View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CWI's Institutional Repository

Metadata categorization for identifying search patterns in a digital library

Tessel Bogaard[1], Laura Hollink[1], Jan Wielemaker[1,2], Jacco van Ossenbruggen[1,2], and Lynda Hardman[1,3]

[1] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{Tessel.Bogaard, L.Hollink, J.Wielemaker, Jacco.van.Ossenbruggen, Lynda.Hardman}@cwi.nl
[2] Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[3] Universiteit Utrecht, Utrecht, The Netherlands

Structured abstract.
Purpose: For digital libraries, it is useful to understand how users search in a collection. Investigating search patterns can help them to improve the user interface, collection management and search algorithms. However, search patterns may vary widely in different parts of a collection. This study demonstrates how to identify these search patterns within a well-curated historical newspaper collection using the existing metadata.
Design/methodology/approach: The authors analyzed search logs combined with metadata records describing the content of the collection, using this metadata to create subsets in the logs corresponding to different parts of the collection.
Findings: The study shows that faceted search is more prevalent than non-faceted search in terms of number of unique queries, time spent, clicks and downloads. Distinct search patterns are observed in different parts of the collection, corresponding to historical periods, geographical regions or subject matter.
Originality/value: First, this study provides deeper insights into search behavior at a fine granularity in a historical newspaper collection, by the inclusion of the metadata in the analysis. Second, it demonstrates how to use metadata categorization as a way to analyze distinct search patterns in a collection.
Keywords: Digital libraries, Newspapers, Information-seeking behavior, Log analysis, Metadata
Article classification: Research paper

1 Introduction

Log analysis is an unobtrusive technique for macro-analysis of user behavior in digital search systems (Hollink et al., 2011; Spink and Jansen, 2006). It contributes to an understanding of the information needs of users and to what extent these needs are met. Results based on log analysis may be used for the evaluation of search algorithms, (re-)design of user interfaces, and to identify potential gaps in the underlying document collection. User behavior in general

web search is well-studied, (Baeza-Yates et al., 2005; Beitzel et al., 2004; Downey et al., 2007; Jansen and Spink, 2006). However, in search engines providing access to a specific type of content or collection ("vertical search engines"), the search functionality is often different, hence, user behavior can be expected to differ. This has been shown, for example, for image archives (Han and Wolfram, 2015; Hollink et al., 2011), a medical knowledge portal (Callahan et al., 2015), a newspaper archive (Gooding, 2016), and in a study of a digital library (Niu and Hemminger, 2015).

Our work is carried out in the context of the online search interface to the historical newspaper collection of the National Library of the Netherlands. The documents in the collection are described with rich, professionally curated bibliographic metadata about their format and origin. The search interface providing access to the documents is typical for a digital library: in addition to regular query input for full text search, users can filter search results based on selected metadata values using *facets* (Hearst et al., 2002). Curators at the National Library of the Netherlands are interested in understanding how users search within their historical newspaper collection. This will allow them to provide improved search features for user groups with specific tasks searching in different parts of the collection. This study therefore addresses the following research question: *How do search patterns differ among users searching in different parts of the collection?*

Previous work has used categorizations of the queries found in logs to find distinct search patterns, for example in the study of religious search relating to five religions (Wan-Chik et al., 2013), or an investigation of different types of learning in search (Eickhoff et al., 2014). Query analysis, however, suffers from various disadvantages. Queries are ambiguous, as they form an uncontrolled vocabulary with little context to interpret the underlying information need. Most queries appear infrequently in the logs. As a consequence, when investigating patterns of queries and clicks, even the most frequently occurring patterns occur infrequently. Furthermore, queries may contain privacy-sensitive information (Jones et al., 2008). We propose to use the metadata instead to investigate different search patterns in a historical newspaper collection. The metadata values of clicked documents and the corresponding facet values come from a controlled vocabulary. We can observe search patterns by grouping individual, unique queries based on facet values. Likewise, (long tail) clicked documents can be grouped by their associated metadata values. Moreover, metadata values of facets and clicked documents are less privacy-sensitive than queries entered by users.

We start with an analysis of faceted search versus non-faceted search to investigate the role of facets in search. Our results show that faceted search (57% of all search) is responsible for the larger part of time spent (median session duration of over an hour versus less than 10 minutes), the majority of unique queries (79%) and documents clicked (78%) and downloaded (72%). We create subsets based on the metadata of facets selected in search, using the selected facet values as a proxy for user interest.

We find distinct search patterns based on the kind of facet selected: publication date, item type, or geographical region. For example, users searching within World War II keep returning to the platform over an extended period of time (median session duration eight days) and click and download many documents (median of 25 clicks, 31% of sessions includes a download). Many users are interested in family announcements (18% of all sessions), with visits that are typically highly focused on the subject matter and contain relatively few clicks. Search for Surinam, though not as popular, is also very focused, with almost all clicks on documents from this part of the collection (84%) in these comparatively shorter visits (median session duration of just under five hours).

The contribution of this paper is twofold. First, we provide detailed insights into user behavior in a historical newspaper collection, observing distinct search patterns within different parts of the collection. Based on our findings, we are able to formulate concrete suggestions for improvement of the online search platform of the National Library: suggestions for improvements to the user interface, recommendations for a different default setting of parameters, and recommendations for prioritization of their ongoing digitization efforts. Second, we illustrate how metadata can be used to analyze behavior in a digital library or archive. As such, it enables us to do a comparative analysis of (1) what users search for (from the faceted query log data), (2) what they find (from click log data), and (3) what is or is not present in the collection (from collection metadata).

