LogCLEF 2011 Multilingual Log File Analysis: Language identification, query classification, and success of a query

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

[1] Department of Information Engineering – University of Padua
Via Gradenigo, 6/a – 35131 Padua – Italy
dinunzio@dei.unipd.it

[2] Centre for Next Generation Localisation (CNGL)
School of Computing, Dublin City University, Dublin 9, Ireland
jleveling@computing.dcu.ie

[3] Information Science,
University of Hildesheim, Germany
mandl@uni-hildesheim.de

Abstract. Since 2009 LogCLEF has been the initiative within the Cross-Language Evaluation Forum which aims at stimulating research on user behavior in multilingual environments and promote standard evaluation collections of log data. During these editions of LogCLEF, different collections of log dataset were distributed to the participants together with manually annotated query records to be used as a training or test set. In this LogCLEF 2011 edition, a Web based interface to annotate log data was designed and realized on the basis on the experience of past participants for different tasks: language identification, query classification, and query drift. The data and the results produced by the participants are analyzed and discussed.

1 Introduction

Interactions between users and information access systems can be analyzed and studied to gather user preferences and to learn what the user likes the most, and to use this information to personalize the presentation of results. Search logs are a means to study user information needs and preferences. The literature of log analysis of information systems shows a wide variety of approaches to learn user preferences by looking at implicit or explicit interaction [1]. However, there has alway been a lack of availability and use of log data for research experiments which makes the verifiability and repeatability of experiments very limited. It is very difficult to find two research works on the same dataset unless by the same author, or where none of the authors worked for a commercial search engine company. This is not only a question of the same data source, but also a problem of using the same period of time for the analysis if the analysis has to be comparable with other works.

Table 1. Log file resources at LogCLEF

Year	Origin	Size	Type
2009	Tumba!	350,000 queries	Query log
2009	TEL	1,900,000 records	Query and activity log
2010	TEL	760,000 records	Query and activity log
2010	TEL	1.5 GB (zipped)	Web server log
2010	DBS	5 GB	Web server log
2011	TEL	950,000 records	Query and activity log
2011	Sogou	1.9 GB (zipped)	Query log

LogCLEF[4] is an evaluation initiative for the analysis of queries and other logged activities used as an expression of user behavior [2, 3]. An important long-term aim of the LogCLEF initiative is to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data. Since 2009, within the Cross-Language Evaluation Forum (CLEF)[5], LogCLEF has been releasing collections of log data with the aim of verifiability and repeatability of experiments. In the three years of LogCLEF editions, different data sets have been distributed to the participants: search engine query and server logs from the Portuguese search engine Tumba![6] and from the German EduServer[7] (Deutscher Bildungsserver: DBS); digital library systems query and server logs from The European Library[8] (TEL); and Web search engine query logs of the Chinese search engine Sogou[9]. Table 1 summarizes the log resources and the relative sizes.

In each edition of LogCLEF, participants are required to:

- process the complete logs;
- make publicly available any resources created based on these logs;
- find out interesting issues about the user behavior as exhibited in the logs; and
- submit results in a structured file.

The public distribution of the datasets as well as the results and the exchange of system components aim at creating of a community in order to advance the state of the art in this research area.

2 Task Definition

The definition of tasks in LogCLEF changed year by year according to the discussions together with the participants. In the first year participants were free to

[4] http://www.promise-noe.eu/mining-user-preference

[5] http://www.clef-campaign.org/

[6] http://www.tumba.pt/ (offline)

[7] http://www.eduserver.de/

[8] http://www.theeuropeanlibrary.org/

[9] http://www.sogou.com/

investigate any hypothesis on the data sets and send their results. In the second year though, the task was a bit more structured with the following suggestions:

1. language identification for the queries;
2. initial language vs. country IP address;
3. subsequent languages used on same search;
4. country of the library vs. language of the query vs. language of the interface.

The LogCLEF 2011 Lab presents four different tasks which tackle some of the issues presented in this work:

- Language identification task: participants are required to recognize the actual language of the query submitted.
- Query classification: participants are required to annotate each query with a label which represents a category of interest. The proposed set of category of interest is:
  - Person (including names, institutions and organizations);
  - Geographic (including geographical entities);
  - Event (historical events);
  - Work title (including work titles and other works such as paintings);
  - Domain specific (technical terms often Latin);
  - Number (including ISBN and dates);
  - Topical (queries which cannot be assigned to the other categories);
  - Undecided.
- 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, actions performed during the browsing of the result list.
- Query re-finding: when a user clicks an item following a search, and then later clicks on the same item via another search; Query refinement: when a user starts with a query and then the following queries in the same session are a generalization, specification, or shift of the original one.

