Search Log Analysis of User Stereotypes, Information Seeking Behavior, and Contextual Evaluation

Junte Zhang[1] Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Faculty of Science, University of Amsterdam

ABSTRACT

Evaluation is needed in order to benchmark and improve systems. In information retrieval (IR), evaluation is centered around the test collection, i.e. the set of documents that systems should retrieve given the matching queries coming from users. Much of the evaluation is uniform, i.e. there is one test collection and every query is processed in the same way by a system. But does one size fit all? Queries are created by different users in different contexts. This paper presents a method to contextualize the IR evaluation using search logs. We study search log files in the archival domain, and the retrieval of archival finding aids in the popular standard Encoded Archival Description (EAD) in particular. We study various aspects of the searching behavior in the log, and use them to define particular searcher stereotypes. Focusing on two user stereotypes, namely novice and expert users, we can automatically derive queries and pseudo-relevance judgments from the interaction data in the log files. We investigate how this can be used for context-sensitive system evaluation tailored to these user stereotypes. Our findings are in line with and complement prior user studies of archival users. The results also show that satisfying the demand of expert users is harder compared to novices as experts have more challenging information seeking needs, but also that the choice of system does not influence the relative IR performance of a system between different user groups.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Search process; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—Systems issues, Users issues

General Terms

Experimentation, Human factors, Measurement, Performance

Keywords

Archives, Evaluation, Search log analysis, User stereotypes

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.

1. INTRODUCTION

Evaluation is fundamental to information retrieval (IR). The dominant Cranfield paradigm [4] focuses on building reusable test collections consisting of "frozen" sets of documents, search requests, and relevance judgments. The Cranfield tests in the 1950's and 1960's have been instrumental, "almost entirely for the good" [16, pp.283], in shaping the view and developing the study of IR systems. Although it was found important to stimulate real life searches, searcher variations were eliminated in order to avoid the sticky issue of relevancy, and also by the need for statistically valid results [16]. As acknowledged by Stephen Robertson [27, pp.460], "we do not know how to simulate a real user's reactions" in a laboratory test, and he added, "operational testing is not easy either." The result is that traditional IR evaluation abstracts away from the specific task and searcher context that are crucially determining the individual searcher's satisfaction.

Instead of starting by collecting source documents with queries, it may be possible to automatically collect these data after a certain period of usage. The study of online usage of websites is done in Transaction Log Analysis (TLA), which is a methodology to "examine the characteristics of searching episodes in order to isolate trends and identify typical interactions between searchers and the system" [13, pp.410]. We can explore search patterns with implicit features that exist in the logs for information retrieval and filtering applications [7].

The domain of our case-study is the archival domain. Crucial in the digital curation of archives is to facilitate reuse, to allow each activation of an archive by a user. Therefore, archives seek to disclose their assets online through their websites, which increasingly often include a search engine. The interactions, both searching and browsing, on websites from archives are automatically logged. Archival access is increasingly shifting online from the 'bricks-and-mortar' archives, yet it often not known how well these digital archives perform, and how to improve them for their different users.

The aim of our investigation is not to study information seeking behavior as in 'browsing,' but on active and directed searching which is explictly recorded in the logs. There are many user studies in the archival domain that qualitatively examine information access to archival materials using 'electronic' finding aids [e.g., 5, 6, 30], though studies that evaluate archival access in a quantitative system-centered manner are scant [12]. Search log files could aid us in quantitatively studying the quality of online archival access, as it previously has been the case for digital libraries [22] and the World

Wide Web at large [13], so as to better understand archival users and improve (archival) information retrieval systems by evaluation. In prior research, it has been shown that search log files can be effectively used to construct domain-specific IR test collections that agree with traditional methods of evaluation [35].

This leads to the following main research question:

- *Can we use the search logs from an archive in order to study different types of users and to contextualize the evaluation for their specific needs?*

Although TLA is a very active research area [13], the archival domain has yet to be explored. We have a massive transaction log covering multiple years of traffic on a high volume site, and the complete collection of archival finding aids that is available. Subsequently, we investigate the first sub-question:

1. Can we use the transaction logs of an archive to get insight into the searching behavior of archival users?

As discussed in [7], implicit measures—this includes for example links, citations, dwell time, scrolling, and viewing—can be used to explore user interests and preferences. An interesting pointer to future work is explore individual or group differences [7]. In terms of IR evaluation [27], we could derive real users' reactions from the log for laboratory evaluation experiments. The study of [33] revealed that there are striking differences between novice and expert users in the archives.

2. Can we identify different user groups—in particular novices and experts—in the log files?

Breaking their search episodes into interaction elements (searching, browsing, etc), do these groups exhibit different information seeking behavior?

From an archival point of view, it has been stated in [5, pp.92] that most "archival information systems have been developed to meet the needs of archivists and historians." These are expert users. But how effective are these information systems really for these users? From an IR point of view, the importance of real life searches is mentioned [16], yet is is not know how to simulate the actions of real users [27], and there is a trade-off between searcher variations and statistical valid results [16]. Can we bridge the two? This leads finally to the third sub-question:

3. Can we use interaction data in the log files to contextually evaluate the IR effectiveness, specifically for novices and expert users?

Are their different needs best served by different systems? Or is still the same system best for all?

The remainder of this paper is organized with the following contributions. In Section 2, we present a literature review regarding the closest related work as outlined in this paper. First, an analysis of our search log is presented in Section 3. Second, we detail how we derive two user groups from the log file—novices and experts—using implicit measures in Section 4. Third, we focus on one aspect within the search logs, namely the online use of digital archives, and contextually evaluate their retrieval performance in Section 5 by tailoring the evaluation to the two user groups. Finally, we draw conclusions in Section 6.

2. RELATED WORK

We give a literature overview of related literature in this section. First, we describe literature about transaction log analysis in libraries and on the World Wide Web. Next, we describe the state-of-the-art of information retrieval evaluation in digital archives archives. We continue by discussing literature on user stereotyping and evaluation. Finally, we focus on IR test collections.

