Regional Effects on Query Reformulation Patterns

Steph Jesper, Paul Clough and Mark Hall

Information School, University of Sheffield, United Kingdom

Abstract. This paper describes an in-depth study of the effects of geographic region on search patterns; particularly query reformulations, in a large query log from the UK National Archives (TNA). A total of 1,700 sessions involving 9,447 queries from 17 countries were manually analyzed for their semantic composition and pairs of queries for their reformulation type. Results show country-level variations for the types of queries commonly issued and typical patterns of query reformulation. Understanding the effects of regional differences will assist with the future design of search algorithms at TNA as they seek to improve their international reach.

1 Introduction

The user's context, including individual differences and search task, are known to affect the way people search for information [1]. In this paper we focus on whether users searching from different countries exhibit different search patterns, in particular when reformulating queries. Query reformulation is a common part of users' information retrieval behavior [2] whereby search queries are adapted until the user fulfills their information need, or they abandon their search. Although query reformulation has been extensively studied, there has been little investigation into the effects of regional variances on query reformulation, even though users' demographics, such as their cultural background and language abilities are known to affect their searching behavior [3, 4]. In this paper we investigate the effects of geographical region (country) on the queries issued and typical patterns of query reformulation for searches at The National Archives (TNA), the UK government's physical and digital repository for all government documents. Understanding how people reformulate queries under different situations can help improve search results [5].

2 Related Work

Query reformulation has been extensively studied in various contexts from web search to library catalogue usage. Approaches to study reformulations are typically based on manually analyzing the transitions between query pairs in a session [2, 6]. Alternatively, automatic techniques have also been used to learn types of query reformulation [5]. Query reformulations have commonly been grouped into three main types: specialization, generalization and parallel moves. The first type reflects the situation in which a user refines a query to be more specific, typically by adding terms to a query. The second type reflects a user generalizing the query, typically by remov-

adfa, p. 1, 2011.

al of query terms. The final type indicates where a user changes to a new aspect of the *same topic*. Findings from previous studies have generally shown that parallel moves are the most common form of reformulation, followed by specializations and then generalizations [4-7]. Various studies have also explored the effects of cultural background on search behaviors. This includes comparing queries originating from different locations [5, 8], as well as users searching with varying language ability [3]. Most relevant to this paper are the studies by Spink et al. [4] and Boldi et al. [5]. Spink et al. analyzed the searching behavior of European users of FAST compared with US users of Excite. Results highlighted clear differences in the topics searched and search behaviors across the two countries, such as the vocabulary of queries used and query lengths. Boldi et al. compared query reformulation patterns identified in query logs from Yahoo! UK and Yahoo! US. Differences in query reformulation patterns could be observed between the UK and US search engine logs, with the UK data displaying higher proportions of specializations and parallel moves.

3 Methodology

Sessions comprising ≥3 queries were extracted from a 2009 search log from TNA containing ~1.9 million queries. These were derived from web logs at TNA that record search interactions with various resources accessed through various search functionalities, including an online catalogue (http://discovery.nationalarchives.gov.uk). Sessions were demarcated as interactions from the same IP address with a time interval between each interaction of <30 mins. The originating country of IP addresses was determined using the Maxmind Geolite geo-location database, which has an accuracy of 99.5% for country-lookup. The first 100 sessions for each region were extracted and analyzed manually providing 1,700 sessions and 9,447 queries for analysis. Queries were analyzed with respect to their linguistic structure and semantic composition (Person, Location, Specific item/object, Organization, Event and Other). Following the analysis of individual queries, the transitions between query pairs, Q_n and Q_{n+1}, were analyzed and categorized regarding type of query reformulation: *New* (N) Q_n and Q_{n+1} have no words in common (or Q_n=Q_0), *Specialization* (S) Q_n and Q_{n+1} are mostly identical but with more specific concepts used or material added, *Generalization* (G) Q_n and Q_{n+1} are mostly identical but with more general concepts used or material removed, *Parallel* (P) at least one phrase in Q_n is exchanged for a different phrase in Q_{n+1}, *Revision* (R) Q_n and Q_{n+1} contain the same information, but re-ordered, re-formatted or spelt differently and *Back* (B): Q_{n+1} is exactly identical to Q_{n-1}.

4 Results

Table 1 summarizes the proportion of queries containing particular semantic entities. Overall, 55.1% of all queries contain a Person element, 30.4% a Location and 21.5% reference to a Specific item/object. Person searches constitute over 50% of all queries submitted, and are the most popular element for all but five of the regions. Latin America shows a particularly strong preference for names of people (92% of South

American searches and 90% of Central American). This corresponds to previous findings showing query topics may vary based on cultural background [4].

