DCU-TCD@LogCLEF 2010: Re-ranking Document Collections and Query Performance Estimation

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

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

Abstract. This paper describes the collaborative participation of Dublin City University and Trinity College Dublin in LogCLEF 2010. Two sets of experiments were conducted. First, different aspects of the TEL query logs were analysed after extracting user sessions of consecutive queries on a topic. The relation between the queries and their length (number of terms) and position (first query or further reformulations) was examined in a session with respect to query performance estimators such as query scope, IDF-based measures, simplified query clarity score, and average inverse document collection frequency. Results of this analysis suggest that only some estimator values show a correlation with query length or position in the TEL logs (e.g. similarity score between collection and query). Second, the relation between three attributes was investigated: the user's country (detected from IP address), the query language, and the interface language. The investigation aimed to explore the influence of the three attributes on the user's collection selection. Moreover, the investigation involved assigning different weights to the three attributes in a scoring function that was used to re-rank the collections displayed to the user according to the language and country. The results of the collection re-ranking show a significant improvement in Mean Average Precision (MAP) over the original collection ranking of TEL. The results also indicate that the query language and interface language have more influence than the user's country on the collections selected by the users.

1 Introduction

LogCLEF at the Conference on Multilingual and Multimodal Information Access Evaluation 2010 is an initiative to analyse search logs and discover patterns

of multilingual search behaviour. The logs come from user interactions with the European Library (TEL)[3], which is a portal forming a single interface for searching across the content of many European national libraries.

LogCLEF started as a task of the CLEF 2009 evaluation campaign [1, 2], in which Trinity College Dublin and Dublin City University participated collaboratively. Our previous participation focused on investigating users' query reformulations and their mental model of the search engine as well as analysing user behaviour for different linguistic backgrounds [3].

In our LogCLEF 2010 participation, two sets of experiments were carried out on TEL logs. The first set of experiments investigated the assumption that query performance predictors not only can be used to estimate performance, but can also be utilised to analyse real user behaviour. For example, users' query modifications generally aim at improving the results and should reflect the fact that the users' queries improve over successive reformulations. This should also result in different values for performance estimates.

The second set of experiments was concerned with studying three attributes related to the users and the queries they submit, namely: the user's country (i.e. the location from which the query was submitted), the language of the submitted query, and the specified interface language by the user. The experiments investigated the influence of the three attributes on the collections (libraries or online resources) selected by the user. In other words, the experiments aimed at answering the question of whether or not the users' choice of collections is influenced by their location and language. The experiments involved re-ranking the list of collections displayed to the user. Typically, result re-ranking involves re-ranking a list of documents. However, in this study, the re-ranking process was applied to the list of collections displayed to the user by TEL. A scoring function that comprises the three attributes was used to determine the relevance of each collection to the current search in terms of country and language. Re-ranking alternatives were investigated by assigning different weights to the three attributes in the scoring function. The results of the collection re-ranking experiments showed a 27.4% improvement in Mean Average Precision (MAP) [4, 5] over the original list of collections displayed by TEL. The results also suggest that the query language and interface language have more influence than the user's country on the collections selected by the users.

The rest of this paper is organised as follows: Section 2 introduces related work; Section 3 describes preprocessing of the log files for session reconstruction; Section 4 introduces query performance predictors and details the analysis carried out with these measures on the query logs; Section 5 describes the collection re-ranking experiments and their results; and finally, conclusions and outlook on future work are given in Section 6.

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

2 Related Work

2.1 Query Performance Estimation

Query performance estimators have been widely used to improve retrieval effectiveness by predicting query performance [6–9], or to reduce long queries [10]. Hauff [11] gives a comprehensive overview over predicting performance of queries and retrieval systems.

Typically, these estimators are applied to user queries for automatic query modification, either before retrieval (i.e. pre-retrieval), based on information available at indexing time, or after an initial retrieval (i.e. post-retrieval), in which case additional information such as relevance scores for the documents can be employed as features for the performance estimators. The main objective is to improve effectiveness of information retrieval systems in general or for specific topics. However, these estimators have – to the best of the authors' knowledge – not yet been applied to real user queries to investigate if real query reformulations actually improve the search results.

