TCD-DCU at LogCLEF 2009: An Analysis of Queries, Actions, and Interface Languages

M. Rami Ghorab [a] Johannes Leveling [b] Dong Zhou [a] Gareth J. F. Jones [b]
Vincent Wade [a]

a: Centre for Next Generation Localisation
Knowledge and Data Engineering Group
Trinity College Dublin
Dublin 2, Ireland
{ghorabm, vincent.wade}@cs.tcd.ie, dongzhou1979@hotmail.com
b: Centre for Next Generation Localisation
School of Computing
Dublin City University
Dublin 9, Ireland
{jleveling, gjones}@computing.dcu.ie

Abstract

This paper describes the collaborative participation of Trinity College Dublin and Dublin City University in the Log Analysis for Digital Societies (LADS) task of LogCLEF 2009 track. An analysis of multilingual search logs was carried out 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. Our 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.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods; Linguistic processing*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation; Search process*

General Terms

Experimentation, Measurement

Keywords

Cross-Language Information Retrieval, Log File Analysis, Query Reformulation, Action Patterns

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 task is based on query logs from The European Library[1] (TEL),

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

which is a portal forming a single interface for searching across the content of many European national libraries. In contrast to most other tasks at CLEF, LADS does not follow the standard procedure to measure the performance of a retrieval system, rather it aims at analysing multilingual search behaviour.

This participation is undertaken as part of ongoing research activities within the: Centre for Next Generation Localisation (CNGL)[2]. The CNGL is investigating novel technologies that address the key localisation challenges of volume, access and personalisation. The Digital Content Management (DCM) track is a subdivision of the CNGL project and is, in part, directed towards advancements in the personalisation of Cross-language Information Retrieval.

The LogCLEF dataset contains log entries for different types of user interactions (hereafter: actions) with the TEL portal, collected between January 2007 and June 2008. A more detailed description of the task and the dataset can be found in [2] and at the LogCLEF web page[3].

We analysed the logs to investigate the following hypotheses:

- Users from different linguistic or cultural backgrounds behave differently in search.
- There are patterns in user actions which could be useful for stereotypical grouping of users.
- User queries reflect the mental model or prior knowledge of a user about a search system.

We believe that the findings of such investigations can be exploited to personalise and improve cross-language search.

The remainder of this paper is organised as follows: Section 2 gives a brief description of the logs and the preprocessing operations performed on them, Section 3 discusses the log analysis along with the obtained results, and the paper ends with conclusions and outlook to future work in Section 4.

2 Brief Description of Logs and Preprocessing Operations

A log entry is created in correspondence with every user action. The log entry contains 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 the study focused on were:

- search_sim: simple text box search.
- search_adv: advanced search by the specific fields of title, creator (i.e. author, composer, etc), subject, type (e.g. text, image, etc), language, ISBN, or ISSN.
- view_brief: clicking on a certain 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 for searching or browsing.

A first analysis of the provided sample log data revealed that the data set had to be preprocessed to solve problems including character encodings, syntactically malformed queries (missing quotation marks, additional parentheses), and actions and attribute values that were not described in the guidelines.

The following were deleted from the dataset: entries having unrecorded session ids (empty or null value), search attempts having empty queries, sessions with missing actions, and sessions having unrecorded or malformed language acronyms. The original number of records was 1,866,330 records, which was reduced to 1,632,044 after the cleaning process (approximately 12.6% of the records were deleted). Furthermore, inconsistencies in the format of the stored queries were dealt with, such as trimming unnecessary brackets, quotations, and white spaces. Moreover, query keywords were extracted and stored separately in an additional table for performing term-based statistics.

[2] http://www.cngl.ie/

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

A major part of data preprocessing was the reconstruction of user sessions. The log entries contain anonymised user IDs and abbreviated IP addresses of the computers used to access the TEL system as well as session IDs. In addition, there is a timestamp attached to each logged action. As the IP address is not sufficient to distinguish between single users and the user ID may be associated with a guest account, session reconstruction was solely based on the session IDs. The session ID was used to reconstruct the actions in single sessions and the timestamp was then used to sort the actions. Session duration was calculated as the time interval between the timestamp of the first action and the timestamp of last action in the session. To identify the first action in a session, a login action was added before the first logged action.

3 Analysis of Log File Entries

3.1 General Statistics

Table 1 and Table 2 present descriptive statistics of the logs. Only a small proportion of the actions were performed by signed-in users (0.76%) compared to the number of actions recorded for guest users (99.34%). This may indicate that users find it easier, and /or perhaps more secure, not to register and sign into a web search system. Such behaviour sets a challenge to fine-grained personalisation (individual user profiling).