2.1 Transaction Log Analysis

Historically, TLA research started and "evolved out of the desire to monitor the performance of computerized IR systems [22, p.42]." In the late 1970s through the mid-1980s, TLA was applied on online public access catalogs (OPACs). The overview of [19] reflected on OPAC research in the UK, and also pointed to computer logging and transaction tape (or log) analysis. As indicated in [19], transaction logs enable the quantification of the use of an OPAC, and show the (changing) patterns of use in time, but there also limitations such as to delineating user sessions to find individual patterns to reveal real user needs.

A historical overview of TLA research in library and information science is presented in [22], where it has been pointed out that due to the development of automated IR systems in general, and transaction logging facilities in particular, TLA research gained ground. This overview shows that TLA research is extensive and diverse, with abundant published work on studies applying TLA on OPACs.

Besides system monitoring, TLA can also be conceptualized as a way to unobtrusively observe human behavior. Studies in a digital library setting have been reported in [18], which focused particularly on the queries that users entered in the system, with the proposition that the analysis can be used to finetune a system for a specific target group of users. It should be noted that currently no TLA research has been conducted yet on online digital archives.

Research on TLA in library and information science preceded the current active research in the World Wide Web, which zooms into IR by analyzing search engines [e.g., 14, 15]. An overview on search log analysis for Web searching, and a methodology, is presented in [13], which shows that literature on TLA for Web-searching is abundant. The logs can also be used to better understand how users search on the Web effectively. An example is the paper of [32], which describes a study about search logs from the three major Web search engines, where the search behavior of advanced and non-advanced search engine users is analyzed by testing which effect several search features, such as query syntax with query operators, have on query-click behavior, browsing behavior, and eventually search success.

2.2 Searching in Archival Finding Aids

The search logs from an archive could be used to evaluate online archival information access. However, currently the evaluation of archival information systems is centered around usability and interface issues, for instance, the user study conducted in [6] obtained user opinions on the content and format of interfaces in these systems. Some backtracking to user information seeking needs and search behavior gives further insight in the merits of these systems, such as with genealogists [5].

However, published research that empirically or experimentally deals with the retrieval side of intellectual access

to archival materials is practically non-existing [12]. The first study in the archival field that empirically tested different subject retrieval methods was [21]. The retrieval experiments of [9] evaluated the retrieval of archival finding aids using keywords, topical subject headings, and Boolean searches. The evaluation consisted of counting the total number of hits for a given query, and the recall in terms of the number of finding aids found in the top 100 hits. Another retrieval experiment was conducted in [31], where they pointed to the effectiveness of phrases for the retrieval of finding aids as full text HTML documents on the World Wide Web in six IR systems. Conversely, they specifically did not investigate archival finding aids represented in the standard *Encoded Archival Description* (EAD, [23]) as these were around that time not indexed by the search engines. Moreover, the context can be used to improve retrieval. For example, extra contextual information can be added by enhancing finding aids with controlled vocabulary terms [24] or metadata of the records on the collection-level [12].

2.3 User Stereotypes and Evaluation

User stereotypes are proposed in [25] as a useful mechanism for building models of individual users based on a small amount of information about them. A user stereotype must accurately characterize the users of a system in order to be useful, and are effective in optimizing the utilization of a system. It is also suggested in [8], where an IR system for social scientists was developed, that the design of IR systems can be based on user models. For evaluating IR systems, certain human characteristics—like the degree of subject knowledge or professional education—affect relevance judgments and their consistency [28]. The importance of this user context is stressed when considering relevance and people [28]. Implicit measures of user interest (e.g. links, citations, etc) can be used for IR applications [7] as these point to information about the users. This means that user stereotypes can be used for developing and evaluating IR systems.

In the archival domain, the importance of users has increased. The archives paid little attention to their users until the 1990s [5]. A user model based on genealogists, who are one of the most frequent users of archives, has been presented [5], and can be used to improve the design of digital archives. The study of [33] pointed to the difference between expert and novice archival users. Additional insight regarding novice and expert (Web) users has been presented in other user studies [3, 11].

2.4 IR Test Collections

Many ideas on evaluation in IR can be traced back to the workprocess of a librarian working with card indexes using library classification schemes [26]. The methodology for IR experimentation has been developed on this observation and has further been defined in the 1950s with the Cranfield experiments [26]. Much of the experimentation focuses on building the ideal 'test collection' (or *qrel*), i.e. compiling the set of documents that was considered relevant and had to be returned by the system [17].

A notable collaborative ('pooling') approach is the Text REtrieval Conference [26]. Another example of building a test collection for a specialized closed domain (biology) was the Cystic Fibrosis database [29] or WT10g for the Web in general [1]. In the field of Focused Retrieval, including XML Retrieval, the Initiative for the Evaluation of XML Retrieval (INEX) constructed test collections from XML files [20]. Doing IR experimentation by re-using existing test collections is also possible [26], but such a test collection does not exist yet in the archival domain.

3. ARCHIVAL SEARCH LOG ANALYSIS

In this section, we offer an analysis of the log files of the website of an archival institution. This analysis offers insight into the search behavior of archival users.

3.1 Log Files

The original ASCII transaction logs were obtained from the National Archives of the Netherlands (NA-NL). The history preserved at this institution goes back to more than 1,000 of years, preserved in archives which stretch more than 93 kilometers (or about 57 miles). It also includes maps, drawings, and photos—much of it is published on the NA-NL website (`http://www.nationaalarchief.nl`). The website provides access by offering a search engine, which includes searching in archival finding aids compiled in Encoded Archival Description (EAD, [23]), image repositories, and separate topic-specific databases.

The logs were 91.1 GB in size, with 39,818,981 unique IP-addresses, and collected from 2004 to a part of 2009.

3.2 Preparation

We use Perl to prepare the original transaction logs for analysis. We start by extracting only the clickthrough data that can be traced to the use of EAD files (archival finding aids) as we have also obtained the matching EAD files for analysis and experimentation. The URLs refer to the filename, frequently also the query terms, and occasionally include parameters like a sub(category) as subject headings.