2.2 Result Re-ranking

Result re-ranking is one of the well-known techniques used for search personalisation. [12–14]. Result re-ranking adapts to the user's needs in that it brings results that are more relevant to him/her to higher positions in the result list. It takes place after an initial set of results have been retrieved by the system, where an additional ranking phase is performed to re-order the results based on various adaptation aspects (e.g. user's language, knowledge, or interests).

Typically, the re-ranking process is applied to the list of retrieved documents. However this study investigates applying the re-ranking process to the list of collections displayed to the user, not the list of documents within a collection.

3 Data Preparation

Two datasets (action log files) were provided at LogCLEF: the first one (L1) contained action logs (queries and user interactions) from January 2007 to June 2008 from the TEL web portal (approximately 1.8 million records), and the second one (L2) contained action logs from January 2009 to December 2009 (approximately 76 thousand records). In addition, a third dataset (HTTP logs) was provided, which contained more details of the search sessions and HTTP connections, including the list of collections searched for each submitted query.

For the first set of experiments, queries from the two datasets of action logs were used. The logs were pre-processed following the approach in our previous experiments [3] where user sessions were reconstructed and consecutive queries on the same topic were determined. Session reconstruction was done by structuring all actions with the same session ID together and sorting them by timestamp. If the time between consecutive actions exceeded 30 minutes, the start of a new

session was assumed. In addition, only English queries and sessions containing at least three queries were considered. The resulting dataset contained 340 sessions from L1 with an average session length of 3.61 queries per session. The queries in this set consisted on average of 3.3 terms. Preprocessing of L2 resulted in 487 sessions with an average session length of 3.61 queries per session. Queries in this set consisted of 4.16 terms on average.

Preliminary experiments on the TEL data showed that only a few URLs from the logs were functional. Retrieving result pages for queries via their URL results in error codes of "not found" (HTTP response code 404). Furthermore, referrer URLs seem to be truncated to a certain number of characters, yielding malformed URLs. Thus, an external resource with reasonable coverage was used to compute query statistics (e.g. term frequencies and collection frequencies). The DBpedia[4] collection of Wikipedia article abstracts was indexed and the queries were evaluated against this index. The Lucene toolkit[5] was employed to index the document collection, using the BM25 model [15] for retrieval (although no results were retrieved for the experiments, because only pre-retrieval query performance estimators have been used).

For the second set of experiments, data from both the action logs and the HTTP logs of TEL 2007 was used. Due to time limitation and some technical errors in the recorded HTTP logs, the experiments were only conducted on a subset of the data. The selected subset comprised the submitted queries and clicked collections during the month of February 2007. This included approximately 1,800 queries from different languages.

4 Experiments on Estimates of Query Performance

4.1 Query Performance Estimators

Several pre-retrieval performance predictors were employed. Pre-retrieval means that the computation does not rely on calculating relevance scores or other post-retrieval information. Performance predictors are described in more detail in [6]. The following is the list of predictors used in our experiments.

Query length. For our experiments, the query length is defined as the number of unique query terms after stopword removal (see Equation 1). Zhai and Lafferty [16] have observed that the query length has a strong impact on the smoothing method for language modeling. Similarly, He and Ounis [17] find that query length strongly affects normalisation methods for probabilistic retrieval models. Intuitively, longer queries should be less ambiguous because they contain additional query terms which provide context for disambiguation.

$$QL = \text{number of unique non-stopword terms} \tag{1}$$

where q is the query.

[4] `http://dbpedia.org/`
[5] `http://lucene.apache.org/`

Query scope (QS). The query scope tries to measure the generality of a query by its result set size [9], i.e. by the cardinality of the set of documents containing at least one query term (see Equation 2). A query matching only very few documents can be considered very specific, while a query containing only high frequency terms will return a larger portion of the document collection.

$$QS = -log(n_q/N) \tag{2}$$

where n_q is result set size and N is total number of documents in the collection.

