USING METADATA TO UNDERSTAND SEARCH BEHAVIOR IN DIGITAL LIBRARIES

Tessel Bogaard

The research presented in this thesis has been carried out at CWI, the national research institute for mathematics and computer science in the Netherlands, within the Human-Centered Data Analytics group.

The research presented in this thesis 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.

SIKS Dissertation Series No. 2023-12
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Cover by Miriam Vieveen

DOI: http://doi.org/10.5463/thesis.213
ISBN: 978-94-93289-23-9

VRIJE UNIVERSITEIT

USING METADATA TO UNDERSTAND SEARCH BEHAVIOR IN DIGITAL LIBRARIES

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. J.J.G. Geurts,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Bètawetenschappen
op woensdag 10 mei 2023 om 13.45 uur
in een bijeenkomst van de universiteit,
De Boelelaan 1105

door

Tessel Bogaard

geboren te Utrecht

promotoren:	prof.dr. J.R. van Ossenbruggen
	prof.dr. L. Hardman
copromotoren:	dr. L. Hollink
	dr. J. Wielemaker
promotiecommissie:	prof.dr. N. Pharo
	dr. C.C.S. Liem
	prof.dr. K.H. van Dalen-Oskam
	prof.dr.ir. A.P. de Vries
	prof.dr. W.J. Fokkink

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many people. I would like to thank all of them, and if I am forgetting some names here, I apologize for the omission.

My colleagues at CWI, an inspiring place to do my research and to learn and grow, with the bonus of making some new friends on the way. And the Human-Centered Data Analytics research group at the CWI, where I have always felt welcome, and where we shared discussions not only about the research but also about what is the best chocolate and the best food. In addition my promotors, Jacco and Lynda, for their scientific guidance during my research. And of course my supervisors, Laura, who always has good feedback on my research, and Jan, who was not only a supervisor but a programming partner in this research as well.

The National Library of the Netherlands for their support in my research, and for giving me the opportunity to study user behavior in the context of a richly annotated, large collection.

Kees Ribbens, who came to me with an interesting problem which led to one the papers found in this thesis, and his colleagues Caroline Schoofs and Koen Smilde at the NIOD Institute for War, Holocaust and Genocide Studies, who helped in the annotation and evaluation for the same the paper.

My colleagues at the newspaper NRC, who have encouraged me to take the time to finish the last stages of my thesis. I would especially like to mention Miriam Vieveen, who was so generous to offer her expertise as a designer and is responsible for the cover of this thesis.

The domain experts and professionals who were kind enough to offer their time as participants in my user study. Since they helped me as anonymous participants I will not name them here.

The members of the doctoral committee, Nils Pharo, Cynthia Liem, Karina van Dalen-Oskam, Arjen de Vries and Wan Fokkink.

Willem, for taking the time to be my paranymph and for being a long-time friend.

Ays, for being my co-author, for taking the time to be my paranymph, and for being a good friend.

My family, especially Gerrie and Jan, for always being interested in my progress, and Martin and Appy, for their support and enthusiasm.

And of course Boris. Thank you for your continued encouragement and for your patience during my journey to finish this work.

CONTENTS

1 INTRODUCTION

Search log analysis is an unobtrusive technique used to better understand user behavior in search systems [7, 8, 22, 23, 31, 42, 48, 51, 83]. It contributes to the understanding of the information needs of users and to what extent these are met. The results of such an analysis can be used to evaluate search algorithms or user interfaces, or to (re-)design systems [13, 41, 85].

Log analysis frequently focuses on queries and click actions on ranked lists of search results [7, 8, 31, 47, 48, 51–53, 70]. This focus poses some disadvantages. First, queries are ambiguous, as they form an uncontrolled vocabulary and have little context to interpret the information need of the user. Second, most queries are in the *long tail*; they occur infrequently, making it hard to find recurring patterns. Third, queries may contain privacy-sensitive information such as names and personal information [26, 55, 56], and thus are seldom shared among researchers.

We study search in "vertical" search engines, as opposed to "horizontal" search engines. We define a vertical search engine as a search engine providing access to professionally curated collections, such as digital libraries, archives and webshops. With the term "horizontal" search engines we refer to search engines on the open web, such as Google, Bing, Baidu etc., not providing access to specific collections or specific types of content or types of media, but providing access to the web in general.

In the context of a vertical search engine providing access to a collection, other rich data is available in addition to the search logs: the annotated documents in the collection, with categorizations presented in professionally curated metadata. This descriptive metadata is often reflected in the search interface in the form of facets, acting as a filter over the search results and as such offering an alternative way to access the collection next to full-text search. Search behavior in vertical search engines can be expected to differ from behavior in horizontal search engines, in part as a reflection of different search functionalities based on the descriptive metadata. This has been shown, for example, for image archives [37, 48], a medical knowledge portal [18], newspaper archives [13, 33], and in a study of a digital library [77].

In this thesis, we focus on how to leverage the descriptive content metadata to further our understanding of search behavior. In our analyses we use both the metadata present in search interactions in the form of selected facets, and the metadata of the clicked documents. We combine search logs collected in a vertical search engine with the metadata records and contents of the searched collection.

The scope of our research is limited to digital libraries within the public domain, supporting collection owners, domain experts and researchers interested in how users search different parts of these collections. At the same time we recognize that the results of our research could likely also be used in other vertical search systems.

In the research we conducted for this thesis we address the following questions: do search patterns differ within different parts of a collection; how can we identify user interests and corresponding search behavior within a collection; how can we retrieve search sessions relating to specific topics; and finally how can we communicate our research results to collection owners, domain experts, and researchers.

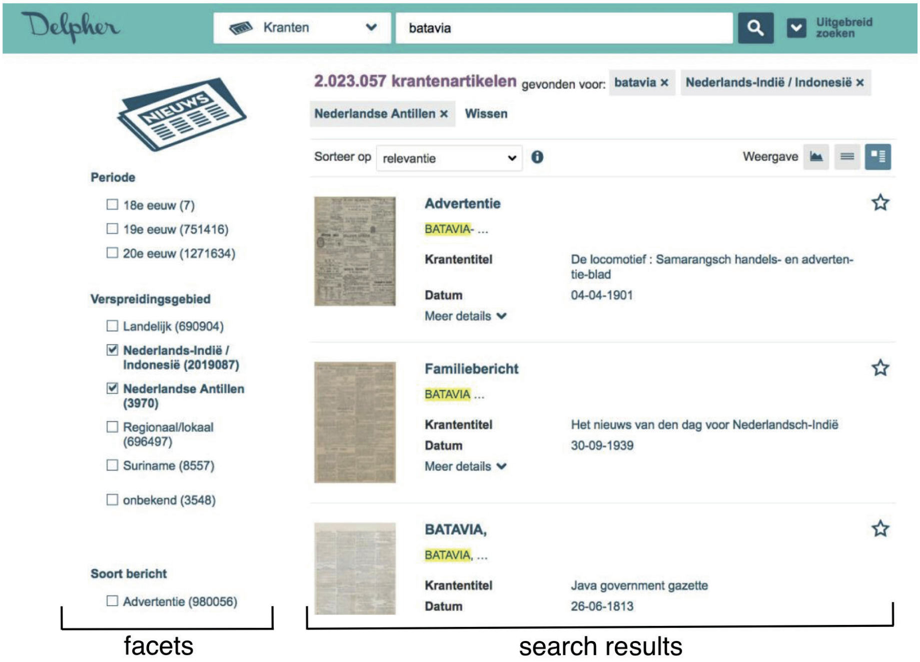
1.1 RESEARCH CONTEXT

The research presented in this thesis was performed at Centrum Wiskunde & Informatica (CWI), and was 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. To conduct the research, search logs and collection data were used from Delpher, a search platform maintained by the National Library of the Netherlands providing digital access to historical newspapers, books, journals and texts of radio news broadcasts from collections of various libraries, museums and other heritage institutions. Within this platform, we focused on search within the historical newspaper collection.

The logs were collected from the Delpher search platform[1], for which access was granted under a strict confidentiality agreement. In addition, for part of our research a collaboration was set up with humanities researchers from the Dutch NIOD Institute for War, Holocaust and Genocide Studies.

COLLECTION The historical newspaper collection 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. 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 item type can be one of the following four types: news article, advertisement, announcement (relating to family such as birth, marriage or death announcements) or image (illustrations or photographs, where search is performed on the caption text). The distribution zone represents the geographical region where the newspaper was distributed,

[1] `http://www.delpher.nl`



batavia

2.023.057 krantenartikelen gevonden voor: batavia × Nederlands-Indië / Indonesië × Nederlandse Antillen ×

Sorteer op relevantie

Periode

- □ 18e eeuw (7)
- □ 19e eeuw (751416)
- □ 20e eeuw (1271634)

Verspreidingsgebied

- □ Landelijk (690904)
- ☑ Nederlands-Indië / Indonesië (2019087)
- ☑ Nederlandse Antillen (3970)
- □ Regionaal/lokaal (696497)
- □ Suriname (8557)
- □ onbekend (3548)

Soort bericht

- □ Advertentie (980056)

Advertentie
BATAVIA- ...
Krantentitel De locomotief : Samarangsch handels- en adverten-tie-blad
Datum 04-04-1901
Meer details

Familiebericht
BATAVIA ...
Krantentitel Het nieuws van den dag voor Nederlandsch-Indië
Datum 30-09-1939
Meer details

BATAVIA,
BATAVIA, ...
Krantentitel Java government gazette
Datum 26-06-1813

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

with – at the time the data was collected – the values “local”, “national”, one of the areas formerly colonized by the Dutch (“Indonesia”, “Suriname”, or the “Antilles”), or, in a few cases, “*unknown*”.

search interface In the Delpher search interface (see Fig. 1) the facets are filters based on the metadata attributes and values of the documents, and these can be used to refine the search results. From a search results page, a user can click on a document in the list of results and, after a click, can download the document.

search logs and documents The web server of the Delpher search platform logs HTTP page requests of its users. The National Library of the Netherlands has provided us with the search 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 URLs (URL where request originated), and the URLs of the requested HTTP pages. 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 selected, and (3) the result ranking method, together representing what we call the user's search interaction.

In addition to the search logs we have received the metadata records of the over 100M documents in the newspaper collection as well as the actual contents of these documents. We have combined these three data sets, making it possible to analyze both the metadata of the search interactions (the facets selected) and the metadata and/or the contents of the clicked documents concurrently.

1.2 RESEARCH GOAL

The main research goal of this thesis is to investigate the use of metadata to analyze search behavior. We include the descriptive content metadata of both the search interactions (the filtering facets selected) and of the clicked documents. First, we study how we can use metadata in three different settings. The settings we study differ in a number of dimensions; on whether we know in advance which topics we want to analyze, and on the availability of relevant metadata. Finally, we investigate how we can communicate the results of these studies to domain experts, collection owners and researchers.

METADATA-BASED ANALYSIS: SETTING 1 In the first setting, we analyze the search log data using specific metadata values defined in advance, in order to study search behavior within specified parts of the collection that we have selected to be relevant. These parts correspond to historical periods, geographical regions or subject matter. This part of the research is presented in Chapter 2 of this thesis.

We performed a descriptive analysis within the historical newspaper collection of the National Library. We combine the search log data with the metadata records describing the contents of the collection, using specific metadata values to create subsets in the logs corresponding to different parts of the collection. When we compare sessions in which users use facets in their search with sessions in which no facets have been selected by the users, we observe that on average users spend more time in the first type of session, and that these sessions contain more clicks, downloads and unique queries. In addition we observed distinct search patterns in different parts of the collection, thus providing deeper insights into search behavior at a fine granularity.

METADATA-BASED ANALYSIS: SETTING 2 In the second setting we study how to leverage the metadata without specifying metadata values of interest beforehand. In this case, we want to gain insights into what subsets of the collection users are showing an interest in, and how they search in these parts, without defining these user interests in advance. The results of this part of the research can be found in Chapter 3.

To find patterns in the data without defining specific metadata values corresponding to the different parts of the collection in advance, we use an unsupervised machine learning algorithm. The algorithm used here is a clustering algorithm, and we cluster the search sessions based on the metadata values present in these sessions. This helps us to find user interests within the collection, and analyze the corresponding search behavior. We observed clusters of users searching in specific parts of the collection. In some parts users were spending little time and few search techniques, in other parts spending a long time and a wide variety of search techniques.

To evaluate the clustering algorithm, we used the stability of the clusters over time. We measured whether the same clusters reappear over a period of six months. Our results showed a good stability for the larger clusters, demonstrating continued user interests in certain parts of the collection. For the smaller clusters, this stability was less strong, showing more variability here.

METADATA-BASED ANALYSIS: SETTING 3 In some cases, professionally curated metadata to identify the relevant parts of the collection we want to study might not be available. Thus, in the third setting we explore how to identify search behavior within specific parts of the collection when no metadata directly describes the topics of interest related to these parts. The results of this part of the research is presented in Chapter 4.

We applied several approaches. First, we defined our topics of interest using a relevant external knowledge resource to create a topic representation. Then, we expanded on the topic representation using local word embeddings on the documents in the collection. In addition, we included manual judgements of the different versions of the topic representations, resulting in a variety of methods to create them. We matched the topic representations to both the user queries and to the clicked documents. We explicitly note that in this setting we do include the user queries, where in the first two settings we did not. We still group the search interactions, but this time not based on any selected facets, as these do not contain our topics. Instead, we group the search interactions based on an analysis of the user queries. To do this, we match the user queries to the topic representations. We applied these different approaches in a double case study including two topics in the same data set, and evaluated the resulting subsets using a ground truth based on an annotated sample of the search sessions.

COMMUNICATING METADATA-BASED ANALYSES Finally, we look into how to communicate the results of these type of analyses to collection owners, domain experts, and professionals in an easy and intuitive way, including the metadata. This part of the research is presented in Chapter 5.

We developed a session visualization which combines a graph visualization of the search interactions in a session, and a coloring representing the metadata related to the search. To evaluate the visualization technique, we conducted a user study to compare our visualization to a baseline session representation in three typical tasks. The participants fill out standard questionnaires for perceived workload and usability of the two visualizations, their activity is logged and they provide us with written comments explaining their answers to the tasks and how they experienced the use of the two visualizations. Our study demonstrates the added value of the visualization. Our design of the session graphs is new in combining both the search interactions and the metadata in a single visualization.

Presenting the search sessions to domain experts, collection owners, researchers, and professionals can help to identify patterns in search by browsing through the sessions at a glance; and it can help to explore the most typical sessions of a clustering.

Our visualization of the search sessions also played a role in the first three research chapters in this thesis. For the research presented in Chapter 2 it provided support in the process of data cleaning and data exploration, and as such it has made the process of data cleaning more transparent and easier to reproduce. For Chapter 3 the visualization technique provided insights into the different search behaviors with respect to the different user interests. And for Chapter 4 the session visualizations have helped to create a ground truth of annotated sessions used to validate our results in the fourth chapter.

1.3 PUBLICATIONS

The following publications formed the basis of this thesis.

CHAPTER 1 is based on the doctoral consortium paper *On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives* presented at the 2018 Conference on Human Information Interaction & Retrieval, March 11–15, 2018, in New Brunswick, NJ, USA, by Tessel Bogaard [10].

CHAPTER 2 is previously published as *Metadata categorization for identifying search patterns in a digital library* published in the Journal of Documentation Volume 75, Number 2, 2019, by Tessel Bogaard, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman [13]. The research for this paper was mainly conducted by Tessel Bogaard. The data cleaning of the raw logs and metadata records was done in collaboration with Jan Wielemaker, the analysis and code for analysis was performed by Tessel Bogaard. Supervision and feedback on the research process was given by the co-authors. All authors contributed to the text.

CHAPTER 3 is previously published as *Searching for Old News: User Interests and Behavior within a National Collection* presented at the 2019 Conference on Human Information Interaction & Retrieval, March 2019, Glasgow, UK, by Tessel

Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen [12]. The paper was awarded an Honorable Mention at the conference. All authors contributed to the text. The idea for using a metadata clustering and the selection of stability as the evaluation method came from Tessel Bogaard, as well as writing the code for the clustering algorithm and the stability measure. Feedback on the process and the text was given by the co-authors.

CHAPTER 4 is previously published as *Comparing Methods for Finding Search Sessions on a Specified Topic: A Double Case Study* presented at the 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, September 13-17, 2021, by Tessel Bogaard, Aysenur Bilgin, Jan Wielemaker, Laura Hollink, Kees Ribbens, and Jacco van Ossenbruggen [11]. All authors contributed to the text. The majority of the code and the analysis was written by Tessel Bogaard. Part of the analytical code, the local word embeddings, was written in collaboration with Aysenur Bilgin. The evaluation method was set up in collaboration with Laura Hollink. The relevance judgments of the topic representations for the WWII topics were provided by Kees Ribbens with the assistance of Caroline Schoofs and Koen Smilde from the NIOD Institute. The topic-relevance annotations of the sessions were performed by Laura Hollink and Tessel Bogaard. Writing contributions and feedback on the text were provided by all authors, especially on the parts they contributed to.

CHAPTER 5 is previously published as *Understanding User Behavior in Digital Libraries Using the MAGUS Session Visualization Tool* presented at the 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, by Tessel Bogaard, Jan Wielemaker, Laura Hollink, Lynda Hardman, and Jacco van Ossenbruggen [14]. All authors contributed to the text. In addition it was based on the demo and paper *SWISH DataLab: A Web Interface for Data Exploration and Analysis* presented at the 28th Benelux Conference on Artificial Intelligence, Amsterdam and published in BNAIC 2016: Artificial Intelligence - 28th Benelux Conference on Artificial Intelligence, Amsterdam, The Netherlands, November 10-11, 2016, Revised Selected Papers [9]. The idea of visualizing sessions in graphs with color to represent the metadata came from Tessel Bogaard. The code was written in collaboration with Jan Wielemaker. The session graphs have been improved through feedback from the other co-authors and feedback in demonstrations in different contexts (DHBenelux conference, ICTOpen and a demo paper at BNAIC conference, demonstrations at the National Library of the Netherlands, plus feedback at the CHIIR doctoral consortium). The user study was set up by Tessel Bogaard with feedback from the co-authors, and with support in coding the forms and logging of the three tasks in the user study by Jan Wielemaker. All authors contributed to the text.

2

METADATA CATEGORIZATION FOR IDENTIFYING SEARCH PATTERNS IN A DIGITAL LIBRARY

2.1 ABSTRACT

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.

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. 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. 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.

2.2 INTRODUCTION

Log analysis is an unobtrusive technique for macro-analysis of user behavior in digital search systems [48, 52]. 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 [6, 8, 29, 52]. 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 [38, 48], a medical knowledge portal [18], a newspaper archive [33], and in a study of a digital library [77].

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* [43]. 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 [88], or an investigation of different types of learning in search [31]. 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 [55]. 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 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 Suriname, 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).

2.3 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 [57]. 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 [67] 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 [80] 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 [33] showed differences between online and offline 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 [77], 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 [77], 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 [27]. 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 [24].

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 [18], or alternatively, an HTTP cookie can be used [33]. 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, Liu, and Wang [36] a session is defined as a single user query and the subsequent clicks; and in Huurnink et al. [51] 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 [4] 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. [31] and [77] and finding its origin in a study of browsing behavior in 1994 [20]. Other examples of sessions defined by a timeout are Hollink, Tsikrika, and Vries [48](15 minutes); Chapelle and Zhang [21] (60 minutes); and Jansen and Spink [52] or [38] 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 [54].

In the context of studying web navigation, the concept of a *clickstream* is often used, as in [89]. 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 [77] 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 [88], or to study different types of learning in search [31]. 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 [56]. This leads to a conflict between protecting the privacy of users and retaining or publishing query logs, as mentioned in Cooper [26]. Techniques such as differential privacy [30] – 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 [60]. Two recent papers [50, 95] 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.

2.4 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).

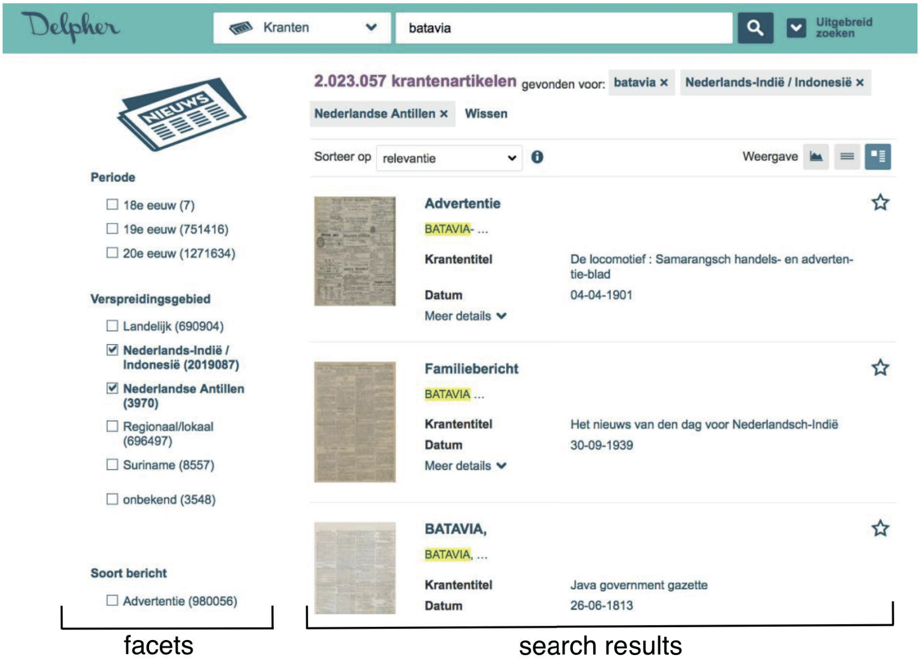
The documents in the collection are described in bibliographic metadata records with the following attributes: a document identifier, the publication date, item type, newspaper title, place of publication, source (the physical location of the original document), and distribution zone. The distribution zone attribute represents the geographical region where the newspaper was distributed, with values "local", "national", one of the former Dutch colonies ("Indonesia", "Suriname", 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. 2) 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

[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/`



Delpher

Kranten

batavia

Uitgebreid zoeken

2.023.057 krantenartikelen gevonden voor: batavia × Nederlands-Indië / Indonesië × Nederlandse Antillen × Wissen

Sorteer op relevantie

Weergave

Periode

- □ 18e eeuw (7)
- □ 19e eeuw (751416)
- □ 20e eeuw (1271634)

Verspreidingsgebied

- □ Landelijk (690904)
- ☑ **Nederlands-Indië / Indonesië (2019087)**
- ☑ **Nederlandse Antillen (3970)**
- □ Regionaal/lokaal (696497)
- □ Suriname (8557)
- □ onbekend (3548)

Soort bericht

- □ Advertentie (980056)

Advertentie

BATAVIA- ...

Krantentitel	De locomotief : Samarangsch handels- en advertentie-blad
Datum	04-04-1901

Meer details

Familiebericht

BATAVIA ...

Krantentitel	Het nieuws van den dag voor Nederlandsch-Indië
Datum	30-09-1939

Meer details

BATAVIA,

BATAVIA, ...

Krantentitel	Java government gazette
Datum	26-06-1813

Figure 2: Search interface for the newspaper collection, with facets to the left and search results to the right.

October 2015 until March 2016. These around 200M records include encoded IP addresses, time of the requests, user agents (identifying client software), referrer 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.

2.5 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.

2.5.1 *Step 1: Session identification in search logs*

As described in the Section 2.3, 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 [27]), 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.

2.5.2 *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 [46]. 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.

2.5.3 *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

[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.

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 $\langle \mathtt{item\_type} = \mathtt{announcement} \rangle$. 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 $\langle \mathtt{item\_type} = \mathtt{article} \rangle$. 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 $\langle \mathtt{item\_type} = \mathtt{announcement} \rangle$ 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.

2.5.4 *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.

2.6 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 Suriname (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.

2.6.1 *Visitor statistics*

The portal is accessed consistently over the days of the week (Fig. 3), in contrast to the observations of Jones et al. [57], Ke et al. [59], and Huurnink et al. [51], 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.

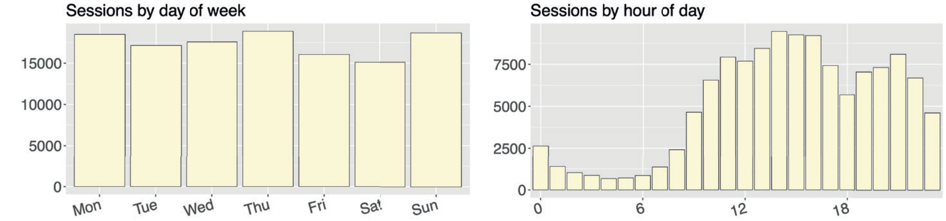


Figure 3: Number of sessions over the days of the week and the hours of the day

2.6.2 *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 [77]. 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.

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 [77]. In a photo archive of a news agency, Hollink, Tsikrika, and Vries found an even lower number of terms in queries (mean of 1.8) [48]. 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.

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

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	Incl. downloads	Incl. quoted query	Incl. 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%

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.

2.6.3 *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% 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. 4). 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 pro-

Table 2: Session subsets by item type facet values

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

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

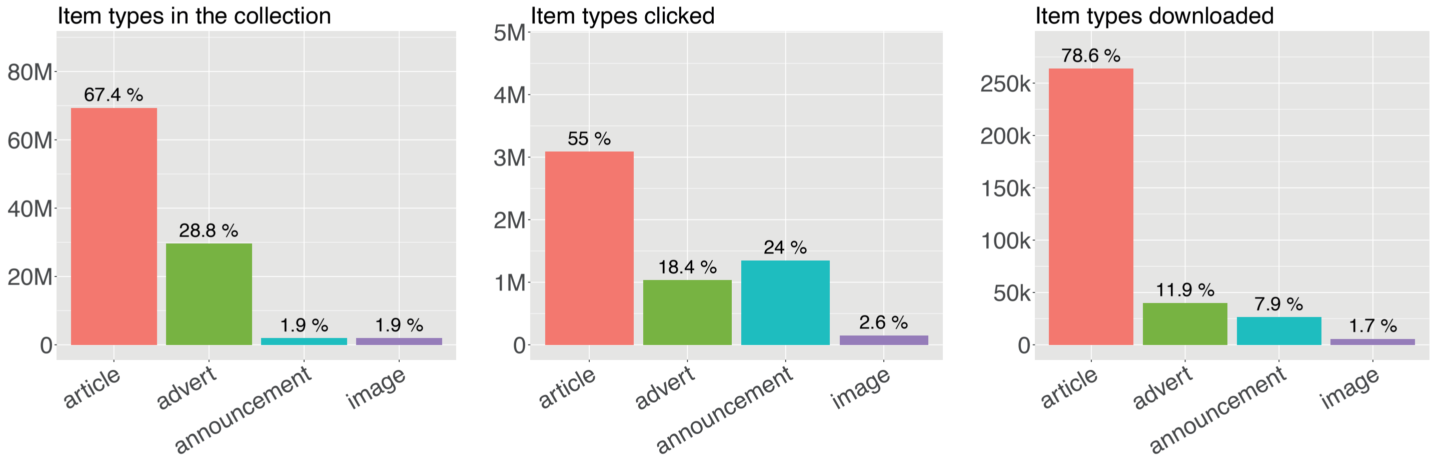


Figure 4: **I**tem types in the collection (103 million documents), of clicks (5.6 million) and of downloads (335 thousand).

portion 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 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.