2 Related work

Diverse studies have used *log analysis* to gain a general understanding of search behavior in digital libraries and archives. In 2000, Jones et al. described the general search behavior in a library of computer science technical reports (Jones et al., 2000). They presented user demographics (multiple countries of origin), discussed use of operators (used in about a third of queries), common terms in queries, number of views per query (mostly zero or one), and length of visits (average of about ten minutes, with more than half around 5 minutes). Mahoui and Cunningham (2001) found similar results in a comparison to a different digital library for computer science researchers in a larger dataset gathered in the same period over a shorter interval. Sfakakis and Kapidakis (2002) distinguished different search patterns for search in various collections – ranging from medical bibliography, and archaeological records, to PhD dissertations – of the Hellenic National Documentation Center in terms of average session length (mostly short sessions with about three interactions) and use of certain search fields, such as any, author, title. More recently, Gooding (2016) showed differences between online and off line search behavior in a Welsh newspaper archive, describing online behavior in terms of number of visits, browsing and viewing content, and time spent on search (about 17 minutes per session and visiting over 20 pages per visit). The data used combine Google analytics with log analysis. Niu and Hemminger analyzed search in a faceted search interface providing access to a digital library (Niu and Hemminger, 2015), combining log analysis with a user

study, observing different search patterns for faceted and non-faceted search, where faceted search, occurring in about 12% of the sessions, correlated with shorter queries (2.6 versus 3.2 terms per query).

User studies have also been used to better understand search behavior in digital libraries, such as in a combination with log analysis as mentioned above (Niu and Hemminger, 2015), where the user study demonstrated that the facets were valued and utilized especially in the context of more exploratory, open-ended search and improved the accuracy of the search. In another user study the focus was on a broader context of search, modeling the search behavior of the growing group of non-professional genealogists and family history researchers in terms of type of search, preferred resources, and different phases of research (Darby and Clough, 2013). Darby and Clough found that these users return to their search and preferred resources frequently, that the search is ongoing and open-ended, and resources such as newspaper collections are used more in a later stage of the research. Another study used pop-up surveys to investigate the motives for search within a cultural heritage site (Clough et al., 2017).

Our analysis of logs of the National Library of the Netherlands, investigating the use of the historical newspaper collection, is different from these studies as it focuses on finding fine-grained search patterns within different parts of a single collection in terms of the metadata descriptions of the collection, as opposed to the more general, over-arching search patterns described above.

To characterize search behavior from log records, individual log records are usually grouped into *sessions*. Session-level analysis captures the context in which individual user actions occurred: it connects search interactions to clicks and partly conveys the user's effort in terms of number of actions and time spent.

Sessions can be defined in several ways, for example using the IP address as a proxy for a user. Even so, using only the IP address can be problematic as there can be multiple users behind a single IP address. In an access-controlled portal a session can be based on login (Callahan et al., 2015), or alternatively, an HTTP cookie can be used (Gooding, 2016). This improves on using only the IP address, as login credentials and HTTP cookies both should uniquely identify a user. Still, login credentials may be shared or the same user may switch devices during a search with different HTTP cookies on each device. Moreover, not all search platforms require login or record HTTP cookies in the logs.

Sessions can also be defined based on queries. For example, in Guo et al. (2009) a session is defined as a single user query and the subsequent clicks; and in Huurnink et al. (2010) a session is dependent on the presence of overlapping terms in consecutive queries. This has the advantage that queries and clicks in succession can be linked. Even so, a single user might interleave several search tasks (Agichtein et al., 2012) and a session might be broken off incorrectly.

Frequently sessions are bounded by a period of inactivity. The length of this timeout is often thirty minutes, mentioned as an established approach in Eickhoff et al. (2014) and (Niu and Hemminger, 2015) and finding its origin in a study of browsing behavior in 1994 (Catledge and Pitkow, 1995). Other examples of sessions defined by a timeout are Hollink et al. (2011)(15 minutes); Chapelle and

Zhang (2009) (60 minutes); and Jansen and Spink (2006) or Han and Wolfram (2015) where sessions were bounded per day. While this is a straightforward method to identify sessions, it does not solve the possibility of joining several users behind a single IP address in a single session. Furthermore, the length of the timeout is hard to choose correctly if the goal is to identify search tasks of a user (Jones and Klinkner, 2008).

In the context of studying web navigation, the concept of a *clickstream* is often used, as in Wang et al. (2016). A clickstream is the navigational path a user follows, consisting of consecutive HTTP requests from a single IP address. The clickstream model can help to untangle multiple users behind a single IP by splitting up separate sequences of interactions occurring (possibly at the same time) from the same IP address. Nevertheless, this could result in wrongly breaking up a session of a single user searching from different tabs in a web browser.

We have identified the sessions based on a clickstream model, as the logs do not contain HTTP cookies and the platform does not require a login.

Grouping sessions makes it possible to find different search patterns. In Niu and Hemminger (2015) sessions are grouped into faceted and non-faceted search sessions. Other studies have used query analysis to find fine-grained search patterns, for example to investigate religious information-seeking related to five main religions (Wan-Chik et al., 2013), or to study different types of learning in search (Eickhoff et al., 2014). However, query analysis has various disadvantages. First, queries can be ambiguous. For example, it is virtually impossible to know whether someone who enters the query "Oudkerk" is interested in stories about the Dutch politician, news related to the Frisian village, or announcements regarding births, deaths or marriages in one of the many Oudkerk families. Second, most queries are in the *long tail*, i.e. they appear infrequently in the query logs. As a consequence, when investigating patterns of queries and clicks, even the most frequently occurring patterns occur infrequently. Finally, queries may contain privacy-sensitive information. Even after removing identifying information users can often still be identified (Jones et al., 2008). This leads to a conflict between protecting the privacy of users and retaining or publishing query logs, as mentioned in Cooper (2008). Techniques such as differential privacy (Dwork, 2006) – a mathematical model for maximizing accuracy while at the same time minimizing chance of identification – do improve the privacy of the user, however the resulting logs do not have the same utility (Korolova et al., 2009). Two recent papers (Hong et al., 2015; Zhang et al., 2016) aim for methods of applying differential privacy to retain the utility of the logs for analysis of query-click pairs while protecting the privacy of the users. Even though these approaches focus on query-click pairs and cannot be transferred to a different dataset, they do recognize the need for privacy protection.

We take a first step towards a more privacy-preserving method of analysis by grouping sessions based on a metadata categorization as present in the facet values instead of a categorization of the queries. The query is only analyzed for

its number of occurrences between sessions, a term count, and use of operators such as AND, OR, NOT, and quotes.

3 National Library of the Netherlands

We present the materials that were used for this study: the library collection and bibliographic metadata, the platform providing online access to the collection, Delpher, and the recorded usage logs.

