LogCLEF 2009: the CLEF 2009 Multilingual Logfile Analysis Track Overview

Thomas Mandl[1], Maristella Agosti[2], Giorgio Maria Di Nunzio[2], Alexander Yeh[3], Inderjeet Mani[3], Christine Doran[3], Julia Maria Schulz[1]

[1] Information Science, University of Hildesheim, Germany
{mandl,schulzju}@uni-hildesheim.de
[2] Department of Information Engineering, University of Padua, Italy
{agosti,dinunzio}@dei.unipd.it
[3] MITRE Corporation, Bedford, Massachusetts, USA
{asy,imani,cdoran}@mitre.org

Abstract

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.

The interest in multilingual log analysis was promoted by the Cross Language Evaluation Forum (CLEF) for the first time with a track named LogCLEF. LogCLEF is an evaluation initiative for the analysis of queries and other logged activities as expression of user behavior. The goal is the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. Two tasks were defined: Log Analysis and Geographic Query Identification (LAGI) and Log Analysis for Digital Societies (LADS). Five groups using a variety of approaches submitted experiments. The data for the track, the evaluation methodology and some results are presented.

Categories and Subject Descriptors

H.1 [MODELS AND PRINCIPLES] H.1.2 User/Machine Systems
H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software.

General Terms

Measurement, Performance, Experimentation

Keywords

Logfile Analysis, Log data, User Behavior, Multilingual Information Retrieval Evaluation Benchmarks.

1 Introduction

Log is a concept commonly used in computer science; in fact, log data are collected by an operating system to make a permanent record of events during the usage of the operating system itself. This is done to better support its operations, and in particular its recovery procedures. Log data are also collected by many applications systems that manage permanent data, among the more relevant ones there are the database management systems (DBMS) that support different types of collection of log data, one of these types is the Write Ahead Log (WAL) that is used for the management and the recovery of transactions. Due to the experience gained in the management of operating systems and the many other application systems that manage permanent data, log

procedures are commonly put in place to collect and store data on the usage of application systems by its users. Initially, these data were mainly used to manage recovery procedures of an application system, but over time it became apparent that they could also be used to study the usage of the application by its users, and to better adapt the system to the objectives the users were expecting to reach.

Like with an operating system and any other software application, log data can be collected during the use of a search engine to monitor its functioning and usage by final and specialized users, which means recording log data to study its use and to consolidate it in a tool that meets end-user search requirements. This means that log data constitute a relevant aspect in the evaluation process of the quality of a search engine and the quality of a multilingual search services; 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.

The interest in multilingual log analysis was promoted by the Cross Language Evaluation Forum (CLEF)[1] for the first time with a track named LogCLEF[2]. LogCLEF is a new track at the Cross Language Evaluation Forum (CLEF) which wants to stimulate research on user behavior in multilingual environments. LogCLEF is developing standard evaluation collections which support long-term research. LogCLEF is an evaluation initiative for the analysis of queries and other logged activities as expression of user behavior. The goal is the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems.

The datasets used in 2009 for the analysis of search logs were derived from the use of the Tumba! Search engine and The European Library (TEL) Web site. Two tasks were defined: Log Analysis and Geographic Query Identification (LAGI) and Log Analysis for Digital Societies (LADS). LAGI required the identification of geographical queries within logs from the Tumba! Search engine and The European Library multilingual information system. LADS intended to analyze the user behavior in the multilingual information system of The European Library. Five groups using a variety of approaches submitted experiments.

The data for the track, the evaluation methodology and some results are presented in this overview paper together with a description of the two sub tasks of LogCLEF 2009.

2 Log Analysis and Geographic Query Identification (LAGI)

The identification of geographic queries within a query stream and the recognition of the geographic component are key problems for geographic information retrieval (GIR). Geographic queries require specific treatment and often a geographically oriented output (e.g. a map). The task would be to (1) classify geographic queries and (2) identify their geographic content. The task design and evaluation measures would be similar to the ones used in the track in 2007 [10, 11].

LAGI 2009 is a task to identify geographic elements in query search logs. Search logs were obtained from the following two search systems:

1. Tumba!, a Portuguese web search engine[3];
2. The on-line catalogue of The European Library (TEL), where an English subset of the TEL queries was used[4].

The Tumba! log files were manually reviewed. All references to domain searches and emails were replaced by 'ZZZ' in order to preserve anonymity. Examples for entries in the log file are shown in Tables 1a to 1d.

