was also recorded. 29.16% of sessions lasted less that one minute (see Figure 3). We assume that in these instances, users are merely investigating the NZDL rather than intending to undertake some querying activity. Over half (54.34%) of user sessions have duration of five or fewer minutes, and two thirds (66.43%) have a duration of ten minutes or less. Some users have long sessions, which leads to an average session length of 10.83 minutes. The length of session and number of queries submitted might be dependent on the user interface and facilities provided by the NZDL, or it might be the case that users make rapid judgements about whether to persevere with the use of such an on-line retrieval system. If this is the case then we must provide for immediate, effective searching.

Table 9 near here

Taking each "session" in the logs as a "visit" to the digital library, we can determine the amount of repeat visitors to the digital library (Table 9). It is disappointing that approximately three-quarters of the users visited the digital library only once in the time period captured in the transaction logs. This is not unexpected, however, as over half of the visitors issued only one or zero queries (Table 8), and they were surely unlikely to return to the NZDL. Again, further investigation is needed to determine why the collection has relatively few repeat user: are they mistakenly visiting a collection that is inappropriate to their needs (a possibility discussed below)? Do the target users tend to have recurring information needs, or do they search at irregular intervals (a hypothesis indirectly supported by Pinelli [24] and Cunningham and Silipigni-Connaway [7])? Is the collection inadequate as a primary resource for computing researchers? Or are we seeing a manifestation of the 80/20 rule? Here, if we factor out the sessions in which no queries were issued, approximately 80% of the users visited the collection only once, with 20% visiting multiple times; this is similar to the proportions of single and multiple visitors that have been observed in studies of conventional, physical libraries.

We are unable to determine from the log data whether users who didn't return to the NZDL had their needs met by alternative on-line information resources, and if so, how those resources were different from the NZDL. Capturing further information about regular users is reasonably straightforward; they may be invited to participate in further studies and to respond to questions during their ongoing use of the system. Fleeting users are equally (perhaps more) interesting, but their very nature makes the elicitation of further information from them a great challenge. Although user registration might be of use here, as we noted in Section 3.1 we do not require users to register so that they might begin searching immediately, and to assure them of their privacy.

4.7 Query Refinement

Table 10 near here

Analysis of users' consecutive queries reveals interesting aspects of query refinement behaviour. A set of 6680 user sessions was analysed that contained a total of 13650 queries. The majority (66.37%) of queries issued by users have at least one term (a word or phrase) in common with the previous query (see Table 10). These figures discount the first query issued by a user. Most often, consecutive queries have one or two terms in common (22.56% and 23.08% respectively). A further 11.34% have three terms in common and 9.39% have four or more in common. This high incidence of term overlap implies that refinement is a common activity. Given that the average number of terms within a query is 2.5, and only a fifth of queries contain four or more terms, we believe that query refinement occurs in small incremental steps. Users are likely to make minor changes by adding a new term, or altering the existing terms.

Table 11 near here

We can look more closely at exactly how queries are refined. In addition to query terms, the logs record how the attributes of consecutive queries change. These include the type of query (Boolean or ranked), the granularity of the search (document level, page level and so on), and the use of stemming and case-sensitivity. Most commonly it is only the terms within the query which are altered. This occurs in 60.68% of cases. The remaining 39.32% of cases contain a variety of combinations of attribute refinement. These are shown in Table 11. We have made a distinction between a query string and the terms within a query. The query string represents the query terms exactly as entered by the user. The query terms are extracted from this string for processing. Two different query strings may contain the same terms, but in a different order. In fact, this was the only change in 5.66% of cases. This may be explained by users amending term ordering in the belief that it would affect the results returned by ranked queries.

As we have noted, users rarely change default settings. This is reflected in the frequency with which settings were changed between consecutive queries. In 3.66% of cases the query type was changed but all other aspects of the query, including the query string and terms, remained the same. We might expect the query string and terms to change because of the insertion or removal of Boolean operators. This perhaps reveals a lack of understanding on the part of users, or in all of these cases only a single query term was involved. This remains to be investigated. In 3.44% of cases a change of query type was accompanied only by a change in query string and terms, which is what we might expect if these are multiple term queries. In 2.72% of cases the search granularity was the only attribute which was changed. An insignificant number of cases involved changes to only the case-sensitivity or stemming.

Table 12 near here

Table 12 shows the percentage of cases in which each of the attributes changed between consecutive queries, including when they changed in conjunction with other attributes. It is worth noting that in 13.75% of cases no aspects of the query changed. That is, exactly the same query was successively submitted. We believe that this is due to the effects of response time. For complex queries, or at times of heavy server loading, the response time might have been such that the users were unsure if their query had been successfully submitted, and tried again.

Overall, although query refinement is a common activity, the nature of refinement is very basic. Users of the CSTR tend to focus on amending query terms rather than attributes of a query. It is possible that the user interface mechanisms for making such changes are not sufficiently evident or intuitive, and we shall investigate this through observational analysis of users. Few users consulted the on-line help documentation—just over 6% of user sessions contained accesses to help—which reinforces the notion that functionality must be as immediately and intuitively accessible as possible.

