UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

System evaluation of archival description and access

Zhang, J.

Publication date
2011

Link to publication

Citation for published version (APA):
Zhang, J. (2011). *System evaluation of archival description and access*. [Thesis, fully internal, Universiteit van Amsterdam]. IR Publications.

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:18 Nov 2026

Chapter 6

User Stereotypes and Evaluation

Chapter 5 investigated how people use EAD finding aids by observing the information search behavior of users within archival finding aids in EAD. This chapter focuses on active and directed search, and also follows up on the whole EAD finding aids evaluation approach as outlined in Chapter 3.

6.1 Introduction

At the heart of IR systems is the idea to aide the (re)use of information by users. It is crucial that these systems are evaluated so as to find out how well they work for that purpose. The next step is to improve these systems based on IR experiments. The Cranfield tests in the 1950s have been instrumental, "almost entirely for the good" (Spärck Jones, 1981, p.283), in shaping the view and developing the study of IR systems. At its core, these laboratory tests involved searching for source documents on which the questions have been based. 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 (Spärck Jones, 1981). As acknowledged by Robertson and Hancock-Beaulieu (1992, p.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." This has profound implications for IR evaluation. How can we deal with this issue?

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 Web sites is done in Web-search Transaction Log Analysis (TLA), which is as a methodology to "examine the characteristics of searching episodes in order to isolate trends and identify typical interactions between searchers and the system" (Jansen, 2006, p.410). We can explore search patterns with implicit features that exist in the logs for information retrieval and filtering applications (Dumais et al., 2003).

In terms of archival access, it is archival description that facilitates reuse, to

allow the activation of an archival record by a user. Therefore, archives seek to disclose their assets online through their websites using archival finding aids (e.g. in EAD), which increasingly often include a search engine. The transactions (or interactions, which includes searches) on websites from archives are automatically logged. Archival access is increasingly shifting online, yet it often not known how well these search systems driven by EAD finding aids perform, and how to improve them for their different users.

In terms of research methodology, there are many user studies in the archival domain that qualitatively examine information access to archival materials using finding aids in EAD (e.g. Daniels and Yakel, 2010; Duff and Johnson, 2003; Duff and Stoyanova, 1998; Duff et al., 2010; Yakel, 2004), though studies that evaluate archival access in a quantitative system-centered manner are scant (e.g. Hutchinson, 1997). 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 (Peters, 1993) and the World Wide Web at large (Jansen, 2006), so as to better understand archival users and improve (archival) information retrieval systems by evaluation. In Chapter 3, it has been shown that log files can be effectively used to construct IR test collections that agree with traditional methods of evaluation. This leads to the following main research question:

- *Q5: Can we use a search log to study different types of users and contextualize the evaluation of their specific needs?*

Although TLA is a very active research area (Jansen, 2006), 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.

7.1 Can we use the search logs of an archive to conduct a transaction log analysis and what does it say about archival access?

As discussed by Dumais et al. (2003), 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 (Dumais et al., 2003). In terms of IR evaluation (Robertson and Hancock-Beaulieu, 1992), we could derive real users' reactions from the log for laboratory evaluation experiments. The study of Yakel and Torres (2003) revealed that there are striking differences between novice and expert users in the archives.

7.2 Can we identify novice and expert archival users in the search log files?

We break down the search episodes into interaction elements (searching, browsing, etc). Do these groups exhibit different information seeking behaviors?

From an archival point of view, it has been stated in (Duff and Johnson, 2003, p.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? Gilliland-Swetland (2001, p.210) makes a distinction between user-centric and materials-centric approaches. The assertion is that the latter approach has increasingly come under fire from users and within the archival profession itself, because it not sufficiently accomodates the needs of diverse user groups. From an IR point of view, the importance of real life searches is mentioned (Spärck Jones, 1981), yet is is not know how to simulate the actions of real users (Robertson and Hancock-Beaulieu, 1992), and there is a trade-off between searcher variations and statistical valid results (Spärck Jones, 1981). Can we bridge the two? This leads finally to the third sub-question:

7.3 Can we use interaction data in the log files to contextualize the evaluation of 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 chapter is organized with the following contributions. In Section 6.2, we present a literature review regarding the closest related work as outlined in this paper. First, an analysis of our Web search log is presented in Section 6.3. Second, we detail how we derive two user groups from the log file—novices and experts—using implicit measures in Section 6.4. Third, we contextually evaluate their retrieval performance in Section 6.5 by tailoring the evaluation to the two user groups. Finally, we draw conclusions in Section 6.6.

