SINAI at LogCLEF 2010

José M. Perea-Ortega, Arturo Montejo-Ráez, Miguel Á. García-Cumbreras,
and L. Alfonso Ureña-López

SINAI research group. Computer Science Department. University of Jaén
Campus Las Lagunillas, Ed. A3, E-23071, Jaén, Spain
{jmperea,amontejo,magc,laurena}@ujaen.es

Abstract. The SINAI[1] research group presents some results obtained after performing a brief analysis to the query logs from The European Library[2](TEL). The objective of the LogCLEF task is to analyze user behavior with a focus on multilingual search. The analysis carried out in this paper is related to the languages used in sessions, the number of interactions per session and the separability of sessions according to the words in the query. As a main conclusion, we can observe that after applying the Principal Component Analysis (PCA), just keeping two components over the different features extracted per session, the 95% of the variance of the data is preserved.

Keywords: Log File Analysis, Log Data, User Behavior, Cross-Language Information Retrieval

1 Introduction

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 expect to reach [2].

This is the first participation of the SINAI research group in the LogCLEF track. The goal of LogCLEF is the analysis of queries from different logs in order to understand the search behavior in multilingual contexts and to improve search systems. In 2010, LogCLEF provides a data collection which consists of a large logfile: **The European Library (TEL)**. This service provides access to several national libraries of Europe. In the TEL service, users and content come from many languages.

The results presented in this paper are related only to The European Library logs. The TEL logs contain entries for different types of user interactions, collected since January 2007 to June 2008, and since January 2009 to December 2009. A more detailed description of the task and the dataset can be found in [2] and at the LogCLEF web page[3].

[1] http://sinai.ujaen.es/

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

[3] http://www.uni-hildesheim.de/logclef/

The rest of the paper is organized as follows: Section 2 gives a brief description of the preprocessing operations performed on the TEL logs, Section 3 discusses the log analysis along with the obtained results. Finally, in Section 4, the paper ends with the conclusions.

2 Preprocessing work

All original TEL log entries have been stored in a MySQL database. A TEL log entry contains some attributes such as the identification number of the session (field *sesid*), the type of action performed by the user (field *action*), the interface language (field *lang*) and the query (field *query*). The experiments carried out in this paper are focused on these attributes.

The first preprocessing work applied to that dataset consisted of two main subtasks related to the field *query*: problems related to character encodings were solved and symbols such as brackets, quotation marks or parentheses were deleted. Then, the following step was to store in an additional table those entries whose fields *sesid, lang, query* and *action* were not empty or null. Therefore, entries having empty queries, interface language or missing actions were deleted. The original number of records (2,628,789) was reduced to 2,417,025 after the cleaning process (approximately 8.1% of the records were deleted).

Finally, in the last preprocessing step, we carried out the reconstruction of user sessions. The reconstructed sessions were stored separately in an additional table with following fields:

- *sesid*: unique identifier of the user session.
- *num_interactions*: number of interactions (entries) for each user session.
- *duration*: it is a field which stores the time difference between the registered *timestamp* of the last entry for the session and the *timestamp* of the first entry for the same session.
- *ip_loc*: this field stores the country of the IP address detected in the session. We have used the *Geo::IP::PurePerl* module for Perl[4].
- *languages*: this field stores all the languages used during the session, separated by commas.

The total number of different reconstructed sessions was 308,938.

3 Analysis of TEL logs

The analysis of TEL logs carried out in this paper focuses on three main aspects: **languages** (showing the percentage of sessions for each language), **interactions** (showing the number of interactions per session) and the study of the separability of sessions according to the words in the query and the actions. We present in the next section the results of this analysis.

[4] http://search.cpan.org/~borisz/Geo-IP-PurePerl-1.25/lib/Geo/IP/PurePerl.pm

3.1 Languages

Figure 1 shows the percentage of sessions per language. Over the total of 308,938 sessions, 95% of all sessions from TEL logs are covered by these 9 languages (with a clear dominance of English as language for the interface). The use of non English languages is not significant.

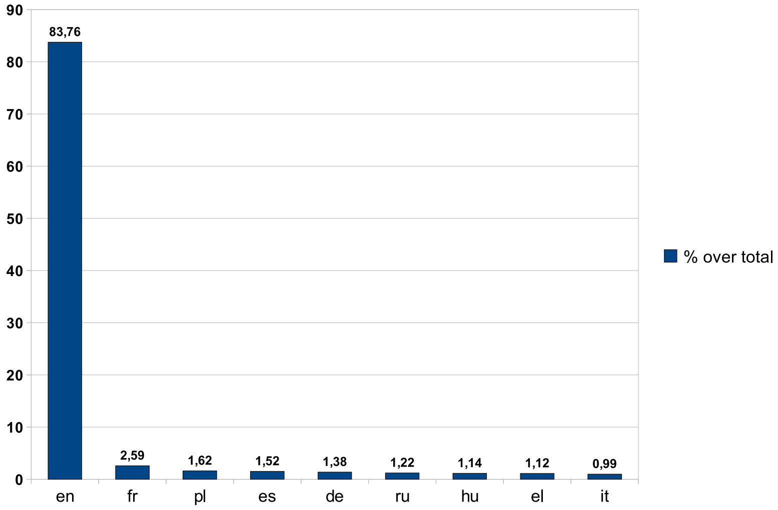

Fig. 1. % sessions per language

3.2 Interactions

Figure 2 shows a known curve in LogCLEF: the frequency of a given number of interactions per session. The vertical axis represents the number of sessions that have the same number of interactions, which is in the horizontal axis. For example, 50,209 sessions show only one interaction. The long tail informs us about the relative low number of interactions that use to be performed in each session. More than 80% over the total of sessions have 10 or less interactions (almost 96% have 30 or less interactions).