IDF based score (IDFmm). $IDFmm$ denotes the distribution of informative amount of terms t in a query. This IDF-based feature is introduced as $\lambda 2$ by He and Ounis [18]. We use INQUERY's IDF formula (Equation 3)

$$idf(t) = \frac{log_2(N + 0.5)/N_t}{log_2(N + 1)} \tag{3}$$

where t is a term and N_t is the number of documents containing t. In the $IDFmm$ formula (see Equation 4), idf_{max} and idf_{min} are the maximum and minimum IDF of terms from the query q.

$$IDFmm = \frac{idf_{max}}{idf_{min}} \tag{4}$$

Simplified query clarity score (SCS). The query clarity is inversely proportional to the ambiguity of the query (i.e. clarity and ambiguity are contrary concepts).

This score measures ambiguity based on the analysis of the coherence of language usage in documents whose models are likely to generate the query [7]. Because the computation of the original query clarity score relies on the availability of relevance scores, its computation would be time-consuming and result in post-retrieval estimates. We employ the definition of a simpler version proposed by He and Ounis [6], the simplified query clarity score (see Equation 5).

$$\begin{aligned} SCS &= \textstyle\sum_{t \in q} P_{ml}(t|q) \cdot log_2 \frac{P_{ml}(t|q)}{P_{coll}(t)} \\ &\simeq \textstyle\sum_{t \in q} \frac{qtf}{ql} \cdot log_2 \frac{qtf/ql}{tf_{coll}/token_{coll}} \end{aligned} \tag{5}$$

where $P_{ml}(t|q)$ is the maximum likelihood of the query model of term t in query q and $P_{coll}(t)$ is the collection model. tf_{coll} is the collection frequency, $token_{coll}$ is the number of tokens in the whole collection, qtf is the term frequency in the query, and ql is the query length (the number of non-stopwords in the query).

Similarity score between collection and query (SCQS). Zhao, Scholer, et al. [8] compute a similarity score between the query and the document collection to estimate pre-retrieval query performance (see Equation 6). For our experiments, we use the sum of the contributions of all individual query terms. As pointed out in [8], this metric will be biased towards longer queries.

$$SCQS = \sum_{t \in q} (1 + ln(tf_{coll})) \cdot ln(1 + \frac{N}{N_t}) \tag{6}$$

where N_t is the number of documents containing term t.

Average inverse document collection term frequency (avICTF). Kwok [19] introduced the inverse collection term frequency ($ICTF$) as an alternative to IDF. As $ICTF$ is highly correlated with search term quality, He and Ounis [6] proposed using the average $ICTF$ ($avICTF$) to predict query performance.

$$avICTF = \frac{log_2(\prod_{t \in q} \frac{token_{coll}}{tf_{coll}})}{ql} \tag{7}$$

The denominator ql is the reciprocal of the maximum likelihood of the query model of SCS in Equation 5. The use of $avICTF$ is similar to measuring the divergence of a collection model (i.e. $ICTF$) from a query model. Thus, $avICTF$ and SCS should have similar query performance estimates.

4.2 Experiments and Results

For the first part of our experiments on the TEL log data, the relation between real user queries and the query performance estimators are investigated. All searches from the reconstructed sessions from L1 and L2 were conducted on the indexed DBpedia abstract collection and the average values for all query performance estimators were calculated. Due to data sparsity, some results were grouped by query length (number of terms in a query) and by position (whether a query is the first one submitted in a session or is a further reformulation) into bins. Moreover, session estimates were analysed by distinguishing between sessions beginning with a short query and those beginning with a long query. This was performed with the aim of investigating if the first query in a session indicates the search behaviour during that session. For example, users may expand short queries and reduce long queries if their initial search is not successful.

For the experiments, short queries were defined as queries containing one to three terms and long queries were defined as queries with more than three terms. This value roughly corresponds to the well-known average number of terms in web queries (2-3 terms) [20] and to the average number of terms in the TEL logs (3-4 terms) [3]. Figures 1-4 show results from the analyses (L1 data on left, L2 data on right in Figures 1-3). Note that there is a difference between QL, which counts the number of unique terms in a query, and the query length in tokens (including duplicates), which is shown on the x-axis in Figure 1.