2.6.4 *Search for Suriname*

In this section we focus on users interested in publications from Suriname, one of the former Dutch colonies. To do this, we will investigate users selecting the Suriname 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 5 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% of the clicks on national and regional articles are from sessions without using the corresponding facets, only 16% of the clicks on articles from Suriname are from sessions not using the Suriname facet value. The Suriname value is selected in 2% of all sessions. This interest in Suriname is higher than is to be ex-

Table 3: Session subsets by distribution zone facet values

Sessions	Frequency		Clicks	
- national	21,325	10%	1,620,113	25%
- local	27,050	13%	1,797,927	28%
- Indonesia	10,930	5%	882,072	14%
- Antilles	2,930	1.4%	289,256	4%
- Suriname	4,004	2%	334,013	5%
- *unknown*	861	0.4%	112,268	2%
	Downloads		Clicks on value	
- national	139,582	24%	639,817	29%
- local	146,184	25%	1,138,093	34%
- Indonesia	71,860	13%	340,384	61%
- Antilles	25,751	4%	43,234	49%
- Suriname	20,857	4%	128,334	84%
- *unknown*	7,385	1%	63	2%
	Median duration	Median queries	Median clicks	
- national	1d 0:24:16	6	12	
- local	17:51:49	6	10	
- Indonesia	21:56:13	7	15	
- Antilles	23:56:51	8	19	
- Suriname	4:44:02	6	14	
- *unknown*	6d 23:48:46	13	37	
	Including downloads	Including quoted query	Including boolean query	
- national	25%	22%	4.3%	
- local	22%	21%	3.1%	
- Indonesia	27%	25%	3.4%	
- Antilles	31%	23%	3.3%	
- Suriname	24%	19%	3.2%	
- *unknown*	33%	24%	2.7%	

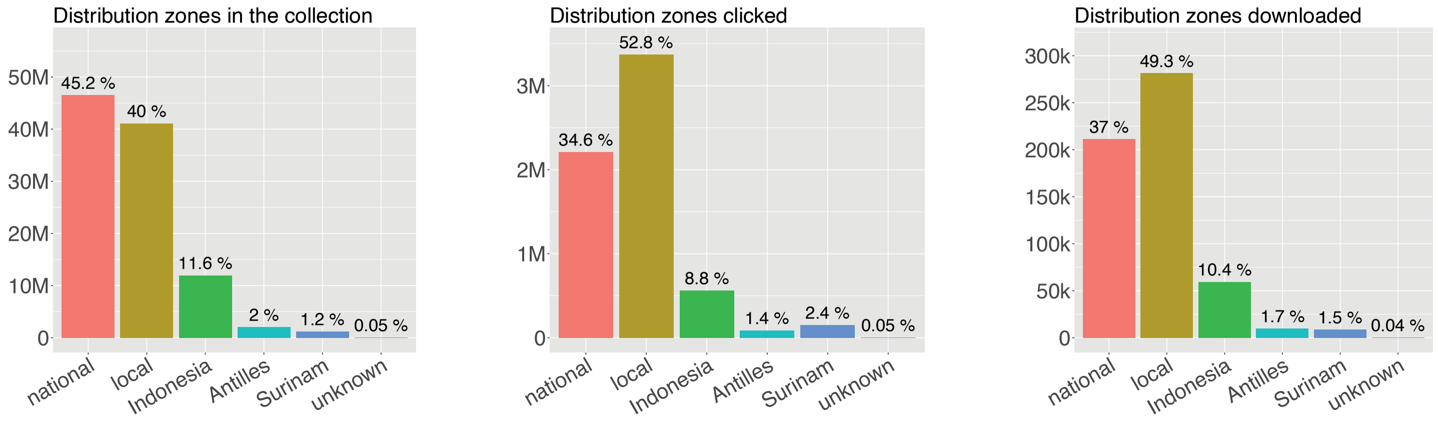


Figure 5: **D**istribution zones in the collection (103 million documents), of clicked documents (6.4 million), and downloads (575 thousand).

pected from the size of the Suriname collection (only 1.2%). The number of clicks on documents from Suriname 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. 5). 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 Suriname (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.

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

RECOMMENDATIONS The relatively low number of clicks and downloads for the Suriname 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.

2.6.5 *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%).

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. 6).

Table 4: Session subset by time facet value

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

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

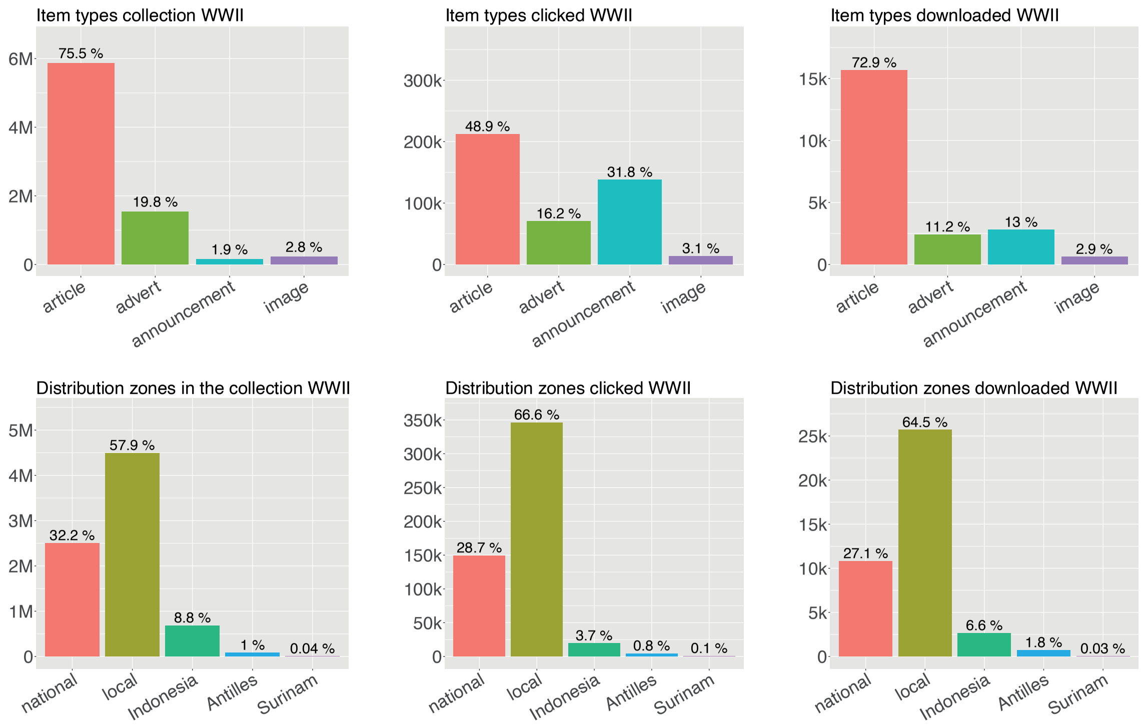


Figure 6: **I**tem types and **D**istribution zones of clicks (434 thousand items, 520 thousand documents) and downloads (40 thousand items and 40 thousand documents) and the collection (7.8 million documents) in World War II.

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.

2.7 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 Suriname. 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.

3

SEARCHING FOR OLD NEWS: USER INTERESTS AND BEHAVIOR WITHIN A NATIONAL COLLECTION

3.1 ABSTRACT

Modeling user interests helps to improve system support or refine recommendations in Interactive Information Retrieval. The aim of this study is to identify user interests in different parts of an online collection and investigate the related search behavior. To do this, we propose to use the metadata of selected facets and clicked documents as features for clustering sessions identified in user logs. We evaluate the session clusters by measuring their stability over a six-month period.

We apply our approach to data from the National Library of the Netherlands, a typical digital library with a richly annotated historical newspaper collection and a faceted search interface. Our results show that users interested in specific parts of the collection use different search techniques. We demonstrate that a metadata-based clustering helps to reveal and understand user interests in terms of the collection, and how search behavior is related to specific parts within the collection.

3.2 INTRODUCTION

Understanding user interests and related search behavior can reveal the different types of system support users need. Collections are not always homogeneous and users may have different information needs depending on which parts of a collection they are interested in. This begs the question of how we can identify different "parts" of a collection – for example through different usage patterns and/or by professionally curated categorizations of the documents in the collection. If we are able to identify different usage patterns corresponding to identifiable parts of the collection then we can better help collection owners in providing support for these.

The research questions addressed in this paper are thus:

- (**RQ1**) *What are the user interests in terms of the different parts of a collection? How can we detect these?*
- (**RQ2**) *What is the related search behavior within these parts?*

Understanding the answers to these questions may lead to more targeted search interfaces, better search algorithms, and a fine-tuning of strategies for collection management.

In a digital library or archive, metadata categorizations of the documents are often reflected in facets, allowing the user to filter the search results. We use the metadata of facets selected and of documents clicked to detect user interests. As we do not know in advance in which (combinations of) metadata categories users are interested, we apply a data-driven partitioning of sessions: a clustering of sessions based on the metadata features of both search (selected facet values) and clicks (document metadata) in each session, and we analyze the behavior in the resulting clusters.

Evaluating the resulting clustering is nontrivial, for different reasons: first, an interpretation of the results is subjective, and second, we have no ground truth available in the data as a way to measure the "correctness" of the clustering. Nevertheless, as we are interested in stable clusters that reappear over different periods, we test the stability of the clusters in each month over a six-month period and interpret stability as an indicator of the quality of the clustering, similar to the cluster stability measured between two periods in [22].

We apply our approach to data from the National Library of the Netherlands, a digital library with a richly annotated historical newspaper collection spanning 400 years, and a faceted search interface. The library has granted us access to both user logs[1] and the metadata descriptions of the documents in the collection.

Our results show that the detected user interests are stable, and that the related search behavior varies within the different parts of the collection. Examples of user interests are: specific types of news items, such as family announcements (relating to births, marriages, deaths), specific periods, such as 1930-49 (including the Great Depression and World War II), or specific regions, such as Suriname (one of the former Dutch colonies). We observe users focusing exclusively on specific parts of the collection, in some parts spending less time and few search techniques, in other parts spending a lot of time and a variety of search techniques. As a result this approach can help to find and investigate these highly-focused users. This can inform the design of more targeted user interfaces, or help to improve search systems or collection management. We contribute to the research field by demonstrating that a partitioning of sessions into clusters based on the metadata of a collection and an investigation of related search behavior reveals specific user needs in specific parts of a collection, where in an overall analysis these patterns would disappear.

3.3 RELATED WORK

To answer our research questions, definitions of user interests and sessions are needed. Additionally, we need a method to group the sessions. In this section we discuss relevant literature with respect to how to detect user interests, define sessions, and what methods can be used to group the identified sessions.

[1] Logs collected from the search platform `http://www.delpher.nl`, access granted under a strict confidentiality agreement.

Detecting user interests

User interests are frequently derived from queries, for example by categorizing user queries in [64], or finding search topics by semantic linking of user queries [48]. Alternatively, interests can be detected in logs using the context of search [92], or search histories [93]; and in [44] mouse hovering is used to help understand user interests within a digital library, in combination with query analysis and the (analyzed) metadata of document clicks in a statistical analysis.

Similar to this research, we use a form of categorization to identify user interests, and similar to [44], we make use of the metadata categories of the collection. However, we use the metadata directly as found in facets selected and documents clicked, rather than the query input, to identify user interests, as we aim for a definition of user interests in terms of parts of the collection.

Defining sessions

Search behavior is often interpreted using a bounded sequence of search actions by a user [52]. Sessions have been studied to understand search in context and to evaluate it in terms of success or failure [52]. Sessions help to provide information about repeated visits [57], to examine query modification [48], to obtain information about learning in search [31], or to find patterns in search behavior [22, 76].

We use sessions to put user interactions in a context and so to enable the detection of user interests and behavior. This requires a computational method for specifying the beginning and end of a session. Sessions can be specified based on query boundaries using the IP address as identifier. For example, in [36] a session is defined as a search query and the following clicks until the next query, and in [51] a session is bounded by the presence of overlapping terms in successive queries until there is no more common term. Sessions are frequently bounded by a timeout, a period of inactivity by a user, e.g. [21, 31, 48, 52]. In the context of studying web navigation, the concept of a *clickstream* is more often used, as in [89]. A clickstream is the navigational path a user follows, consisting of consecutive HTTP requests from a single IP address. We adopt this definition of a session, as it enables the identification of multiple users behind a single IP and we want to avoid breaking up longer sessions by using a timeout.

Grouping user logs

Several approaches exist to group user logs in order to find patterns, for example logs can be classified or clustered. To classify different types of behavior, queries have been grouped into *why* versus *what* questions [31], into DBpedia concepts [70], or into categorizations based on a thesaurus related to the collection [51]. Alternatively, Niu and Hemminger have provided an analysis of faceted versus non-faceted search, grouping the logs based on user actions, showing in their work that facets play an important role in search [77]. In our study, we not only include the facets, we also enrich the clicked documents with their metadata descriptions, and use this metadata explicitly to group the user logs for the detection of user interests.

Clustering techniques can also be used to detect patterns in logs. For example, Wang et al. use unsupervised hierarchical clustering to detect user behavior patterns in social networks [89]. In our work, we also use unsupervised clustering and not supervised classification, for similar reasons: we do not have a ground truth available in the data, nor do we know in advance which patterns we want to detect. However, since our data is skewed we use a different algorithm that is more robust to outliers.

Clustering techniques have been used before in the context of a digital library. Chen and Cooper applied a hybrid clustering technique to detect different types of users in the logs, combining an initial clustering using k-means with hierarchical clustering to get to the final clusters [22]. In this research, sessions are represented using a set of features based on user interactions with the search system. More recently, Niu and Hemminger have reproduced this research with an added focus on the facets present in more recent search interfaces [76]. In our study the goal is different, as we aim to find the user interests in terms of the collection and relate these user interests to search behavior. Nevertheless, we use a similar clustering technique and a similar representation of the sessions to be clustered as in [22] and [76], even though we focus exclusively on the bibliographic metadata features of search and clicks.

To evaluate the clustering we look at stability [86], similar to the approach in [22]. This approach was more recently investigated as a validation method for a clustering in a log analysis of a digital library in [34].

3.4 METHOD

In this study we use a clustering algorithm to detect the user interests and investigate the relation between these user interests and search behavior in the collection. For the clustering of the sessions, we base the features on the metadata of facets and clicked documents (the metadata of the facets are the selected values used in search). To do this we need both user logs and metadata records of the collection being searched.

3.4.1 *Session Identification and Representation*

We identify sessions in the logs based on a *clickstream* model, using the IP address as identifier and connecting sequential HTTP requests to follow the user navigating the search platform.

We represent the sessions based on the metadata values of the search interactions, where available in the facets selected, and clicked documents, linked to the metadata records of the collection. We include all values of the (main) categories in the metadata (such as publication date, origin or type of document). These values are proportional to the number of search interactions or the number of clicked documents per session, and are used as features for the clustering.

To detect the user interests, we apply a clustering algorithm representing the sessions using a metadata feature set. As the features are likely to be correlated, principal component analysis is applied for dimensionality reduction before clustering with a standardized feature set. We retain the principal components with a standard deviation equal to or higher than 1 for the clustering.

In addition, we collect interaction variables based on user interactions within the search interface to analyze the search behavior. These variables include typical variables, such as the total duration of a session, the number of HTTP requests, the proportions of actions that are search or clicks, and specific variables dependent on the search interface, such as facets or reordering of results.

3.4.2 *Clustering*

We use an unsupervised clustering algorithm, as we have no ground truth available and do not know in advance what kind of patterns are present in the data. Since we cannot assume the data adheres to a normal distribution, we have chosen a k-medoids method [58], partitioning the data into k clusters, as k-medoids is more robust against outliers than k-means is, it is to k-means what the median is to the mean. As we have a high number of sessions and many dimensions in the clustering, we apply the CLARANS algorithm [74], a k-medoids variant optimized for large datasets. We use the Manhattan distance as distance metric for the clustering, because it is suitable for data represented in a high dimensional space [2]. To choose the number of clusters k, we apply the silhouette method [79], which measures the separation between the clusters with values ranging from -1 to 1, the higher values indicating a better clustering. We cluster the sessions repeatedly with different values for k and select the k with highest average silhouette width. We use a statistical summary of user behavior in each resulting cluster to analyze differences in behavior between the clusters based on the user interests.

3.4.3 *Evaluation of Clustering*

Our goal is to find stable patterns that reoccur in different period, so we evaluate the stability of the clustering over time, using this as an indication for clustering quality [86]. For this purpose, we cluster logs collected in separate periods, similar to the approach in [22]. We use a six-month period as it is the maximum period user logs can be retained according to Dutch law and as is common practice to protect the privacy of users. The size of each period is a month, as the sample size used in the collection of the logs was a month and some sessions have a duration longer than two weeks (12% of the sessions).

The stability of the clusters between two periods, the previous period and the target period, is measured as follows:

(1) We cluster the sessions in the previous period using the same value for *k* as was used for the target period.

(2) For each cluster in the previous period we determine a "center" by taking the original metadata features of the sessions and computing the median for each feature, resulting in a set of medians.

(3) For each session in the target period, we compute the Manhattan distance to each of the centers in the previous period based on the original metadata features.

(4) We assign each session in the target period to the cluster from the previous period with the shortest Manhattan distance, the nearest "center".

(5) For each of the *k* clusters in the target period, we compute the percentage of sessions in each of the *k* clusters of the previous period, resulting in *k x k* percentages .

(6) We define the stability of a cluster in the target period as the highest of the *k* percentages, the best match.

(7) The stability of a clustering as a whole is the average stability of all its clusters, weighted by cluster size.

We inspect in detail the overlap between the clusters between two periods. We do this with a "stability matrix", that shows the amount of matching (i.e. the percentages per cluster as assigned in step 5) between each of the clusters of the two periods. In the stability matrix, the clusters of the target period are the columns (percentages in the columns sum to 100%), and the previous period the rows.

We remark that cluster stability and silhouette widths measure different things: the first consistency between clusterings over time and the second consistency within a clustering.

3.5 THE NATIONAL LIBRARY OF THE NETHERLANDS

We apply our method to a library that is representative for digital libraries in general, with a richly annotated collection of digitized historical documents and a faceted search interface. The National Library of the Netherlands has granted us access to user logs from their search platform[2], our focus is on the historical newspaper collection, amounting to more than 90% of all HTTP page requests to the library's search platform.

From this collection, users can retrieve full newspaper issues, pages, or individual items on a newspaper page. The documents in the collection are an-

[2] `http://www.delpher.nl` provides access to collections from the National Library of the Netherlands and other heritage institutions, comprising newspapers, magazines, radio bulletins, and books.

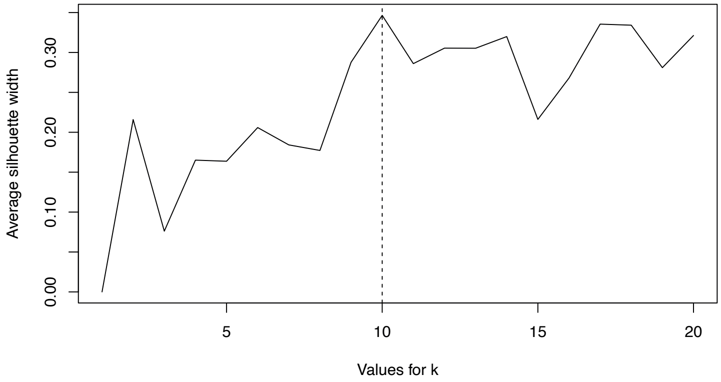


Figure 7: Average silhouette widths for k in March

notated with bibliographic metadata records, including a publication date, distribution zone and type of newspaper item. The distribution zone of a document is the geographical region where the newspaper was distributed, and can be one of the following values: "local", "national", one of the former Dutch colonies ("Indonesia", "Suriname", or the "Antilles"), or, in a few cases, "unknown". The available newspaper item types are: news articles, advertisements, announcements (relating to births, marriages, deaths) or images (illustrations or photographs, search on the caption text). Table 6 shows the percentages for each metadata value in the collection.

The search interface combines full-text search with facets. The facets are filters based on the metadata attributes of the collection, and include time facets, indicating the publication date, item type facets, and distribution zone facets. In addition, users may change the relevance ranking of the results on a results page to alphabetical or chronological ordering. From a results page, a user can click on a document and, after viewing a document, download it.

The logs used in this experiment were collected between October 2015 and March 2016 (raw data 200M records). In addition, we received the full text digitalization and metadata records of the historical newspaper collection (103M documents at the time), making it possible to link the clicked documents in the logs to the metadata records of all the documents in the collection.

3.5.1 *Session Identification and Representation*

The user logs contain all HTTP requests to the server. This includes the requested URL, the referrer URL (the origin of the request), the IP address of the client, the browser agent and a timestamp.

Table 5: Session features based on metadata

	Publication date clicks
1	percentage of clicks published between 1600 and 1899
2	percentage of clicks published between 1900 and 1929
3	percentage of clicks published between 1930 and 1949
4	percentage of clicks published between 1950 and 1995
	Item types clicks
5	percentage clicked articles
6	percentage clicked family announcements
7	percentage clicked advertisements
8	percentage clicked images
	Distribution zone clicks
9	percentage of clicks with a local distribution zone
10	percentage of clicks with a national distribution zone
11	percentage of clicks with an Indonesian distribution zone
12	percentage of clicks with a Suriname distribution zone
13	percentage of clicks with an Antilles distribution zone
14	percentage of clicks with an unknown distribution zone
	Search with time facets
15	percentage of search with time facets
16	percentage of search with time facets within 1600 and 1899
17	percentage of search with time facets within 1900 and 1995
18	percentage of search with time facets within 1900 and 1929
19	percentage of search with time facets within 1930 and 1949
20	percentage of search with time facets within 1950 and 1995
	Search with item facets
21	percentage of search with item facets
22	percentage of search with article facets
23	percentage of search with family announcement facets
24	percentage of search with advertisement facets
25	percentage of search with image facets
	Search with distribution zone facets
26	percentage of search with distribution zone facets
27	percentage of search with local facets
28	percentage of search with national facets
29	percentage of search with Indonesian facets
30	percentage of search with Suriname facets
31	percentage of search with Antilles facets
32	percentage of search with unknown facets

Table 6: Collection metadata

Publication date	**Percentage**
between 1600 and 1899	12%
between 1900 and 1929	27%
between 1930 and 1949	26%
between 1950 and 1995	35%
Item type	
articles	67%
family announcements	2%
advertisements	29%
images	2%
Distribution zone	
local	40%
national	45%
Indonesia	12%
Suriname	1%
Antilles	2%
unknown	0.05%

Table 7: Clusters March - part 1

Number of sessions	Silh. width	Label	Description
8667 (19%)	0.66	no metadata	Sessions with little to no metadata values.
7238 (16%)	0.42	recent national	At least half the sessions include 100% clicks between 1950-95 with a national distribution zone. About 25% sessions additionally include other clicks.
7549 (16%)	0.25	recent local	At least half the sessions include 100% local clicks, of which 85% or more are between 1950-95. About 25% sessions additionally include other clicks.
6837 (15%)	0.13	1930-49	At least half the sessions in this cluster include 100% clicks between 1930-49. About 25% of sessions additionally include clicks after 1949.
5537 (12%)	0.02	1900-29	At least half the sessions include 100% clicks between 1900-29. In the sessions more clicks have a national distribution zone than a local one. About 25% of sessions additionally include a minority of clicks between 1930-49s.
4156 (9%)	0.19	historical	At least 75% sessions include facets or (a majority of) clicks from before 1900. About half the sessions also include clicks on adverts, about 25% clicks on announcements. There are more clicks in the sessions on documents with a local distribution zone than with a national one.

Table 8: Clusters March - part 2

Number of sessions	Silh. width	Label	Description
2701 (6%)	0.62	family	At least 75% sessions include announcement facets and a majority of clicks on announcements. In addition, more clicks in the sessions are local than national, and published in the 20th century. About 25% of sessions additionally include clicks on adverts; 25% include clicks on Indonesian documents; 25% clicks on pre-1900 documents; and 25% include time facets or distribution zone facets.
2101 (5%)	0.37	article	All sessions include item facets. At least 75% include article facets, 25% advertisement facets. Most of the sessions include a majority of article clicks; some sessions additionally include advertisement clicks. About 25% include time facets between 1900-95; and 25% include national distribution zone facets.
850 (2%)	0.64	Suriname	At least 75% sessions include a majority of Suriname clicks, and about half the sessions include Suriname facets. Additionally, about half the sessions include clicks between 1950-95; and about 25% sessions include announcement facets; 25% include some announcement clicks; and 25% include some advertisement clicks.
208 (0.5%)	0.0	Antilles	All sessions include Antilles facets; at least 75% sessions also a majority of Antilles clicks.

We identified sessions from these logs using a *clickstream* model, following the navigational path of a user on the search platform. We removed all logs stemming from web crawlers based on browser agents or a request for robots.txt, redirects, and the loading of style sheets and images. Sequential requests for the same URL right after each other are removed as well, as these are likely reloads of the browser and do not represent a new user interaction.

We group the records by IP address, using the referrer URL to link subsequent requests. Since we are interested in search behavior in relation to the metadata of facets used and documents clicked, we kept only sessions where the sequence consists of more than one search interaction or clicked document in the newspaper collection. This brings the total number of sessions to 255,175 in six months.

