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 3

Construction of an Archival Test Collection for Evaluation

This chapter investigates how to evaluate the IR effectiveness of search systems driven by EAD finding aids like the README system in Chapter 2. We have the system, but we do not know how well it works. We restrict our attention in this chapter to the evaluation of the document ranking approach. In Chapter 4, we also conduct evaluation on the retrieval of descriptions within an EAD finding aid by testing the element relevance ranking and aggregation-based approaches.

3.1 Introduction

Studies can investigate the user needs in advance, such as the study of Yakel (2004), or in conjunction with a system. When the latter is the case, there is also a form of post hoc performance measurement of that system. Fox and Sornil (1999) think that the serious challenge for digital libraries, such as archival information systems, is to have usable systems. Testing the usability of a system is also evaluating a system. An example is the focus group study of Duff and Stoyanova (1998) that investigates the content and user interfaces of archival information systems, and finds suggestions to design more usable displays.

Increasingly more common, the evaluation of systems deals with IR system performance measurements. Robertson (2008) describes early experiments that were formative in the understanding of IR. The Cranfield experiments in the 1950s introduced empirical experimentation to assess the effectiveness of several indexing schemes, and defined the methodology of IR experimentation. The method evolved from known-item search—start from a known document and formulate a question to which that document is a suitable answer—to include judgements of relevance with the aim to find all relevant documents in the collection. The Cranfield experiments measured information retrieval performance with metrics, which later was to become known as *precision* and *recall*. Robertson and Hancock-Beaulieu (1992) also point more in detail to method-

ological issues, such as laboratory and operational experiments. While the former deals with control over variables, observability, and replicability, the latter deals with realism. Keeping the balance between both methodological approaches is difficult (Robertson and Hancock-Beaulieu, 1992, p.460).[1]

So in IR, the dominant approach to evaluation uses a test collection: a set of documents, a set of search requests (topics), and a set of relevance judgments for each topic (*qrels*). Researchers create such test collections collaboratively on generic (artificial) set of documents, such as newspaper corpora or Wikipedia, which are useful to study generic aspects of the algorithms that drive search systems (i.e. retrieval models). Such test collections provide part of the answers, but fail to address the unique collection and types of search requests of an individual archive. Creating a test collection with this conventional Cranfield approach, for each archive, is simply too expensive in time and effort.

A major challenge for search systems driven by EAD finding aids, is how to evaluate the information retrieval effectiveness given the domain-specifity of their collections, their particular user groups, and how to use this crucial evaluation step to improve a system. We propose a search log-based approach to the IR evaluation of systems driven by EAD finding aids. Nowadays, almost every EAD finding aid system is Web-based, and the interaction between the system and the user is logged in so-called search logs, often hidden deeply away or primarily used to generate descriptive statistics about the general Web traffic of a website. This includes the search requests that have been entered by users, and what and where it was clicked, and so on. Is it reasonable to assume this interaction data can be reused for automated system evaluation? This results in this chapter's research question:

- *Q2: Can we use an online archive's search log to derive a domain-specific test collection?*

We deal with this research question with the following sub-questions:

2.1 How can we derive a test collection from an archival search log?
2.2 How can we validate a log-based test collection?

The clicks of users in the search log may have been greatly influenced by the system that they have used.

2.3 Is there a bias from the original system when using the log-based test collection?

The envisioned test collection is tailored to the archive at hand, representative to both its document collection and its search requests. As a test collection,

[1] Tague-Sutcliffe (1992) also points to methodological issues for carrying out an IR experiment, such as determining the need for testing, choosing the type of test (laboratory or operational), defining the variables, collecting and analyzing the data, and presenting the results.

it can be (re)used for comparative testing under the same experimental conditions. Performance is topic-dependent and this avoids comparing over different topic sets. We apply this approach to a particular domain-specific collection of documents, in a special genre, namely archival finding aids in EAD that describe the archives of persons, families, and corporations.

The remainder of this chapter is structured as follows. Section 3.2 describes the related work. We explain the interaction occurring in the *Nationaal Archief* EAD system and subsequently derive a domain-specific test collection from this search log in Section 3.3, using a range of retrieval models for further testing in Section 3.4. In order to validate the log-based evaluation, we construct a set of topics judged by human experts—based on a set of e-mail reference questions sent to the *Nationaal Archief* and responses from their archivists. The results are analyzed and discussed in Section 3.5. Moreover, we derive a test collection from a research guide compiled by the *Nationaal Archief* and repeat the IR experimentation in Subsection 3.5.4. Finally, we conclude with our main findings and discuss pointers to future research in Section 3.6.

3.2 Background and Related Work

In this section we discuss three strands of related work: search log analysis for IR evaluation, IR test collections, and archival (metadata) retrieval evaluation.

3.2.1 Log Analysis for Information Retrieval Evaluation

Historically, the analysis of log files started and "evolved out of the desire to monitor the performance of computerized IR systems" (Peters, 1993, p.42). The focus has been to analyze how systems are used. Besides system monitoring, it can also be conceptualized as a way to unobtrusively observe human behaviors. Studies in a digital library setting have been reported in (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, but it did not specifically investigate the IR effectiveness.

Research on log analysis in library and information science precedes the research in the World Wide Web, where the latter zooms into IR by analyzing search engines. An overview on search log analysis for Web searching, and a methodology, is presented in (Jansen, 2006), which shows that literature on log analysis for Web-searching is abundant. The logs can be analyzed to better understand how users search on the Web effectively (also see Chapters 5 and 6 of this dissertation for EAD finding aids). An example is the paper of White and Morris (2007), which describes a study about search logs, where the search behavior of advanced and non-advanced users is analyzed by testing the effects of query syntax with query operators on query-click behavior, browsing behavior, and search success.

There has been substantial interest in using click-through data from search logs as a form of implicit feedback (Dumais et al., 2003). A range of implicit feedback techniques have been used for query expansion and user profiling in information retrieval tasks (Kelly and Teevan, 2003; Kim et al., 2001). Joachims et al. (2005, p.160) conclude that "the implicit feedback generated from clicks shows reasonable agreement with the explicit judgments of the pages." There is active research on building formal models of interaction from logs to infer document relevance (Dupret and Liao, 2010).

3.2.2 IR Test Collections

IR evaluation can be traced back to the work-process of a librarian working with card indexes using library classification schemes (Robertson, 2008). The basic methodology for IR experimentation has been developed in the 1950s with the Cranfield experiments, focusing on retrieval effectiveness by the comparative evaluation of different systems (then indexing languages, nowadays retrieval algorithms). Much of the experimentation focuses on building a test collection consisting of a document collection, a set of topics and judgments on which documents are relevant for each topic (Spärck Jones and van Rijsbergen, 1976).

Test collections can be reused by evaluating new or adapted systems or ranking algorithms under the exact same experimental conditions. A notable collaborative ('pooling') approach is the Text REtrieval Conference (Voorhees and Harman, 2005). There are test collections for a variety of domains. Examples included the Cystic Fibrosis database in the medical domain (Shaw et al., 1991), and WT10g for the Web in general (Bailey et al., 2003). In the field of XML retrieval, the Initiative for the Evaluation of XML Retrieval (INEX) constructed test collections based on XML files (Gövert and Kazai, 2002b; Lalmas, 2010). Creating a traditional test collection requires substantial (collaborative) human effort in topic creation and relevance judgments (domain knowledge), and such a test collection does not exist yet in the archival domain.

