A Search Log-Based Approach to Evaluation

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

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

Abstract. Anyone offering content in a digital library is naturally interested in assessing its performance: how well does my system meet the users' information needs? Standard evaluation benchmarks have been developed in information retrieval that can be used to test retrieval effectiveness. However, these generic benchmarks focus on a single document genre, language, media-type, and searcher stereotype that is radically different from the unique content and user community of a particular digital library. This paper proposes to derive a domain-specific test collection from readily available interaction data in search logs files that captures the domain-specificity of digital libraries. We use as case study an archival institution's complete search log that spans over multiple years, and derive a large-scale test collection. We manually derive a set of topics judged by human experts—based on a set of e-mail reference questions and responses from archivists—and use this for validation. Our main finding is that we can derive a reliable and domain-specific test collection from search log files.

1 Introduction

A digital library (DL) is created for domain-specific collections, whether it be in cultural heritage or about scientific articles. But how good are DLs for disclosing their collections for their particular user groups? Anyone who is interested in DLs probably has already asked this question [8]. A major challenge for DLs is how to evaluate the information retrieval (IR) effectiveness given the domain-specifity of their collections, and how to use this crucial evaluation step to improve a DL.

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*). Such test collections are created collaboratively on generic (artificial) set of documents, such as newspaper corpora or Wikipedia, and are useful to study generic aspects of 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 DL. Creating a test collection with this conventional approach, for each DL, is simply too expensive in time and effort.

We propose a log-based approach to IR evaluation of DLs. Nowadays, almost every DL 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

M. Lalmas et al. (Eds.): ECDL 2010, LNCS 6273, pp. 248–260, 2010.

includes information that has been entered by the user, and what and where it was clicked, and so on. Is it reasonable to assume that such data can be reused for evaluation? This results in this main research question:

- *Can we use a digital library's search log to derive a domain-specific test collection?*

The envisioned test collection is tailored to the DL at hand, representative to both its document collection and its search requests. As a test collection, 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 for archives of persons, families, and corporations. These archival finding aids are created in electronic form to provide online archival access, using the Encoded Archival Description (EAD, [17]) standard based on Extensible Markup Language (XML, [4]).

The remainder of this paper is structured as follows. Section 2 describes related work. An archival institution's complete search log that spans over multiple years is used in our experimentation. We derive a domain-specific test collection from a search log in Section 3. We deploy the resulting domain-specific test collection for evaluation using a range of retrieval models in Section 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 and responses from archivists. The results are analyzed and discussed in Section 5. Finally, we conclude with our main findings and discuss pointers to future work in Section 6.

2 Background and Related Work

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

2.1 Log Analysis for Information Retrieval

Historically, the analysis of log files started and "evolved out of the desire to monitor the performance of computerized IR systems" [16, 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 DL setting have been reported in [13], 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 investigate the IR effectiveness.

Research on log analysis in library and information science preceded 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 [11], 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. An example is the paper of [23], 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 clickthrough data from search logs as a form of implicit feedback [5]. Joachims et al. [12], 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 [6].

2.2 IR Test Collections

IR evaluation can be traced back to the workprocess of a librarian working with card indexes using library classification schemes [19]. 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 (indexing languages in the 1950, retrieval algorithms nowadays). 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 [21]. Test collections can be reused by evaluating new or adapted systems or ranking algorithms under the exact same experimental conditions. There exist test collections for a variety of domains. Examples included the Cystic Fibrosis database [20], and WT10g for the Web in general [2]. In the field of Focused Retrieval, the Initiative for the Evaluation of XML Retrieval (INEX) constructed test collections from XML files [14].

2.3 Evaluating Archival Metadata Retrieval

Published research that empirically or experimentally deals with the evaluation of archival metadata retrieval is scant [10]. 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 [17] 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 [15]. Subsequently, there were a few studies that tested the effects of some external context knowledge on retrieval, such as controlled vocabulary terms [18] or document-collection granularity [10]. The retrieval of online archival finding aids (not in EAD) have been examined in the study of [7] by counting the number of finding aids returned by search engines using different types of query reformulations, i.e. keyword, phrase, and Boolean searches using the topical subject and names headings as queries. The retrieval experiments of [22] 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) in six IR systems. The only study so far that focused on the use of EAD on the XML element level was [24], which tested the ranking based on relevance and archival context.