4.8 Result Viewing

Table 13 near here

In almost 90% of queries the default result set size of 50 documents was retained (see Table 13). Intermediate sizes of 100 and 200 were each requested in approximately 2.5% of queries, and a size of 500 was requested in almost 6% of queries. Again users seem content with default settings. However, we find a distinction when ranked and Boolean queries are considered separately. 95.6% of ranked queries, but only 77.4% of Boolean queries used the default setting. A substantial number of users require larger result sets to be returned when Boolean queries are used. With reflection, this seems sensible. Ranked queries imply that the most useful documents will be presented first, and consequently there may be little need to look past the first 50 resulting documents. With Boolean queries there is no ranking of the result set, and therefore users might retrieve and be more prepared to browse larger result sets to find interesting documents.

Table 14 near here

Disappointingly, the majority of queries (64.2%) do not lead to users viewing document content (see Table 14). Just over 19% of queries result in the viewing of one document, 12.7% result in the viewing of two, three or four documents, with around 4% resulting in the viewing of 5 or more. The distributions of the number of documents viewed for ranked and Boolean queries are very similar. The document summaries provided in query result lists appear to effectively support users in determining that they are *not* interested in particular documents. The queries that users form may be too simplistic to produce result lists which appropriately match their needs. However, when we compare the mean length of queries that led to no documents being viewed to that of those that led to at least one document being viewed we find virtually no difference. In both cases the mean length is approximately 2.5 terms—the mean for all queries. This indicates that query length is not necessarily a factor in whether users view any of the result documents. An alternative explanation is that the results returned may not be displayed at the appropriate granularity. For example, an uninteresting document title may hide the presence of a highly relevant subsection within the document. We are investigating the effects of passage level indexing and retrieval for this collection [32].

Figure 4 near here

When users do view documents they are most likely to view those which are at the start of the result list (see Figure 4). 12.7% of all viewed documents were located at the first position in the result list. The next most common location was the second position (6.8% of viewed documents). Nearly three- quarters (73.2%) of all documents viewed were in the first 25 positions in the list. The similar document viewing distribution between ranked and Boolean queries implies that the effect is not attributable to ranking of query results. Consequently, the presentation order of result sets lists must be carefully considered.

4. 9 Server Loading

Figure 5 near here

Figure 5 shows a representative two month extract of the logs from 30 September 1996 to 1 December 1996. The number of queries issued on each day in this period is shown. A pattern for access over each week can clearly be seen, and is repeated throughout the full logs. Each vertical bar is placed at a Monday (in New Zealand, the location of the NZDL server). The peaks and troughs of the graph correspond directly to weekdays and weekends. Although a reduction in usage might be expected throughout the weekend, access from North America on Friday (New Zealand Saturday) ensures that there is only one day per week when usage substantially drops. Such information can support planning of system maintenance and upgrading to cause minimum disruption to globally distributed users.

5 Manual analysis of transaction logs

To gain a finer grained appreciation of the types of searches that users conduct, the 30,000 queries were manually examined. While statistical analysis gives broad overviews of trends in usage, these summary tables cannot convey details of the semantic intent of the user queries. In this section, we discuss more qualitative information garnered from the transaction logs.

5.1 Spelling issues

Mis-spellings are relatively rare in the search terms; only approximately 240 of the searches contained incorrectly spelled words or typographical errors (although since the unusual number of acronyms in the computing field sometimes makes it difficult to determine whether a term is mis-spelled or mis-typed, this estimate should be regarded as a lower bound). The major problem detected with spelling was that few users took into account the differences between UK and American spellings when constructing their queries. While these differences may only cause minor losses in recall or precision for some disciplines, in computing the affected words are sometimes crucial parts of descriptive phrases: for example, “information seeking behavior/behaviour”, or “data visualization/visualisation”. Interestingly, some users appear to attempt to perform their own stemming, rather than setting the stemming option in the search interface (for example, searching for “chain & topology & algebrai & simpli”).

Another difficult issue with computing terms is that many product names, protocols, program names, and so forth contain special characters. The underlying search engine for the CSTR collection, like many retrieval engines, strips special characters and retains only alphanumerics. While this situation is not problematic in many disciplines, for computing it means that, for example, it is difficult to locate documents about the language “C++” as distinct from the language “C”. Additionally, it is unclear to users how they should represent strings containing special characters: for example, should they type “modula-3”, “modula_3”, or “modula 3”? Since most users do not read the system documentation, they must discover by trial and error that “modula-3” and “modula 3” map to the same internal query string.

5.2 Sub-collection choice

The New Zealand Digital Library architecture is designed as a collection of collections. Rather than a single, homogenous digital library covering all subjects, it is instead seen by the users as a set of sub-collections each focussing on a different subject area. At the startup query page, the user must select the sub-collection to search as s/he enters the initial query. The transaction log was examined for indications that users were inappropriately searching the CSTR collection—that is, directing a query to the CSTR when another sub-collection would have been more suited to filling the user's apparent information need. Since computing is very much an applied field, it is difficult to categorically state that a given query is not related to computer science; for example, the query "berrypicking" may refer to a particular model of the interaction between a user and an information retrieval system, and "snake" appears as a technical term in a surprising number of theoretical computing and computer vision documents. However, upon examination of the logs we noted 149 queries that appeared highly unlikely to be pertinent to the CSTR collection. These queries fell into three categories: the searcher (not unnaturally) appeared to believe that a service with the title "New Zealand Digital Library" would contain general information about New Zealand ("kiwi bird", "1080 poison" [a possum poison very much in the news in NZ at that time]); the user believed that the digital library was a general search engine ("Anarcist [sic] cook book", "the civil war of america", "gay marriage"); or the search seemed to be aimed at retrieving documents held in the Gutenberg or Oxford Text Archive collections ("Animal Farm by George Orwell", "I Know Why The Caged Bird Sings", "social satire");

