Working Paper Series
ISSN 1170-487X

An Analysis of Usage of a Digital Library

by Steve Jones, Sally Jo Cunningham and Rodger McNab

Working Paper 98/13
June 1998

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

An Analysis of Usage of a Digital Library

Steve Jones, Sally Jo Cunningham, Rodger McNab

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}@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 conduct 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.

KEYWORDS

transaction log analysis, search interface, usage analysis.

1 INTRODUCTION

There is extensive literature on transaction log analysis of OPACs (see Peters, 1993 for an overview). However, only recently have these techniques has been applied to digital libraries-likely because many digital libraries have only just attained a usage level suitable for log analysis (Jansen et al, 1998; Spink et al, 1998). 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 apply transaction log analysis techniques to the New Zealand Digital Library (http://www.nzdl.org), focussing in this paper on the Computer Science Technical Reports (CSTR) collection. The CSTR contains nearly 46,000 publicly available computing-related technical reports harvested from over 300 research institutions from around the world. 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.

In the following section 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 3. A manual analysis of the logs is presented in section 4. In section 5 we summarise our observations.

2 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. Data from local users is not included in this analysis.

2.1 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 (Samuelson, 1998)).

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

Number of searches: 29,041 Time span: 30 April 1996 - 2 July 1997

Table 1: CSTR usage statistics by domain.

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.

3 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.

3.1 User Acceptance of Default Settings

The logs reveal that users rarely amend default settings for query and result display options (Table 2). With respect to query type (Boolean or ranked), only 33% of queries use non-default settings. This is consistent regardless of the default setting. Also, only 21% of queries changed the default term proximity setting. Default settings for case-sensitivity and stemming were changed even less frequently-in only 5% and 6% of queries respectively. The default result set size was changed in only 10.5% of user queries.

	Boolean as default 46 week period	Ranked as default 15 week period	Total 61 week period
Number of queries	24687	8115	32802
Boolean queries	16333 (66.2%)	2693 (33.2%)	19026 (58%)
Ranked queries	8354 (33.8%)	5420 (66.8%)	13774 (42%)

Table 2: Frequency of Boolean and ranked 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 3.2), 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; an additional user study is needed to confirm this default setting.

3.2 Query Complexity

The CSTR collection supports both ranked and Boolean querying (including intersection, union and negation operators and compound expressions formed through inclusion of parentheses). Queries tend to be short and simple-the average number of search terms in a query is 2.5 and just under 80% of queries contained one, two or three terms (see Table 3). Given this extreme query brevity the choice of each search term becomes crucial. We are investigating techniques to support users in selecting terms which accurately and concisely represent their information needs (Nevill-Manning et al, 1997).

No. of terms in query	0	1	2	3	4	5	6	>6
Frequency (total=32796)	492	8788	11095	6505	2926	1477	692	821
Percentage	1.5	26.79	33.83	19.83	8.92	4.50	2.11	2.5

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

Just over a quarter of Boolean queries contained at least one intersection operator, only 2.5% contained at least one union operator and only 1% included the negation operator. Only 4.5% of Boolean queries contain compound expressions. By far the majority of Boolean queries use no Boolean operators at all (see Table 4). Consequently we might surmise that the underlying search engine need not be further optimised to process complex queries.

	Boolean as default 46 week period	Ranked as default 15 week period	Total 61 week period
Number of Boolean queries containing			
intersection	3731 (22.8%)	1178 (43.7%)	4909 (25.8%)
union	345 (2.1%)	122 (4.5%)	467 (2.5%)
negation	181 (1.1%)	35 (1.3%)	215 (1.1%)
compound expressions	682 (4.2%)	187 (6.9%)	869 (4.6%)

Table 4: Frequency of operators in Boolean queries.

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 (Borgman, 1986; Borgman, 1996; Greene et al, 1990). 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. Similarly, OR tends to be exclusive in English, but inclusive in Boolean logic. 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 (Davies & Willie, 1995; Halpin, 1989). For this reason we are investigating alternative interface metaphors for Boolean querying (Jones & McInnes, 1997).

