TCD-DCU at LogCLEF 2009

An Analysis of Queries, Actions, and Interface Languages

M. Rami Ghorab **Johannes Leveling** **Dong Zhou**

Gareth J. F. Jones **Vincent Wade**

Centre for Next Generation Localisation

NDP National Development Plan 2007 - 2013

sfi science foundation ireland

DCU Dublin City University

UCD DUBLIN University College Dublin

University of Limerick

Trinity College Dublin

Motivation

- We analysed the TEL logs with the aim of investigating 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.

Pre-processing & General Statistics

- Original number of records in the TEL logs was: 1,866,330.
- Reduced to 1,632,044 after some cleaning and pre-processing operations (approximately 12.6% of the records were deleted).

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

1. Linguistic & Cultural Differences

Users from different linguistic or cultural backgrounds behave differently in search

- Investigating relation between search behaviour and interface language selected by users.
- Recorded actions were distributed among 30 languages.
- Top five languages in terms of the number of actions: English (86.47%), French (3.44%), Polish (2.17%), German (1.48%), and Italian (1.39%).

1. Linguistic & Cultural Differences

- Actions & queries per session:

Interface Language	Average No. of actions/session	Average No. of queries/session
English	7.97	2.7
French	9.2	3.01
Polish	8.63	3.14
German	9.37	3.03
Italian	11.3	3.73
Slovenian	*27.43*	*6.82*

1. Linguistic & Cultural Differences

- Distribution of common actions:

Interface Language	Action					
	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%

1. Linguistic & Cultural Differences

- Average number of terms per query:
  - German also exhibited the largest distribution of queries made up of just one term, while English exhibited the smallest.
  - Probably because the German language allows noun compounds written as single words.

Interface Language	Simple Search	Advanced Search
English	2.38	3.05
French	2.09	2.85
Polish	1.89	2.59
German	1.77	2.6
Italian	2.09	3.17

