

2001

Search Behavior in a Research Oriented Digital Library

Malika Mahoui

Sally Jo Cunningham

Report Number:

01-007

Mahoui, Malika and Cunningham, Sally Jo, "Search Behavior in a Research Oriented Digital Library" (2001). *Computer Science Technical Reports*. Paper 1505.

<http://docs.lib.purdue.edu/cstech/1505>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

**SEARCH BEHAVIOR IN A RESEARCH-ORIENTED
DIGITAL LIBRARY**

**Malika Mahoui
Sally Jo Cunningham**

**CSD TR #01-007
May 2001**

Search behavior in a research-oriented digital library

Malika Mahoui¹, Sally Jo Cunningham²

¹Department of Computer Science, Purdue University,
West Lafayette, Indiana, USA 47907
mmahoui@cs.purdue.edu

²Department of Computer Science, University of Waikato,
Private Bag 3105, Hamilton, New Zealand
sallyjo@cs.waikato.ac.nz

Abstract: This paper presents a transaction log analysis of ResearchIndex, a digital library for computer science researchers. ResearchIndex is an important information resource for members of this target group, and the collection sees significant use worldwide. Queries from over six months of usage were analyzed, to determine patterns in query construction and search session behavior. Where appropriate, these results are compared to earlier studies of search behavior in two other computing digital libraries.

1. Introduction

Understanding the information behavior of digital library users is central to creating useful, and usable, digital libraries. One particularly fruitful area of research involves studying how users interact with the current library interface, with a view to using the insights gained from the study to improve the library's interface or the collection's contents.

Many different techniques exist to study the behavior of library users: focus groups, talk-aloud protocols, and post-search interviews. These techniques are rich sources of data for gaining insight into users' search intentions and high level strategies, but they are also highly intrusive—and so the data gathering itself may skew the search/browsing tasks, or it may be subject to faulty memories or retroactive re-interpretation of search behavior.

Transaction log analysis—examining information behavior through the search artifacts automatically recorded when a user interacts with a library search system—offers an unobtrusive means for finding out *what* users are doing in a digital library. Although log analysis cannot provide insight into the *why* of search behavior, this method supports examination of very large numbers of search sessions and queries, on a scale that more qualitative studies cannot match.

Although transaction log analysis (TLA) has been applied extensively to the study of search behavior on conventional library OPACs, few studies of digital libraries (for example, ([2], [5]) or other large-scale WWW-based document collections (for example, [8]) exist. Presumably few log analyses exist because digital libraries have only recently seen usage levels warranting analysis. Other search interfaces, such as

WWW search engines, tend to be commercial enterprises, and are generally reluctant to allow research access to their usage logs.

In this paper, we use TLA techniques to study usage patterns in the ResearchIndex (formerly known as CiteSeer) digital library (<http://www.researchindex.org/cs>). ResearchIndex (RI) has been developed and maintained by the NEC Research Institute, Inc. It is a large digital library; at the time of the data collection, it provided access to over 290,000 full text documents. This analysis is compared with results from previous studies of two other digital libraries: the Computer Science Technical Reports (CSTR) collection developed by the New Zealand Digital Library project¹; and the Computer Science Bibliographies² (CSBIB) maintained by Alf-Christian Achilles at Karlsruhe University. The CSTR log analysis statistics described in this paper are presented in more detail in [2]; the CSBIB results were previously published in [5].

All three digital libraries are intended to support the same type of user: computer science researchers and tertiary computing students. The comparison of log analysis results is of significance, then, as it highlights the common search behavior shown by this group. Differences in behavior across the three systems can, in some cases, be traced to differences in the search interfaces.

In the following section, we describe the ResearchIndex digital library, and briefly outline the interface and collection characteristics of the CSTR and CSBIB digital libraries. Section 3 describes the collection of the usage data from these three digital libraries. Sections 4 – 6 present the results of the analysis of the ResearchIndex logs, describing user demographics, user session lengths, query complexity, and query refinement patterns. Where applicable, these results are compared to previous results from analysis of CSTR and CSBIB usage logs.

2. Three Computer Science digital libraries: RI, CSTR, CSBIB

ResearchIndex (RI), previously known as CiteSeer ([3], [4]), is a digital library focusing on computer science research documents. During the period in which the transaction logs were collected, the collection included more than 290,000 documents. These documents are not assumed to have any bibliographic record available; instead, the document's text is extracted and then parsed to extract the document's bibliographic details and its list of references to other documents. This information is then used to build a citation index and a full text index. Given a search query, ResearchIndex retrieves either the documents (document option) for which the content match best the query terms, or the citations (citation option) that best matches the query terms.