Evidence for the first two cases is supported by noting a similar set of off-target queries having been posed to the CSTR reference librarian, who fielded requests for help from patrons having difficulty locating information in the CSTR collection [6]. The latter problem—users selecting the wrong sub-collection to search—appears to be due at least in part to the fact that the CSTR collection is pre-selected as the default in the radio button list of sub-collections on the initial query screen. Additionally, if the user pages back to the initial query screen after performing a query, then the default is automatically reset to the CSTR—and the user must notice this and change the target collection again. Currently, the CSTR is the largest sub-collection in the NZDL, and consequently receives the lion's share of usage. As the other collections grow, this problem in locating a relevant collection may be expected to grow as well, necessitating a re-design of the initial system page to direct users to appropriate document sets.

A further problem may be that users simply do not understand the differences between the documents covered in each of the sub-collections. In addition to the 149 queries that were almost certainly not applicable to the CSTR, others were noted that seemed more appropriately directed at the NZDL's two sub-collections that include popular computing topics rather than to the more strictly academically focussed CSTR ("Microsoft Access 7.0", "pentium processor"). Again, this misunderstanding of the CSTR focus is supported by reference help requests for information on current popular computing topics (Cunningham, 1998). This problem indicates a need to include more information on sub-collection focus in the initial system page, rather than storing these details in subsidiary information pages (as is currently the case).

5.3 Additional search strategies

As noted in the introduction, the CSTR collection is uncatalogued; users are limited to keyword searches. The system documentation suggests work-arounds for approximating some types of fielded searches (for example, limiting a search on an author's name to the first page, as most technical reports list the authors there). As noted in the previous section, few users consult the documentation or use the "first page only" option. However, examination of the logs reveals a significant number of queries that appear to be attempting to search on what would, in a formally catalogued system, be fielded document access points: author names, full document titles, technical report numbers, date of publication, author contact details (institute, email address), and journal or proceedings title. Users searching under publication details (such as journal or proceedings name) appear not to realize that the CSTR contains unpublished technical reports, and that these searches would be better directed to a different sub-collection containing a bibliography of published works.

We have recently introduced emulation of fielded searching into some collections of the NZDL. For example, entries in the Human Computer Interaction Bibliography collection contain bibliographic information about published research papers. When results of queries to this collection are returned, elements of each document description such as author names and journal/proceedings titles are presented as links. When the user clicks on a link the name or title is submitted as a query, simulating a fielded search. In another collection, the Computer Science Bibliographies, we redirect users to the CSTR where they may uncover documents relating to the bibliographic entries that they have found. Each result item returned by the Computer Science Bibliography collection has an associated link, which when selected submits the document title and abstract as query to the CSTR to reveal related documents. Ideally, for the CSTR, we might automatically determine queries in which users are attempting to carry out fielded searching, and redirect them to the appropriate bibliographic collection.

However, it is difficult to identify this type of search. Only an approximate measure of the number of appearances of them can be taken; for example, it can be difficult to distinguish an uncapitalized author's name from a lowercase acronym, and some searches are undoubtedly intended as keyword searches for mention of a technique ("texture Fourier" is most likely intended to retrieve documents on the use of Fourier transforms in recognizing/rendering textures, rather than papers by Fourier). Given these caveats, roughly 17% of searches appear to include a name in the search string—a significant minority. Informal discussions with local (New Zealand) users indicate that some searches are indeed intended to retrieve documents where the referenced individual is an author, while other searches take advantage of the fact that the entirety of the document is indexed and are attempting to locate matches in the reference sections—thereby retrieving documents that cite that author. In either case, the simple keyword search approach appears in some cases to be insufficiently precise, as evidenced by successive queries presenting the name in different formats (apparently in an attempt to guess the "correct" form in the collection index).

A handful of queries appeared to recognize that although by far the majority of documents in the CSTR are in English, the collection also contains technical reports in other languages. These user sessions included queries in German, sometimes with German translations of queries following the English terms ("heterogenous databases", "heterogene datenbanken"). Our current strategy for dealing with a multi-language document collections includes a multilingual interface (with help screens and query construction pages available in five languages); however, the NZDL does not currently support language-specific stemming over more than one language per collection, and does not permit the restriction of queries to a single language (primarily because the CSTR documents are not tagged by language). These issues remain to be incorporated into our digital library architecture.

5.4 Graffitti

Various examinations of public access retrieval systems have noted small numbers of "graffiti" queries—terms, sentences, or even paragraphs of obscene or irrelevant entries [10]. When scanning the lists of query terms, we noted a few examples in the CSTR transaction logs. Some graffiti, particularly the bursts of expletives, may have been entered to vent frustration at the system or at search results. Or perhaps some users were curious as to whether various naughty words appeared in the collection? Other queries seem examples of playfulness; a particular favorite of the authors' is the query, "why statisticians fart".

