A Complete Year of User Retrieval Sessions in a Social Sciences Academic Search Engine

Philipp Mayr(✉) and Ameni Kacem

GESIS - Leibniz Institute for Social Sciences, Cologne, Germany
{philipp.mayr,ameni.sahraoui}@gesis.org

Abstract. In this paper, we present an open data set extracted from the transaction log of the social sciences academic search engine sowiport. The data set includes a filtered set of 484,449 retrieval sessions which have been carried out by sowiport users in the period from April 2014 to April 2015. We propose a description of interactions performed by the academic search engine users that can be used in different applications such as result ranking improvement, user modeling, query reformulation analysis, search pattern recognition.

Keywords: Whole session retrieval · Information behavior · Session log analysis · User session data · Social sciences users

1 Introduction

Every Digital Library (DL) system generates huge amounts of usage data and DL operators often face the problem of not being able to report about the real usage on an expressive level that is moreover understandable for laymen. Reporting average statistics like number of unique sessions, page impressions, amount of actions and even click-through rates is not enough because these numbers cannot represent and explain the underlying pattern of the information behavior of DL users. Exploratory search in DLs and academic search engines [1] is a rewarding research environment for interactive IR researchers because evolving searches with complex search tasks can be observed much easier compared to web search where searchers often jump into different websites. In DLs, users typically stay in the system and work with the variety of facilities it offers. This is due to the fact that state-of-the-art DLs offer dozens of possibilities to navigate and interact with the search system [2,3]. Our motivation in proposing this data set is grounded in the observation that in the field very few open data sets which support whole session investigation exist. To the best of our knowledge there is no open data set available from academic search engines or DLs with full coverage of whole session information. Among the available data sets, we find the most famous evaluation campaign TREC (Text REtrieval Conference) which proposed TREC Session[1] [4] and Interactive[2] tracks. In fact, one way to

[1] http://trec.nist.gov/data/session.html.

[2] http://trec.nist.gov/data/interactive.html.

J. Kamps et al. (Eds.): TPDL 2017, LNCS 10450, pp. 560–565, 2017.
DOI: 10.1007/978-3-319-67008-9_46

enhance the development and evaluation of information-seeking systems is to propose shareable data sets in order to facilitate the collaboration within an interdisciplinary team including developers, computer scientists, and behavioral experts who work together in order to explore new ideas and propose improvements [5].

Consequently, with the proposed data set we want to support DL developers and IR researchers to work on the analysis of whole retrieval sessions. These practitioners need such data sets to propose methods and techniques which allow us to examine search steps, analyze usage data, understand the underlying information behavior covered in search sessions that are performed by geographically distributed persons.

2 Related Work

Interactive information retrieval (IIR) refers to a research discipline that studies the interaction between the user and the search system. In fact, researchers have moved from considering only the current query to consider the user's past interactions. Research approaches aim to understand the user search behavior in order to improve the ranking of results after submitting a query and enhance the user experience with an IR system. Thus, they study concepts such as search strategies [1,6], search term suggestions [7], communities' detection [8], personalization of search results, recommendation's impact [7], users information needs frequency and change. Many interactive IR models have been proposed in the literature (e.g. [9]) that describe the user's behavior by different steps (stages) of information seeking and interacting with an information retrieval system. In order to evaluate and analyze such models and approaches log analysis has been introduced. In [10], the authors proposed a detailed overview of the history and development of transaction log analysis by examining possible applications and features analysis. Jones et al. [11] investigated transaction logs for the Computer Science Technical Reports Collection of the New Zealand DL. The authors analyzed query complexity, query terms change, sessions frequency and length.

3 Dataset

Sowiport[3] is a DL for the Social Sciences that contains more than nine million records, full texts and research projects included from twenty-two different databases whose content is in English and German [2]. This data set **Sowiport User Search Sessions Data Set (SUSS)**[4] [12] contains individual search sessions extracted from the transaction log of sowiport. The data was collected over a period of one year (between 2nd April 2014 and 2nd April 2015). The web server log files and specific JavaScript-based logging techniques were, first, used to capture the user behavior within the system. Then, the log was heavily filtered to

[3] http://www.sowiport.de.

[4] To download the dataset: http://dx.doi.org/10.7802/1380.

exclude transactions performed by robots and short interactions limited to one action per session. After that, all transaction activities are mapped to a list of 58 different user actions which cover all types of activities and pages that can be carried out/visited within the system (e.g. typing a query, visiting a document, selecting a facet, exporting a document, etc.). For each action, a session id, the date stamp and additional information (e.g. query terms, document ids, and result lists) are stored. Based on the session id and date stamp, the step in which an action is conducted and the length of the action is included in the data set as well. The session id is assigned via browser cookies and allows tracking user behavior over multiple search sessions. Session boundaries were specified after a threshold period indicating a period of inactivity and thus the end of the session. In our data set this threshold is equal to 20 min. Thus, in the data set we find 484,449 individual search sessions and a total of 7,982,427 log entries.

