Navigating the User Query Space

Ronan Cummins[1], Mounia Lalmas[2], Colm O'Riordan[3] and Joemon M. Jose[1]

[1]School of Computing Science, University of Glasgow, UK
[2]Yahoo! Research, Barcelona, Spain
[3]Dept. of Information Technology, National University of Ireland, Galway, Ireland
ronan.cummins@nuigalway.ie

Abstract. Query performance prediction (QPP) aims to automatically estimate the performance of a query. Recently there have been many attempts to use these predictors to estimate whether a perturbed version of a query will outperform the original version. In essence, these approaches attempt to navigate the space of queries in a guided manner.
In this paper, we perform an analysis of the *query space* over a substantial number of queries and show that (1) users tend to be able to extract queries that perform in the top 5% of all possible user queries for a specific topic, (2) that *post-retrieval* predictors outperform *pre-retrieval* predictors at the high end of the query space. And, finally (3), we show that some post retrieval predictors are better able to select high performing queries from a group of user queries for the same topic.

1 Introduction

Query performance prediction (QPP) (or estimating query difficulty) has become a vibrant research area in the last decade. Predicting the performance of a query is a useful task for many reasons. For example, search engines may wish to augment queries in different ways depending on their estimated performance. In fact, if query performance prediction becomes good enough [6], the space of all possible queries for a given topic may be able to be navigated efficiently, so that an initial query can be perturbed effectively. Furthermore, such techniques might be effective for creating good queries when a large number of terms are available. Query performance predictors can be used in conjunction with information extraction techniques to be able to extract good queries from these longer information needs. These approaches may ultimately help in shifting the cognitive load of query creation from the user to the system.

In this paper, we analyse the space of possible user queries (under some assumptions) over a range of topics and collections. In particular, we show that (1) while there are a number of queries which are extremely effective, humans create queries which peform within the top 5% of all possible user queries that can be extracted from a given *information need* (IN) under certain assumptions. Furthermore, (2) we show that post retrieval predictors are more effective than pre-retrieval predictors for predicting the performance of user queries (i.e. high

performing queries for a topic). Finally, (3) we demonstrate that some post-retrieval predictors are very successful at selecting high performing queries from a set of user queries (for the same topic).

The remainder of the paper is organised as follows: Section 2 presents background and related research that is relevant to this work. Section 3 comprises three parts. In section 3.1, we perform an analysis of the query space for a number of topics and collections. In section 3.2, we conduct a study which outlines the correlation of numerous pre- and post-retrieval predictors on sets of user queries for the same topic. Section 3.3 demonstrates a practical application of using predictors to select good user queries. Finally, section 4 outlines our conclusions.

2 Background and Related Research

Fundamentally, retrieval predictors can be divided into two classes: *pre-retrieval* [7, 6, 12] and *post-retrieval* [3, 10, 11] predictors. Pre-retrieval predictors use features from the query, document and collection before a query has been processed in order to ascertain its performance. Conversely, post-retrieval predictors analyse the result list, scores and complex features to create predictors that have a higher overhead in terms of computation [2]. One of the earliest approaches to QPP has been that of the clarity score [3], which measures the KL-divergence between the query and collection model in a language modelling framework. Recent research has shown that the standard deviation (σ) of scores in a ranked list is a good predictor of query performance [10, 11] for the traditional QPP task. It has also been shown [10] that even better prediction can be obtained if a variable cut-off point is used (i.e. different cut-off points for different queries). A relatively new predictor has also been introduced where the standard deviation of the first N documents is calculated, where N is the number of documents in the head of the list that are within a certain a percentage (i.e. 50%) of the top score [5].

Recently work has been conducted into combining retrieval predictors with the aim of improving performance by reducing queries that may contain noisy terms (e.g. noisy terms in the description field of topics) [9, 1]. Some work similar to the research outlined herein has been conducted [8]. However, we place the problem of selecting user queries in a query prediction framework, and review a substantial number of high performing pre-retrieval and post-retrieval methods. We also conduct an analysis of how effective users are at the task of query extraction.

3 Experimental Analysis

In this section, we conduct and analysis of the query space. Firstly, we show that the ranked performance of queries follows a power law distribution, and that user create queries that lie within the fifth percentile of such a distribution. Then, we perform an analysis of a number of pre-retrieval and post-retrieval predictors

and show that post-retrieval predictors can more easily predict high performing queries.

