Multilingual Log Analysis: LogCLEF

Giorgio Maria Di Nunzio[1], Johannes Leveling[2], and Thomas Mandl[3]

[1] Department of Information Engineering – University of Padua – Italy
dinunzio@dei.unipd.it
[2] Centre for Next Generation Localisation (CNGL), Dublin City University, Ireland
jleveling@computing.dcu.ie
[3] Information Science, University of Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. The current lack of recent and long-term query logs makes the verifiability and repeatability of log analysis experiments very limited. A first attempt in this direction has been made within the Cross-Language Evaluation Forum in 2009 in a track named LogCLEF which aims to stimulate research on user behaviour in multilingual environments and promote standard evaluation collections of log data. We report on similarities and differences of the most recent activities for LogCLEF.

1 Introduction

Log data containing interactions between users and information systems is important for different research communities. Computer science researchers could study and analyze new algorithms via a common benchmark search log, learn about user information needs and query formulation approaches. Social scientists could investigate the use of language in queries as well as discrepancies between user interests as revealed by their queries versus their interests expressed in face-to-face surveys. Advertisers could use interaction logs to better understand how users navigate to their pages and improve keyword advertising campaigns [1].

Recently, researchers have been addressing the problem of the availability and use of log data: how log files should be made publicly available to researchers, whether log data should be gathered for specific tasks, whether there is value in general log data, and how additional information can be gathered and correlated with query log data [2]. The current lack of recent and long-term data makes the verifiability and repeatability of experiments very limited. It is practically impossible to find two works on the same dataset unless by the same author, or at least one of the authors worked for a commercial search engine company.

2 LogCLEF

A first attempt to release a collection of log data with the aim of verifiability and repeatability was done within the Cross-Language Evaluation Forum (CLEF) in 2009 in a track named LogCLEF which is an evaluation initiative for the analysis

Table 1. Log file resources at CLEF

Year Origin	Size Type	Year Origin	Size Type
2009 Tumba!	350K queries query log	2010 TEL	2.6M records activity log
2009 TEL	1.87M records activity log	2010 TEL	1.5 GB (zipped) web server log
		2010 DBS	5 GB web server log

of queries and other user activities [3]. An important long-term aim of the track is to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data. In the first two LogCLEF editions, different data sets have been distributed to the participants: search engine query and server logs from the Portuguese search engine Tumba! and from the German EduServer (Deutscher Bildungsserver (DBS)); and digital library systems query and server logs from The European Library (TEL). Table 1 summarizes the log resources and the relative sizes.

In particular, the analyses of the TEL logs are challenging given the nature of the the service. TEL is a free service that offers access to the resources of 48 national libraries of Europe in 35 languages. It aims to provide a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage. Resources can be both digital (e.g. books, posters, maps, sound recordings, videos) and bibliographical and the quality and reliability of the documents are guaranteed by the 48 collaborating libraries.

LogCLEF 2009 participation and results. Four groups participated in LogCLEF 2009. A thorough analysis of query reformulation, query length and activity sequence was carried out by Ghorab et al. [4]. The group showed that many query modification operations concern the addition or the removal of stopwords. These actions only have an effect for the language collection in which the word is a stop word. The ultimate goal is the understanding of the behavior of users from different linguistic or cultural backgrounds. The application of activity sequences for the identification of communities is also explored. The analysis revealed the most frequent operations as well as problems with the user interface of TEL. Lamm et al. analyzed sequences of interactions within the log file. These were visualized in an interactive user interface which allows the exploration of the sequences [5]. In combination with a heuristic success definition, this system lets one identify typical successful activity sequences. This analysis can be done for users from one top level domain. A few differences for users from different countries were observed but more analysis is necessary to reveal if these are real differences in behavior. In addition, issues with the logging facility were identified.

LogCLEF 2010 participation and results. In 2010, seven groups participated (five of which were newcomers).The major topics of interest at LogCLEF 2010 were named entities in queries, language identification (LI) of queries, determining successful searches, and comparing search behaviour between web search and