Comparing all six query performance estimators investigated (i.e. QL, QS, $IDFmm$, SCS, $avICTF$, and $SCQS$) to the query length, we find that, unsurprisingly, there is a high correlation between the query length (non-stopword terms) and QL (unique non-stopword terms). Query scope (QS) increases slightly for queries containing more than 3 terms, but remains almost constant for longer queries. $IDFmm$ slightly increases for longer queries. SCS has its maximum for the shortest queries (length 1-3 terms) decreases for medium queries, and increases again for most longer queries. Finally, $avICTF$ also decreases for medium-length queries, with higher values for short and long queries (see Figure 1). All estimators show similar behaviour on the sessions of L1 and L2.

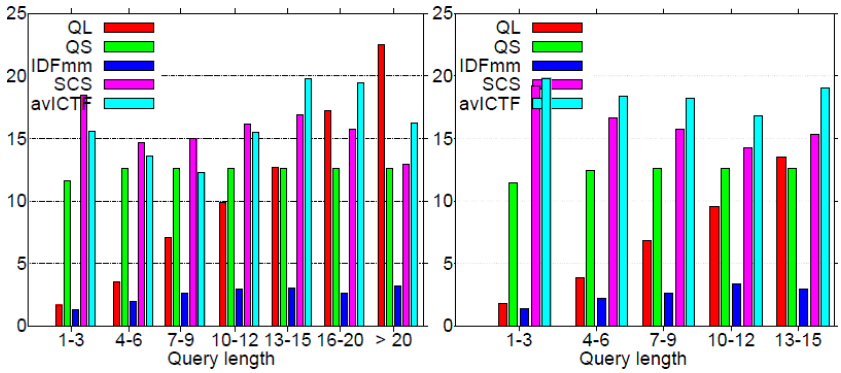

Fig. 1. QL, QS, $IDFmm$, SCS, and $avICTF$ vs. query length.

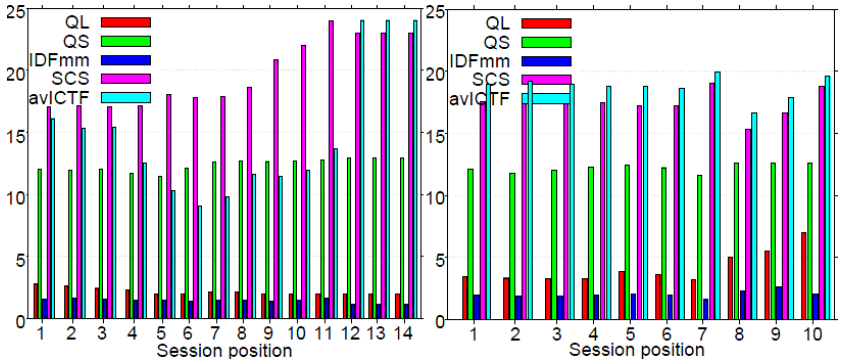

Fig. 2. QL, QS, $IDFmm$, SCS, and $avICTF$ vs. query position in session.

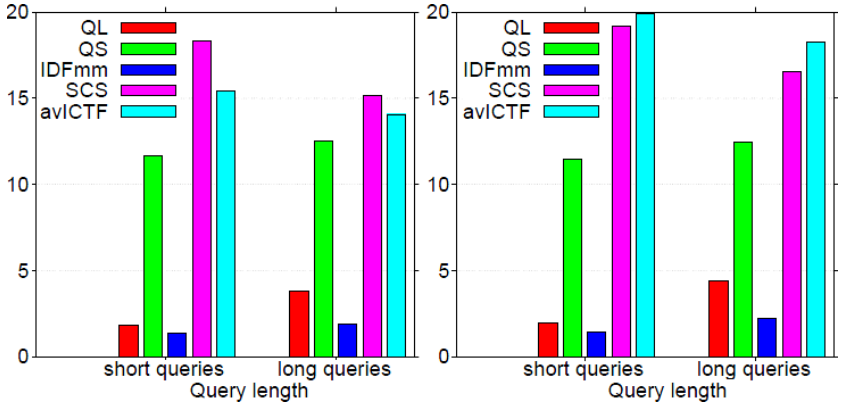

Fig. 3. QL, QS, $IDFmm$, SCS, and $avICTF$ vs. sessions with initial short and long query.