Library collection and metadata The National Library of the Netherlands curates a historical newspaper collection[1]. This collection is – as self-described on the platform – targeted at researchers of any type, such as scholars, students, journalists and genealogists. It contains over 100 million newspaper documents published in about 1500 newspaper titles between 1618 and 1995. These documents have been scanned and digitized for online access. Users can retrieve entire newspapers, newspaper pages, or individual items on the page, where the last can be one of four types: news articles, advertisements, announcements (relating to family such as birth, marriage or death announcements) or images (illustrations or photographs, where search is done on the caption text).

The documents in the collection are described in bibliographic metadata records with the following attributes: a document identifier, the publication date, item type, newspaper title, place of publication, source (the physical location of the original document), and distribution zone. The distribution zone attribute represents the geographical region where the newspaper was distributed, with values "local", "national", one of the former Dutch colonies ("Indonesia", "Surinam", or the "Antilles"), or, in a few cases, "*unknown*".

Online access The newspaper collection is accessible through the Delpher platform[2]. In the Delpher search interface (see Fig. 1) the facets are filters based on metadata attributes and values of the documents. The facets visible in the figure, from top to bottom, are time facets ("Periode"), where a user can refine search by century, then by decade and by year, up to an exact date; distribution zone ("Verspreidingsgebied"); and type of newspaper item ("Soort bericht"). Users may change the default relevance ranking of results ("Sorteer op: relevantie") to alphabetical ordering by item title or by newspaper title, or to chronological ordering (ascending or descending). From a search results page, a user may click on a document in the result list and, after a click, may decide to download the document. A download can be a scanned image, a digitized text, or a bibliographic reference of the document.

Search logs The web server of the Delpher search platform logs HTTP page requests of its users. Under a strict confidentiality agreement the National Library of the Netherlands has provided us with the log records collected from October 2015 until March 2016. These around 200M records include encoded IP addresses, time of the requests, user agents (identifying client software), referrer

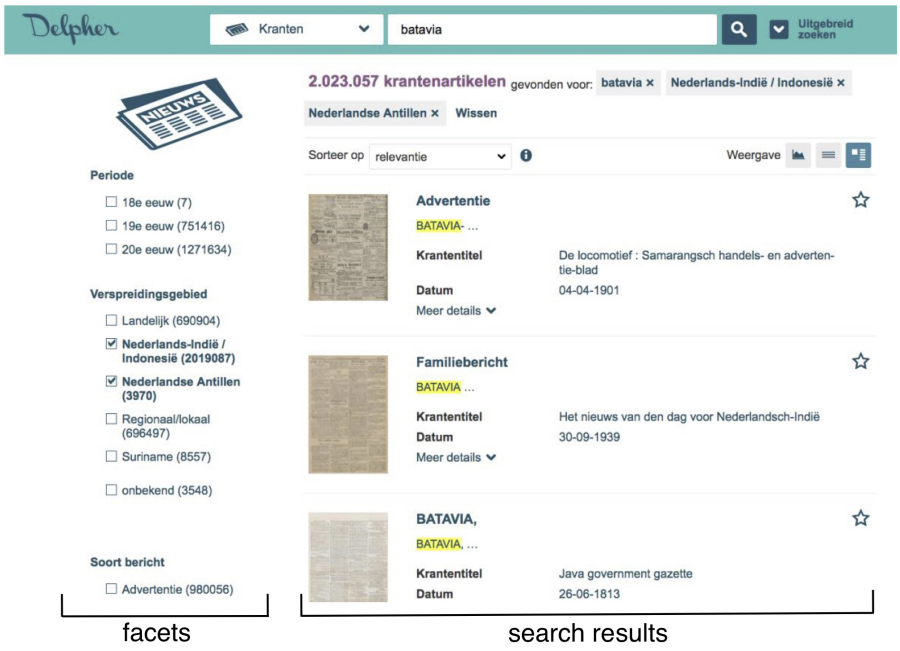

Fig. 1. Search interface for the newspaper collection, with facets to the left and search results to the right.

URLs (URL where request originated), and the URLs of the requested HTTP pages. The IP addresses are hashed (obfuscated) to protect the privacy of users, and have only been used to help define sessions. The URL of a requested page contains a document identifier in case the request was for a document. In case the requested URL is a search results page, we extract from it (1) the query string, (2) any facets used, and (3) the result ranking method, together representing what we call the user's search interaction.

4 Method

To be able to discover search patterns in different parts of the collection, we start with identifying sessions in the logs, then we add session properties, and finally we create subsets of sessions based on the bibliographic metadata values. We use these subsets to compare and analyze specific search patterns.

4.1 *Step 1: Session identification in search logs*

As described in the Section 2, a session can be defined in different ways, depending on the information available in the logs. For this study we have chosen

a clickstream-based model, using the (hashes of) IP addresses and the referrer URLs to combine individual interactions into a session. The referrer URL helps to connect records, matching the referrer URL to a (previously) requested URL found in the records. We have selected this approach for a few reasons. First, we expect a possibly large proportion of users to be engaged in exploratory, open-ended search (as is the case, for example, for genealogists and family historians as described in Darby and Clough (2013)), thus using a timeout might result in breaking up visits that occur with long pauses. Second, the historical newspaper collection is accessible without login, and the server does not log HTTP cookies. Third, as our focus is not on the query this is not an obvious choice for our session definition. Finally, using the referrer URLs to link interactions is a relatively straightforward way to define sessions, trying to avoid combining multiple users into a single session and keeping sessions of users returning to their search over a longer period intact, even if we might break up sessions of users searching in different tabs: an HTTP request using "open in new tab" might still be connected to the previous user interactions, however a copy-paste of a URL is not.

Search log data cleaning Consecutive visits of the same URL are removed as this is likely a reload of the web browser and not a new action by the user. Thus, a reload of a document is not counted as a second click. As we are interested in user behavior, we remove all records stemming from web crawlers[3]. Web crawlers are identified based on the user agent or a request for robots.txt, and records with matching IP address are filtered out.

Since our aim is to analyze search behavior, we only analyze sessions that include a search interaction within the newspaper collection. This means we exclude sessions that contain only clicks (following deep links, for example), or visits to the homepage. Additionally, sessions that consist of only a single interaction are also discarded. The remaining 204,125 sessions consist of 17,053,823 search interactions (of which 6,000,589 search interactions include facets); 6,430,674 clicks on documents and 574,831 downloads.