3.3 Queries and actions

We have studied the separability of the TEL log sessions according to three main aspects: the words in the query strings, the actions present in a session and the number of changes (related to the words) from one interaction to the next one.

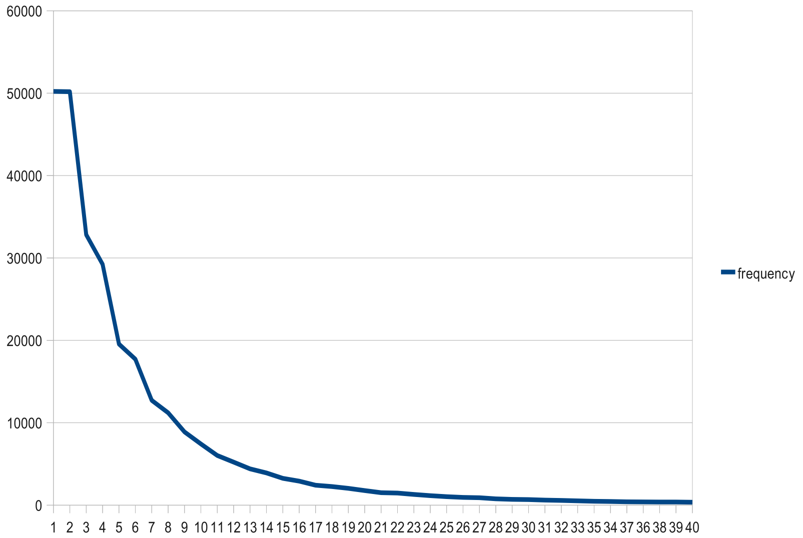

Fig. 2. Frequency of the number of interactions per session

After applying the Principal Component Analysis (PCA) [1], just keeping two components (used words and languages) over the different features extracted per session (average number of words per query, total number of different words used in queries in that session, number of successful actions, etc.), the 95% of the variance of the data is preserved. In our case, we considered successful action when the user completes the search with the actions *view_full* or *view_brief*. As we can observe in the Figure 3, some outliers are identifed, due to that small number of sessions with high number of interactions. But a big cloud can be identified, and some variability in its geometry may be guessed. This would need further analysis in order to discover possible classes of sessions according to the queries.

4 Conclusions

This is the first participation of SINAI group in LogCLEF track. Initially, the main motivation for participating in LogCLEF was in the Log Analysis and Geographic Query Identification (LAGI) subtask proposed in the previous year. The identification of geographic queries within a query stream and the recognition of the geographic component are key problems for Geographic Information Retrieval (GIR). But finally this year, the organizers of LogCLEF decided not to take into account any subtask related to geographic query identification.

Nevertheless, we decided to participate in LogCLEF providing a brief analysis and statistics of TEL logs. As main conclusions, we can observe a clear domi-

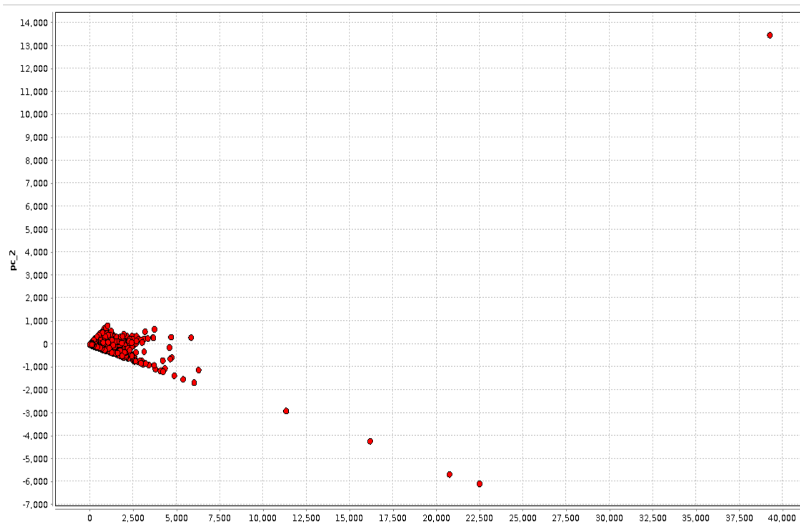

Fig. 3. Scatter matrix for two principal components about used words and languages

nance of English as language for the interface and more than 80% over the total of sessions have 10 or less interactions. In addition, after applying the Principal Component Analysis, just keeping two components over the different features extracted per session (used words and languages), the 95% of the variance of the data is preserved.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project TEXT-COOL 2.0 (TIN2009-13391-C04-02), a grant from the Andalusian Government, project GeOasis (P08-TIC-41999), and a grant from the University of Jaen, project RFC/PP2008/UJA-08-16-14 and project UJA2009/12/14.

References

1. Lam, S.L.Y., Lee, D.L.: Feature reduction for neural network based text categorization. In: DASFAA '99: Proceedings of the Sixth International Conference on Database Systems for Advanced Applications. pp. 195–202 (1999)
2. Mandl, T., Agosti, M., Nunzio, G.M.D., Yeh, A., Mani, I., Doran, C., Schulz, J.M.: LogCLEF 2009: the CLEF 2009 Multilingual Logfile Analysis Track Overview. In: Working Notes of the CLEF 2009 workshop (2009)