For the clustering we create a feature set relating to the metadata values of (i) the clicked documents and of (ii) the facets used in search (Table 5), proportional to the number of clicks or search interactions in that session. For the time facets and publication dates of clicked documents, we split the values into four bins based on equal proportions over all clicked documents and rounded to decades. This leads to a single bin for the period before 1900 and three bins in the 1900-1995 period (Table 6 for the distribution of these values within the collection). The values for the time facets are based on dates within the indicated years, using the same bins as for clicks. We add an extra time facet for the period 1900-1995 to capture those facets that cross the boundaries of the bins in the period 1900-1995.

We define additional session variables influenced by the user interactions in the search interface, and not used for clustering; these are the duration of a session, the number of search interactions and clicks, the use of facets or multiple facets in an interaction, the use of quotes in queries and the reranking of the results by time. We compute these variables – except the total duration and length (number of interactions) – proportional to the length of the session or the number of search interactions.

3.5.2 *Clustering Sessions*

We applied principal component analysis on the metadata features in each month separately, in March this led to 15 principal components with an explained variance in the data of 75%. These 15 principal components are used for the clustering, reducing the number of dimensions for the clustering from 32 to 15. We have chosen the number of clusters *k* based on the average silhouette widths for this month, the highest average silhouette width under twenty is for *k* equals 10 with a value of 0.35, the first silhouette width above 0.3 (not a high silhouette width but this is not unexpected considering the 15 dimensions – the principal components – used to cluster). We have clustered the sessions from the month March (45,845 sessions) into ten clusters, and using the same value for *k* as for

March we have also clustered the sessions from the previous months to evaluate the stability of the patterns found in March.

3.6 RESULTS

We describe the resulting ten clusters from March ($k = 10$ based on the average silhouette widths as mentioned in section 3.5.2) in terms of the original values of the metadata features used for clustering, and investigate the stability of this clustering over time. Then, we analyze the search behavior within the clusters.

3.6.1 *Clusters*

We have labeled the clusters using the most distinctive values of the session features present in a cluster (Tables 7, 8), and provided short descriptions of the clusters. The clusters show focused sessions centered around dedicated metadata categories.

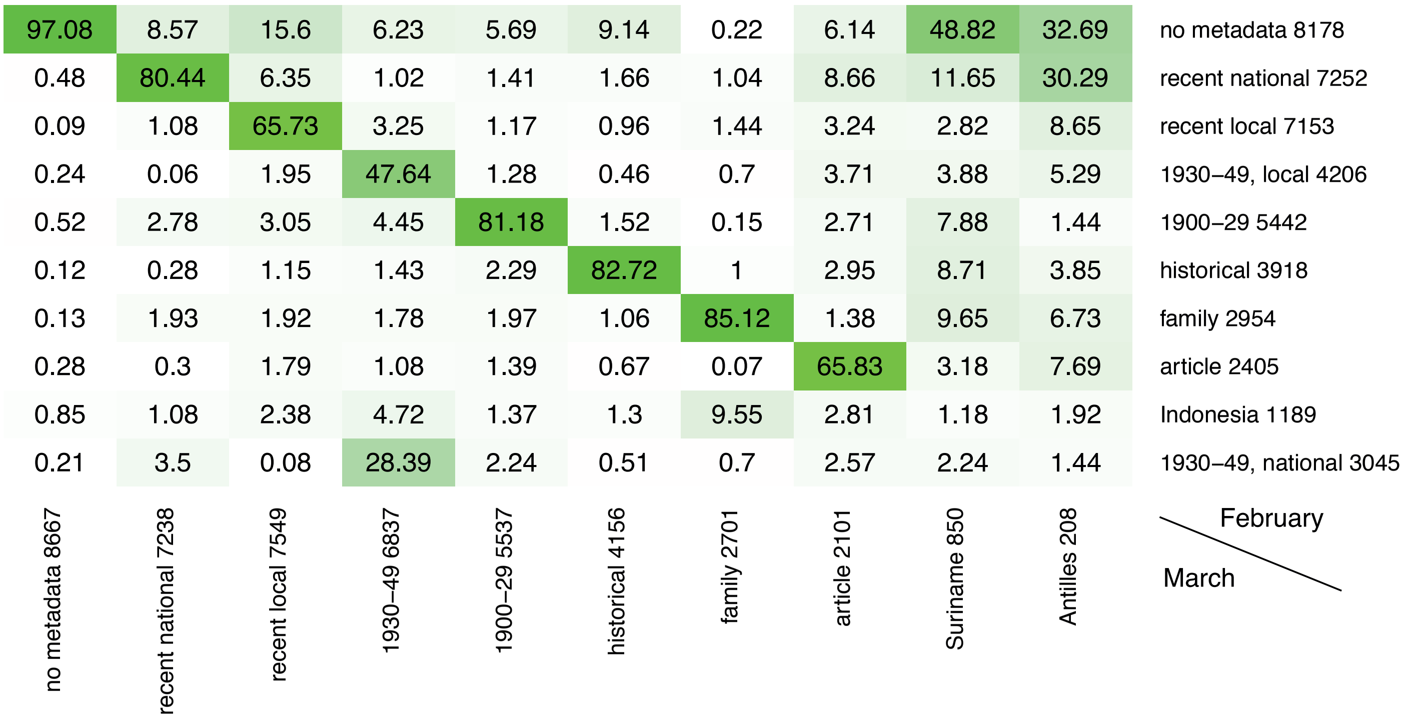
For example, one of the larger clusters, the *recent national* cluster (16%), is exclusively centered around the recent national documents in the collection. In most sessions in this cluster, all the clicked documents are published between 1950-95 and have a national distribution zone. Similarly, most sessions in the *recent local* cluster (16%) contain only clicked documents with a local distribution zone of which the large majority is published between 1950-95. This indicates that users searching in the recent parts of the collection are mainly searching for documents with either a local or a national distribution zone and not both, resulting in two separate clusters.

Other clusters are likewise focused, either on a specific period, such as the *1930-49* cluster (15%) with the clicks on documents published during the Great Depression and World War II in the Netherlands, the *1900-29* and the *historical* cluster; or on a specific item type, such as the *family* cluster, where in addition to a majority of announcement clicks most sessions also include announcement facets, and the *article* cluster. For the two smallest clusters, based on a distribution zone, the *Suriname* cluster and the *Antilles* cluster, most sessions include the distribution zone facet next to a majority of clicks from the distribution zone.

The largest cluster (19%), however, is the cluster with sessions without distinct metadata, labeled *no metadata*. Despite leaving out the sessions of length 1 in the data preparation, there is still a relatively large cluster of sessions where hardly any facets are used or documents clicked, leading to a sessions without any representative metadata values.

Table 9: Stability testing over time (March)

clusters	freq	Oct	Nov	Dec	Jan	Feb
combined	100%	74%	68%	75%	72%	75%
no metadata	19%	93%	96%	95%	96%	97%
recent national	16%	80%	78%	81%	75%	80%
recent local	16%	60%	62%	63%	59%	66%
1930-49	15%	72%	49%	77%	50%	48%
1900-29	12%	58%	26%	79%	67%	81%
historical	9%	82%	82%	68%	81%	83%
family	6%	88%	86%	86%	84%	85%
article	5%	61%	69%	19%	67%	66%
Suriname	2%	49%	55%	54%	38%	49%
Antilles	0.5%	35%	29%	27%	30%	33%



no metadata 8667	recent national 7238	recent local 7549	1930–49 6837	1900–29 5537	historical 4156	family 2701	article 2101	Suriname 850	Antilles 208	February / March
97.08	8.57	15.6	6.23	5.69	9.14	0.22	6.14	48.82	32.69	no metadata 8178
0.48	80.44	6.35	1.02	1.41	1.66	1.04	8.66	11.65	30.29	recent national 7252
0.09	1.08	65.73	3.25	1.17	0.96	1.44	3.24	2.82	8.65	recent local 7153
0.24	0.06	1.95	47.64	1.28	0.46	0.7	3.71	3.88	5.29	1930–49, local 4206
0.52	2.78	3.05	4.45	81.18	1.52	0.15	2.71	7.88	1.44	1900–29 5442
0.12	0.28	1.15	1.43	2.29	82.72	1	2.95	8.71	3.85	historical 3918
0.13	1.93	1.92	1.78	1.97	1.06	85.12	1.38	9.65	6.73	family 2954
0.28	0.3	1.79	1.08	1.39	0.67	0.07	65.83	3.18	7.69	article 2405
0.85	1.08	2.38	4.72	1.37	1.3	9.55	2.81	1.18	1.92	Indonesia 1189
0.21	3.5	0.08	28.39	2.24	0.51	0.7	2.57	2.24	1.44	1930–49, national 3045

Figure 8: Stability matrix, tracing how the sessions in the clusters of March (columns, next to the label the size of the cluster is given) are matched to the cluster centers of February (rows, next to the label the size of the February cluster is given). The percentages in the columns (each column totaling 100 percent) signify the percentages of sessions in the March cluster closest to a February cluster in the rows (distance measured using the Manhattan distance of the metadata features for each session in March to the median values of the metadata features of the clusters of February).

Table 10: Search behavior in metadata clusters March

clusters	duration	length	search	clicks	facets		multiple facets		quotes	reranking re- sults
	median	median	median	median	median	q3*	median	q3*	q3*	q3*
combined	00:13:22	16	81%	18%	0%	46%	0%	3%	0%	0%
no metadata	00:01:43	5	100%	0%	0%	33%	0%	0%	0%	0%
recent national	00:13:30	17	68%	29%	0%	7%	0%	0%	0%	0%
recent local	00:13:16	17	71%	25%	0%	16%	0%	0%	0%	0%
1930-49	00:18:04	20	75%	23%	0%	39%	0%	0%	0%	0%
1900-29	00:17:17	16	71%	25%	0%	31%	0%	0%	0%	0%
historical	00:44:02	36	84%	15%	43%	67%	0%	13%	0%	3%
family	46:07:41	70	83%	17%	52%	77%	9%	36%	29%	19%
article	01:14:03	41	82%	17%	48%	85%	13%	59%	0%	17%
Suriname	00:39:27	30	80%	19%	38%	68%	0%	29%	0%	0%
Antilles	01:03:30	40	81%	18%	61%	87%	11%	48%	0%	28%

* q3, or third quartile, is the middle value between the median and maximum

3.6.2 *Cluster Stability*

To evaluate the clustering, we check the stability of the clusters, matching the sessions in the clusters to the cluster centers of the previous five months. Table 9 shows, per cluster and for all clusters combined, the percentage of sessions in the clusters of March that falls in the highest matching cluster of each of the previous months.

We observe that overall the clustering is stable, with an average stability of 73%. In particular, the *recent national*, *historical* and *family* clusters are stable every month, as is the *no metadata* cluster. (Note that, even while the percentage of family announcements in the collection is low at 2% (Table 6), there is stable user interest in this part.) Nevertheless, not all clusters in March can be traced back in the previous months. For example, the two smallest clusters in March, *Suriname* and the *Antilles*, do not match well in most of the previous months. Furthermore, the *1930-49* and *1900-29* clusters match well in most but not all months.

The silhouette widths (measuring consistency within and between the clusters) of the clusters show no direct connection to whether a cluster is stable over time. The *family* cluster, for example, has a relatively high silhouette width of 0.62, but the *historical* cluster, similarly stable, has a lower silhouette width of 0.19. On the other hand, the *Suriname* cluster also has a relatively high silhouette width of 0.64 but a low stability, as for the *Antilles* cluster, both the silhouette width and the stability are low. This can be explained by the fact that cluster stability and silhouette width measure different things: consistency between clusterings over time and consistency within a clustering respectively.

To better understand the stability measurements in detail, we show a single month of the stability results in Figure 8, comparing March 2016 to February 2016. The clusters in February (on the rows) have been labeled in the same manner as the clusters in March. Here we observe good matching scores on the diagonal for the *no metadata*, *recent national*, *recent local*, *1900-29*, *historical*, *family* and *article* clusters. The *1930-49* cluster, however, does not match to a single cluster, but to two with 48% in one and 28% in another cluster of February. A closer inspection shows that in February the period 1930-49 is split up into two separate clusters, one with mainly local clicks, and a second cluster with mainly national clicks within the same time period. On the other hand, the smallest clusters, the *Suriname* and *Antilles* clusters, have no good match in February at all. The highest matches here are with the *no metadata* cluster. This is because frequently for these sessions the Manhattan distance to the *no metadata* cluster is smaller than to the other clusters, resulting in these cases in an assignment to the *no metadata* cluster.

3.6.3 *Search Behavior*

We observe a split between the first five clusters in March (*no metadata*, *recent national*, *recent local*, *1930-49* and *1900-29*), and the last five clusters (*historical*, *family*, *article* and *Suriname* and *Antilles*) in Table 10. The first five clusters are shorter, use fewer advanced search techniques, and – with the exception of the first cluster – are more click-oriented; the last five clusters are much longer in time spent and pages visited, and use more advanced search techniques such as facets or reranking of results.

Among the first five clusters, the *no metadata* cluster is different. The sessions in this cluster are the shortest, with the majority less than 2 minutes, and consist of only search interactions, no clicks. Nevertheless, users do spend time and effort (median of 5 interactions), possibly we observe users that completed their search using only the snippets on the results page, or these might be examples of failed search. Of the four more click-oriented clusters, all focus on documents published in the 20th century, with the *recent national* on average the highest percentage of clicks per session. The majority of sessions in these clusters does not make much use of the facets or other more advanced search techniques, but show a more "browsing" behavior where users click through results instead of refining their search. This could in part be explained by the collection, these clusters represent larger parts of the collection (Table 6), the digitization of these documents is likely better (the paper of the newspapers are not aged as much, the language in the documents easier to digitize), and fewer search techniques may be needed to find the desired document.

Next, we have five clusters where users spend a long time and visit many pages. The sessions in these clusters contain a lower percentage of clicks, and the majority of the sessions uses facets. Note that, apart from the *article* cluster, these clusters correlate with smaller parts of the collection (Table 6), and thus likely require more effort from the user. Of these, the *family* cluster contains on average the longest sessions, the majority is longer than a day and the number of interactions is by far the highest, in line with previous research into genealogists and family historians [27], and this cluster likely represents in large part this user group. (Sessions longer than a day are unlikely to be sessions where a user continuously searches, but sessions where a user returns to the same search a day later.) This cluster contains just 6% of the number of the sessions in the month, but the number of interactions is high with a median of 70, resulting in a lot of traffic on the search platform even while the percentage of announcements in the collection is just 2%, and suggesting the users in this cluster are highly engaged in their search. In this cluster we also observe the most frequent use of quotes for the queries, this is not unexpected as search within the family announcements are likely to include search for personal names with respect to genealogy and family histories.

3.7 DISCUSSION

Our results demonstrate that patterns of user behavior can be correlated with document metadata in a way that provides clusters that can be described in a meaningful way to collection curators.

METADATA DEPENDENCY Our clustering using the metadata of search and clicks is dependent on by the existing metadata categories the curators have given; however, this is inherent to any curated online collection. It is possible to (additionally) use query analysis and link the query to the metadata of the collection, for example by using a relevant ontology or thesaurus as was done in [48, 51]. Query analysis, however, suffers from several disadvantages: queries can be ambiguous as they form an uncontrolled vocabulary, and queries may include privacy-sensitive information.

SESSION IDENTIFICATION To identify the sessions to be clustered we have chosen a clickstream model, as it can help to split possible multiple users behind a single IP address, and to find complete searches. This approach leads in some cases to shorter or longer sessions than when a timeout is used, think for example when a user continues their search in a new tab thereby breaking off a clickstream-based session, or the opposite case when a user continues the next day with their search, this would lead to a break in a timeout-based session. For example, the sessions in the *family* cluster last for longer than a day in the clickstream-based sessions, when using the timeout-based session definition these sessions would be broken up into multiple sessions. An alternative to a purely clickstream-based session definition could be a combination of clickstream and query-term overlap, even though query analysis can introduce another sort of bias, and also for this reason we have chosen to keep the session definition simple.

EXPLORING K We have clustered the sessions into ten clusters based on the best average silhouette width under twenty, however, the number of clusters *k* can also be used as a parameter of how fine-grained the analysis of the user interests is going to be. As the average silhouette widths for *k* values under twenty illustrate (Fig. 7), higher values of *k* can have similar average silhouette widths, making it possible to first set *k* low for an overview, and then higher to investigate more detailed user interests. The extent to which the value of *k* should be manipulated is dependent on, among other things, the existing metadata categories and the level of detail deemed appropriate by curators. Also, a higher value for *k*, might solve the disappearance of clusters like the *Suriname* and *Antilles* clusters in previous periods, which in the stability matrix merged into the *no metadata* cluster in the month of February (Fig. 8).

FUZZY CLUSTERING Even though the large majority of sessions in each cluster are highly focused, we do find sessions on the edges of the clusters that are a bit more "mixed" with respect to user interests, such as the *1900-29* cluster where some of the sessions also include a minority of clicks from between 1930-49 (Table 7). The clustering algorithm we applied, however, is binary, in the sense that a session belongs to a single cluster, even if in some cases it is possible that it has characteristics matching more than one. For future work, it could be interesting to look into more fuzzy or soft clustering techniques, where a session can belong to multiple clusters.

CLUSTERING SEARCH BEHAVIOR It is possible to cluster the same sessions using interaction features describing search behavior, such as session duration or number of clicks. These "behavior" clusters can then be mapped to the identified user interests, as opposed to a simple statistical summary, making it possible to find more than a single search pattern for each user interest. However, a first attempt using the same clustering method but with interaction features based on the search interface did not lead to more detailed insights than the statistical analysis provided: the overall overview remained the same. Possibly a search task analysis, such as presented in [42] is more effective here.

3.8 CONCLUSION

By applying a clustering algorithm we were able to identify user interests and investigate the relation between them and search behavior within the historical newspaper collection of the National Library of the Netherlands. The user interests we identified are stable over a six-month period. Our approach can be used to find relations between user interests and behavior in any collection described by metadata, such as digital libraries and archives.

Using the clustering based on the metadata features of search and clicks, we were able to observe users focusing on specific parts of the collection, in some parts spending less time and few search techniques, in other parts spending a large amount of time and a variety of search techniques. This method can help to find and investigate these highly-focused users. These findings can inform the design of more targeted user interfaces providing better access to specific parts of the collection, or help to improve search systems or collection management.

4 COMPARING METHODS FOR FINDING SEARCH SESSIONS ON A SPECIFIED TOPIC: A DOUBLE CASE STUDY

4.1 ABSTRACT

Users searching for different topics in a collection may show distinct search patterns. To analyze search behavior of users searching for a specific topic, we need to retrieve the sessions containing this topic. In this paper, we compare different topic representations and approaches to find topic-specific sessions. We conduct our research in a double case study of two topics, World War II and feminism, using search logs of a historical newspaper collection. We evaluate the results using manually created ground truths of over 600 sessions per topic. The two case studies show similar results: The query-based methods yield high precision, at the expense of recall. The document-based methods find more sessions, at the expense of precision. In both approaches, precision improves significantly by manually curating the topic representations. This study demonstrates how different methods to find sessions containing specific topics can be applied by digital humanities scholars and practitioners.

4.2 INTRODUCTION

Analysis of search logs is an unobtrusive technique for large-scale investigations into user behavior in digital libraries. Users interested in different topics might display different search behaviors. For example, the work presented in [88] demonstrated different search patterns of users searching for five major religions. In a previous study, we observed a distinct search pattern for users searching for documents related to World War II (WWII) [13]. For these types of studies, we need to be able to retrieve those user interactions from the search logs that relate to a user interest in a specified topic. In this paper, we propose and compare generally applicable methods to find user interactions that relate to a specified topic from a larger set of logged search interactions. We work at the level of sessions (coherent sequences of user interactions with the collection) as they capture the context in which individual user actions occurred and connect search actions to clicks on documents. We address two research questions:

(**RQ1**) *How can we represent a specified topic?*

(**RQ2**) *How can we use the topic representation to retrieve relevant sessions?*

To answer the first research question, we look into different, consecutive ways to build a term list as a representation of a topic: i) using semantic relations in an explicit knowledge resource, ii) applying local word embeddings trained on the documents in the collection, and iii) in each step, by manual curation of the term lists by domain experts. To answer the second research question, we look into matching the different term lists to user sessions. We match them to either a) the user queries, or b) the contents of the clicked documents. We compare and discuss the combined methods in terms of number of retrieved sessions as well as estimated precision scores. We conduct our research using data from the National Library of the Netherlands, focusing on search in their historical newspaper collection[1]. In previous work [12, 13], the search logs of the digital library were already split into user sessions, and we consider this session identification step outside the scope of this paper. We present a double case study in the context of two historical topics with societal relevance: WWII (a pivotal period in Dutch and global history), and feminism (a movement that has had and still has an impact on Dutch society). We evaluate our methods on a ground truth of over 600 manually assessed sessions per topic.

This study contributes insights into how different topic representations and matching approaches perform when retrieving topic-specific sessions. Our results show that when sessions are retrieved based solely on user queries, the precision is high, however, the set of sessions remains small. When the document-based matching approach is used, the set of sessions retrieved increases, but at the expense of precision. Moreover, we find that by manually curating the term lists we improve precision while still preserving a larger set of sessions. The two topics investigated in this paper show similar general patterns in their results, however, we observe a higher overall precision for the more popular topic (WWII). Finally, our study demonstrates how different methods can be applied and combined by digital humanities scholars and practitioners to retrieve topic-specific sessions.

4.3 RELATED WORK

We discuss work on detection and analysis of user interests; and how knowledge resources and word embeddings have been used to enrich queries and documents.

4.3.1 *Topic-specific Search Log Analysis*

Search behavior in digital libraries and archives has been studied frequently, e.g., [15, 24, 47, 61, 77, 87]. Topics have been detected in search logs for various reasons; for example, to determine user interests [44, 48, 65, 70], to uncover topic-specific search patterns [12, 88], or to recognize changes in topic within a session [45]. Other studies observe topic-specific search patterns by analyzing

[1] The National Library of the Netherlands has granted us access to user logs from their search platform `https://www.delpher.nl`, providing access to collections from the National Library of the Netherlands and other heritage institutions

logs from a specific search interface, such as a health portal [18], or a media archive [48, 51].

In most cases, topics are detected in search logs by investigating the queries that users entered. Sometimes, in addition to the query, the contents of what was clicked in sessions is also taken into account. For example, query analysis has been combined with mouse-fixation behavior and the metadata of clicked documents [44]. In previous work, we used the metadata of clicked documents, as well as the use of facets to filter search results, to understand search behavior in different parts of a digital library collection [12, 13]. In this study, we investigate and compare how query-based and content-based approaches perform.

We represent a topic as a list of terms. This is similar to the work presented in [88], where users searching for five large religions were identified by matching queries to five respective lists of professionally curated terms. The authors of [31] used a list of terms and phrases that signify specific types of questions, and matched these to queries in order to analyze how people learn within sessions.

4.3.2 *External Resources to Enrich Queries or Documents*

In previous work, knowledge resources have been used to classify documents in collections, e.g., by finding relevant Wikipedia categories [90]; or by finding relevant concepts [68] for the documents in the collection. In other cases, knowledge resources have provided a semantic enrichment of user queries, e.g., to categorize queries [51, 70, 94]; or for query expansion during search, e.g., by searching related concepts in Wikipedia [1, 3, 35]. In the present study, we use Wikipedia as a knowledge resource to expand a single term topic representation. Wikipedia is widely used, publicly available and has a broad coverage, making it applicable to many use cases beyond the ones studied in this paper. This makes Wikipedia an attractive option, even though we are aware of the fact that Wikipedia is biased both with respect to which topics are represented in the articles and the contents of the articles [19, 81].

Word embeddings have been used by researchers in several query expansion applications, such as search, text classification, plagiarism detection [5]. In this type of distributed representation, words with similar meanings are more likely to be close together [73]. The semantic associations between words that thus emerge, have been shown to be effective in tackling the query-document vocabulary mismatch problem [32]. We use word embeddings to expand on the terms representing a topic, and as such to be able to increase the number of sessions found. Specifically, we use local embeddings, following [28], where it was demonstrated that corpus-specific embeddings perform better than global embeddings for query expansion.

4.4 DATA

We use collection and log data from the National Library of the Netherlands. As a knowledge resource for the topic representations we use Wikipedia.

4.4.1 *Document Collection and Search Logs*

Our research is conducted using a document collection and search logs from the National Library of the Netherlands. The library maintains a number of digitized historic collections, our focus is on the historical newspaper collection spanning almost four hundred years (1618-1995). Within this collection, users can search using full-text search and facets (filters based on the metadata attributes of the collection). The logs used in this study were collected between October 2015 and March 2016 (raw data 200M records). They record the user interactions with the search system. These interactions have previously been grouped in sessions, to be able to study search behavior in context. The log records have been cleaned and processed, and sessions have been identified based on a *clickstream* model as described in [12, 13], using the IP address as identifier and connecting sequential HTTP requests to follow a user navigating the search system. For this study, we have retained all sessions which include clicked documents within the newspaper collection, resulting in a total of 204,266 sessions over the six month period. In addition, we received the full text digitization and metadata records of the historical newspaper collection (103M documents at the time).

4.4.2 *Knowledge Sources for Topic Representations*

In this double case study, the topics of interest are WWII ("Tweede Wereldoorlog" in Dutch), and feminism ("feminisme" in Dutch). These topics are selected based on their societal relevance, and thus their value to digital humanities scholars. For example, professional historians from the Dutch NIOD Institute for War, Holocaust and Genocide Studies are interested in understanding how people search for topics related to World War II (WWII) in the media, and how this changes over time. We represent the two topics using lists of relevant terms. In the first expansion of the list of relevant terms, we use Wikipedia. As this is a publicly available knowledge resource, with many possible applications in different domains for different topics, it contributes to the general applicability of our methods. Our topics of interest correspond to the existing Wikipedia categories for WWII[2] and for feminism[3]. We have selected the top-300 Wikipedia articles in these categories, based on the popularity within the same period as the logs (October 2015-March 2016). To collect these Wikipedia articles, we have used the

[2] `https://nl.wikipedia.org/wiki/Categorie:Tweede_Wereldoorlog`

[3] `https://nl.wikipedia.org/wiki/Categorie:Feminisme`

tool Massview Analysis[4], including the subcategories. The top-300 most popular Wikipedia articles within the WWII category counted to of a total of 4.7 million views, compared to 1.2 million views within the feminism category. The assessed Wikipedia articles for the topics are available online[5]. We use the popularity ranking as an indicator for public interest in the topics described in the articles as the use of Wikipedia is a strong indicator for how this interest is composed in a country such as the Netherlands.