3 From Search Log to Test Collection

In this section, we study how to derive a test collection from a search log. We perform a case study of an archival institution, and use its search log to create domain-specific test collections, tailored to the collection and users.

3.1 Search Log Files and Document Collection

We have obtained the search logs of the National Archives of the Netherlands (NA-NL). The history preserved at this institution goes back to more than 1,000 of years, preserved in archives which stretch more than 93 kilometers or 58 miles. It also includes maps, drawings, and photos—much of it is published on the NA-NL website (`www.nationaalarchief.nl`). The website provides access by offering a search engine, which includes searching in archival finding aids compiled in EAD [17], image repositories, and separate topic-specific databases.

The logs were 91.1 GB in size, with 39,818,981 unique IP-addresses, and collected from 2004 to a part of 2009 on a Microsoft IIS server. This illustrates that the NA-NL 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 included a date, a timestamp of a hit, unique identifier for the user, the URL of the link that was visited, the query string, and a browser identifier. In the ELF format, it also included a referral, and transactions were recorded in detail within each second.

In our experiments, we focus on clickthrough data of online archival finding aids in EAD, where each click contains the filename of a result and a corresponding query. The reason is that we have also obtained these matching EAD files for analysis and further experimentation. Each EAD file describes the contents of an entire archival collection. We use 4,885 EAD files in XML—651 MB of data obtained and mostly written in the Dutch language—from the National Archives of the Netherlands, which were also found in the log files. The mean length of the text-only content of these files is 40,608 bytes (median = 9,119), the mean count of the number of XML pair tags is 2,334 (median = 540), thus some of the archival finding aids are exceptionally long in content and complexly and deeply structured in XML.

3.2 Information Extraction from Logs

A DL's search log contains both searching and browsing behavior, with complete sessions starting from an initial query. Given the massive size of the log, we preprocessed it by extracting the clickthrough data that consist of the subset of clicks to EAD URLs. The query string, clicked URL, and the IDs of 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

Table 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

sessions. Table 1 presents the extracted interaction data on an aggregated level. This is used to derive a test collection.

When we focus on Table 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 file "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. Henceforth, the EAD file "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 paper we are interested in the potential of log-based evaluation, and a relatively naive click model is sufficient for that purpose.

3.3 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 archival Web 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 National Archives of the Netherlands 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 were in total 465,089 clicks with 91,009 unique topic-click pairs. Since the collection consists of 4,885 EAD files, numerous topics matched with these files. Table 2 depicts the 8 most popular topics. The queries have a long-tail distribution, where the majority of the topics were unique queries with only 1 hit. This is also

Table 2. Top 8 most popular used query strings, where the total number of clicks with query is 465,089 with 50,424 unique 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)

typical in the archival domain, for example genealogists looking for (unique) family names, and this was also the case in the NA-NL 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 in our experiments are:

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 filtered by the agreement among multiple users on the same clicked documents for a given topic. For agreement 2, we only retain documents clicked by at least two users which restricts it to 4,855 topics.

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

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.

4.1 IR Models and Systems

Our system [25] uses MonetDB with the XQuery front-end Pathfinder [3] and the IR module PF/Tijah [9]. All of our EAD files in XML are indexed into a single main memory XML database without stopword removal, and with the Dutch snowball stemmer. To test the effectiveness of the two types of test collections, we use four retrieval models used in PF/Tijah as independent variables. These are controlled by using the default parameter values, the collection λ is set to 0.15—which we find to be working optimal—and we set the threshold of the ranking for each topic to 100.

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

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

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

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

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

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

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.

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

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

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

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.

4.2 Complete Log Test Collection

In our evaluation, we use 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 by different results as a form of graded relevance judgment.