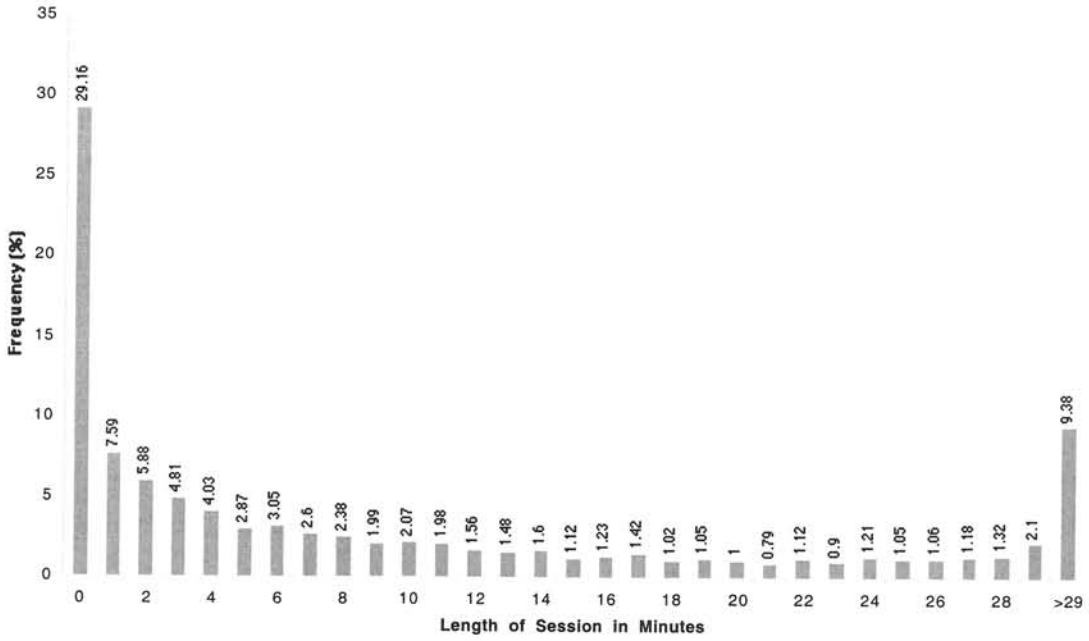
3.3 User Sessions

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 5. 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, 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 to determine just why so many users seem to abandon their query process prematurely.

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

Table 5: Frequency of queries issued in user session.

The length of user sessions was also recorded. 29.16% of sessions lasted less that one minute (see Figure 1). 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 session 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.

Figure 1: Distribution of user session lengths.

3.4 Query Refinement

Analysis of users' consecutive queries reveals interesting aspects of query refinement behaviour. A set of 6680 user sessions was analysed and 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 6). 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. User are likely to make minor changes by adding a new term, or altering the existing 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 6: Frequency with which consecutive queries contain common terms.

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 7. 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.

Changes to Query Components						Percentage 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 7: Frequency with which refinements are made within consecutive 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 8 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.

Changed Query Attribute	Percentage of Consecutive Query Pairs With Change
Query String	76.94
QueryTerms	69.46
Query Type	11.74
Search Granularity	8.75
Stemming	3.38
Case Sensitivity	1.35

Table 8: Frequency with which refinements are made within consecutive queries (including changes to more than one attribute).

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.

3.5 Result Viewing

In almost 90% of queries the default result set size of 50 documents was retained (see Table 9). 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, with 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.

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 9: Frequency with which result list size options are selected.

Disappointingly, the majority of queries (64.2%) do not lead to users viewing document content (see Table 10). 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. However, the queries that users form may be too simplistic to produce result lists which appropriately match their needs. Alternatively, 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 [Williams, 1998].

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	204	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 10: Distribution of the number of documents viewed per query.

When users do view documents they are most likely to view those which are at the start of the result list (see Figure 2). 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 retrieved 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.

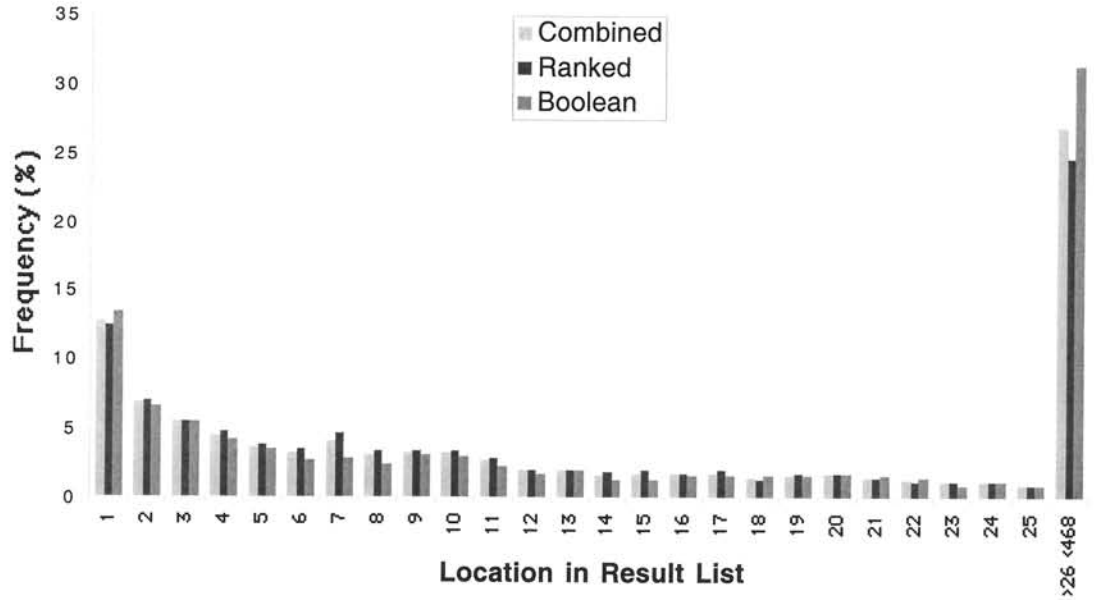


Figure 2: Distribution of the result list location of viewed documents.