4.5 METHOD

We describe the different methods we compare to find topic-specific sessions. First, we explain the consecutive steps to build term lists representing the topics. Second, we describe how to use the term lists to find topic-specific sessions in a larger set of sessions. Third, we explain how we evaluate the different methods.

4.5.1 *Creating Term Lists*

We compare five ways of creating terms lists to represent the topics, where each list builds on the previous list.

List 1. Single term: List 1 contains a single term or phrase to represent the topic, in our case "Tweede Wereldoorlog" (WWII) or "feminisme" (feminism).

List 2. Wikipedia: For this list, we leverage the semantic relations in Wikipedia to find additional terms to represent the topic. First, we match the term in List 1 to their corresponding Wikipedia category and add them to List 2. Then, we take the article titles of pages within that category or any of its subcategories. To increase the likelihood that these article titles are indeed relevant terms, we select only the top-300 most popular titles based on Wikipedia page view data. Some Wikipedia article titles require preprocessing. Where the title only consists of a named entity, it is used as-is. In the case of a title consisting of a named entity and a class between parentheses, for example, "The Color Purple (film)", we separate the class from the named entity. In the case of a title consisting of a classifying noun, preposition, named entity title phrase, for example "Bombardement op Rotterdam" (Bombing of Rotterdam), we leave out the preposition when it is not part of a named entity.

List 3. Wikipedia curated: For List 3, we ask domain experts to manually assess the terms in List 2 and remove those that are less relevant, in the assumption that this will improve the quality of the terms on the list and thus improve the precision of the matched sessions. For the WWII terms, experts from the NIOD Institute for War, Holocaust and Genocide Studies were involved in the assessment; for the feminism terms, two of the authors of this paper familiar with the

[4] `https://pageviews.toolforge.org/massviews/`, by MusikAnimal, Kaldari, and Marcel Ruiz Forns.
[5] `https://edu.nl/4arxw` and `https://edu.nl/9qbfr`

topic. The assessment is based on the question whether it is plausible that someone with a specific interest in WWII or in feminism would consult the subject described in the corresponding Wikipedia article. Articles in which our main topic of interest (WWII or feminism) is only of minor importance – for example, in biographies of politicians, actors and professional sportsmen for whom the WWII period was not pivotal in their lives – were removed from the lists. Similarly, articles referring to a topic occurring outside the time period of the historical newspaper collection (1618-1995) – for example, movies or books published after 1995 – were also removed. We note that in the case of the WWII topics, most of these are topics from the war period itself or from the period leading to the war, but also included are issues that are part of the post-war remembrance culture and therefore refer to the period after WWII.

List 4. Wikipedia expanded: We expand the terms in List 3 using local word embeddings to create the larger List 4. We describe this process in detail in Section 4.5.2.

List 5. Wikipedia expanded and curated: To create List 5, we ask domain experts to asses the terms in List 4, using the same process as for List 3.

This results in five term lists for our topics (see Table 1).

4.5.2 *Term Expansion Based on Local Word Embeddings*

To expand the term lists, we employ a widely used technique based on word embeddings [72], vector representations of words where words that appear close together in the vector space are likely to have a similar meaning. We use local embeddings instead of global embeddings, training on a selected set of topically relevant documents, as we expect term similarity to be highly dependent on the context of the topic, as was shown in [28]. For this purpose, we query the library's newspaper collection for documents that contain the terms in List 3, and use those as a topically-constrained training corpus. We work with the Indri search engine [84], using default Dirichlet smoothing [82]. The terms are translated to Indri queries, searching for an exact phrase match, or in the case of a title and a class description an exact phrase match and a Boolean AND for the class. We use the gensim library[6] for both preprocessing and to train the embeddings. To preprocess the digitized text in the training corpus, we first identify the combination of symbols and characters that mark the beginning and end of each article, and remove them. Next, we extract the sentences to be broken down into tokens, and lowercase the text. For the configuration of the hyper-parameters of gensim's word embedding algorithm, we refer to the *set expansion* solution proposed by [69] where the authors suggest setting the word vector size to 100 and the window size to 10[7]. The reason to use a window size as large as 10,

[6] https://radimrehurek.com/gensim/

[7] https://github.com/NervanaSystems/nlp-architect/tree/master/nlp_architect/solutions/set_expansion

Table 1: Number of terms in each term list

	1: single	2: wiki	3: wiki curated	4: wiki expanded	5: wiki exp&cur
WWII	1	300	200	728	364
Feminism	1	300	199	703	327

is the empirical evidence that larger window sizes are good at providing more topical similarity [63]. Since we are interested in identifying phrases that can be made up of multiple words (e.g., "Nationaal-Socialistische Beweging", "Tweede Wereldoorlog"), we instruct the model to learn bigrams and trigrams (phrases that contain two and three words). With these settings the model is expected to find associations for the single or multi-word target phrase, and suggest related words (made up of phrases consisting of one or more words). Once the model is trained, we query it using the terms in List 3 as seeds. We retain the top-3 most similar words for each term, and add them as expanded terms to List 4.

4.5.3 *Matching Terms to Sessions*

We match the terms of the five lists to sessions in two ways: matching the terms to (a) the user queries and to (b) the clicked documents.

In the **query-based** approach, user queries in the sessions are compared to the terms in the lists using exact phrase matching. As there is little context in a user query, we only include the named entity and not any information included in brackets (such as a class or publication year for the terms based on the Wikipedia article titles). Sessions are considered relevant to a topic if they contain at least one query that contains words matching a term from the topical term list.

In the **document-based** approach, we leverage the contents of the documents clicked in the sessions. For the matching of a term with the content of clicked documents, we include – where present – the class or the noun in the set of terms. This results in for example, the terms *"A Bridge Too Far" AND film*, or *Bombing AND Rotterdam*. For the WWII matching we include an extra step: we remove all matched clicked documents published before 1920, as WWII is a topic based on a historical period, and any documents from before 1920 are considered not relevant. Thus, we retrieve all sessions in which at least one matching document has been clicked.

4.5.4 *Manually Evaluating Retrieved Sessions*

The different methods provide us with sets of sessions for each of the five term lists based on either the query matching, and the document matching, with a total of ten sets of sessions for each topic. To estimate the precision of the resulting ten sets of sessions for each topic, human raters assess samples drawn from these sets. The raters judge whether one of the information needs of the user in that session is to find newspaper documents about a topic that is directly related to the topic of interest. To do this, the rater can inspect the session, using a visualization that includes the search interactions with the queries and selected facets [14], and the clicked documents and their metadata and content. We use inter-rater reliability to check the agreement among the raters.

4.6 RESULTS

We apply the ten methods described in Section 4.5 – five ways to represent a topic as a term list, combined with two approaches to match the terms to a session – to the full set of sessions. This results in ten retrieved sets of sessions per topic.

To estimate precision, we draw samples from each set and manually assess a total of 1243 sessions. We compute the inter-rater agreement using Cohen's κ [25] based on a dual assessment of about 50 sessions per topic. We observe a κ of 0.90 for WWII and 0.84 for feminism, demonstrating good agreement.

SIZE AND PRECISION Figure 2 shows the number of sessions and the precision of each set. As expected, the use of a single term to represent a topic (List 1) results in almost perfect precision but a low number of retrieved sessions. Precision remains high (97% to 100%) when using the longer, curated lists (Lists 3 and 5) in a query-based matching. This method increases the number of retrieved sessions significantly (5 to 13 times as many, in our case). When these lists (3 and 5) are used for document-based matching, the number of retrieved documents increases even more; however, precision is lower. On the WWII topic, precision of this method may still be acceptable (74% to 81%) but on feminism it is probably not (56% and 63%). For the expanded term lists in their un-curated form (Lists 2 and 4), precision drops depending on matching method and topic. When List 2 is used for query-based matching, precision is 83% on the WWII topic, which may still be acceptable. For document-based matching, and/or when applied to feminism, precision will be too low for most applications (37% to 66%). List 4 results in low precision in all cases (9% to 52%).

Note that Lists 4 and 5 were created by expanding List 3. In theory, the same local embedding-based expansion method could be applied to Lists 1 and 2. However, in practice, this is not promising, as List 1 consists of a single term and List 2 has relatively low precision. For that reason, expansions of List 1 and 2 were not included in our experiment.

COMBINING TWO MATCHING METHODS Table 3 shows the number of sessions that appear in both the query-based and document-based session sets, i.e., the intersection of the two sets. For List 1, the intersection is relatively small: e.g., for WWII, only 23 of the 89 sessions retrieved with the query based method are also in the document-based set. We conclude that when using a single term topic representation, a combination of query-based and document based matching is a good way to increase the number of retrieved sessions. For List 2, 3, 4 and 5, on the other hand, the intersection is relatively large; the majority of the sessions retrieved with the query-based methods are also retrieved with the document-based methods. Combining the two methods is less worthwhile here.

Figure 2: Size (count) and precision (percentage) for the retrieved session sets

topic representation	query-based matching				document-based matching			
	WWII	Feminism	WWII	Feminism	WWII	Feminism	WWII	Feminism
1: single	89	26	100%	100%	116	37	95%	100%
2: wiki	667	702	83%	44%	10,001	8,260	66%	37%
3: wiki curated	471	222	100%	100%	7,434	5,341	81%	56%
4: wiki expanded	4,977	6,064	52%	20%	103,833	126,656	22%	9%
5: wiki exp&cur	626	339	100%	97%	12,044	10,519	74%	63%

Table 3: Intersection of query- and document-based session sets

topic	**WWII**			**Feminism**		
representation	**query only**	**both**	**doc only**	**query only**	**both**	**doc only**
1: single	66	**23**	93	9	**17**	20
2: wiki	142	**525**	9,476	386	**316**	7,944
3: wiki curated	116	**355**	7,079	39	**183**	5,158
4: wiki expanded	107	**4,870**	98,963	390	**5,674**	120,982
5: wiki exp&cur	114	**512**	11,532	45	**294**	10,225

ERROR ANALYSIS Our manual annotation effort gave us insight into the types of errors that occur. Some sessions were incorrectly retrieved because of terms in the term lists that are not unique to the two topics, WWII and feminism. For example "concentration camps" may also occur in documents about the Indonesian Independence War; "emancipation" may occur in documents about slavery or religion; the term "gas chamber" is now almost uniquely associated with WWII, but had a different meaning historically; Anne Frank's last name is common in the Netherlands and appears in sessions of family historians unrelated to Anne Frank, and in many documents throughout the collection that are not related to WWII. This type of error happens with all methods but is more frequent when using the document-based matching. We hypothesize that users act as a "smart filter", as they are less likely to use generic or ambiguous query terms without adding meaningful modifying terms. A future direction of research could be to investigate if only selecting terms that users used in their queries might increase precision for a document-based matching.

Another cause for errors in the document-based matching is brought on by mistakes in the digitization process, such as incorrectly set document boundaries, or when the newspaper document contains multiple topics, such as articles summarizing local news or presenting a cultural calendar.

4.7 LESSONS LEARNED

In general, query-based matching results in higher precision than document-based matching. Document-based matching, on the other hand, results in more retrieved sessions (up to 20 times more, in our experiments) at the loss of precision. We have experimented with a more narrow inclusion of document-based matched sessions (e.g., matching more than one document in a session), but a preliminary inspection did not seem to increase precision. Future work, though, could investigate this further. A combination of query- and document-based matching is useful when a topic is represented as a single term. In this case, the combination retrieves significantly more sessions without loosing precision. We hypothesize that the combination is similarly worthwhile when topics are represented as relatively short terms lists.

When a topic representation is expanded to a longer term list, manual curation of the terms is key. This holds for both our expansion methods (using a knowledge resource or local word embeddings). Curation is especially critical for document-based matching. In this work, we leveraged the category structure of Wikipedia. Future work will have to determine how other knowledge resources and other semantic relations perform.

All our methods perform better on the WWII topic than on the feminism topic. This could in part be due to the fact that WWII is a less abstract topic than feminism and as such may be easier to detect. Even so, we hypothesize the prevalence of the topic in our data plays a large part as well: WWII is not only the more popular topic on Wikipedia, the retrieved session sets (both query- and document-based) are larger than the respective sets containing topics related to feminism. It would be interesting to investigate this further using topics at different abstraction and popularity levels. In general, we expect that both knowledge-based and corpus-based expansion methods work better on more popular topics.

4.8 conclusion

Understanding search behavior for topics with societal relevance can provide digital humanities scholars insights into the interest in these topics within a collection, and the research presented in this paper supports this objective. We compared different methods on how to retrieve user sessions containing specified topics, using different term lists to represent the topics and applying term matching to user queries and to clicked documents. We observed that when retrieving sessions is based solely on user queries, the precision is high, but the number of sessions retrieved small. Using the document-based matching approach, more sessions are retrieved, at the expense of precision. We found that manual curation is essential, without this step the expanded lists (using a knowledge resource or local word embeddings) perform poorly in terms of precision. This effect was particular strong for the document-based matching. Furthermore, we observed a higher overall precision for the more popular, WWII topic. In conclusion, we believe this research helps to pave the way for a better understanding and communication of topic-specific user interests within collections for digital humanities scholars as well as collection owners and practitioners.

Acknowledgements

We would like to thank Lynda Hardman (Centrum Wiskunde & Informatica) for their support. The Wikipedia articles related to WWII were assessed by Kees Ribbens with the assistance of Caroline Schoofs and Koen Smilde.

5 UNDERSTANDING USER BEHAVIOR IN DIGITAL LIBRARIES USING THE MAGUS SESSION VISUALIZATION TOOL

5.1 ABSTRACT

Manual inspection of individual user sessions provides valuable information on how users search within a collection. To support this inspection we present a session visualization tool, Metadata Augmented Graphs for User Sessions (MAGUS), representing sessions in a digital library. We evaluate MAGUS by comparing it with the more widely used table visualization in three representative tasks of increasing complexity performed by 12 professional participants. The perceived workload was a little higher for MAGUS than for the table. However, the answers provided during the tasks using MAGUS were generally more detailed using different types of arguments. These answers focused more on specific search behaviors and the parts of the collection users are interested in, using MAGUS's visualization of the (bibliographic) metadata of clicked documents and selected facets. MAGUS allows professionals to extract more, valuable information on how users search within a collection.

5.2 INTRODUCTION

Many studies on large-scale analyses of search logs in digital libraries [12, 51, 57, 87] provide a high-level view of user behavior through methods that report descriptive statistics over groups of sessions, such as demographics, average session duration or number of clicks. Less is known, however, about how search logs can be presented to a researcher or library professional to understand the behavior of individual users. Manual inspection of user sessions (coherent sequences of interactions of an individual user within the search system) provides valuable information on how a user searches within a collection. System developers, for example, inspect sessions to assess whether user behavior on their platform conforms to the system's design. And library professionals are interested in understanding how users search in different parts of the collection to improve search features.

In our research we inspect and interpret user behavior within a historical collection, for instance how users search within different time periods. In the context of a digital library, the documents in the collection are frequently described with rich, professionally curated bibliographic metadata, which can be used to identify users with specific interests [12].

Frequently, a table visualization is used to inspect individual sessions [40]. A table is uncomplicated, typically consisting of a list of queries and URLs of corresponding clicked documents. This, however, has some disadvantages. As an example, in a table it is not directly visible in which part of the collection a user searched; if this is within a specific period, such as World War II (WWII), or for a specific type of document, such as newspaper adverts or family announcements. Also, it can be difficult to recognize specific interaction patterns, such as a user returning to an earlier query, especially in longer sessions.

We present MAGUS (Metadata Augmented Graphs for User Sessions), a tool for visualizing a session in a meaningful way. We describe the design of MAGUS, and discuss in what ways it can overcome the limitations of a table visualization. For example, MAGUS visualizes the facets selected during search and the metadata of clicked documents, providing a visualization of the specific parts of the collection a user is interested in. We evaluate the MAGUS visualization by comparing it with a table representation in three representative tasks completed by 12 participants from diverse professional backgrounds. The questions we address in the evaluation are: (**RQ1**) *Is session inspection easier in terms of time and effort spent when using MAGUS?*; and (**RQ2**) *Are the answers provided better in terms of accuracy and level of detail when using MAGUS?* For transparency, we report all measurements taken, including those that gave negative or inconclusive results, such as agreement between participants or the perceived workload.

5.3 RELATED WORK

LOG ANALYSIS IN DIGITAL LIBRARIES Search logs collected from digital libraries and archives has been studied frequently [12, 13, 24, 51, 57, 71, 77, 87]. In some cases, studies focus on the detection and analysis of (topical) user interests, for example to categorize search topics [51, 71], or to identify usage patterns in different parts of the collection [12, 13]. These studies focus on a statistical analysis of search logs. However, manual inspection of individual sessions can also provide valuable information on how users search in a search system. For example, in [40], individual search behavior is studied to train and develop machine learning algorithms to be able to predict whether a user is demonstrating struggling or exploring search.

VISUALIZATION OF USER BEHAVIOR Frequently used visualisations such as the Behavior Flow in Google Analytics show results aggregated over all users, providing a bird's eye view of search behavior. Similarly aggregated graph visualizations have been used in earlier work, e.g. [16, 49]. To visualize a single session, a simple table format is frequently used, e.g. [40]. Alternatively, single sessions have been represented as linear sequences of colored blocks, with the colors denoting the type of interaction or page visited [62, 66, 91, 96]. In [75], this idea is applied to the search logs of a digital library, with the colors also denoting typical interactions such as adding or removing facets during the search. In this work, we aim to gain more insights into individual user behavior by visualizing single user sessions. We use a directed graph to represent a complete session, and use color and shape of the graph nodes to represent the search and click interactions. The directed graph representation allows the visualization of both the complete navigational path of a user and the repeated user interactions in a single node.

user studies In a meta-review of empirical studies focusing on user experience, Pettersson et al., [78], report that in 26% of the studies standardized questionnaires are used, and in 31% user activity is logged, often in combination with other methods, with most studies combining quantitative and qualitative data. In our user study, we similarly combine methods, using activity logging and standardized questionnaires, the NASA-TLX [39] and the System Usability Scale [17], combined with open questions and analysis of answers provided to the tasks.

5.4 session visualization

To visualize a session, we need to specify the start and end of the session, record the queries, facets, and search options submitted during the session and collect information about the documents clicked by the user. For our study, we identify sessions from search logs based on the concept of a clickstream, following the navigational path of a user. The queries, facets, and search options represent the user's search interactions on the platform, and are logged by the search system. Documents in a digital library are frequently described using bibliographic metadata. Clicked documents can be annotated with this metadata, providing insights into the parts of the collection the user searched [13].

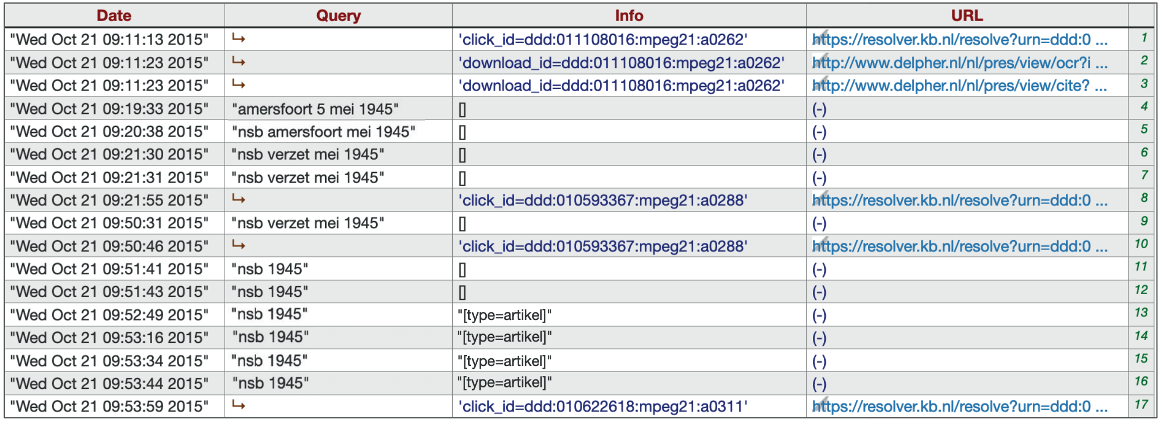
5.4.1 *Session as a table*

Sessions are frequently visualized using a table format, typically containing the user queries and URLs of clicked results sequentially, Table 4 and [40]. The format is uncomplicated, providing an overview of user queries and clicked results. For our table visualization, we adapt the example for the open web, [40], to the context of a digital library. Our table consists of four columns (see Fig. 3): (i) the timestamps of the interactions; (ii) the user query, or in the case of a click or download, an arrow; (iii) additional information on the search interactions, such as selected facets or search options, or a document identifier for clicks and downloads; and (iv) a link to a clicked or downloaded document.

A table visualization suffers from a number of disadvantages. *Issue 1*: it is difficult to see the connection among interactions other than their time sequence. *Issue 2*: it is not easy to recognize repeated interactions, for example, it is not

Table 4: Example table format used by Hassan et al. adapted from [40]

5:55:48 PM	**Query**		employment issues articles
5:55:52 PM		**-Click**	`http://jobseekeradvice.com/category/employment...`
6:01:02 PM	**Query**		professional career advice
6:01:05 PM		**-Click**	`http://ezinearticles.com/?Career-Advice-and-Pro...`
6:03:09 PM		**-Click**	`http://askville.amazon.com/buy-version-Tax-soft...`
6:03:35 PM	**Query**		what is a resume
6:04:21 PM		**-Click**	`http://en.wikipedia.org/wiki/R%C3%A9sum%C3%A9...`
6:07:15 PM			**END OF SESSION**

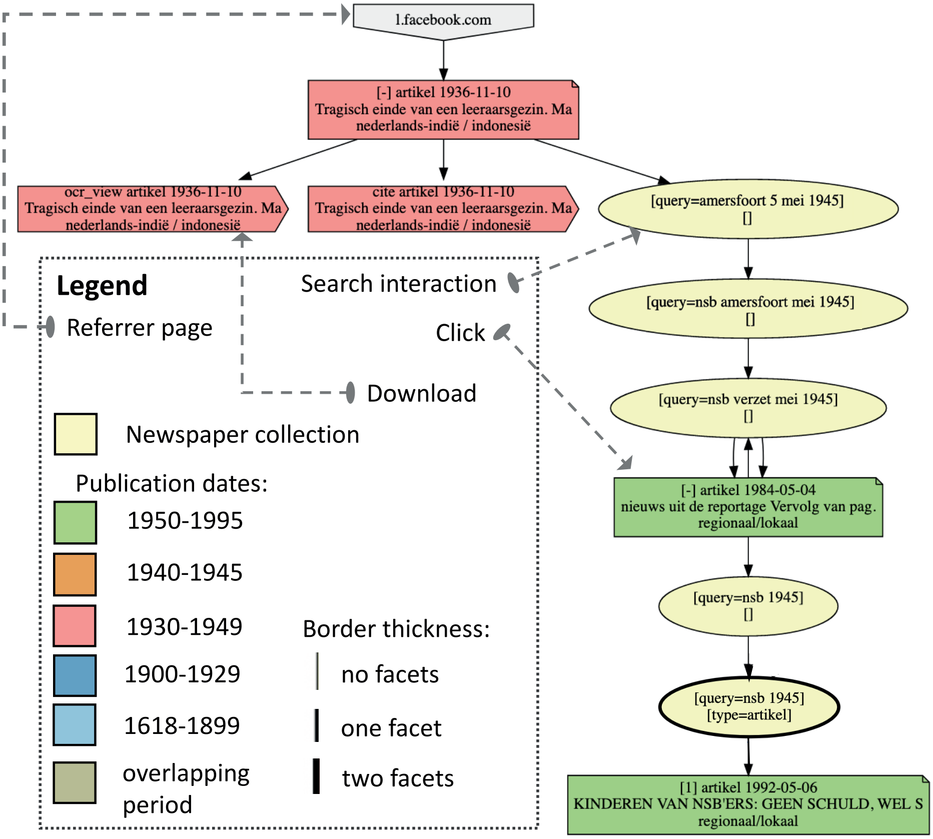


Date	Query	Info	URL	
"Wed Oct 21 09:11:13 2015"	↳	'click_id=ddd:011108016:mpeg21:a0262'	https://resolver.kb.nl/resolve?urn=ddd:0 ...	1
"Wed Oct 21 09:11:23 2015"	↳	'download_id=ddd:011108016:mpeg21:a0262'	http://www.delpher.nl/nl/pres/view/ocr?i ...	2
"Wed Oct 21 09:11:23 2015"	↳	'download_id=ddd:011108016:mpeg21:a0262'	http://www.delpher.nl/nl/pres/view/cite? ...	3
"Wed Oct 21 09:19:33 2015"	"amersfoort 5 mei 1945"	[]	(-)	4
"Wed Oct 21 09:20:38 2015"	"nsb amersfoort mei 1945"	[]	(-)	5
"Wed Oct 21 09:21:30 2015"	"nsb verzet mei 1945"	[]	(-)	6
"Wed Oct 21 09:21:31 2015"	"nsb verzet mei 1945"	[]	(-)	7
"Wed Oct 21 09:21:55 2015"	↳	'click_id=ddd:010593367:mpeg21:a0288'	https://resolver.kb.nl/resolve?urn=ddd:0 ...	8
"Wed Oct 21 09:50:31 2015"	"nsb verzet mei 1945"	[]	(-)	9
"Wed Oct 21 09:50:46 2015"	↳	'click_id=ddd:010593367:mpeg21:a0288'	https://resolver.kb.nl/resolve?urn=ddd:0 ...	10
"Wed Oct 21 09:51:41 2015"	"nsb 1945"	[]	(-)	11
"Wed Oct 21 09:51:43 2015"	"nsb 1945"	[]	(-)	12
"Wed Oct 21 09:52:49 2015"	"nsb 1945"	"[type=artikel]"	(-)	13
"Wed Oct 21 09:53:16 2015"	"nsb 1945"	"[type=artikel]"	(-)	14
"Wed Oct 21 09:53:34 2015"	"nsb 1945"	"[type=artikel]"	(-)	15
"Wed Oct 21 09:53:44 2015"	"nsb 1945"	"[type=artikel]"	(-)	16
"Wed Oct 21 09:53:59 2015"	↳	'click_id=ddd:010622618:mpeg21:a0311'	https://resolver.kb.nl/resolve?urn=ddd:0 ...	17

Figure 3: Session from Fig. 4 visualized as a table

directly visible when a user returns to an earlier query, for example rows 9 and 10 are equal to rows 7 and 8, Figure 3. *Issue 3*: it can be hard to view all interactions in a session at once, to see how often each type of interaction occurs, especially for longer sessions. In the context of a digital library, it is difficult to see *issue 4*: which facets users selected during the search; and *issue 5*: the (bibliographic) metadata of the clicked results which can provide meaningful information about the different parts of the collection users are interested in. To address these disadvantages we have developed a session visualization tool, the Metadata Augmented Graphs for User Sessions (MAGUS).

