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

Philipp Mayr, Ameni Kacem

GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany,
firstname.lastname@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 the data set features that can be used as ground truth for 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 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 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 [1,2].

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] [3] and Interactive[2] tracks. With the proposed data set we want to support DL developers and IR researchers to work on the analysis of whole sessions and the underlying information behavior covered in the provided user session data.

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 [4], search term suggestions [5], communities' detection [6], personalization of search results [7], recommendation's impact [5], user's information needs frequency and change.

Many interactive IR models have been proposed in the literature (e.g. [8]) 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 [9], 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. [10] investigated transaction logs for the Computer Science Technical Reports Collection of the New Zealand Digital Library. 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 [1].

This data set "Sowiport User Search Sessions Data Set (SUSS)"[4] [11] 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 used to capture the user behavior within the system. The log was heavily filtered to exclude transactions performed by robots. 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.).

[1] http://trec.nist.gov/data/session.html
[2] http://trec.nist.gov/data/interactive.html
[3] http://www.sowiport.de
[4] http://dx.doi.org/10.7802/1380

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 searches. 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 descriptive analysis of the SUSS data set regarding sessions, users and searches.

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. In fact, we grouped the main actions into two categories: "Query"-related and "Document"-related actions. Another categorization of actions was proposed in [5] by specifying search interactions and successive positive actions.

Table 1. Main actions performed by users in sowiport

Category	Action	Description
Query	query_form	Formulating a query
	search	A search result list for any kind of search
	search_advanced	A search with the advanced settings that can limit the search fields, information type, provider/database, language: or time (year, recent only)
	search_keyword	A search for a keyword
	search_thesaurus	Usage of the thesaurus system
	search_institution	A search for an institution
	search_person	A search for a specific person (author/editor)
Document	view_record	Displaying a record in the result list after clicking on it
	view_citation	View the document's citation(s)
	view_references	View the document's references
	view_description	View the document's abstract
	export_bib	Export the document through different formats
	export_cite	Export the document's citations list
	export_mail	Send the document via email
	to_favorites	Save the document to the favorite list

In order to analyze the actions' impact on the whole session search performance, we present an overview regarding the frequency of use of some actions in the next section.

4.2 Users and Sessions

Given the data set described in Section 3, we first analyze the user types. In fact, a user can perform a search and submit a query to sowiport without signing up. However, the actions will be limited. Registered users can keep the search history, add 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 Actions Investigation

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.

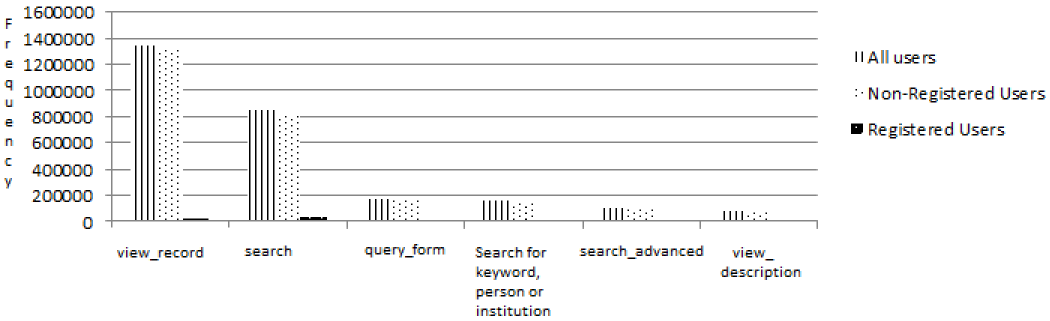

Fig. 1. Frequency distribution of the six most performed actions

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 doing some navigational actions. After getting the result list, labeled as *resultlistids*, 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.

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 16:08:47	goto_favorites	21
	2014-10-28 16:09:08	goto_home	2
	2014-10-28 16:09:13	query_form	22
	2014-10-28 16:09:35	search	10
	2014-10-28 16:09:35	searchterm_2	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 16:17:07	goto_google_scholar	0

In Figure 2 we display the amount 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).

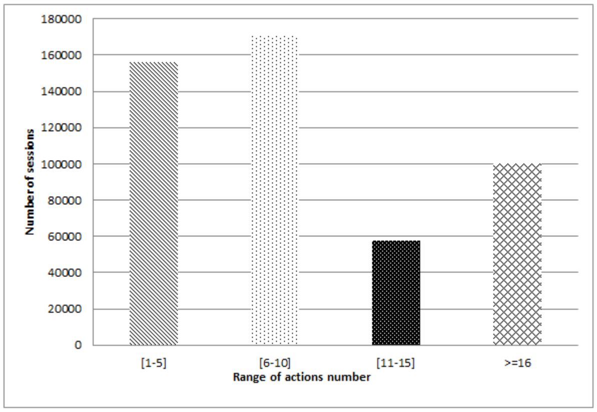

Fig. 2. Distribution of the amount of actions contained in a session

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.

6 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. Hienert, D., Sawitzki, F., Mayr, P.: Digital Library Research in Action – Supporting Information Retrieval in Sowiport. D-Lib Magazine **21**(3/4) (2015)
2. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C.P., Kovács, L., Landoni, M., Micsik, A., Papatheodorou, C., Peters, C., Sølvberg, I.: Evaluation of digital libraries. International Journal on Digital Libraries **8**(1) (2007) 21–38
3. Carterette, B., Clough, P., Hall, M., Kanoulas, E., Sanderson, M.: Evaluating Retrieval over Sessions: The TREC Session Track 2011-2014. In: Proceedings of SIGIR '16, ACM Press (2016) 685–688
4. Carevic, Z., Mayr, P.: Survey on high-level search activities based on the stratagem level in digital libraries. In: Proceedings of TPDL 2016. (2016) 54–66
5. Hienert, D., Mutschke, P.: A usefulness-based approach for measuring the local and global effect of iir services. In: Proceedings of CHIIR '16, ACM (2016) 153–162
6. Akbar, M., Shaffer, C.A., Fox, E.A.: Deduced social networks for an educational digital library. In: Proceedings of JCDL '12, ACM (2012) 43–46
7. Liu, C., Belkin, N.J., Cole, M.J.: Personalization of search results using interaction behaviors in search sessions. In: Proceedings of SIGIR '12, ACM (2012) 205–214
8. Ellis, D.: A behavioural approach to information retrieval system design. Journal of Documentation **45**(3) (1989) 171–212
9. Peters, T.A.: The history and development of transaction log analysis. Library Hi Tech **11**(2) (1993) 41–66
10. Jones, S., Cunningham, S.J., McNab, R.: An analysis of usage of a digital library. In: Proceedings of ECDL 1998. (1998) 261–277
11. Mayr, P.: Sowiport User Search Sessions Data Set (SUSS) (2016)