4 Preliminary Analysis

In this section, we present first descriptive analysis of the SUSS data set regarding sessions, users and searches. These analyses are not following concrete research questions but are intended to show the richness of this open data set.

4.1 Description of Actions

Searching sowiport can be performed through an *All fields* search box (default search without specification), or through specifying one or more field(s): title, person, institution, number, keyword or year. The users' main actions are described in Table 1. We grouped the main actions into two categories: “Query”-related and “Document”-related actions. Another categorization of actions was proposed in [7] by specifying search interactions and successive positive actions.

4.2 Users and Sessions

Given the data set described in Sect. 3, we first analyze the user types. A user can perform a search and submit a query to sowiport without signing up. Registered users can keep the search history, add a document to favorites and create favorite lists according to their interests. We found 1,509 registered users who performed 3,372 unique sessions (0.69%). The rest of the sessions in sowiport were performed by non-registered users (99.31%).

4.3 Investigation of Actions

Main user actions as described before can be categorized into actions regarding either search queries or documents. These actions are used in different scales in the data set. Query-related actions represent 29.84% while document-related actions represent 35.79% of the total amount of actions. The rest of actions contain navigational interactions such as logging in the system, managing favorites, and accessing the system pages.

Table 1. Main actions performed by users in sowiport

Category	Action	Description	Frequency
Query	query_form	Formulating a query	179,964
	search	A search result list for any kind of search	848,556
	search_advanced	A search with the advanced settings that can limit the search fields, information type, etc	103,432
	search_keyword	A search for a keyword	43,608
	search_thesaurus	Usage of the thesaurus system	71,599
	search_institution	A search for an institution	13,104
	search_person	A search for a specific person (author/editor)	93,083
Document	view_record	Displaying a record in the result list after clicking on it	1,344,361
	view_citation	View the document's citation(s)	24,994
	view_references	View the document's references	2,086
	view_description	View the document's abstract	86,752
	export_bib	Export the document through different formats	27,229
	export_cite	Export the document's citations list	27,385
	export_mail	Send the document via email	10,987
	to_favorites	Save the document to the favorite list	5,431

Figure 1 shows the frequencies of the top six most used actions by the users in the data set. We notice that the actions *"view_record"* and *"search"* are the most used ones before *"query_form"* and *"search_keyword, person, institution"*.

In Table 2, we show a specific session, the user's ID and the actions' label and length in seconds. In this session, the user with ID *41821* started with logging into the system and then submitted a query describing his/her information need (*query_form*). After getting the result list, labeled as *resultlistids* and viewing a document, the user performed additional searches (*searchterm_2*), and displayed some results' content (*view_record*). Finally, he/she checked the external availability of a result (*goto_google_scholar*). We notice that the user spent more than 40% of the time reading documents' content.

In Fig. 2, we display the number of actions per session. We note that the average number of actions per session is 16 and only sessions with a minimum of one action are considered in this data set. We conclude, from this figure, that the number of sessions with less than 16 actions (n = 384,087) is much larger than the number of sessions having over 16 actions (n = 100,360).

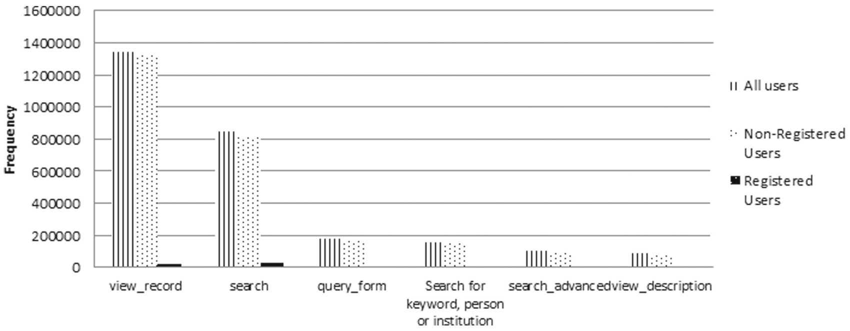

Fig. 1. Frequency distribution of the six most performed action groups

Table 2. Sample of a session search for a specific user

User ID	Date	Action label	Action length (s)
41821	2014-10-28 16:08:46	goto_login	1
	2014-10-28 16:09:13	query_form	22
	2014-10-28 16:09:35	search	10
	2014-10-28 16:09:35	resultlistids	10
	2014-10-28 16:09:45	view_record	31
	2014-10-28 16:09:45	docid	31
	2014-10-28 16:10:16	view_record	392
	2014-10-28 16:16:48	search	10
	2014-10-28 16:16:48	searchterm_2	10
	2014-10-28 16:16:48	resultlistids	10
	2014-10-28 16:16:58	view_record	9
	2014-10-28 16:17:07	goto_google_scholar	0

