Identifying Common User Behaviour in Multilingual Search Logs

M. Rami Ghorab[1], Johannes Leveling[2], Dong Zhou[1], Gareth J. F. Jones[2], and Vincent Wade[1]

[1] Centre for Next Generation Localisation
Knowledge and Data Engineering Group
Trinity College Dublin
Dublin 2, Ireland
{ghorabm, dong.zhou, vincent.wade}@scss.tcd.ie
[2] Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
{jleveling, gjones}@computing.dcu.ie

Abstract. The LADS (Log Analysis for Digital Societies) task at CLEF aims at investigating user actions in a multilingual setting. We carried out an analysis of search logs with the objectives of investigating how users from different linguistic or cultural backgrounds behave in search, and how the discovery of patterns in user actions could be used for community identification. The findings confirm that users from a different background behave differently, and that there are identifiable patterns in the user actions. The findings suggest that there is scope for further investigation of how search logs can be exploited to personalise and improve cross-language search as well as improve the TEL search system.

1 Introduction

The Log Analysis for Digital Societies (LADS) task is part of the LogCLEF track at the Cross-Language Evaluation Forum (CLEF) 2009. The LADS dataset contains log entries of user interactions with the TEL[3] portal. The logs were analysed to investigate the following hypotheses: (1) users from different linguistic or cultural backgrounds behave differently in search; (2) there are patterns in user actions which could be useful for stereotypical grouping of users; (3) user queries reflect the mental model or prior knowledge of a user about a search system.

The remainder of this paper is organised as follows: Sect. 2 gives a brief description of the logs, Sect. 3 discusses the log analysis and results, and the paper ends with conclusions and outlook to future work in Sect. 4.

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

2 Brief Description of Logs and Preprocessing Operations

A log entry is created for every user interaction with the TEL portal. Log entries contain the type of action performed, together with attributes such as the interface language, query, and timestamp. The experiments focused on the following attributes: *lang* (interface language selected by the user), *action*, and *query*. The main actions that our study investigated were:

- search_sim: searching via a simple text box.
- search_adv: advanced search by the specific fields of title, creator (e.g. author or composer), subject, type (e.g. text or image), language, ISBN, or ISSN.
- view_brief: clicking on a library's collection to view its brief list of results.
- view_full: clicking on a title link in the list of brief records to expand it.
- col_set_theme: specifying a certain collection to search within.
- col_set_theme_country: specifying multiple collections to search or browse.

An important part of preprocessing the logs was session reconstruction. Each action was associated with a session ID and a timestamp. Actions of the same session were grouped together by session ID and then were sorted by timestamp. Details of the dataset and the preprocessing can be found in [1] and [2].

3 Analysis of Log File Entries

3.1 General Statistics

Table 1 presents statistics from the log analysis. Only a small proportion of the actions were performed by signed-in users (0.76%) compared to the number of actions recorded for guests (99.24%). This may indicate that users find it easier, and/or perhaps more secure, not to register in a web search system. Such behaviour sets a challenge to individualised personalisation.

Table 1. Descriptive Statistics.

Item	Frequency
Actions by guests	1,619,587
Actions by logged-in users	12,457
Queries by guests	456,816
Queries by logged-in users	2,973
Sessions	194,627
User IDs	690

User actions were classified into four categories: *Search*, *Browse* (browsing or navigating result pages), *Collection* (limiting the search scope by selecting a collection or subject), and *Other*. Table 2 shows the distribution of actions along the categories. A considerable number of user actions (11.34%) were performed before attempting the search, such as specifying collections for search. This indicates the diversity of user preferences, where users seek to customise the search environment according to their needs. User profiling may help to save user effort by automatically adjusting the search environment upon user identification.

Table 2. Broad classification of actions.

Classification	Percentage
Search	28.17
Browse	56.79
Collection	11.34
Other	3.70

With regards to searching, it was found that there was a great inclination towards using simple search (16.14% of total actions) compared to using advanced search (4.35% of total actions). Another inclination was found in the pre-selection of a single collection for search, which occurred more frequently than the pre-selection of multiple collections (col_set_theme: 7.13% of actions; col_set_theme_country: 2.72%). This suggests that users seeking to limit their search tend to be very specific in selecting a collection. This may come from previous experience with the portal, where users found that certain collections had a higher degree of satisfying their information needs.