5.5 Boolean queries

As noted above, the query preprocessor catches syntactically incorrect Boolean queries before they are entered into the transaction logs. However, a manual examination of Boolean query strings revealed a significant number of Boolean queries that—from a semantic point of view—were problematic. Specifically, we noted:

- *improper use of and, or, and not:* Users substituted the terms *and, or*, and *not* for the Boolean operators required by the MG search engine (&, |, and !, respectively). In most cases this would not appreciably affect search results, since the most commonly used term is "and", and in the absence of correct Boolean operators the default is to "and" all terms together (so, in effect, the user's query has the word "and" and-ed to the search string). When we re-examined the instructions appearing on the "advanced search" page, the description of the correct forms for the Boolean operators appears ambiguous; this likely accounts for some of these errors. Additionally, users may simply overlook the instructions and use the more standard terms.
- *use of syntax from other search engines*: Some queries included syntax or operators that were applicable to other common retrieval systems, but not to the NZDL. For example, some users appear to have added a "+" symbols

before terms to require that those terms appear in the documents retrieved or a "-" to negate the term, as is permitted by several web search engines (such as Excite). Other users apparently had significant experience with library online catalogs, which usually require that author names be inverted when querying (eg, "Bell, T"). Library catalog systems have addressed user difficulties in the face of incompatible query syntax by settling on a standard (Z39.50); perhaps the digital library search engines will also converge on a common syntax.

- *natural language queries:* Some queries are obviously constructed as natural language descriptions of information needs, and should have been of ranked rather than Boolean type (*How Pascal uses records*, *why c++ has no automatic garbage collection*).
- *extraneous parentheses and quotes:* Parentheses are intended for controlling operator precedence. Some queries contain redundant parentheses that do not contribute to the logic of the query (for example, the query *(Distributed) & (Transaction) & (database)*). Other users appear to confuse the function of parentheses and quotes, and use parentheses when forming phrases. This mistake can have a significant impact on query results: for example, the query *Bogoni & (Discrete Event)* is equivalent to *Bogoni & Discrete & Event*—and the two words "discrete" and "event" are likely to generate a larger number of hits than the phrase "discrete event" in collection of computing documents. Still other queries contain single term phrases (eg, *"Monica" & "cluster"*)—the inclusion of which may increase processing time considerably.

These semantic irregularities in Boolean query strings provide additional supporting evidence for the hypothesis that users of the CSTR collection experience difficulties in correctly translating their information needs to Boolean form. We are currently investigating graphical interfaces that may address these problems in Boolean query construction [17].

6 Comparison with other studies

A substantial number of studies have investigated use of electronic library catalogues (see [22] for a literature review). Others have investigated use of

World Wide Web (WWW) based search engines [15, 16, 29]. There are similarities and differences between these contexts and that of a digital library. In each of these contexts the mechanisms by which users search for information are similar—keyword based queries using ranked or Boolean retrieval. Although the syntax and range of query operators available may vary, as may the options which control query processing and result display, there is a high degree of commonality between the basic functions of the query interfaces. However, some systems support fielded searches and this is not the case in CSTR. The context that we have studied differs from WWW search engines and library catalogues in that our CSTR collection is targeted at a relatively narrow field of study (computer science) and a particular level of use (primarily for research). The majority of the earlier studies have concentrated on collections aimed at the general public.

Table 15 near here

Few studies have looked such a large number of interactions with a retrieval system over such a substantial period of time as we have considered here. The most closely related work [15, 16, 29] looked at 51,453 transactions from a single day's use of the Excite internet search engine (http://www.excite.com), and it is to this study that we most closely compare our findings. The Excite interface studied by Jansen et al is similar to the CSTR in that the default query type is ranked and there is a particular syntax for specification of Boolean operators. Unlike the CSTR, the maximum number of terms that Excite users can enter is restricted to ten (although a thesaurus is used to supplement the user's query with additional terms). Results are displayed ten per page and include a short summary of each result document. Table 15 shows findings from the two studies where direct comparisons are possible.

There is a noticeable level of similarity between the results from the two studies. Both find that queries tend to be short, agreeing that approximately four out of five queries contain fewer than four terms (~86% for Excite against ~82% for the CSTR). A higher proportion of Excite queries contained zero terms (~5%) than for the CSTR (~1.5%). This is perhaps explicable by the fact that Excite contains a "More like this" button, the use of which is recorded as a zero term query. For both systems the average query length was found to be between two and three terms (2.21 for Excite, 2.5 for CSTR). A higher proportion of Excite queries contained zero or one terms (36.83%) than in the CSTR (28.29%), but for all other frequencies the proportions were slightly higher for the CSTR. However, the term frequency distributions for the two studies are strikingly similar.

Both studies also agree that users tend to enter between two and three queries in a session, although the average for Excite was higher than for the CSTR (2.8 queries against 2.04 queries). The biggest disparities between the systems are at zero queries per session (26.6% for Excite against 21.5% for CSTR) and at one query per session (56% for Excite against 34.5% for CSTR). Higher proportions of CSTR sessions were observed to have all other query frequencies than for Excite, although these proportions are small and do not impact substantially on the overall average number of queries per session.