Using the document option, the user can browse through each document; information displayed includes the first lines of the documents, the list of references cited in the paper, the list of papers citing the document and the list of other related documents. The user may select any of these entries for further browsing. He/she also may download the paper or choose to display further extracted text.

¹ <http://www.nzdl.org>

² <http://liinwww.ira.uka.de/bibliography/index.html>

references. Approximately 9% of the references included a link to an online version of the corresponding paper. However, the full text of the online papers was not searchable. CSBIB also offers a simple and an advanced search interface. The simple search screen for the CSBIB is similar to the advanced search option of the CSTR; users can select a number of options, including stemming, number of documents in the result set, etc. The CSBIB advanced search supports the simple search options, and also allows a user to limit searches by bibliographic field (author, title, date, etc.).

3. Data Collection

User activity was automatically logged on all three digital libraries. At the times that the log files were collected, the CSTR and the RI systems were undergoing testing by the digital library developers. For that reason, local queries for these two collections were excluded from this analysis, as during the period studied many local queries were submitted as system tests.

The total number of queries and the time period of study are summarized in Table 1. For all three digital libraries, user activities are timestamped and include the machine identifier (IP address) from which the query was issued, the query text, and all query options selected (for example, ranked or boolean). The users themselves remain anonymous. Since users do not log in or out of the system, it is problematic to identify the beginning/end of a session. A simple heuristic was used to approximate session limits: a session is assumed to be a series of queries containing the same machine identifier, and with no more than a 30 minute lapse between consecutive queries.

The logs from all three systems were taken over significant periods of time, allowing us to view user activities across more than one session. This longer time period also reduces the possibility that the logs represent an atypical or unrepresentative set of queries and usage patterns.

Table 1. Summary of data collection

Digital Library	Period of study	No. of weeks	No. of queries/accesses	No. of user sessions	Average no. of queries per week
RI	Aug 99 – Feb 2000	29	1,541,148	46,486	53,143
CSTR	Apr 96 – Jul 97	61	32,802	26,128	428
CSBIB	Sept – Dec 99	17	251,878	54,671	14,816

4. ResearchIndex user demographics

Since the RI collection is freely accessible—users do not register for the collection—the only information held on a user is the IP address through which that user accessed the RI digital library. This is one significant drawback to studying search behavior

that is shared by many digital collections: it is not possible to incorporate detailed user demographics into the transaction log analysis.

However, this user anonymity has its advantages: anonymous access appears likely to prove attractive to computing digital library users, and to increase the appeal of a particular library. In all three collections studied in this paper, users appear to prefer brief interactions with the search systems—and so would likely prefer a system that allows them to immediately begin searching, without spending time registering or verifying their account. Other research suggests that digital library users may be concerned about privacy [7]—and so users may prefer a system that prevents user interest profiles from being linked to a particular individual.

Examination of user domain codes indicates that educational (.edu) institutions form the largest identifiable group of users—suggesting that the RI digital library is indeed reaching its intended users in tertiary institutions. A similar proportion of commercial (.com) users presumably indicates that the RI collection is seeing use in corporate research and development units.

The remaining domain codes primarily indicate national origin, with the highest proportion of use by country coming from users located in Europe (particularly Germany, France, and the UK). RI is truly receiving worldwide attention: the top 24 domains are drawn from such linguistically diverse and geographically dispersed countries as Japan, Brazil, Israel, and Greece.

Table 2. RI usage by domain.

Domain	Sessions (%)	Domain	Sessions (%)
cdu	17.04	pt (Portugal)	1.31
com	16.76	gr (Greece)	1.27
net	10.29	br (Brazil)	1.17
de (Germany)	4.27	se (Sweden)	1.13
fr (France)	4.18	ch (China)	0.94
uk (United Kingdom)	3.18	gov (Government)	0.79
ca (Canada)	2.84	es (Spain)	0.78
it (Italy)	2.52	at (Austria)	0.58
nl (Netherlands)	1.97	fi (Finland)	0.57
au (Australia)	1.94	il (Israel)	0.56
jp (Japan)	1.69	All others	22.54

5. User Sessions

The 1,541,148 queries or browsing accesses logged for the RI digital library are divided into 46,486 sessions. The first, startling, result from the analysis of these

sessions is only about 6% of the total number of sessions started with a citation/document search query—that is, from the main search page for the digital library! 4.17% of the total number of user sessions began with a citation search query, and 1.85% started with a document search query; the vast majority of sessions began with a search that bypassed the main query screen. If the users don't enter the digital library through the initial search page, then how do they get in?