5.4.2 *Introducing MAGUS*



l.facebook.com
[-] artikel 1936-11-10
Tragisch einde van een leeraarsgezin. Ma
nederlands-indië / indonesië
ocr_view artikel 1936-11-10
Tragisch einde van een leeraarsgezin. Ma
nederlands-indië / indonesië
cite artikel 1936-11-10
Tragisch einde van een leeraarsgezin. Ma
nederlands-indië / indonesië
[query=amersfoort 5 mei 1945]
[]
[query=nsb amersfoort mei 1945]
[]
[query=nsb verzet mei 1945]
[]
[-] artikel 1984-05-04
nieuws uit de reportage Vervolg van pag.
regionaal/lokaal
[query=nsb 1945]
[]
[query=nsb 1945]
[type=artikel]
[1] artikel 1992-05-06
KINDEREN VAN NSB'ERS: GEEN SCHULD, WEL S
regionaal/lokaal
Legend
Search interaction
Referrer page
Click
Download
Newspaper collection
Publication dates:
1950-1995
1940-1945
1930-1949
1900-1929
1618-1899
overlapping period
Border thickness:
no facets
one facet
two facets

Figure 4: Session from Fig. 3 visualized with MAGUS

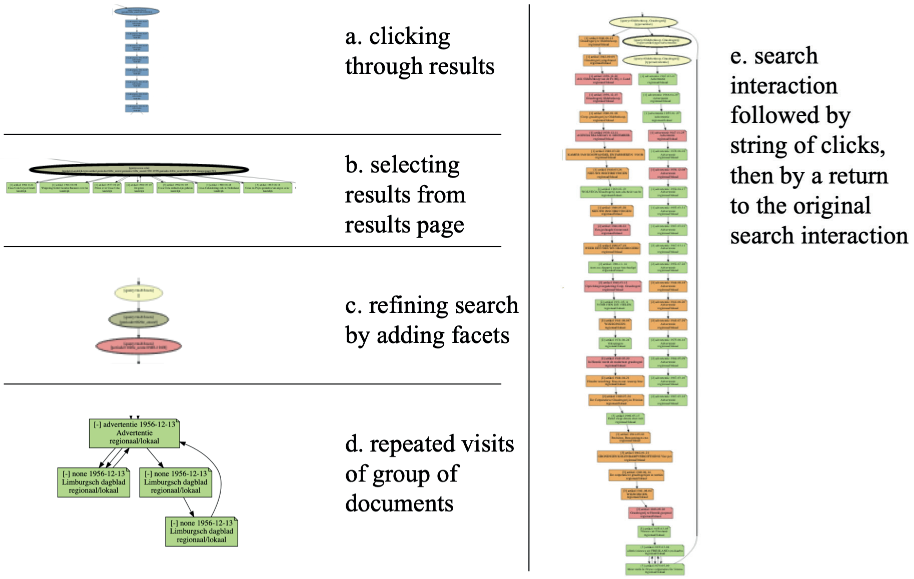

a. clicking through results
b. selecting results from results page
c. refining search by adding facets
[-] advertentie 1956-12-13 Advertentie regionaal/lokaal
[-] none 1956-12-13 Limburgsch dagblad regionaal/lokaal
[-] none 1956-12-13 Limburgsch dagblad regionaal/lokaal
[-] none 1956-12-13 Limburgsch dagblad regionaal/lokaal
d. repeated visits of group of documents
e. search interaction followed by string of clicks, then by a return to the original search interaction

Figure 5: Multiple graph segments in small size showing different types of user behavior.

In MAGUS[1], a session is visualized as a directed graph where the nodes represent the user interactions, and the arrows the navigational path of the user (addressing *issue 1*). MAGUS is built in the SWISH DataLab environment [9], where Graphviz[2] was used for graph visualization. Figure 4 visualizes a relatively small user session. The session starts at the top, where the gray shape indicates that the user arrived by following a link from an external website, in this case a link from a Facebook post. Through the link, the user arrives directly on a specific article (rectangle). From here, the user performed three interactions, temporally ordered from left to right. The user downloads the OCR text of the article, followed by its citation (both indicated by a block arrow shape), then leaves the entry page by initiating a new query and navigating to the search results page (indicated by the yellow ellipse, addressing *issue 3*).

From there, a series of interactions follows: two searches with query refinements (ellipses) and a click on an article (rectangle) are followed by a brief return to the previous page, and back to the article (indicated by the back and forth arrows above the first green rectangle, *issue 2*). To understand the user's search intent it is useful to, in addition to the query, also know which facets were selected (*issue 4*). The user initially used no facets (indicated by the empty square brackets [] in the ellipses), but later added a [type=article] facet, constraining the document type to article (indicated by the thicker line for the last ellipse).

[1] Demo and source code available at `https://swish.swi-prolog.org/p/magus.swinb`
[2] `http://www.graphviz.org/`

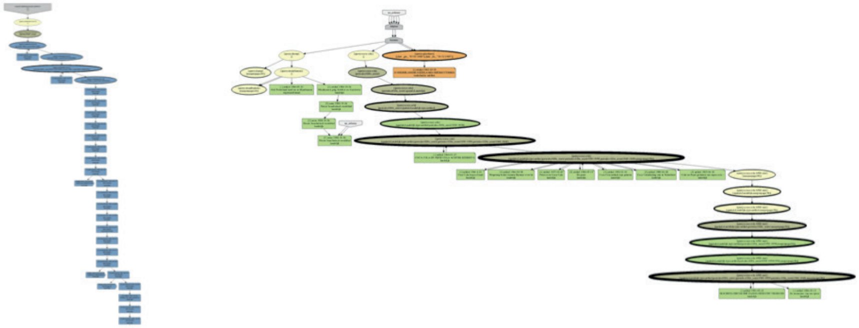
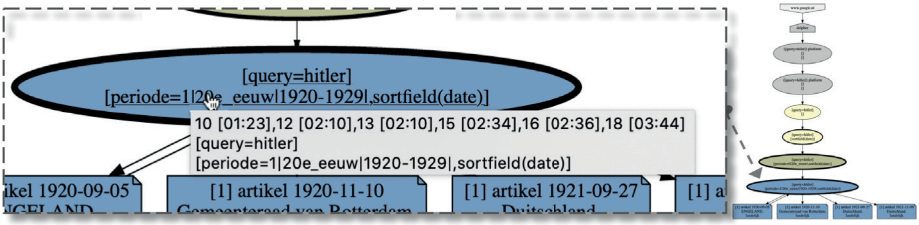


Figure 6: Two small session graphs. The user on the left was browsing through documents published in the 1900-29 period (succession of blue rectangles). The user on the right was using faceted (thick borders) search interactions (ellipses) after 1950 (green).



10 [01:23],12 [02:10],13 [02:10],15 [02:34],16 [02:36],18 [03:44]
[query=hitler]
[periode=1|20e_eeuw|1920-1929|,sortfield(date)]

Figure 7: Hovering over a node displays timestamps with a counter relative to the start of the session, clicking on the node links to the visited web page.

In the historical collection where the example is taken from, it helps library professionals and historians to understand in which period the user is interested. MAGUS allows specific metadata fields to be used to color the nodes in the graph. In this example, we use the publication date from the library's metadata records to color the click nodes. The light red used on the top left indicates documents published in the period around WWII, while the green on the bottom right indicates documents published after 1950 (addressing *issue 5*).

Users exhibit many different interaction patterns, Figure 5, some of which can be more easily distinguished in MAGUS than in a table. For example, a user clicking from one results page to another using the "next button", or a user selecting multiple results from the results page and opening them in a new tab, result in deep vertical versus broad horizontal graphs respectively, Figure 6. Even when the graphs have been reduced in size to a small scale, the difference between the typical "click" behavior of the user on the left can be easily distinguished from the more search-oriented behavior of the user on the right (*issue 3*): the session on the left is dominated by clicks (rectangles) while the session on the right has alternates searches (ellipses) with clicks (*issue 3*). The use of facets is easy to recognize (*issue 4*) in the session on the right by the thick lines used to draw the ellipses of the search nodes, while their color indicates the use of time facets in the post-1950 period (green). The use of the publication dates from the metadata records (*issue 5*) to color the click nodes also immediately conveys that the user on the left is focusing on the 1900-29 period (blue) while the user on the right is more interested in the post-1950 period (green). Additional information about the interaction is displayed in each node. The click and download interactions include the document metadata values, the document title, and the page number of the results page of the click. The search interactions include the query, selected facets, and search options used. In addition to the visualizations, hovering over a node will display timestamps relative to the start of the session, and a link to the web page visited (see Fig. 7).

5.5 EVALUATION SETUP

In a small-scale experiment we evaluate MAGUS and compare it with a table visualization, Fig. 3. We recruited 12 participants (of which 5 men) among historians, computer scientists, library collection specialists and data scientists. We asked them to perform three tasks and measured the time spent, perceived workload, usability scores (widely used for user studies, [78]). In addition, we measured the certainty of and agreement among the answers given, and performed an analysis of their free-text answers. The experiments were performed on HTTP server logs from the National Library of the Netherlands[3]. The search platform provides access to historical newspaper documents using a faceted interface, with the facets based on the (bibliographic) metadata describing the documents within the collection (such as the publication date). We cleaned and split the logs into sessions as described in [13].

TASKS The study includes three tasks of increasing complexity. The sessions we selected to be visualized in the tasks all relate to one specific subject–WWII–in the sense that they contain queries and/or clicks on documents about topics related to WWII. This choice is inspired by an ongoing collaboration with the NIOD Institute for War, Holocaust and Genocide Studies[4].

TASK 1 IDENTIFY INFORMATION NEEDS: *Inspect a session and assess if one of the information needs of the user is to find documents about a topic directly related to WWII.* This task is relevant, for example, to historians who are interested in users searching for WWII-related documents, to understand how users search and which topics they search for. Such a task can also be relevant to manual label sessions for a training and test set. For example, [54] created

[3] Logs collected from the search platform `https://www.delpher.nl/`, access granted under a strict confidentiality agreement.

[4] `https://www.niod.nl/en`

such a training set for automatic segmentation of search topics in logs. Each participant performed this task 4 times (subtasks 1.1 - 1.4).

TASK 2 DISTINGUISH STRUGGLING FROM EXPLORING: *Inspect a session and assess whether the user was struggling or exploring.* This kind of task could be performed by a library professional who seeks to understand if users find what they are looking for in the library collection. It is also relevant when building a training set for a classifier, as was done by crowd workers in [40]. Disambiguation between struggling and exploring sessions is important both for understanding search success and when providing real-time user support [40]. Participants performed this task 4 times (subtasks 2.1-2.4).

TASK 3 DESCRIBE A CLUSTER OF SESSIONS: *Provide fitting labels and descriptions for four clusters of sessions, by inspecting four sessions per cluster.* In this task, we study to what extent inspection of a few (in this case four) individual sessions allows a professional to see shared, high-level usage patterns and distinguish different types of uses.

For tasks 1 and 2, we manually selected sessions that we judged to be suitable for the tasks and that demonstrate a user interest in WWII topics, based on a list of WWII-related terms provided by the NIOD. For task 3, we clustered sessions including WWII topics using a k-medoids algorithm as described in [12]. This resulted in four distinct clusters. Table 5 provides median values of the clustering features, serving as a high-level overview of the sessions in each cluster. Cluster 1 contains sessions with mainly clicked documents and little search interactions; cluster 2 sessions with clicked documents followed by downloads; cluster 3 sessions with faceted search, focusing on the 1930-49 period; and cluster 4 faceted search with the focus outside the 1930-49 period. In task 3, participants of the study were not shown the session statistics, but were presented with the four most typical sessions of each cluster, i.e. the sessions with the shortest Manhattan distance to the set of medians of the session features in a cluster.

TWO VISUALIZATIONS We use a within-subjects design where each participant is exposed to both visualizations. We always present tasks and sessions in the same order. However, we present the visualizations in different orders to avoid measuring a learning effect for either visualization. One group uses MAGUS for subtasks 1.1 and 1.2, the table for subtasks 1.3, 1.4, 2.1 and 2.2, and then MAGUS for subtasks 2.3, 2.4 and task 3, the other group swaps the visualization tools. Participants are randomly spread over the groups.

PROCEDURE, DATA COLLECTION AND DATA PREPARATION First, each participant receives a short training in the use of both visualizations. Then, the participant performs the three tasks. Finally, the participant fills out the System

Table 5: Median values of all clustering features for the four clusters.

					search	search	search		
				search	WWII	1930-49	time	clicks	clicks
cluster	clicks	downloads	search	facets	facets	facets	ranking	WWII	1930-49
1	**88%**	0%	11%	0%	0%	0%	0%	1%	20%
2	33%	**64%**	0%	0%	0%	0%	0%	0%	0%
3	22%	0%	**76%**	**46%**	0%	**11%**	0%	**50%**	**91%**
4	23%	0%	**74%**	**44%**	0%	0%	6%	3%	18%

Usability Scale (SUS) questionnaire [17] for both visualizations, and provides further written comments on the use of both visualizations. For the sessions in tasks 1 and 2, the participants select an answer (yes/no on task 1; struggling/-exploring on task 2) and provide a free-text justification of their answer. For the clusters in task 3, they provide a free-text label and description. After each session or cluster, we ask participants to assign a measure of their certainty on a five-point Likert scale. After each task and for each visualization method, the participants fill out a NASA TLX questionnaire [39]. All tasks were timed.

We manually annotate the free-text answers to record whether the participants' arguments contain one or more of eight categories of information about a session: (1) queries (for example, a participant writes "hitler as search term"); (2) clicks (for example, "left [...] without clicking"); (3) downloads ("the user didn't download"); (4) links ("possibly saved links"); (5) specific content or metadata values in documents or search facets ("all post-war phenomena" or "time range around ww2 (30-49)"); (6) search behavior ("doesn't use facets", or "click through the results"); (7) blacklist notice, a warning page shown before accessing Nazi-propaganda ("he/she clicked on the blacklist consent"); (8) time ("he/she spent not too much time"). Subjective arguments are left out, such as "he/she seems knowledgeable", "I wonder if they can find it", "couldn't find what he/she was looking for", or "feels more frustrated".

5.6 EVALUATION RESULTS

FREE-TEXT ANSWERS We analyze the manually annotated free-text answers by counting how many times each argument-category was used by participants. Table 6 shows the number of arguments in total and of each category separately, for the three tasks as well as overall. It also lists the mean word count of the free-text answers. We notice that only slightly more arguments were used with MAGUS than with the table visualization (187 for MAGUS vs. 171 for the table), and on average the same number of words (20). Only in task 2 participants clearly use more arguments when using MAGUS. However, the type of arguments used is different between the two visualizations. When using the table, participants use the query more frequently as an argument (53 times with MAGUS vs. 69 with the table). With MAGUS, the focus is more strongly on specific content and metadata (41 times with MAGUS vs. 22 with the table), and on search behavior (36 vs. 28). This suggests that MAGUS indeed focuses participants' attention not only on the query but also on other aspects present in the sessions, such as the metadata and the search techniques used.

The free-text cluster descriptions given by participants in task 3, show a difference between MAGUS and the table. As discussed in Section 5.5, cluster 3 focuses on WWII, while cluster 4 does not. Five out of six participants who used MAGUS for task 3 mention this in their description of cluster 3 and/or cluster 4. Only one of the participants that used the table does, labeling cluster 4

Table 6: Argument analysis of participants' free-text explanations and descriptions.

		mean						spec.	search		
	visuali-	word	arg.					content/	techn./		
task	zation	count	count	query	click	downl.	link	metadata	behavior	blackl.	time
1	MAGUS	14	57	25	3	0	0	25	1	3	0
	table	13	54	37	0	0	0	16	1	0	0
2	MAGUS	26	75	26	16	6	2	6	16	2	1
	table	25	59	25	12	4	0	2	12	0	4
3	MAGUS	20	55	2	12	7	5	10	19	0	0
	table	22	58	7	16	10	6	4	15	0	0
combi	MAGUS	20	**187**	53	31	13	7	**41**	**36**	5	1
	table	20	171	**69**	28	14	6	22	28	0	4

as "advanced search after WWII". This demonstrates how MAGUS can improve the quality of answers for tasks where it is important to understand how users search in different parts of the collection.

AGREEMENT BETWEEN THE PARTICIPANTS For tasks 1 and 2, we do not consider answers as correct or incorrect, but rather check whether participants agreed on their answers. The number of participants that agreed with each other is exactly the same among participants that used MAGUS and among those that used the table, showing that the visualization method does not impact the agreement. Agreement is different for the different tasks, with almost perfect agreement on task 1 and moderate disagreement on task 2.

CERTAINTY OF THE ANSWERS We find no differences between MAGUS and the table with respect to how certain participants are of their answers (Fig. 8).

TIME SPENT The participants need, on average, more time when using MAGUS than when using the table for task 1 and especially task 2. There is no clear difference on task 3. The observed difference in time spent between the two visualizations is small compared with the variation among participants and the difference between tasks, with task 3 requiring considerably more time. (Fig. 9).

WORKLOAD Table 7 presents the perceived workload for both session visualizations. Workload is measured through the NASA TLX questionnaire on six dimensions. For task 1, the perceived workload is lower for MAGUS than for the table on all dimensions. For task 2, on the other hand, all workload dimensions are scored slightly higher for MAGUS, and for task 3 the workload is even considerably higher for MAGUS. However, again, standard deviations are high on all questions; variation among participants is generally higher than the difference between the table and MAGUS.

USABILITY In terms of the reported usability (Fig. 10), the differences are small. MAGUS is liked a bit more than the table. Some participants find the table cumbersome. On the other hand, the participants feel that MAGUS is a bit more difficult to use, as can be seen from the slightly better scores of the table visualization on complexity, ease of use, and the need for support. While the majority of participants reported that there was little need to learn how to use the two visualizations, multiple participants comment on this. For example, participant 1, an information professional, writes: "You need to learn how to read a graph and understand what is happening in it. But if you inspect it (more) carefully with a legend, then it provides a wealth of information!" We find no conclusive differences with respect to the usability aspects ‘well integrated”, “inconsistant”, ”understand quickly” and “felt confident.”

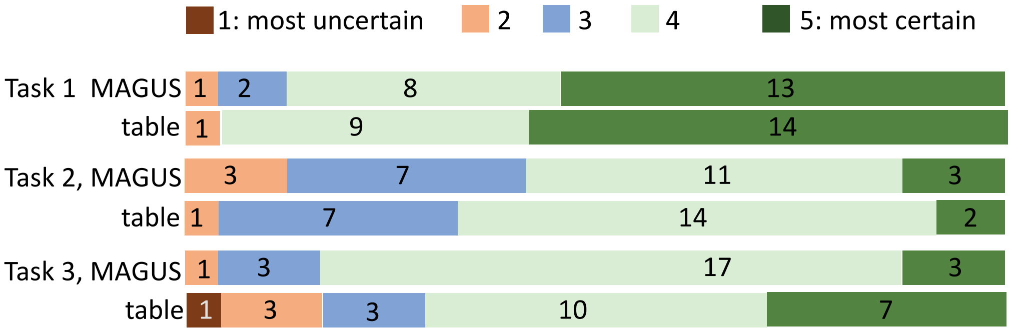


Figure 8: Certainty: number of times each point on a Likert scale from uncertain to certain was selected.

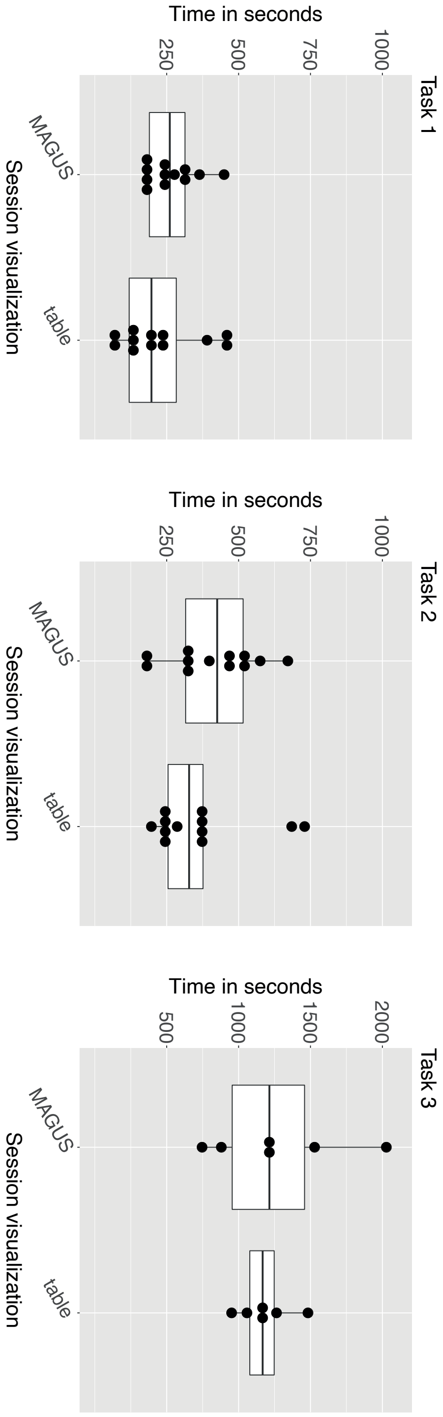


Figure 9: Time spent per task. Dots represent participants. (Different scale on Task 3.)

Table 7: Perceived workload measure, on a scale from 0 to 100, lower is better.

	task 1				task 2				task 3			
	MAGUS		table		MAGUS		table		MAGUS		table	
	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
Mental demand	20	37	34	36	51	17	47	22	67	14	52	21
Physical demand	8	9	10	18	11	6	10	5	19	23	13	6
Temporal demand	15	19	21	26	31	18	26	16	41	24	32	23
Performance	20	30	27	28	45	17	40	18	53	12	40	13
Effort	17	32	31	33	43	16	41	18	62	13	34	14
Frustration	14	23	14	18	27	23	23	18	27	18	26	10

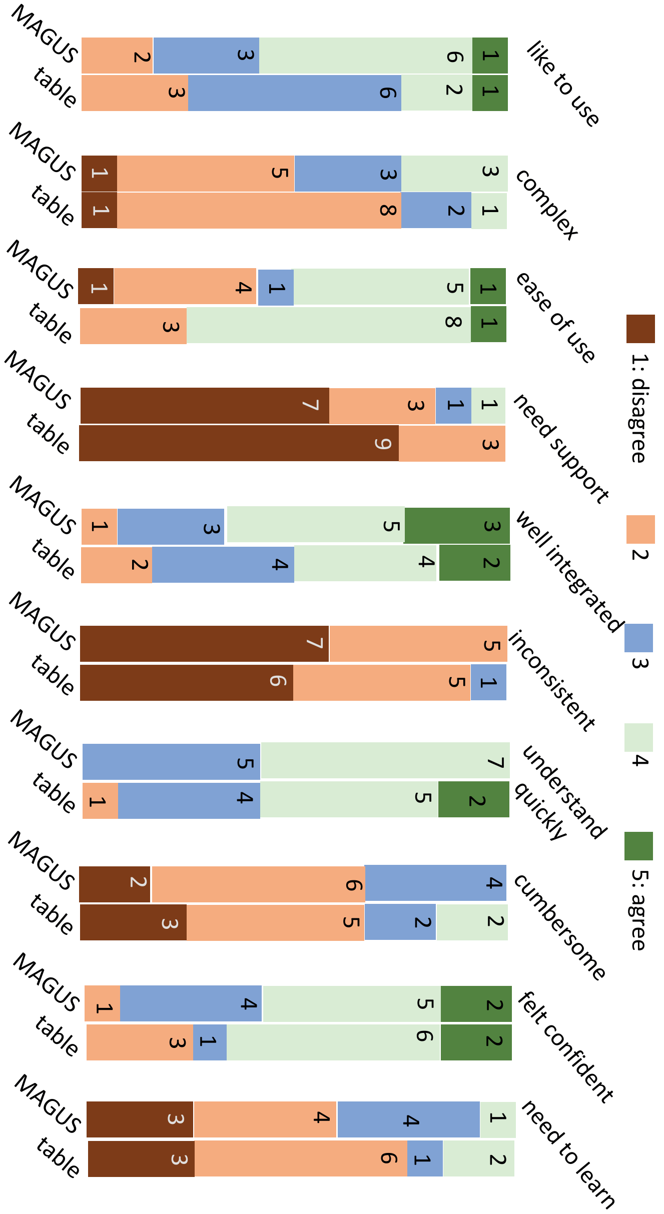


Figure 10: Usability: number of times each point on a Likert scale from uncertain to certain was selected in the System Usability Scale (SUS) questionnaire.

5.7 CONCLUSION

We have developed MAGUS, a tool for visualizing individual user sessions. MAGUS visualizes a user's navigational path as a directed graph, mapping repeated interactions onto a single node. Our tool highlights the different types of user interactions such as searches, clicks and downloads, the use of search facets, and relevant metadata of the clicked documents. In this way, MAGUS allows researchers and library professionals to recognize different interaction patterns and provides insights into the parts of the collection a user is interested in.

We have evaluated our tool on three tasks performed by 12 professionals in a comparison with a standard table visualization. An analysis of the free-text answers demonstrated that MAGUS indeed enabled participants to identify the part of the collection a user is interested in, and that it helps to distinguish different types of search behavior. Further empirical research into specific aspects of the session visualization separately, such as the metadata coloring, could provide more insights into the benefits of each aspect. The results of the workload questionnaire and activity logging suggest that participants find MAGUS more difficult to use than the table, even though the participants do like the tool. MAGUS may be perceived as more difficult due to the steeper learning curve associated with our tool, and it would be interesting to do a follow-up study to confirm this. A larger follow-up study could also include an investigation of the different professional backgrounds of participants, for example, to compare whether data professionals and domain experts use the tool differently. Furthermore, we would like to investigate which types of tasks specifically benefit from MAGUS, and for which types of sessions the tool works best, as several participants mentioned the benefit of MAGUS for long, complicated sessions.