In the Excite study 61% of users viewed only the first 10 results for a query, with a further 19.73% viewing 20 results. We are unable to do a direct comparison because result list size and segmentation are under user control in the CSTR. However, a good comparative indicator is the position of documents that CSTR users viewed the details of. 49.7% of such documents were within the first 10 results, and a further 17.8% were within the next ten documents.

The Excite study reports that just over a third of queries are repetitious (37.73%), whereas for the CSTR this figure is just over a quarter (26.98%). Here we have taken repetition to mean that the terms within consecutive queries were the same, although they might have a different ordering. Although at first look, repetitious queries seem to be much more likely for Excite, some of this difference may be explained that viewing a further page of results in Excite causes a query repetition. We are unable to factor out this effect from the reported Excite results to do a more direct comparison. The percentage of unique queries (those that were not refinements of previous queries) are very similar for both studies—just over a third of all queries.

The main observed difference between the two studies is in the use of Boolean operators in queries. A higher proportion of queries in the CSTR contained Boolean operators than those in Excite. This is true for individual operators—just over twice as many for intersection, five and a half times as many for union, almost three times as many for negation and five times as many for parentheses. This is likely attributable to the fact that the target audience of the CSTR is computer scientist who we might consider more likely to be knowledgeable in Boolean logic than the general search engine user. However, the overall frequencies for Boolean operator use are still low.

Most research on transaction log analysis has concentrated on library online public access catalog systems (OPACS). This work is not directly comparable with our study, as OPACS support searching in fielded collections of bibliographic records, whereas we are examining searching patterns on a collection of unfielded full documents. Differences between results from transaction log analysis of OPACs and of digital libraries may illuminate the impact that fielding and full text has on searching—particularly if the systems being compared target similar user groups.

The most significant commonality between results from OPAC studies and our work is that OPAC users also appear, in general, to perform relatively simple queries—both Sandore [26] and Peters [22] reference studies reporting an average query length of between one and two terms, whereas CSTR users average 2.5 terms per query. Studies reporting significantly longer average queries ([28] for example) tend to be based on smaller sample sizes or on mediated searches . The reported average length of online or OPAC search sessions varies widely: for example, MELVYL OPAC users averaged approximately seven minutes per session [20], while MEDLINE sessions lasted thirty minutes [4]. This range may reflect different system response times, or differences in user needs; in either case, these results cannot be directly compared to session times for the WWW-based CSTR. Unfortunately, most transaction log studies have reported time lengths on searches, rather than a more easily compared measure such as the average number of queries per session.

Users searching public libraries or general purpose collections tend to examine only the first few returned citations (for example, MELVYL catalog searchers displayed on average only the first 15 citations returned by a query [20]). A study of University of Colorado OPAC use reports a much higher persistence level among their users, possibly because tertiary students and researchers require a higher level of recall [31]. CSTR search result viewing patterns appear to bear a closer resemblance to the UC results than to those reported for general collections, as might be expected; however, differences in measurement techniques and in system interfaces make exact comparison difficult.

OPAC studies report that significant proportions of searches contain errors such as incorrect syntax, mis-spelled query terms, typographical errors, and logical errors [22]. The level of searching errors discovered in the CSTR logs is not unusual, although it was perhaps surprising for a user group that would be expected to be highly familiar with Boolean logic and computing interfaces. There appears to be a universal need for a search interface that will support the user in creating correctly formed queries.

7 Summary

The target user group for the CSTR collection—computer science researchers—might be expected to exhibit a propensity towards active exploration of new software and its functions. However, we have observed that the majority of users discriminate little when provided with tailorable querying options. Most accept the default settings, regardless of what those settings are. Very few investigate the system through supporting online documentation, or by experimentation with alternative settings and actions. Since this user group might be considered a 'best case' group for voluntary investigation of software this low level of interaction with the system indicates that initial default settings must be given full consideration.

Overall, user sessions are very short, few queries are submitted in those sessions, and the queries themselves are very simple. This strongly suggests that users wish to invest minimal time and effort in forming detailed specifications of their information needs. When refinement to queries does occur, users tend to make relatively small changes, most likely to involve addition or rearrangement of query terms. Little investigation of result sets occurs. Most user queries do not result in documents being viewed or retrieved, and it seems that users focus on only the first few returned documents. Consequently we must support users in converging rapidly on effective query terms and search options. Precision might be emphasised over recall in retrieving documents, given that exploration of result sets appears to be minimal.

Many users seem to be familiar with fielded searching, as evinced by their attempts to use cataloguing information such as title or author in their keyword searches. We should be working towards capitalizing on this familiarity by focussing on soft-parsing or heuristic techniques for extracting bibliographic information from uncatalogued documents [1].