3.1 Sampling the User Query Space

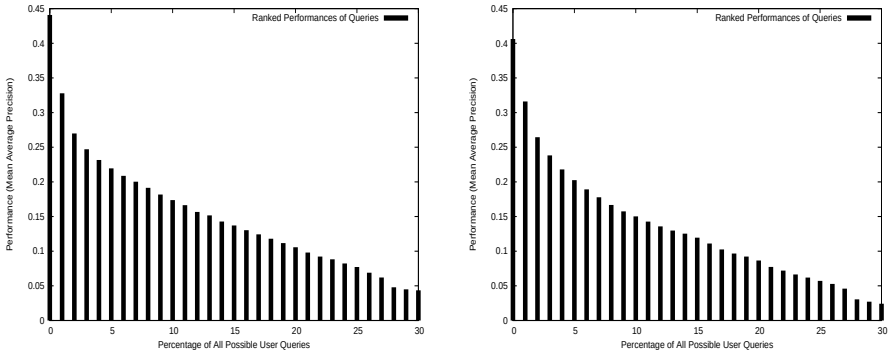
First we outline two assumptions that constrain this work; (1) We assume user queries consist of queries of not longer than six terms (research has indicated that the vast majority of user queries are indeed less that this). And (2) we assume that user queries comprise terms that appear somewhere in the TREC topic statement (i.e. *title*, *desc*, and *narrative*), as these topic statements model actual *information needs*).

Now, to analyse the user query space in a thorough manner, we wish to sample a large number of the high performing queries that a user might possibly generate (when prompted with an *information need*). Given that there are 2^N possible queries for an *information need* of N terms, we cannot exhaustively evaluate and analyse all possible queries. Therefore, we create a sample of queries for a topic in the following manner; (1) We extract the top 20 (i.e. $N = 20$) most discriminative terms (*idf*) from the topic (this is all the terms for some topics) to be used in our sample user queries. (2) We submit *all* queries of length one and two terms, and record their performance (average precision). Then (3), for all other queries from length three to six terms (in that order), if a query has an estimated[1] performance within 66% of the best query found thus far for that information need, we submit it to the system and record its performance. Therefore, we are quite confident that by the end of the process we have a large selection of queries within the high end of the query space. Figure 1 shows that the distribution of queries when ranked by performance (mean average precision) follows a power law (i.e. there are few high performing queries and many poorly performing ones).

We asked human annotators to extract keyword type queries from the *desc* and *narr* fields in a topic (similarly to previous research [4]). This resulted in four sets of short keyword type queries for each topic. Table 1 shows the percentage of queries found in our sampling approach that outperform the actual user queries [2]. The analysis shows that users perform around the fifth percentile of all possible queries for the query extraction task. Another important point to note is that users do not simply extract the same *good* queries. We analysed all possible pairs of user queries (within each topic) and found that 89% of all possible query pairs are unique. This indicates that user queries for the same topic (even when prompted with the actual *desc* and *narr*) are quite varied.

[1] When estimating a query of length Q, we find the performance of one of its sub-queries of length $Q - 1$ and aggregate this with the performance of the single query term remaining. This is very generous estimation of a query and is likely to over-estimate the performance of a query.

[2] We created another sample of the query space by exhaustively evaluating all queries of length one, two and three, and obtained nearly identical statistics.



(a) FT collection (b) WSJ collection

Fig. 1. Performance of All Queries

Table 1. Percentile Report (and Standard Deviation) for User Queries

Collection	Topic Range	# Topics	User1	User2	User3	User4
AP	051-200	149	4.0 (4.0)	3.4 (3.5)	3.0 (3.1)	3.3 (3.5)
FBIS	301-450	116	4.8 (7.8)	4.8 (7.8)	5.3 (7.6)	4.8 (7.6)
FT	250-450	188	5.2 (8.6)	5.3 (8.5)	5.7 (8.4)	5.7 (9.3)
WSJ	051-200	150	5.7 (7.8)	4.6 (6.6)	5.0 (7.6)	4.1 (6.0)

3.2 Correlation of User Extracted Queries

In this section, we report the performance of a number of representative pre-retrieval and post-retrieval performance predictors from the literature[3] on the task of query performance prediction within each topic. We increase the number of queries per topics to five by including the original *desc* field. Although, we have only five queries for each topic, we have a large amount of topics across which to average the correlation coefficients. Furthermore, we know that the five queries for each topic are high performing queries, and we can confirm that for over 75% of the topics the full five queries are unique.