3.2.3 Evaluating Archival Metadata Retrieval

Published research that empirically or experimentally deals with the evaluation of archival metadata retrieval is scant (Hutchinson, 1997). Experiments that specifically examine the retrieval performance potential of archival finding aids in specifically EAD is almost non-existent, despite the emergence of EAD in 1997 (Pitti, 1999) and its increasing adoption and popular use in archives.

The first study in the archival field that empirically tested different subject retrieval methods was Lytle (1980). Subsequently, there were a few studies that tested the effects of some external context knowledge on retrieval, such as controlled vocabulary terms (Ribeiro, 1996) or document-collection granularity (Hutchinson, 1997). Hutchinson (1997) outlines and tests several strategies

for searching in EAD finding aids, namely as a whole unit, only the introductory components of a finding aid before the actual inventory descriptions, introductory components enhanced with controlled access terms extracted from the catalogue records, and catalogue records only. The retrieval of online archival finding aids (not in EAD) have been examined in the study of Feeney (1999) by counting the number of finding aids returned by search engines using different types of query formulations, i.e. keyword, phrase, and Boolean searches using the topical subject and names headings as queries. The retrieval experiments of Tibbo and Meho (2001) on finding aids as full text HTML documents on the World Wide Web pointed to the effectiveness of phrases for the retrieval of finding aids (not in EAD). Moving away from the system-centered view, Daniels and Yakel (2010) present a user-centric experiment, where they observe what users do when they search, how they search, and how often their recall-oriented search was succesful.

3.3 From Search Log to Test Collection

In this section, we study how to derive a test collection from a search log. We look at the search interaction possible with the system of the *Nationaal Archief* and use its search log to create domain-specific test collections for IR evaluation, tailored to their EAD finding aids collection and its users.

3.3.1 *Nationaal Archief* System Interaction

The flow charts in Figure 3.1, 3.2, and 3.3 depict the work-flow of the *Nationaal Archief* system as of 2009, and we have numbered each step possible during that work-flow. This work-flow consists of starting the Web site's frontpage **(1)**, then moving to search in **(2)** by formulating a query (in this example query *Juliana*, the name of a former Dutch queen, 1909-2004) and entering it in the search box. The NA system returns found results in **(3)** from different sources, such as EAD finding aids, the image repository, several thematic databases, and remaining resources. We zoom in on the list of EAD finding aids in **(4)**. From this point, the actual search interaction starts. Figure 3.2 on page 39 and Figure 3.3 on page 40 show divergent search paths stemming from **(4)**.

1. The first search path as illustrated in Figure 3.2 is an example of full-text (or document) search, where the returned results in **(5a)** consist of links to the archive overview of an EAD finding aid in **(6a)**, and then to the beginning of an EAD finding aid in the interface of **(7a)**. The user has to start to use the table of contents on the left side of the screen to browse through the whole EAD finding aid.
2. The second search path in Figure 3.3 is an example of focused (or sub-document) search, where a user can directly go to an archival description of a component in an EAD finding aid by clicking on it in **(5b)**, and then views the

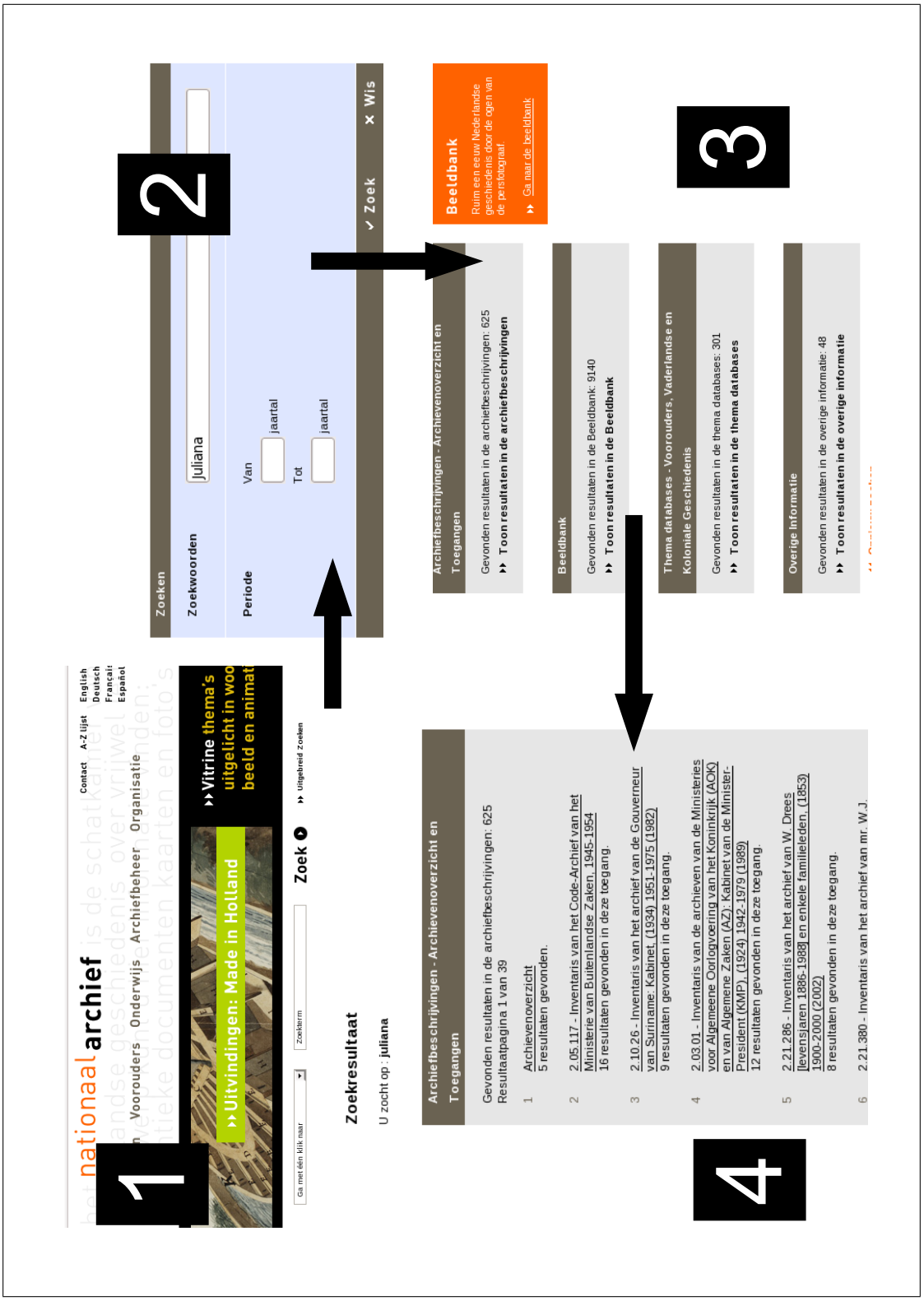

Figure 3.1: *Nationaal Archief* System Interaction (1). Image courtesy of the *Nationaal Archief* (rotated clockwise).

