Building Interaction Profiles for Better Search Tools in DLs

Maram Barifah
Università della Svizzera italiana (USI)
Faculty of Informatics, Lugano, Switzerland
maram.barifah@usi.ch

This research starts by considering users of a digital library (DL) and aims at using the data extracted from logged files, including search strategies, and queries, to build effective user-interaction profiles. and use them to guide designers and systems developers in the production of more usable, useful and effective interaction profiles. It is important to stress that the proposed interaction profiles are built by extracting a number of features from the log files. Thus, they contain information about real searching experiences, including usage patterns, user familiarity with the system, and time intervals.

This study is conducted in collaboration with RERO Doc digital library[1]. RERO Doc is the network of the libraries of Western Switzerland. Users from different parts of the world can search on different domains: Nursing, Economics, Computer Science and others. RERO Doc provides various document types such as books, articles, theses, periodicals,etc. Thus, the research questions are:

(1) *What are the most suitable techniques to produce rich/realistic groups of data extracted from log files in order to build interaction profiles?*
(2) *What are the main features to characterise interaction profiles?*
(3) *What is the minimum size of data to produce robust groups to use for building interaction profiles?*

Data preparing and processing:
Interface analysis: the interface is inspected in order to understand different search options.
Preprocessing phase: we follow the framework of [2] as the following:
Data loading: the dataset consists of 59 million records 20 GB collected over a six-month period.
Data cleaning: including users identifications hidden, elimination of the erroneous, and corrupted records.
Data parsing: consists of sessions recognition and removing the non-human sessions e.g. Googlebot, SemanticScolarBot. The session "is a common unit of interaction that is used in search log analysis" [3]. Session recognition depends on the user interactions and on the features of the interface. This phase is crucial for identifying distinct classes of searching patterns [2, 3]. We identify a session by the combinations of user IP, time stamp, and user agent extracted from the log files. Also, the non-human requests are removed in this stage and so data is reduced to 9 GB.
Data coding: in this phase, the URL requests were analysed and divided into meaningful parts including user IP, time stamp, request, referrer, user agent, session IP. Researchers follow different strategies to analyse the URLs embodied on the log files. For example,[1] build a hierarchical taxonomy of the website, and [2] define a code schema of the all types of the interactions based on analysing and understanding the structure of the website. In this research we consider both strategies.
Features engineering: Based on the interface and log files analysis, the meaningful features are identified.
Mining user behaviour: The remaining sessions are further analysed and grouped based on the available variables in the records. We started the analysis of the first million record which consists of 125000 session records. 72095 sessions were detected with only one record. The aim of this phase is to identify different usage patterns among user interactions and group them accordingly. Two main different grouping techniques were used: topic modelling and K-Means. For 6 topic model, the Coherence Score of the topic modelling is 0.35. For K-Means the estimated number of clusters is 6 with Silhouette Coefficient of 0.95. So far, six different usage patterns have been identified and interpreted qualitatively:

(1) Single sessions or known-item, where searchers visit RERO for downloading documents without any interactions.
(2) Complicated sessions, where the usage pattern is characterised by heavily interactions including submitting queries, browsing, and using different functions.
(3) Light navigators who navigate the library for navigating without using different functions on the interface.
(4) Advance navigators whose navigations are characterised by using different functions and many iterations.
(5) Light browsers whose searching is simple, short without using different functions.
(6) Advanced browsers, their interactions is long and including advance search functions.

In conclusion, log file analysis is an unobtrusive method to detect usage patterns of digital library. The aim of this research in progress is to build interactions profiles in the digital library context in order to gain more insights into users searching experiences. We plan more experiments to test and compare the effectiveness of the techniques used to group data for preparing interaction profiles. Then, we will involve experts to assess the quality of these interaction profiles and how effective these are in assisting designers and system developers in the production of more usable, useful and effective tools to support searchers.

[1] https://doc.rero.ch

DESIRES, August 2018, Bertinoro, Italy

REFERENCES

[1] Hui-Min Chen and Michael D Cooper. 2002. Stochastic modeling of usage patterns in a web-based information system. *Journal of the Association for Information Science and Technology* 53, 7 (2002), 536–548.
[2] Yu Chi, Tingting Jiang, Daqing He, and Rui Meng. 2017. Towards an integrated clickstream data analysis framework for understanding web users' information behavior. *iConference 2017 Proceedings* (2017).
[3] Tony Russell-Rose, Paul Clough, and Elaine G Toms. 2014. Categorising search sessions: some insights from human judgments. In *Proceedings of the 5th Information Interaction in Context Symposium*. 251–254.