DAEDALUS at LogCLEF 2010: Analyzing the Success of Search Queries

Sara Lana-Serrano[1,3], Julio Villena-Román[2,3], José Carlos González-Cristóbal[1,3]

[1] Universidad Politécnica de Madrid
[2] Universidad Carlos III de Madrid
[3] DAEDALUS - Data, Decisions and Language, S.A.
slana@diatel.upm.es, jvillena@it.uc3m.es, josecarlos.gonzalez@upm.es

Abstract. This paper describes the participation of DAEDALUS at the LogCLEF task. The focus of our experiments was to study if the difference between the native language of the user and the interface language could affect the way in which the user interacts with the search application and the success of the search queries. First, the provided log data was parsed into 194,040 sessions containing the set of sequential actions carried out by the same user. Then, only those sessions that include at least one search query were selected, 16% of the total number of sessions. Within that session set, a total number of 388,272 queries have been run, only 6.45% of which were successful, i.e. return any result, thus resulting in 10.6% of successful sessions. After a statistical correlation analysis of these figures, the main conclusion that can be drawn is that, in the general case, the fact that the native language is used or not as the interface language doesn't seem to affect to the success rate of the search queries.

Keywords: LogCLEF, log file analysis, The European Library, user language, native language, interface language, action patterns.

1 Introduction

The basic goal of the LogCLEF track [1] at CLEF 2010 is to perform any kind of analysis over The European Library (TEL) [2] logs to research on the effects that the language adopted by users may have on the search operations, in order to understand user search behavior in multilingual contexts and ultimately to improve search systems.

Specifically, in this research, three involved languages are considered: language in which the user has set up the search tool interface, language of the collections of information on which the user makes his/her queries and/or navigates through the results, and the inherent language of the user (his/her native language), inferred based on the browser IP.

Our research group is led by and named after DAEDALUS, a small private company in the field of Information and Telecommunication Technologies and a leading provider of language-based solutions in Spain, and research groups of two universities, Universidad Politécnica de Madrid and Universidad Carlos III de Madrid. We have taken part in CLEF since 2003 in many different tracks and tasks, as part of the MIRACLE team till last year. This paper describes our participation at the LogCLEF track.

The aim of our research is to study 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. The idea is to determine if this difference may condition the way in which the user interacts with the search application.

As our analysis involves the identification and analysis of a sequence of actions carried out by the same user, only those entries in the log files for which it was possible to extract a session identifier have been considered, so as to be able to associate them to a set of related actions.

2 Log Analysis and Information Modelling

Based on the analysis of the data existing both in the log files and the action file provided with The European Library data [2], we defined a data model to represent the information associated to the following logical entities:

- **Query**: set of sequential actions by the user in which a query is involved.
- **Session**: set of sequential actions carried out by a given user. A session may involve zero, one or several queries. In our study, only sessions with at least one query have been considered.

In addition, each query has been modelled by a series of properties:

- **Action that has triggered the query**: we have considered that a query is triggered when the user makes any of the following actions: “search_sim”, “search_adv”, “search_res”, “search_url”, and also when the text of the query is modified.
- **Primary language**: language selected in the user interface at the beginning of the session.
- **Secondary languages**: list of languages, different to the primary language, which the user has selected in the interface, without any modification of the query.
- **Query language**: inherent language of the query, inferred from the IP address of the user.
- **Number of filtering actions**: a filtering action (“search_res_rec_any”, “search_res_rec_all”) is one that allows the user to refine the results associated to the query.
- **Number of browsing actions**: a browsing action (“view_brief”, “jump_to_page”, “page_brief”) represents an interaction by the user on the search results, which is not any successful action.
- **Number of collections**: number of different collections on which the user has carried out any action.

- Number of different collections in which the language matches the language in which the user interface is configured.
- Number of different collections in which the language matches the user language inferred from his/her IP address (native language).
- Number of times that the user has carried out a view detail action ("view_full"). This action is very important because it gives access to actions identified as successful actions.
- Number of unsuccessful queries after the last successful query in the same session.
- **Successful query**: a query is successful if it involves at least one of the following actions: "available_at", "see_online", "option_save_session_favorite", "option_send_email".
- Number of times that each successful action has been run.

Moreover, for each session in which a previous selection of the search collections has been made (by means of the "col_set_theme_country" action), the relationships existing among the language inferred by the IP address, the language in which the user interface is configured and the language associated to the selected collections, has been considered in the data model.

3 Results

Once filtered and modelled the information in the log files according to the described model, 194,040 sessions are selected (i.e., those including significant information for our analysis) out of the 225,358 total sessions, which means that 16% of the started sessions don't involve any search operation. In those selected actions, a total of 388,272 queries have been made, 6.45% of which are successful, corresponding to a 10.6% of successful sessions.

The following **Table 1** shows the average value of the main features in a session, considering if the interfaz language matches the language inferred from the IP (Lang=1) or not (Lang=0). The columns include *Sessions* (number of sessions), *Queries* (average number of queries per session), *Jumps* (average number of jumps), *Filters* (average number of filtering actions), *Detail* (average number of *view_full* actions), *NotSuccess* (average number of queries between two successful queries); *actionSuccess* (average number of successful actions).

Table 1. Average values of session features.

Lang	Sessions	Queries	Success	Jumps
0	130,123	1.9883	0.1249	1.1736
1	67,018	1.9331	0.1314	1.0698

Lang	Filters	Detail	NotSuccess	ActionSuccess
0	0.0093	1.4587	0.45704	0.2448
1	0.0076	1.4222	0.48345	0.2515

After a correlation analysis of these figures, we could affirm that, in general, the fact that the native language of the user matches or not the interface language, doesn't have apparently any impact on the success rate of the search queries. Another conclusion that can be drawn is that the filtering option in the interface doesn't receive a high interest from the users.

If we analyze the way the users carry out different types of queries, it can be noticed that there is no direct relation between the involved languages and the query type. Only 15% of queries make use of the advanced search form in the web page, and only 4.32% of them are successful as compared to the 6.45% of the rest of queries.

Another interesting result is that the original query is modified in the interface only in 62 queries. In addition, 4,593 queries have involved more than one query with different settings for the interface language. Both of them turn out to be negligible values.

The analysis on how the users select the collections on which they want to search shows that this possibility has been used only in 16.26% of the sessions. Collections with the same language as the one associated to the user IP address were selected in 27% of operations, whereas 30.13% of the cases selected collections with the same language as the interface. In 9.5% of the queries, the three involved languages (describe in section 1) were the same.

4 Conclusions and Future Work

The aim of our research was to study 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. Based on the results obtained, the main conclusion that can be drawn is that, in the general case, the fact that the native language is used or not as the interface language doesn't seem to affect to the success rate of the search queries. In other words, whether this difference in languages conditions or not the way in which users interact with the search application doesn't have any significant impact on the success rate.

There are still a lot of open questions and possible valuable analyses using the provided log files for future participations in the task. In particular, we were initially interested in researching on the actual semantic content of the query and its relation (if there is any) with any of the involved languages or the success of the query, but unfortunately we had to abandon this approach due to lack of time and resources. We may be able to carry out these types of analyses in future years.

Acknowledgements

This work has been partially supported by the Spanish Center for Industry Technological Development (CDTI, Ministry of Industry, Tourism and Trade),

through the BUSCAMEDIA Project (CEN-20091026). Authors would like to thank all BUSCAMEDIA partners for their knowledge and contribution.

References

1. Overview of the LogCLEF track at CLEF 2010. Working Notes of CLEF 2010. Padova. Italy. 2010.
2. The European Library (TEL). http://search.theeuropeanlibrary.org/.