The next step is to partition the log files into smaller subsets by identifying user sessions. A *user session* is defined in [15, p.862] from "a contextual viewpoint as a series of interactions by the user toward addressing a single information need." We define a session as a subset of n clicks from the same IP address, if and only if the difference between i and $i+1 < 30$ minutes, where i is 1 click, hence it is possible that a user has multiple sessions.

3.3 Analysis

We analyze the logs to illuminate the search behavior of archival users in general. We mainly look at the queries used, the session length, and session duration as has been done in previous studies with log files in other domains [14, 18].

3.3.1 Query Terms

We count the frequencies of all query terms (keywords). Table 1 shows the top 10 most frequently used query terms for searching in the archival finding aids. This count is interpolated over URLs that did not have a query included. That is, we assign the last known query to a hit without query. There are in total 464,932 hits with a query found in the complete log. The distribution of the query terms has a long-tail shape, which means that most users, like genealogists, entered unique keywords—mostly names—when interacting with the NA-NL system. This distribution complements previous Web search log studies [14, 18], as well as findings in archival studies [5]. We also see that the users of the system searched for the popular archives—mostly from the Dutch colonial past—of the NA-NL. At position 1

Table 1: Top 10 most popular used query strings aggregated and interpolated over each hit from 2004-2009, where the total number of (interpolated) queries is 465,089 with 50,424 unique type of queries.

Position	Query String	Count (%)
1	voc	4,383 (0.94)
2	suriname	4,277 (0.92)
3	knil	2,785 (0.60)
4	knvb	2,506 (0.54)
5	wic	1,891 (0.41)
6	hof	1,633 (0.35)
7	hof van holland	1,567 (0.34)
8	arbeidsdienst	1,541 (0.33)
9	2.10.01	1,510 (0.32)
10	drees	1,334 (0.29)

stands the archive about the *Vereenigde Oostindische Compagnie* (VOC; in English: Dutch East India Company). The users also used other acronyms as queries, such as *Koninklijke Nederlandse Voetbal Bond* (KNVB; in English: Royal Netherlands Football Association). At position 9, we see the query *2.10.01*, which is the UUID belonging to the archive of the "Ministry of Colonial Affairs (1814-1849)." These UUIDs are often used, which implies that there is known-item search, i.e. the user used the search engine as a bookmark tool.

3.3.2 Session Lengths

We explore the session length [13], i.e. the number of queries used in a session. We fine-tune this by looking at two aspects to explore the session length: (1) the number of unique queries used in a session (i.e. query revision), and (2) the number of clicks in a session. The results presented in Table 2 and 3 are grouped by frequency.

There are 78,190 sessions with a known query, while there are 194,138 sessions in total. This means that for the majority of the sessions (115,948), no query could be found, or have an 'empty query.' The majority of the sessions with a query have only one type of query (81%), while there are 538 sessions with more than 10 unique queries (see Table 2). This implies that users—even when they visit frequently—mostly search for one query during a session. But how often do users click on different results using these queries?

As Table 3 shows, in 45% of all queries, only one result has been clicked. However, for almost 15% of all queries there were more than 10 clicks on different EAD files. The former could mean that a user directly found the desired result, or discovered that further search with a query would not be effective and stopped. The latter could mean that a certain query yielded many relevant results, or the user decided to continue searching regardlessly. What does this mean for the time spent per session?

3.3.3 Session Duration and Repeated Visits

We check the time (in seconds) that was spent in a session, which we call the *dwell time*. In case of one-click sessions, this time is set to 0. Table 4 shows the distribution of the dwell time grouped per bin. Most of the sessions consisted of a interactions that consisted of just one click. In case there are more clicks, the session lasted no longer than 500 seconds (about 8 minutes). There are 2,363 instances of sessions where a user would search for more than 3,000 seconds (or 50 minutes). This distribution is similar to the ones found in previous studies [15, 18]. It can be imagined that a user continues searching after a break, so how often do users revisit and thus re-use the search engine?

Table 5 depicts the maximum number of repeated visits per user. It shows that the majority of the users searched in the archival finding aids only 1 time, and 1,391 users reused the files more than 10 times. It is interesting to note that

Table 2: Distribution of the aggregated (bins) number of unique known queries used, where the documents have been clicked in total 464,932 times in 78,190 sessions with a known query.

Queries Per Session	Session Count	
	N	%
1	63,549	(81.28)
2	9,524	(12.18)
3	2,516	(3.22)
4	941	(1.20)
5	471	(0.60)
6	229	(0.29)
7	159	(0.20)
8	113	(0.14)
9	96	(0.12)
10	54	(0.07)
> 10	538	(0.69)
	78,190	100

Table 3: Distribution of the aggregated number of documents clicked and viewed per non-empty query, where the documents have been clicked in total 465,089 times.

Clicks Per Query	Query Count	
	N	%
1	22,444	(44.51)
2	6,686	(13.26)
3	3,552	(7.04)
4	2,474	(4.91)
5	1,877	(3.72)
6	1,457	(2.89)
7	1,151	(2.28)
8	992	(1.97)
9	898	(1.78)
10	762	(1.51)
> 10	8,131	(16.13)
	50,424	100

Table 4: Session duration: distribution of the dwell time grouped per bin.

Time (s)	Count	
	N	%
0	118,564	61.07
0-500	54,109	27.87
500-1,000	9,336	4.81
1,000-1,500	4,639	2.39
1,500-2,000	2,924	1.51
2,000-2,500	1,315	0.68
2,500-3,000	888	0.46
$> 3{,}000$	2,363	1.22
	194,138	100

Table 5: Repeated visits: distribution of the maximum number of sessions and number of users (IP).

N Visits	Users		*N* Sessions
	N	%	
1	88,539	79.91	88,539
2	11,660	10.52	23,320
3	41,02	3.70	12,306
4	1,903	1.72	7,612
5	1,119	1.01	5,595
6	717	0.65	4,302
7	478	0.43	3,346
8	389	0.35	3,112
9	295	0.27	2,655
10	212	0.19	2,120
> 10	1,391	1.26	41,231
	110,805	100	194,138

41,231 sessions could be traced back to 1,391 users, so there are on average about 30 sessions per user in this group.

4. DERIVING USER GROUPS

In this section, we will try to uncover specific and interesting groups of users in the log, and analyze their information seeking behavior.