3.2 Query Reformulation

There are several types of reformulation of successive user queries: focusing on search terms and disregarding Boolean operators, a term can be added, deleted, or modified. For advanced search, in addition, a field can be added, deleted, or changed (some of the latter actions co-occur with operations on search terms). We defined four types of query reformulation, depending on the way query terms are affected: term addition, term deletion, term modification, and term change. Term modifications are changes to single-term queries.[4] No differentiation was made between queries submitted under different interface languages, because (i) the major part of the queries were submitted under English, and thus, the data for other interface languages might not be sufficient, and (ii) some query changes were manually observed as changing a query to another language.

As some users switch from the simple to the advanced search interface, related queries are difficult to identify if different types of queries are considered. For the following experiment, search terms were extracted from queries in order to identify how users typically modify a query. Only successive searches on the same topic were considered. To identify queries about the same topic, the following

[4] The distinction between term changes and term modifications originates from the definition of successive queries for queries with one and with more search terms.

approach was used: consecutive queries must have at least one term in common (if the query contains more than one search term) or a term in the query must have a Levenshtein distance [3] less than three to one in the other query. A query parser was implemented to extract the terms from the query log and identify the type of query modification and the most frequent changes.

Table 3 shows some of the reformulation classes based on the top 50 reformulations. A hyphen in Table 3 indicates operations which are not observable. It was found that 16% of term additions (add), 24% of term deletions (del), and 28% of term changes (chg) affected stopwords. Such changes might make sense under the assumption that users sometimes copy and paste text into a search box, and they might have just mistakenly inserted unwanted stopwords into the TEL search box. However, if the underlying indexing/retrieval system of TEL ignores stopwords, then adding or changing them will have no effect on search results, and would be considered a waste of effort for TEL users. A quick test reveals that stopword removal is handled inconsistently by the libraries in TEL, e.g. a search for *“the”* returns zero hits for the Austrian and French national library, but several thousand for the German and Belgian national library.

Table 3. Top 50 changes to terms in successive related queries.

			Percentage			
Type	Brief description	Example	add	del	mod	chg
ST	use of stopwords	*“a”* → *“the”*	16	24	6	28
BL	use of Boolean operators	*“AND”* → *“OR”*	4	6	0	12
CC	change of lowercase or uppercase	*“europe”* → *“Europe”*	0	0	8	0
SC	spelling change	*“wolrd”* [!] → *“world”*	0	6	4	4
CH	use of special characters	*“*”* at the end of term	6	0	0	4
LC	language code change	*“ita”* → *“eng”*	2	2	0	20
RT	related terms	*“triangulum”* → *“quadratum”*	–	–	2	4
MO	morphologic variant	*“city”* → *“cities”*	–	–	26	2
TR	translation or transliteration	*“power”* → *“kraft”*	–	–	24	4
PN	change proper noun/name	*“mozart”* → *“amadeus”*	42	26	20	8
PI	single character (initials)	*“elzbieta”* → *“e”*	20	20	0	2
DT	date/number change	*“1915”* → *“1914”*	4	6	0	6
OT	unknown change/other	*“test”* → *“toto”*	6	10	10	6

Proper nouns and single characters (mostly denoting initials of names) made up 62% of term additions, 46% of deletions, 20% of modifications (mod), and 10% of changes. In contrast, term modification mostly affect morphological variations and translations (26% and 24%, respectively). Such modifications would not have any effect on the search results, because the TEL system does not seem to perform stemming.

Special characters (e.g. wildcards) were rarely used. Moreover, a small number of changes involved the use of semantically related terms (including narrower terms or broader terms). Also, only a small number of changes involved changing Boolean operators (e.g. *"AND"* → *"OR"*). This behaviour implies that some users are familiar with different search operators supported by the TEL portal.