As we stated in the introduction to this paper, transaction logs provide detailed descriptions of user actions for a particular system over a specific period of time. We have presented a study which has characterised more than a year's use of a digital library collection of technical reports, and suggested a range of issues to be addressed in the design of user interfaces to this and other types of collection. The period covered by the study was of particular interest because of the changes that occurred in the NZDL interface during that time. However, our studies are ongoing, and transaction logs for the NZDL are continuously captured. For the period from January 1998 to May 1999 our initial analysis shows that again more than two thirds (~73%) of queries adopt the default query type. Boolean queries display almost identical characteristics to those shown in Table 4: 1.4% of them use negation, 3.6% use union, 6.9% use compound expressions and 41% use intersection. Queries tend to be short—again approximately 80% of queries contained one, two or three terms. Once more, few queries contain amended query options although we do observe a little variation. The case sensitivity default was changed less often in the more recent period (in 1.5% of queries as opposed to 5% of queries in the earlier logs), and the stemming default was changed more often in the more recent period (in 14% of queries compared to 6% of queries for the earlier period). The topics of interest to users are also similar in both periods—38 of the 50 most frequent query terms are common to both periods of analysis. From this recent analysis we believe that the characterisations of use presented here are usefully generalisable beyond the period of the study and reflect ongoing use of the NZDL.

References

1. Bollacker, K.D, Lawrence, S. and Giles, C.L. CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In: Proceedings of the Second International Conference on Autonomous Agents, Minneapolis, St.Paul, May 9-13, 1998. ACM New York, pp 116-123, 1998.

2. Borgman, C.L. The user's mental model of an information retrieval system: an experiment on a prototype online catalog. International Journal of Man-Machine Studies, 24:1, 47-64 (1986).

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

4. Brown, R.N., and Agawala, A.K. On the behavior of users of the MEDLINE system. In: Changing Patterns in Information Retrieval: Tenth Annual National Information Retrieval Colloquium (Washington, DC: American Society for Information Science), pp 36-38, 1974.

5. Croft, W.B., Cook, R., and Wilder, D. Providing government information on the internet: experiences with THOMAS. In: Proceedings of Digital Libraries '95 , Austin, TX, June 11-13, 1995, pp 19-24, 1995.

6. Cunningham, S.J. Providing internet reference service for the New Zealand Digital Library: gaining insight into the user base for a digital library. In: Proceedings of the 10th International Conference on New Information Technology, Hanoi, Vietnam, March 24-26, 1998, pp 27-34, 1998.

7. Cunningham, S.J., and Silipigni-Connaway, L. Information searching preferences and practices of computer science researchers. In: Proceedings of OZCHI '96, Hamilton, New Zealand, November 24-27, 1996. IEEE Press, pp 294-299, 1996.

8. Davies, T. and Willie, S. The efficacy of a venn-based query interface: an evaluation. In: Proceedings of QCHI95 Symposium, Bond University, Queensland, Australia, August 21, 1995, pp 41-50, 1995.

9. Dickson, J. Analysis of user errors in searching an online catalog. Cataloging & Classification Quarterly, 4, 19-38 (1984).

10. Flaherty, P. Transaction logging systems: a descriptive summary. Library Hi Tech, 41, 11:2, 67-78 (1993).

11. Frank, E., Paynter, G.W, Witten, I.H., Gutwin, C. and Nevill-Manning, C.G. Domain-specific keyphrase extraction. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999. Morgan Kaufmann Publishers, San Francisco, CA. In Press.

12. Greene, S.L., Devlin, S.J., Cannata, P.E. and Gomez, L.M. No IFs, ANDs or ORs: a study of database querying. International Journal of Man-Machine Studies, 32:3, 303-326, (1990).

13. Halpin, T.A: Venn diagrams and SQL queries. The Australian Computer Journal, 21:1, 27-32 (1989).

14. Hunter, Rhonda N. Successes and failures of patrons searching the online catalog at a large academic library: a transaction log analysis. RQ, 30, 395-402 (1991).

15. Jansen, B.J., Spink, A., and Saracevic, T. Failure analysis in query construction: data and analysis from a large sample of web queries. In: Proceedings of Digital Libraries '98, Pittsburgh, PA, USA, June 24-27, 1998. ACM Press, New York, pp 289-290, 1998a.

16. Jansen, B.J., Spink, A., Bateman, J., and Saracevic, T. Real life information retrieval: a study of user queries on the web. Sigir Forum, 32:1, 5-17 (1998b).

17. Jones, S. Graphical query specification and dynamic result previews for a digital library. In: Proceedings of UIST '98: 11th Annual Symposium on User Interface Software and Technology. San Francisco, CA, USA, November 1-4, 1998. ACM Press, New York, pp 143-151, 1998.

18. Jones, S., and McInnes, S. A graphical user interface for boolean query specification. Working Paper 97/31, Dept. of Computer Science, University of Waikato, New Zealand, 1997.

19. McNab, R. J., Witten, I. H. and Boddie, S. J. A distributed digital library architecture incorporating different index styles. In: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries, Santa Barbara, CA, April 22-24, 1998. IEEE Press, pp. 36-45, 1998.

20. Millsap, L., and Ferl, T.E. Search patterns of remote users: an analysis of OPAC transaction logs. Information Technology and Libraries, September, 321-343, (1993).

21. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. Browsing in digital libraries: a phrase-based approach. In: Proceedings of ACM Digital Libraries '97, Philadelphia, July 25-28, 1997. ACM Press, New York, pp230-236, 1997.

22. Peters, T.A. The history and development of transaction log analysis. Library Hi Tech 42, 11:2, 41-66 (1993).

23. Peters, Thomas A. When smart people fail: ana anlysis of the transaction log of an online public access catalog. Journal of Academic Librarianship, 15, 267-273 (1989).

