DAEDALUS at LogCLEF 2011: Analyzing Query Success and User Context

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 lab in CLEF 2011. This year, the objectives of our participation are twofold. The first topic is to analyze 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. The second topic is to analyze the user context and his/her interaction with the system in the case of successful queries, to discover out 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. Only 6.89% of queries are successful out of the 628,607 queries in the 320,001 sessions with at least one search query in the log. The main conclusion that can be drawn is that, in general for all languages, whether the native language matches the interface language or not does not seem to affect the success rate of the search queries. On the other hand, the analysis of the strategy adopted by users when looking for a particular resource shows that people tend to use the simple search tool, frequently first running short queries build up of just one specific term and then browsing through the results to locate the expected resource.

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

1 Introduction

This paper describes the participation of DAEDALUS team at the LogCLEF lab [1], part of CLEF 2011. The main goal of this lab is to carry out 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 behaviour in multilingual contexts and ultimately to improve search systems.

Specifically, three involved languages are considered in this research: 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.

After our participation in the previous edition of LogCLEF [3], this year we decided to focus on two specific objectives. On the one hand, we are very interested in 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. The idea is to determine if this difference may condition the way in which the user interacts with the search application. On the other hand, we wanted to study in detail the user context and his/her interaction with the system in the case of sessions with a successful operation (*available_at*, *see_online*) over the same resource. Our final objective was to try to discover out 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.

In the following sections we will fully describe our analysis and the results and conclusions that can be drawn from this work.

2 Log Analysis and Information Modelling

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.

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

- **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.

This model is similar to the one that we defined for our previous participation in LogCLEF [3].

In order to deal with the first of our objectives, each query is modelled by a series of properties:

- **Action that 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 user IP address.

- **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 as it leads 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 these 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.

In addition to this information, to deal with our second objective, the following information has been extracted for each of the resources that have been requested by means of the *available_at* or *see_online* actions within a query and a session:

- **Successful action**: one of the following actions: *available_at*, *see_online*.
- **Successful language**: language of the interface when the action was run.
- **Resource**: URL of the requested resource.
- **Number of filtering actions** after the last successful operation, or, for the first successful action, the number of filtering actions from the first run of the query.
- **Number of jumping actions:** a jumping action (*jump_to_page*) represents a navigation (or browsing) operation over the result listing.

3 Results

Once the information in the log files has been filtered and organized according to the previously described model, 367,348 sessions are kept (i.e., those including significant information for our analysis) out of the 320,001 total sessions. This means that 12.88% of the started sessions do not involve any search operation. In those selected actions, a total of 628,607 queries have been made, 6.89% of which are successful, corresponding to a 11.11% of successful sessions.

The following Table 1 shows the average value of the main features in a session, considering whether the interface language matches the language inferred from the IP (Lang=1) or not (Lang=0). Parameters in rows include *Sessions* (number of sessions), *Queries* (average number of queries per session), *Jumps* (average number of navigation operation over the result listing), *Filters* (average number of filtering actions), *Detail* (average number of *view_full* actions), *NotSuccess* (average number of queries between two successful queries), and *ActionSuccess* (average number of successful actions).

Table 1. Average values of session features**.**

Parameter	Not match (Lang=0)	Match (Lang=1)	Difference
Sessions	208,384	116,270	-44.2%
Queries	1.9641	1.8863	-4.0%
Success	0.1327	0.1348	+1.9%
Jumps	1.2678	1.1668	-8.0%
Filters	0.0058	0.0044	-24.1%
Detail	1.2943	1.1755	-9.18%
NotSuccess	0.4450	0.4500	+1.1%
ActionSuccess	0.2614	0.2530	-3.2%

Table 2 shows the number of sessions and queries aggregated by the language in which users setup the interface, for the 10 most-frequent languages. The last column shows the percentage of queries in which the interface language matches the user native language.

Table 2. Language of the interface.

Language	Sessions	Queries	Match (Lang=1)
en (English)	273,936	520,337	26.1%
fr (French)	8,206	15,929	83.5%
pl (Polish)	5,339	11,630	77.8%
de (German)	4,935	10,311	78.1%
ru (Russian)	4,726	9,496	65.2%
es (Spanish)	4,530	8,046	90.2%
pt (Portuguese)	3,636	7,181	87.1%
it (Italian)	3,152	7,071	89.6%
hu (Hungarian)	2,499	5,419	78.2%
tr (Turkish)	2,385	4,340	94.9%

It can be easily observed that the most frequent language for the interface is English, although it only matches the user language in 26.1% of queries.

Table 3 shows a detailed analysis of some selected parameters similar to Table 1 for the 5 main languages.

Table 3. Average values of session features, by interface language.

Parameter	Language	Not match (Lang=0)	Match (Lang=1)	Difference
Sessions	en	197,387	76,582	-61.2%
	fr	1,394	6,813	+388.7%
	pl	1,350	3,991	+195.6%
	de	1,166	3,769	+223.2%
	ru	1,668	3,058	+83.3%
Queries	en	1.9471	1.7759	-8.8%
	fr	1.8802	1.9533	+3.9%
	pl	1.9659	2.2491	+14.4%
	de	1.9391	2.1358	+10.1%
	ru	1.9790	2.0258	+2.4%
Success	en	0.1289	0.1161	-9.9%
	fr	0.2260	0.2456	+8.7%
	pl	0.1570	0.1541	-1.8%
	de	0.1655	0.1995	+20.5%
	ru	0.1481	0.1298	-12.4%
NotSuccess	en	0.4365	0.3578	-18.0%
	fr	0.6926	0.5575	-19.5%
	pl	0.5474	0.6718	+22.7%
	de	0.7734	0.6767	-12.5%
	ru	0.5430	0.5576	+2.7%
ActionSuccess	en	0.2553	0.2084	-18.4%
	fr	0.4527	0.5073	+12.1%
	pl	0.2733	0.3693	+35.1%
	de	0.2899	0.4105	+41.6%
	ru	0.3135	0.2822	-10.0%

After a correlation analysis of these figures, we could affirm that, in general for all languages (as shown in Table 1), 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.