The query reformulation analysis supports the hypothesis that a large group of users has little knowledge of the system, as they include stopwords and even change them (assuming TEL ignores stopwords as is commonly done by search engines). This group corresponds to novice users. On the other hand, a small group, corresponding to experienced users, used advanced query operators such as wildcards.

3.3 Interface Languages

In an attempt to investigate the relation between language and search behaviour, several variables were studied across the interface language selected by users of the portal. Actions were distributed among 30 languages. Hereafter, the study focuses on the top five languages in terms of the number of actions. The top language was English (86.47% of the actions), followed by French (3.44%), Polish (2.17%), German (1.48%), and Italian (1.39%). It is to be noted that the interface language does not necessarily imply the language of the query. One possible cause for the bias towards English, aside from its inherent popularity, is that it is the default language in the portal. Due to such anticipated bias, we will not include English (as an interface language, not as a query language) in further comparative discussions against other interface languages in this study. Nevertheless, we will show its associated percentages in subsequent tables for the sake of completeness. Possible ways to avoid this bias in the future would be to ask the user to specify a language before attempting the search, or to have the default language automatically specified according to the client's IP address.

The frequency distribution of the six main actions across the five languages is shown in Table 4. It was observed that users of the Polish language seemed to have a higher rate than others in using the feature of specifying a single collection before attempting the search. This finding may support the hypothesis that users from different linguistic or cultural backgrounds behave differently in search. However, we cannot rule out the fact that such observation may have been specifically governed by the amount of available collections in TEL.

3.4 Term Frequencies and Categories

As part of our analysis, the number of terms per query and the top queried terms for simple and advanced search were studied. Table 5 shows the mean and median of the number of terms per query across interface languages. It can be seen that German showed the lowest mean in both types of search. Moreover, part of the analysis revealed that German exhibited the largest distribution of queries made up of just one term. This may be because German noun compounds, which can express complex topics, are written as a single word.

Table 4. Action distribution across languages.

Lang	search_sim	search_adv	view_brief	view_full	col_set_theme	col_set_theme_country
English	16.48%	4.32%	25.79%	30.65%	6.79%	2.66%
French	14.27%	4.46%	27.34%	23.55%	10.86%	3.12%
Polish	15.18%	4.23%	26.99%	21.95%	13.58%	3.39%
German	14.75%	4.31%	28.96%	23.53%	9.46%	2.93%
Italian	14.44%	6.16%	24.81%	28.39%	9.35%	2.78%

A comparison was made between the mean of the number of terms per query in simple search and the results reported in [4], which was a similar study applied on logs from AlltheWeb.com[5] (a European search engine that allows limiting the search to documents in a language of choice). With the exception of English, the means were approximately the same, despite the fact that the former is a library search system and the latter is a general search engine.

Table 5. Number of terms per query across interface languages.

	Simple Search		Advanced Search	
Language	Mean	Median	Mean	Median
English	2.38	2	3.05	3
French	2.09	1	2.85	2
Polish	1.89	1	2.59	2
German	1.77	1	2.6	2
Italian	2.09	2	3.17	2

Part of the log analysis involved the extraction of the top 20 occurring search terms for each interface language, excluding stopwords. A term was only counted once in a session. This was done to avoid bias towards terms that were repeatedly searched for in the same session. Furthermore, terms were divided into five categories: *creator* (author, composer, artist, etc.), *location* (cities, countries, etc.), *subject* (as per Dewey Decimal Classification), *title* (including proper nouns and common nouns), and *type* (document types, e.g. text, image, sound). These categories were mostly based on the fields of the advanced search in TEL.