4.2 *Step 2: Computing session properties*

Next, we add for each session a set of properties that we use in the analysis:

1. session duration (computed as the time interval between the first and last interaction in a session)
2. number of queries, number of queries using quotes, and number of queries using boolean operators[4]
3. number of clicked documents and their metadata values
4. number of downloaded documents and their metadata values
5. number of facets selected and their metadata values

We report aggregate session properties for the entire dataset and for specific subsets of the data. As most of our data has a skewed distribution, with high

outliers, we report the median instead of the mean (Hoaglin, 1983). The median values are session duration and number of queries and clicks per session. In addition, we report a percentage of sessions with at least one download, sessions with a query using quotes, and with a query using boolean operators. We use percentages for these last three, as they occur in less than half of all sessions and a median value would always be zero. In addition, we include absolute numbers of clicks and downloads.

4.3 *Step 3: Grouping and analyzing sessions*

To study the different information-seeking behaviors, we create subsets in the dataset. First, we compare the session properties of sessions with and without facet use, to investigate whether the use of facets plays an important role in search. In addition, we analyze how often queries reoccur in different sessions, and in which subsets of sessions the unique, *long tail* queries occur.

Next, we use metadata values to create subsets in the sessions and compare the resulting subsets. We can do this based on (1) metadata values of facets selected in a session, or based on (2) the metadata values of clicked documents, depending on the results of the previous step, whether faceted search plays a sufficiently important role in search. The aim here is to discover whether search patterns are different for users interested in different parts of the collection. For example, we can investigate behavior of users searching for family histories by taking the subset of sessions that include a search interaction with the facet $\langle item\_type = announcement \rangle$. We compare the session properties in this subset with those found in other subsets, e.g. we compare them to the session properties of the subset of sessions that include $\langle item\_type = article \rangle$. Note that subsets may overlap as one session can contain multiple facets.

Lastly, we compare the popularity of the various metadata values to how often documents with the corresponding values are clicked on, downloaded, and how often they appear in the collection. For example, to put the sessions with the facet value $\langle item\_type = announcement \rangle$ into perspective, we compare their number to the number of clicks on announcements, downloads of announcements and the number of announcements in the entire collection.

4.4 Limitations

While log analysis is a good technique for obtaining a general understanding of user behavior identified in search patterns, it cannot explain *why* users follow these patterns. Further research would be needed to uncover their reasons and motivations.

We have focused on session level analysis to bring the user interactions into a context, as opposed to providing an analysis at the level of the individual interactions. However, any session definition has limitations as well. We have chosen a clickstream-model session definition, and while this might keep the interactions together of a user continuing a search over multiple days, it still could in some cases break up the search of a user searching in multiple tabs.

Moreover, the dataset puts some constraints on what we can analyze. The hashing of the IP addresses makes it impossible to provide demographics over who visits the historical newspaper archive. The ranking of clicked results is not logged, thus an analysis of the depth of clicked results is not possible. The subsets we create are bounded by the metadata categories available in the collection, possibly other categorizations could be of interest as well. In addition, we have made the choice not to analyze the query in detail.

5 Results

We first provide some general statistics of visits to the newspaper collection. Then, we investigate faceted search and look at the frequency of use of the three main search facets presented on the platform, to determine how the use of facets correlates with other search behavior. Finally, we analyze user behavior in more detail by focusing on a few specific use cases, the information-seeking behavior of users interested in genealogy and family history, in Surinam (one of the former Dutch colonies), or in World War II (WWII). To find these search patterns, we use the relevant metadata values present in sessions as a proxy for user interest in that specific part of the collection to create subsets within the sessions. Based on the observed search patterns we give concrete recommendations to the National Library which are included at the end of each subsection, demonstrating the effectiveness of extending log analysis with facet usage and collection metadata.

5.1 *Visitor statistics*

The portal is accessed consistently over the days of the week (Fig. 2), in contrast to the observations of Jones et al. (2000), Ke et al. (2002), and Huurnink et al. (2010), where there was a significant drop in usage in the weekend. When we plot session start times, we see that usage starts to peak in office hours, and continues into the evening with only a small drop around 18:00 (Dutch dinnertime). Both findings suggest a mix of professional and amateur researchers visiting the platform.

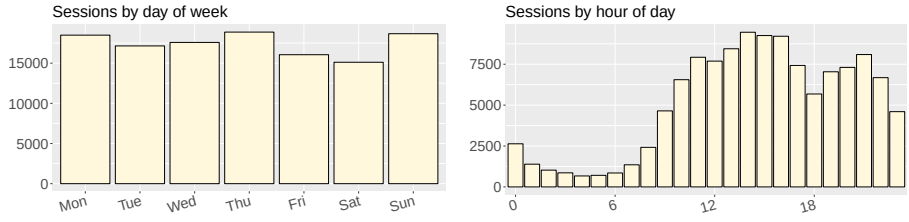

Fig. 2. Number of sessions over the days of the week and the hours of the day

5.2 *Faceted search*

Table 1 summarizes session properties and facet use. Facets are used in 57% of the sessions, higher than the 12% Niu and Hemminger observed in a university library catalog (Niu and Hemminger, 2015). Time facets are most popular (40%), followed by item type facets (31%) and distribution zone facets (26%). We observe that sessions in which facets are used are much longer (median of 1:05:32 versus 9:32 without facets), and contain more queries, clicks and downloads. The 57% sessions including facets contain 78% of all clicked and 72% of all downloaded documents. Moreover, 80% of sessions with faceted search lead to clicks, whereas this is 69% of sessions without facets. In total 75% of all sessions include a click on a document. The 25% of sessions not leading to a click are very short sessions (a median duration under 2 minutes), and on average consist of a single query.