Acknowledgements

We would like to thank the National Library of the Netherlands, the NIOD Institute for War, Holocaust and Genocide Studies, and BNNVARA (public broadcasting company) for their support.

6 CONCLUSIONS

In this thesis, we have reported on our research into how to use metadata to improve our understanding of search behavior in digital libraries. First, we investigated this research theme in three different settings. In setting 1, we analyzed search log data using specific metadata values which we defined in advance, in order to investigate search behavior in specific parts of the collection. In setting 2, we did not define the relevant metadata values in advance, so as to be able to identify in which parts of the collection users are showing an interest. In setting 3, we looked into how to proceed when existing metadata does not directly correlate with the topic for which we want to investigate the corresponding search behavior. And finally, we examined how to communicate results of these types of analyses to domain experts, collection owners and researchers, including both search behavior and relevant metadata.

The previous chapters described our research on each of these topics. In this chapter, we provide a concise, high level summary on the contributions and the results presented in the four research chapters. We will discuss our findings and provide an outlook on possible future directions of research could take.

6.1 MAIN FINDINGS

CHAPTER 2: METADATA CATEGORIZATION In Chapter 2, we leveraged a selected set of metadata values to analyze search behavior in different parts of a historical newspaper collection. The research question we addressed was the following:

(RQ) *How do search patterns differ among users searching in different parts of the collection?*

To identify these different parts, we analyzed search log data in combination with metadata records describing the contents of the collection. We have used both the metadata present in the search interactions in the form of selected filtering facets, and the metadata of the clicked documents. This metadata helped to create subsets corresponding to different parts of the collection.

We found that metadata in the form of filtering facets over the search results is used more often than not by users searching the historical newspaper collection. Furthermore, sessions that include facets are typically longer, and contain more clicks and downloads and more, unique, shorter keyword queries. And we observed distinct search patterns in different parts of the collection, with these patterns corresponding to historical periods, geographical regions or subject matter.

The contribution of this chapter is twofold. First, the chapter has provided detailed insights into user behavior in a historical newspaper collection, observing distinct search patterns within different parts of the collection. Based on these search patterns, we were 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, the study demonstrated how metadata can be used to analyze search behavior within specific parts of a digital library. As such, an analysis leveraging the metadata enables researchers to do a comparative analysis of (1) what users search for (using the selected facets in the search interactions), (2) what they find (using the metadata of clicked documents), and (3) what is or is not present in the collection (using the collection metadata).

CHAPTER 3: USER INTERESTS In this chapter, we investigated user interests within the collection. To do this, we did not define in advance which metadata values were relevant. The research questions we have addressed in this chapter are the following two:

(RQ1) *What are the user interests in terms of the different parts of a collection? How can we detect these?*

(RQ2) *What is the related search behavior within these parts?*

To detect user interests within different parts of the collection we applied a clustering algorithm to partition the search sessions based on the metadata of search interactions and clicked documents. This helped us to identify user interests within the collection; investigate the relation between them; and analyze the related search behavior within the different parts of the historical newspaper collection users showed an interest in. Examples of user interests we detected corresponded to specific types of news items, such as family announcements (relating to births, marriages, and deaths), specific periods, such as 1930-1949 (including the Great Depression and World War II), or specific regions, such as Suriname (one of the former Dutch colonies). To evaluate the clusters resulting from the clustering algorithm, we measured the stability of these clusters over a six-month period. The results showed that detected user interests are stable, with the same user interests reoccurring over the six-month period in the analysis. We found that related search behavior varied within the different parts of the collection, with users spending little time and few search techniques in some parts of the collection, in other parts using a wide variety of search techniques and spending a lot of time. As a result, this approach also facilitates the detection and investigation of highly-focused user groups using many search techniques and spending a long time. In addition, it can help to inform the design of more targeted user interfaces, or to improve search systems or collection management.

Our contribution to the research field is the demonstration that a partitioning of sessions based on the metadata of a collection and an investigation of the related search behavior reveals specific user needs in specific parts of a collection, where in an overall analysis these patterns would disappear.

CHAPTER 4: TOPICS For the next chapter, we have explored how to proceed in a setting when the specific topic for which we want to investigate the user interest and related search behavior is not captured in the available metadata. In this exploration we have addressed the following two research questions:

(**RQ1**) *How can we represent a specified topic?*

(**RQ2**) *How can we use the topic representation to retrieve relevant sessions?*

The first of these research questions was answered by looking into different, consecutive ways we can use to build a term list as our topic representation. We investigated term lists created (i) with the use of semantic relations in an explicit knowledge resource, (ii) with the application of local word embeddings trained on the documents in the collection, and (iii) by manual curation of the term lists in each step, with the help of domain experts with curating the terms[1]. To answer the second research question we looked into how to match the different term lists to user sessions. We matched the terms in the lists to either a) the user queries, or b) the contents of the clicked documents. We compared and discussed the combined methods in terms of number of retrieved sessions as well as estimated precision scores. These estimated precision scores were computed using manually created ground truths of over 600 sessions per topic, using the MAGUS visualization tool to assess the sessions.

The results of the exploration showed that the precision is high when retrieving sessions based solely on user queries, however, the number of sessions retrieved remains small. When the document-based matching approach was used, the number of sessions retrieved increases, but at the expense of precision. Additionally, we found that manually curating the term lists improved precision while still preserving a larger number of sessions, and without this step the expanded lists (using a knowledge resource or local word embeddings) perform poorly in terms of precision. This effect was particular strong for the document-based matching. We investigated two topics, WWII and feminism, and both showed similar general patterns in their results, yet we observed a higher overall precision for the more popular topic (WWII).

This study provided insights into how different topic representations and matching approaches perform when retrieving topic-specific sessions. It demonstrates how different methods can be applied and combined by digital humanities scholars and practitioners to retrieve topic-specific sessions. Understanding search behavior for topics with societal relevance can provide insights into the interest in these topics. We believe this research helps to pave the way for a better understanding and communication of topic-specific user interests within

[1] The assessed term lists for the topics are available online `https://edu.nl/4arxw` and `https://edu.nl/9qbfr`

collections for digital humanities scholars as well as collection owners and practitioners.

CHAPTER 5: SESSION VISUALIZATION In the final research chapter, we have examined how to communicate the results of these types of analyses, where the metadata is explicitly included, to domain experts, collection owners, and researchers. We have developed MAGUS, a session visualization technique combining graphs to visualize search behavior and colors corresponding to the relevant metadata values of the search[2]. Our tool highlights different types of user interactions such as searches, clicks and downloads; and the relevant metadata of search facets and clicked documents. In this way, MAGUS allows researchers and library professionals to recognize different interaction patterns and at the same time it provides insights into the parts of the collection a user is interested in. To evaluate our tool, we conducted a user study, where we have addressed the following research questions:

(RQ1) *Is session inspection easier in terms of time and effort spent when using MAGUS?*

(RQ2) *Are the answers provided better in terms of accuracy and level of detail when using MAGUS?*

We have evaluated MAGUS in a comparison with a table representation in three representative tasks completed by 12 participants from diverse professional backgrounds. For the evaluation we used a mixed method approach: we timed the users in their tasks, we analyzed the answers given during the tasks, and we added both a usability and a perceived workload questionnaire, as well as leaving space for the participants to comment on the tools. We analyzed the answers given to the tasks and these demonstrated that MAGUS indeed enabled participants to identify the part of the collection a user is interested in, and that it helps to distinguish different types of search behavior.

6.2 DISCUSSION AND FUTURE WORK

With this thesis we provide insights into how we can leverage descriptive metadata to better understand search behavior within different parts of a digital library collection. In this section, we will discuss some limitations and, where relevant, how possible directions of future work may help to resolve them.

DEPENDENCY ON A SINGLE DATA SET We have conducted our research with the use of a single combined data set from a single digital library. This is a limitation of the work, as it makes it harder to judge the generalizability of our findings. But data sets combining both search log data and metadata records and the contents of the searched collection are hard to obtain. And we are able

[2] Demo and source code for MAGUS available at `https://swish.swi-prolog.org/p/magus.swinb`

to make assumptions about the generalizability of the research. We fully expect that the same techniques can be used for other collections with professionally curated metadata, particularly if the available metadata is of a high standard. In order to prove this assumption, we would need to repeat our research with data from a different digital library, from which we would need both search log data, and metadata records and the contents of the collection. In such a case, this type of approach could be a starting point for inter-collection comparison of search behavior within different parts of the collections among digital libraries sharing similar metadata categories. Alternatively, we could compare user interests among different collections, either in a detection of the most popular user interests, or by looking into specific user interests within collections. For example, we could compare search for specific topics, such as WWII, between the collection maintained by the NIOD institute and the historical newspaper collection maintained by the National Library of the Netherlands. A limitation of a comparison is of course the existence and comparability of the metadata of the collections. Naturally, different vertical search systems will have different search functionalities. This notwithstanding, a commonality among digital libraries is the functionality to access the collection using metadata describing the contents. With this type of search functionality and metadata describing the contents of the collection, we can apply the methods described in this thesis. We can then analyze how search patterns differ in the different parts of the collection, after we select metadata values of interest. And we can investigate the user interests within the collection, based on the metadata available. We can analyze search for specific topics. And we can use the MAGUS session visualization technique, after adapting the tool to the specific search system.

DATA PREPARATION AND AVAILABILITY It is complex to go from HTTP logs to information about user behavior. For one, it is not simple to identify sessions from search log data. Moreover, the search system influences both the structure and contents of the log data and corresponding search behavior. How to best approach the preparation of the search logs for an analysis of search behavior is dependent on the type of questions one wants to answer in the research. The chosen approach also directly impacts the possibilities to compare the results with other studies.

To identify sessions from search log data, we have to decide on how to define a session. Researchers have chosen different definitions for a session, for example, based on session cookies, on overlapping queries, or on a timeout. This choice also depends on the type of research to perform, and no session definition will suit all types of research. In this thesis, we have defined sessions using a clickstream model, following the flow of connected interactions. A clickstream model can help to split possible multiple users behind a single IP address, and to find complete searches over longer time periods. We have selected this approach for a few reasons. First, the historical newspaper collection is accessible without login and the server does not log cookies, making an approach using session cookies not possible. Second, as our focus is not on the query, the query would not be an obvious choice for our session definition. Third, we expect a possible large proportion of users to be engaged in exploratory, open-ended search, and thus a timeout would result in breaking up visits that occur with long pauses. Using the clickstream model, we observed users searching in the collection for a single topic over long periods with long pauses in between, and identifying these users was valuable to the National Library. Nevertheless, a clickstream-based approach may lead to shorter or to longer sessions than when a timeout is used. For example, when a user continues their search in a new tab, this breaks off a clickstream-based session; or, when a user continues the next day with their search, this leads to a break in a timeout-based session.

The search system influences both the search log data and corresponding search behavior. A different search interface will naturally lead to differences in behavior. For example, the way facets are or are not employed in the search interface will affect the possible interactions with the system, and thus user behavior. In addition, the search log data available for analysis is influenced by how the system logs user interactions on the server, for example by what is and what is not logged. These design choices of the search system under investigation impacts the possibilities of the research.

Our tool MAGUS is also affected by this. Building a session graph is impacted by the way the log records are recorded. As there are many different ways to build a search interface and log the behavior within the search system, it is difficult to develop a software package to map any set of log records to MAGUS' graph visualization. This influenced our decision to develop MAGUS specifically for the logs we received from the National Library, and not as a ready-to-use software package. It would be possible to do this, of course, but to do so was outside the scope of this thesis.

CLUSTERING In Chapter 3 we clustered sessions based on their metadata values in the clicked documents and selected facets, and we then analyzed the corresponding search interactions within the metadata-based clusters. We believe there is still more to investigate with respect to the clustering of sessions.

It would be interesting to cluster on the interactions in the sessions as well, using features describing search behavior, such as session duration or number of clicks. These "behavior" clusters can then be mapped to the identified user interests, instead of the simple statistical summary, making it possible to find more than a single search pattern for each user interest. Or we could do this for the complete set of sessions and map the "behavior" clusters to the metadata clusters, creating a double clustering and finding the relations between these clusters. However, a first attempt to do just this using the same clustering method with interaction features did not lead to more detailed insights into the behavior within the metadata clusters than the statistical summary did. Another research approach could be to apply a fuzzy clustering. Even though a majority of sessions within the clusters are highly focused, we also find, in all the clusters, sessions on the edges of their clusters. With the k-medoids algorithm, such sessions are assigned to a single cluster. However, these sessions are more mixed with respect to the user interests, and they could have matched to more than one metadata cluster. Thus, assigning a value for how well they match to these metadata clusters can give better overall results than simply assigning these sessions to a single cluster. For example, in the *1900-29* cluster some of the sessions also include a minority of clicks from between 1930-49 (Table 7, Chapter 3). A clustering algorithm that allows these sessions on the 'edges' to belong to more than one cluster would be an interesting step in future research.

user studies To evaluate our session visualization tool MAGUS, we conducted a user study among domain experts and data professionals. In addition to this study, we would also like to know what the differences and similarities are between the different groups in the original user study. However, we need a higher number of participants from each of the groups separately to be able to draw conclusions about differences and similarities between the groups. A larger follow-up study could include an investigation of the different backgrounds of participants, for example, in a comparison between data professionals and domain experts. In addition, as several participants mentioned the benefit of MAGUS for long, complicated sessions, we would like to study for which types of sessions MAGUS performs better. Similarly, we would like to look into which types of tasks can benefit from MAGUS to what degree.

Apart from conducting a user study among the potential users of the techniques presented and discussed here, a study among the users of the digital library can provide insights into the motivation behind search behavior within a collection. We already mentioned some limitations of log analysis. An additional limitation of research based on log data is that it mainly studies how people search, and not why they do so. In future research we could complement this by conducting both qualitative and quantitative user studies among the users searching within different parts of a collection in a combination of log analysis, thus providing a why to the how.

"a wealth of information" Leveraging metadata in the analysis enhanced our understanding of how users are searching within the different parts of a digital library. We were able to provide collection owners recommendations about how to improve access to the collection. For example, by using a timeline representation of the search results in a historical newspaper collection. At the same time, domain experts gained more insights into what kind of topics people are searching for in a collection, whether these user interests were defined in advance (such as search for WWII) or not. In addition, our session graph visualization offered information professionals "a wealth of information", as one of the participants stated.

Vertical search engines and its use are ubiquitous and the search interfaces providing access to these systems are complex. As such, leveraging metadata in an analysis of the use of these systems will help to continue to improve both our understanding of search and to improve the design of these systems. An added benefit is the fact that using metadata instead of user queries for an analysis contributes to a better balance between privacy and the knowledge needed about users searching to better support them. However, there are potential situations where query-based analysis could have an advantage over a metadata-based approach. Clearly, if a vertical search system has low-quality metadata, leveraging this metadata would not be the best approach. But also, in a vertical search system where the metadata is of high-quality, imagine those sessions where the metadata is sparse and/or widely varied. For example, sessions where no facets have been selected and the metadata of the clicked documents – if there are any – is diverse. In these cases, a query-based analysis might be able to find a better categorization of such sessions than a metadata-based approach could. Possibly a form of query expansion or topic modeling over the queries in the sessions and/or over the clicked documents can find a better categorization for these sessions than the metadata-based approach would be able to. As such, this could be an interesting direction of future research, and – in a sense – the fourth analytical setting: a setting where we do not necessarily have metadata readily available (which we had in Chapters 2 and 3); nor have predefined the parts of the collection or topics of interest to study (which we had in Chapters 2 and 4).

Be that as it may, with this thesis we have shown that it is possible to understand search behavior without using queries directly, but by using metadata instead.

SUMMARY

Search log analysis is an unobtrusive technique used to better understand online search behavior. In this thesis, we study search in "vertical" search engines that provide access to curated collections. In this context, other data is available in addition to search logs: the documents in the collection, categorized with professionally curated metadata. This metadata is often reflected in the search interface in the form of facets, acting as a filter over the results. Our research focuses on how to leverage this metadata to improve our understanding of search behavior. To do this, we use both the metadata present in the selected facets and the metadata of the clicked documents, combining search logs with metadata records and the contents of the collection. First, we investigate how to leverage this metadata in three different analytical settings. After that, we examine how to communicate results of such an analysis including search behavior and relevant metadata. The research is conducted using data from the National Library of the Netherlands, a typical digital library with a richly annotated historical newspaper collection and a faceted search interface.

In the first setting (Chapter 2), we analyze the search logs using metadata values defined in advance, in order to study search behavior within specified parts of the collection that we have selected to be relevant. The analysis shows that faceted search is common, that sessions including facets are typically longer, contain more clicks and downloads and more unique queries. We observe distinct search patterns in different parts of the collection, and we are able to formulate concrete suggestions for improvement of the search system and collection management, showing how metadata can be used to analyze search behavior in specific parts of a collection.

In the second setting (Chapter 3), the goal is to detect user interests within the collection without defining metadata values in advance. We apply a clustering algorithm grouping sessions based on the metadata of selected facets and clicked documents. To evaluate resulting clusters, their stability over a six-month period is measured. The results show that user interests are stable, with the same interests reoccurring. The related search behavior varies per cluster, with users spending little time and few search techniques in some clusters, in others using a wide variety of search techniques and spending a lot of time. This demonstrates that a partitioning of sessions based on metadata, and an investigation of the related search behavior reveals specific user needs in specific parts of a collection, where in an overall analysis these patterns would disappear.

In the third setting (Chapter 4) we explore how to identify search for specific topics when no metadata directly describes these. We look into different, consecutive ways to build a term list as a topic representation: (i) using a knowledge resource, (ii) using local word embeddings trained on the collection, and (iii) by manual curation. Then we look into how to match the different term lists to search sessions: matching the terms to a) user queries, or b) clicked documents. We investigate two topics of societal relevance, WWII and feminism, and compare and discuss the combined methods in terms of number of retrieved sessions as well as estimated precision scores computed using manually created ground truths. With this work we provide insights into how different topic representations and matching approaches perform when retrieving topic-specific sessions.

Finally, we examine how to communicate the results of these types of analyses (Chapter 5). We introduce MAGUS, a session visualization tool combining graphs to visualize search behavior with colors to visualize relevant metadata. Our design is new in combining both search interactions and metadata in a single visualization, allowing researchers and professionals to recognize different interaction patterns while at the same time providing insights into the parts of the collection a user is interested in. For the evaluation we conduct a user study comparing MAGUS with a table representation in three tasks completed by 12 participants from diverse backgrounds. In the study we use mixed methods, combining quantitative and qualitative measures, such as timing the users, standardized questionnaires, as well as analyzing comments and written explanations given during the tasks. Our study demonstrates that MAGUS enables participants to identify the part of the collection a user is interested in, and that it helps to distinguish different types of search behavior.

We expect that the presented methods can be used for other collections. Vertical search engines are ubiquitous and the search interfaces providing access to these systems are complex. We have shown how leveraging metadata in an analysis of search behavior can enhance our understanding of how users are searching within the different parts of a digital library, and we were able to provide collection owners with recommendations about how to improve access to the collection. An added benefit is the fact that using metadata instead of user queries for an analysis contributes to a better balance between privacy and the knowledge needed about users to better support them. With this thesis we show that it is possible to understand search behavior without using queries directly, but by using metadata instead.

SAMENVATTING

Een analyse van de gelogde interacties met de zoekmachine is een discrete techniek die gebruikt kan worden om online zoekgedrag beter te begrijpen. In dit proefschift bestuderen wij zoekgedrag in 'verticale' zoekmachines die toegang geven tot gecureerde collecties. In deze context is andere data beschikbaar naast de logs: de documenten in de collectie gecategoriseerd met professioneel gecureerde metadata. Deze metadata is vaak beschikbaar in de zoekinterface als facetten die de resultaten filteren. Onze research focust op hoe we deze metadata kunnen gebruiken om het zoekgedrag beter te begrijpen. We gebruiken zowel de metadata van de door de gebruiker geselecteerde filterfacetten, als de metadata van de documenten waarop geklikt is, waarbij we de zoeklogs, metadata en documenten van de collectie combineren. Eerst onderzoeken we hoe we deze metadata kunnen aanwenden in drie verschillende analytische scenario's. Daarna bekijken we hoe we de resultaten van een dergelijke analyse kunnen delen, waarbij we én het zoekgedrag én de relevante metadata betrekken. Voor deze research hebben we data van de Koninklijke Bibliotheek gebruikt, een typerende digitale bibliotheek met een rijk geannoteerde historische krantencollectie en een zoekinterface met filterfacetten.

In het eerste scenario (Hoofdstuk 2) analyseren we de zoeklogs waarbij we van te voren bepalen welke metadata we bekijken, zodat we zoekgedrag in specifieke, door ons als relevant bestempelde delen van de collectie kunnen bestuderen. Onze analyse toont aan dat gebruikers vaak facetten selecteren, en dat sessies met facetten in het algemeen langer zijn met meer kliks en downloads en meer unieke zoekopdrachten. We zien specifieke zoekpatronen in verschillende delen van de collectie, en we zijn in staat om concrete aanbevelingen te geven om het zoeksysteem en collectiebeheer te verbeteren. Dit toont aan hoe metadata gebruikt kan worden om zoekgedrag in specifieke delen van de collectie te analyseren. In het tweede scenario (Hoofdstuk 3) willen we gebruikersinteresses binnen de collectie ontdekken zonder specifieke metadata van te voren te definiëren. Met een clusteringsalgoritme groeperen we sessies gebaseerd op de metadata van geselecteerde facetten en geklikte documenten. Om de resulterende clusters te evalueren, meten we hun stabiliteit over een periode van zes maanden. De resultaten laten zien dat de gebruikersinteresses stabiel zijn, met maandelijks terugkerende interesses. Het gerelateerde zoekgedrag verschilt tussen de clusters. In sommige clusters besteden de gebruikers weinig tijd en passen ze weinig zoektechnieken toe; in andere clusters passen ze een verscheidenheid aan zoektechnieken toe en besteden ze veel tijd. Dit demonstreert dat het clusteren van sessies gebaseerd op de metadata, in combinatie met een inspectie van het gerelateerde zoekgedrag, specifieke gebruikersbehoeftes onthult die in een analyse over alle sessies uit het zicht zou verdwijnen. In het derde scenario (Hoofdstuk 4) exploreren we hoe we gebruikers kunnen identificeren die zoeken naar specifieke onderwerpen waarbij we geen direct passende metadata hebben. We bekijken verschillende, opeenvolgende manieren om een termenlijst te creëren die het onderwerp representeert, gebruikmakend van: (i) een kennisbron, (ii) 'local word embeddings' getraind op de collectie, en (iii) een handmatige beoordeling. Daarna bekijken we hoe we deze verschillende termenlijsten kunnen koppelen aan de sessies, via een match met a) de zoekopdrachten, of met b) de geklikte documenten. We onderzoeken twee maatschappelijk relevante onderwerpen, WOII en feminisme, en vergelijken en bespreken de gecombineerde methodes in termen van aantallen opgehaalde sessies en geschatte precisiescores berekend op basis van handmatig geannoteerde sessies. Met deze studie geven we inzicht in hoe de verschillende manieren om het onderwerp te representeren en de verschillende matchingsmethodes presteren bij het vinden van sessies over specifieke onderwerpen.

Tenslotte onderzoeken we hoe we de resultaten van dit soort analyses kunnen overbrengen (Hoofdstuk 5). We introduceren MAGUS, een visualisatie tool die een sessie als graaf verbeeld, waarbij de vorm het zoekgedrag en kleuren de metadata visualiseren. Ons ontwerp is nieuw in het combineren van zowel het zoekgedrag als de metadata in één visualisatie. Hiermee kunnen onderzoekers en professionals verschillende interactiepatronen herkennen en tegelijkertijd ook inzicht verkrijgen over de delen van de collectie waarin de gebruiker geïnteresseerd is. We evalueren MAGUS in een gebruikersstudie met 12 participanten met diverse achtergronden die drie taken uitvoeren, waar we onze visualisatie techniek vergelijken met een tabelrepresentatie van een sessie. We gebruiken verschillende onderzoeksmethodes, kwantitatief en kwalitatief: het meten van de benodigde tijd, gestandaardiseerde vragenlijsten en een analyse van de commentaren en uitgeschreven toelichtingen bij de opdrachten. De gebruikersstudie laat zien hoe MAGUS het de participanten mogelijk maakt om de delen van de collectie te identificeren waarin de gebruikers geïnteresseerd zijn, en om de verschillende soorten zoekgedrag te herkennen.

We verwachten dat de gepresenteerde methodes gebruikt kunnen worden voor andere collecties. Er zijn veel verticale zoekmachines en de interfaces die toegang bieden tot dit soort collecties zijn complex. We hebben laten zien hoe de metadata in een analyse van zoekgedrag ons begrip en onze kennis van hoe gebruikers zoeken in verschillende delen van een digitale bibliotheek kunnen verbeteren. Daarnaast konden we aanbevelingen geven aan de collectiebeheerders over hoe de toegankelijkheid van hun collectie verbeterd kan worden. Een bijkomend voordeel is dat het gebruik van metadata in de analyse in plaats van zoekopdrachten een bijdrage levert aan een betere balans tussen privacy en de kennis die nodig is om de gebruikers betere ondersteuning te verlenen. In dit proefschrift demonstreren we hoe het mogelijk is om zoekgedrag te begrijpen zonder de zoekopdrachten maar in plaats hiervan de metadata te gebruiken.

PUBLICATIONS

[1] T. Bogaard, J. Wielemaker, L. Hollink, and J. van Ossenbruggen. „SWISH DataLab: A web interface for data exploration and analysis." In: *BNAIC 2016: Artificial Intelligence: 28th Benelux Conference on Artificial Intelligence, Amsterdam, The Netherlands, November 10-11, 2016, Revised Selected Papers*. Ed. by Tibor Bosse and Bert Bredeweg. Vol. 765. Cham: Springer, 2017. Chap. 13, pp. 181–187. ISBN: 978-3-319-67468-1. DOI: `10.1007/978-3-319-67468-1{\_}13`.

[2] Tessel Bogaard. „On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives." In: *Proceedings of the 2018 Conference on Human Information Interaction &amp; Retrieval*. CHIIR '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 339–341. ISBN: 9781450349253. DOI: `10.1145/3176349.3176350`. URL: `https://doi.org/10.1145/3176349.3176350`.