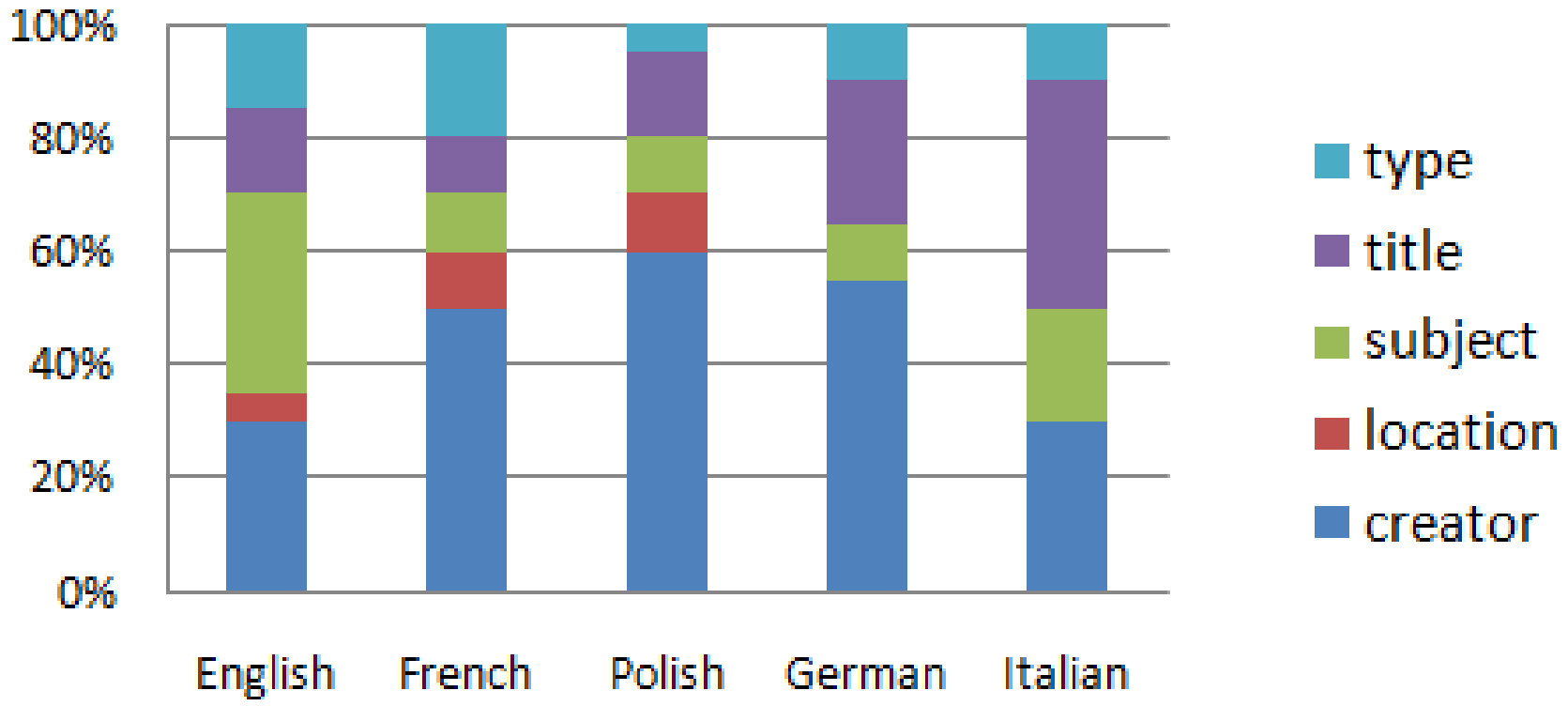
1. Linguistic & Cultural Differences

- Top 20 search terms for each interface language were analyzed, and were divided into five categories:
  - **Creator:** author, composer, artist, etc.
  - **Location:** cities, countries, etc.
  - **Subject:** History, Art, Science, etc (as per Dewey Decimal Classification).
  - **Title:** book/document titles, proper nouns, and common nouns
  - **Type:** document types, such as text, image, sound, etc.

- Mostly based on the fields of the advanced search in TEL portal.

1. Linguistic & Cultural Differences

- Distribution of term categories:



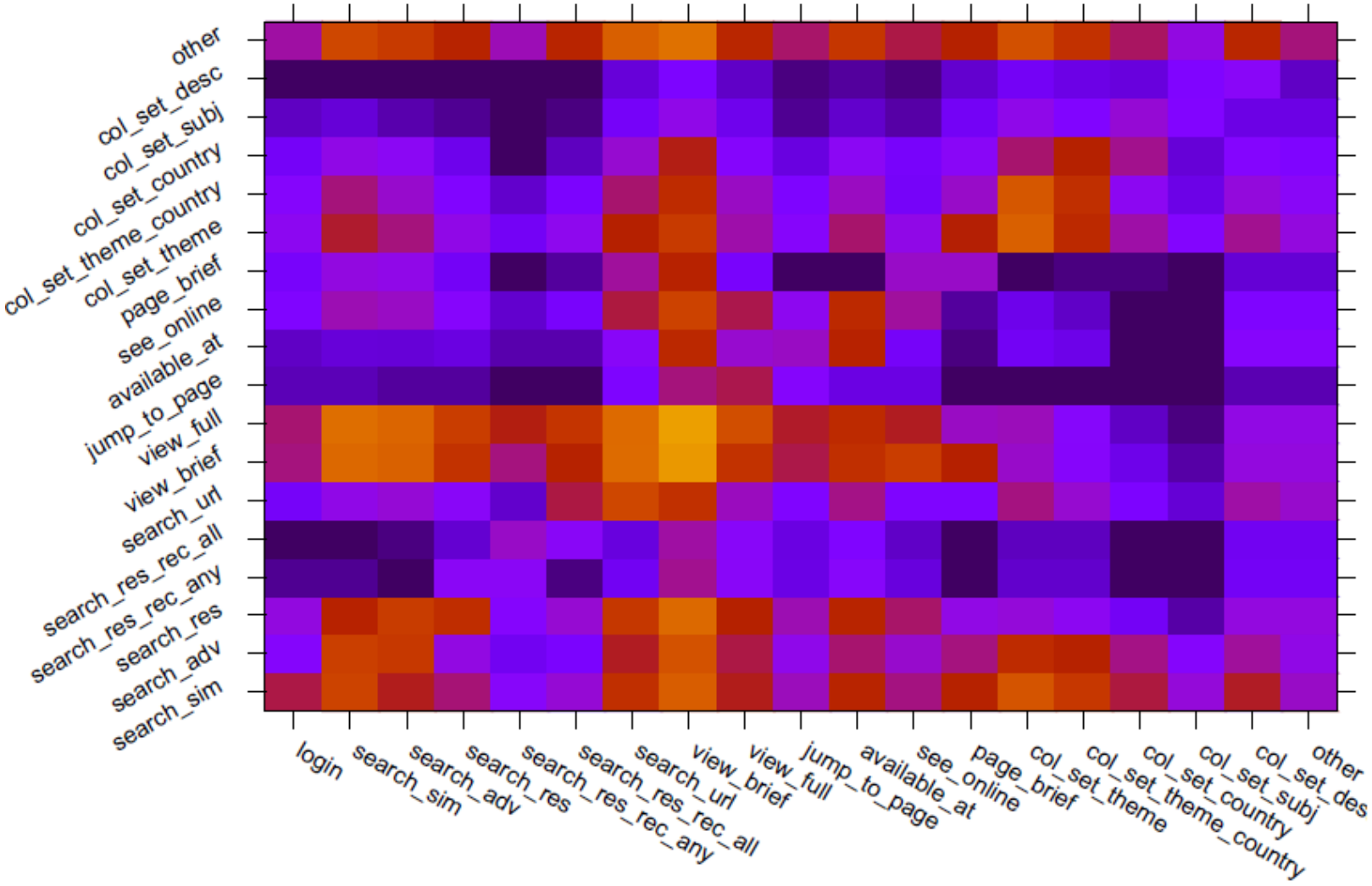
2. Sequence Patterns in User Actions

There are patterns in user actions which could be useful for stereotypical grouping of users.

- Some interesting two & three action sequences:

Action 1	Action 2	Action 3	Frequency
search_sim	view_full	-	112,562
search_sim	view_brief	-	86,625
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

2. Sequence Patterns in User Actions



3. Queries Reflect Users' Mental Model

User queries reflect the mental model or prior knowledge about a search system.

- To identify successive related queries about the same topic in the same session, the following approach was used:
  - Multiple term queries: must have at least one search term in common.
  - Single term queries: must have a Levenshtein distance less than three.
- Query reformulations were classified into term addition, term deletion, term change, and term modification (changes to single-term queries).

3. Queries Reflect Users' Mental Model

type	Descriptioon	Example	Add	Del	Mod	Chg
ST	Use of stopwords	"a" → "the"	16%	24%	6%	28%
CC	Change of case	"europe" → "Europe"	0%	0%	6%	0%

- Some users have little knowledge of the search system, as they modify stopwords and change letter case.
- It can be inferred that the query edit behaviour of such users is focused more on topic, rather than on IR.

3. Queries Reflect Users' Mental Model

type	Description	Example	Add	Del	Mod	Chg
BL	Use of Boolean operators	"AND" → "OR"	4%	6%	0%	12%
CH	Use of special characters	"*" at the end of term	6%	0%	0%	4%

- A small number of users used advanced query operators such as wildcards in their queries.
- Such behaviour reflects that they are users with good experience with search systems.

3. Queries Reflect Users' Mental Model

- Users had an inclination of specifying the language (interface or in adv. search) in combination with a collection specification.
- May indicate that users were not generally aware of the purpose of the features concerning the change of the language.
- Example: concerning the language field in adv. search, users may have interpreted it as a means of automatically translating query terms into a different language instead of a means of filtering out documents which were not written in the specified language

Conclusion

- Different user behaviour was exhibited in the logs for users from different linguistic and cultural backgrounds.
  - We can adopt a different query expansion strategy for each language or group of languages.
  - We can pre-select certain collections and/or re-rank their order in the results page.
- User action patterns and common specification of preferences were observed.
  - User profiling may help save user effort by automatically adjusting the search environment where the user or group can be identified.
  - We can provide interactive help/suggestions with the search box depending on the type of user (novice/expert).

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

NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD DUBLIN University College Dublin
University of Limerick
Trinity College Dublin


CNGL
Centre for Next Generation Localisation
Thank you
for your attention
NDP
National Development Plan 2007 - 2013
sfi
science foundation ireland
DCU
Dublin City University
UCD DUBLIN
University College Dublin
University of Limerick
Trinity College Dublin