Next, the estimator values were calculated and compared across different positions in the sessions (see Figure 2). The positions correspond to users' query reformulations. The sessions from L1 are longer than sessions from L2 (14 queries maximum for L1 vs. 10 queries maximum for L2). With a few exceptions, most estimators have almost constant values for the different reformulations. One of the exceptions is $avICTF$, which drops after position 3 and has a sudden, but constant maximum after position 12. However, SCS increases after position 7 in both data sets. $SCQS$ compared to the query position shows that with a higher number of queries in a session (i.e. more query reformulations), $SCQS$ drops consistently (see Figure 4 left).

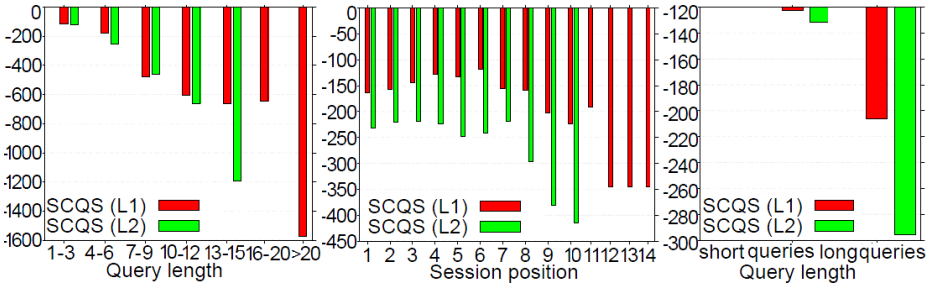

Fig. 4. $SCQS$ vs. query length, position in session, and sessions with initial short and long query.

Finally, dividing the data into sessions starting with a short query (less than 4 terms) and sessions starting with longer queries reveals the most interesting results. QL, QS, and $IDFmm$ are higher for longer queries in both data sets investigated, while SCS and $avICTF$ are higher for shorter queries (see Figure 3). The highest difference can be observed for $SCQS$ (see Figure 4 right) which shows much lower values for sessions with initial long queries.

4.3 Discussion

Most of the query performance predictors we experimented with the aim of measuring clarity, specificity, or ambiguity of query terms. For example, query length is presumed to be an indicator of result quality because the more terms a query contains, the more specific the query is, i.e. the more precise the results should be. However, our experiments did not show that longer queries are consistently less ambiguous, or more specific (Figure 1).

Our next analysis (Figure 2) was concerned with query reformulations in a single session. We investigated whether queries become more successful after reformulation, i.e. if a user's query reformulation will achieve better performance (estimates). We also have not observed a simple or direct relationship between the query position and the performance estimates. Instead, the extracted user

sessions might contain data from users with different search strategies or data from users changing their search strategy in the middle of a session. This may be an explanation for sudden drops and increases in some performance estimators (see, for instance $avICTF$ in Figure 2 on the L1 data on the left).

Finally, $SCQS$ showed the most interesting behaviour. For longer queries, its value decreases consistently for both data sets. With higher query positions, its value typically decreases, and the differences for sessions beginning with short and long queries are huge.

A small sample of sessions was manually investigated in more detail. Several users reformulate their queries by using the same search terms in different fields (for the advanced field-based search) or to search in different indexes (e.g. "lima licinio" repeated for different fields). This type of modification might not actually improve the retrieved results. However, the current TEL portal already supports a catch-all field ("any field"), i.e. a field covering a search in all indexes. Possibly this field was not available at the time the query logs were collected.

Some users modify their original query, examine the results, and submit the original query again (e.g. "environmental comunication" is modified into "comunication" which is modified back to "environmental comunication" – note the spelling error in all queries). A similar behaviour can be observed for longer sessions. After several queries, the users seem to revert to previously submitted queries, probably because the TEL portal does not provide query suggestions and the users do not find relevant or new information allowing them to change their query formulation.