Figure 1 shows the category distribution of the top 20 search terms for each language. Differences were observed in user behaviour between different languages. For example, in simple search, 20% of the terms under French were subjects and 25% were creators, while under Italian, only 5% of the terms were subjects, while 40% of the terms were creators. Such findings reflect the differences between users of different languages and may contribute towards further

[5] http://www.alltheweb.com/

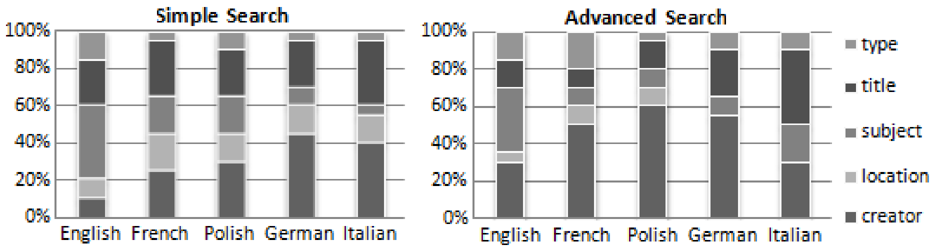

Fig. 1. Distribution of term categories across languages.

research in multilingual query adaptation, perhaps suggesting a different adaptation strategy for each language or group of languages.

3.5 Action Sequences

Table 6 shows patterns of two and three successive user actions. It points out the top most occurring patterns, as well as some other interesting patterns that have a high frequency. Related patterns are grouped together. It is observed that more users, after performing a search action, seem to directly view a full record (click for expansion) rather than clicking on a collection first (view_brief) before clicking to view full. The reason for this may be that the collection they wanted was already highlighted (TEL automatically highlights the top most collection in alphabetical order). This may indicate that more people prefer to specify collections before they perform the search so as to directly jump to view full without having to click on a collection.

Table 6. Selected sequential action patterns for two and three successive actions.

Action 1	Action 2	Action 3	Frequency
view_full	view_full	–	153,952
search_sim	view_full	–	112,562
search_sim	view_brief	–	86,625
col_set_theme	search_sim	–	40,044
col_set_theme_country	search_sim	–	12,397
view_full	view_full	view_full	79,346
col_set_theme	col_set_theme_country	col_set_theme	4,735
col_set_theme_country	col_set_theme	search_sim	3,159

It can also be observed that users seem to get confused between two features (both accessible from TEL web site main page), which are: col_set_theme (choose a single collection) and col_set_theme_country (browse collections/choose multiple collections, which redirects the user to another page). This was observed as user actions subsequently alternated between the two features. Based on the pattern frequencies and the findings presented in Sect. 3.1, it can be inferred that users prefer the feature of choosing a single collection. Perhaps deeper analysis of such patterns may introduce certain changes to the TEL portal's GUI.[6]

4 Summary and Outlook

We have described an analysis of multilingual search logs from TEL for the LADS task at CLEF 2009. The results of the analysis support the hypotheses that: (1) users from different linguistic or cultural backgrounds behave differently in search; (2) the identification of patterns in user actions could be useful for stereotypical grouping of users; and (3) user queries reflect the mental model or prior knowledge of a user about a search system.

The results suggest that there is scope for further investigation of how search logs can be exploited to improve cross-language search personalisation. Furthermore, the results imply that there is scope for improving the TEL system in a number of ways: (1) integrating a query adaptation process into TEL, where queries can be automatically adapted to retrieve more relevant results; (2) offering focused online help if a user spends an uncharacteristically long time between some actions or if a user performs a sequence of logically inconsistent actions; (3) highlighting elements in the TEL GUI as a default action or a typical next action; and (4) identifying the type of user for the sake of search personalisation.

Acknowledgments

This research is supported by the Science Foundation of Ireland (grant 07/CE/-I1142) as part of the Centre for Next Generation Localisation at Trinity College Dublin and Dublin City University (http://www.cngl.ie/).

References

1. 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)
2. Ghorab, M., Leveling, J., Zhou, D., Jones, G., Wade, V.: TCD-DCU at LogCLEF 2009: An analysis of queries, actions, and interface languages. In Borri, F., Nardi, A., Peters, C., eds.: Results of the CLEF 2009 Cross-Language System Evaluation Campaign, Working Notes of the CLEF 2009 Workshop. (2009)
3. Wagner, R.A., Lowrance, R.: An extension of the string-to-string correction problem. JACM **22**(2) (1975) 177–183
4. Jansen, B.J., Spink, A.: An analysis of web searching by European AlltheWeb.com users. Information Processing & Management **41**(2) (2005) 361–381

[6] Comments on the TEL GUI refer to the version accessible on the web in July 2009.