On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives

Tessel Bogaard
2nd Year PhD
Supervisors: Lynda Hardman, Laura Hollink, Jacco van Ossenbruggen, Jan Wielemaker
Centrum Wiskunde & Informatica (CWI)
Amsterdam, The Netherlands
Tessel.Bogaard@cwi.nl

ABSTRACT

Log analysis is an unobtrusive technique used to better understand search behavior and evaluate search systems. However, in contrast with open web search, in a vertical search system such as a digital library or media archive the collection is known and central to its purpose. This drives different, more collection-oriented questions when studying the logs. For example, whether users need different support in different parts of the collection.

In a digital library, the collection is categorized using professionally curated metadata. We conjecture that using this metadata can improve and extend the methods and techniques for log analysis. We investigate how to identify different types of search behavior using the metadata explicitly, how to explain and predict user interactions for the different types of behavior found, and finally how to communicate our research results to domain experts.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; *Query log analysis*; Search interfaces; • **Human-centered computing** → *Visual analytics*;

KEYWORDS

Log analysis, Search behavior, Faceted search, Metadata

ACM Reference format:
Tessel Bogaard. 2018. On the Interplay Between Search Behavior and Collections in Digital Libraries and Archives. In *Proceedings of 2018 Conference on Human Information Interaction & Retrieval, New Brunswick, NJ, USA, March 11–15, 2018 (CHIIR '18),* 3 pages.
https://doi.org/10.1145/3176349.3176350

1 MOTIVATION

Search log analysis is an unobtrusive technique used to better understand user behavior in search systems [1, 3, 5, 6, 8, 9, 11, 12, 19]. It can be used to evaluate search algorithms or user interfaces, or to (re-)design systems.

Traditional log analysis focuses on queries and clicks [1, 3, 8, 10–13, 16, 19]. This focus poses some disadvantages. First, queries are ambiguous, as they form an uncontrolled vocabulary and have little context to interpret the information need. 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 [7, 14, 15], and thus are seldom shared among researchers.

In a closed, vertical search system such as a digital library or archive other data is available in addition to the logs: the documents in the collection and their categorizations using professionally curated metadata. This metadata is often reflected in the search interface in facets, acting as a filter over the search results. This extra information is normally not an integral part of a log analysis.

We conjecture that using the metadata of the collection explicitly can improve and extend analytical methods for logs collected in such search systems, in order to be able to examine detailed types of search behavior in relation to specific subsets in the collection. The collection inspires some of the questions relevant here: What parts of the collection have a high user interest? Do users search differently in different parts of the collection? Do they need different support for different parts of the collection? Are there potential gaps in the collection or are some parts of the collection harder to find? A focus on the metadata of the collection, both with respect to the facets used in search and the metadata of clicked documents, makes it possible to answer these type of questions. For example, we can identify gaps in the collection where people search for certain categories of documents but have difficulties finding them. Or we may discover that search behavior within specific subsets of the collection is different, suggesting the need for a different kind of support from the search system. Additionally, with this shift away from the query to the metadata of facet use and clicked documents, we alleviate the disadvantages previously mentioned. First, facet and document metadata is not ambiguous, as it forms a controlled vocabulary. Second, we can group infrequent queries based on shared metadata. Third, the metadata is less privacy-sensitive.

In our research we investigate how to identify different types of search behavior based on metadata of search and clicked documents, how to explain and predict user interactions for these types of behavior, and how to communicate our research results to the domain experts.

2 RESEARCH GOALS AND METHODOLOGY

The main goal of this research is to improve the understanding of different types of user search behavior in vertical search systems where the content is known, by studying the interplay between search behavior and the collection in digital libraries and media

CHIIR '18, March 11–15, 2018, New Brunswick, NJ, USA

ACM ISBN 978-1-4503-4925-3/18/03.
https://doi.org/10.1145/3176349.3176350

archives. We investigate specific and detailed types of behavior and their relation to the subsets present in the collection. We expect this to lead to better support for the user (such as dynamic facet presentation or generation, or the development of different search interfaces for different types of use), and help the curators of a collection to provide better access to their documents.

We address the following research questions:

(1) How can we identify different types of search behavior in relation to the metadata of facets used and documents clicked?

(2) How can we explain and predict user interactions for each type of search behavior to gain a better understanding of the different types of search behavior?

(3) How can we communicate the research results of **RQ1** and **RQ2** to domain experts?

Central to our methodology is the investigation of correlations between metadata-based subsets in the collection and certain types of user search behavior. In the next sections we describe the methods and techniques that we (plan to) use to answer these questions: (1a) a descriptive, comparative analysis, (1b) clustering, (2) sequential modeling and (3) graph visualization.

We do our research in the context of the search interface, logs, and content of the search platform curated by the National Library of the Netherlands. This collection can be accessed using an advanced, faceted search interface[1]. We have been given access to data from the National Library. This includes ten months of log records (October 2015-March 2016, April-July 2017, and we expect to receive more logs in the future), and the complete historical newspaper collection, described in metadata records (400 years, over 100M documents) with full text available as well.