4.1 Implicit Features of User Interest

Can we identify different user groups—novices and experts—in the log files? A reasonable assumption is that archival experts—like genealogists or historians–use the archives more frequently than novice users. This supposition is supported by previous user studies [5, 33]. Hence our operational definition of "archival experience" is in terms of frequency of visits. We experiment with categorizing the interaction data extracted from the log by visit counts (i.e. number of sessions per user). The search engine of the NA-NL website presents links to archival finding aids in HTML that appear to be relevant to a query. The number of visits by a user to these finding aids suggests the amount of experience that a user has with working with the search engine.

The complete log has been processed and partitioned in sessions. We use these sessions to create 11 groups (or bins)—aggregated over all years—by the maximum session count. We pay special attention to 2 groups:

First group This group stands for *bin 1*, i.e. the set of sessions that correspond to the one-visit sessions (see Table 5).

Last group This group is *bin > 10*, i.e. the set of sessions that can be traced back to users who used the archives more than 10 times in different sessions.

We have identified and extracted the following implicit features that could point to user interest for each bin of sessions. Can we use the following implicit features to identify user groups?

Dwell time The amount of time in seconds that a user spends interacting with a system in a session, where the time-out between two interactions is set to 1800 seconds (30 minutes). A one-click session is a session with a dwell time of 0 seconds.

Table 6: Statistics about the dwell time and one-click sessions (0 dwell time) found in the log over all bins.

Bin	Dwell Time		One-Click	
	M (SD)	N	Count	%
1	105.07 (347.78)	88,539	60,341	68.15
2	179.83 (481.32)	23,320	13,551	58.11
3	218.15 (570.75)	12,306	6,815	55.38
4	256.66 (646.34)	7,612	4,071	53.48
5	271.46 (704.45)	5,595	2,933	52.42
6	262.18 (660.32)	4,302	2,343	54.46
7	301.32 (743.13)	3,346	1,730	51.70
8	279.76 (688.18)	3,112	1,727	55.50
9	288.15 (736.97)	2,655	1,441	54.28
10	265.71 (682.25)	2,120	1,173	55.33
> 10	520.80 (1,773.87)	41,231	22,436	54.42

Query Revision The number of queries used in a session. The Query Revision has a value of 0 when there are no queries found in a session.

Repeated Queries The number of times the first query of a session is repeated later in that session.

Query Length The number of terms in a query.

Deep Linking The number of times the user clicks on an anchor value that links to a part of a document.

Full-text Linking The number of times the user clicks on an anchor value that links to a full-text document.

Additionally, users of the NA-NL website searched in the archives using topical (sub)categories. Therefore, we also extract and count the instances of the use of (sub)categories. The results are shown in Tables 6, 7, 8, 9, and 10.

4.2 Results

We see that for the *first group*—set of one-visit sessions—the dwell time is on average the least. We see a clear divide in Table 6. It is the highest—on average almost 5 times as long—for sessions belonging to users who have visited more than 10 times (*last group*). We test whether there is a significant difference between the mean scores of the dwell time (independent variable) of the *first group* and *last group* using the independent samples t-test. We find a significant difference for the first group (M = 105.07, SD = 347.78) and last group (M = 520.80, SD = 1773.87; t(42713) = -47.17, $p < .000$, two-tailed). We also notice that as the number of visit count is increased, the dwell time also tend to increase. We check whether this is significant using a one-way between-groups ANOVA. We find statistical significant differences at the $p < 0.01$ level for the eleven groups according to the dwell time (F(10, 194) = 591.68).

Table 7 show the results related to the query properties: the query revision, repeated queries, and query length. Regarding the query revision and repeated queries, we again see strong differences between the first group and the last group. The former group has on average a query revision value of 0.4810, while the latter group revise the query significantly three times more often (t(42364) = -39.23, $p < .000$). Queries are not often repeated, but when they were, the group which used the archives most frequently also reused their queries most often. This is surprising, since we expected that if a query is revised less often, the same query

Table 7: Statistics about the queries found in the log over all bins.

Bin	Query Revision M (SD)	N	Repeated Queries M (SD)	N	Query Length (SD)	N
1	0.4810 (0.9404)	88,539	0.0787 (0.4100)	88,539	1.7295 (1.1629)	42,599
2	0.6907 (1.2810)	23,320	0.1414 (0.7049)	23,320	1.7727 (1.2362)	16,108
3	0.7817 (1.5929)	12,306	0.1491 (0.5799)	12,306	1.7648 (1.2575)	9,619
4	0.8543 (1.5900)	7,612	0.1797 (0.8556)	7,612	1.8630 (1.6133)	6,505
5	0.9040 (1.6048)	5,595	0.1735 (0.6818)	5,595	1.7404 (1.2587)	5,058
6	0.8459 (1.4115)	4,302	0.1690 (0.6408)	4,302	1.8063 (1.2905)	3,639
7	0.9800 (1.7542)	3,346	0.2047 (0.7084)	3,346	1.7557 (1.3177)	3,279
8	0.8969 (1.6782)	3,112	0.1951 (0.8083)	3,112	1.7191 (1.1519)	2,791
9	0.9571 (1.9779)	2,655	0.1992 (0.8294)	2,655	1.6934 (1.1308)	2,541
10	0.9557 (2.4104)	2,120	0.1632 (0.6181)	2,120	1.7493 (1.2046)	2,026
> 10	1.5468 (5.4793)	41,231	0.2113 (1.3042)	41,231	1.5400 (1.2493)	63,778

is repeated more often. Overall, fewer interaction is found in the first group than in the last group.

Interestingly, we observe that the last group used on average shorter queries. These query length values are lower than reported in a previous study on digital libraries and on the Web [18]. A reason could be the particular use of acronyms, as Table 1 depicts, which we treated as singleton queries.

The logs also recorded the navigation path between web pages. We use the search engine of the NA-NL to discover the sequence of the different types of links. The interaction flow is as follow. After the user enters a query in the search engine, an overview of the results is presented with two options.