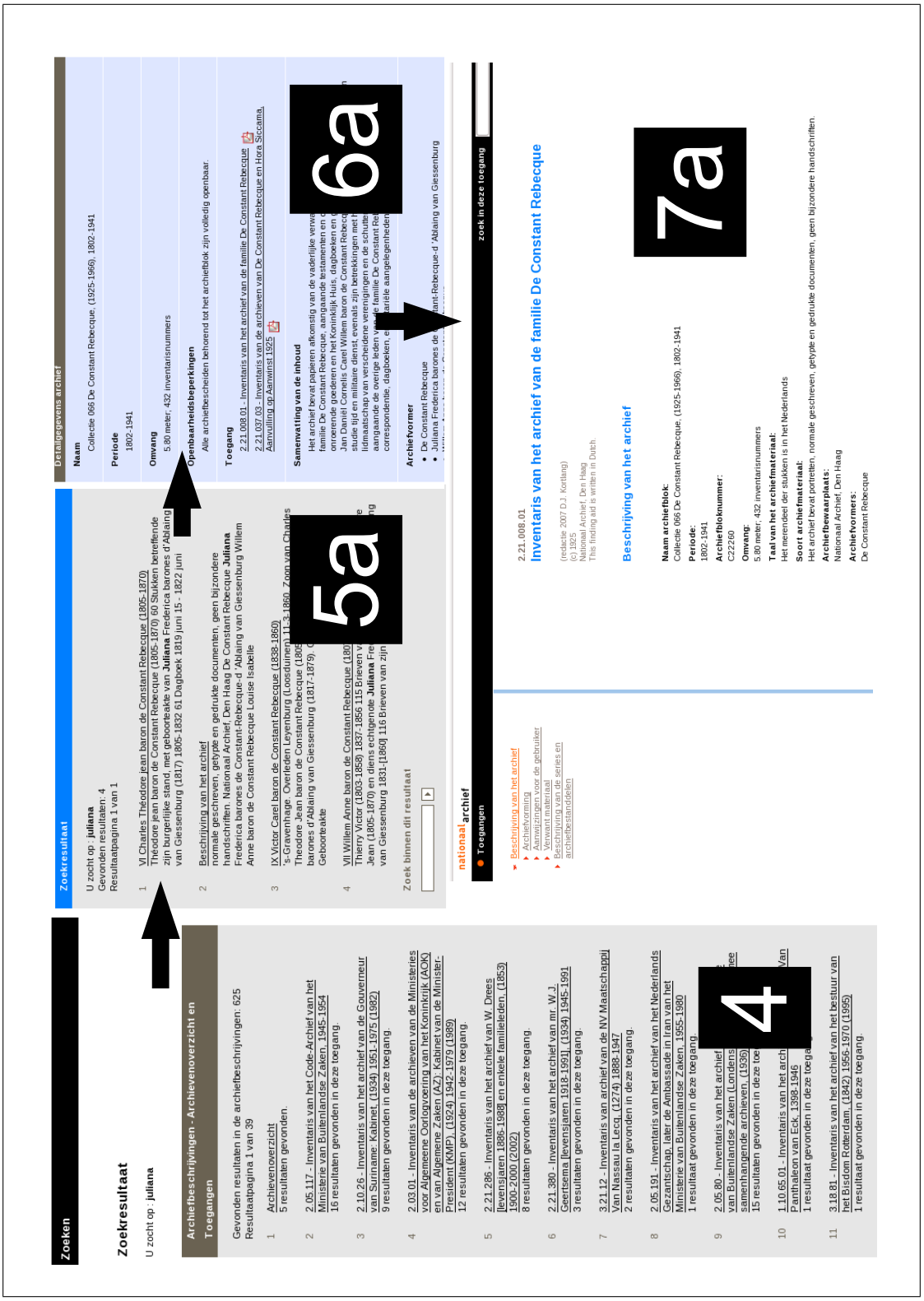

Figure 3.2: *Nationaal Archief* System Interaction (2). Image courtesy of the *Nationaal Archief* (rotated clockwise).

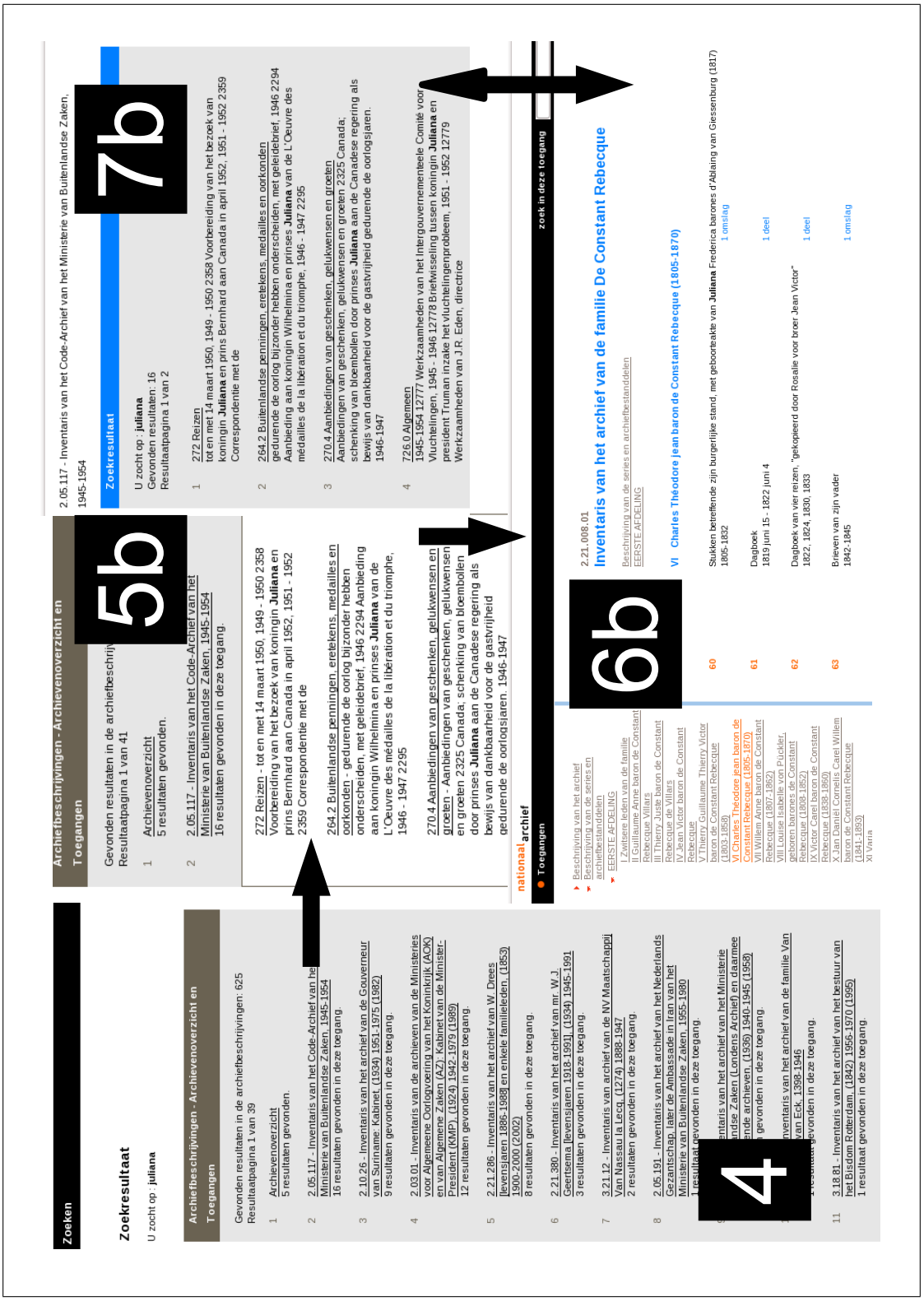

Figure 3.3: *Nationaal Archief* System Interaction (3). Image courtesy of the *Nationaal Archief* (rotated clockwise).

interface that zooms into the clicked archival component in **(6b)**. Further search in the EAD finding aid is possible by entering a query in the search box on the top-right side of the screen, which directs the user to the list of archival descriptions of components in **(7b)**, and eventually back to the interface of archival descriptions in **(6b)**.