We suggest two possible explanations for this situation: either technique that we use for identifying the start of a session is not appropriate for ResearchIndex data, or that the majority of the sessions have been initiated by linking through the results of a previously executed query from a search engine such as Altavista or Google. Setting the timeout between two user sessions to 30 minutes is a heuristic that is plausible from a commonsense point of view, and this heuristic has been adopted by most of the community working on TLA and Web mining (see, for example, [9]). Further, an earlier study of computing researchers indicated that many of these researchers used general purpose search engines to locate research papers more frequently than they used 'formal' computing subject indexes [1]. We therefore tend to the second conjecture, particularly as an examination of the results from popular search engines for queries containing computing-related term reveals the frequent presence of links to RI search result pages.

This observation is emphasized by the total number of sessions including either citation or document search queries as shown in Table 3 (53.31%). When combined with the number of sessions that started with citation/document search queries, we conclude that about 47.31% of the sessions originated by loading results of a 'ready made' search query, and then included at least one citation or document search query later in the session. This suggests that links from general purpose search engines are an effective way to draw users into a digital library, as nearly half of the sessions are initiated in this way and then include further exploration of the RI collection.

Table 3. Summary of session activity

Total number of sessions	% sessions <i>not including</i> search queries	% sessions <i>including</i> search queries
46,486	46.69	53.31

Table 4. frequency of the query types in sessions including search queries

% sessions starting with a search query	% sessions including both citation and document queries	% sessions not including document queries	% sessions not including citation queries
6.02	19.87	24.12	9.4

Table 4 shows the percentage of search sessions not including citation search queries (9.4%) compared to the percentage of search sessions not including document search queries. Recall that 4.17% of the total number of user sessions began with a citation search query, and 1.85% started with a document search query. Taken together, these results indicate that users tend to explicitly change the default search type (citations search) and prefer to run a document type search.

This is an interesting observation, since the CSTR and CSBIB users, in the overwhelming majority of cases, do *not* change default settings ([1], [5]). The

movement of the RI users from the default citation search to the (full text) document search therefore gains significance: the changing of the default is unlikely to occur unless a clear, strong preference exists for full text search. Perhaps the common usage of full text search through general purpose search engines such as Google or AltaVista when conducting a literature survey plays a part in this preference for document search [1]. Or perhaps researchers do not normally begin a search with citation links: the computing researchers studied in the Cunningham and Connaway [1] investigation used citation links, but only by following links within documents that they had read and found relevant. Again, a limitation of transaction log analysis is that it can tell us *what* occurs in a search session, but not *why* those actions occurred; we must therefore be cautious in ascribing motivations to the patterns of action that we observe. On the other hand, the volume of data that is analyzed in these transaction logs, and the length of time over which the logs were gathered, gives confidence that the observed pattern is not a product of coincidence or chance.

Table 5. Frequency distribution of the number of queries issued in user session for RI

Number of search queries issued in user sessions	Adjusted number of search queries issued in user sessions	Sessions (%)
0	1	46.69
1	2	11.51
2	3	7.48
3	4	5.42
4	5	3.86
5	6	3.10
6	7	2.63
7	8	2.07
8	9	1.63
9	10	1.50
10	11	1.21
10 < x < 31	11 < x < 32	9.59
>30	>31	3.24

The analysis of the frequency distribution of queries issued in user sessions for ResearchIndex (Table 5) presented challenges, mainly because of the large portion of sessions that did not include a search query (46.69%). One way to compute the percentages is to discard the sessions that strongly suggest the presence of outliers. This category refers not only to sessions not including any search query (46.69%), but also those that present an extraordinarily large number of queries (3.24%).

A second approach to creating a frequency distribution would be to consider a session that didn't initiate any search query as being a result of a query made by a

third party (i.e., a search engines) on behalf of the user. So, from the user's point of view, the session includes a search query, even though this query hasn't been explicitly created by the user through a RI query page. Furthermore, as the number of sessions that include search queries but didn't start with an explicit search is high (47.31%), compared to the number of sessions (6.02%) that started with an explicit search query, it is reasonable to include a 'third party' query as one of the series of queries issued in user sessions. We choose to work with the second approach; it is shown in the 'adjusted' column in Table 5. A final advantage of this approach is that it allows us to easily compare ResearchIndex results with those from the CSBIB and CSTR collections.