[3] Tessel Bogaard, Aysenur Bilgin, Jan Wielemaker, Laura Hollink, Kees Ribbens, and Jacco van Ossenbruggen. „Comparing Methods for Finding Search Sessions on a Specified Topic: A Double Case Study." In: *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 189–201. ISBN: 978-3-030-86323-4. DOI: `10.1007/978-3-030-86324-1{\_}23`. URL: `https://doi.org/10.1007/978-3-030-86324-1_23`.

[4] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen. „Searching for Old News: User Interests and Behavior Within a National Collection." In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. CHIIR '19. New York, NY, USA: ACM, 2019, pp. 113–121. ISBN: 978-1-4503-6025-8. DOI: `10.1145/3295750.3298925`.

[5] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. „Metadata categorization for identifying search patterns in a digital library." In: *Journal of Documentation* 75.2 (2019), pp. 270–286. DOI: `10.1108/JD-06-2018-0087`.

[6] Tessel Bogaard, Jan Wielemaker, Laura Hollink, Lynda Hardman, and Jacco van Ossenbruggen. „Understanding User Behavior in Digital Libraries Using the MAGUS Session Visualization Tool." In: *Digital Libraries for Open Knowledge - 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, Proceedings*. Ed. by Mark M Hall, Tanja Mercun, Thomas Risse, and Fabien Duchateau. Vol. 12246.

Lecture Notes in Computer Science. Springer, 2020, pp. 171–184. DOI: 10.1007/978-3-030-54956-5{_}13. URL: https://doi.org/10.1007/978-3-030-54956-5_13.

BIBLIOGRAPHY

[1] Mohannad ALMasri, Catherine Berrut, and Jean-Pierre Chevallet. „Wikipedia-based Semantic Query Enrichment." In: *Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval*. ESAIR '13. New York, NY, USA: ACM, 2013, pp. 5–8. ISBN: 978-1-4503-2413-7. DOI: 10.1145/2513204.2513209. URL: http://doi.acm.org/10.1145/2513204.2513209.

[2] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. „On the surprising behavior of distance metrics in high dimensional space." In: *Database Theory – ICDT 2001* (2001). ISSN: 0956-7925. DOI: 10.1007/3-540-44503-X{_}27.

[3] Nitish Aggarwal and Paul Buitelaar. „Query Expansion Using Wikipedia and Dbpedia." In: *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012*. Ed. by Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker. Vol. 1178. CEUR Workshop Proceedings. CEUR-WS.org, 2012. URL: http://ceur-ws.org/Vol-1178/CLEF2012wn-CHiC-AggarwalEt2012.pdf.

[4] Eugene Agichtein, Ryen W White, Susan T Dumais, and Paul N Bennet. „Search, Interrupted: Understanding and Predicting Search Task Continuation." In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12. New York, NY, USA: ACM, 2012, pp. 315–324. ISBN: 978-1-4503-1472-5. DOI: 10.1145/2348283.2348328. URL: http://doi.acm.org/10.1145/2348283.2348328.

[5] Hiteshwar Kumar Azad and Akshay Deepak. „Query expansion techniques for information retrieval: A survey." In: *Information Processing and Management* (2019). ISSN: 03064573. DOI: 10.1016/j.ipm.2019.05.009.

[6] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. „Query Recommendation Using Query Logs in Search Engines." In: *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*. Ed. by Wolfgang Lindner, Marco Mesiti, Can Türker, Yannis Tzitzikas, and Athena I Vakali. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 588–596. ISBN: 978-3-540-30192-9. DOI: 10.1007/978-3-540-30192-9{_}58. URL: http://dx.doi.org/10.1007/978-3-540-30192-9_58http://link.springer.com/10.1007/978-3-540-30192-9_58.

[7] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. „Query Recommendation Using Query Logs in Search Engines." In: *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*. Ed. by Wolfgang Lindner, Marco Mesiti, Can Türker, Yannis Tzitzikas, and Athena I Vakali. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 588–596. ISBN: 978-3-540-30192-9. DOI: 10.1007/978-3-540-30192-9{_}58. URL: http://dx.doi.org/10.1007/978-3-540-30192-9_58.

[8] S Beitzel and M. „Hourly analysis of a very large topically categorized web query log." In: *In SIGIR'* 04pages (2004). Ed. by Kalervo Järvelin, James Allan, Peter Bruza, and Mark Sanderson, pp. 321–328. ISSN: 15322882. DOI: 10.1145/1008992.1009048. URL: http://doi.acm.org/10.1145/1008992.1009048.

[9] T. Bogaard, J. Wielemaker, L. Hollink, and J. van Ossenbruggen. „SWISH DataLab: A web interface for data exploration and analysis." In: *BNAIC 2016: Artificial Intelligence: 28th Benelux Conference on Artificial Intelligence, Amsterdam, The Netherlands, November 10-11, 2016, Revised Selected Papers*. Ed. by Tibor Bosse and Bert Bredeweg. Vol. 765. Cham: Springer, 2017. Chap. 13, pp. 181–187. ISBN: 978-3-319-67468-1. DOI: 10.1007/978-3-319-67468-1{_}13.

[10] Tessel Bogaard. „On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives." In: *Proceedings of the 2018 Conference on Human Information Interaction &amp; Retrieval*. CHIIR '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 339–341. ISBN: 9781450349253. DOI: 10.1145/3176349.3176350. URL: https://doi.org/10.1145/3176349.3176350.

[11] Tessel Bogaard, Aysenur Bilgin, Jan Wielemaker, Laura Hollink, Kees Ribbens, and Jacco van Ossenbruggen. „Comparing Methods for Finding Search Sessions on a Specified Topic: A Double Case Study." In: *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2021, pp. 189–201. ISBN: 978-3-030-86323-4. DOI: 10.1007/978-3-030-86324-1{_}23. URL: https://doi.org/10.1007/978-3-030-86324-1_23.

[12] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Lynda Hardman, and Jacco van Ossenbruggen. „Searching for Old News: User Interests and Behavior Within a National Collection." In: *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. CHIIR '19. New York, NY, USA: ACM, 2019, pp. 113–121. ISBN: 978-1-4503-6025-8. DOI: 10.1145/3295750.3298925.

[13] Tessel Bogaard, Laura Hollink, Jan Wielemaker, Jacco van Ossenbruggen, and Lynda Hardman. „Metadata categorization for identifying search patterns in a digital library." In: *Journal of Documentation* 75.2 (2019), pp. 270–286. DOI: 10.1108/JD-06-2018-0087.

[14] Tessel Bogaard, Jan Wielemaker, Laura Hollink, Lynda Hardman, and Jacco van Ossenbruggen. „Understanding User Behavior in Digital Libraries Using the MAGUS Session Visualization Tool." In: *Digital Libraries for Open Knowledge - 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25-27, 2020, Proceedings.* Ed. by Mark M Hall, Tanja Mercun, Thomas Risse, and Fabien Duchateau. Vol. 12246. Lecture Notes in Computer Science. Springer, 2020, pp. 171–184. DOI: 10.1007/978-3-030-54956-5{_}13. URL: https://doi.org/10.1007/978-3-030-54956-5_13.

[15] Christine L Borgman, Laura J Smart, Kelli A Millwood, Jason R Finley, Leslie Champeny, Anne J Gilliland, and Gregory H Leazer. „Comparing faculty information seeking in teaching and research: Implications for the design of digital libraries." In: *Journal of the American Society for Information Science and Technology* 56.6 (2005), pp. 636–657. DOI: 10.1002/asi.20154. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20154.

[16] Jeffrey Brainerd and Barry Becker. „Case Study: E-Commerce Clickstream Visualization." In: *IEEE Symposium on Information Visualization 2001 (INFOVIS'01), SanDiego, CA, USA, October 22-23, 2001.* Ed. by Keith Andrews, Steven F Roth, and Pak Chung Wong. IEEE Computer Society, 2001, pp. 153–156. ISBN: 0-7695-1342-5. DOI: 10.1109/INFVIS.2001.963293.

[17] John Brooke and others. „SUS-A quick and dirty usability scale." In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[18] Alison Callahan, Igor Pernek, Gregor Stiglic, Jure Leskovec, Howard R Strasberg, and Nigam Haresh Shah. „Analyzing Information Seeking and Drug-Safety Alert Response by Health Care Professionals as New Methods for Surveillance." In: *Journal of medical Internet research* 17.8 (Jan. 2015), e204. ISSN: 1438-8871. DOI: 10.2196/jmir.4427. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4642796&tool=pmcentrez&rendertype=abstract.

[19] Ewa S Callahan and Susan C Herring. „Cultural bias in Wikipedia content on famous persons." In: *Journal of the American Society for Information Science and Technology* 62.10 (2011), pp. 1899–1915. DOI: 10.1002/asi.21577. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21577.

[20] Lara D Catledge and James E Pitkow. „Characterizing browsing strategies in the World-Wide web." In: *Computer Networks and ISDN Systems* 27.6 (1995), pp. 1065 –1073. ISSN: 0169-7552. DOI: http://dx.doi.org/10.1016/

0169-7552(95)00043-7. URL: http://www.sciencedirect.com/science/article/pii/0169755295000437.

[21] Olivier Chapelle and Ya Zhang. „A Dynamic Bayesian Network Click Model for Web Search Ranking." In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. New York, NY, USA: ACM, 2009, pp. 1–10. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526711. URL: http://doi.acm.org/10.1145/1526709.1526711.

[22] Hui Min Chen and Michael D Cooper. „Using clustering techniques to detect usage patterns in a Web-based information system." In: *Journal of the American Society for Information Science and Technology* 52.11 (2001), pp. 888–904. ISSN: 15322882. DOI: 10.1002/asi.1159.

[23] Hui-Min Chen and Michael D. Cooper. „Stochastic modeling of usage patterns in a web-based information system." In: *Journal of the American Society for Information Science and Technology* (2002). ISSN: 1532-2882. DOI: 10.1002/asi.10076.

[24] Paul Clough, Timothy Hill, Monica Lestari Paramita, and Paula Goodale. „Europeana: What Users Search for and Why." In: *Research and Advanced Technology for Digital Libraries*. Ed. by Jaap Kamps, Giannis Tsakonas, Yannis Manolopoulos, Lazaros Iliadis, and Ioannis Karydis. Cham: Springer International Publishing, 2017, pp. 207–219. ISBN: 978-3-319-67008-9.

[25] Jacob Cohen. „A Coefficient of Agreement for Nominal Scales." In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104. URL: https://doi.org/10.1177/001316446002000104.

[26] Alissa Cooper. „A Survey of Query Log Privacy-enhancing Techniques from a Policy Perspective." In: *ACM Trans. Web* 2.4 (Oct. 2008), 19:1–19:27. ISSN: 1559-1131. DOI: 10.1145/1409220.1409222. URL: http://doi.acm.org/10.1145/1409220.1409222.

[27] Paul Darby and Paul D Clough. „Investigating the information-seeking behaviour of genealogists and family historians." In: *J. Information Science* 39.1 (2013), pp. 73–84. DOI: 10.1177/0165551512469765. URL: https://doi.org/10.1177/0165551512469765.

[28] Fernando Diaz, Bhaskar Mitra, and Nick Craswell. „Query Expansion with Locally-Trained Word Embeddings." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 367–377. DOI: 10.18653/v1/P16-1035. URL: https://www.aclweb.org/anthology/P16-1035.

[29] Doug Downey, Susan Dumais, and Eric Horvitz. „Models of Searching and Browsing: Languages, Studies, and Applications." In: *Proceedings of the 20th International Joint Conference on Artifical Intelligence*. IJCAI'07. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007, pp. 2740–2747. URL: http://dl.acm.org/citation.cfm?id=1625275.1625716.

[30] Cynthia Dwork. „Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II." In: ed. by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. Chap. Differential Privacy, pp. 1–12. ISBN: 978-3-540-35908-1. DOI: 10.1007/11787006{_}1. URL: http://dx.doi.org/10.1007/11787006_1.

[31] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. „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. New York, NY, USA: ACM, 2014, pp. 223–232. ISBN: 978-1-4503-2351-2. DOI: 10.1145/2556195.2556217. URL: http://doi.acm.org/10.1145/2556195.2556217.

[32] Francis C. Fernández-Reyes, Jorge Hermosillo-Valadez, and Manuel Montes-y Gómez. „A Prospect-Guided global query expansion strategy using word embeddings." In: *Information Processing & Management* 54.1 (Jan. 2018), pp. 1–13. ISSN: 0306-4573. DOI: 10.1016/J.IPM.2017.09.001. URL: https://www.sciencedirect.com/science/article/abs/pii/S0306457317301140?via%3Dihub.

[33] Paul Gooding. „Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis." In: *Journal of Documentation* 72.2 (2016), pp. 232–246. DOI: 10.1108/JD-10-2014-0149. URL: http://dx.doi.org/10.1108/JD-10-2014-0149.

[34] Daniel Grech and Paul Clough. „Investigating Cluster Stability when Analyzing Transaction Logs." In: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. JCDL '16. New York, NY, USA: ACM, 2016, pp. 115–118. ISBN: 978-1-4503-4229-2. DOI: 10.1145/2910896.2910923. URL: http://doi.acm.org/10.1145/2910896.2910923.

[35] Joan Guisado-Gámez, Arnau Prat-Pérez, and Josep Lluis Larriba-Pey. „Query Expansion via structural motifs in Wikipedia Graph." In: *CoRR* abs/1602.07217 (Feb. 2016). URL: http://arxiv.org/abs/1602.07217.

[36] Fan Guo, Chao Liu, and Yi Min Wang. „Efficient Multiple-click Models in Web Search." In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. WSDM '09. New York, NY, USA: ACM, 2009, pp. 124–131. ISBN: 978-1-60558-390-7. DOI: 10.1145/1498759.1498818. URL: http://doi.acm.org/10.1145/1498759.1498818.

[37] Hye-Jung Han, Soohyung Joo, and Dietmar Wolfram. „Using transaction logs to better understand user search session patterns in an image-based digital library." In: *Journal of the Korean BIBLIA Society for library and Information Science* 25.1 (2014), pp. 19–37.

[38] Hyejung Han and Dietmar Wolfram. „An exploration of search session patterns in an image-based digital library." In: *Journal of Information Science* 42.4 (2016), pp. 477–491. DOI: 10.1177/0165551515598952. URL: https://doi.org/10.1177/0165551515598952.

[39] Sandra G Hart and Lowell E Staveland. „Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research." In: *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[40] Ahmed Hassan, Ryen W. White, Susan T. Dumais, and Yi-Min Wang. „Struggling or Exploring? Disambiguating Long Search Sessions." In: *Seventh ACM WSDM*. 2014. ISBN: 9781450323512. DOI: 10.1145/2556195.2556221.

[41] Jiyin He, Marc Bron, Arjen de Vries, Leif Azzopardi, and Maarten de Rijke. „Untangling Result List Refinement and Ranking Quality: A Framework for Evaluation and Prediction." In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '15. New York, NY, USA: ACM, 2015, pp. 293–302. ISBN: 978-1-4503-3621-5. DOI: 10.1145/2766462.2767740. URL: http://doi.acm.org/10.1145/2766462.2767740.

[42] Jiyin He, Pernilla Qvarfordt, Martin Halvey, and Gene Golovchinsky. „Beyond actions: Exploring the discovery of tactics from user logs." In: *Information Processing & Management* 52.6 (2016), pp. 1200 –1226. DOI: http://dx.doi.org/10.1016/j.ipm.2016.05.007. URL: http://www.sciencedirect.com/science/article/pii/S0306457316301625.

[43] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. „Finding the Flow in Web Site Search." In: *Commun. ACM* 45.9 (Sept. 2002), pp. 42–49. ISSN: 0001-0782. DOI: 10.1145/567498.567525. URL: http://doi.acm.org/10.1145/567498.567525.

[44] Daniel Hienert and Dagmar Kern. „Term-Mouse-Fixations As an Additional Indicator for Topical User Interests in Domain-Specific Search." In: *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*. ICTIR '17. New York, NY, USA: ACM, 2017, pp. 249–252. ISBN: 978-1-4503-4490-6. DOI: 10.1145/3121050.3121088. URL: http://doi.acm.org/10.1145/3121050.3121088.

[45] Daniel Hienert and Dagmar Kern. „Recognizing Topic Change in Search Sessions of Digital Libraries Based on Thesaurus and Classification System." In: *19th ACM/IEEE Joint Conference on Digital Libraries, JCDL 2019, Champaign, IL, USA, June 2-6, 2019*. Ed. by Maria Bonn, Dan Wu, J. Stephen

Downie, and Alain Martaus. IEEE, 2019, pp. 297–300. DOI: 10.1109/JCDL.2019.00049. URL: https://doi.org/10.1109/JCDL.2019.00049.

[46] David C Hoaglin. „Letter Values: A Set of Selected Order Statistics." In: *Understanding robust and exploratory data analysis*. Ed. by David C Hoaglin, Frederick Mosteller, and John Wilder Tukey. Wiley New York, 1983. Chap. 2, pp. 33–57.

[47] Laura Hollink, Peter Mika, and Roi Blanco. „Web Usage Mining with Semantic Analysis." In: *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: ACM, 2013, pp. 561–570. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488438. URL: http://doi.acm.org/10.1145/2488388.2488438.

[48] Vera Hollink, Theodora Tsikrika, and Arjen P de Vries. „Semantic search log analysis: A method and a study on professional image search." In: *Journal of the American Society for Information Science and Technology* 62.4 (June 2011), pp. 691–713. ISSN: 1532-2890. DOI: 10.1002/asi.21484. URL: http://dx.doi.org/10.1002/asi.21484.

[49] Jason I Hong and James A Landay. „WebQuilt: A Framework for Capturing and Visualizing the Web Experience." In: *Proceedings of the 10th International Conference on World Wide Web*. WWW '01. New York, NY, USA: ACM, 2001, pp. 717–724. ISBN: 1-58113-348-0. DOI: 10.1145/371920.372188.

[50] Yuan Hong, Jaideep Vaidya, Haibing Lu, Panagiotis Karras, and Sanjay Goel. „Collaborative search log sanitization: Toward differential privacy and boosted utility." In: *IEEE Transactions on Dependable and Secure Computing* 12.5 (2015), pp. 504–518. DOI: 10.1109/TDSC.2014.2369034. URL: http://dx.doi.org/10.1109/TDSC.2014.2369034.

[51] Bouke Huurnink, Laura Hollink, Wietske Den Van Heuvel, and Maarten De Rijke. „Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis." In: *Journal of the American Society for Information Science and Technology* (2010). ISSN: 15322882. DOI: 10.1002/asi.21327.

[52] Bernard J. Jansen and Amanda Spink. *How are we searching the World Wide Web? A comparison of nine search engine transaction logs*. 2006. DOI: 10.1016/j.ipm.2004.10.007.

[53] Bernard J Jansen, Amanda Spink, and Tefko Saracevic. „Real life, real users, and real needs: a study and analysis of user queries on the web." In: *Information Processing and Management* 36.2 (2000), pp. 207–227.

[54] Rosie Jones and Kristina Lisa Klinkner. „Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs." In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. New York, NY, USA: ACM, 2008, pp. 699–708. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458176.

[55] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. „"I Know What You Did Last Summer": Query Logs and User Privacy." In: *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. New York, NY, USA: ACM, 2007, pp. 909–914. ISBN: 978-1-59593-803-9. DOI: 10.1145/1321440.1321573. URL: http://doi.acm.org/10.1145/1321440.1321573.

[56] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. „Vanity Fair: Privacy in Querylog Bundles." In: *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. CIKM '08. New York, NY, USA: ACM, 2008, pp. 853–862. ISBN: 978-1-59593-991-3. DOI: 10.1145/1458082.1458195. URL: http://doi.acm.org/10.1145/1458082.1458195.

[57] Steve Jones, Sally Jo Cunningham, Rodger Mcnab, and Stefan Boddie. „A Transaction Log Analysis of a Digital Library." In: *International Journal on Digital Libraries* 3.2 (2000), pp. 152–169. DOI: 10.1007/s007999900022.

[58] L Kaufman and P J Rousseeuw. *Clustering by means of medoids*. 1987.

[59] Hao Ren Ke, Rolf Kwakkelaar, Yu Min Tai, and Li Chun Chen. „Exploring behavior of E-journal users in science and technology: Transaction log analysis of Elsevier's ScienceDirect OnSite in Taiwan." In: *Library and Information Science Research* (2002). ISSN: 07408188. DOI: 10.1016/S0740-8188(02)00126-3.

[60] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. „Releasing Search Queries and Clicks Privately." In: *Proceedings of the 18th International Conference on World Wide Web*. WWW '09. New York, NY, USA: ACM, 2009, pp. 171–180. ISBN: 978-1-60558-487-4. DOI: 10.1145/1526709.1526733. URL: http://doi.acm.org/10.1145/1526709.1526733.

[61] Bill Kules and Robert Capra. „Designing Exploratory Search Tasks for User Studies of Information Seeking Support Systems." In: *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*. JCDL '09. New York, NY, USA: ACM, 2009, pp. 419–420. ISBN: 978-1-60558-322-8. DOI: 10.1145/1555400.1555492. URL: http://doi.acm.org/10.1145/1555400.1555492.

[62] Heidi Lam, Daniel M Russell, Diane Tang, and Tamara Munzner. „Session Viewer: Visual Exploratory Analysis of Web Session Logs." In: *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2007, Sacramento, California, USA, October 30-November 1, 2007*. IEEE Computer Society, 2007, pp. 147–154. DOI: 10.1109/VAST.2007.4389008.

[63] Omer Levy and Yoav Goldberg. „Dependency-Based Word Embeddings." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 302–308. DOI: 10.3115/v1/P14-2050.

[64] Fang Liu, Clement Yu, and Weiyi Meng. „Personalized Web Search by Mapping User Queries to Categories.” In: *Proceedings of the Eleventh International Conference on Information and Knowledge Management*. CIKM ’02. New York, NY, USA: ACM, 2002, pp. 558–565. ISBN: 1-58113-492-4. DOI: 10.1145/584792.584884. URL: http://doi.acm.org/10.1145/584792.584884.

[65] Haibin Liu and Vlado Kešelj. „Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users’ future requests.” In: *Data and Knowledge Engineering* 61.2 (2007), pp. 304–330. DOI: 10.1016/j.datak.2006.06.001.

[66] Zhicheng Liu, Yang Wang, Mira Dontcheva, Matthew Hoffman, Seth Walker, and Alan Wilson. „Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths.” In: *IEEE Trans. Vis. Comput. Graph.* 23.1 (2017), pp. 321–330. DOI: 10.1109/TVCG.2016.2598797.

[67] Malika Mahoui and Sally Jo Cunningham. „Search Behavior in a Research-Oriented Digital Library.” In: *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001 Darmstadt, Germany, September 4-9, 2001 Proceedings*. Ed. by Panos Constantopoulos and Ingeborg T Sølvberg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 13–24. ISBN: 978-3-540-44796-2. DOI: 10.1007/3-540-44796-2{_}2. URL: http://dx.doi.org/10.1007/3-540-44796-2_2.

[68] Pekka Malo, Ankur Sinha, Jyrki Wallenius, and Pekka Korhonen. „Concept-based document classification using Wikipedia and value function.” In: *Journal of the American Society for Information Science and Technology* 62.12 (Dec. 2011), pp. 2496–2511. ISSN: 15322882. DOI: 10.1002/asi.21596.

[69] Jonathan Mamou, Oren Pereg, Moshe Wasserblat, Ido Dagan, Yoav Goldberg, Alon Eirew, Yael Green, Shira Guskin, Peter Izsak, and Daniel Korat. „Term Set Expansion based on Multi-Context Term Embeddings: an End-to-end Workflow.” In: *arXiv preprint arXiv:1807.10104* (2018).

[70] Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke. „Mapping queries to the Linking Open Data cloud: A case study using DBpedia.” In: *Web Semantics: Science, Services and Agents on the World Wide Web* 9.4 (2011), pp. 418 –433. DOI: http://dx.doi.org/10.1016/j.websem.2011.04.001. URL: http://www.sciencedirect.com/science/article/pii/S1570826811000187.

[71] Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke. „Mapping queries to the Linking Open Data cloud: A case study usingDBpedia.” In: *J. Web Semant.* 9.4 (2011), pp. 418–433. DOI: 10.1016/j.websem.2011.04.001.

[72] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. „Efficient Estimation of Word Representations in Vector Space." In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. Jan. 2013. URL: http://arxiv.org/abs/1301.3781.

[73] Aida Nematzadeh, Stephan C Meylan, and Thomas L Griffiths. „Evaluating Vector-Space Models of Word Representation, or, The Unreasonable Effectiveness of Counting Words Near Other Words." In: *CogSci*. 2017.

[74] Raymond T Ng and Jiawei Han. „CLARANS: A Method for Clustering Objects for Spatial Data Mining." In: *IEEE Trans. Knowl. Data Eng.* 14.5 (2002), pp. 1003–1016. DOI: 10.1109/TKDE.2002.1033770. URL: https://doi.org/10.1109/TKDE.2002.1033770.

[75] Xi Niu and Bradley M Hemminger. „A method for visualizing transaction logs of a faceted OPAC." In: *Code4Lib Journal* 12 (2010).

[76] Xi Niu and Bradley M. Hemminger. „Beyond text querying and ranking list: How people are searching through faceted catalogs in two library environments." In: *Proceedings of the ASIST Annual Meeting*. 2010. DOI: 10.1002/meet.14504701294.

[77] Xi Niu and Bradley Hemminger. „Analyzing the interaction patterns in a faceted search interface." In: *Journal of the Association for Information Science and Technology* (2015). ISSN: 23301643. DOI: 10.1002/asi.23227.

[78] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. „A Bermuda Triangle?: A Review of Method Application and Triangulation in User Experience Evaluation." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: ACM, 2018, 461:1–461:16. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174035.

[79] Peter J Rousseeuw. „Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." In: *Journal of Computational and Applied Mathematics* 20.Supplement C (1987), pp. 53 –65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

[80] Michalis Sfakakis and Sarantos Kapidakis. „User Behavior Tendencies on Data Collections in a Digital Library." In: *Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002 Rome, Italy, September 16–18, 2002 Proceedings*. Ed. by Maristella Agosti and Costantino Thanos. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 550–559. ISBN: 978-3-540-45747-3. DOI: 10.1007/3-540-45747-X{_}41. URL: http://dx.doi.org/10.1007/3-540-45747-X_41.

