Working Paper Series
ISSN 1170-487X

A Comparative Transaction Log Analysis of Two Computing Collections

by Malika Mahoui
and Sally Jo Cunningham

Working Paper 00/12
July 2000

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

A Comparative Transaction Log Analysis of Two Computing Collections

Malika Mahoui and Sally Jo Cunningham

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Telephone: +64 7 838 4021 Fax: +64 7 838 4155 Email: {mmahoui, sallyjo}
@cs.waikato.ac.nz

Abstract. Transaction logs are invaluable sources of fine-grained information about users' search behavior. This paper compares the searching behavior of users across two WWW-accessible digital libraries: the New Zealand Digital Library's Computer Science Technical Reports collection (CSTR), and the Karlsruhe Computer Science Bibliographies (CSBIB) collection. Since the two collections are designed to support the same type of users-researchers/students in computer science-a comparative log analysis is likely to uncover common searching preferences for that user group. The two collections differ in their content, however; the CSTR indexes a full text collection, while the CSBIB is primarily a bibliographic database. Differences in searching behavior between the two systems may indicate the effect of differing search facilities and content type.

1 Introduction

Transaction log analysis has been extensively applied to the study of searching behavior in OPACs (for an overview of this body of literature, see [3]). Log analysis has the advantage of providing a descriptive summary of large numbers of user interactions, and as such is a valuable research tool for investigating user search behavior. As a caveat, it is difficult to generalize for all users, at all times-the value of the technique lies in describing a particular system's users, searching specific collections. The contribution of this paper is to provide an opportunity for generalization by examining the logs for two digital libraries created to support the same type of user: researchers and tertiary students in computer science.

Digital libraries usage logs have only recently attained a usage level appropriate for log analysis (see, for example [1, 4]). In a previous paper [2], we applied transaction log analysis techniques to the Computer Science Technical Reports (CSTR) collection of the New Zealand Digital Library (http://www.nzdl.org). At that time, the CSTR indexed nearly 46,000 publicly available computing-related technical reports harvested from over 300 research institutions from around the world. As no formal bibliographic metadata is available for the CSTR documents, the search interface is based solely on keyword searching, and the full text of the documents have been indexed. When searching identifies a relevant

document, the document can be immediately retrieved; although the CSTR itself does not store copies of the collection documents, the collection provides links to the original sites from which the documents were identified for indexing.

In contrast, the Computer Science Bibliographies (CSBIB) collection primarily contains bibliographic metadata: the collection holds over a million references in BIBTEX format, donated as single references, subject bibliographies, or single-author bibliographies (http://liinwww.ira.uka.de/bibliography/index.html). Over 90,000 of the references contain a link to an online version of the corresponding paper; however, the full text of the papers is not searchable, and less than 1% of the references include links to papers. The emphasis of the CSBIB collection is the provision of references, rather than immediate access to the documents themselves.

In the following section we describe how the transaction log data has been collected. The usage logs are automatically processed for analysis. The results of this analysis are presented in Sections 3 - 5: we describe query complexity and query types, analyze user session lengths, and examine user acceptance of system default settings. These statistics are compared across the two collections, and (where relevant) across different search options within each system.

2 Data collection

All user activity in the CSBIB and CSTR collections is automatically logged. Although actions can be grouped into user sessions, individual users cannot be identified—the users of the two digital libraries remain anonymous. For the CSTR collection, local users were excluded from this analysis; during the period studied many local queries were submitted as system tests.

The CSBIB log totals 251878 queries submitted in the period 1st September 1999 to 26th December 1999. CSBIB queries are of two types: simple search (keyword search over all bibliographic record fields) and advanced search (in which users specify which fields to search). For the CSTR, over 30,000 queries were recorded from April 1996 to July 1997. User activities in both logs are timestamped. The log attributes used in this analysis include the query text and query options (ie, simple or advanced for CSBIB, ranked or Boolean for CSTR).

