Log File Analysis and Mining

Maarten de Rijke

ISLA, University of Amsterdam, Science Park 904
1098 XH Amsterdam, The Netherlands
derijke@uva.nl

Information retrieval is no longer just about matching the content of queries to the content of documents. For nearly two decades, links and link structure have been brought to bear on the information retrieval problem in a web setting. During the past five years, as part of the development of information retrieval algorithms, content and link analysis are increasingly being complemented with insights gleaned from observation of people and of people's interactions with information and the search engine.

One of the ways in which user search behaviors may be analyzed is through a transaction log analysis, which over the years has proved an apt method for the characterization of user behavior. Its strengths include its non-intrusive nature — the logs are collected without questioning or otherwise interacting with the user — and the large amounts of data that can be used to generalize over the cumulative actions taken by large numbers of users. It is important to note that transaction log analysis faces limitations: not all aspects of the search can be monitored by this method, for example, the underlying information need. It can also be difficult to compare across transaction log studies of different systems due to system dependencies and varying implementations of analytical methods. Comparability can be improved to some extent by providing clear descriptions of the system under investigation and the variables used.

Information retrieval has a long history of transaction log analysis, from early studies of the logs created by users of library online public access catalog systems to later studies of the logs of Web search engines. This was followed by the analysis of more specialized search engines and their transaction logs. For instance, authors have studied the behavior of users of a blog search engine through a log file analysis and examined the difference between the vocabularies of queries, social bookmarking tags, and online documents. Three frequently used units of analysis have emerged from the body of work: the *session*, the *query*, and the *term*, though the definition of each unit may vary across studies.

In the tutorial, I will provide a number of examples of log file studies as well as the type of knowledge that can be obtained by studying log files: about people's information behavior, about experimental evaluation of search engines, and about online optimizations of search engines.

The tutorial is based on joint work with Richard Berendsen, Katja Hofmann, Bouke Huurnink, Bogomil Kovachev, Edgar Meij, Gilad Mishne, Evangelia-Paraskevi Nastou, Wouter Weerkamp, and Shimon Whiteson.

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,

Nicola Ferro and Letizia Tanca (Eds.): SEBD 2012, Edizioni Libreria Progetto, Padova, Italia
ISBN: 978-88-96477-23-6,