However, there are noticeable differences in the detailed analysis for each language (Table 3**)**, especially for German (20.5% increment in success when languages match). These differences have yet to be explained.

Another conclusion that can be drawn from Table 1 is that the filtering option in the interface does not 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 14.27% of queries make use of the advanced search form in the web page, and only 4.73% are successful as compared to the 6.89% of the rest of queries.

So as to explore the way in which users interact with the system when they are looking for a given resource, we have carried out a set of studies that focus on the resources that have been accessed by a given user after a search process and the queries that such user has run to locate them.

Regretfully, as dynamic parameters in the URL that identify the resources are not currently stored in the logs, the information provided was useful only for resources whose URL is static. Thus, this analysis is only possible for .jpg, .pdf, .txt and .doc resources.

For this analysis, we only have considered queries that allowed to access any of those resource types by means of an *available_at* or *see_online* action.

Assuming those criteria, 2,391 different queries have been identified, 6,002 requested resources and 6,884 different query-resource combinations.

Table 4 shows some statistics associated to the most frequent queries. Columns include *Query* (the user query), *Queries* (number of times that the query has been run), *Sessions* (number of different sessions), *Resources* (number of different requested resources), *LangU* (number of different user languages involved), *LangI* (list of different interface languages involved), and the number of matches between the interface and user language.

Table 4. Most frequent queries.

Query	Queries	Sessions	Resources	LangU	LangI	Match (Lang=1)
mozart	297	87	110	28	4 (de,es,en,fr)	77
"france"	293	2	272	2	1 (en)	1
weltkrieg	150	1	148	1	1 (en)	0
winterspaß	133	18	105	11	3 (en,fr,es)	114
"paris"	127	9	119	6	2 (en,fr)	123
"hitler"	125	11	74	9	2 (en,fr)	41
galizien	102	2	97	2	2 (pl,en)	99
"bosnien"	80	5	78	2	1 (en)	0
einstein	78	30	22	15	5 (en,de,nl,el,fr)	51
"warsaw"	72	8	40	1	2 (pl,en)	60

Table 5 shows similar statistics for the most requested resources. In this case, *Hits* represents the number of times the resource has been requested.

Table 5. Most frequent resources.

Resource	Hits	Sessions	Queries	LangU	LangI	Match (Lang=1)
#1	16	13	11	6	2 (en,fr)	4
#2	13	10	10	6	2 (en,fr)	4
#3	9	9	9	7	1 (en)	1
#4	9	8	7	8	2 (en,fr)	4
#5	11	11	7	5	2 (bg,en)	3

#6	10	7	7	3	2 (en,es)	2
#7	7	7	7	7	2 (en,ru)	5
#8	6	6	6	5	1 (en)	1
#9	24	18	6	8	3 (en,sl,de)	7
#10	13	12	6	7	2 (en,bg)	2

Again, it can be observed that the relation between the interface language and the user language does not have a strong effect on the success of the query. We believe that the main reason for such lack of correlation is due to the fact that most queries are composed up of just one search term, which typically are very specific queries containing a given proper name (such as the examples shown in Table 4). Thus, in this scenario, only 43 of the 304 queries that are formulated in more than one session (14%) contain more than one search term, and 29 of them (68%) correspond to a multiword proper noun (such as “da vinci”) .

4 Conclusions

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 achieved, 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 does not apparently affect 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 does not have any significant impact on the success rate.

On the other hand, we have analyzed the strategy adopted by users when they are looking for a particular resource. People tend to use the simple search tool, frequently first running short queries build up of just one specific term and then browsing through the results to locate the expected resource.

For future participations in the task, we are still 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. Unfortunately we had to abandon this idea due to lack of time and resources, but we may be able to carry it out in future years.

Acknowledgements

This work has been partially supported by several Spanish research projects: MA2VICMR: Improving the access, analysis and visibility of the multilingual and multimedia information in web for the Region of Madrid (S2009/TIC-1542), MULTIMEDICA: Multilingual Information Extraction in Health domain and application to scientific and informative documents (TIN2010-20644-C03-01) and BUSCAMEDIA: Towards a semantic adaptation of multi-network-multiterminal

digital media (CEN-20091026). Authors would like to thank all partners for their knowledge and support.

References

1. Overview of the LogCLEF track at CLEF 2011. *CLEF 2011 LABs and Workshops, Notebook Papers*. 19-22 September, Amsterdam, The Netherlands, 2011.
2. The European Library (TEL). http://search.theeuropeanlibrary.org/.
3. Lana-Serrano, Sara; Villena-Román, Julio; González-Cristóbal, José Carlos. DAEDALUS at LogCLEF 2010: Analyzing the Success of Search Queries. *CLEF 2010 LABs and Workshops, Notebook Papers*. 22-23 September, Padua Italy, 2010. ISSN 2038-4963.