3 Data description

3.1 Log datasets

Three different log datasets were distributed to the participants in this LogCLEF edition:

- search engine query and server logs from the German EduServer (Deutscher Bildungsserver: DBS);
- digital library systems query and server logs from The European Library (TEL);
- Web search engine query logs of the Chinese search engine Sogou.

The summary of these resources in terms of size and number of records is included in Table 1.

EduServer. The DBS EduServer logs are server logs in standards format in which the searches and the results viewed can be observed and the data have been anonymized by partially obscuring the IP addresses of users. The two upper levels of server names or IP addresses have been hashed. This allows the reconstruction of sessions within the data. Note that accesses by search engine bots are still contained within the logs. 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);
- queries within the DBS (see query parameters in metasuche/qsuche).

The logs also allow to observe the browsing behavior within the DBS server structure and to access two types of content and compare them to the queries: the descriptions of the educational web sites within DBS, the content of the educational web sites themselves (which might have changed since the logs have been collected) in those cases where the user might have accessed them.

The logs were collected in the time between September and November of 2009.

TEL dataset. The TEL search/action logs are stored in a relational table and contain different types of actions and choices of the user. Each record represents a user action and the most significant fields: A numeric id, for identifying registered users or “guest” otherwise;

- User’s IP address;
- An automatically generated alphanumeric, identifying sequential actions of the same user (sessions);
- Query contents;
- Name of the action that a user performed;
- The corresponding collection’s alphanumeric id;
- Date and time of the action’s occurrence.

Three years and a half of log data will be released:

- January 2007-June 2008, 1,900,000 records (distributed at LogCLEF 2009)
- January 2009-December 2009, 760,000 records (distributed at LogCLEF 2010)
- January 2010-December 2010, 950,000 records (to be distributed at LogCLEF 2011)

Sogou dataset. The Sogou query logs (SougouQ) contain queries to the Chinese search engine Sogou[10] and were provided by the Tsinghua-Sogou Joint lab of Search Technology. The data contains:

- a user ID,
- the query terms,
- URL in the result ranking, and
- user click information.

The data covers one month of web search logs from June 2008.

[10] http://www.sogou.com/labs/dl/q.html

3.2 Annotated data

Another important aim of LogCLEF is to distribute ground truth generated manually or automatically by participants themselves. In CLEF 2010 the research teams of Humboldt University of Berlin and CELI s.r.l. prepared annotations for a small subset of the TEL query logs. The annotated data contains the following data:

- manual annotations for 510 query records about query language and category of the query;
- automatic annotations for 100 query records about query language.

In the current LogCLEF edition at CLEF 2011, an interface for query log annotation was designed and implemented by University of Padua[11] by gathering requirements from both LogCLEF participants and organizers (University of Padua, Dublin City University, University of Hildesheim, University of Amsterdam, Humboldt University of Berlin). The aim of this interface is to involve participants and researchers in the creation of manually annotated datasets that can be used by to test automatic systems.

A short guide for annotating query records was given to the participants. The guide consisted in the following four points which corresponded to the four steps of the query annotation interface:

1. annotate the language of the query; use undecided for a query whose language is ambiguous (example, mozart), use unknown if you don't know/recognize the language at all
2. annotate the language of the query knowing the language of the interface; in most cases the default language is English. Does the information of the language of the interface of the user change your mind or help to understand the language of the query?
3. annotate the change of the query/topic within a session; use "same query" if the text of the query didn't change at all, use "generalization" if the user changed the initial query to a broader query (mozart piano sheets → mozart music), use "specification" if the user changed the query from a wider one to a more narrow query (beethoven → beethoven sonata and symphony), use "drifting" if the user changed the topic of the initial query (mozart childhood → mozart musical style), use "more than two different queries" if the session contains many different queries, use "not applicable" if none of the options are applicable.
4. annotate the query with one or more categories (Person, Geographic, etc.).

During the LogCLEF 2011 the following manually annotated data have been produced and distributed to the research teams:

- 723 annotated query record with language, query session, and query category.

[11] http://ims.dei.unipd.it/websites/LogCLEF/Logs/login.php

Moreover, a baseline for comparing the systems developed by the participants has been generated using an automatic open source software for language recognition[12]. A total of 940,957 annotated query records with languages have been created and distributed to the research teams.