3 Query complexity

Queries to both collections are short; over 90% of the CSBIB simple search queries and approximately 80% of CSTR queries contain three or fewer terms (Table 1). The average number of search terms in a CSBIB simple search query is 1.8%, compared to 2.5% in the CSTR collection. This difference is slightly misleading, in that the CSBIB syntax permits author names to be entered with initials appended to the family name (for example, as SmithJ rather than Smith J or John Smith)—which will have the effect of reducing the number of query terms in many author queries.

Table 1. Distribution of the number of query terms (CSBIB simple search, CSTR)

No of terms in query	0	1	2	3	4	5	6	>7
CSBIB (simple search)	0%	52.72%	28.10%	10.8%	4.18%	1.75%	1.02%	1.41%
CSTR	1.59%	27.06%	34.04%	19.76%	8.98%	4.26%	2.06%	2.25%

In both systems the default Boolean operator is an OR, and the majority of queries contain no explicit Boolean operator. The most common user-specified Boolean operator is the AND (Table 2). Interestingly, in the CSBIB collection

Table 2. Frequency of operators in Boolean queries.

percentage of queries containing	CSTR	CSBIB
no Boolean operators	66.0%	84.1%
at least one intersection operator	25.8%	14.18%
at least one union operator	2.5%	1.69%
parentheses for compound expressions	4.6%	0.01%

only14% of the queries included the AND operator compared to 26% in the CSTR collection–despite help text on the CSBIB simple search page that concisely enumerates the logic operators and gives a small example on how to use them. The CSTR interface does not contain syntax help on the search page itself.

4 User sessions

Each query in the transaction logs contains a user id (although, as noted in Section 2, these ids cannot be traced back to collection users). A simple heuristic was used to identify user 'sessions': a session is assumed to be a series of queries containing the same user id, and with no more than a 30 minutes lapse between consecutive queries.

For both the CSTR and the CSBIB collection, users tend to issue relatively few queries per session (Table 3); over 80% of user sessions include 5 or fewer queries. The interesting difference between the CSTR and CSBIB collections is the relatively larger number of CSBIB sessions that include more than 9 queries. It appears that CSBIB users tend to persevere in query refinement to a greater extent than CSTR users. We can only hypothesize about the basis for this difference. It may indicate that the CSBIB interface encourages a greater degree of exploration; or that the users of the CSTR tend to require more exhaustively complete results (a plausible hypothesis, if CSBIB users are more likely to be students/academics seeking to complete literature reviews); or that the ability to easily view the full text of search results (as in the CSTR) permits a user to more quickly home in on relevant documents in the collection; or perhaps

Table 3. Frequency distribution of the number of queries issued in user sessions.

No queries issued in a user session	CSTR % of sessions	CSBIB (simple search) % of sessions	CSBIB (advanced search) % of sessions
1	43.89	35.97	29.95
2	21.95	20.02	20.43
3	12.1	12.19	12.88
4	7.76	8.51	8.46
5	4.88	5.84	5.82
6	2.90	3.83	4.22
7	1.92	2.68	3.14
8	1.53	2.13	2.35
>9	2.41	7.32	10.79

there are other explanations. This point indicates a weakness of transaction log analysis: it can indicate patterns of user behaviors, but cannot explain those behaviors. At this point, we must engage in an interview-based or ethnographic study to further explore the users' motivations.

Tracing the individual user ids across the time periods captured in the transaction logs, we can determine the number of repeat visitors to the CSTR and CSBIB digital libraries (Table 4).

Table 4. Distribution of repeat visits to the CSTR and CSBIB collections.

number of visits	CSTR (%)	CSBIB (%)
1	72.82	72.48
2	14.36	11.24
3	4.31	4.24
4	2.19	2.99
5	1.42	1.65
6	1.06	1.49
> 6	3.84	5.88

Again, the two logs indicate similar behavior. Disappointingly, nearly three-quarters of the users of both collections visited the collection only once during the extensive time period covered in this analysis. The result is perhaps tied to the relatively large proportion of users who issue only one or two queries during a search session—and who presumably either have very straightforward information needs that are quickly satisfied, or who decide that the collection cannot fulfil their information needs.