- The first option is to click on an overview view which presents potentially relevant links to summary views (*Summary*)—these summary views link to the start of a file (*Page View*) and present contextual information (e.g. title, summary). On a Page View, users can continue the search within an EAD file by deeplinking.
- The second option is to click directly to a part of a document (*Direct To File*) and skip the Summary.

We focus on the number of times the users clicked on a deep link, or to a full-text EAD file (thus starting from the beginning). We see in Table 8 that the users clicked more often on a deep link than a full-text link. This is a feature of the EAD files, which provide access to information to a part of a document. The first group has the fewest number of clicks, whereas the last group has the most ($p < .000$). Again, we see that there is more interaction in terms of clicks coming from users who use search in the archives more frequently.

Table 9 shows that the majority of clicks link to the page views, which includes deep links. Then comes the summary views, and finally clicks within a file. This suggests that users more often start searching at the summary views—and narrow down their search by browse and click within a file—rather than clicking directly to a part of a document. This is the case for all groups. Again, we see that the number of clicks is least frequent for the first group, and the most for the last group, though the search pattern is the same.

Finally, let us focus on the use of (sub)categories by the users. A distinction between novice and expert users of archives is the search for names [33], e.g. users looking for ancestral information by using their names as query—and this happens particularly often in archives. The transition

Table 8: Average number of deep links and full-text links found in the log over all bins.

Bin	Deep Link M (SD)	Full-text Link M (SD)	N
1	2.03 (5.13)	0.74 (1.23)	88,539
2	3.10 (10.82)	0.94 (1.69)	23,320
3	3.47 (8.55)	1.07 (2.27)	12,306
4	3.88 (10.48)	1.14 (2.62)	7,612
5	4.10 (13.28)	1.17 (2.28)	5,595
6	3.93 (9.63)	1.07 (1.98)	4,302
7	4.36 (10.36)	1.28 (2.47)	3,346
8	4.11 (9.78)	1.08 (2.04)	3,112
9	4.05 (8.85)	1.16 (2.56)	2,655
10	3.87 (10.67)	1.16 (3.29)	2,120
> 10	5.36 (19.37)	2.01 (7.03)	41,231

Table 10: Use of Family and Personal names as categories.

Category	Bin 1 Rank	Bin 1 *N* (%)	Bin > 10 Rank	Bin > 10 *N* (%)
Fam. and Pers.	1	805 (18.56)	5	269 (5.30)
Pers. and Fam.	2	799 (18.42)	7	221 (4.35)

from tracing personal and organization names (novices) to a particular research project (experts) is an essential part of distinguishing both user groups as this enables more effective information retrieval [33]. Table 10 shows the rank order of the use of categories *Family and Persons* and *Persons and Families* both containing personal records. It shows that for the first group, these categories were the most popular, and used less frequently in the last group.

4.3 Novices and Experts

The results show two clearly different interaction stereotypes. On the one hand, we see a group of users which spends the least amount of time to search of all groups, has most one-click sessions, revises and repeats queries least often, clicks less often on results given their queries, and mostly seem to search for names. On the other hand, we have a group of users which spends more time to search than any other group, revises and repeated queries most often, clicks more than the other groups, and did not primarily search by looking for names. Can we match both interaction stereotypes with certain user groups?

In a previous study [11], a finding was that a user with considerable knowledge in a certain domain spends signifi-

Table 9: Average number of types of links found in the log over all bins.

Bin	Summary M (SD)	Page View M (SD)	Direct to File M (SD)	N
1	0.3655 (0.8313)	2.2966 (5.2208)	0.2079 (0.6334)	88,539
2	0.5027 (1.1739)	3.4172 (10.5553)	0.3061 (0.8072)	23,320
3	0.5791 (1.6434)	3.8861 (8.9876)	0.3489 (0.8505)	12,306
4	0.6171 (1.8375)	4.3427 (11.0165)	0.3866 (0.9099)	7,612
5	0.6272 (1.5163)	4.5458 (13.5902)	0.4152 (0.9856)	5,595
6	0.5700 (1.2754)	4.3708 (10.0514)	0.4193 (0.9924)	4,302
7	0.7113 (1.8031)	4.8715 (10.8487)	0.4441 (1.0608)	3,346
8	0.5993 (1.4655)	4.5508 (10.1379)	0.4296 (1.1041)	3,112
9	0.6354 (1.7361)	4.5168 (9.4326)	0.4648 (1.2595)	2,655
10	0.6429 (2.2847)	4.3137 (11.2965)	0.4175 (1.0109)	2,120
> 10	1.1896 (4.6813)	6.1017 (21.3702)	0.5252 (1.8338)	41,231

cantly less time to read documents in that domain. In other words, domain experts have a better performance as they search more efficiently and spend less time. However, we have to note that archival users are different than Web users with different information tasks. Archival finding aids are also complex document representations, which differ from normal web pages, particularly by the length of content and depth of document structure. Expert archival users are doing research, and have problem-solving tasks.

A study on information problem solving processes of novices and experts—e.g. identifying information needs, locate information sources, etc—revealed that experts spend significantly more time to complete a task than novices [3]. This study showed that experts would spend the maximum available time to try to solve a problem. This is a match with expert archival users, such as genealogists, who continue searching until they have found the information they needed [5].

Regarding the query properties, we see a match with a finding of [11], namely that users with little domain knowledge (novices), used longer queries than experts. A reason could be that domain experts know more effective query terms, and needed fewer terms to formulate a query. Another matching finding of [11] with our results is that experts were more inclined to select a target document for assessment than novices (see Table 9). This is also in line with [3], who found out that experts elaborate more often on the content and judge the information more often. Moreover, a similar finding is that experts process information more often than novices [11].

In summary, we can assert that the first group shares traits (or stereotypes) that can be matched with novice users. The second group can be matched with expert archival users. Moreover, our analysis in subsection 4.2 showed that there are statistical significant differences between the mean values using the implicit features as independent variables.