When we use all topics for evaluation, and look at all measures, we see in Table 3 that *BOOL* is obviously the worst performing system. We note that the 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.

Table 3. 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)

Table 4. Distribution (in percentages) of topics over query length for all topics compared to when filtering on agreements, and e-mail references, resulting in N topics

# Tokens	All	Agree 2	Agree 3	Agree 4	E-mail
1	37.66	77.65	80.90	79.45	17.81
2	33.43	15.16	12.62	13.27	19.18
3	16.97	5.33	4.89	5.29	30.14
4	6.68	1.15	0.98	1.23	19.18
N	50,424	4,855	2,147	1,304	73

When looking at the recall over all topics, we see that *BOOL* and *LM* retrieved 48,096 relevant results out of 87,057 (55.25%), *LMS* returned 57,935 of 89,906 (64.44%), *NLLR* had a recall of 65.18%, and *OKAPI* returned most relevant results with 65.69%. It shows that the system using the Okapi model performs best with our Dutch document collection, and that exact matching using *BOOL* and *LM* 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-occuring names.

In summary, these results are in line with our expectations, namely that *BOOL* would be the worst-performing system, then *LM*, with *LMS* improving over *LM*, and that the differences among *LMS*, *NLLR*, and *OKAPI* would be 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.

4.3 Test Collection Based on Agreements

The log-based test collection has many more topics than a traditional test collection with 25-200 topics. While having thousands of topics opens us new uses, such as focusing on various breakdowns of the topic set even on relatively rare phenomena, it also presents an efficiency challenge: many DLs crumble under thousands of queries. We take into account the agreement that exists among different searchers. For example, when we pay attention to Table 1, this means that only EAD file “1.05.11.16” is included as a relevance judgement in the test collection, and the rest is discarded. We see in Table 4 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

Table 5. System-ranking of runs over topics with agreement

	Agreement	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)

size decreases to 4,855 from 50,424, and when we set the threshold to 4, only 1,304 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 5. We focus on the differences of the MAP scores when we take an agreement between two different IPs. The *BOOL* is significantly performing worst, and *OKAPI* is performing the best compared to either *LMS* with a significant improvement of 3.61% (t(4835) = 5.50, $p < 0.01$, one-tailed), or similarly an 3.35% significant improvement over *NLLR*. Although the difference between *LM* and *LMS* was only 0.25%, it was also significant (t(4835) = 2.40, $p < 0.01$, 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 *BOOL* is again significantly the worst performing system ($p < 0.01$), and *OKAPI* is significantly performing better on a 1% significance level. Interestingly, we see that *LM* is slightly performing better than *LMS* ($p < 0.05$). As Table 4 shows that when we increase the agreement threshold, there are only 2,147 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 *BOOL* being the worst, and *OKAPI* the best ($p < 0.01$). The findings are also consistent with the MRR and nDCG scores.

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

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

Table 6. System-ranking of runs over e-mail topics

	BOOL	**LM**	**LMS**	**NLLR**	**OKAPI**
MAP	0.1521 (5)	0.2632 (4)	0.3135 (3)	0.3147 (2)	0.3478 (1)
MRR	0.1732 (5)	0.3040 (4)	0.3550 (2)	0.3550 (2)	0.3907 (1)
nDCG	0.2430 (5)	0.3396 (4)	0.4386 (3)	0.4394 (2)	0.4623 (1)

```
<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.xml ; 2.05.19.xml ; 2.21.253.xml</files>
</topic>
```

Fig. 1. An example of a topic based on an e-mail reference request

human experts: responses of archivists to e-mail reference questions. The system rankings of the log-based test collection are compared to this ground-truth.

5.1 Test Collection Based on E-Mail Reference Requests

We analyze a subset of e-mails that the NA-NL 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. Eventually, we manually derive 73 different topics (and recommended EAD links) from the e-mail files. A typical example is the information request in Fig. 1.