24. Pinelli, T.E. The information-seeking habits and practices of engineers. Science and Technology Libraries 12, 5-16 (1991).

25. Samuelson, P. Legally speaking: encoding the law into digital libraries, Communications of the ACM 41:4, 13-18, (1998).

26. Sandore, B. Applying the results of transaction log analysis. Library Hi Tech, 42, 11:2, 87-97, (1993).

27. Shneiderman, B. Designing the user interface, Addison-Wesley, 1999.

28. Spink, A., and Saracevic, T. Interactive information retrieval: Sources and effectiveness of search terms during mediated online searching". Journal of the American Society for Information Science, 48:8, 741-761, (1997).

29. Spink, A., Bateman, J., and Jansen, B.J. Searching heterogeneous collections on the web: behaviour of Excite users. Information Research: an Electronic Journal, 4,:2 (1998).

30. Tonta, Yasar. Analysis of search failures in document retrieval systems: a review. The Public-Access Computer Systems Review, 3,:1, 4-53 (1992).

31. Wallace, P.M. How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. RQ, 33:2, 239-252, (1993).

32. Williams, M. An evaluation of passage-level indexing strategies for a full-text document archive. LIBRES 8:1 (1998). Electronic publication, available at http://aztec.lib.utk.edu/libres/libre8n1/.

33. Witten, I.H., McNab, R., Jones, S., Cunningham, S.J., Bainbridge, D., and Apperley, M. Managing Multiple Collection, Multiple Languages, and Multiple Media in a Distributed Digital Library. *IEEE Computer*, 32, 2, 74-79 (1999).

34. Witten, I.H., Moffat, A., and Bell, T.C. Managing gigabytes: compressing and indexing documents and images, Van Nostrand Reinhold, New York, 1994.

35. Zink, Steven D. Monitoring user search success through transaction log analysis: the WolfPac example. Reference Services Review, 19, 49-56 (1991).

Figure 1: (a) showing the entry page to the NZDL, which supports collection selection and simple queries; (b) the advanced search page for the CSTR which allows control over query type and other query options.

Figure 2: the result display page for the CSTR. Query options are summarised, further queries can be issued and a summary passage is shown for each document.

Figure 3:Distribution of the length of user sessions

Figure 4: Frequency distribution of the location in the result list of viewed documents

Figure 5: Two month sample of number of queries issued per day

Table 1: by-domain breakdown of CSTR usage between April 1996 and July 1997. The total number of accesses was 28359.

		Accesses				Accesses	
Domain	Country	frequency	%	Domain	Country	frequency	%
edu		3515	12.39	th	Thailand	246	0.87
com		3406	12.01	pt	Portugal	237	0.84
de	Germany	3102	10.94	gr	Greece	231	0.81
nz	New Zealand	1957	6.90	ru	Russia	197	0.69
fr	France	1381	4.87	tw	Taiwan	193	0.68
au	Australia	1308	4.61	be	Belgium	185	0.65
ca	Canada	1307	4.61	il	Israel	185	0.65
kr	South Korea	1224	4.32	org		170	0.60
net		1197	4.22	at	Austria	151	0.53
uk	United Kingdom	1051	3.71	dk	Denmark	143	0.50
fi	Finland	918	3.24	id	Indonesia	128	0.45
jp	Japan	822	2.90	hk	Hong Kong	124	0.44
it	Italy	662	2.33	ch	Switzerland	89	0.31
other	Other Countries	567	2.00	pl	Poland	83	0.29
es	Spain	559	1.97	za	South Africa	77	0.27
br	Brazil	480	1.69	ar	Argentina	74	0.26
si	Slovenia	335	1.18	mx	Mexico	69	0.24
se	Sweden	310	1.09	uy	Uruguay	66	0.23
ie	Ireland	309	1.09	ph	Philippines	65	0.23
my	Malaysia	286	1.01	no	St. Pierre & Miquelon	63	0.22
sg	Singapore	274	0.97	mil		51	0.18
gov		267	0.94	arpa		20	0.07
nl	Netherlands	259	0.91	lk	Sri Lanka	16	0.06

Table 2: Frequency of Boolean and ranked queries for each default query type period.

	Boolean as default 46 week period		Ranked as default 15 week period		Total 61 week period	
Number of queries	24687		8113		32800	
Number of Boolean queries	16333	(66.2%)	2693	(33.2%)	19026	(58%)
Number of ranked queries	8354	(33.8%)	5420	(66.8%)	13774	(42%)

Table 3: Distribution of the number of terms in queries.

Number of terms in query		0	1	2	3	4	5	6	>6
Percentage of queries with given number of terms	Boolean queries	1.7	21.79	34.46	21.86	10.21	5.01	2.54	2.42
	ranked queries	1.42	34.36	33.47	16.85	7.28	3.21	1.40	2.01
	all queries	1.59	27.06	34.04	19.76	8.98	4.26	2.06	2.25

Table 4: Frequency of operators in Boolean queries.

	Boolean as default (total=16333)		Ranked as default (total=2693)		Full period (total=19026)	
Number of queries containing						
no Boolean operators	11394	(67.8%)	1171	(43.5%)	12565	(66.0%)
at least one intersection operator	3731	(22.8%)	1178	(43.7%)	4909	(25.8%)
at least one union operator	345	(2.1%)	122	(4.5%)	467	(2.5%)
at least one negation operator	181	(1.1%)	35	(1.3%)	216	(1.1%)
parentheses for compound expressions	683	(4.2%)	187	(6.9%)	869	(4.6%)

Table 5: The fifty most commonly occurring query terms.