6.2 Related Work

We give a literature overview of related literature. First, we describe literature about transaction log analysis in libraries and on the World Wide Web. We continue by discussing literature on user stereotyping and evaluation.

6.2.1 Transaction Log Analysis

Historically, TLA research started and "evolved out of the desire to monitor the performance of computerized IR systems" (Peters, 1993, p.42). In the late 1970s through the mid-1980s, TLA was applied on online public access catalogs (OPACs), such as the work of Tolle (1983) with a study that focused on the (then) current use of OPACs with early implications for IR. The overview of Kinsella and Bryant (1987) reflected on OPAC research in the UK, and also pointed to computer logging and transaction tape (or log) analysis. As indicated by Kinsella and Bryant (1987), transaction logs enable to quantify 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 by Peters (1993), 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. For example, besides system monitoring, TLA can also be conceptualized as a way to unobtrusively observe human behavior. Fang (2007) experiments with using Google Analytics to analyze library Web sites and subsequently improve the design of these sites.

Other studies in a digital library setting have been reported by Jones et al. (2000), 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. 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. Jansen et al. (1998, 2007)). An overview on search log analysis for Web searching, and a methodology, is presented by Jansen (2006), which shows that literature on TLA for Web-searching is abundant. In the archival domain, Prom (2011) makes the case that Web analytics can be used as method to measure user actions, to understand some aspects of user behavior, and to subsequently improve online services.

6.2.2 User Stereotypes and Evaluation

Stereotypes are proposed by Rich (1979) as a useful mechanism for building models of individual users based on a small amount of information about them, which is also called a *persona* (Chapman et al., 2008). 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 by Ellis (1989), 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 (Saracevic, 1975). The importance of this user context is stressed when considering relevance and people (Saracevic, 1975). Implicit measures of user interest (e.g. links, citations, etc) can be used for IR applications (Dumais et al., 2003) as these point to information about the users. So 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 (Duff and Johnson, 2003). A user model based on genealogists, who are one of the most frequent users of archives, has been presented by Duff and Johnson (2003), and can be used to improve the design of online archival search systems. The study of Yakel and Torres (2003) points to the difference between expert and novice archival users.

Additional insight regarding novice and expert (Web) users has been presented in other user studies (Brand-Gruwel et al., 2005; Holscher and Strube, 2000).

6.2.3 Improving IR Systems and Search Behavior

Wilbur (1998, p.526) notes that it remains to be seen whether there are any simple principles that characterize human performance and which can serve to improve machine performance, or whether we must await the development of machines that can learn by accumulated experience an array of responses, essentially irreducible in its complexity, such as the one that may underlie human behavior (Wilbur, 1998, p.527). The idea to define simple principles that characterize human performance is—in a related way—done by Davis (1989).

Davis (1989) develops and validates new scales for two specific variables, perceived usefulness and perceived ease of use, which are hypothesized to be fundamental determinants of user acceptance of computers (e.g. programs and technologies). Although human performance does not equal to user acceptance of computers, it starts with it. Moreover, the more a user accepts a computer by perceiving it to be useful and of ease of use, the likely the performance will improve. Davis (1989) observes that perceived usefulness and ease of use were significantly correlated with usage behavior, usefulness had a significantly greater correlation with usage behavior than did ease of use, and usefulness could be a causal antecedent of easy of use. In other words, when a system is regarded as useful, the user may perceive it also as easy to use.

Going a step further, and in the IR context, the objective of characterizing human performance into a form of simple principles to improve machine performance is not a farfetched idea either. Ellis (1989) develops a behavioral model—which attempts to move beyond the interface but on the interaction with a database itself—based on empirically investigating the search behavior of social science researchers and uses this model to inform the design of information retrieval systems. Ellis (1989) proposes to replicate this model in other domains. Since then, the behavioral model of Ellis has been applied to compare researchers in the physical sciences with social scientists (Ellis et al., 1993), in the context of an engineering company (Ellis, 1997), or with nursing science, literature/cultural studies, history and ecological environmental science (Talja and Maula, 2003).