Table 1a: Original queries from the TEL log file

```
875336 & 5431 & ("central europe")
828587 & 12840 & ("sicilia")
902980 & 482 &  (creator all "casanova")
196270 & 5365 & ("casanova")
906474 & 15432 & casanova
528968 & 190 &  ("iceland*")
470448 & 8435 & ("iceland")
712725 & 5409 & ("cavan county ireland 1870")
```

[1] http://www.clef-campaign.org/
[2] http://www.uni-hildesheim.de/logclef/
[3] http://www.tumba.pt/
[4] http://www.theeuropeanlibrary.org/

```
671397 & 14093 & ("university")
```

Table 1b: Annotated queries from the TEL log file

```
875336 & 5431 & ("<place>central europe</place>")
828587 & 12840 & ("<place>sicilia</place>")
902980 & 482 &  (creator all "casanova")
196270 & 5365 & ("casanova")
906474 & 15432 & casanova
528968 & 190 &  ("<place>iceland</place>*")
470448 & 8435 & ("<place>iceland</place>")
712725 & 5409 & ("<place>cavan county ireland</place> 1870")
671397 & 14093 & ("university")
```

Table 1c: Original queries from the Tumba! log file

```
4333825 @ 4777 @ "administração escolar"
4933229 @ 7888 @ "escola+hip+hop"
39283 @ 62180 @ chaves
2290106 @ 19398 @ CHAVES
1420489 @ 20564 @ Chaves
6971716 @ 106342 @ jornais de leiria
8403308 @ 83318 @ escolas de marinheiro
```

Table 1d: Annotated queries from the Tumba! log file

```
4333825 @ 4777 @ "administração escolar"
4933229 @ 7888 @ "escola+hip+hop"
39283 @ 62180 @ <place>chaves</place>
2290106 @ 19398 @ <place>CHAVES</place>
1420489 @ 20564 @ <place>Chaves</place>
6971716 @ 106342 @ jornais de <place>leiria</place>
8403308 @ 83318 @ escolas de marinheiro
```

The geographic elements to be marked are either those found in a gazetteer or places related to those found in the gazetteer. So if a city is listed in a gazetteer, then related places include hospitals, schools, etc. associated with that city. The gazetteer used is a static version of Wikipedia (Portuguese version for Tumba!, English version for the TEL subset), due to its coverage and availability to participants. A static version is used because the live Wikipedia (http://pt.wikipedia.org, http://en.wikipedia.org) is constantly changing, and so altering what places are listed, etc.

Many query terms have both geographic and non-geographic senses. Examples include “casanova” and “ireland” in English and “chaves” in Portuguese. Queries have inconsistent capitalization and are often short. So it may not be clear which sense to use for such terms. Wikipedia is used to disambiguate such terms by preferring the first sense returned by a Wikipedia look-up. In our examples, a look-up[5] of “casanova” initially returns an article on a person, so it is deemed a non-place in ambiguous cases. A look-up of “Ireland” initially returns an article on an island, so it is deemed a place in ambiguous cases. Sometimes, the initial article returned does not have a clear preference. For example, a look-up of “chaves” returns a disambiguation page/article which lists both place (including a city) and non-place (including a television show) senses. In such situations, the term is deemed a place in ambiguous cases for the purposes of this evaluation. This method of disambiguation is used when a query has no indicated preference for which sense to use. But if the query indicates a preference for a sense, then that sense is what is used. An example is the query 'casanova commune'. A search for 'casanova commune' in the English Wikipedia does not return an article. Rather it returns a 'search' page (instead of being an article for some term, the page gives a ranked list of articles that contain parts of the candidate place term somewhere in the articles' text), which is ignored in this evaluation. For 'casanova', the English Wikipedia returns an article on a person named Casanova, so that is the default predominant sense. But that article has a link to a

[5] To look-up a term (not search for a term), type the term into the Wikipedia “search” (English) or “busca” (Portuguese) box and then click “Go” (English) or “Ir” (Portuguese).

disambiguation page, and the disambiguation page has a link to the place 'Casanova, Haute-Corse', which is a commune. This query indicates that this sense of 'casanova' is the preferred one for the query, so this overrides the default preferred sense based on the initial page returned by the Wikipedia.

There are still complications in look-ups. For one thing, it turns out that many queries have misspelled words, and judgments need to be made about these. Also, many terms exist in queries that turn out not to have a Wikipedia article about them. In addition, Wikipedia will sometimes prefer an unusual sense of a word over a more usual sense. Two examples are "de" in Portuguese and "Parisian" in English. "de" is a common preposition meaning something like "of" . For example, "jornais de leiria" loosely means "periodicals of leiria". A look-up of "de" returns a disambiguation page that mentions "de" possibly standing for Delaware or Deutschland (Germany), but does not mention "de" being a preposition. Various common meanings for "Parisian" include a person from Paris or something in or associated with Paris. But a look-up of "Parisian" first returns an article about a chain of department stores in the US with the name "Parisian". We dealt with these complications by adding to the task guidelines and removing queries that could not be handled by the guidelines.