4.4 Correlations

We check the correlations between the variables dwell time (*DWELL*), query revision (*QREV*), deep linking (*DEEPL*), and full-text linking (*FTL*) for both groups using Pearson product-moment correlation coefficient. The correlation values with the novices are presented in Table 11, and with the experts in Table 12. These values are significant ($p < 0.01$, 2-tailed). Regarding the novices, we observe a strong correlation between dwell time and deep linking. There is medium correlation between dwell time and query revision,

Table 11: Correlation matrix with the novices, where (**) is significant at the 0.01 level (2-tailed).

	DWELL	*QREV*	*DEEPL*	*FTL*
DWELL	—	—	—	—
QREV	0.527(**)	—	—	—
DEEPL	0.625(**)	0.318(**)	—	—
FTL	0.450(**)	0.516(**)	0.252(**)	—

Table 12: Correlation matrix with the experts, where (**) is significant at the 0.01 level (2-tailed).

	DWELL	*QREV*	*DEEPL*	*FTL*
DWELL	—	—	—	—
QREV	0.837(**)	—	—	—
DEEPL	0.743(**)	0.675(**)	—	—
FTL	0.824(**)	0.933(**)	0.668(**)	—

dwell time and full-text linking, and between query revision and full-text linking. There is weak correlation between query revision and deep linking, and between deep linking and full-text linking.

Interestingly, we see strong correlations between all variables in the expert user group (see Table 12). This means that the variables are statistical dependent on each other in this group. There is very strong correlation (0.933) between query revision and full-text linking, in other words, each time a query is revised, the search procedure is in fact re-started as indicated by the full-text links in the summary and page views of the system.

In this section, we identified different groups of searchers corresponding to "novice" and "expert" stereotypes, and saw that these groups exhibit significantly different information seeking behavior: where "novices" follow a hit and run approach, the "experts" actively and interactively explore the information available. In the next section, we try to determine what is the best search system for these different searcher stereotypes.

5. IR EVALUATION IN CONTEXT

In this section, we use the interaction data of particular groups of users for contextual evaluation, trying to answer what type of system is best for their types of queries and their choice of results to inspect in detail. In the previous section, we saw significant differences between the interac-

tions of "novices" and "experts" in the archives. Are they also served best by different systems? Or is the same system best for all types of users?

5.1 Experimental Setup

We now describe retrieval experiments that use the extracted interaction data for a search log-based context-sensitive IR evaluation.

5.1.1 Collection

We study the transactions between December 31 2008 till January 31 2009—this is a month of data or 3.8 GB in size—and focus on the use of EAD files in particular. The information contained in this search log has been recorded in the *W3C Extended Logfile Format* (ELF), which includes a date, a timestamp of a hit, unique identifier for the user, the URL of the link that was visited, the query string, a browser identifier, a referral, and hits were recorded in detail within each second. Moreover, we index 4,885 EAD files—obtained from the NA-NL and mostly written in Dutch—that could be found in these logs.

5.1.2 Systems

Test collections are used for comparative system ranking, so we need several retrieval systems in order to study their similarities and differences. For our first exploration, we opt for several familiar retrieval models, with varying degrees of expected retrieval effectiveness. The system builds on the previously explained system framework [34]. We use MonetDB with the XQuery front-end Pathfinder [2] and the information retrieval module extension PF/Tijah [10]. All of our finding aids in EAD are indexed into a single main memory XML database that completely preserves the XML structure of the EAD files and allows powerful XQuery querying. We indexed the collection without stopword removal, and used the Dutch snowball stemmer.

To test the effectiveness of the different types of test collections, we use five retrieval models as dependent variables to compute the scores used in the ranking. These are controlled by using the default parameter values, the collection λ is set to 0.15, and the threshold of the ranking is set to 100.

BOOL is the Boolean model, where there is no ranking, but a batch retrieval of exact matching results. The query is interpreted as AND over all query terms, and the resulting set is ordered by document id.

LM is standard language modeling without smoothing, which means that all keywords in the query need to appear in the result.

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)} \quad (1)$$

where $n(t, q)$ is the number of times term t is present in query q.

LMS is an extension of the first model by applying smoothing, so that results are also retrieved when at least one of the keywords in the query appears.

$$P(t|d) = (1 - \lambda) \cdot P_{mle}(t|d) + \lambda \cdot P_{mle}(t|C) \quad (2)$$

where $P_{mle}(t|C) = \frac{df_t}{\sum_t df_t}$, df_t is the document frequency of query term t in the collection C.

Table 13: Properties of the test collections: number of topics and the number of relevant results.

	N **Topics**	N **Relevant**
'Novice'	1,388	1,775
'Expert'	1,701	3,053

NLLR is the NLLR or length-normalized logarithmic likelihood ratio, is also based on a language modeling approach. It normalizes the query and produces scores independent of the length of a query.

$$NLLR(d, q) = \sum_{t \in q} P(t|q) \cdot \log \left(\frac{(1 - \lambda) \cdot P(t|d) + \lambda \cdot P(t|C)}{\lambda \cdot P(t|C)} \right) \quad (3)$$

OKAPI is Okapi BM25, which incorporates several more scoring functions to compute a ranking, such as also the document length as evidence.

$$BM25(d, q) = \sum_{t \in q} IDF(t) \cdot \left(\frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \right) \quad (4)$$

where we set $k_1 = 2.0$ and $b = 0.25$. We use $IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5}$, where N is the total number of documents in the collection, and $n(t)$ is the function that counts the number of documents that contains query term t.

5.2 Filtering Assessments From User Groups

In a previous study [35], it is explained how the log files can be used to construct a massive test collection for IR evaluation, which leads to very similar system rankings as a set of humanly judged topics. But can we use interaction data in the log files to evaluate the IR effectiveness of specific user groups? The step is to construct the test collections using the subsets of sessions that have been identified, namely (1) all one-visit sessions, and (2) all sessions that can be traced back to a user with more than 10 visits. We have asserted that the former group can be related to novice users, and the latter group to archival experts.

We have large sets of queries and corresponding clicks (both from the selection of search results, and from further browsing within the selected results). We make the reasonable assumption that clicks correspond to results that a user purposefully wants to inspect in full detail, which is related to the relevance of the result (although not necessarily in a strict sense of topically relevant). In short, we treat clicks as pseudo-relevance judgments, and assume that a system that ranks "clicked" results higher is a better system. Using the two lists of sessions, we can derive two types of test collections from the log file to evaluate differentially (or in context). Table 13 shows that both test collections are large enough [17], and also that the 'expert' qrels consist of more clicks than the 'novice' variant.