[81] By Shane Greenstein and Feng Zhu. „Is Wikipedia Biased?" In: *American Economic Review* 102.3 (2012), pp. 343–48. DOI: 10.1257/aer.102.3.343.

[82] Mark D Smucker and James Allan. *An Investigation of Dirichlet Prior Smoothing's Performance Advantage*. Tech. rep. The University of Massachusetts, The Center for Intelligent Information Retrieval, 2006.

[83] Amanda Spink and Bernard J Jansen. *Web search: Public searching of the Web*. Vol. 6. Springer Science & Business Media, 2006.

[84] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. „Indri: A language model-based search engine for complex queries (extended version)." In: *CIIR Technical Report* (2005).

[85] Jaime Teevan, Meredith Ringel Morris, and Steve Bush. „Discovering and using groups to improve personalized search." In: *Proceedings of the Second International Conference on Web Search andWeb Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*. Ed. by and, Paolo Boldi and, Berthier A. Ribeiro-Neto and, and Ricardo A Baeza-Yates Berkant Barla Cambazoglu. ACM, 2009, pp. 15–24. ISBN: 978-1-60558-390-7. DOI: 10.1145/1498759.1498786. URL: http://doi.acm.org/10.1145/1498759.1498786.

[86] Robert Tibshirani and Guenther Walther. „Cluster Validation by Prediction Strength." In: *Journal of Computational and Graphical Statistics* 14.3 (Feb. 2005), pp. 511–528. ISSN: 1061-8600. DOI: 10.1198/106186005X59243.

[87] David Walsh, Paul D Clough, Mark M Hall, Frank Hopfgartner, Jonathan Foster, and Georgios Kontonatsios. „Analysis of Transaction Logs from National Museums Liverpool." In: *Digital Libraries for Open Knowledge*. Ed. by Antoine Doucet, Antoine Isaac, Koraljka Golub, Trond Aalberg, and Adam Jatowt. Vol. 11799. Lecture Notes in Computer Science. Springer, 2019, pp. 84–98. ISBN: 978-3-030-30759-2. DOI: 10.1007/978-3-030-30760-8{_}7.

[88] Rita Wan-Chik, Paul Clough, and Mark Sanderson. „Investigating religious information searching through analysis of a search engine log." In: *Journal of the American Society for Information Science and Technology* 64.12 (2013), pp. 2492–2506. DOI: 10.1002/asi.22945. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.22945.

[89] Gang Wang, Xinyi Zhang, Shiliang Tang, Haitao Zheng, and Ben Y Zhao. „Unsupervised Clickstream Clustering for User Behavior Analysis." In: *Proceedings of the 2016 CHI Conference on Human Factors in ComputingSystems, San Jose, CA, USA, May 7-12, 2016*. Ed. by Allison Druin and, Cliff Lampe and, Dan Morris and, Juan Pablo Hourcade and, and Jofish Kaye. ACM, 2016, pp. 225–236. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858107. URL: http://doi.acm.org/10.1145/2858036.2858107.

[90] Timothy Weale. *Utilizing Wikipedia Categories for Document Classification*. 2006.

[91] Jishang Wei, Zeqian Shen, Neel Sundaresan, and Kwan-Liu Ma. „Visual cluster exploration of web clickstream data." In: *2012 IEEE Conference on Visual Analytics Science and Technology, VAST 2012, Seattle, WA, USA, October 14-19, 2012*. IEEE Computer Society, 2012, pp. 3–12. DOI: 10.1109/VAST.2012.6400494.

[92] Ryen W White, Peter Bailey, and Liwei Chen. „Predicting User Interests from Contextual Information." In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 363–370. ISBN: 978-1-60558-483-6. DOI: 10.1145/1571941.1572005. URL: http://doi.acm.org/10.1145/1571941.1572005.

[93] Ryen W White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. „Enhancing Personalized Search by Mining and Modeling Task Behavior." In: *Proceedings of the 22Nd International Conference on World Wide Web*. WWW '13. New York, NY, USA: ACM, 2013, pp. 1411–1420. ISBN: 978-1-4503-2035-1. DOI: 10.1145/2488388.2488511. URL: http://doi.acm.org/10.1145/2488388.2488511.

[94] Yang Xu, Gareth J F Jones, and Bin Wang. „Query Dependent Pseudo-relevance Feedback Based on Wikipedia." In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '09. New York, NY, USA: ACM, 2009, pp. 59–66. ISBN: 978-1-60558-483-6. DOI: 10.1145/1571941.1571954. URL: http://doi.acm.org/10.1145/1571941.1571954.

[95] Sicong Zhang, Hui Yang, and Lisa Singh. „Anonymizing Query Logs by Differential Privacy." In: *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '16. New York, NY, USA: ACM, 2016, pp. 753–756. ISBN: 978-1-4503-4069-4. DOI: 10.1145/2911451.2914732. URL: http://doi.acm.org/10.1145/2911451.2914732.

[96] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. „MatrixWave: Visual Comparison of Event Sequence Data." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23,2015*. Ed. by Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo. ACM, 2015, pp. 259–268. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702419.

LIST OF SIKS DISSERTATIONS

2016

01 Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

02 Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

03 Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

04 Laurens Rietveld (VU), Publishing and Consuming Linked Data

05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers

06 Michel Wilson (TUD), Robust scheduling in an uncertain environment

07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training

08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data

09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts

10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms

11 Anne Schuth (UVA), Search Engines that Learn from Their Users

12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems

13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach

14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization

15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments

16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward

17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms

18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web

19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data

20 Daan Odijk (UVA), Context & Semantics in News & Web Search

21 Alejandro Moreno Célleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground

22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems

23 Fei Cai (UVA), Query Auto Completion in Information Retrieval

24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach

25 Julia Kiseleva (TU/e), Using Contextual Information to Understand Searching and Browsing Behavior

26 Dilhan Thilakarathne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains

27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media

28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control

29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning

30 Ruud Mattheij (UvT), The Eyes Have It

31 Mohammad Khelghati (UT), Deep web content monitoring

32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations

33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example

34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment

35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation

36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies

37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry

38 Andrea Minuto (UT), Materials that Matter - Smart Materials meet Art & Interaction Design

39 Merijn Bruijnes (UT), Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect

40 Christian Detweiler (TUD), Accounting for Values in Design

41 Thomas King (TUD), Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance

42 Spyros Martzoukos (UVA), Combinatorial and Compositional Aspects of Bilingual Aligned Corpora

43 Saskia Koldijk (RUN), Context-Aware Support for Stress Self-Management: From Theory to Practice

44 Thibault Sellam (UVA), Automatic Assistants for Database Exploration

45 Bram van de Laar (UT), Experiencing Brain-Computer Interface Control

46 Jorge Gallego Perez (UT), Robots to Make you Happy

47 Christina Weber (UL), Real-time foresight - Preparedness for dynamic innovation networks

48 Tanja Buttler (TUD), Collecting Lessons Learned

49 Gleb Polevoy (TUD), Participation and Interaction in Projects. A Game-Theoretic Analysis

50 Yan Wang (UVT), The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains

2017

01 Jan-Jaap Oerlemans (UL), Investigating Cybercrime

02 Sjoerd Timmer (UU), Designing and Understanding Forensic Bayesian Networks using Argumentation

03 Daniël Harold Telgen (UU), Grid Manufacturing; A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines

04 Mrunal Gawade (CWI), Multi-core Parallelism in a Column-store

05 Mahdieh Shadi (UVA), Collaboration Behavior

06 Damir Vandic (EUR), Intelligent Information Systems for Web Product Search

07 Roel Bertens (UU), Insight in Information: from Abstract to Anomaly

08 Rob Konijn (VU) , Detecting Interesting Differences:Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery

09 Dong Nguyen (UT), Text as Social and Cultural Data: A Computational Perspective on Variation in Text

10 Robby van Delden (UT), (Steering) Interactive Play Behavior

11 Florian Kunneman (RUN), Modelling patterns of time and emotion in Twitter #anticipointment

12 Sander Leemans (TUE), Robust Process Mining with Guarantees

13 Gijs Huisman (UT), Social Touch Technology - Extending the reach of social touch through haptic technology

14 Shoshannah Tekofsky (UvT), You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior

15 Peter Berck (RUN), Memory-Based Text Correction

16 Aleksandr Chuklin (UVA), Understanding and Modeling Users of Modern Search Engines

17 Daniel Dimov (UL), Crowdsourced Online Dispute Resolution

18 Ridho Reinanda (UVA), Entity Associations for Search

19 Jeroen Vuurens (UT), Proximity of Terms, Texts and Semantic Vectors in Information Retrieval

20 Mohammadbashir Sedighi (TUD), Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility

21 Jeroen Linssen (UT), Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)

22 Sara Magliacane (VU), Logics for causal inference under uncertainty

23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces

24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning

25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search

26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch

27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors

28 John Klein (VU), Architecture Practices for Complex Contexts

29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT"

30 Wilma Latuny (UvT), The Power of Facial Expressions

31 Ben Ruijl (UL), Advances in computational methods for QFT calculations

32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives

33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity

34 Maren Scheffel (OU), The Evaluation Framework for Learning Analytics

35 Martine de Vos (VU), Interpreting natural science spreadsheets

36 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging

37 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy

38 Alex Kayal (TUD), Normative Social Applications

39 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR

40 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems

41 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle

42 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets

43 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval

44 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering

45 Bas Testerink (UU), Decentralized Runtime Norm Enforcement

46 Jan Schneider (OU), Sensor-based Learning Support

47 Jie Yang (TUD), Crowd Knowledge Creation Acceleration

48 Angel Suarez (OU), Collaborative inquiry-based learning

2018

01 Han van der Aa (VUA), Comparing and Aligning Process Representations

02 Felix Mannhardt (TUE), Multi-perspective Process Mining

03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction

04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks

05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process

06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems

07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems

08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems

09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations

10 Julienka Mollee (VUA), Moving forward: supporting physical activity behavior change through intelligent technology

11 Mahdi Sargolzaei (UVA), Enabling Framework for Service-oriented Collaborative Networks

12 Xixi Lu (TUE), Using behavioral context in process mining

13 Seyed Amin Tabatabaei (VUA), Computing a Sustainable Future

14 Bart Joosten (UVT), Detecting Social Signals with Spatiotemporal Gabor Filters

15 Naser Davarzani (UM), Biomarker discovery in heart failure

16 Jaebok Kim (UT), Automatic recognition of engagement and emotion in a group of children

17 Jianpeng Zhang (TUE), On Graph Sample Clustering

18 Henriette Nakad (UL), De Notaris en Private Rechtspraak

19 Minh Duc Pham (VUA), Emergent relational schemas for RDF

20 Manxia Liu (RUN), Time and Bayesian Networks

21 Aad Slootmaker (OUN), EMERGO: a generic platform for authoring and playing scenario-based serious games

22 Eric Fernandes de Mello Araújo (VUA), Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks

23 Kim Schouten (EUR), Semantics-driven Aspect-Based Sentiment Analysis

24 Jered Vroon (UT), Responsive Social Positioning Behaviour for Semi-Autonomous Telepresence Robots

25 Riste Gligorov (VUA), Serious Games in Audio-Visual Collections

26 Roelof Anne Jelle de Vries (UT),Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology

27 Maikel Leemans (TUE), Hierarchical Process Mining for Scalable Software Analysis

28 Christian Willemse (UT), Social Touch Technologies: How they feel and how they make you feel

29 Yu Gu (UVT), Emotion Recognition from Mandarin Speech

30 Wouter Beek, The "K" in "semantic web" stands for "knowledge": scaling semantics to the web

2019

01 Rob van Eijk (UL),Web privacy measurement in real-time bidding systems. A graph-based approach to RTB system classification

02 Emmanuelle Beauxis Aussalet (CWI, UU), Statistics and Visualizations for Assessing Class Size Uncertainty

03 Eduardo Gonzalez Lopez de Murillas (TUE), Process Mining on Databases: Extracting Event Data from Real Life Data Sources

04 Ridho Rahmadi (RUN), Finding stable causal structures from clinical data

05 Sebastiaan van Zelst (TUE), Process Mining with Streaming Data

06 Chris Dijkshoorn (VU), Nichesourcing for Improving Access to Linked Cultural Heritage Datasets

07 Soude Fazeli (TUD), Recommender Systems in Social Learning Platforms

08 Frits de Nijs (TUD), Resource-constrained Multi-agent Markov Decision Processes

09 Fahimeh Alizadeh Moghaddam (UVA), Self-adaptation for energy efficiency in software systems

10 Qing Chuan Ye (EUR), Multi-objective Optimization Methods for Allocation and Prediction

11 Yue Zhao (TUD), Learning Analytics Technology to Understand Learner Behavioral Engagement in MOOCs

12 Jacqueline Heinerman (VU), Better Together

13 Guanliang Chen (TUD), MOOC Analytics: Learner Modeling and Content Generation

14 Daniel Davis (TUD), Large-Scale Learning Analytics: Modeling Learner Behavior & Improving Learning Outcomes in Massive Open Online Courses

15 Erwin Walraven (TUD), Planning under Uncertainty in Constrained and Partially Observable Environments

16 Guangming Li (TUE), Process Mining based on Object-Centric Behavioral Constraint (OCBC) Models

17 Ali Hurriyetoglu (RUN),Extracting actionable information from microtexts

18 Gerard Wagenaar (UU), Artefacts in Agile Team Communication

19 Vincent Koeman (TUD), Tools for Developing Cognitive Agents

20 Chide Groenouwe (UU), Fostering technically augmented human collective intelligence

21 Cong Liu (TUE), Software Data Analytics: Architectural Model Discovery and Design Pattern Detection

22 Martin van den Berg (VU),Improving IT Decisions with Enterprise Architecture

23 Qin Liu (TUD), Intelligent Control Systems: Learning, Interpreting, Verification

24 Anca Dumitrache (VU), Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

25 Emiel van Miltenburg (VU), Pragmatic factors in (automatic) image description

26 Prince Singh (UT), An Integration Platform for Synchromodal Transport

27 Alessandra Antonaci (OUN), The Gamification Design Process applied to (Massive) Open Online Courses

28 Esther Kuindersma (UL), Cleared for take-off: Game-based learning to prepare airline pilots for critical situations

29 Daniel Formolo (VU), Using virtual agents for simulation and training of social skills in safety-critical circumstances

30 Vahid Yazdanpanah (UT), Multiagent Industrial Symbiosis Systems

31 Milan Jelisavcic (VU), Alive and Kicking: Baby Steps in Robotics

32 Chiara Sironi (UM), Monte-Carlo Tree Search for Artificial General Intelligence in Games

33 Anil Yaman (TUE), Evolution of Biologically Inspired Learning in Artificial Neural Networks

34 Negar Ahmadi (TUE), EEG Microstate and Functional Brain Network Features for Classification of Epilepsy and PNES

35 Lisa Facey-Shaw (OUN), Gamification with digital badges in learning programming

36 Kevin Ackermans (OUN), Designing Video-Enhanced Rubrics to Master Complex Skills

37 Jian Fang (TUD), Database Acceleration on FPGAs

38 Akos Kadar (OUN), Learning visually grounded and multilingual representations

2020

01 Armon Toubman (UL), Calculated Moves: Generating Air Combat Behaviour

02 Marcos de Paula Bueno (UL), Unraveling Temporal Processes using Probabilistic Graphical Models

03 Mostafa Deghani (UvA), Learning with Imperfect Supervision for Language Understanding

04 Maarten van Gompel (RUN), Context as Linguistic Bridges

05 Yulong Pei (TUE), On local and global structure mining

06 Preethu Rose Anish (UT), Stimulation Architectural Thinking during Requirements Elicitation - An Approach and Tool Support

07 Wim van der Vegt (OUN), Towards a software architecture for reusable game components

08 Ali Mirsoleimani (UL),Structured Parallel Programming for Monte Carlo Tree Search

09 Myriam Traub (UU), Measuring Tool Bias and Improving Data Quality for Digital Humanities Research

10 Alifah Syamsiyah (TUE), In-database Preprocessing for Process Mining

11 Sepideh Mesbah (TUD), Semantic-Enhanced Training Data AugmentationMethods for Long-Tail Entity Recognition Models

12 Ward van Breda (VU), Predictive Modeling in E-Mental Health: Exploring Applicability in Personalised Depression Treatment

13 Marco Virgolin (CWI), Design and Application of Gene-pool Optimal Mixing Evolutionary Algorithms for Genetic Programming

14 Mark Raasveldt (CWI/UL), Integrating Analytics with Relational Databases

15 Konstantinos Georgiadis (OUN), Smart CAT: Machine Learning for Configurable Assessments in Serious Games

16 Ilona Wilmont (RUN), Cognitive Aspects of Conceptual Modelling

17 Daniele Di Mitri (OUN), The Multimodal Tutor: Adaptive Feedback from Multimodal Experiences

18 Georgios Methenitis (TUD), Agent Interactions & Mechanisms in Markets with Uncertainties: Electricity Markets in Renewable Energy Systems

19 Guido van Capelleveen (UT), Industrial Symbiosis Recommender Systems

20 Albert Hankel (VU), Embedding Green ICT Maturity in Organisations

21 Karine da Silva Miras de Araujo (VU), Where is the robot?: Life as it could be

22 Maryam Masoud Khamis (RUN), Understanding complex systems implementation through a modeling approach: the case of e-government in Zanzibar

23 Rianne Conijn (UT), The Keys to Writing: A writing analytics approach to studying writing processes using keystroke logging

24 Lenin da Nóbrega Medeiros (VUA/RUN), How are you feeling, human? Towards emotionally supportive chatbots

25 Xin Du (TUE), The Uncertainty in Exceptional Model Mining

26 Krzysztof Leszek Sadowski (UU), GAMBIT: Genetic Algorithm for Model-Based mixed-Integer opTimization

27 Ekaterina Muravyeva (TUD), Personal data and informed consent in an educational context

28 Bibeg Limbu (TUD), Multimodal interaction for deliberate practice: Training complex skills with augmented reality

29 Ioan Gabriel Bucur (RUN), Being Bayesian about Causal Inference

30 Bob Zadok Blok (UL), Creatief, Creatieve, Creatiefst

31 Gongjin Lan (VU), Learning better – From Baby to Better

32 Jason Rhuggenaath (TUE), Revenue management in online markets: pricing and online advertising

33 Rick Gilsing (TUE), Supporting service-dominant business model evaluation in the context of business model innovation

34 Anna Bon (MU), Intervention or Collaboration? Redesigning Information and Communication Technologies for Development

35 Siamak Farshidi (UU), Multi-Criteria Decision-Making in Software Production

2021

01 Francisco Xavier Dos Santos Fonseca (TUD),Location-based Games for Social Interaction in Public Space

02 Rijk Mercuur (TUD), Simulating Human Routines: Integrating Social Practice Theory in Agent-Based Models

03 Seyyed Hadi Hashemi (UVA), Modeling Users Interacting with Smart Devices

04 Ioana Jivet (OU), The Dashboard That Loved Me: Designing adaptive learning analytics for self-regulated learning

05 Davide Dell'Anna (UU), Data-Driven Supervision of Autonomous Systems

06 Daniel Davison (UT), "Hey robot, what do you think?" How children learn with a social robot

07 Armel Lefebvre (UU), Research data management for open science

08 Nardie Fanchamps (OU), The Influence of Sense-Reason-Act Programming on Computational Thinking

09 Cristina Zaga (UT), The Design of Robothings. Non-Anthropomorphic and Non-Verbal Robots to Promote Children's Collaboration Through Play

10 Quinten Meertens (UvA), Misclassification Bias in Statistical Learning

11 Anne van Rossum (UL), Nonparametric Bayesian Methods in Robotic Vision

12 Lei Pi (UL), External Knowledge Absorption in Chinese SMEs

13 Bob R. Schadenberg (UT), Robots for Autistic Children: Understanding and Facilitating Predictability for Engagement in Learning

14 Negin Samaeemofrad (UL), Business Incubators: The Impact of Their Support

15 Onat Ege Adali (TU/e), Transformation of Value Propositions into Resource Re-Configurations through the Business Services Paradigm

16 Esam A. H. Ghaleb (UM), Bimodal emotion recognition from audio-visual cues

17 Dario Dotti (UM), Human Behavior Understanding from motion and bodily cues using deep neural networks

18 Remi Wieten (UU), Bridging the Gap Between Informal Sense-Making Tools and Formal Systems - Facilitating the Construction of Bayesian Networks and Argumentation Frameworks

19 Roberto Verdecchia (VU), Architectural Technical Debt: Identification and Management

20 Masoud Mansoury (TU/e), Understanding and Mitigating Multi-Sided Exposure Bias in Recommender Systems

21 Pedro Thiago Timbó Holanda (CWI), Progressive Indexes

22 Sihang Qiu (TUD), Conversational Crowdsourcing

23 Hugo Manuel Proença (LIACS), Robust rules for prediction and description

24 Kaijie Zhu (TUE), On Efficient Temporal Subgraph Query Processing

25 Eoin Martino Grua (VUA), The Future of E-Health is Mobile: Combining AI and Self-Adaptation to Create Adaptive E-Health Mobile Applications

26 Benno Kruit (CWI & VUA), Reading the Grid: Extending Knowledge Bases from Human-readable Tables

27 Jelte van Waterschoot (UT), Personalized and Personal Conversations: Designing Agents Who Want to Connect With You

28 Christoph Selig (UL), Understanding the Heterogeneity of Corporate Entrepreneurship Programs

2022

01 Judith van Stegeren (UT), Flavor text generation for role-playing video games

02 Paulo da Costa (TU/e), Data-driven Prognostics and Logistics Optimisation: A Deep Learning Journey

03 Ali el Hassouni (VUA), A Model A Day Keeps The Doctor Away: Reinforcement Learning For Personalized Healthcare

04 Ünal Aksu (UU), A Cross-Organizational Process Mining Framework

05 Shiwei Liu (TU/e), Sparse Neural Network Training with In-Time Over-Parameterization

06 Reza Refaei Afshar (TU/e), Machine Learning for Ad Publishers in Real Time Bidding

07 Sambit Praharaj (OU), Measuring the Unmeasurable? Towards Automatic Co-located Collaboration Analytics

08 Maikel L. van Eck (TU/e), Process Mining for Smart Product Design

09 Oana Andreea Inel (VUA), Understanding Events: A Diversity-driven Human-Machine Approach

10 Felipe Moraes Gomes (TUD), Examining the Effectiveness of Collaborative Search Engines

11 Mirjam de Haas (UT), Staying engaged in child-robot interaction, a quantitative approach to studying preschoolers’ engagement with robots and tasks during second-language tutoring

12 Guanyi Chen (UU), Computational Generation of Chinese Noun Phrases

13 Xander Wilcke (VUA), Machine Learning on Multimodal Knowledge Graphs: Opportunities, Challenges, and Methods for Learning on Real-World Heterogeneous and Spatially-Oriented Knowledge

14 Michiel Overeem (UU), Evolution of Low-Code Platforms

15 Jelmer Jan Koorn (UU), Work in Process: Unearthing Meaning using Process Mining

16 Pieter Gijsbers (TU/e), Systems for AutoML Research

17 Laura van der Lubbe (VUA), Empowering vulnerable people with serious games and gamification

18 Paris Mavromoustakos Blom (TiU), Player Affect Modelling and Video Game Personalisation

19 Bilge Yigit Ozkan (UU), Cybersecurity Maturity Assessment and Standardisation

20 Fakhra Jabeen (VUA), Dark Side of the Digital Media - Computational Analysis of Negative Human Behaviors on Social Media

21 Seethu Mariyam Christopher (UM), Intelligent Toys for Physical and Cognitive Assessments

22 Alexandra Sierra Rativa (TiU), Virtual Character Design and its potential to foster Empathy, Immersion, and Collaboration Skills in Video Games and Virtual Reality Simulations

23 Ilir Kola (TUD), Enabling Social Situation Awareness in Support Agents

24 Samaneh Heidari (UU), Agents with Social Norms and Values - A framework for agent based social simulations with social norms and personal values

25 Anna L.D. Latour (LU), Optimal decision-making under constraints and uncertainty

26 Anne Dirkson (LU), Knowledge Discovery from Patient Forums: Gaining novel medical insights from patient experiences

27 Christos Athanasiadis (UM), Emotion-aware cross-modal domain adaptation in video sequences

28 Onuralp Ulusoy (UU), Privacy in Collaborative Systems

29 Jan Kolkmeier (UT), From Head Transform to Mind Transplant: Social Interactions in Mixed Reality

30 Dean De Leo (CWI), Analysis of Dynamic Graphs on Sparse Arrays

31 Konstantinos Traganos (TU/e), Tackling Complexity in Smart Manufacturing with Advanced Manufacturing Process Management

32 Cezara Pastrav (UU), Social simulation for socio-ecological systems

33 Brinn Hekkelman (CWI/TUD), Fair Mechanisms for Smart Grid Congestion Management

34 Nimat Ullah (VUA), Mind Your Behaviour: Computational Modelling of Emotion & Desire Regulation for Behaviour Change

35 Mike E.U. Ligthart (VUA), Shaping the Child-Robot Relationship: Interaction Design Patterns for a Sustainable Interaction

2023

01 Bojan Simoski (VUA), Untangling the Puzzle of Digital Health Interventions

02 Mariana Rachel Dias da Silva (TiU), Grounded or in flight? What our bodies can tell us about the whereabouts of our thoughts

03 Shabnam Najafian (TUD), User Modeling for Privacy-preserving Explanations in Group Recommendations

04 Gineke Wiggers (UL), The Relevance of Impact: bibliometric-enhanced legal information retrieval

05 Anton Bouter (CWI), Optimal Mixing Evolutionary Algorithms for Large-Scale Real-Valued Optimization, Including Real-World Medical Applications

06 António Pereira Barata (UL), Reliable and Fair Machine Learning for Risk Assessment

07 Tianjin Huang (TU/e), The Roles of Adversarial Examples on Trustworthiness of Deep Learning

08 Lu Yin (TU/e), Knowledge Elicitation using Psychometric Learning

09 Xu Wang (VUA), Scientific Dataset Recommendation with Semantic Techniques

10 Dennis J.N.J. Soemers (UM), Learning State-Action Features for General Game Playing

11 Fawad Taj (VUA), Towards Motivating Machines: Computational Modeling of the Mechanism of Actions for Effective Digital Health Behavior Change Applications