Beyond look-up complications, there were also complications with Wikipedia software. It turns out that installing a static version of Wikipedia is hard. One group made an unsuccessful attempt and another group could install it well enough to support the evaluation, but there were still complications. The successful installation was served over the Internet for others. Besides being hard to install, a Wikipedia is somewhat slow as it takes quite a bit of computational resources to run and Internet congestion can considerably add to the response time.

These complications of Wikipedia combined with difficulties in obtaining the search logs until late in the CLEF campaign delayed the dataset annotation and constrained its size: there was only enough annotated data to produce a small test set (and no training set). The TEL test set has 108 queries with 21 places annotated. The Tumba! test set has 146 queries and 30 to 35 places annotated[6].

A total of two runs were submitted, both by the same group at the "Alexandru Ioan Cuza" University in Romania [8]. The two runs used different resources (1. GATE, 2, Wikipedia) for finding places. Overall, precision in finding places turned out to more of challenge than recall. The recall scores ranged from 33% to 76%, while the precision scores were 26% or less.

Table 2: Results of the LAGI sub task for the test set

Resource	TEL	Tumba! version A	Tumba! version B
Gate	R 33%, P 24%	R 51%, P 26%	R 50%, P 22%
Wikipedia	R 76%, P 16%	R 37%, P 9%	R 40%, P 8%

With both TEL and Tumba!, the Wikipedia resources produced much lower precision than the GATE resources. Using the Wikipedia resources often resulted in an entire query being annotated as a place. For Tumba!, the Wikipedia resources also produced worse recall, but for TEL, Wikipedia resources produced better recall.

In summary, this is a first run of the LAGI task for finding places in search queries. We obtained the use of two search query sets, Tumba! and TEL, and used Wikipedia as a gazetteer and for disambiguating between place and non-place senses of a term. Delays in obtaining the data sets and complications with using Wikipedia resulted in a delay in producing a dataset and also only a small test set being produced (no training set).

3 Log Analysis for Digital Societies (LADS)

The Log Analysis for Digital Society (LADS) task deals with logs from The European Library (TEL) and intends to analyze user behavior with a focus on multilingual search. TEL is a free service that offers access to the resources of 48 national libraries of Europe in 35 languages. Resources can be both digital (e.g. books, posters, maps, sound recordings, videos) and bibliographical. Quality and reliability are guaranteed by the 48 collaborating national libraries of Europe[7]. TEL aims to provide a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage.

[6] The Tumba! test set had a number of queries that could be annotated in two different ways. We made two versions where version A was annotated in one way (for 35 places) and version B the other way (for 30 places).

[7] http://www.theeuropeanlibrary.org/portal/organisation/about_us/aboutus_en.html

The quality of the services and documents TEL supplies are very important for all the different categories of users of a digital library system. Log data constitute a relevant aspect in the evaluation process of the quality of a digital library system and of the quality of interoperability of digital library services [1].

It is worth underlining that the access to each service a digital library system provides is usually supplied through a Web browser, and not through a specifically designed interface. This means that the analysis of user interaction with a system that has a Web-based interface requires the forecasting of ways that support the reconstruction of sessions in a setting, like the Web, where sessions are not naturally identified and kept [2].

3.1 Goals

Potential targets for experiments are query reformulation, multilingual search behavior and community identification. This task was open to diverse approaches, in particular data mining techniques in order to extract knowledge from the data and find interesting user patterns.

Suggested sub-tasks for the analysis of the log data are:

1. user session reconstruction; this step needs to be considered as a prerequisite to the following ones [3];
2. user interaction with the portal at query time; e.g. how users interact with the search interface, what kind of search they perform (simple or advanced), and how many users feel satisfied/unsatisfied with the first search and how many of them reformulate queries, browse results, leave the portal to follow the search in a national library;
3. multilinguality and query reformulation; e.g. what are the collections that are selected the most by users, how the language (country/portal interface) of the user is correlated to the collections selected during the search, how the users reformulate the query in one language or in a different language;
4. user context and user profile; e.g. how the study of the actions in the log can identify user profiles, how the implicit feedback information recorded in the logs can be exploited to create the context in which the user operates and how this context evolves.