This search interaction has been stored in the NA log files. From **(4)**, we can trace each and every click on an EAD finding aid back to a query that has been entered in the search box in **(2)**. This resembles a user request with desired answers to that request.

3.3.2 Search Log Files and Document Collection

We use the search logs of the *Nationaal Archief*. The history preserved at this institution goes back to more than 1,000 of years, with materials preserved in archives which stretch more than 93 kilometers or 58 miles. It also includes maps, drawings, and photos—much of it is published on the *Nationaal Archief* website (`www.nationaalarchief.nl`). The website provides access by offering a search engine, which includes searching in archival finding aids compiled in EAD, image repositories, and separate topic-specific databases.

The NA logs are 91.1 GB in size, with 39,818,981 unique IP-addresses, and collected from 2004 to a part of 2009 on a Microsoft IIS server. This illustrates that the *Nationaal Archief* attracts high traffic. The information contained in the search logs were recorded from 2004-2006 in the *Common Logfile Format* (CLF), and from 2007 to 2009 in the *W3C Extended Logfile Format* (ELF). The information in the CLF format includes a date, a time-stamp of a hit, unique identifier representing a user, the URL of the link that was visited, the query string, and a browser identifier. In the ELF format, it also includes a referral, and transactions have been recorded in detail within each second.

In our experiments, we only focus on click-through data of user interaction with EAD finding aids, where each click contains the filename of an EAD finding aid, which can be traced back to a corresponding query. The reason for this selection is that we have also obtained the corresponding EAD finding aids—also supplied by the *Nationaal Archief* —for analysis and further experimentation. Each EAD finding aid encodes an archival description that describes the content and context of archival material. We use 5,934 EAD finding aids in XML—937 MB of data—which are mostly written in the Dutch language.

3.3.3 Information Extraction from Logs

An archival search log contains both searching and browsing behavior, with complete sessions starting from an initial query. Given the massive size of the log, we pre-processed it by extracting the click-through data that consist of the subset of clicks to EAD URLs. The query string, clicked URL, and the IDs of

Table 3.1: Example of information in sessions extracted from the log.

Query (Topic)	File	Session ID	#
burgerlijke stand suriname	1.05.11.16	504d2bbe246d877bda09856ecc300612.5	28
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.1	6
burgerlijke stand suriname	3.223.06	22fe3a65b0c9223280f2dd576c57a012.35	1
burgerlijke stand suriname	1.05.11.16	2b844140ef7cfd438300da7ec6278de0.147	1
burgerlijke stand suriname	2.05.65.01	3784a93938e29a6aef8f50baa845a6f3.1	1
burgerlijke stand suriname	1.05.11.16	8b21ec51722f3a52cfaf35d320dfacb0.3	1
burgerlijke stand suriname	1.05.11.16	212de7cab1c3709be3a95ac1a37a7873.2	1
burgerlijke stand suriname	1.05.11.16	9235756a6dbdcffba9179d75108cd220.433	1
burgerlijke stand suriname	3.231.07	3c34072bef0d505467ca9394c392888d.2	1

the user are extracted. It is further processed by aggregating the clicks for each query in a session and keeping track of the count. 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 (or 1,800 seconds), where i is a click. This results eventually in 194,138 sessions. Table 3.1 presents the extracted interaction data on an aggregated level. This is used to derive a test collection.

When we focus on Table 3.1, we notice that for query "burgerlijke stand suriname" (in English, "registry of births, deaths and marriages suriname") clicks exist in 9 different sessions, coming from 8 different IPs. There were 28 clicks in one session for EAD finding aid "1.05.11.16," and the same file was clicked in total 6 different sessions. The same file was re-clicked from an IP address in the next session. So "1.05.11.16" could be regarded as "relevant."

Although we regard here "clicked pages" as pseudo-relevant, we make no particular claims on the interpretation of clicks. We make the reasonable assumptions that searchers found these pages of sufficient interest—for whatever reason—to consult them more closely, and that a more effective ranking algorithm will tend to rank such pages higher than those that do not receive clicks. In this chapter, we are interested in the potential of log-based evaluation, and a relatively naive click model is sufficient for that purpose.

3.3.4 Types of Test Collections

A subset of the search log files is used, namely the clicks on archival finding aids in EAD, which is rapidly growing in use for online archival access. We notice that the usage of EAD started to take off in 2006 (19.9MB out of 9.8GB; 0.20%), and this trend in popularity was upward, as it also increased in 2007 (1.5GB out of 31.5GB; 4.8%), and in 2008 (2.8GB out of 41.2GB; 6.8%), and a part of 2009 (304.4MB out of 3.8GB; 7.8%). Hence, the Web traffic of the *Nationaal Archief* is increasingly consisting of the use of EAD, although the amount of EADs published online has increased as well.

We extract from the search log files in total 50,424 unique topics (after string

processing, i.e. squeezing white spaces, conversion to lowercase, removal of punctuation), which have been created by 110,805 unique IP-addresses. There are in total 441,575 clicks with 91,009 unique topic-click pairs. Since the collection consists of 5,934 EAD files, numerous topics matched with these files. The queries have a long-tail distribution, where the majority of the topics consist of unique queries that occurred just once. This is also typical in the archival domain, for example genealogists looking for (unique) family names, and this was also the case in the *Nationaal Archief* logs.

We derive different test collections from the logs. We use clicks on the file-level in order to evaluate full-text retrieval. The two types of test collections used:

Complete Log Test Collection The set of 50,424 unique topics, and their corresponding clicks to EAD files, where each and any click is treated as a pseudo-relevance judgment.

Test Collections Based on Agreement Subsets of the search log filtered by the agreement reached among multiple users when they clicked on the same EAD finding aids for a given topic. For example, for agreement 2, we only retain documents clicked by at least two users, which scales down the test collection to 2,455 topics.

We test the two types of test collections separately in the next section.

3.4 Log-Based Evaluation in Action

In this section, we use the log-based test collections to determine the retrieval effectiveness of different ranking methods. We look at both the complete log, as well as on smaller subsets based on agreement. Recall that test collections are used for the comparative evaluation of systems or ranking algorithms, hence we need a number of variant systems in order to show their retrieval effectiveness.

3.4.1 IR Models and Systems

We use the README system framework, which has been described in the previous chapter. To test the effectiveness of the two types of test collections, we regard the sets of queries of the test collections and the retrieval system as independent variables, with the retrieval effectiveness as dependent variable.

We measure the retrieval effectiveness using three IR measures, namely Mean Average Precision (MAP), which is the most frequently used summary measure for a set of ranked results, Mean Recipropal Rank (MRR), and Normalized Discounted Cumulative Gain (nDCG). The MRR is a static measure that looks at the rank of the first relevant result for each topic, and the nDCG measure that uses the number of clicks on each result as a form of graded relevance judgment.

The systems use the default parameter values, the collection λ is set to 0.15—and we set the threshold of returned results for each topic to 100.