Table 1. Session subsets overview

Sessions	Frequency		Clicks		Downloads	
all	204,125		6,430,674		574,831	
- without facets	87,348	43%	1,410,385	22%	159,400	28%
- with facets	116,777	57%	5,020,289	78%	415,431	72%
- - time facets	81,321	40%	3,480,966	54%	281,750	49%
- - item type facets	64,272	31%	3,748,762	58%	309,294	54%
- - distr. zone facets	52,927	26%	3,064,239	48%	254,689	44%
- without clicks	50,226	25%	0	0%	46	0.008%
- with clicks	153,899	75%	6,430,674	100%	574,785	100%

	Median duration	Median queries	Median clicks	Including downloads	Including quoted query	Including boolean query
all	24:50	3	3	12%	19%	2.3%
- without facets	9:32	2	2	11%	15%	1.7%
- with facets	1:05:32	4	6	18%	21%	2.9%
- - time facets	1:17:38	4	6	18%	22%	2.8%
- - item type facets	9:35:59	6	10	21%	25%	3.6%
- - distr. zone facets	3:26:51	5	9	21%	20%	3.1%
- without clicks	1:35	1	0	0.04%	11%	1.6%
- with clicks	1:11:10	4	7	20%	21%	2.6%

Queries Queries are short, mostly two terms. We observe a slight difference between the queries with and without facets: with facets the mean number of terms in a query is 2.2; without the mean is 2.4. Similarly, Niu and Hemminger observed a lower mean for faceted search, 2.6 terms versus 3.2 for non-faceted search (Niu and Hemminger, 2015). In a photo archive of a news agency, Hollink et al. found an even lower number of terms in queries (mean of 1.8) (Hollink

et al., 2011). In contrast, in open web search an average of four terms per query is not uncommon[5]. This suggests a different type of usage in specialized search engine, and especially news archives with a higher likelihood of search for named entities and fewer natural language queries.

Another indication for named entity search is the relatively frequent use of quotes (19% of sessions include a query with quotes, see Table 1). Boolean operators are less frequently used (in only 2.3% of all sessions). The use of quotes and of boolean operators again occurs more often in faceted than in non-faceted search (21% versus 15% of sessions uses quotes, and 3.6% versus 1.7% boolean operators). This is even stronger for the 31% sessions using an item type facet value leading to 58% of all clicks. 25% of these sessions use quotes, and 3.6% boolean operators. When we analyze the number of occurrences of queries, we find that 96% of queries occur only in a single session. Moreover, 79% of these queries occur in faceted search. These findings demonstrate the importance of faceted search in this historical newspaper collection.

Reranking of results The search interface default setting is to rank search results by relevance. We observe that in 24% of all sessions, at some point, the user selects the option to rerank the results by time. This option is used more often in sessions using facets (29% of these sessions) than in sessions not using facets (16%). The frequent use of this option suggests that the default relevance ranking alone does not suffice for a large group of users.

Overall, we observe that most actions come from sessions using faceted search, the sessions are longer, contain more complex, and unique queries, use search options more often and generate the majority of the clicks and downloads. Thus, we will create subsets in the sessions based on the metadata of the facets used.

Recommendations Since many users reorder the results by time, a suggestion would be remembering the preference within a session or providing an option in user preference settings for a default ranking by time. Another suggestion could be a timeline visualization of the results. As a matter of fact, such a visualization of results has become part of the search interface since June 2016.

5.3 *Genealogy and family history search*

In this section we focus on users selecting the family announcement facet value, to gain insight into the behavior of users interested in genealogy and family history in the collection. We use a comparative analysis of the sessions subsets by item type. The item values are one of article, advert, announcement, and image. Table 2 summarizes the session properties per item type.

Search behavior The announcement value is the most frequent item type value selected, in 18% of all sessions. The sessions are shorter than the other sessions using item type facets, and generate fewer clicks and downloads. The number of distinct queries per session is not high with a median of 7 queries. However, 47%

Table 2. Session subsets by item type facet values

Sessions	Frequency		Clicks		Downloads		Clicks on value	
- article	28,442	14%	2,252,398	35%	214,957	37%	1,106,441	36%
- advert	16,045	8%	1,695,772	26%	139,949	25%	386,900	37%
- announcement	37,733	18%	2,554,849	40%	169,166	29%	1,074,282	80%
- image	7,461	4%	875,964	14%	68,249	12%	70,200	47%

	Median duration	Median queries	Median clicks	Including downloads	Including quoted query	Including boolean query
- article	1d 15:08:41	7	14	28%	25%	4.7%
- advert	5d 16:53:46	10	22	30%	30%	5.3%
- announcement	1d 6:27:38	7	11	21%	30%	3.4%
- image	7d 7:12:28	12	28	34%	29%	5%

of the long tail, single-occurrence queries are found in these sessions. Quotes are used relatively frequently in family search, even if boolean operators are not used as often as for the other item type values. Interestingly, these sessions have about the same number of queries per session as the sessions using the article facet value, even while fewer results are clicked or downloaded. This could be because the relevance and content of the short announcements can often be assessed from the result page snippets, without actually clicking a document. In sessions where the announcement facet value was selected, many clicks are on announcements (1M of the 2,6M clicks), making these sessions more focused than most sessions involving the other types. For comparison, in sessions that include the image value, less than 10% of the clicks are on images (70k of the 876k clicks). This indicates that users searching explicitly for announcements have less interest in results of other types. At the same time, relatively few announcements (20%) are found in sessions not using that facet. For comparison, 64% of all articles are clicked in sessions not using the article facet at all (see the "Clicks on value" column in Table 2). This suggests that announcements could be hard to find unless the corresponding facet has been selected, while articles are also found and clicked without the help of the corresponding facet.

An analysis of the documents that were clicked and downloaded confirms that search for announcements follows a different pattern than search for the other items. Where announcements are just 2% of the collection, the percentage of clicks on announcements is much higher at 24% (Fig. 3). When we investigate downloads, on the other hand, most notable are the high proportion of article downloads and the low proportion of announcement downloads. This low proportion may be due to their short length, making it easy to write them down or copy and paste them.

Altogether, this part of the collection receives a high user interest. The frequent visits are comparatively quick and strongly focused on announcements. Many of the unique queries appear here, and a high number of sessions uses quotes for

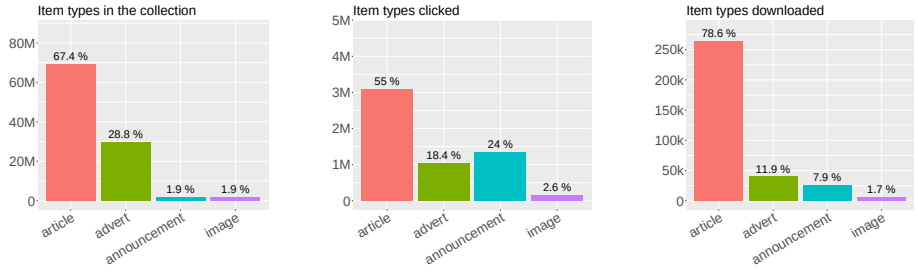

Fig. 3. Item types in the collection (103 million documents), of clicks (5.6 million) and of downloads (335 thousand).

the queries. Nevertheless, these sessions have fewer clicks on average than some of the others, and only few of the clicks are downloaded.