Finally, some queries contain more than one change at once (e.g. "music manuscripts anton bruckner" is changed to "bruckner anton sinfonie"). Thus, changes in performance predictors might also incorporate several changes (i.e. increases and decreases in quality of results) at once.

5 Re-ranking the List of Collections

5.1 Overview

When users submit a search to the TEL portal, they are presented with a list of collections on the left side of the screen and a list of results from the selected collection on the right side of the screen. A collection is either a library or an online resource that is associated with a certain country. The list of collections is presented to the user in alphabetical order of the countries' acronyms (two letters according to the ISO 3166-1 standard).

The users' queries and interactions with the portal were recorded in two datasets called: action logs and HTTP logs. This set of experiments involved studying different aspects regarding three attributes exhibited in the logs: user's country (location from which the query was submitted), query language, and interface language. In our experiments, the user's country was determined from the IP address recorded in the logs. The query language was detected using the Google AJAX Language API[6], which returns a confidence level associated with

[6] `http://code.google.com/apis/ajaxlanguage/`

the detected language of a query. Only queries that had a minimum value of 10% confidence level were used in the experiments, which reduced the number of queries subject to experimentation to 566 queries (from the 1,800 queries of February 2007).

Furthermore, the experiments involved studying the list of collections that were searched for each query, and the collections that the user clicked on. The study investigated re-ranking the list of collections displayed to the user with the aim of increasing the retrieval precision (across the collections, not the actual result documents) by bringing collections that match the user's country or language to higher ranks in the list. In other words, the re-ranking process focused on studying if users have a higher tendency to click on collections that belong to their country or language. For this part of the experiments, languages were assigned to a collection based on the official languages that are spoken in the country associated with that collection.

5.2 Descriptive Statistics

The following statistics are drawn from the selected subset of the logs. Figure 5 shows the distribution of users' countries on the left (i.e. countries that the queries came from) and the language distribution of queries on the right. It is noted that a large percentage (50%) of the queries were in English although the total percentage of queries coming from English-speaking countries (United States, United Kingdom, and Canada) was only 16%. This suggests that many users from non-English countries do not submit queries in their native language.

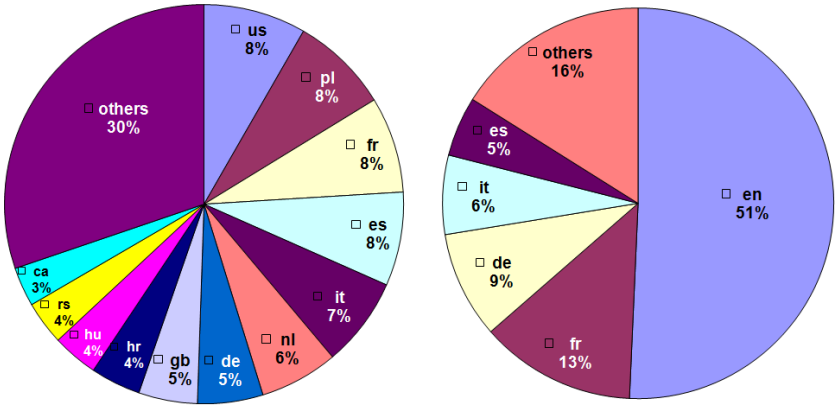

Fig. 5. Country distribution (left) and language distribution (right).

Figure 6 shows the relation between user's country, query language, and interface language for all queries (left) and for non-English queries (right). Surprisingly, it is noted that only 24% of the queries coming from a country are

in languages associated with that country. This is not necessarily an inclination towards using English in search, as similar results were observed when studying non-English queries, where less than 30% of the queries were submitted in a language that matched the country. Therefore, this suggests that the country attribute may not be very reliable to base the collection re-ranking decision upon. This was confirmed by further experiments (discussed below).