BOOL is the Boolean model, where there is no ranking, but a batch retrieval of exact matching results (e.g. Belkin and Croft, 1992). 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, all terms in the query need to appear in the result (e.g. Ponte and Croft, 1998).

$$P(q|d) = \prod_{t \in q} P(t|d)^{n(t,q)} \tag{3.1}$$

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 (e.g. Chen and Goodman, 1996).

$$P(t|d) = (1 - \lambda) \cdot P_{mle}(t|d) + \lambda \cdot P_{mle}(t|C) \tag{3.2}$$

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.

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 (e.g. Kraaij, 2004).

$$NLLR(d, q) = \sum_{t \in q} P(t|q) \cdot \log \left(\frac{(1 - \lambda) \cdot P(t|d) + \lambda \cdot P(t|C)}{\lambda \cdot P(t|C)} \right) \tag{3.3}$$

OKAPI is Okapi BM25, which incorporates several more scoring functions to compute a ranking, such as also the document length as evidence (e.g. Robertson et al., 1994).

$$BM25(d, q) = \sum_{t \in q} IDF(t) \cdot \left(\frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \right), \tag{3.4}$$

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.

3.4.2 Study 1: Complete Log Test Collection

When we use all topics for evaluation, and look at all measures, we see in Table 3.2 that *BOOL* is obviously the worst performing system. We note that the

Table 3.2: System-ranking of runs over all topics.

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1808 (5)	0.2493 (4)	0.2548 (3)	0.2591 (2)	0.2631 (1)
MRR	0.2015 (5)	0.2940 (4)	0.2980 (3)	0.3024 (2)	0.3077 (1)
nDCG	0.2659 (5)	0.3289 (4)	0.3547 (3)	0.3605 (2)	0.3652 (1)

differences among the other systems are modest, but these differences are all significant (1-tailed) using the paired-samples t-test on a 1% significance level.

When looking at the recall over all topics, we see that the BOOL and LM systems retrieve 48,096 relevant results out of 87,057 (55.25%), the LMS system returns 57,935 of 89,906 (64.44%), the NLLR system has a recall of 65.18%, and the OKAPI system returns most relevant results with 65.69%. It shows that the system using the OKAPI model performs best with our document collection, and that exact matching using the BOOL and LM systems both do not pay off for the early rank (MRR), and as expected hurts the recall. The recall values can be clarified by the long-tail distribution of query terms, which contains many non-occurring names.

In summary, the BOOL system is the worst-performing system, then the LM system, with the LMS system improving over the LM system, and that the differences among the LMS, NLLR, and OKAPI systems are modest (but significant). We will validate the relative system ranking against a set of humanly judged topics in the next section, but first we will look at the system ranking induced by smaller subsets of topics based on agreement.

3.4.3 Study 2: Test Collection Based on Agreements

The search log-based test collection has considerably more topics than a traditional test collection with 25-200 topics. While having thousands of topics opens up new uses, such as focusing on various breakdowns of the topic set even on relatively rare phenomena, it also presents an efficiency challenge: many online archival search systems may crumble under thousands of queries, preventing fast and efficient evaluation. Therefore, we take into account the agreement that exists among different searchers. For example, when we pay attention to Table 3.1, this means that only EAD file "1.05.11.16" is included as a relevance judgement in the test collection.

We see in Table 3.3 that as we increase the threshold of agreement, the number of topics decreases significantly. Take for example notice that in the case that if the agreement is set to 2, the topic set size decreases to 2,455 from 50,424, and when we set the threshold to 4, only 533 topics are left over. What does this mean for evaluating a system with such a set size? The results of this experiments are presented in Table 3.4, where we focus on the MAP scores.

Table 3.3: Distribution (in percentages) of N topics over query length for all topics compared to when filtering on agreements, and e-mail references.

# Tokens	All	Agree 2	Agree 3	Agree 4	E-mail
1	37.66	76.82	82.38	81.99	17.81
2	33.43	15.76	12.02	12.76	19.18
3	16.97	5.87	4.87	4.69	30.14
4	6.68	1.14	0.73	0.56	19.18
N	50,424	2,455	965	533	73

Table 3.4: System-ranking of runs over topics with agreement.

	Agree	BOOL	LM	LMS	NLLR	OKAPI
MAP	**2**	0.1522 (5)	0.3605 (4)	0.3620 (3)	0.3629 (2)	0.3751 (1)
	3	0.1120 (5)	0.3891 (3)	0.3888 (4)	0.3894 (2)	0.3991 (1)
	4	0.1071 (5)	0.3637 (4)	0.3639 (3)	0.3641 (2)	0.3793 (1)
MRR	**2**	0.1629 (5)	0.4020 (4)	0.4030 (3)	0.4039 (2)	0.4157 (1)
	3	0.1188 (5)	0.4253 (2)	0.4247 (4)	0.4253 (2)	0.4356 (1)
	4	0.1132 (5)	0.3943 (3)	0.3942 (4)	0.3945 (2)	0.4110 (1)
nDCG	**2**	0.2734 (5)	0.4521 (4)	0.4564 (3)	0.4578 (2)	0.4726 (1)
	3	0.2384 (5)	0.4750 (4)	0.4767 (3)	0.4778 (2)	0.4913 (1)
	4	0.2315 (5)	0.4520 (4)	0.4552 (3)	0.4560 (2)	0.4735 (1)

We focus on the differences of the MAP scores when we take an agreement between two different IPs. The BOOL system is significantly performing worst, and the OKAPI system is performing the best compared to either the LMS system with a significant improvement of 3.62% (t(4835) = 5.50, $p < 0.01$, one-tailed), or similarly an 3.36% significant improvement over the NLLR. Although the improvement between the LM and LMS systems was only 0.42%, it was also significant (t(4835) = 3.76, $p < 0.01$, one-tailed). This is completely in line with our findings when using the full set of topics.

What happens when we take an agreement of a click among 3 different IPs? Again we focus on the MAP scores. We see that the BOOL system is again significantly the worst performing system, and the OKAPI is significantly performing better on a 1% significance level. The LM and LMS systems also swap from position. However, we did not see significant differences among the other three models. This is partly clarified due to the distribution of the number of keywords in a query as presented in Table 3.3, because the effect of smoothing is leveled out when the topics become in majority singleton queries. As Table 3.3 shows, when we increase the agreement threshold, there are only 965 queries are left, which are predominantly very short (limiting the impact of smoothing) and many of them having having only a single relevant document.

Finally, what happens when we take an agreement among 4 different IPs?

We still see the same pattern as the previous runs, with the BOOL being the worst, and the OKAPI system the best (+4.17%, $t(1300) = 3.21$, $p < 0.01$, one-tailed) compared to the runner up system NLLR. The findings are also consistent with the nDCG scores, though we see that with MRR, the LM and LMS systems swap position with a minor non-significant difference.

In summary, there are two implications. First, deriving a test collection using agreement of 2 seems a viable alternative for using the whole log file. Second, the system rankings are similar when treating the clicks as binary pseudo-relevance judgements (MAP, MRR) and as graded judgements (nDCG).

3.5 External Validation

We investigate the validity of the log-based test collection in terms of the resulting system ranking. As ground-truth, we use a test collection constructed by human experts: responses of archivists to e-mail reference questions. We assume that the 'manual approach' of archivists providing answers to the questions of users are of high quality, but these are difficult to obtain, hence the comparison with the log-based approach. The system rankings of the search log-based test collection are compared to this ground-truth. Additionally, we compare the search log-based test collections with the original ranking of the *Nationaal Archief* system. This is the original system used by users in the log-based approach, and archivists in the e-mail references.