search in TEL data. Bosca et al. [6] experimented on LI in queries. They found that LI is a difficult task because of missing context in queries and that named entities can lead to misclassifications. For example, "Mozart" may be classified as a German query, but can also be a query in many other different languages. They concluded that LI for queries should be different from LI for documents. The experiments were performed on a manually annotated subset of 100 queries. Stiller et al. [7] analyzed and manually annotated a subset of 510 queries from the TEL data. They found that more than half of the queries are for named entities, which has a huge impact on correct LI of queries. The query language could often not even be manually determined or disambiguated, because many proper nouns are not translated between different languages. They report that seven of the ten most frequent queries contain named entities and that 167 out of 279 named entity queries are ambiguous. Takaku et al. [8] performed an analysis of search sessions and click ranks. They viewed sessions as sequences of actions and durations and compare actions with web search log actions. Leveling et al. [9] investigated the relation between query language, interface language and user IP address. They showed that these aspects correlate and this information can be used to automatically generate a ranking of document collections that better reflects user preferences. In addition they examined query performance indicators for web search and applied them to queries in sessions to find out if performance of user queries increases over time. They found that there are only few consistent changes in consecutive queries on the same topic. However, the first query in a session seems to indicate behaviour in the remainder of the search session: long initial queries seem to be improved by removing terms, while initially short queries will be expanded, Lana-Serrano et al. [10] defined successful queries as queries with results and user interactions on these results. They reported that choosing the native language as the interface language does not affect the success rate of queries. Verberne et al. [11] investigated search behaviour for users of the TEL portal in comparison to ad-hoc searchers using MSN services. The queries do not differ much in average length, but they differ in the topics of interest and in the diversity of languages (mono- vs. multilingual search). In contrast to the TEL data, Web search logs contain a high fraction of navigational and transactional queries while the most frequent TEL queries contain named entities. They also investigated intra-session search behaviour. Perea-Ortega et al. [12] performed a brief analysis of the TEL data, reconstructing user sessions. They analyze TEL queries with a focus on multilingual search and report that nine major European languages cover 95% of all sessions, with 84% of the queries in English.

Acknowledgments

This work has been partially supported by the PROMISE network of excellence (contract n. 258191) project, as part of the 7th Framework Program of the European Commission.

References

1. Korolova, A., Kenthapadi, K., Mishra, N., Ntoulas, A.: Releasing search queries and clicks privately. In Quemada, J., León, G., Maarek, Y.S., Nejdl, W., eds.: WWW, ACM (2009) 171–180
2. Clough, P., Berendt, B.: Report on the TrebleCLEF query log analysis workshop 2009. SIGIR Forum **43** (2009) 71–77
3. Mandl, T., Agosti, M., Di Nunzio, G.M., Yeh, A.S., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: The CLEF 2009 multilingual logfile analysis track overview. [13] 508–517
4. Ghorab, M.R., Leveling, J., Zhou, D., Jones, G.J.F., Wade, V.: Identifying common user behaviour in multilingual search logs. [13] 518–525
5. Lamm, K., Mandl, T., Koelle, R.: Search path visualization and session performance evaluation with log files. [13] 538–543
6. Bosca, A., Dini, L.: Language identification strategies for cross language information retrieval. [14] 100
7. Stiller, J., Gäde, M., Petras, V.: Ambiguity of queries and the challenges for query language detection. [14] 99
8. Takaku, M., Egusa, Y., Saito, H., Kando, N., Terai, H., Miwa, M.: CRES at LogCLEF 2010: Towards understanding the user behaviors through an analysis of search sessions, search units and click ranks. [14] 103
9. Leveling, J., Ghorab, M.R., Magdy, W., Jones, G.J.F., Wade, V.: DCU-TCD@LogCLEF 2010: Re-ranking document collections and query performance estimation. [14] 101
10. Lana-Serrano, S., Villena-Román, J., Cristóbal, J.C.G.: DAEDALUS at LogCLEF 2010: Analyzing the success of search queries. [14] 102
11. Verberne, S., Hinne, M., van der Heijden, M., Hoenkamp, E., Kraaij, W., van der Weide, T.P.: How does the library searcher behave? A contrastive study of library search against ad-hoc search. [14] 99
12. Perea-Ortega, J.M., Ráez, A.M., Cumbreras, M.A.G., López, L.A.U.: SINAI at LogCLEF 2010. [14] 100
13. Peters, C., Nunzio, G.M.D., Kurimo, M., Mostefa, D., Peñas, A., Roda, G., eds.: Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers. In Peters, C., Nunzio, G.M.D., Kurimo, M., Mostefa, D., Peñas, A., Roda, G., eds.: CLEF. Volume 6241 of Lecture Notes in Computer Science., Springer (2010)
14. Braschler, M., Harman, D., Pianta, E., eds.: CLEF 2010 Labs and Workshops. Abstracts Notebook Papers . 22-23 September 2010, University of Padua, Italy. In Braschler, M., Harman, D., Pianta, E., eds.: CLEF 2010. (2010)