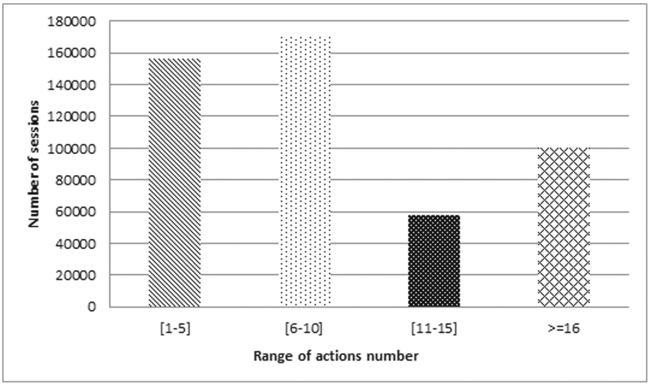

Fig. 2. Distribution of the Number of actions contained in a session

5 Future Work

For academia there is a need for open data sets which provide information about the variety of retrieval sessions and help to study and understand the abstract information behavior and common scan paths of academic users in a DL. In fact, session log provision and investigation open opportunities to enhance DLs' systems and to offer new services. Some possible future work based on our proposed data set can be outlined as follows: finding and studying abstract user groups like exhaustive or effective users; modeling academic users; analyzing reformulation and refining strategies; identifying various search phases like starting; chaining, browsing and differentiating; task characterization and prediction; personalization of search results according to the user behavior within search sessions.

Acknowledgement. This work was funded by Deutsche Forschungsgemeinschaft (DFG), grant no. MA 3964/5-1; the AMUR project at GESIS together with the working group of Norbert Fuhr. The AMUR project aims at improving the support of interactive retrieval sessions following two major goals: improving user guidance and system tuning.

References

1. Carevic, Z., Lusky, M., van Hoek, W., Mayr, P.: Investigating exploratory search activities based on the stratagem level in digital libraries. Int. J. Dig. Libr. (2017). https://link.springer.com/article/10.1007/s00799-017-0226-6
2. Hienert, D., Sawitzki, F., Mayr, P.: Digital Library Research in Action Supporting Information Retrieval in Sowiport. D-Lib Mag. **21**(3/4) (2015). doi:10.1045/march2015-hienert, http://www.dlib.org/dlib/march15/hienert/03hienert
3. Fuhr, N., et al.: Evaluation of digital libraries. Int. J. Dig. Libr. **8**(1), 21–38 (2007)
4. Kacem, A., Boughanem, M., Faiz, R.: Emphasizing temporal-based user profile modeling in the context of session search. In: SAC, pp. 925–930. ACM (2017)
5. Kelly, D., Dumais, S.T., Pedersen, J.O.: Evaluation challenges and directions for information-seeking support systems. IEEE Comput. **42**(3), 60–66 (2009)
6. Carevic, Z., Mayr, P.: Survey on high-level search activities based on the stratagem level in digital libraries. In: Fuhr, N., Kovács, L., Risse, T., Nejdl, W. (eds.) TPDL 2016. LNCS, vol. 9819, pp. 54–66. Springer, Cham (2016). doi:10.1007/978-3-319-43997-6_5
7. Hienert, D., Mutschke, P.: A usefulness-based approach for measuring the local and global effect of IIR services. In: Proceedings of CHIIR 2016, pp. 153–162. ACM (2016)
8. Akbar, M., Shaffer, C.A., Fox, E.A.: Deduced social networks for an educational digital library. In: Proceedings of JCDL 2012, pp. 43–46. ACM (2012)
9. Ellis, D.: A behavioural approach to information retrieval system design. J. Documentation **45**(3), 171–212 (1989)
10. Peters, T.A.: The history and development of transaction log analysis. Libr. Hi Tech. **11**(2), 41–66 (1993)
11. Jones, S., Cunningham, S.J., McNab, R.: An Analysis of Usage of a Digital Library. In: Nikolaou, C., Stephanidis, C. (eds.) ECDL 1998. LNCS, vol. 1513, pp. 261–277. Springer, Heidelberg (1998). doi:10.1007/3-540-49653-X_16
12. Mayr, P.: Sowiport User Search Sessions Data Set (SUSS) (2016)