3.5.1 Test Collection Based on E-Mail Reference Requests

We analyze the e-mails that the *Nationaal Archief* received from users, and with replies from archivists that referred explicitly to EAD files. We look at all correspondence (4.1GB of data). The e-mails are converted from *PST* file format to *mbox* format, which comes in readable textual form. Eventually, we manually select 73 different topics (and recommended EAD links) from the e-mail files that have a clear information need expressed by the user and a clear recommendation for at least one EAD finding aid given by an archivist.

A typical example is the information request in Figure 3.4. The explanation of the information request is included in `<narrative>`, the topic in `<title>`, and the relevant files for that topic in `<file>`. We select typical replies from an archivist who linked to EAD files using the query, or recommended the EAD finding aids which are relevant.

3.5.2 System Rank Correlations

First, the recall values of the e-mail runs. We find for the BOOL and LM systems that there are 61 topics with relevant results, where 46 out of 90 (51.1%) EADs have been found. For the LMS and NLLR systems, there are 63 valid topics that return 66 relevant EADs out of 92 (71.7%). For the OKAPI system, there are 63

```
<topics>
 <topic nr="1">
  <title> wateringen </title>
  <narrative> I. Purpose: Query for Commissioned Research Support II. Topic: Historical
Water and Flood Management in Holland III. Geographic Location: My research focuses on
two locations: 1. Along the river channels, uiterwaarden, and polders of the Nederrijn and
Lek Rivers; and 2. along the Oude Rhine (region), between about Woerden and Katwijk. This
includes the “waterschaapen" of Rhineland and Delftland. IV. Type of materials: maps, sur-
veys, and tables of data that pertain to land use (agriculture types), water/flood management
(dijk, sluice, drainage ditches, pumping, etc...) I am especially interested in archival materials
that relate to the early period (before 1300), but including up to about 1925.</narrative>
  <files>3.01.04.01; 4.VTH; 4.KIVI; 4.ZHPB4; 2.16.91; 2.16.06</files>
 </topic>
 <topic nr="2">
  <title>ministerie van verkeer en waterstaat ptt draadloos telegrafie</title>
  <narrative>Do you know whether the “Ministerie van Verkeer en Waterstaat" has an own
archive which includes PTT documents relating to wireless telegraphy?</narrative>
  <files>2.16.09; 2.16.93</files>
 </topic>
 <topic nr="34">
  <title>Frans Beelaerts Blokland Peking Beijing</title>
  <narrative>I am writing a book about foreigners in Beijing from the Boxer Rising in
1900 to the Communist takeover 1949. Jonkheer Frans Beelaerts van Blokland was the Dutch
Minister in Peking during the World War One. I am very interested in seeing any papers that
you may hold relating to his years in Peking.</narrative>
  <files>2.05.90 ; 2.05.19 ; 2.21.253</files>
 </topic>
</topics>
```

Figure 3.4: Examples of three topics based on e-mail reference requests, where the <title> is used as query, and if necessary, translated to Dutch.

topics with relevant results, i.e. 67 out of 92 EADs (72.8%). We note that there are 73 topics, but the systems could not find results for all topics.

The results of Table 3.5 show that the BOOL system performs worst, and the OKAPI system is performing best. We again use the paired-samples t-test to check for significance. We look at the MAP scores. When we rank with the LM system, there is also a significant improvement of 65.49% over the BOOL system ($t(60) = 3.03$, $p < 0.05$, one-tailed). When we use the LMS system, we see a 7.45% significant improvement over LM without smoothing ($t(62) = 2.20$, $p < 0.05$, one-tailed). However, the difference between the LMS and NLLR models was only 0.47%, and is not significant. Moreover, the OKAPI system performed 8.77% better than *NLLR*, but it is not significantly better. The findings are similar using the MRR and nDCG measures, though we do not see a difference between the LMS and NLLR systems.

How reliable are our log-based test collections when compared to the test collection semi-automatically derived from the e-mails and their experts' replies? When we compare the system rankings of the test collection from the whole

Table 3.5: System-ranking of runs over e-mail topics.

	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.1678 (5)	0.2777 (4)	0.2984 (3)	0.2998 (2)	0.3261 (1)
MRR	0.1896 (5)	0.3197 (4)	0.3313 (2)	0.3313 (2)	0.3589 (1)
nDCG	0.2646 (5)	0.3624 (4)	0.4167 (2)	0.4176 (2)	0.4361 (1)

log (Table 3.2) with the e-mail topics (Table 3.5) using the MAP scores, we see a complete agreement with a Kendall's Tau value of 1. Overall, we see full agreement between the log-based evaluation and the reference requests, and among the test collections of the log-based evaluation approach.

The results show that the *OKAPI* performs best for EAD finding aids written in the Dutch language. It confirms earlier findings by Savoy (2003) who showed that *OKAPI* also works best for information retrieval with Dutch texts.

These results are promising, but we further research the rank correlations between the system rankings of both approaches by adding the *Nationaal Archief* system itself in the next experiments. For the e-mail topics, archivists have been using the *Nationaal Archief* system in their answers. Is there any effect of this? Are there any limitations in the log-based approach? A limitation could be the position-bias due to the *Nationaal Archief* system, where users tend to click on the first results they see.

3.5.3 Comparison with the *Nationaal Archief* System

Is there a bias from the original system when using the log-based test collection? The clicks of users in the search log may have been greatly influenced by the system that they have used. We compare the results with the ranking of the *Nationaal Archief* system (NA) itself. We use each and every topic of a part of the test collection, namely a month (January) of 2009, to query the *Nationaal Archief* system. The hit lists that this system return for every topic are downloaded. There are 4,110 topics in this month. The returned lists are another type of run. We use the same set of topics to query our systems. See Table 3.6.

Table 3.6: System-ranking of log-based approach compared to the *Nationaal Archief* system over January 2009.

	NA	BOOL	LM	LMS	NLLR	OKAPI
MAP	0.4283 (1)	0.1501 (6)	0.2703 (5)	0.2752 (4)	0.2787 (3)	0.2896 (2)
MRR	0.5112 (1)	0.1627 (6)	0.3100 (5)	0.3142 (4)	0.3177 (3)	0.3319 (2)
nDCG	0.5020 (1)	0.2473 (6)	0.3542 (5)	0.3728 (4)	0.3773 (3)	0.3896 (2)

There is a change. We find here that the *Nationaal Archief* system heavily outperforms the other systems and has a system rank of 1. This illustrates the

click-bias existing in the log-based approach. When looking at the MAP score, the difference between the *Nationaal Archief* system and the OKAPI system, we find an improvement of the *Nationaal Archief* system of 47.89% ($t(4162) = 8.17$, $p < 0.01$, one-tailed). The differences among the other systems are also significantly on a 1% significance level.

We also show the contrast between the topics of the e-mail reference requests with our IR models and the *Nationaal Archief* system. The results for the performance of these topics with our systems have already been shown in Table 3.5, but now we append the results of the *Nationaal Archief* system. For the *Nationaal Archief* system, we find a MAP score of 0.2738, a MRR score of 0.3039, and a nDCG score of 0.3432. Here, we see that the system ranking of the *Nationaal Archief* system is positioned between the BOOL and LM systems, but the actual difference between the *Nationaal Archief* and LM systems is 1.42% and not significant. So the e-mail test collection does not have the same click-bias as the log-based approach has. Now, when we look at the Kendall's tau, we see a value of 0.467 over the six systems between the log-based and e-mail reference request approaches, so adding just one extra system shakes up the rank correlation of the system rankings between the two types of test collections and approaches.