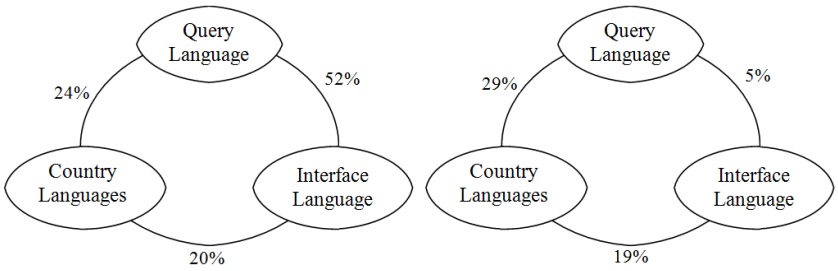

Fig. 6. Relation between attributes: all queries (left) and non-English queries (right).

5.3 Experimental Setup and Re-ranking Function

In order to evaluate the retrieval precision over the list of collections presented to the user, we used the collections that the user clicked on as implicit relevance judgements (i.e. binary relevance judgements where the clicked collections are assumed to be the relevant ones, and non-clicked collections are assumed to be irrelevant). This follows on the method adopted in [12]. However, the sense of relevance in our experiments is in terms of matching country and language.

Mean Average Precision (MAP) was used for evaluation as it is a common evaluation metric in information retrieval that rewards relevant items being ranked higher in the list [4, 5]. MAP was calculated across the queries in the selected subset of the data. The original ranked list of collections (i.e the one presented to the user by TEL) was used as the baseline for evaluation of retrieval precision (MAP score = 0.580). Several alternative re-ranked lists were investigated and compared to the baseline.

A result (collection) scoring function was used to re-rank the list of collections. The function was based on matching the three attributes with the collection's country and language as follows (where matching=1 and non-matching=0):

1. M_c: matching the user's country with the collection's country.
2. M_q: matching the query's language with the collection's language (i.e. matching with any of the official languages spoken in the corresponding country).
3. M_i: matching the interface language with the collection's language.

Each of the above attributes is multiplied by a scalar weight (W_c, W_q, W_i respectively) so as to control (and test) the degree of contribution of each attribute

in the function. Thus the collection scoring function becomes Equation 8

$$NewScore = (W_c \cdot M_c) + (W_q \cdot M_q) + (W_i \cdot M_i) \quad (8)$$

Finally, the collections are re-ranked based on descending order of the new score.

Unlike other collections, collection #a0000 had no country associated with it (a virtual collection of various European digitised books, images, maps, etc.). This would have caused it to be permanently placed at the end of the re-ranked list (i.e. get a zero score from the scoring function) because of not matching any country or language. Therefore, an exception was made regarding this collection; it was placed at the top of the re-ranked list (similar to TEL's ranking).

5.4 Experimental Results

Table 1 shows the MAP value for some selected re-ranking runs with alternative combinations of weights that ranged from 0.0 to 1.0. The results showed a significant improvement of 27.4% in retrieval precision for the re-ranked collection lists (with weights: W_c=0.1, W_q=0.3, W_i=0.6) over the baseline ranking. This improvement is statistically significant as per the T-test (with p=0.01) and Wilcoxon test (with confidence=99%). The results also suggest that there is a relation between the query language and interface language on the one hand and the clicked collections on the other hand. Moreover, these two attributes seem to have more influence than the user's country (location from which the query was submitted) on the selected collections by the users.

Table 1. Re-ranking MAP scores for alternative weight combinations.

Weight				
W_c	W_q	W_i	MAP score	Improvement
Baseline			0.580	-
0.1	0.3	0.6	0.739	27.4%
0.1	0.6	0.3	0.727	25.3%
0.6	0.1	0.3	0.719	23.9%
0.6	0.3	0.1	0.709	22.2%
0.0	0.0	1.0	0.709	22.2%
0.0	1.0	0.0	0.685	18.1%
1.0	0.0	0.0	0.600	3.4%

6 Conclusions and Future Work

The experiments using query performance predictors show that *SCQS* has a high correlation with query length and query position. Distinguishing between

sessions with an initial short or long query seems to indicate a different subsequent user behaviour. For example, users starting with a short query are more likely to expand their query to obtain more specific results, while users starting with a longer query probably reduce their query because the retrieved result set is too small. To obtain more conclusive results, we want to perform more experiments on larger query logs. The results might also differ from results obtained from performing the same analysis on web search logs, because of the different application domain (bibliographic search in the TEL portal), the slightly longer queries and possibly longer sessions, which include more query reformulations. Query performance estimators have been developed for web search and may thus be optimised for shorter queries and fewer search iterations. In conclusion, more consistent results may be found if the methods are applied on a larger data set.