Koenemann and Belkin (1996) study interactive information retrieval behavior and effectiveness, and describe an experiment investigating the information seeking behavior of novice searchers who used one of four versions of an IR system to formulate routing queries for two given search topics. The independent variable is here relevance feedback, which is a query formulation support mechanism that modifies an existing query based on available relevance judgements for previously retrieved documents. The results show that relevance feedback appears to be working effectively, especially when it is made trans-

parent ('penetrable') in a system, as users needed fewer iterations to achieve results comparable to, or better than the other, less interactive, feedback conditions.

Wilson et al. (2009) propose a framework to evaluate advanced search interfaces using established information-seeking models through the quantification of the strengths and weaknesses of the interfaces in supporting user tactics and varying user conditions. Rose (2006) considers several characteristics of user search behavior, e.g. the variety of information-seeking goals, the cultural and situational context of search, and the iterative nature of the search task, and suggests ways that interfaces can be redesigned to make searching more effective for users.

Moving beyond the evaluation of interfaces, Yuan and Belkin (2010b) investigate IR support techniques for different information-seeking strategies (ISS). There are four conditions in their within-subjects experiment, where each condition is represented by an experimental system that supports a particular ISS. The findings are that it is possible to support different behaviors with different combinations of IR support techniques, and systems that support these techniques work better. Following up on that study, Yuan and Belkin (2010a) evaluate an integrated system that adapts to support different ISS, which include both scanning and searching behaviors. It is found that combining and supporting different ISS in one system works better than the baseline of a standard IR system. In fact, (Bates, 1990) already stated that users can search more effectively when search activities associated with information searching are incorporated into the system. So Bates (1990, p.590) concludes:

> "A really good information retrieval system that allows us to exercise strategic search choices quickly and easily may, in like manner, lead us to explore knowledge and research our information needs in far more powerful and creatively stimulating ways than we ever imagined in the days of the manual library or the simple online bibliographic database."

On the one hand, a system can be designed that enables search behaviors, assuming that by enabling certain interactions, a system may work better. On the other hand, off-the-shelf data on search behaviors can be (re-)used to improve systems. An instance of the latter case is the study of Beg and Ahmad (2007) that proposes a personalized Web search system by maintaining the search profile of each user, which is the basis of determing the search results of that system. The Web search quality is a component of their framework. This quality is computed by the sequence in which a user picks up the results, the time spent at those documents and whether or not the user prints, saves, bookmarks, e-mails to someone or copies-and-pastes a portion of that document. Interacting with the stored profile of each user, this integrated system is simulated and found to give improved Web search results. Likewise, in the controlled laboratory IR evaluation experiment of Agichtein et al. (2006), it is also shown that incorporating user behavior data can significantly improve the

ranking of top results in a real Web search setting compared to the baseline of the original system performance.

For archives, the research consists of studying search behaviors to improve systems. An example is the proposal of Gilliland-Swetland (2001) to enhance archival access by 'popularizing' the archival finding aid in EAD for diverse user groups by re-fitting the model of Ellis (1989). A user group are genealogists, and Duff and Johnson (2003) conceptualize an access system driven by finding aids based on interviews. No research exists in the archival domain that formally and quantifiably analyzes archival search behavior based on off-the-shelf interaction data derived from search logs that comprise of 'real' usage of archival finding aids in EAD, and conducts contextualized system evaluation based on that.

6.3 Archival Search Log Analysis

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

6.3.1 Preparation

We use Perl to prepare the original logs for analysis. We start by extracting only the click-through data that can be traced to the use of EAD finding aids as we have also obtained the matching EAD finding aids 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 (Jansen et al., 2007, 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.

6.3.2 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 (e.g. Jansen et al., 1998; Jones et al., 2000).

Origin of Users

Table 6.1 gives an overview of the origin of all users who used the *Nationaal Archief* Web site, where we only list the origin of the IPs that could be traced back. It states the official short names in English as given in ISO 3166-1 and

the corresponding ISO 3166-1-alpha-2 code elements.[1] Not surprisingly, we see that Dutch users top the list. American users are the second most popular group, perhaps due to the number of Dutch emigrants and their offspring.[2] We observe that users coming from the former Dutch colonies Surinam and Indonesia are in the third and fifth position respectively.