3.6 Server Loading

Figure 3 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.

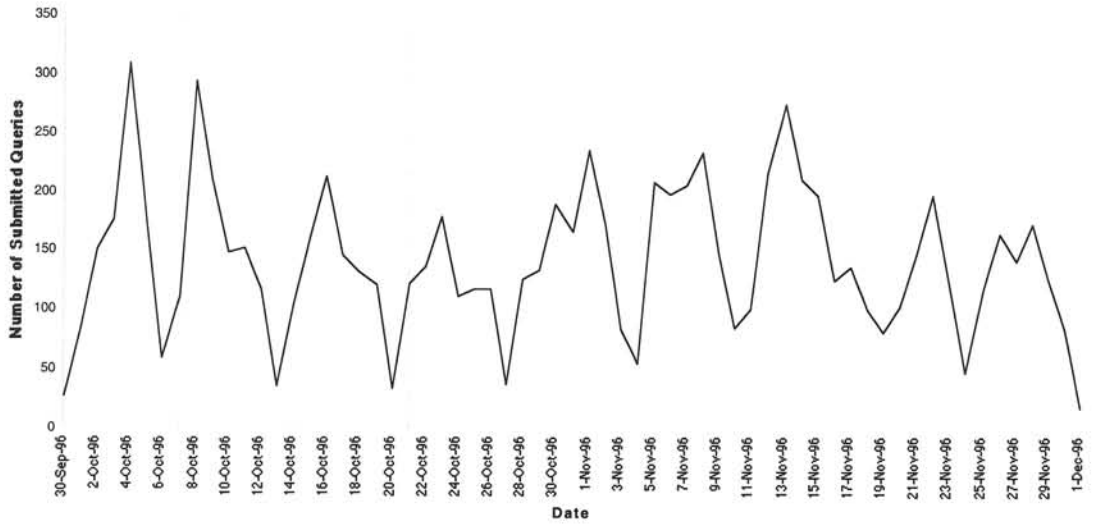


Figure 3: Two month sample of server loading showing number of submitted queries.

4 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.

4.1 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.

4.2 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 categorically 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 in 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 Initiative 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 (Cunningham, 1998). 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 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).

4.3 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.

Again, only an approximate measure of the number of appearances of these types of search 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 indicates that some searches are indeed intended to retrieve documents where 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 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 (Bollacker et al, 1998).

6 REFERENCES

Bollacker, K.D, Lawrence, S. and Giles, C.L. CiteSeer: an Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. *Proceedings of the Second International Conrerence on Autonomous* Agents, Minneapolis, St.Paul, May 9-13, 1998.

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 (1986), 47-64, 1986.

Borgman, C.L. Why Are Online Catalogs Still Hard to Use? *Journal of the American Society for Information Science*, 47, 7 493-503, 1996.

Cunningham, S.J. Providing internet reference service for the New Zealand Digital Library: Gaining insight into the user base for a digital library. *Proceedings of the 10th International Conference on New Information Technology*, pp. 27-34, 1998.

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, pp 41-50, 1995.

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.

Halpin, T.A: Venn Diagrams and SQL Queries. *The Australian Computer Journal*, 21, 1, 27-32, 1989.

Jansen, B.J., Spink, A., and Saracevic, T. Failure Analysis in query construction: data and analysis from a large sample of web queries. *Proceedings of Digital Libraries '98*, Pittsburgh, 1998.

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.

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

Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. Browsing in Digital Libraries: a Phrase-based Approach. *Proceedings of ACM Digital Libraries '97*, Philadelphia, 230-236, 1997.

Peters, T.A. The History and Development of Transaction Log Analysis. *Library Hi Tech* 42 (11:2), 41-66, 1993.

Spink, A., Bateman, J., and Jansen, B.J. Searching heterogeneous collections on the web: behavior of EXCITE users. *Proceedings of Digital Libraries '98*, Pittsburgh, 1998.

Williams, M. An Evaluation of Passage-Level Indezing Strategies for a Full-Text Document Archive. *LIBRES* 8, 1, 1998. Electronic publication, available at http://aztec.lib.utk.edu/libres/libre8n1/.