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

Table 2: Central Tendencies.

Item	Average	Median
Actions per session	8.39	4
Queries per session	2.81	2
Session duration (mm:ss)	17:20	01:35

The logs exhibited outliers, such as the existence of sessions with either a very large number of actions or just a single action (max: 1,093; min: 1), sessions with extremely long or short duration (max: 116 days; min: 1 second), and sessions with a large number of queries (max: 179; min: 1).

User actions were classified into four broad categories: *Search* (query actions), *Browse* (browsing / navigating result pages of TEL web site, excluding following links leading to the browsing of external web sites), *Collection* (actions involving limiting the search scope by the selection of a collection, theme or subject), and *Other* (all other actions).

Table 3 shows the distribution of actions along the broad classification. It was observed that search actions formed nearly half of the browsing actions. Further experimentation will look into the effect of query adaptation on the ratio between the number of search and browsing actions.

A large number of user actions (11%) were performed before attempting the search, including the specification of certain collections or subjects 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 where the user or group can be identified.

The study focused on the six actions mentioned in the previous section, as they had a high frequency, which indicates typical behaviour for library search. In regards to searching, it was found that there was

Table 3: Broad classification of actions.

Classification	Percentage
Search	28.17
Browse	56.78
Collection	11.34
Other	3.70

a great inclination towards using the simple search feature of the TEL portal (16.14% of total actions) compared to using the advanced search feature (4.35% of total actions). Further investigation is required to determine the degree of success of using the simple search feature compared to the advanced search feature in terms of satisfying the search with a fewer number of queries.

Another inclination in user actions was found for the pre-selection of a single collection for search, which occurred considerably more frequently than the pre-selection of multiple collections (col_set_theme was 7.13% of total actions and col_set_theme_country was 2.72%). This suggests that users who seek to limit their search tend to be very specific in selecting a designated collection. This may arise from their previous experience with the search portal, where users found that certain collections have a higher degree of satisfying their information needs. This finding may suggest further research towards performing a re-ranking operation for the list of collections depending on collection selection history in the logs.

3.2 Query Reformulation

Query reformulations, within each session, were classified into term addition, term deletion, term modification, and term change. Term modifications are term changes for queries containing a single search term. Query reformulation can also be classified by the type of the term(s) it affects or the type of transformation between two terms (e.g. translation). For this analysis, no differentiation was made between queries submitted under different interface languages of the portal, because i) the major part of the queries were submitted under English, and so, 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 (translation).

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 (of course, some of the latter actions co-occur with modifying search terms). As some users switch from the simple to the advanced search interface of the TEL portal, related queries are difficult to identify if different types of queries are considered. For the following experiment, search terms were extracted from all queries in the logs in order to identify how users typically modified a query. Only successive searches on the same topic were considered. To identify queries that were about the same search topic, the following approach was used: consecutive queries must have at least one search term in common (if the query contains more than one search term) or the search term in the queries must have a Levenstein distance [3] less than three. A query parser was implemented to extract the search terms from the query log and identify the type of query modification and the most frequent changes.

Table 4 shows the reformulation classes based on the top-50 reformulations. It was observed that 16% of term additions, 24% of term deletions, and 28% of term changes were stopwords or changes to stopwords (e.g. prepositions). Such changes might make sense under the assumption that people sometimes do copy and paste to directly insert a number of search terms in a search box, and so they might have just pasted some unwanted stopwords into the TEL search box by mistake. 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.

It was observed that proper nouns and single characters (mostly denoting initials of names) made up 62% of term additions, 46% of deletions, 20% of modifications, and 10% of changes. In contrast, term modification mostly affect morphological variations (e.g. plural forms, derivation, etc) and translations (26% and 24%, respectively). Such modifications would not have any effect on the search results if the TEL system performs stemming.

Table 4: Top-50 changes to terms in subsequent related queries.

type	brief description	example	add	del	mod	chg
ST	use of stopwords	*"a"* → *"the"*	8	12	3	14
BL	use of Boolean operators	*"AND"* → *"OR"*	2	3	0	6
CC	change of lowercase or uppercase	*"europe"* → *"Europe"*	0	0	3	0
SC	spelling change	*"wolrd"* [!] → *"world"*	0	3	2	2
CH	use of special characters	"*" at the end of term	3	0	0	2
LC	language code change	*"ita"* → *"eng"*	1	1	0	10
RT	related terms	*"triangulum"* → *"quadratum"*	-	-	1	2
MO	morphologic variant	*"city"* → *"cities"*	-	-	13	1
TR	translation or transliteration	*"power"* → *"kraft"*	-	-	12	2
PN	change proper noun/name	*"mozart"* → *"amadeus"*	21	13	10	4
PI	single character (initials)	*"elzbieta"* → *"e"*	10	10	0	1
DT	date/number change	*"1915"* → *"1914"*	2	3	0	3
OT	unknown change/other	*"test"* → *"toto"*	3	5	5	3