5.3 Results

Tables 14 and 15 show the results of our experiments. In our evaluation, we used three IR measures. We first treat every click as a binary relevance judgment. This is captured

Table 14: System-ranking of runs evaluated against judgments from 'novices.'

	MAP	MRR	nDCG
BOOL	0.1539 (5)	0.1609 (5)	0.2532 (5)
LM	0.2615 (4)	0.2785 (4)	0.3486 (4)
LMS	0.2648 (3)	0.2815 (3)	0.3670 (3)
NLLR	0.2705 (2)	0.2872 (2)	0.3739 (2)
OKAPI	0.2791 (1)	0.2969 (1)	0.3825 (1)

Table 15: System-ranking of runs evaluated against judgments from 'experts.'

	MAP	MRR	nDCG
BOOL	0.0971 (5)	0.1113 (5)	0.1985 (5)
LM	0.2295 (4)	0.2821 (4)	0.3188 (4)
LMS	0.2333 (3)	0.2858 (3)	0.3305 (3)
NLLR	0.2343 (2)	0.2869 (2)	0.3324 (2)
OKAPI	0.2520 (1)	0.3106 (1)	0.3519 (1)

by Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, and Mean Recipropal Rank (MRR). The MRR is a static measure that looks at the rank of the first relevant result for each topic. Moreover, we can also use the number of clicks on each result by the different searchers as a form of graded relevance judgment using the Normalized Discounted Cumulative Gain (nDCG).

We observe that the MAP scores are higher when evaluating the derived set of topics from novices than experts. IR evaluation scores depend on the topic set at hand, but since we deal with a large set of representative topics, the score difference may indicate underlying differences between the two sets. This can be clarified by one of the conclusions in [28, pp.163] that the level of subject knowledge impacts the relevance judgments, i.e. "the less the subject knowledge, the more lenient are their judgments."

Test collections are used to study comparative system ranking, and we see that the system rankings are completely in line for the two groups. *OKAPI* is the best performing, *BOOL* is the worst performing run—which was expected beforehand, and LM smoothing helps the retrieval performance. The MRR scores are also higher than the MAP values, and the difference is greater with experts than novices. The MRR scores are also close between both evaluation sets. Let us focus on using the number of clicks on a result as graded relevance judgments, i.e. when users clicked more often on a result, it is treated as more relevant. This is represented in the nDCG scores, and we see the same system rankings for both groups.

For the MAP scores, we also checked for statistical significance using the paired samples t-test. We start with the results in Table 14. *BOOL* is significantly performing worst. There is a minor but significant improvement of 1.26% of *LMS* over *LM* ($t(1,387) = 4.02$, $p < 0.01$, one-tailed). There is also a significant improvement of 2.15% of *NLLR* over *LMS* ($t(1,387) = 3.13$, $p < 0.01$, one-tailed). Finally, we see an improvement of 3.18% of the Okapi system over the NLLR system on a 5% significance level. We now turn our attention to the results of Table 15. *LMS* has a significant 1.66% improvement over *LM* ($t(1,700) = 3.40$, $p < 0.01$, one-tailed). *LMS* and *NLLR* have very close MAP scores, with only a 0.43% difference, and is not significant. The best performing system *OKAPI* has a 7.55% significant improvement over the second-best performing system *NLLR* ($t(1,700) = 5.14$, $p < 0.01$, one-tailed).

In this section, we used the interaction data of particular groups of users for contextual evaluation. In the previous section, we saw significant differences between the interactions of "novices" and "experts" in the archives. There is an open debate in archival science whether the currently used systems, which are tailored to archival experts, are also suitable for novices like incidental visitors to archival web sites. Our results show that, in terms of retrieval effectiveness, the system ranking over the "experts" is identical to the system ranking over the "novices." Even though this result is limited to the options under consideration—we only explored five variants of the ranking method—it is a reassuring result. It can also be viewed as a proof of concept of the approach, and further experiments could consider other document representations (such as user tags or queries), recommendation and (pseudo-)relevance feedback, or even experiments with interface changes in the wild.

6. CONCLUSIONS

We investigated the complete search logs from an archival institution covering six years. These logs represented the full searches of archival visitors who sought online archival access. The general question is whether we can derive context from the logs. If so, how we can use this for a context-sensitive IR evaluation? Like in other domains, we can use the transaction logs to give insight into the search behavior of archival users. We looked at several generic properties that can be extracted from the logs. These are query terms, session length, and session duration. The log files were iteratively processed in order to record these statistics. Our main finding is that the logs give insight in the searches of archival users, which can be used to answer currently open questions on the effectiveness of archival access with currently available information and systems.

There is an open debate in archival science whether the currently used systems, which are tailored to archival experts, are also suitable for novices like incidental visitors to archival web sites. We experimented with the visit count of a user to group user sessions, which is the maximum number of sessions that can be traced back. Our assumption was that more experienced users use the archives more frequently than novice users. Using implicit features that point to user interest, we have observed two very different interaction stereotypes. Our assertion is that we can match these to novice and expert user stereotypes. Our main finding is that novice and expert searchers exhibit a significantly different information seeking behavior.

The results from this study helped us in constructing two test collections with each group. We can treat each click to a file—one which can (in)directly be traced back to a query—as a pseudo-relevance judgment. The system rankings over the two sets of topics corresponding to "novice" and "expert" searchers were identical. Our main finding is that, despite significantly different search episodes reflected by their specific information requests and choice of results to inspect in detail, both the experts and the novices are best served by the same type of system.