In the next section, we further research the scope of the search log-based approach by exploring another third source, namely a research guide of the *Nationaal Archief* consisting of twenty-five questions and answers, which is recall-oriented.

3.5.4 The War Research Guide

In this section, we present a study on using the list of twenty-five questions of the *Nationaal Archief* (see Figure 3.5) called *De Oorlogsgids* (In English, *The War Guide*)[2] to construct an archival IR test collection. We present the experiment, compare and validate its results with the previous experiments.

Experimental Setup

There are twenty-five questions (topics or requests). These questions are example case-studies that are aimed to teach archival users to do research in archives, and refer to archives related to the Second World War (WWII, 1939-1945). These includes archives of Dutch governmental institutions or of persons/families, which were created or changed as a result of the German occupation of the Netherlands during WWII or the Japanese occupation of the former Dutch East Indies (nowadays Indonesia). The majority of the archival materials are publicly accessible.

[2] Retrieved 2011/03/15 from `http://www.nationaalarchief.nl/collectie/ondersteuning/onderzoeksgids/default.asp?ComponentID=11865&SourcePageID=16483`

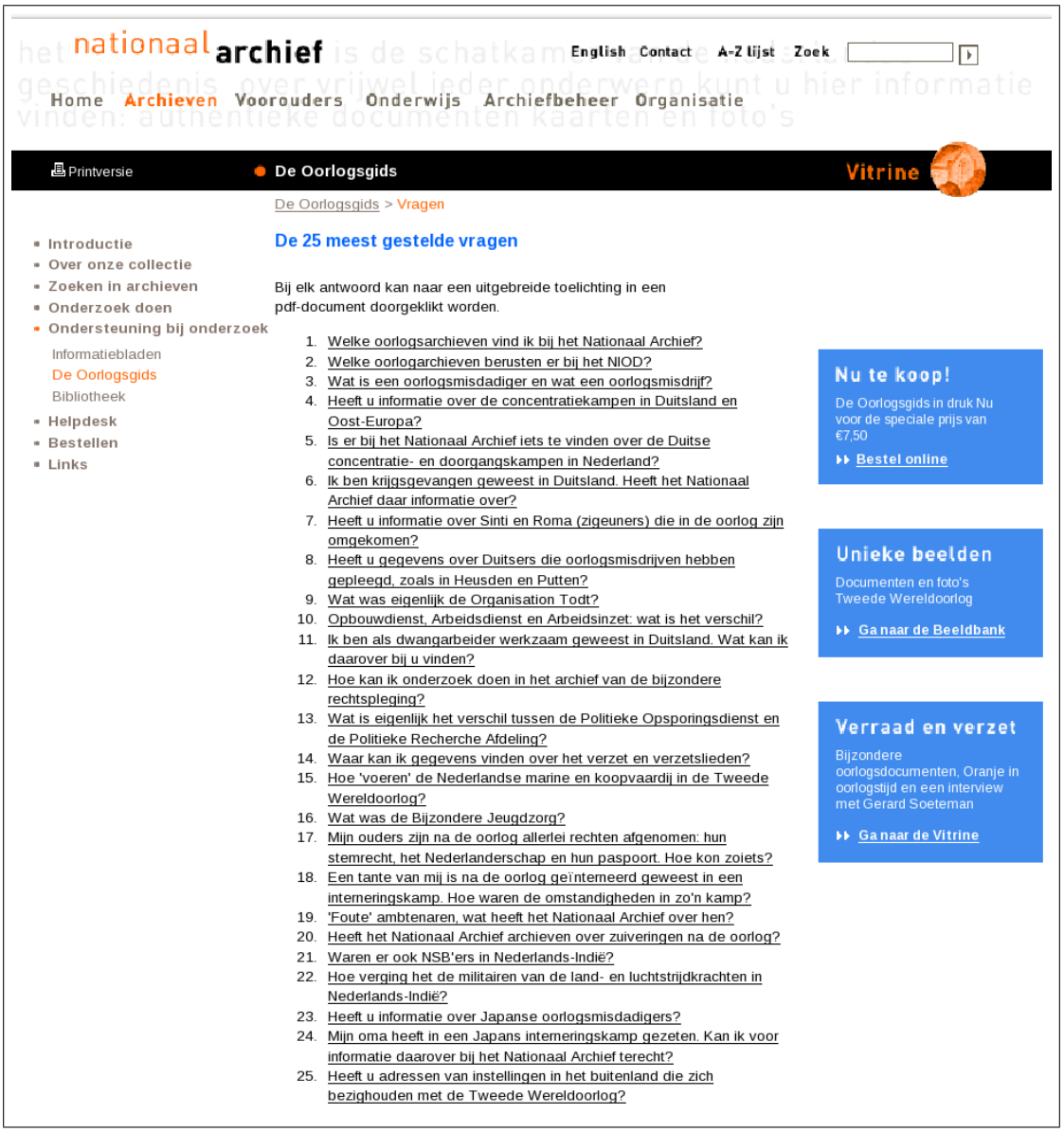

Figure 3.5: Research guide with twenty-five questions on the archives about the Second World War. Image courtesy of the *Nationaal Archief*.

Each of these topics also refers to a PDF file that answers a question, i.e. describes the context of the question and refers to relevant EAD finding aids for researching the answer to this question (relevance judgements and document collection). Here, there are 22 questions that also refer to EAD finding aids, so we have 22 topics. Therefore, we have the building blocks necessary to construct an IR test collection. Figure 3.6 illustrates these building blocks with two example topics. We refer to this as the "Research Guide" approach. A difference with the other approaches, is that the Research Guide approach is more oriented towards recall, as the lists of EAD finding aids suggested by the *Na-*

```
<topics>
 <topic nr="1">
  <title> war archives </title>
  <description> Which archives on the Second World War can be found at the Nationaal
Archief? </description>
  <narrative> The War Guide consists of 25 most frequently asked questions on archives
related to the Second World War. This guide tells us which archives have to be researched in
order to get an answer to the questions. </narrative>
  <files> refers to 241 EADs (not listed to conserve space) </files>
 </topic>
 <topic nr="4">
  <title> concentration and internment camps </title>
  <description> Do you have information on the concentration and internment camps in
Germany and Eastern Europe? </description>
  <narrative> Yes, the Nationaal Archief has archival material on that topic. It mostly touches
upon information after the period of the closure of these camps. You will find data on the
preparation for the relief of the war victims and—not always complete—lists of names of
prisoners, survivors, and of those who have passed away. </narrative>
  <files> refers to 23 EADs (not listed to conserve space) </files>
 </topic>
</topics>
```

Figure 3.6: Example of topics based on frequently asked questions posed to archivists of the *Nationaal Archief* (here translated to English from Dutch).

tionaal Archief are exhaustive. For the log-based and e-mail approaches, there is a focus on precision. There are also shallow clicks in the log, for example, where users click once on an EAD finding aid.