Query Terms

We count the frequencies of all query terms (keywords). Table 6.2 shows the top 10 most frequently used query terms for searching in the archival finding aids in EAD. 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 441,575 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 *Nationaal Archief* system. This distribution complements previous search log studies (Jansen et al., 1998; Jones et al., 2000), as well as findings in archival studies (Duff and Johnson, 2003). We also see that the users of the system searched for the popular archives of the *Nationaal Archief* —mostly from the Dutch colonial past. At position 1 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 8, 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.

Session Lengths

We explore the session length (Jansen, 2006), 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 are presented in Table 6.3 and 6.4 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 6.3). This implies that users—even

[1] Retrieved 2011/04/14 from `http://www.iso.org/iso/english_country_names_and_code_elements`

[2] According to the US census of 2006 (`http://factfinder.census.gov/servlet/ADPTable?_bm=y&-geo_id=01000US&-ds_name=ACS_2006_EST_G00_`), there are 5,187,864 Americans with Dutch roots, which accounts for 1.6% of the US population.

Table 6.1: Origin of the users of the *Nationaal Archief* Web site.

Origin	Count	GR	6,176	LA	663	SC	97
NL	7,837,158	FI	5,614	JO	654	GA	85
US	2,739,877	CO	4,665	RW	640	AD	85
SR	335,818	HK	4,550	MK	640	MC	72
BE	181,473	MX	4,452	CI	607	MD	67
ID	132,745	LV	4,262	BS	593	ZW	66
GB	118,259	CR	4,104	GE	566	MP	59
AU	105,590	VE	3,807	IS	538	BI	59
IN	80,445	TN	3,351	DO	537	HN	57
DE	78,529	KW	3,265	MO	465	GU	55
RU	75,603	QA	3,145	TT	444	LI	55
CN	70,673	EG	2,641	NC	443	GM	54
TR	67,974	LU	2,570	BO	441	PG	43
JP	63,082	KZ	2,307	ME	439	CG	41
FR	63,044	PR	2,166	GT	434	DJ	39
CA	62,775	LT	2,166	AZ	399	KG	39
SE	52,973	BH	2,113	SY	397	MQ	37
ZA	49,069	MA	2,065	FO	393	BM	35
CH	41,446	PK	2,060	TZ	377	CD	33
SG	32,504	BY	1,965	PF	363	VU	32
AN	32,148	BG	1,949	BJ	353	MR	32
IT	31,718	OM	1,846	IQ	350	YU	32
ES	30,361	BB	1,741	AO	346	AS	31
KR	28,355	SK	1,657	ML	340	ZM	30
AE	25,407	PA	1,553	SD	322	SL	26
MY	24,912	GH	1,527	MM	315	**	25
BR	23,302	HR	1,525	JM	306	SO	25
AT	19,328	MV	1,520	AP	304	TJ	24
UA	18,842	AG	1,357	CU	295	TO	24
PT	18,107	NA	1,336	AW	292	AX	24
IL	17,791	EC	1,328	BW	251	BZ	21
LK	17,702	CY	1,314	AL	251	GQ	16
IE	16,760	NG	1,287	CM	248	TC	15
NO	16,744	BD	1,188	TG	234	HT	12
CZ	12,776	LB	1,173	CV	221	CK	10
TH	12,365	MU	1,148	UZ	209	GN	8
EU	11,446	SN	1,029	AF	202	VA	8
RO	11,296	SI	967	UG	200	GY	8
PH	10,551	KE	941	AM	196	PW	7
PL	9,389	EE	935	UY	186	BT	7
NZ	8,931	BN	913	NP	174	TD	6
DK	8,292	PY	855	LY	164	AI	5
CL	7,859	PS	843	MW	161	WS	4
AR	7,566	YE	823	MG	156	LS	4
DZ	7,559	SV	817	MN	150	NF	4
SA	7,424	ET	800	GI	149	CF	3
TW	7,176	MZ	789	VI	143	SM	3
IR	7,061	MT	751	KY	119	NE	2
HU	6,551	BA	730	FJ	119	IO	2
PE	6,507	KH	691	BF	100	FK	1
VN	6,190	RS	671	NI	98	GP	1

Table 6.2: 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)

Table 6.3: 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

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 6.4 shows, in almost 45% of all queries, only one result has been clicked. However, for more than 16% 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?