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

5.2 System Rank Correlations

We again use the Paired Samples T-test to check for significance by looking at the MAP scores. The results of Table 6 show that *BOOL* performs worst as well ($p < 0.01$, one-tailed). When we rank with LM without smoothing, there is also a significant improvement of 73.04% over *BOOL* (t(67) = 3.22, $p < 0.01$, one-tailed). When we use LM extended with smoothing, we see a 19.11% significant improvement. However, the difference between the LMS and NLLR models was only 0.38%, and is not significant. Moreover, *OKAPI* performed 10.52% better than *NLLR*, but is not significantly better. The findings are similar using the MRR and nDCG measures.

How reliable are our test collections derived from log files when compared to the test collection manually derived from e-mails and experts' replies? When we compare the system rankings of the test collection from the whole log (Table 3) with the e-mail topics (Table 6) 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 high agreement among the test collections of the log-based evaluation approach.

6 Conclusions and Future Work

This paper investigated a search log-based approach to the evaluation of digital libraries. By using the DLs 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 set of EAD documents and search logs of an archival institution. This resulted in a test collection to evaluate the retrieval of digital archives. This extends initial experiments using a museum's log file to create a domain-specific test collection [1], by using a massive archival collection from the National Archives of the Netherlands, and a massive search log covering several years of this high-profile website. 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 determine the viability of the approach. We validated the results against 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.

In our future work, we will further refine the log-based approach to evaluation, by using more advanced click models and by filtering out interesting categories of search requests. We will also use our test collections for improving archival access, by developing retrieval models for archival descriptions. These methods will be applicable to all archival institutions publishing their finding aids in EAD, the *de facto* standard. In addition, we are currently extracting the navigation within the archival finding aids, allowing us to locate particular document-components or parts of the archive of user interest. This allows us to build an evaluation set for focused or sub-document retrieval.

Acknowledgments. We gratefully thank Henny van Schie of the National Archives of the Netherlands for providing the data, and for valuable discussion and feedback on early versions of this paper. This research is supported by the Netherlands Organisation for Scientific Research (NWO) under project # 639.072.601.

References

[1] Arampatzis, A., Kamps, J., Koolen, M., Nussbaum, N.: Deriving a domain specific test collection from a query log. In: LaTeCH 2007, pp. 73–80. ACL (2007)

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

[3] Boncz, P.A., Grust, T., van Keulen, M., Manegold, S., Rittinger, J., Teubner, J.: MonetDB/XQuery: A Fast XQuery Processor Powered by a Relational Engine. In: SIGMOD 2006, pp. 479–490. ACM, New York (2006)

[4] Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible markup language (XML) 1.0, 5th edn.(2008)

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

[6] Dupret, G., Liao, C.: A model to estimate intrinsic document relevance from the clickthrough logs of a web search engine. In: WSDM 2010. ACM Press, New York (2010)

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

[8] Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.-P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Sølvberg, I.: Evaluation of digital libraries. Int. J. on Digital Libraries 8(1), 21–38 (2007)

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

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

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

[12] Joachims, T., Granka, L., Pan, B., Hembrooke, H., Gay, G.: Accurately interpreting clickthrough data as implicit feedback. In: SIGIR 2005, pp. 154–161 (2005)

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

[14] Lalmas, M.: XML Information Retrieval. Encycl. of Lib. and Inf. Sciences (2009)

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

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

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

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

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

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

[21] Spärck Jones, K., van Rijsbergen, C.J.: Information retrieval test collections. J. of Documentation 32(1), 59–75 (1976)

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

[23] White, R.W., Morris, D.: Investigating the querying and browsing behavior of advanced search engine users. In: SIGIR 2007, pp. 255–262. ACM, New York (2007)

[24] Zhang, J., Kamps, J.: Searching archival finding aids: Retrieval in original order? In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 447–450. Springer, Heidelberg (2009)

[25] Zhang, J., Kamps, J.: Focused search in digital archives. In: Vossen, G., Long, D.D.E., Yu, J.X. (eds.) WISE 2009. LNCS, vol. 5802, pp. 463–471. Springer, Heidelberg (2009)