4 Participation and Results

As shown in Table 2, a total of 4 groups out of 17 registered participants submitted results for LogCLEF. The results of the participating groups are reported in the following section and elaborated in the papers of the participants. All groups analyzed the TEL logs, one participants analyzed the DBS logs, none presented analyses on the Sogou logs.

Table 2. LogCLEF 2011 participants.

Participant	Institution	Country
DAEDALUS	Universidad Politécnica de Madrid Universidad Carlos III de Madrid DAEDALUS - Data, Decisions and Language, S.A.	Spain
UBER-UvA	Humboldt University of Berlin University of Amsterdam	Germany The Netherlands
CUZA	"Alexandru Ioan Cuza" University	Romania
ESSEX	University of Essex	United Kingdom

DAEDALUS [4] focused on the following specific objectives: analyzing if there is any measurable effect on the success of the search queries if the native language and the interface language chosen by the user are different; to study in detail the user context and his interaction with the system in the case of sessions with a successful operation over the same resource; to discover any relation among the user native language, the language of the resource involved and the interaction strategy adopted by the user to find out such resource. The analysis of the data showed that, in general for all languages, the fact that the native language of the user matches or not the interface language does not have apparently any impact on the success rate of the search queries.

UBER and UvA [5] investigated multilingual user behavior in terms of different aspects such as the native language of the user, the preferred retrieval language of the user, the interface language, the query language, and so on. They also presented some practical issues concerning collecting language indicators from the IP address and the text of the query. Some of the analysis

[12] http://code.google.com/p/language-detection/

concerned the study of the success rate of a search compared to the language of the interface and the nationality of the user. A different analysis was also performed by studying the interface language switch (from the default English language to another language). By comparing the actions conducted before and after the first interface language change they observed that the frequency of any particular action related to success increases after the language change, however, the frequency distribution of actions does not change in general.

CUZA [6] presented a study of the applicability for language identification tasks in which the text is very short like a query, and they also discussed some issues and some methods to overcome the problems related to short queries. A first issue was the significant number of queries for which the language was unknown or undecided. They experimented language identification by using an N-grams probabilistic classifier together with alphabet diacritics recognition to partially solve the problem of noisy data.

ESSEX [7] is the only group in this edition of LogCLEF who analyzed two different datasets: the DBS EduServer logs and the TEL logs. They first presented a comparison of the two datasets in terms of the number of total and distinct queries, the number of sessions and the single query sessions. Then, they discussed a method for query suggestion named Ant Colony Optimisation to build query association graphs from the query logs. The directed association graph is used for query recommendation by starting from the query node in question, and then traversing the graph edges to identify and rank associated query nodes using the weights on the edges. The authors also explored the effect of query suggestions in reducing the number of steps required by the user to achieve their goals.

5 Conclusions

- For LogCLEF 2011, annotated logs files have been made available to participants and interested researchers.
- Although the number of registered participants has reached a new high for this edition of LogCLEF, four groups participated in LogCLEF. This may be due to the fact that compared to previous editions, this time the task was more restrictive. The difficulty of the tasks will be discussed during the lab to understand better how to design tasks more accurately in the future.

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

References

1. Agosti, M., Crivellari, F., Di Nunzio, G.: Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. Data Mining and Knowledge Discovery (2011) 1–34
2. Mandl, T., Agosti, M., Di Nunzio, G., Yeh, A., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: the CLEF 2009 Cross-Language Logfile Analysis Track Overview. In Peters, C., Di Nunzio, G., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G., eds.: Multilingual Information Access Evaluation Vol. I. Text Retrieval Experiments: Proceedings 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece. LNCS, Springer (2010)
3. Di Nunzio, G.M., Leveling, J., Mandl, T.: Multilingual log analysis: LogCLEF. In Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Murdoch, V., eds.: Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings. Volume 6611 of Lecture Notes in Computer Science., Springer (2011) 675–678
4. Lana-Serrano, S., Villena-Román, J., González-Cristóbal, J.C.: DAEDALUS at LogCLEF 2011: Analyzing Query Success and User Context. In: This volume. (2011)
5. Gäde, M., Stiller, J., Berendsen, R., Petras, V.: Interface Language, User Language and Success Rates in The European Library. In: This volume. (2011)
6. Gînscă, A.L., Boroş, E., Iftene, A.: Adapting Statistical Language Identification Methods for Short Queries. In: This volume. (2011)
7. Albakour, M.D., Kruschwitz, U.: University of Essex at LogCLEF 2011: Studying query refinement. In: This volume. (2011)