It was also observed that special characters (e.g. wildcards, which are used for more complex query operations) were rarely used. Moreover, a small number of changes involved the use of related terms (including narrower terms or broader terms). Also, only a small number of changes involved changing Boolean operators (e.g. *"AND"* → *"OR"*), dates (*"2005"* → *"2006"*), or numerals (*"i"* → *"1"*).

Furthermore, it was observed that 20% of term changes involved changing the language code. It seemed that users had an inclination of specifying the language (interface language of the whole portal and/or the language field of the advanced search page) in combination with the specification of a collection to search within. Such behaviour may indicate that users were not generally aware of the purpose of the features concerning the change of the language. For example, concerning the language field in advanced search, it might be the case that they interpreted it as a means of automatically translating query terms into a different language instead of a means of filtering out books which were not written in the specified language.

The analysis of query reformulations supports our hypothesis that some users have little knowledge of the search system, as they include stopwords and even change them (assuming TEL ignores stopwords as is commonly done by search engines). It can be inferred that the query edit behaviour of such users is focused more on domain, rather than on IR. This group will correspond to novice users. On the other hand, a small group of users used advanced query operators such as wildcards in their queries, which corresponds to experienced users.

3.3 Interface Languages

In an attempt to investigate the relation between language and search behavior, several variables were studied across the interface language selected by users of the portal. Recorded 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 worth mentioning here that the selection of an interface language does not necessarily imply the language of the query that the user inputs. One possible cause for the bias towards English, aside from its inherent popularity, is that it is the default language in the portal. Possible ways to avoid this bias would be to force the user to specify a language before attempting the search, or to have the default language automatically specified according to client machine's IP address.

Table 5 states the average and median for the number of actions and queries per session. Users of the English language exhibited the lowest average in both measures. This may suggest that users who used the TEL portal in languages other than English had to submit more queries to satisfy their information needs.

Among the rest of the languages, the Slovenian language stood out as an exceptional case where the average number of actions per session was 27.43 (median: 13) and the average number of queries per session was 6.82 (median: 3). Further investigation is required to determine the cause of this observation.

Table 5: Actions and queries per session across interface languages.

	Actions per session		Queries per session	
Language	Average	Median	Average	Median
English	7.97	4	2.7	2
French	9.2	5	3.01	2
Polish	8.63	5	3.14	2
German	9.37	5	3.03	2
Italian	11.3	6	3.73	2

The frequency distribution of the six main actions across each of the five interface languages is shown in Table 6. It was found that for Italian, the ratio between the number of simple search actions and advanced search actions was 2.34, while the ratio for the other four languages was 3.51 on average. A probable cause for this may be that a greater number of queries submitted under the Italian language were not satisfied through simple search, and users had to reformulate their queries through advanced search. Further investigation is needed to validate this assumption.

It was also found that users of the Polish language seem to have a higher rate than others in using the feature of specifying a single collection before attempting the search. On the other hand, English was found to have the lowest rate of usage of this feature. This finding supports our hypothesis that users from different linguistic or cultural backgrounds behave differently in search.

Table 6: Actions distribution across languages.

Language	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%

3.4 Term Frequencies and Categories

As part of the analysis, the number of terms per query and the top queried terms, for both, simple search and advanced search were studied. Table 7 shows the number of terms per query, starting at queries made up of one term and up to queries made up of six or more terms. The percentage of queries made up of three terms or less was 83.12% in simple search and 69.42% in advanced search. For both types of search, the frequency of query length was inversely proportional to the number of terms per query, with only one exceptional value for advanced search at three terms per query. This trend of users entering fewer search terms increases the ambiguity of the query, and thus sets challenges for query disambiguation.

In advanced search, the percentages of queries made up of three or more terms surpass those of simple search. This may suggest that users are encouraged to enter more search terms by the availability of multiple input fields. However, it is important to point out here that throughout the experiments, Boolean connectors were not removed from queries. This might be another reason behind the difference in percentages as the advanced search feature automatically adds connectors between search fields; thus it would naturally incur more connectors than simple search. Nevertheless, part of the analysis revealed that users still used connectors in simple search, although they have no effect (treated as normal terms).