Table 6.4: 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 6.5: 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

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 6.5 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 (e.g. Jansen et al., 2007; Jones et al., 2000). It can be imagined that a user continues searching after a break, so how often do users re-visit and thus re-use the search system?

Table 6.6 depicts the maximum number of repeated visits per user. It shows

Table 6.6: Repeated visits: distribution of the maximum number of sessions and number of users (IP), which shows that the majority of users have interacted in one session.

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

that the majority of the users searched in the archival finding aids in EAD only one time, and 1,391 users reused the files more than ten times. It is interesting to note that 41,231 sessions could be traced back to 1,391 users, so there are on average about 30 sessions per user in this group.

6.4 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.

6.4.1 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 (Duff and Johnson, 2003; Yakel and Torres, 2003). 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 system of the *Nationaal Archief* website presents links to archival finding aids in EAD in a HTML interface that appear to be relevant to a query. The number of visits by a user to these finding aids in EAD suggests the amount of experience that user has with working the search system.

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 two groups:

First group This group stands for *bin 1*, i.e. the set of sessions that correspond to the single-visit sessions (see Table 6.6).

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 2 interactions is set to 1,800 seconds (30 minutes). A one-click session is a session with a dwell time of 0 seconds.

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 *Nationaal Archief* website searched in the archives using topical (sub)categories. Therefore, we also extract and count the use of (sub)categories. The results are shown in Tables 6.9, 6.10, 6.11, 6.12, and 6.13.

6.4.2 Vocabulary Use

We also look at the vocabulary use of the first group and last group in order to see the differences. When users clicked on an EAD finding aid without having used a query in the session, it is represented as '—'. We see that this the most popular in the last group, which suggests that the last group browses more. Both *suriname* (Surinam) and *voc* (Dutch East India Company) are the most popular used queries in both groups. We see that the last group makes more use of know-item search when they retrieve an EAD finding aid using the number of that finding aid (e.g. *2.10.01*), which is not known unless the user already knows it. So we see similarities and differences in the search requests between both groups.

6.4.3 Results

We see that for the *first group*—set of single-visit sessions—the dwell time is on average the least. We see a clear divide in Table 6.9. It is the highest—on average almost 5 times as long—for sessions belonging to users who have

Table 6.7: Top 10 most popular used query strings of the first group ($N = 14,368$).

Query String	Count (%)
suriname	295(2.05)
voc	293(2.04)
knil	202(1.41)
nsb	85(0.59)
knvb	73(0.51)
arbeidsdienst	69(0.48)
wic	61(0.42)
—	60(0.42)
oorlog	60(0.42)
drees	59(0.41)

Table 6.8: Top 10 most popular used query strings of the last group ($N = 14,866$).

Query String	Count (%)
—	719(4.84)
suriname	506(3.40)
voc	295(1.98)
hof van holland	222(1.49)
curacao	198(1.33)
staten generaal	197(1.33)
2.10.01	181(1.22)
2.10.02	180(1.21)
1.04.02	178(1.20)
drees	153(1.03)

Table 6.9: 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

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 6.10 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

Table 6.10: Statistics about the queries found in the log over all bins.

Bin	Query Revision		Repeated Queries		Query Length	
	M (SD)	N	M (SD)	N	(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

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 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 the World Wide Web (Jones et al., 2000). A reason could be the particular use of acronyms, which we treated as singleton queries.

The logs also recorded the navigation path among different EAD finding aids. We use the search system of the *Nationaal Archief* to discover the sequence of the different types of links. The interaction flow is as follow (also see page 37). After the user enters a query in the search system, an overview of the results is presented with 2 options.

1. 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.
2. 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 6.11 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. Again, we see that there is more interaction in terms of clicks coming from users who search in the archives more frequently.

Table 6.11: Average number of deep links and full-text links found in the log over all bins.

Bin	Deep Linking M (SD)	Full-text Linking 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 6.12 shows that the majority of clicks link to the page views, which includes deep links. Then comes the summary views, and finally *Direct To File* clicks. This suggests that users more often start searching at the summary views—and narrow down their search by browsing and clicking within a file—rather than using *Direct To File* links. 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.