Analysis of consecutive queries in a session reveals an interesting difference in query refinement behaviour: more than half (58.38%) of the queries in CSBIB

sessions have no term in common with the previous query, compared to only 33% of the consecutive queries in the CSTR collection (see Table 5).

Table 5. Frequency with which consecutive queries contain common terms (CSBIB simple search, CSTR).

No of terms in consecutive queries	0	1	2	3	4	5	>5
CSBIB (simple search)	58.38%	25.85%	10.70%	2.98%	1.11%	0.45%	0.51%
CSTR	33.53%	22.56%	23.08%	11.34%	4.71%	2.22%	2.25%

These figures discount the first query in a session. This low incidence of term overlap in CSBIB collection indicates either that CSBIB users tend to attempt to satisfy more than one distinct information need in a session, or that the CSBIB users refine consecutive queries more radically than CSTR searchers.

5 User acceptance of default settings

The logs from both collections show that users tend to settle for the system's default settings. For the CSTR, during the log collection period the default search type was changed from Boolean to Ranked–and approximately 66% of queries used the default setting, no matter what the setting actually was (Table 6). Approximately 80% of queries to the CSBIB collection were submitted through

Table 6. Frequency of default query types.

CSTR	Boolean as default 46 week period	Ranked as default 15 week period		CSBIB	
number of queries	24687	8115		simple queries	202947 (80.57%)
Boolean queries	16333 (66.2%)	2693 (33.2%)		advanced queries	48931 (19.43%)
Ranked queries	8354 (33.8%)	5420 (66.8%)			

the default standard search (keyword search) interface. Given the strong tendency of this user group to accept system defaults, it is important that these defaults be set 'correctly'–that is, so as to maximize the opportunities of the users in satisfying their information needs. Given that users tend to submit short, simple queries (Section 3), a useful interface strategy may be to maximize search recall for those queries. To that end, the choice of ranked output as default for the CSTR and simple search (keyword, over all bibliographic record fields) for the CSBIB appear to be sensible interface design decisions.

6 Conclusions

This paper updates a previous study on transaction log analysis ([2]) by comparing the searching behavior of users across CSTR and CSBIB, two WWW-accessible digital libraries for computer science researchers. Commonalities in the log analysis results indicates that this user group prefers to issue relatively few, brief queries in a session. While computing researchers might be expected to actively explore software and its functions, these users tend to accept default settings-no matter what those settings are. Differences in search behavior across the two systems include a tendency of CSBIB users to issue more queries in a session, and for consecutive queries in a session to show a lesser degree of term overlap. Further study (including qualitative, interview-based examinations of small groups of users) is indicated to determine the basis for these differences.

Acknowledgements

Alf-Christian Achilles has developed and maintained the CSBIB collection since 1995; he generously provided the CSBIB transaction logs analyzed in this paper.

References

1. Jansen, B.J., Spink, A., Saracevic, T.: Failure analysis in query construction: data and analysis from a large sample of web queries. *Proceedings of ACM Digital Libraries*, Pittsburgh (1998) 289-290
2. Jones, S., Cunningham, S.J. and McNab, R.: An analysis of usage of a digital library. *European Conference on Digital Libraries*, Heraklion, Crete, Greece, August. Lecture Notes in Computer Science **1513** (1998) 261-277
3. Peters, T.A.: The history and development of transaction log analysis. *Library Hi Tech* **42**(11:2) (1993) 41-66
4. Spink, A., Batement, J., and Jansen, B.J.: Searching heterogeneous collections on the web: behaviour of Excite users. *Information Research: an electronic journal* (1998) 4:2