Recommendations Snippets of announcements as they appear in the results set have added value. Announcements receive a lot of user interest, so our recommendation for the library is to give snippets of these short items extra attention. For articles, which are typically much longer, this is probably not as useful since people are more likely to click on a result to scan or read the full text.

Another suggestion would be to consider prioritizing post-correction of the digitized announcements: user interest is high; the total volume is low at 2% of the collection; and announcements are potentially more impacted by OCR mistakes since entity names can have unique spelling variations.

5.4 *Search for Surinam*

In this section we focus on users interested in publications from Surinam, one of the former Dutch colonies. To do this, we will investigate users selecting the Surinam distribution zone facet value. The distribution zone is the geographical region where a newspaper is distributed. This facet is selected in 26% of all sessions. Table 3 summarizes the session properties and Figure 4 compares the occurrence of the relevant metadata values in the collection, in clicked results and in downloaded documents. The most popular value here is the local distribution zone facet value, used in 13% of all sessions. This may be connected to the relatively high user interest in family announcements discussed in the previous section, which frequently appear in local newspapers. The *unknown* facet value is least popular, and appears in very long sessions with many queries, clicks and downloads, in combination with other facet values. However, only 2% of the clicks on the *unknown* value occur in these sessions, and most clicks here are on the other values.

Search behavior The distribution zone facet appears to be needed to retrieve documents from particular, smaller subsets of the collection. While well over 60%

Table 3. Session subsets by distribution zone facet values

Sessions	Frequency		Clicks		Downloads		Clicks on value	
- national	21,325	10%	1,620,113	25%	139,582	24%	639,817	29%
- local	27,050	13%	1,797,927	28%	146,184	25%	1,138,093	34%
- Indonesia	10,930	5%	882,072	14%	71,860	13%	340,384	61%
- Antilles	2,930	1.4%	289,256	4%	25,751	4%	43,234	49%
- Surinam	4,004	2%	334,013	5%	20,857	4%	128,334	84%
- *unknown*	861	0.4%	112,268	2%	7,385	1%	63	2%

	Median duration	Median queries	Median clicks	Including downloads	Including quoted query	Including boolean query
- national	1d 0:24:16	6	12	25%	22%	4.3%
- local	17:51:49	6	10	22%	21%	3.1%
- Indonesia	21:56:13	7	15	27%	25%	3.4%
- Antilles	23:56:51	8	19	31%	23%	3.3%
- Surinam	4:44:02	6	14	24%	19%	3.2%
- *unknown*	6d 23:48:46	13	37	33%	24%	2.7%

of the clicks on national and regional articles are from sessions without using the corresponding facets, only 16% of the clicks on articles from Surinam are from sessions not using the Surinam facet value. The Surinam value is selected in 2% of all sessions. This interest in Surinam is higher than is to be expected from the size of the Surinam collection (only 1.2%). The number of clicks on documents from Surinam is in line with the number of sessions including this facet value (only 2.4%), but the percentage of downloads is quite low (only 1.5%, see Fig. 4). The total number of clicks in these sessions is not as high as for some of the other values, nevertheless the focus is on documents from Surinam (with 128k of the 334k clicks). The queries are a bit shorter than average, with a mean query length of 1.97 terms, and fewer sessions include quoted queries (19%). We find 5% of the single-occurrence queries in these sessions.

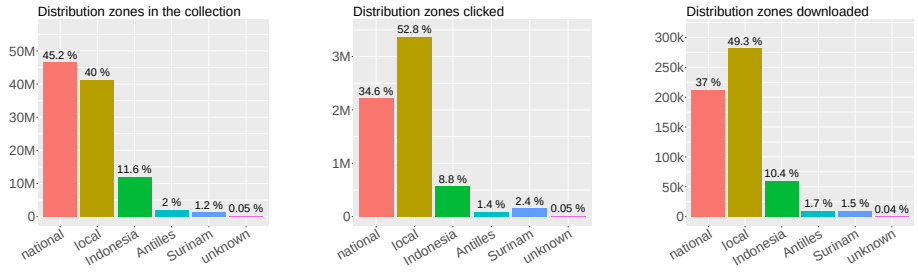

Fig. 4. Distribution zones in the collection (103 million documents), of clicked documents (6.4 million), and downloads (575 thousand).

Overall, we find that search for Surinam occurs in relatively short and not very complex sessions. Hardly any documents from Surinam are clicked outside these sessions, suggesting the facet is needed to find the documents. We hypothesize that users interested in Surinam have more difficulty finding what they are looking for.

Recommendations The relatively low number of clicks and downloads for the Surinam value – despite a user interest – could reflect a problem. A suggestion to the National Library here would be to investigate potential causes. It could be that user expectations need to be moderated. The relevance ranking could be performing non-optimally here. Or OCR quality could be more problematic for this part of the collection and OCR post-correction is needed.

5.5 *Search within World War II*

Time facets are the most popular, selected in 40% of all sessions. Since WWII was a pivotal time in Dutch history that the National Library of the Netherlands prioritizes, for example in digitization of the resistance's illegal press, we zoom in on this period to investigate how users search for these documents.

Search behavior Sessions with time facets are not as long as sessions with item type or distribution zone facets (a bit over one hour versus more than nine and three hours respectively, see Table 1). However, sessions with time facet values within the years of WWII (1940 to 1945 in the Netherlands) are much longer with a median of more than eight days (Table 4). These sessions contain more queries, clicks and downloads. In these 3% of all sessions, we find 26% of all clicks on WWII documents. In addition, 13% of the single-occurrence queries occur here. Quotes are used frequently (in 30% of the sessions), as are boolean operators (4.1%).