The experiments of collection re-ranking based on matching the query language and the interface language with the collection's language showed a significant improvement in precision. This suggests that there is opportunity for improving the user's experience with multilingual search in the TEL portal if the user's language is taken into consideration. Future work will involve using machine learning techniques to learn a ranking function that optimises the weights of the three attributes: country, query language, and interface language. Moreover, we will attempt to work around the technical errors present in the dataset of TEL HTTP logs, which will enable us to conduct the experiments on the full dataset.

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 Dublin City University and Trinity College Dublin.

References

1. Mandl, T., di Nunzio, G.: Overview of the LogCLEF track. 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)
2. 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. Revised Selected Papers. LNCS, Springer (2010)
3. 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)
4. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. 1st edn. (1999)
5. Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. 1st edn. Cambridge University Press (2008)

6. He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In Apostolico, A., Melucci, M., eds.: String Processing and Information Retrieval, 11th International Conference, SPIRE 2004, Padova, Italy, October 5-8, 2004, Proceedings. Volume 3246 of Lecture Notes in Computer Science., Springer (2004) 43–54
7. Cronen-Townsend, S., Zhou, Y., Croft, B.W.: Predicting query performance. In: SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland, ACM (2002) 299–306
8. Zhao, Y., Scholer, F., Tsegay, Y.: Effective pre-retrieval query performance prediction using similarity and variability evidence. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W., eds.: Advances in Information Retrieval, 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings. Volume 4956 of Lecture Notes in Computer Science., Springer (2008) 52–64
9. Plachouras, V., Ounis, I., van Rijsbergen, C.J., Cacheda, F.: University of Glasgow at the Web Track: Dynamic application of hyperlink analysis using the query scope. In: Proceedings of TREC 2003, Gaithersburg, MD, NIST (2003) 646–652
10. Kumaran, G., Carvalho, V.R.: Reducing long queries using query quality predictors. In Allan, J., Aslam, J.A., Sanderson, M., Zhai, C., Zobel, J., eds.: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009, ACM (2009) 564–571
11. Hauff, C.: Predicting the Effectiveness of Queries and Retrieval Systems. PhD thesis, University of Twente. CTIT (2010)
12. Speretta, M., Gauch, S.: Personalized search based on user search histories. In: IEEE/WIC/ACM International Conference on Web Intelligence. (2005) 622–628
13. Pretschner, A., Gauch, S.: Ontology based personalized search. In: 11th IEEE International Conference on Tools with Artificial Intelligence. (1999) 391–398
14. Agichtein, E., Brill, E., Dumais, S.: Improving web search ranking by incorporating user behavior information. In: 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006). (2006) 19–26
15. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M.M., Gatford, M.: Okapi at TREC-3. In Harman, D.K., ed.: Overview of the Third Text Retrieval Conference (TREC-3), Gaithersburg, MD, USA, National Institute of Standards and Technology (NIST) (1995) 109–126
16. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems **22**(2) (2004) 179–214
17. He, B., Ounis, I.: A study of parameter tuning for term frequency normalization. In: Proceedings of the 2003 ACM CIKM International Conference on Information and Knowledge Management, New Orleans, Louisiana, USA, November 2-8, 2003, ACM (2003) 10–16
18. He, B., Ounis, I.: Query performance prediction. Information Systems **31**(7) (2006) 585–594
19. Kwok, K.L.: A new method of weighting query terms for ad-hoc retrieval. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM (1996) 187–195
20. Jansen, B.J., Spink, A.: An analysis of web searching by european alltheweb.com users. Information Processing & Management **41**(2) (2005) 361–381