We test this test collection with the following experimental setup. We use the ranking (or 'archival overview') that the *Nationaal Archief* system provides as ground truth. We use the systems as described in Subsection 3.4.1 on page 43. Additionally, given the topic specification in Figure 3.6, we have five types of runs: (**R1**) `<title>`, (**R2**) `<description>`, (**R3**) `<title>` + `<description>`, (**R4**) `<narrative>`, and (**R5**) `<title>` + `<narrative>`. So we use the (combination of the) content of these XML elements as queries to retrieve EAD finding aids. These runs stand for the amount of keywords used to express an information need into a query, where R1 is the most concise query possible and R5 is the most elaborate. We measure the results with the same performance measures as in the previous experiments. We return the top 1000 results for each topic.

Results

Now, the results. Table 3.7 shows the numbers. When we look at the MAP scores, we see that for run R1 and run R2, which consist of shorter content, the LM system is working best—this is different compared to the log-based approach. The OKAPI system comes second. We did not find significant improvements among the systems. This system ranking suggests that the keywords in

<title> and <description> are far from being optimal for recall, and very centered towards early high precision, though the suggested EADs in the qrels are exhaustive.

Table 3.7: System-ranking of five types of runs over the Research Guide topics.

	Run	NA	BOOL	LM	LMS	NLLR	OKAPI
MAP	**R1**	0.1095 (5)	0.0806 (6)	0.1272 (1)	0.1165 (4)	0.1167 (3)	0.1206 (2)
	R2	0.0871 (3)	0.0516 (6)	0.1041 (1)	0.0805 (5)	0.0817 (4)	0.0971 (2)
	R3	0.0970 (3)	0.0516 (5)	0.0886 (4)	0.0024 (6)	0.1037 (2)	0.1099 (1)
	R4	0.0297 (3)	0 (5)	0 (5)	0.0020 (4)	0.0962 (2)	0.1474 (1)
	R5	0.0740 (3)	0 (5)	0 (5)	0.0021 (4)	0.1034 (2)	0.1562 (1)
MRR	**R1**	0.3250 (5)	0.2934 (6)	0.4172 (1)	0.3746 (4)	0.3748 (3)	0.4081 (2)
	R2	0.2514 (6)	0.3274 (2)	0.4107 (1)	0.2813 (5)	0.2987 (4)	0.3217 (3)
	R3	0.3104 (5)	0.3274 (4)	0.3810 (2)	0.0241 (6)	0.3794 (3)	0.3840 (1)
	R4	0.1250 (3)	0 (5)	0 (5)	0.0024 (4)	0.3747 (2)	0.5095 (1)
	R5	0.4114 (2)	0 (5)	0 (5)	0.0024 (4)	0.3969 (3)	0.5418 (1)
nDCG	**R1**	0.2523 (5)	0.2224 (6)	0.2745 (4)	0.3150 (3)	0.3211 (2)	0.3382 (1)
	R2	0.2546 (2)	0.1441 (6)	0.1970 (5)	0.2381 (4)	0.2476 (3)	0.3179 (1)
	R3	0.2721 (3)	0.1441 (5)	0.1823 (4)	0.0492 (6)	0.2965 (2)	0.3449 (1)
	R4	0.2818 (3)	0 (5)	0 (5)	0.0581 (4)	0.3142 (2)	0.4114 (1)
	R5	0.2843 (3)	0 (5)	0 (5)	0.0554 (4)	0.3257 (2)	0.4227 (1)

When we add more text, or evidence, to the query (i.e. for runs R3, R4, and R5), we see that the OKAPI system still gives the best results. The LMS system is not working here as the MAP scores are just 2%, but when we remove the smoothing in the LM system, the MAP scores even plummet to 0, because there is no EAD finding aid that contains all the keywords of the query. In run R4, we note that the difference between the *Nationaal Archief* system and the OKAPI system is the greatest, where the performance almost quadruples with 396.30% ($t(20) = 6$, $p < 0.01$, one-tailed). In run R5, the improvement of OKAPI over the NA system more than doubles with 111.08% ($t(20) = 5.64$, $p < 0.01$, one-tailed). We note that the OKAPI system performs better overall in all runs in comparison with the standard *Nationaal Archief* system.

When we look at the rank correlations between the system rankings of these runs and the system ranking of the log-based approach over January 2009 in Table 3.6 using Kendall's tau, we see for R1 a tau value of 0.067, for R2 a value of 0.333, for R3 a value of 0.467, and for R4 and R5 a value of 0.733. For the rank correlation between the system rankings of the e-mail reference request approach and the Research Guide approach, we see for R1 a tau value of 0.6, for R2 and R3 a value of 0.467, and for R4 and R5 a value of 0.7333. When we omit the *Nationaal Archief* run for R4 and R5, we still see a full agreement between the log-based and e-mail reference request approaches with the Research Guide approach, as both then have a rank correlation of 1. However, the rank correlations of the R1, R2 and R3 runs do not show such consistency.

Moreover, we note that the test collection derived from the Research Guide does not have the same impact of the click-bias as the test collections derived from the log-based approach seem to have. Another finding is that the OKAPI system is the most effective for (very) long queries across the compared systems, and also for shorter queries. This is a similar finding as with the log-based approach, where it also performed best, given our experimental setup with the tested systems. However, there are also limitations to the Research Guide approach, because we find that the topics here do not seem to be optimally formulated, and focused towards exact matching, while the qrels are oriented towards measuring recall. When we add more evidence of the information need to the query, we see that the system rankings become similar as with the other test collections. As the Research Guide qrels are oriented towards recall, we see that the MAP scores are considerably lower than the results of the other approaches.

3.6 Conclusion

This chapter investigated a search log-based approach to the evaluation of access systems driven by EAD finding aids. Information retrieval has developed standard benchmarks that can be used to evaluate the retrieval effectiveness. However, these generic benchmarks focus on a single document genre, language, media-type, and searcher profile that is radically different from the unique content and user community of a particular archive. One of the main challenges in the evaluation is to develop contextualized evaluation methods that closely capture the unique setting at hand. Therefore, we have proposed the following. By using an archive's own collection and exploiting readily available interaction data in search logs, we can create a domain-specific test collection tailored to the case at hand. That is, having a representative document collection and representative sets of search requests. As a test collection, it can be used and reused for comparative testing under the same experimental conditions.

We conducted a large case study using a large set of EAD finding aids and search logs of the *Nationaal Archief* EAD search system. These logs are massive as these cover several years of interaction with this system. This resulted in a test collection to evaluate the full-text retrieval of archival finding aids in EAD. We presented generic methods to derive a domain-specific test collection, and used a range of retrieval models to determine the effectiveness of the test collections. Our extraction methods are naive—we treat every clicked document as pseudo-relevant—but suffice to illustrate the viability of the approach.

We compared the results by repeating the experiment using a set of traditional topics derived from email requests to the archive and the archivist's responses. We found complete agreement between the log-based evaluation and the traditional topics. However, when we extended the experimentation by adding one extra system, namely the *Nationaal Archief* system itself, we have

to nuance our findings, as the rank correlations then changes due to the bias of this system. When we repeated the experimentation by deriving another test collection from a *Nationaal Archief* research guide, we also saw similarities and differences, which may due to the recall-based orientation of this test collection, compared to the log-based approach that focused on precision. Still, our initiative of a log-based approach to IR evaluation is a step forwards towards automated system-centered evaluation of access to EAD finding aids.

So far, we looked at full-text retrieval only, but we can also evaluate on an XML element level by looking at descriptions within an EAD finding aid. This is what we will investigate in Chapter 4.