For the evaluation of the techniques and methods we are looking for a second dataset. We are currently investigating the possible use of logs of the collections of Europeana[2].

3 PROGRESS

3.1 Comparative Log Analysis

In the first year, an approach for a comparative log analysis using metadata to observe usage patterns has been developed and executed. The main research question addressed here is **RQ1**. Concretely we focus on the following question:

How can we discover specific usage patterns by comparing (1) subsets of sessions, in which certain facets were used, to (2) clicked documents, and (3) the collection?

In addition, we address **RQ3**, in particular the question whether we can provide recommendations to the curators of a digital library based on our results.

The applied methodology is to automatically label the sessions identified in the logs with the (different categories of) metadata of facets used, and then to explore if and how subsets, based on those metadata labels, correlate with specific usage patterns.

We found distinct usage patterns based on the metadata. For example, the results showed that the family announcement facet

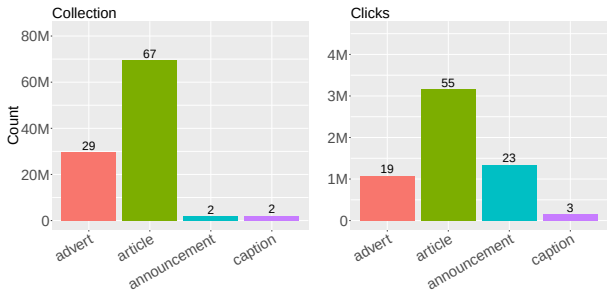

Figure 1: Item types in the collection, and of clicks in search sessions. Percentages given above the bars.

(relating to birth, death and marriage announcements) was the most frequent choice of the item type facets. The facet was used in 19% of all search sessions, with the percentage of clicks on announcements at 23%, even though announcements represent only 2% of the collection (Fig. 1). Sessions using the announcement facet are comparatively short, with fewer clicked results per session than for the other facets, even if the number of search interactions is similar. Perhaps users can more easily assess the relevance of the documents from their snippets on the results page. A recommendation given to the National Library is to give the snippets for these items extra attention.

3.2 Clustering Search Behavior

Following this, ongoing work started in the second year, we approach **RQ1** again, this time from another angle, addressing the question:

How can we identify different types of search behavior by clustering sessions based on facet use and metadata of clicked documents, as well as traditional query and click features?

We want to explore if and how characteristics of a clustering of different types of behavior correlate with search within certain subsets in the collection.

The sessions are represented using a set of features based on the interactions within the search interface, similar to the variables in [5]. In addition, the sessions are represented using a second set of features based on the metadata of the facets used and clicked documents. We then cluster the sessions two times, based on (1) the interactions with the search interface and based on (2) the metadata of facets used and clicked documents. Investigating the two resulting clusterings makes it possible to explore if and how general search behavior correlates with specific facets and metadata of clicked documents.

Since the normality assumption does not hold for our data, the CLARANS algorithm is chosen for the clustering [17]. The silhouette method is used to choose the number of clusters [18], and the stability of the clusters over different samples of the dataset to evaluate the validity of the clustering [20]. For the description and labeling of the clusters we use both feature sets for each clustering. In addition, we believe it will be insightful to investigate what type of information tactics are used within the clusters [2], and to label them according to an existing information seeking behavior model, such as those described in [21].

[1] 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. Our focus is on the well-curated historical newspaper collection, which amounts to more than 90% of all HTTP page requests on Delpher.

[2] https://www.europeana.eu/portal/en provides access to different collections from European cultural heritage institutions

4 FUTURE PLANS

After identifying different types of behavior using the clustering technique, we plan to address the second research question, **RQ2**: How can we explain and predict user interactions for each type of search behavior to gain a better understanding of the different types of search behavior? We will use a data mining technique to find the common sequential patterns within the clusters. The methodology envisioned here is Markov chain analysis to find the sequential patterns for each cluster. We will use the facets and metadata of clicked documents, labeling the type of search interaction or click. We are planning to use two evaluation methods for this stage of the research. First, we will use the predictive accuracy of the patterns discovered in the analysis. Second, we will use a test for statistical significance of the differences found between the patterns, like the Chi-squared test used for the work in [6].

In parallel to the first two research questions, we focus on **RQ3**: How can we communicate the results of **RQ1** and **RQ2** to domain experts? The domain experts are both curators and developers of digital libraries. In an ongoing development, started in the first year, we created what we call a *session graph*, a graph visualization that represents the user interactions in their search session. This visualization is already part of a method of an iterative, transparent data cleaning process, where the session graphs function as a sanity check as to whether the processed logs make sense and represent valid user interactions [4]. We will investigate how this visualization can help interpret and understand search behavior and how it can improve the analysis of search logs. An important aspect of this visualization technique is to try and create prototype session graphs. This type of graph will visualize a virtual session graph that is most typical for a cluster, creating what amounts to an aggregated central graph for the cluster. This graph visualization will – like the session graphs – be designed to help the curators of a collection, to visually inspect aggregated common search behavior in their search system. The visualization techniques we develop will be evaluated in a user study among curators and developers of a digital library.