Table 6. Frequency distribution of the number of queries issued in user sessions

No queries issued in a user session	ResearchIndex % of sessions	CSTR % of sessions	CSBIB (advanced search) % of sessions	CSBIB (simple search) % of sessions
1	46.69	43.89	35.97	29.95
2	11.51	21.95	20.02	20.43
3	7.48	12.1	12.19	12.88
4	5.42	7.76	8.51	8.46
5	3.86	4.88	5.84	5.82
6	3.10	2.90	3.83	4.22
7	2.63	1.92	2.68	3.14
8	2.07	1.53	2.13	2.35
>8	17.17	2.41	8.81	12.71

The majority of ResearchIndex sessions (74.96%) include fewer than six queries. This behavior is similar to that of CSBIB and CSTR users (Table 7). However, the RI query frequency distribution contains a far longer 'tail' of than the CSTR and CSBIB distributions (Table 6). In particular, RI sessions including between 9 and 30 queries account for 12.3% of the total number of sessions. The largest number of queries issued in a single session is 18,359—surely beyond the limits of even the most dedicated human researcher!

Table 7. Percentage of sessions including fewer than six search queries

RI % of sessions	CSTR % of sessions	CSBIB (advanced search) % of sessions	CSBIB (basic search) % of sessions
74.96	90.58	82.53	77.54

An examination of the user session lengths in minutes tells a similar story: the majority of RI sessions are relatively brief, and the distribution for sessions lasting less than 10 minutes is strikingly similar to the distributions for the CSTR and CSBIB collections. Users for all three digital libraries tend to run short sessions containing relatively few queries; presumably these users either quickly find relevant documents to satisfy their information need, or quickly decide that the digital library will not provide useful documents for this need. The exceptionally long 'tail' for the RI sessions includes a maximum session length of nearly 25 days.

Table 8. Session lengths in minutes

Number of minutes	RI sessions (%)	CSTR sessions (%)	CSBIB (advanced search) sessions (%)	CSBIB (simple search) sessions (%)
<1	46.90	29.16	47.10	44.18
1	10.57	7.59	7.39	9.07
2	6.48	5.88	4.97	5.38
3	4.87	4.81	3.41	3.68
4	3.70	4.03	2.71	2.58
5	2.89	2.87	2.04	2.05
6	2.61	3.05	1.67	1.68
7	2.08	2.60	1.54	1.29
8	1.93	2.38	1.26	1.15
9	1.50	1.99	1.23	0.91
10	1.46	2.07	1.02	0.90
10<x<30	15.00	24.19	12.37	7.79
>= 30	28.54	9.38	13.30	19.28

A manual examination of the transaction logs supports the conjecture that the majority of lengthy sessions including a large number of queries are the results of robot actions, submitting non related queries to satisfy a broad range of topics. We intend to pursue our tests to assess the validity of this conjecture or find evidence of more convincing explanations for this behavior.

6. Query Complexity

The analysis of the distribution of the number of query terms for ResearchIndex confirms previous results gathered from CSBIB and CSTR collections: user queries are short. For each collection, at least 80% of users queries contain three or fewer terms (Table 9). The average number of query terms is 2.32% in ResearchIndex queries, compared to 2.5% in the CSTR collection and 1.8% in the CSBIB collection. The distribution of the number of query terms in ResearchIndex is closer to that of CSTR collection than to that of CSBIB collection. The number of query terms in CSBIB may have been affected by a quirk in the CSBIB syntax, which enters author names as initials appended to the family name (for example, as the one term SmithJ rather than the two terms J Smith)—which will have the effect of reducing the number of query terms in many author queries. An alternative explanation is that users tend to enter more query terms when searching full text systems (such as RI and CSTR) than when searching a bibliographic database (such as CSBIB). This hypothesis is supported by many OPAC transaction log studies, which report extremely brief queries as the norm (see, for example, [6]).

The total percentage of queries including three or four queries is more evenly distributed in ResaerchIndex collection than in CSTR collection; more precisely, there are more queries with four terms in ResearchIndex. The analysis of a sample of these queries revealed that many of these queries are in the form "Lee w/2 Giles OR L w/2 Giles". In this example the query includes four terms used in combination with the union and search proximity (i.e., w/n or within *n* words) logical operators.

Table 9. Distribution of the number of query terms (RI, CSBIB simple search, CSTR)

No of terms in query	0	1	2	3	4	5	6	>6
RI	0.07%	37.02%	30.98%	12.44%	13.34%	2.16%	1.37%	2.67%
CSTR	1.59%	27.06%	34.04%	19.76%	8.98%	4.26%	2.06%	2.25%
CSBIB (simple search)	0%	52.72%	28.10%	10.8%	4.18%	1.75%	1.02%	1.41%

In all three systems the default Boolean operator is the union operator; that is, if no operator is explicitly specified in a Boolean search, then the union operator is assumed. Overall, RI users tend to use more operators than CSTR and CSBIB users (Table 10). The search proximity operator is available in the ResearchIndex system, but not in the CSTR and CSBIB interfaces. Over 9% of RI Boolean queries include at least one search proximity operator. The relative popularity of this operator is likely due to the prominent message explaining the operator, which is positioned prominently on search result pages for searches that yield no or few matches. It appears that the users are taking into account this search refinement strategy proposed by the RI interface.