Table 4. Session subset by time facet value

Sessions	Frequency	Clicks	Downloads	Clicks on value
- WWII facets	5,563 3%	694,989 11%	52,395 9%	133,231 26%

	Median duration	Median queries	Median clicks	Including downloads	Including quoted query	Including boolean query
- WWII facets	8d 0:22:19	12	25	31%	30%	4.1%

The relatively high user interest in announcements that we observed in the overall collection is even more pronounced for the WWII period: announcements receive almost 32% of the clicks while they still make up only 2% of the collection (Fig. 5).

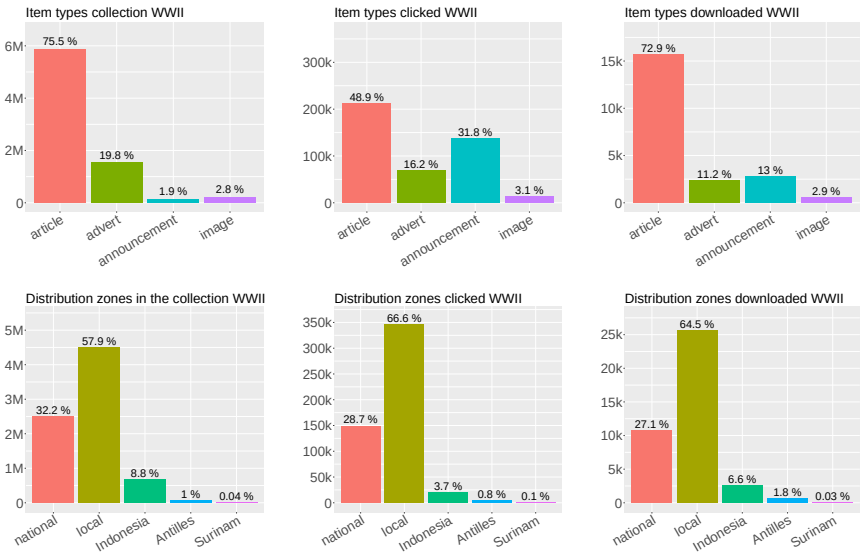

Fig. 5. Item types and **Distribution zones** of clicks (434 thousand items, 520 thousand documents) and downloads (40 thousand items and 40 thousand documents) and the collection (7.8 million documents) in World War II.

On the whole, search within WWII is more complex, with a high number of unique queries and many sessions including quoted queries or boolean queries. Moreover, the sessions have a long duration and the number of clicks is high.

Recommendations If we take the long session duration with many clicks and downloads as an indication that users are highly engaged and perform successful searches, this would suggest that the National Library's prioritization of the WWII period pays off. A further extension of the collection with documents from the postwar period would probably interest users.

Since there is clear user interest in the WWII period, a suggestion to the National Library would be to consider using special time facets to easily filter for specific periods in history; a WWII facet value might very well be of interest to the users.

As for the even more pronounced user interest in announcements, this strengthens our earlier recommendation to consider improving snippets for these items.

6 Conclusions

We have presented an analysis of fine-grained search patterns within a historical newspaper collection using metadata categorizations. The analysis method deploys metadata as a shared vocabulary to compare the logged (faceted) search behavior, the clicked results and the collection. Focusing on the metadata of

facets and clicked results instead of on the query, we alleviate the disadvantages of query-level analysis. Facets are not ambiguous like queries. We are able to isolate and observe search patterns by grouping long-tail queries based on shared facet use. Finally, facets are less privacy-sensitive than user-entered queries.

We have observed distinct search patterns that are not visible from overall usage statistics. Faceted search is more prevalent than non-faceted search and follows a different pattern: sessions that include facets are typically longer, contain more clicks and downloads and more unique, shorter keyword queries. Some parts of the collection stand out with an increased user interest. Documents from WWII, for example, are frequently searched and appear in very long sessions with many clicks and a high proportion of unique queries, signifying highly engaged users. The family announcements are also disproportionately popular in search, confirming the assumption of the National Library that genealogists and family historians constitute a high proportion of their user base. Smaller parts of the collection are hard to find without using the corresponding facets. This applies, for example, to the family announcements and to documents from Surinam. Based on the observed patterns, we were able to give concrete recommendations to the Library about improvements to the user interface, a different default setting of search parameters, and for prioritization of their ongoing digitization efforts.

We expect that this approach can be used for any faceted search system for collections with curated metadata. Also, this approach could potentially be a starting point for inter-collection comparison of user search behavior for digital libraries or archives sharing similar metadata categories. Future work will concentrate on a more data-driven method to find fine-grained search patterns in a curated collection.

Acknowledgments This research was partially supported by the VRE4EIC project, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 676247. The computational part of the research has been carried out on the SWISH DataLab software infrastructure developed within the VRE4EIC project (Bogaard et al., 2017). We thank the National Library of the Netherlands for providing access to their data and feedback on earlier drafts of this paper.

Notes

[1]More information about the National Library of the Netherlands can be found at the following URL: `https://www.kb.nl/en`

[2]The Delpher search platform can be accessed using the following URL: `https://www.delpher.nl/`

[3]A web crawler is an internet bot that automatically 'crawls' the web to collect information, e.g. for a search engine.

[4]Boolean operators in a query, such as AND, OR, NOT and PROX, can be used to broaden or narrow a search. For example, term A PROX term B searches for documents that contain the two terms in close proximity.

[5]Two blogs reporting on the trend of increasing query length: `https://tinyurl.com/y9eja22b`, and `https://tinyurl.com/y8twrjhv` (accessed 29 May 2018)

References

Agichtein, E., White, R. W., Dumais, S. T., and Bennett, P. N. (2012). Search, interrupted: understanding and predicting search task continuation. In Hersh, W. R., Callan, J., Maarek, Y., and Sanderson, M., editors, *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 315–324. ACM.

Baeza-Yates, R., Hurtado, C., and Mendoza, M. (2005). Query recommendation using query logs in search engines. In Lindner, W., Mesiti, M., Türker, C., Tzitzikas, Y., and Vakali, A. I., editors, *Current Trends in Database Technology - EDBT 2004 Workshops: EDBT 2004 Workshops PhD, DataX, PIM, P2P&DB, and ClustWeb, Heraklion, Crete, Greece, March 14-18, 2004. Revised Selected Papers*, pages 588–596, Berlin, Heidelberg. Springer Berlin Heidelberg.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. A., and Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In Sanderson, M., Järvelin, K., Allan, J., and Bruza, P., editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 321–328. ACM.