Table 8 shows the average and median of the number of terms per query across interface languages. It can be seen that German showed the lowest average in both types of search (simple search: 1.77; advanced search: 2.6). Moreover, part of the analysis revealed that German exhibited the largest distribution of queries made up of just one term, while English exhibited the smallest. This may be because the German

Table 7: Number of terms per query.

	Simple Search		Advanced Search	
Terms	Frequency	Percentage	Frequency	Percentage
1	108,049	41.03%	20,038	28.23%
2	76,993	29.24%	13,867	19.54%
3	33,855	12.86%	15,364	21.65%
4	17,893	6.79%	7,595	10.70%
5	10,267	3.90%	6,086	8.58%
6+	16,295	6.19%	8,020	11.30%

Table 8: Number of terms per query across interface languages.

	Simple Search		Advanced Search	
Language	Average	Median	Average	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

language allows noun compounds written as single words, which can express complex topics as a single word. Such differences between languages forms an important point of focus for our ongoing research.

We compared the average number of terms per query of simple search with the results reported in [1], which was a similar study applied on search logs from AlltheWeb.com[4] (a European web search engine that allows limiting the search to documents in a language of choice). With the exception of English, the averages for the languages were found to be approximately the same for both, TEL and AlltheWeb.com logs, in spite of the fact that the former is a library search system and the latter is a general search engine.

Part of the log analysis, involved the extraction of the top twenty occurring search terms for each interface language, excluding stopwords. A term was only counted once in a session, even if it appeared multiple times in the 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, such as: text, image, sound, etc). These categories were mostly based on the fields of the advanced search in TEL portal, except for the location category.

Figure 1 shows the average category distribution of the five languages combined. In simple search, most of the search terms came under the creator and title categories (30% and 28% respectively). The same was exhibited for advanced search, though with a greater inclination towards the creator category (45%). This may indicate that user searches were better satisfied by including document creator in the query.

Figure 2 shows category distribution of the top twenty search terms for each of the five languages in simple search and advanced search. A large difference was observed in user search behaviour between different languages. For example, in English, 40% of the terms were subjects and 10% were creators, while in German, rather contrasting values were observed where 45% of the terms were creators, and only 10% of the terms were subjects. Such findings reflect the differences between users of different languages and will contribute towards further research in multilingual query adaptation, perhaps suggesting a different adaptation strategy for each language or group of languages.

[4] `http://www.alltheweb.com/`

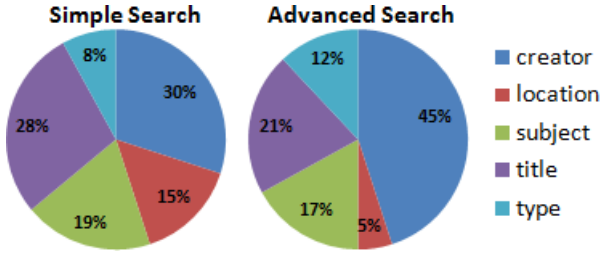

Figure 1: Distribution of term categories.

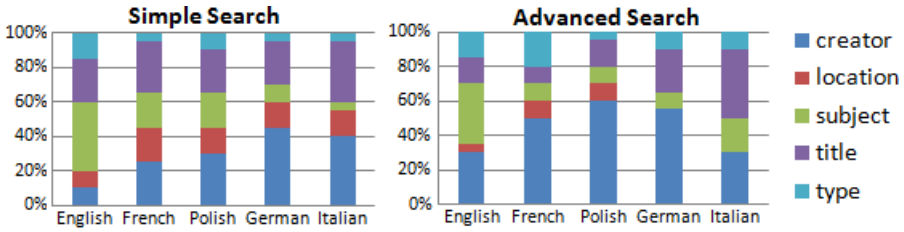

Figure 2: Distribution of term categories across languages.

3.5 Action Sequences and Durations

Figure 3 shows the average time between actions. Most of the time is spent before changing search options (e.g. col_set_desc, col_set_subj, etc.), which is illustrated by the bright areas in the diagram. The dark areas correspond to actions taken almost immediately after another (e.g. search_sim as the first action after login), or to two actions which never follow each other.

Figure 4 shows the frequency of two subsequent actions taken by the users. The most frequent action sequences consist of searching and viewing results, searching and changing options, and switching between results views (e.g. view_brief-view_full).