How to interpret this outcome? One explanation is that

the search log merely reflects the ranking of the operational system, and that we are in fact measuring the click-bias in the ranking. This explanation is unlikely since we validated the resulting log-based evaluation against a human judged topic set and obtained very similar system rankings [35]. We will extend the straightforward interpretation of click-as-judgments to advanced click models in future research, also linking the click model to the information seeking behavior of groups of searchers. Another explanation is that despite broad overall agreement between the two sets, there may be interesting upsets in between closely ranked systems or in other aspects than the ranking component. This still leaves the general conclusion intact that the relative effectiveness of different retrieval models is similar. Another explanation is that the groups of least and most frequent visitors don't correspond well to degrees of archival experience. There is considerable evidence from visitors to the archive, email questions, and contacts with historians and genealogists, on the existence of a large group of experienced archival users that regularly visit the web site, and are captured in the search log. Hence the separation based on visit frequency is a reasonable approximation, and helps us understand what differences in searcher competencies are affecting the system rankings, and which are not.

Acknowledgments.

We thank the reviewers for their insightful feedback and comments, Henny van Schie of the National Archives of the Netherlands for providing the data and for valuable discussion, and Henning Rode for his support on PF/Tijah. This research is supported by the Netherlands Organisation for Scientific Research (NWO) under project # 639.072.601.

REFERENCES

[1] Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for web retrieval experiments. *Inf. Process. Manage.*, *39*(6), 853–871.

[2] Boncz, P. A., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., & Teubner, J. (2006). MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine. In *SIGMOD '06*, (pp. 479–490). ACM.

[3] Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behavior*, *21*(3), 487 – 508.

[4] Cleverdon, C. W. (1967). The Cranfield tests on index language devices. *Aslib*, *19*, 173–192.

[5] Duff, W. M., & Johnson, C. A. (2003). Where is the list with all the names? information-seeking behavior of genealogists. *The American archivist*, *66*, 79–95.

[6] Duff, W. M., & Stoyanova, P. (1998). Transforming the Crazy Quilt: Archival Displays from a User's Point of View. *Archivaria*, *45*(Spring), 44–79.

[7] Dumais, S., Joachims, T., Bharat, K., & Weigend, A. (2003). SIGIR 2003 workshop report: implicit measures of user interests and preferences. *SIGIR Forum*, *37*(2), 50–54.

[8] Ellis, D. (1989). A behavioral approach to information retrieval system design. *Journal of Documentation*, *45*(3), 171–212.

[9] Feeney, K. (1999). Retrieval of archival finding aids using world-wide-web search engines. *The American Archivist*, *62*(2), 206–228.

[10] Hiemstra, D., Rode, H., van Os, R., & Flokstra, J. (2006). PF/Tijah: text search in an XML database system. In *OSIR '06*, (pp. 12–17).

[11] Holscher, C., & Strube, G. (2000). Web search behavior of internet experts and newbies. *Computer Networks*, *33*(1-6), 337 – 346.

[12] Hutchinson, T. (1997). Strategies for Searching Online Finding Aids: A Retrieval Experiment. *Archivaria*, *44*(Fall), 72–101.

[13] Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, *28*(3), 407–432.

[14] Jansen, B. J., Spink, A., Bateman, J., & Saracevic, T. (1998). Real life information retrieval: a study of user queries on the web. *SIGIR Forum*, *32*(1), 5–17.

[15] Jansen, B. J., Spink, A., Blakely, C., & Koshman, S. (2007). Defining a session on web search engines. *J. Am. Soc. Inf. Sci. Technol.*, *58*(6), 862–871.

[16] Jones, K. (1981). The cranfield tests. In K. Jones (Ed.) *Information Retrieval Experiment*, (pp. 256–284). Butterworth.

[17] Jones, K., & van Rijsbergen, C. (1976). Information retrieval test collections. *Journal of Documentation, 32*, (pp. 59–75).

[18] Jones, S., Cunningham, S. J., McNab, R. J., & Boddie, S. J. (2000). A transaction log analysis of a digital library. *Int. J. on Digital Libraries*, *3*(2), 152–169.

[19] Kinsella, J., & Bryant, P. (1987). Online public access catalog research in the united kingdom: An overview. *Library Trends*, *35*(4), 619–630.

[20] Lalmas, M. (2009). XML Information Retrieval. *Encycl. of Library and Information Sciences*.

[21] Lytle, R. H. (1980). Intellectual Access to Archives: I. Provenance and Content Indexing Methods of Subject Retrieval. *American Archivist*, *43*(Winter), 64–75.

[22] Peters, T. (1993). The history and development of transaction log analysis. *Library Hi Tech*, *42*(11), 41–66.

[23] Pitti, D. V. (1999). Encoded Archival Description: An Introduction and Overview. *D-Lib Magazine*, *5*(11).

[24] Ribeiro, F. (1996). Subject Indexing and Authority Control in Archives: The Need for Subject Indexing in Archives and for an Indexing Policy Using Controlled Language. *Journal of the Society of Archivists*, *17*(1), 27–54.

[25] Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*(4), 329–354.

[26] Robertson, S. (2008). On the history of evaluation in IR. *J. Inf. Sci.*, *34*(4), 439–456.

[27] Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Inf. Process. Manage.*, *28*(4), 457–466.

[28] Saracevic, T. (1975). Relevance: a review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, *26*(6), 321–343.

[29] Shaw, W. M., Wood, J. B., Wood, R. E., & Tibbo, H. R. (1991). The cystic fibrosis database: content and research opportunities. *Library and Information Science Research*, *13*, 347–366.

[30] Tibbo, H. R. (2002). Primarily history: historians and the search for primary source materials. In *JCDL '02*, (pp. 1–10). New York, NY, USA: ACM.

[31] Tibbo, H. R., & Meho, L. I. (2001). Finding finding aids on the world wide web. *The American Archivist*, *64*(1), 61–77.

[32] White, R. W., & Morris, D. (2007). Investigating the querying and browsing behavior of advanced search engine users. In *SIGIR '07*, (pp. 255–262). ACM.

[33] Yakel, E., & Torres, D. A. (2003). AI: Archival Intelligence and User Expertise. *The American Archivist*, *66*(1), 51–78.

[34] Zhang, J., & Kamps, J. (2009). Focused search in digital archives. In *WISE*, LNCS, (pp. 463–471).

[35] Zhang, J., & Kamps, J. (2010). A search log-based approach to evaluation. In *ECDL*, LNCS. Springer.