Table 6.12: 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

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 (Yakel and Torres, 2003), e.g. users looking for ancestral information by using their

names as query—and this happens particularly often in archives. The transition 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 (Yakel and Torres, 2003). Table 6.13 shows the rank order of the use of categories *Families and Persons* and *Persons and Families*. Both are in fact the same, but this division may be due to an artifact of the system. It shows that for the first group, these categories were the most popular, and used less frequently in the last group. This is another indication that the last group exhibits traits of expert users, as experts move away from searching for names, and to a particular research project.

Table 6.13: On the use of Family and Personal names as categories, which shows that these categories were most popular for first group, whereas the last group used other category subjects more frequently.

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

6.4.4 Novices and Experts

The results show two clearly different interaction stereotypes. On the one hand, we see a group of users which spends least amount of time 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 types of interaction stereotype with certain user groups?

In a previous study (Holscher and Strube, 2000), a finding was that a user with considerable knowledge in a certain domain spends significantly 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 in EAD 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, and use the document structure to search for relationships.

Lazonder et al. (2000) examine differences between novice and experienced users in searching information on the World Wide Web. They state that seeking information on the Web is similar to the work of a detective. It involves tracing

relevant information, one has to ask the right questions, consult proper sources of information, and creatively combine search outcomes (Lazonder et al., 2000, p.576). Their controlled experiment uses tasks and questionnaires. The participants are divided in two groups, experts and novices, using a set of questionnaires. In terms of finding relevant Web sites, their findings show that experts perform tasks faster and more successfully with fewer interaction than novices. In terms of finding information in Web sites, there is no significant difference.

However, 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 (Brand-Gruwel et al., 2005). 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 (Duff and Johnson, 2003).

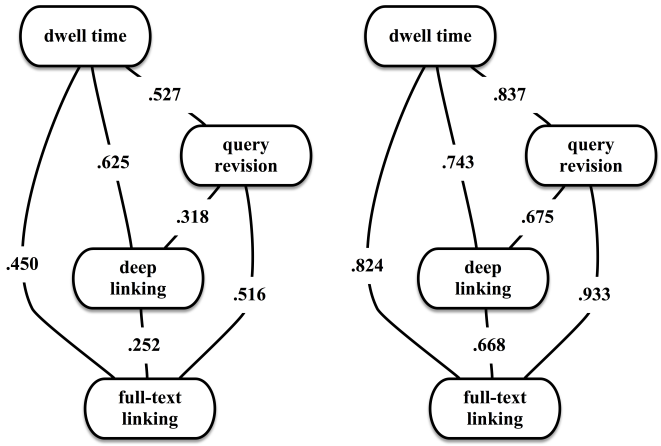
Regarding the query properties, we see a match with a finding of Holscher and Strube (2000), namely that users with little domain knowledge (novices), used longer queries than experts. A reason could be that domain experts know more effective query terms, and needed fewer terms to formulate a query. Another matching finding of Holscher and Strube (2000) with our results is that experts were more inclined to select a target document for assessment than novices (see Table 6.12). This is also in line with the results of Brand-Gruwel et al. (2005), 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 (Holscher and Strube, 2000).

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 6.4.3 showed that there are statistical significant differences between the mean values using the implicit features as independent variables.

6.4.5 Correlations

We check the correlations between the dwell time, query revision, deep linking, and full-text linking for both groups using Pearson product-moment correlation coefficient. The correlation values with the novices are presented in Figure 6.1(a), and with the experts in Figure 6.1(b). 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, 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 Figure 6.1(b)). This means that the variables are statistical de-



(a) Novice group. (b) Expert group.

Figure 6.1: Correlation ($p < 0.01$) between the variables in both the novice and expert groups: dwell time, query revision, deep linking and full-text linking.

pendent on each other in this group. There is very strong correlation (0.933) between query revision and full-text linking, which may be due to the fact that 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.

6.5 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 interactions 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?

6.5.1 Experimental Setup

Next, we describe retrieval experiments that use the extracted interaction data for a search log-based context-sensitive IR evaluation. We study the transactions between December 31 2008 till January 31 2009—this is a month of data—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.

The log file—of the month January 2009—is 3.8 GB in size, and after filtering transactions of only EAD hits, the object of study in our experiments is 304.4 MB of data. Moreover, we index 5,934 EAD files—937 MB of data obtained from the *Nationaal Archief* —that could be found in the log. The experimental setup is replicated from the study as described in Chapter 3.

