LogCLEF: Enabling Research on Multilingual Log Files

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

[1] Centre for Next Generation Localisation (CNGL)
School of Computing, Dublin City University, Dublin 9, Ireland
jleveling@computing.dcu.ie
[2] Department of Information Engineering – University of Padua
Via Gradenigo, 6/a – 35131 Padua – Italy
dinunzio@dei.unipd.it
[3] Information Science,
University of Hildesheim, Germany
mandl@uni-hildesheim.de

1 Overview

Interactions between users and information access systems can be analyzed and studied to gather user preferences and to learn what a user likes the most, and to use this information to adapt the search to users and personalize the presentation of results. The LogCLEF lab - "A benchmarking activity on Multilingual Log File Analysis: Language identification, query classification, success of a query" deals with information contained in query logs of search engines and digital libraries from which knowledge can be mined to understand search behavior in multilingual context. LogCLEF has created the first long-term standard collection for evaluation purposes in the area of log analysis. The LogCLEF 2011 lab is the continuation of the past two editions: as a pilot task in CLEF 2009, and a workshop in CLEF 2010. The Cross-Language Evaluation Forum (CLEF) promotes research and development in multilingual information access and is an activity of the PROMISE Network of Excellence.

2 Topic and Goal

Log data constitute a relevant aspect in the evaluation process of the quality of a search engine and the quality of a multilingual search service; log data can be used to study the usage of a search engine, and to better adapt it to the objectives the users were expecting to reach. Research on log files is an area with obstacles. Most search services, especially large search engines do exploit their search logs but do not grant public access to them. A standard evaluation resource will make systems comparable and make research more transparent. More groups who do currently not have access to real world log files can work in the field of log file analysis.

The research goal of LogCLEF is the analysis and classification of queries, the definition of success of a search in order to understand search behaviour in multilingual contexts. One of the aims of LogCLEF is to foster the exchange of systems and components as well as the sharing of heterogeneous annotation of log data in order to advance the state of the art in this research area.

3 Tasks

LogCLEF 2011 offers three tasks, based on the exchange of ideas and proposals among the participants during the last LogCLEF 2010 workshop:

1. Language identification: participants are required to identify the actual language of queries. Annotated resources manually generated by participants of previous editions form a basic set of ground-truth data. This ground truth will be used to evaluate the automatic language recognition algorithms.
2. Query classification: participants are required to annotate each query with a label which represents a category of interest, e.g. PERSON (i.e. Leonardo Da Vinci) or EVENT (i.e. Revolución francesa).
3. Success of a query: participants are required to study the trend of the success of a search. The success can be defined in terms of time spent on a page, number of clicked items, or actions performed on the result list.

4 Log Datasets

LogCLEF 2011 offers three datasets to participants.

1. The European Library (TEL) dataset, consisting of three years and a half of log data, with more than three million records of user interactions with the TEL portal. The TEL portal provides search across all national libraries and other digital libraries in Europe. Naturally, the dataset is multilingual because TEL has users from all European countries and thus, queries to the TEL portal cover all major European languages.
2. The Sogou dataset, which contains queries to the Chinese Sogou search engine. The data contains a user ID, the query, URL in the result ranking, and user click information. The queries are mostly in Chinese, but certain parts can be written in other languages. For example, proper nouns can be transcribed into English.
3. A dataset from the German Education Server (Deutscher Bildungserver, DBS). The DBS is a clearinghouse for educational resources on the Web. It also contains content as well as descriptions and reviews on Web sites on education. The DBS web logs are server logs in standard format in which the searches and the results viewed can be observed. The logs allow to observe two types of user queries: queries in search engines (in the referrer when DBS files were found using a search engine) and queries within the DBS. This data allows studying search behaviour on a single web server and in a specific domain.

LogCLEF 2011 is one of the benchmarking activities of CLEF 2011 that will take place in Amsterdam in September 2011.

Acknowledgements

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, and by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (`http://www.cngl.ie`).