Participants were required to:

- apply some algorithm to process the complete logs;
- resources created based on the logs (e.g. annotations of a small subsets) need to be made publicly available;
- find out interesting issues about the user behavior as exhibited in the logs;
- submit results in a structured file.

3.2 Data

The data used for the LADS task are search logs of The European Library portal; those logs are usually named "action logs" in the context of TEL activities. In order to better understand the type of those action log that have been distributed to the participant, an example of the possible usage of the portal is described in the following.

In TEL portal's home page, a user can initiate a simple keyword search with a default predefined collection list presenting catalogues from national libraries. From the same page, a user may perform an advanced search with Boolean operators and/or limit search to specific fields like author, language, and ISBN. It is also possible to change the searched collection by checking the theme categories below the search box. After search button is clicked the result page appears, where results are classified by collections and the results of the top collection in the list are presented with brief descriptions. Then, a user may choose to see result lists of other collections or move to the next page of records of current collection's results. While viewing a result list page a user may also click on a specific record to see detailed information about the specific record. Additional services may be available according to the record selected.

All these type of actions are logged and stored by TEL in a relational table, where a table record represents a user action. The most significant columns of the table are:

- 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.

Action logs distributed to the participants of the task cover the period from 1st January 2007 until 30th June 2008. The log file contains user activities and queries entered at the search site of TEL. Examples for entries in the log file are shown in Table 3.

Table 3: Examples from the TELlog (date has been deleted for readability)

```
id;userid;userip;sesid;lang;query;action;colid;nrrecords;recordposition;sboxid;objurl;date
892989;guest;62.121.xxx.xxx;btprfui7keanue1u0nanhte5j0;en;("plastics mould");view_brief;a0037;31;;;
893209;guest;213.149.xxx.xxx;o270cev7upbblmqja30rdeo3p4;en;("penser leurope");search_sim;;0;-;;;
893261;guest;194.171.xxx.xxx;null;en;(“magna carta”);search_url;;0;-;;;
893487;guest;81.179.xxx.xxx;9rrrtrdp2kqrtd706pha470486;en;("spengemann");view_brief;a0067;1;-;;;
893488;guest;81.179.xxx.xxx;9rrrtrdp2kqrtd706pha470486;en;("spengemann");view_brief;a0000;0;-;;;
893533;guest;85.192.xxx.xxx;ckujekqff2et6r9p27h8r89le6;fr;("egypt france britain");search_sim;;0;-;;;
```

3.3 Participants and Experiments

As shown in Table 4, a total of 4 groups submitted results for the LADS task. The results of the participating groups are reported in the following section.

Table 4. LogCLEF 2009 participants

Participant	Institution	Country
Sunderland	University of Sunderland	UK
TCD-DCU	Trinity College Dublin	Ireland
Info Science	University of Hildesheim	Germany
CELI s.r.l	CELI Research, Torino	Italy

3.4 Results of the LADS Task

The CELI research institute tried to identify translations of search queries [4]. The result is a list of pairs of queries in two languages. This is an important step in observing multilingual user behavior. Combined with session information, it is possible the check whether users translate their query within a session.

The group from the University of Sunderland argues that users rarely switch the query language during their sessions. They also found out that queries are typically submitted in the language of the interface which the user selects [12].

A thorough analysis of query reformulation, query length and activity sequence was carried out by the Trinity College, Dublin [6]. 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 University of Hildesheim analyzed sequences of interactions within the log file. These were visualized in an interactive user interface which allows the exploration of the sequences [9]. In combination with a heuristic success definition, this system lets one to identify typical successful activity sequences. This analysis can be done for users from one top level domain.

The design of future tasks is encouraged by a position paper from the University of Amsterdam. The authors argue that the limited knowledge about the user which is inherent in log files needs to be tackled in order to gain more context information. They argue for the semantic enrichment of the queries by linking them to digital objects [7].

4 Conclusions and Future Work

Studies on log files are essential for personalization purposes, since they implicitly capture user intentions and preferences in a particular instant of time. There is an emerging research activity about log analysis which tackles cross-lingual issues: extending the notion of query suggestion to cross-lingual query suggestion studying search query logs; leveraging click-through data to extract query translation pairs.
LogCLEF has provided an evaluation resource with log files of user activities in multilingual search environments: the Tumba! Search engine and The European Library (TEL) Web site. With these two different datasets, one related with searches in Web sites of interest to the Portuguese community and the other with searches for library catalogues in many European Libraries, it was possible to define two sub-tasks: Log Analysis and Geographic Query Identification (LAGI) and Log Analysis for Digital Societies (LADS).
A total of 4 groups submitted very diversified results: identify list of pairs of queries in two languages combined with session information; correlation between language of the interface and language of the query; activities at query time to study different user backgrounds.
The results and approaches of the participants to the 2009 campaign will be helpful to define a more formal task in another editions of the track. All participants of LogCLEF 2009 are invited to participate in the discussion of the future of LogCLEF.

