

# Usage Analysis 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

We analyse transaction logs for a large full-text document collection for Computer Science researchers. We report insights gained from this analysis and identify resulting search interface design issues.

**KEYWORDS:** transaction log analysis, search interface, usage analysis.

## INTRODUCTION

There is extensive literature on transaction log analysis of OPACs (see [3] for an overview). However, little work of this nature has been applied to digital libraries—likely because many digital libraries have only recently attained a usage level suitable for log analysis. Since log analysis provides insight into user search behaviour it is useful in the design and consideration of query interfaces.

We apply transaction log analysis techniques to the New Zealand Digital Library (<http://www.nzdl.org>). We focus on the Computer Science Technical Reports (CSTR) collection which contains almost 46000 publically available Computer Science technical reports from around the world. Because the collection is not formally catalogued users carry out keyword searches within the *full* texts of the documents. Both ranked and Boolean querying are supported.

## DATA COLLECTION AND ANALYSIS

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.

	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 1: Frequency of Boolean and ranked queries**

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 2: Distribution of the number of terms in queries**

## RESULTS

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.

### User Acceptance of Default Settings

The logs reveal that users rarely amend default settings for query and result display options (Table 1). 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. There are two possible interpretations of these observations. First, the default settings are appropriate to the requirements of the majority of users. However this 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.

### 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 simplistic. 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 2). Therefore we are investigating techniques to support users in selecting terms which accurately and concisely represent their information needs [2]. 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

	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 3: Frequency of operators in Boolean queries**

contain compound expressions. By far the majority of Boolean queries use no Boolean operators at all (see Table 3). Consequently we might surmise that the underlying search engine need not be further optimised to process complex queries. We might expect the target users of the CSTR to be conversant with Boolean logic, yet they appear unwilling to apply it when searching. Human-computer interaction literature shows that user querying behaviour is related more closely to the query model user interface than to the query model itself. We are investigating alternative interface metaphors for Boolean querying [1].

### Query Refinement

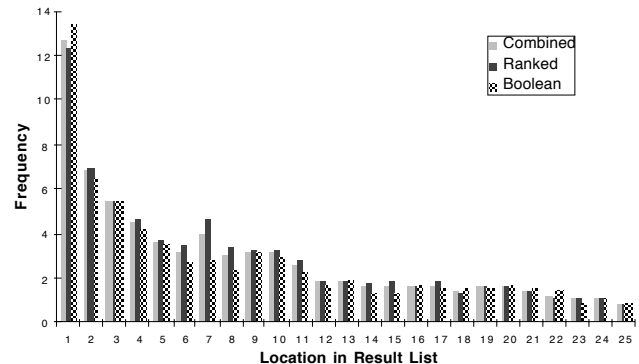
Analysis of consecutive queries indicates that more than half of all queries build on the previous query (discounting the first query in each user's session), having at least one query term in common. This high incidence of term overlap implies query refinement is a common activity. Iterative query refinement (generally only addition or removal of terms) requires supportive interface mechanisms which appropriately direct query refinement [1].

### Result Viewing

In almost 90% of queries the default result set size of 50 documents was retained. Intermediate sizes of 100 and 200 were each requested in approximately 2.5% of queries, and a size of 500 was requested almost 6% of the time. 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. Boolean queries are more likely to require larger result sets to be returned. The majority of queries (64%) do not lead to users viewing document content (see Table 4). The distributions of the number of documents viewed for ranked and Boolean queries are strikingly 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 [4]. Documents are more likely to be viewed the closer they are to the beginning of the result list (see Figure 1). 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.

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 4: Number of documents viewed per query**



**Figure 1: Location of viewed documents in result lists**

## DISCUSSION

Transaction log analysis, as applied to OPACs, has yielded a diversity of results [3]. 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 system, for a particular document collection. In this paper we have suggested way that these fine-grained details can then be used to tailor our system to its target user group.

## REFERENCES

1. 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.
2. Nevill-Manning, C.G., Witten, I.H. and Paynter, G.W. Browsing in Digital Libraries: a Phrase-based Approach. *Proc ACM Digital Libraries '97*, Philadelphia, 230-236, 1997.
3. Peters, T.A. The History and Development of Transaction Log Analysis. *Library Hi Tech* 42 (11:2), 41-66, 1993.
4. Williams, M. An Evaluation of Passage-Level Indexing Strategies for a Full-Text Document Archive. *Working Paper 98/7*, Dept. of Computer Science, University of Waikato, New Zealand, 1998.