ACKNOWLEDGMENTS

I would like to thank the National Library of the Netherlands for their support.

This research is 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.

REFERENCES

[1] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2005. 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*, Wolfgang Lindner, Marco Mesiti, Can Türker, Yannis Tzitzikas, and Athena I Vakali (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 588–596. https://doi.org/10.1007/978-3-540-30192-9{_}58

[2] Marcia J. Bates. 1979. Information search tactics. *Journal of the American Society for Information Science* 30, 4 (7 1979), 205–214. https://doi.org/10.1002/asi.4630300406

[3] Steven M Beitzel, Eric C Jensen, Abdur Chowdhury, David A Grossman, and Ophir Frieder. 2004. Hourly analysis of a very large topically categorized web query log. In *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIRConference on Research and Development in Information Retrieval, Sheffield,UK, July 25-29, 2004*, Kalervo Järvelin, James Allan, Peter Bruza, and Mark Sanderson (Eds.). ACM, 321–328. https://doi.org/10.1145/1008992.1009048

[4] T. Bogaard, J. Wielemaker, L. Hollink, and J. van Ossenbruggen. 2017. 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*, Tibor Bosse and Bert Bredeweg (Eds.). Vol. 765. Chapter 13, 181–187. https://doi.org/10.1007/978-3-319-67468-1_13

[5] Hui-Min Chen and Michael D Cooper. 2001. Using Clustering Techniques to Detect Usage Patterns in a Web-Based Information System. *Journal of the American Society for Information Science and Technology* 52, 11 (2001), 888–904. https://doi.org/10.1002/asi.1159

[6] Hui-Min Chen and Michael D. Cooper. 2002. Stochastic modeling of usage patterns in a web-based information system. *Journal of the American Society for Information Science and Technology* (2002). https://doi.org/10.1002/asi.10076

[7] Alissa Cooper. 2008. A Survey of Query Log Privacy-enhancing Techniques from a Policy Perspective. *ACM Trans. Web* 2, 4 (10 2008), 19:1Â¢ÂĂÂŞ19:27. https://doi.org/10.1145/1409220.1409222

[8] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan Dumais. 2014. Lessons from the Journey: A Query Log Analysis of Within-session Learning. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM '14)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/2556195.2556217

[9] Jiyin He, Pernilla Qvarfordt, Martin Halvey, and Gene Golovchinsky. 2016. Beyond actions: Exploring the discovery of tactics from user logs. *Information Processing & Management* 52, 6 (2016), 1200 – 1226. https://doi.org/10.1016/j.ipm.2016.05.007

[10] Laura Hollink, Peter Mika, and Roi Blanco. 2013. Web Usage Mining with Semantic Analysis. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 561–570. https://doi.org/10.1145/2488388.2488438

[11] Vera Hollink, Theodora Tsikrika, and Arjen P de Vries. 2011. Semantic search log analysis: A method and a study on professional image search. *Journal of the American Society for Information Science and Technology* 62, 4 (6 2011), 691–713. https://doi.org/10.1002/asi.21484

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

[13] Bernard J Jansen, Amanda Spink, and Tefko Saracevic. 2000. Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing and Management* 36, 2 (2000), 207–227.

[14] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2007. "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)*. ACM, New York, NY, USA, 909–914. https://doi.org/10.1145/1321440.1321573

[15] Rosie Jones, Ravi Kumar, Bo Pang, and Andrew Tomkins. 2008. Vanity Fair: Privacy in Querylog Bundles. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, New York, NY, USA, 853–862. https://doi.org/10.1145/1458082.1458195

[16] Edgar Meij, Marc Bron, Laura Hollink, Bouke Huurnink, and Maarten de Rijke. 2011. Mapping queries to the Linking Open Data cloud: A case study using DBpedia. *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 4 (2011), 418 – 433. https://doi.org/10.1016/j.websem.2011.04.001

[17] Raymond T Ng and Jiawei Han. 2002. CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE Trans. Knowl. Data Eng.* 14, 5 (2002), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770

[18] Peter J Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, Supplement C (1987), 53 – 65. https://doi.org/10.1016/0377-0427(87)90125-7

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

[20] Robert Tibshirani and Guenther Walther. 2005. Cluster Validation by Prediction Strength. *Journal of Computational and Graphical Statistics* 14, 3 (2005), 511–528. http://www.jstor.org/stable/27594130

[21] Peiling Wang. 2011. Information Behavior and Seeking. In *Interactive Information Seeking, Behaviour and Retrieval*, Ian Ruthven and Diane Kelly (Eds.). Facet Publishing, London, Chapter 2, 15–41.