Bogaard, T., Wielemaker, J., Hollink, L., and van Ossenbruggen, J. (2017). Swish datalab: A web interface for data exploration and analysis. In Bosse, T. and Bredeweg, B., editors, *BNAIC 2016: Artificial Intelligence*, pages 181–187, Cham. Springer International Publishing.

Callahan, A., Pernek, I., Stiglic, G., Leskovec, J., Strasberg, R. H., and Shah, H. N. (2015). Analyzing information seeking and drug-safety alert response by health care professionals as new methods for surveillance. *J Med Internet Res*, 17(8):e204.

Catledge, L. D. and Pitkow, J. E. (1995). Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065 – 1073.

Chapelle, O. and Zhang, Y. (2009). A dynamic bayesian network click model for web search ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 1–10, New York, NY, USA. ACM.

Clough, P. D., Hill, T., Paramita, M. L., and Goodale, P. (2017). Europeana: What users search for and why. In Kamps, J., Tsakonas, G., Manolopoulos, Y., Iliadis, L. S., and Karydis, I., editors, *Research and Advanced Technology for Digital Libraries - 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*, volume 10450 of *Lecture Notes in Computer Science*, pages 207–219. Springer.

Cooper, A. (2008). A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web*, 2(4):19:1–19:27.

Darby, P. and Clough, P. D. (2013). Investigating the information-seeking behaviour of genealogists and family historians. *J. Information Science*, 39(1):73–84.

Downey, D., Dumais, S., and Horvitz, E. (2007). Models of searching and browsing: Languages, studies, and applications. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2740–2747, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Dwork, C. (2006). Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I., editors, *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, pages 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Eickhoff, C., Teevan, J., White, R., and Dumais, S. (2014). Lessons from the journey: A query log analysis of within-session learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, WSDM '14, pages 223–232, New York, NY, USA. ACM.

Gooding, P. (2016). Exploring the information behaviour of users of Welsh newspapers online through web log analysis. *Journal of Documentation*, 72(2):232–246.

Guo, F., Liu, C., and Wang, Y. M. (2009). Efficient multiple-click models in web search. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 124–131, New York, NY, USA. ACM.

Han, H. J. and Wolfram, D. (2015). An exploration of search session patterns in an image-based digital library. *J. Information Science*, 42(4):477–491.

Hearst, M., Elliott, A., English, J., Sinha, R., Swearingen, K., and Yee, K.-P. (2002). Finding the flow in web site search. *Commun. ACM*, 45(9):42–49.

Hoaglin, D. C. (1983). Letter values: A set of selected order statistics. In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, *Understanding robust and exploratory data analysis*, chapter 2, pages 33–57. Wiley New York.

Hollink, V., Tsikrika, T., and Vries, A. P. d. (2011). Semantic search log analysis: A method and a study on professional image search. *Journal of the American Society for Information Science and Technology*, 62(4):691–713.

Hong, Y., Vaidya, J., Lu, H., Karras, P., and Goel, S. (2015). Collaborative search log sanitization: Toward differential privacy and boosted utility. *IEEE Trans. Dependable Sec. Comput.*, 12(5):504–518.

Huurnink, B., Hollink, L., van den Heuvel, W., and de Rijke, M. (2010). Search behavior of media professionals at an audiovisual archive: A transaction log analysis. *Journal of the American Society for Information Science and Technology*, 61(6):1180–1197.

Jansen, B. J. and Spink, A. (2006). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Inf. Process. Manage.*, 42(1):248–263.

Jones, R. and Klinkner, K. L. (2008). Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In Shanahan, J. G., Amer-Yahia, S., Manolescu, I., Zhang, Y., Evans, D. A., Kolcz, A., Choi, K.,

and Chowdhury, A., editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 699–708. ACM.

Jones, R., Kumar, R., Pang, B., and Tomkins, A. (2008). Vanity fair: Privacy in querylog bundles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 853–862, New York, NY, USA. ACM.

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

Ke, H.-R., Kwakkelaar, R., Tai, Y.-M., and Chen, L.-C. (2002). Exploring behavior of e-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan. *Library & Information Science Research*, 24(3):265–291.

Korolova, A., Kenthapadi, K., Mishra, N., and Ntoulas, A. (2009). Releasing search queries and clicks privately. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 171–180, New York, NY, USA. ACM.

Mahoui, M. and Cunningham, S. J. (2001). Search behavior in a research-oriented digital library. In Constantopoulos, P. and Sølvberg, I., editors, *Research and Advanced Technology for Digital Libraries, 5th European Conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, Proceedings*, volume 2163 of *Lecture Notes in Computer Science*, pages 13–24. Springer.

Niu, X. and Hemminger, B. M. (2015). Analyzing the interaction patterns in a faceted search interface. *JASIST*, 66(5):1030–1047.

Sfakakis, M. and Kapidakis, S. (2002). User behavior tendencies on data collections in a digital library. In Agosti, M. and Thanos, C., editors, *Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002 Rome, Italy, September 16–18, 2002 Proceedings*, pages 550–559, Berlin, Heidelberg. Springer Berlin Heidelberg.

Spink, A. and Jansen, B. J. (2006). *Web search: Public searching of the Web*, volume 6. Springer Science & Business Media.

Wan-Chik, R., Clough, P. D., and Sanderson, M. (2013). Investigating religious information searching through analysis of a search engine log. *JASIST*, 64(12):2492–2506.

Wang, G., Zhang, X., Tang, S., Zheng, H., and Zhao, B. Y. (2016). Unsupervised Clickstream Clustering for User Behavior Analysis. In Allison Druin and, Cliff Lampe and, Dan Morris and, Juan Pablo Hourcade and, and Jofish Kaye, editors, *Proceedings of the 2016 CHI Conference on Human Factors in ComputingSystems, San Jose, CA, USA, May 7-12, 2016*, pages 225–236. ACM.

Zhang, S., Yang, H., and Singh, L. (2016). Anonymizing query logs by differential privacy. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 753–756, New York, NY, USA. ACM.