Acknowledgments

The organization of LogCLEF was mainly volunteer work. At MITRE, work on this task was funded by the MITRE Innovation Program (public release approval case# 09-3188). We want to thank The European Library (TEL) and the Tumba! search engine for providing the log files. Many thanks especially to Nuno Cardoso from the XLDB Research Team at the University of Lisbon, who coordinated the anonymisation and the manual reviewing of the original logfile. At the University of Padua, work on this task was partially supported by the TELplus Targeted Project for digital libraries[8], as part of the eContent*plus* Program of the European Commission (Contract ECP-2006-DILI-510003) and by the TrebleCLEF Coordination Action[9], as part of the 7th Framework Programme of the European Commission, Theme ICT-1-4-1 Digital libraries and technology-enhanced learning (Grant agreement: 215231).

References

[1] Agosti, Maristella (2008): Log Data in Digital Libraries. In: Agosti, Maristella; Esposito, Floriana; Thanos, Costantino (Eds.): *Post-proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008)*. Padova, DELOS: an Association for Digital Libraries, July 2008, 115-121.

[2] Agosti, Maristella; Di Nunzio, Giorgio Maria. Gathering and Mining Information from Web Log Files. In: Thanos, Costantino; Borri, Francesca; Candela, Leonardo (Eds.). *Digital Libraries: Research and Development, First Int. DELOS Conference*, Pisa, Italy, February 13-14, 2007, Revised Selected Papers. Springer, Berlin/Heidelberg, Germany, LNCS 4877, ISBN 978-3-540-77087-9, 2007, 104-113.

[3] Agosti, Maristella; Di Nunzio, Giorgio Maria. Web Log Mining: A Study of User Sessions. *10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL 2007).* Corfù, Greece, 2007, 70-74.

[4] Bosca, Alessio; Dini, Luca (2009): CACAO project at the LogCLEF Track. *In this volume*.

[5] Di Nunzio, Giorgio Maria (2009): LogCLEF 2009 2009/03/02 v 1.0 Description of the The European Library (TEL) Search Action Log Files. http://www.uni-hildesheim.de/logclef/Daten/LogCLEF2009_file_description.pdf

[6] Ghorab, M. Rami; Leveling, Johannes; Zhou, Dong; Jones, Gareth; Wade, Vincent (2009): TCD-DCU at LogCLEF 2009: An Analysis of Queries, Actions, and Interface Languages. *In this volume*.

[7] Hofmann, Katja; de Rijke, Maarten; Huurnink, Bouke; Meij, Edgar (2009): A Semantic Perspective on Query Log Analysis. *In this volume*.

[8] Iftene, Adrian (2009): UAIC: Participation in LAGI Task. *In this volume*.

[9] Lamm, Katrin; Mandl, Thomas; Kölle, Ralph (2009): Search Path Visualization and Session Performance Evaluation with Log Files from The European Library (TEL) *In this volume*.

[8] http://www.theeuropeanlibrary.org/telplus/
[9] http://www.trebleclef.eu/

[10] Li, Zhisheng; Wang, Chong; Xing, Xie; Ma, Wie-Ying (2007): Query Parsing Task for GeoCLEF 2007 Report. In *Cross Language Evaluation Forum (CLEF) 2007 Working Notes*. http://www.clef-campaign.org/2007/working_notes/LI_OverviewCLEF2007.pdf

[11] Mandl, Thomas; Gey, Fredric; Di Nunzio, Giorgio; Ferro, Nicola; Larson, Ray; Sanderson, Mark; Santos, Diana; Womser-Hacker, Christa; Xing, Xie: GeoCLEF 2007: the CLEF 2007 Cross-Language Geographic Information Retrieval Track Overview. In: Peters, Carol; Jijkoun, Valentin; Mandl, Thomas; Müller, Henning Oard, Doug; Peñas, Anselmo; Petras, Vivien; Santos, Diana (Eds.): *Advances in Multilingual and Multimodal Information Retrieval: 8th Workshop of the Cross-Language Evaluation Forum. CLEF 2007, Budapest, Hungary, Revised Selected Papers*. Berlin et al.: Springer [Lecture Notes in Computer Science 5152] 2008

[12] Oakes, Michael; Xu, Yan (2009): LADS at Sunderland. *In this volume*.