6.5.2 Filtering Assessments from User Groups

In Chapter 3, it is explained how the log files can be used to construct a massive and reliable test collection for IR evaluation. 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 single-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 judgements, and assume that a system that ranks "clicked" results higher is a better system. Using the two lists of sessions, we can derive 2 types of test collections from the log file. Table 6.14 shows that both test collections are large enough (Spärck Jones and van Rijsbergen, 1976).

Table 6.14: Properties of the test collections: number of topics and the number of relevant results.

	#Topics	#Relevant
Novice	1,374	1,758
Expert	1,687	3,027

6.5.3 Results

Figure 6.2(a) shows the line graph of the 'novice' group. Figure 6.2(a) shows the line graph of the 'expert' group. Tables 6.15 and 6.16 further summarize the results of our experiments. In our evaluation, we used three IR measures. We first treat every click as a binary relevance judgement. This is captured by

Table 6.15: System-ranking of runs evaluated against judgments from 'novices.'

	MAP	MRR	nDCG
BOOL	0.1716 (5)	0.1791 (5)	0.2686 (5)
LM	0.2753 (4)	0.2935 (4)	0.3607 (4)
LMS	0.2763 (3)	0.2941 (3)	0.3761 (3)
NLLR	0.2817 (2)	0.2995 (2)	0.3827 (2)
OKAPI	0.2912 (1)	0.3102 (1)	0.3920 (1)

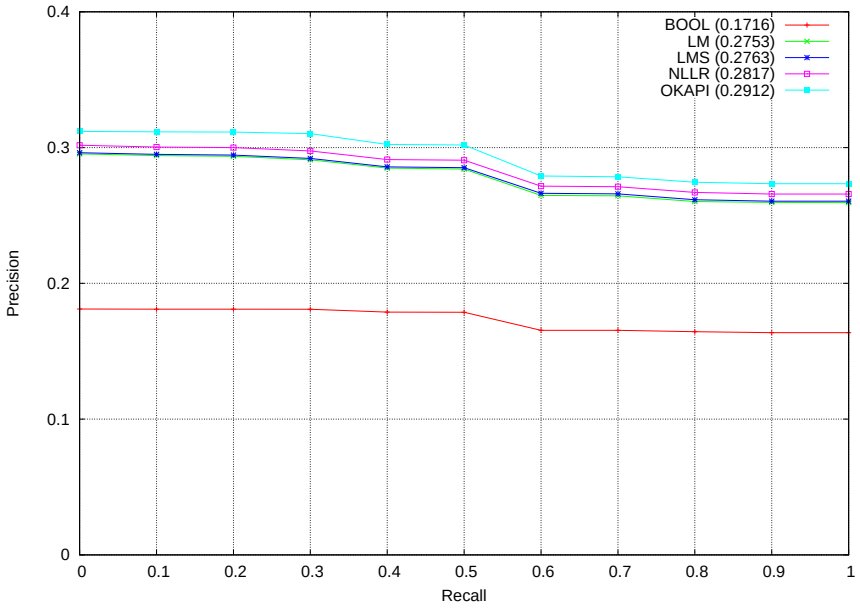
Table 6.16: System-ranking of runs evaluated against judgments from 'experts.'

	MAP	MRR	nDCG
BOOL	0.1070 (5)	0.1217 (5)	0.2041 (5)
LM	0.2436 (4)	0.2982 (4)	0.3272 (4)
LMS	0.2480 (3)	0.3025 (3)	0.3398 (3)
NLLR	0.2492 (2)	0.3037 (2)	0.3419 (2)
OKAPI	0.2666 (1)	0.3272 (1)	0.3616 (1)