Surprisingly, the union operator is explicitly included in 12.78% of ResearchIndex Boolean queries, despite being the default operator. A further analysis revealed that more than 8% of the total queries included both the union operator and search proximity operator. Note that this percentage also accounts for most of the queries

including the search proximity operator—so the bulk of the union operators are included in support of the proximity operator.

Table 10. Frequency of operators in Boolean queries

Percentage of queries containing	RI	CSTR	CSBIB
at least one intersection operator	14.73%	25.8%	14.18%
at least one union operator	12.78%	2.5%	1.69%
parentheses for compound expressions	0.75%	4.6%	0.01%
at least one proximity (NEAR) operator	9.32%	Nil	Nil

7. Conclusions

This paper examines user search behavior in the ResearchIndex digital library. This library is a significant resource for researchers and tertiary students working in computing, and it indeed sees significant usage worldwide. Usage of ResearchIndex is compared to usage in two other digital libraries intended to support this same user group: the Computer Science Technical Reports collection, and the Computer Science Bibliographies. For all three systems, user activities were logged over an extended period, to allow us to examine user behavior over time, and also to minimize the possibility that the period of study is in some way uncharacteristic.

Results from the log analysis of the RI collection indicates that RI users prefer relatively brief queries (fewer than 3 words), and relatively short search sessions (measured both in clock time and in number of queries per session). This pattern of behavior is also noted in the CSTR and CSBIB collections.

Most RI user sessions appear to have been initiated through links from general search engine result pages—indeed, only about 6% of users enter ResearchIndex through the ‘front door’ of the digital library, so to speak. The links from search engine result pages are extremely effective in bringing searchers into RI; nearly half of the sessions begin with a link from a search engine and then continue with one or more additional queries.

The RI search refinement hint about use of the proximity operator appears to be highly effective: this operator is not common in general search engines and digital libraries, but is used in one-eighth of the RI queries.

Acknowledgements

We are greatly indebted to the NEC Research Institute, Inc., who have created and maintained the ResearchIndex digital library, for providing access to ResearchIndex transaction logs and for their immense help in interpreting those raw logs. Alf-Christian Achilles has developed and maintained the CSBIB collection since 1995; he generously provided the CSBIB transaction logs described in this paper. We gratefully acknowledge Steve Jones, Roger McNab, and Stefan Boddie, who have

worked with us in earlier analysis of CSTR logs. The New Zealand Digital Library project members have inspired us with their enthusiasm and ideas. Zayed University (Abu Dhabi, UAE) provided support and resources for the second author during the final stages of this work.

References

1. Cunningham, S.J., and Connaway, L.S. Information searching preferences and practices of computer science researchers. Proceedings of OZCHI '96 (Hamilton, New Zealand) (1996) 294-299.
2. Jones, S., Cunningham, S.J., McNab, R.J., and Boddie, S. A transaction log analysis of a digital library. *International Journal on Digital Libraries* 3(2) (2000) 152-169.
3. Lawrence, S., Bollacker, K., and Giles, L.C. Indexing and retrieval of scientific literature. Proceedings of the Eighth ACM International Conference on Information and Knowledge Management (Kansas City MO, November) (1999) 139-146.
4. Lawrence, S., Giles, L.C., and Bollacker, K. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32: 6 (1999) 67-71.
5. Mahoui, M., and Cunningham, S.J. A Comparative Transaction Log Analysis of Two Computing Collections. Research and Advanced Technology for Digital Libraries: Proceedings of the 4th European Conference, ECDL (Lisbon, Portugal, Sept.) (2000) 418-423.
6. Peters, T.A. The history and development of transaction log analysis. *Library Hi Tech* 42(11:2) (1993) 41-66.
7. Samuelson, P. Legally speaking: encoding the law into digital libraries. *Communications of the ACM* 41:8 (1998) 13-18.
8. Spink, A., Bateman, J., and Jansen, B.J. Searching heterogeneous collections on the web: behavior of EXCITE users. *Information Research: an electronic journal*, 4:2 (1998). <http://www.shef.ac.uk/~is/publications/infres/paper53.html>
9. Tan, Pang-Ning, and Kumar, Vipin. Discovery of Web Robot Sessions based on their navigational patterns. Technical Report, University of Minnesota (2001). Available at <http://www.cs.umn.edu/~ptan/dmkd.ps>.