	Per. (%)	Loc. (%)	Item (%)	Org. (%)	Event (%)	Other (%)
Australia & New Zealand	64.4	20.0	24.8	15.3	4.8	11.2
British Isles	65.0	22.4	20.3	13.9	3.2	11.2
Caribbean	51.4	49.3	26.1	14.2	9.0	16.1
Central Africa	39.2	48.7	28.6	16.4	6.9	22.4
Central America	90.1	9.0	4.2	2.9	2.4	5.9
East Asia	32.1	43.3	32.0	16.6	11	14.7
Eastern Europe	56.8	11.4	20.1	24.1	2.4	13.4
Middle East	40.7	44.8	23.6	22.3	17	24.0
Nordic Countries	64.5	16.6	15.2	15.4	2.6	15.6
North America	63.6	21.7	17.4	13.7	2.6	11.3
Northern Africa	31.9	56.6	19.6	14.4	3.6	24.9
South America	92.1	4.0	3.5	7.0	0.2	1.2
South-Central Asia	49.8	30.1	33.3	16.3	6.1	25.2
Southeast Asia	49.8	38.7	24.5	19.4	7.9	14.1
Southern Africa	69.2	21.2	13.0	18.2	10	17.5
Southern Europe	54.4	29.5	23.7	15.0	7.6	14.8
Western Europe	51.0	22	24.2	27.9	4.4	16.0
OVERALL	55.1	30.4	21.5	16.4	6.5	15.8

Table 1 Percentage of queries containing semantic entities

	N (%)	S (%)	G (%)	P (%)	R (%)	B (%)	Modal path
Australia & New Zealand	24.4	19.1	15.2	18.9	9.7	12.8	N→ S→B
British Isles	24.7	22.6	10.0	13.9	15.8	13.0	N→ S→B
Caribbean	19.9	19.9	7.6	22.1	15.6	14.9	N→ R→R
Central Africa	25.3	18.9	6.9	25.8	12.3	10.8	N→ R→R
Central America	23.7	15.4	9.4	12.1	25.9	13.6	N→ R→R→R→R
East Asia	20.7	17.2	10.7	20.7	15.2	15.6	N→ S→B
Eastern Europe	22.7	13.6	3.7	22.1	31.4	6.5	N→ R→R
Middle East	18.1	16.7	8.8	26.8	20.0	9.6	N→ S→P
Nordic Countries	24.6	15.0	5.8	14.4	24.6	15.8	N→ R→R
North America	23.6	16.7	10.6	23.6	16.9	8.5	N→ S→B
Northern Africa	21.3	18.3	10.3	22.3	15.4	12.4	N→ S→P
South America	23.5	6.5	7.9	8.8	43.0	10.2	N→ R→R
South-Central Asia	22.6	18.3	7.7	21.3	18.1	12.0	N→ S→B
Southeast Asia	19.9	19.3	7.0	24.2	20.0	9.6	N→ S→P
Southern Africa	20.3	21.5	11.3	20.8	11.7	14.3	N→ S→P
Southern Europe	22.1	19.0	8.5	16.5	21.2	12.6	N→ R→R→R→R
Western Europe	22.4	17.8	10.1	23.3	15.2	11.2	N→ S→B
OVERALL	22.2	17.6	9.0	20.4	18.9	12.0	N→ R→R

Table 2 Percentage of reformulation types and most frequent reformulation path

Table 2 shows the proportion of reformulation types for each of the regions and across all queries. Ignoring the results for New, overall the most common query reformulation is Parallel (20.4%), Revision (18.9%), Specialization (17.6%), Back (12%) and Generalization (9%). The results are similar to previous findings where results showed that parallel is the most common form of reformulation and specializa-

tions are performed more often than generalizations [2]. In this dataset we observe a high proportion of revision reformulations. This is partly explained by the high number of person searches, which commonly exhibit revision moves. Revisions are also generally higher for regions where English is not the first language. The final column in Table 2 shows the most frequent reformulation paths taken through the sessions. In most cases the session length consists of 3 queries and the most frequent path is N→ R→R due to the high number of revisions. Another dominant pattern is to start with a more general query and then specialize.

5 Conclusions

In this paper we investigate differences in search patterns that arise from users around the world searching for information from the UK National Archives. Similar to previous findings, regional variations are shown to exist in the semantic composition of queries and reformulation types. For example, visitors from the Middle East use longer queries and sessions, and seek locations more than they seek people; visitors from Latin America have a particularly high interest in people. Analyzing query logs enables a better understanding of how people interact with search engines and can be used to improve retrieval performance and enhance user interaction. For example, in this study the high proportion query revisions, particularly for countries that frequently search for person names, highlights the need for improvements in name matching.

6 References

1. Ford, N., Miller, D., & Moss, N. (2001) The role of individual differences in Internet searching: an empirical study, *Journal of the American Society for Information Science and Technology*, 52(12), 1049-1066.
2. Rieh, S.Y., & Xie, H. (2006) Analysis of multiple query reformulations on the web: The interactive information retrieval context, *Information Processing & Management*, 42(3), 751-768.
3. Zoe, L.R., & DiMartino, D. (2000) Cultural diversity and end-user searching: An analysis by gender and language background, *Research Strategies*, 17(4), 291-305.
4. Spink, A., Ozmutlu, S., Ozmutlu, H. C., & Jansen, B.J. (2002). US versus European Web searching trends. ACM SIGIR Forum, 36(2).
5. Boldi, P., Bonchi, F., Castillo, C., & Vigna, S. (2008). "From "Dango" to "Japanese Cakes": Query Reformulation Models and Patterns" In: *Proceedings of the 2009 IEEE/WIC/ACM international conference on Web intelligence, WI 2009,* 183-190.
6. Jansen, B.J., Zhang, M., & Spink, A. (2007) Patterns and transitions of query reformulation during web searching, *International Journal Web Information Systems*, 3(4), 328-340.
7. Huang, J., & Efthimiadis, E.N. (2009) Analyzing and Evaluating Query Reformulation Strategies in Web Search Logs, In: *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, 77-86.
8. Weber, I., & Castillo, C. (2010) The demographics of web search. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 523-530.