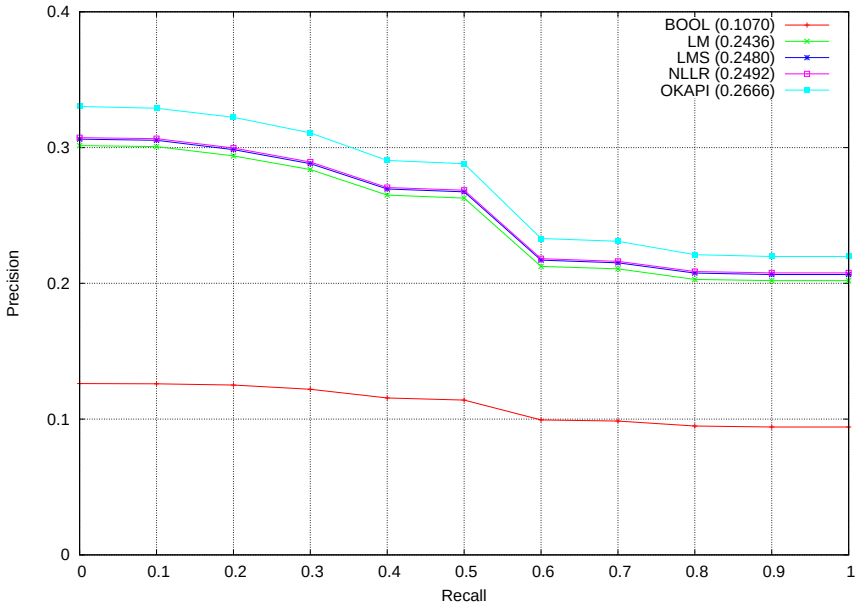
Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, and Mean Recipropal Rank (MRR). The MRR is a static measure that looks at the rank of the first relevant result for each topic, and focuses on precision. Moreover, we can also use the number of clicks on each result by different results 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. This can be clarified by one of the conclusions of Saracevic (1975, p.163) that the level of subject knowledge varies inversely with the number of relevant judged documents, i.e. "the less the subject knowledge, the more lenient are their judgements."

We take notice 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, the difference is greater with experts than novices, and the scores are also close between both evaluation sets. We check for statistical significance using the paired samples t-test. We start with Table 6.15. We focus on the MAP scores for the significance tests. BOOL is significantly performing worst. There is a minor but significant improvement of 0.36% for LMS over LM (t(1373) = 3.81, $p < 0.0005$, one-tailed). There is also a significant improvement of 1.95% of NLLR over LMS (t(1373) = 3.00, $p < 0.005$, one-tailed). We also see a significant improvement of 3.37% of OKAPI over NLLR ($p < 0.01$, one-tailed). Next, we focus on the results of Table 6.16. LMS has a significant 1.81% improvement over LM (t(1686) = 3.78, $p < 0.0005$, one-tailed). The difference between LMS and NLLR is only 0.48%,



(a) Precision and recall graphs for the 'novices.'



(b) Precision and recall graphs for the 'experts.'

Figure 6.2: Precision and recall line graphs.

but significant ($p < 0.05$). OKAPI has a 6.98% improvement over NLLR, and this was also significant ($p < 0.0005$). These findings are the same as compared to our retrieval experiments in Chapter 3.

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 that is best for the "experts" is also the best system for the "novices." Even though this result is limited to the options under consideration—we only explored 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, or even experiments with interface changes in the wild.

6.6 Conclusion

We have investigated the complete Web search logs from the *Nationaal Archief* from 6 years. These logs represented the full searches of archival visitors who sought online archival access with EAD finding aids. 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?

Our results show that we can use the search logs to give more insight into the search behavior of archival users, and we have looked at several generic properties that can be extracted from the logs. These are query terms, session length, and session duration. The main finding is that the search log gives insight in the searches of archival users, which can be used to answer the currently open question on the effectiveness of archival access to 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. Our assumption was that more experienced users consult the EAD finding aids 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 helped us in constructing a test collection for 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 scores for experts were lower than for experts, but the system rankings were the same for the test collections. 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 system.

To conclude, the results show that the same IR system is working the best for two sets of relevance judgments coming from two different pools of users. The

MAP scores were considerable lower when evaluated using the set of relevance judgments. This could mean that the topics coming from these users are harder, i.e. to deliver the relevant results. This explains the higher dwell time. On the one hand, expert archival users search exhaustively for information, and it would benefit them to improve the IR effectiveness of a system in terms of the MAP scores. However, for the early precision, the difference between both groups is minor using the same systems. This implies, on the other hand, that we could also avoid complicated system personalization issues to fine-tune the IR effectiveness of a system for different users—and use one archival search system for everyone.