Term	Occurrences	Term	Occurrences
and	570	robot	87
compression	232	atm	84
object	224	algorithm	83
software	199	video	81
database	194	recognition	80
network	181	text	79
neural	166	architecture	75
computer	154	graph	73
system	149	processing	72
information	139	for	70
oriented	138	a	69
data	137	networks	69
systems	130	java	68
image	128	memory	67
of	122	internet	66
distributed	113	learning	66
time	108	real	66
design	107	c	65
trie	99	multimedia	65
or	98	based	63
parallel	92	digital	63
security	91	in	62
model	89	the	60
retrieval	88	performance	59
analysis	87	code	58

Table 6: The number of documents matched by query terms issued by users.

Number of documents matching query terms	Number of query terms
0	280 (5.61%)
1	80 (1.60%)
2-10	333 (6.67%)
11-100	808 (16.18%)
101-200	394 (7.89%)
201-300	264 (5.29%)
301-500	320 (6.41%)
501-1000	439 (8.79%)
1001-5000	1058 (21.19%)
5001+	1017 (20.37%)
Total number of unique terms	4993

Table 7: Frequency of query error types.

Error type	Frequency	Percentage of all queries
No query string supplied	492	1.49
Query contains more than one phrase	163	0.50
Boolean operator appears in ranked query	1292	3.94
Mixed case terms appear with casefolding turned on	6717	20.48
Total errors	2365	29.14

Table 8: Frequency distribution of the number of queries issued in user sessions.

Number of queries issued in a user session	Frequency % of sessions
0	21.51
1	34.45
2	17.23
3	9.49
4	6.09
5	3.83
6	2.28
7	1.51
8	1.20
9	0.54
10	0.49
>10	1.38

Table 9: Distribution of the number of repeat visits to the digital library.

Number of visits	Number of users	Percentage
1	1993	72.82%
2	393	14.36%
3	118	4.31%
4	60	2.19%
5	39	1.42%
6	29	1.06%
7	16	0.58%
8	18	0.66%
9	10	0.37%
10	14	0.51%
11	7	0.26%
12	6	0.22%
13 - 122	34	1.24%
Total:	2737	100.00%

Table 10: Frequency with which consecutive queries contain common terms.

Number of common terms in consecutive queries	Frequency %
0	33.53
1	22.56
2	23.08
3	11.34
4	4.71
5	2.22
6	1.15
7	0.76
8	0.32
9	0.19
10	0.04

Table 11: Frequency with which combinations of refinements are made within consecutive queries.

Changes to query components						% of consecutive query pairs with change
Query string	Query terms	Query type	Search granularity	Stemming	Case sensitivity	
√	√					60.68
						13.75
√						5.66
		√				3.66
√	√	√				3.44
			√			2.72
√	√		√			2.56
√		√				1.19
√	√	√	√			1.03
					Other	5.31

Table 12: Summary of frequency with which refinements are made within consecutive queries (including changes to more than one attribute).

Changed query attribute	Percentage of consecutive query pairs with change
Query string	76.94
Query terms	69.46
Query type	11.74
Search granularity	8.75
Stemming	3.38
Case sensitivity	1.35

Table 13: Frequency with which result list size options are selected.

Maximum number of documents to be returned	Ranked		Boolean		Total	
	Frequency	%	Frequency	%	Frequency	%
50	5515	95.61	2293	77.36	7808	89.42
100	54	0.94	176	5.94	230	2.63
200	42	0.73	162	5.47	204	2.34
500	157	2.72	333	11.23	490	5.61

Table 14: Distribution of the number of documents viewed per query.

Documents viewed per query	Ranked		Boolean		Total	
	Frequency	%	Frequency	%	Frequency	%
0	3700	64.2	1909	64.4	5609	64.2
1	1103	19.1	573	19.3	1676	19.2
2	404	7.0	2.4	6.9	608	7.0
3	192	3.3	107	3.6	299	3.4
4	143	2.5	61	2.1	204	2.3
5	65	1.1	36	1.2	101	1.2
6	40	0.7	20	0.7	60	0.7
7	30	0.5	12	0.4	42	0.5
8	19	0.3	7	0.2	26	0.3
9	16	0.3	6	0.2	22	0.3
10	16	0.3	4	0.1	20	0.2
11-67	40	0.7	25	0.8	65	0.7

Table 15: Comparison of CSTR measures against those of Jansen et al for the Excite WWW search engine.

Measure	Excite (Jansen et al)	CSTR
average number of search terms per query	2.21	2.5
percent of queries with fewer than 4 terms	~86	~82
average number of queries per user session	2.8	2.04
percent of all queries using boolean operators	intersection 6.27	intersection 14.97
	union 0.26	union 1.42
	negation 0.23	negation 0.66
	parentheses 0.53	parentheses 2.65
result browsing	80% of users did not browse beyond the first 20 results	67.5% of of documents viewed were in the first 20 results
percent of repeated queries (same terms)	~38	~27
percent of unique queries (no term overlap)	35.2	33.5