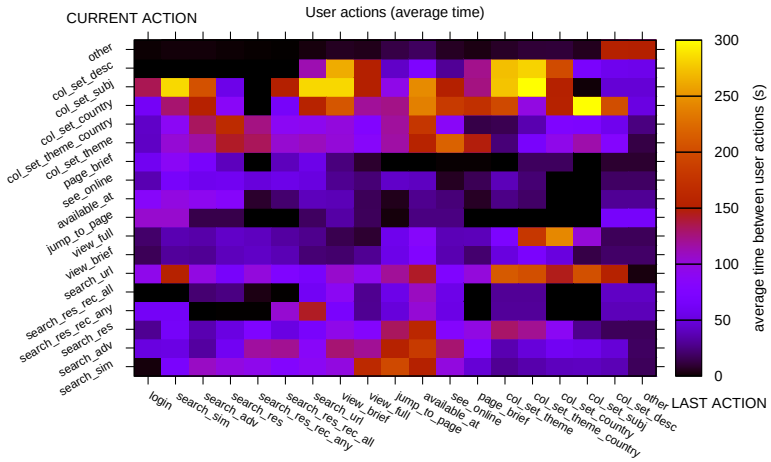

Figure 3: Analysis of time to next user action.

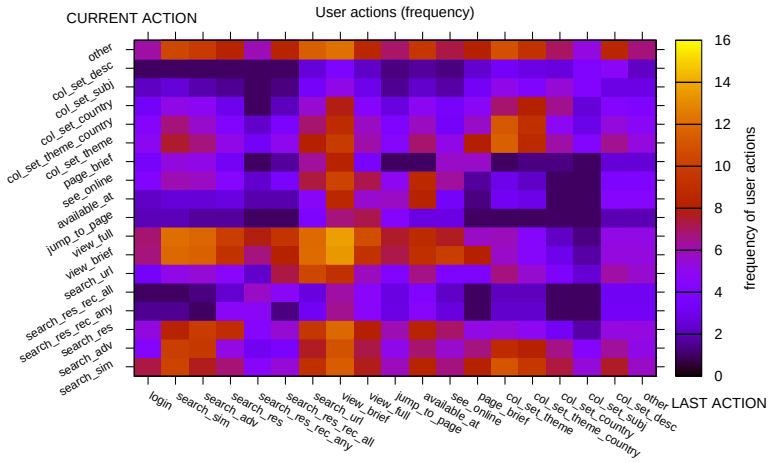

Figure 4: Analysis of most frequent next user action.

Table 9 and Table 10 show patterns of two and three subsequent actions. Each table points out the top most occurring pattern, as well as some other interesting patterns that have a rather high frequency. Related patterns are grouped together in the tables. It is observed that more users, after performing a search action, seem to go directly to view a full record (click for expansion) rather than clicking on a collection first (view_brief) then 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.

It can also be observed that users seem to get confused between two features (available as combo boxes) that both appear on the main page of the TEL web site. The two features 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 subsection 3.1 it can be inferred that users prefer the feature of choosing a single collection (col_set_theme). Perhaps deeper analysis of such patterns may introduce certain changes to the TEL portal's GUI.

Table 9: Sequential action patterns for two subsequent actions.

Action 1	Action 2	Frequency
view_full	view_full	153,952
search_sim	view_full	112,562
search_sim	view_brief	86,625
search_adv	view_full	32,356
search_adv	view_brief	28,732
col_set_theme	search_sim	40,044
col_set_theme_country	search_sim	12,397

Table 10: Sequential action patterns for three subsequent actions.

Action 1	Action 2	Action 3	Frequency
view_full	view_full	view_full	79,346
col_set_theme	search_sim	view_full	18,562
col_set_theme	search_sim	view_brief	16,446
col_set_theme_country	search_sim	view_brief	2,530
col_set_theme_country	search_sim	view_full	8,458
col_set_theme	col_set_theme_country	col_set_theme	4,735
col_set_theme_country	col_set_theme	search_sim	3,159

4 Summary and Outlook

This paper has described an analysis of the multilingual search logs from TEL. The results of the analysis support our 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 personalise and improve cross-language search. One suggestion concerning the logs would be to include the results that the users viewed. Such logs would be more informative and thus would contribute to a more thorough analysis.

Furthermore, the results also suggest 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 in order to retrieve more relevant results (term expansion, deletion, or modification); (2) offering focused online help if a user spends an uncharacteristically long time between some actions while using the TEL system or if a user performs a sequence of actions that may logically be inconsistent or opposite to each other; (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.

Acknowledgements

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

References

[1] Bernard J. Jansen and Amanda Spink. An analysis of web searching by European AlltheWeb.com users. *Information Processing & Management*, 41(2):361–381, 2005.

[2] Thomas Mandl and Giorgio di Nunzio. Overview of the LogCLEF track. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *CLEF 2009 Working Notes*, 2009.

[3] Robert A. Wagner and Roy Lowrance. An extension of the string-to-string correction problem. *JACM*, 22(2):177–183, 1975.