The best pre-retrieval predictors from the literature are the simplified clarity score (scs), the average idf of query terms (idf_{avg}), the maximum idf of the query terms (idf_{max}), the scq score, the normalised scq ($nscq$) score, and the maximum contributing term to the scq score (scq_{max}) [12]. Some of the highest performing post-retrieval predictors from the literature are query clarity ($clarity$), standard deviation at 100 documents (σ_{100}), a normalised version of standard deviation at 100 documents (i.e. the ncq predictor [11]), the maximum standard deviation in the ranked-list (σ_{max}) [10]. We also use two new predictors that calculate

[3] While we have not included, nor conducted experiments on, an exhaustive list of pre-retrieval and post-retrieval predictors, we have included the highest performing predictors from the literature.

the standard deviation using a variable cut-off point ($\sigma_{50\%}$), and a query length normalised version of that predictor ($n(\sigma_{50\%})$) [5].

Table 2 shows the average correlation between the output of each predictor and the performance of the user queries (i.e. those at the high performing end of the query space). Firstly, we can see that the post-retrieval predictors (bottom half of the table) outperform the pre-retrieval predictors (top half of the table) for this part of the query space. For example, idf_{max}, a high performing predictor in other studies [6], performs poorly at the high end of the query space. This is because users will often choose the same highly discriminating term when creating a query for the same topic. Therefore, it should be noted that many proposed predictors (especially *pre-retrieval* predictors), may not be able to distinguish between high performing queries. The highest correlated pre- and post-retrieval predictors are outlined in bold.

Table 2. Correlations (ρ and r) for User Queries Averaged Over All Topics

Coll.	AP		FBIS		FT		WSJ	
Predictor	r	ρ	r	ρ	r	ρ	r	ρ
scs	0.073	0.086	0.069	0.108	0.133	0.142	0.000	0.035
idf_{avg}	0.063	0.067	0.086	0.109	0.163	0.153	0.035	0.051
idf_{max}	-0.022	-0.040	0.015	0.123	0.107	0.131	0.045	-0.042
scq	-0.033	0.017	-0.06	-0.018	-0.013	0.005	-0.015	0.001
scq_{max}	0.043	0.024	0.024	0.155	0.107	0.131	0.101	0.008
$nscq$	**0.122**	**0.093**	**0.128**	**0.193**	**0.139**	**0.159**	**0.073**	**0.106**
$clarity$	0.118	0.135	0.216	0.259	0.208	0.217	0.103	0.138
σ_{100}	0.185	0.157	0.196	0.237	0.265	0.211	0.263	0.253
σ_{max}	0.134	0.162	0.178	0.227	0.227	0.225	0.173	0.208
ncq	0.115	0.136	0.235	0.283	0.247	0.232	0.267	0.246
$\sigma_{50\%}$	0.250	0.238	0.188	0.211	0.215	0.214	0.271	0.260
$n(\sigma_{50\%})$	**0.368**	**0.328**	**0.280**	**0.340**	**0.255**	**0.269**	**0.405**	**0.398**

3.3 Usability of Predictors for Query Selection

We now conduct an experiment to investigate the usefulness of the query performance predictors at selecting the best query among a group of high performing queries[4]. Within each topic, we use each predictor in turn to select the best query (as predicted by the predictor) and then measure the MAP of the set of queries chosen (i.e. the predictor selects one of five queries for each topic). Table 3 shows the performance (MAP) of each predictor for such a task. We deem a predictor to be useful when it consistently[5] outperforms the performance of the best single user. The best predictors tend to be the ones based on standard deviations (i.e. ncq, $\sigma_{50\%}$, and $n(\sigma_{50\%})$). Many of these predictors significantly outperform the average query for a topic. Overall, the best predictor for selecting good user queries are the $n(\sigma_{50\%})$ predictor [5]. The predictor can outperform the best single performing set of queries.

[4] Such a scenario may have applications in an collaborative search scenario.

[5] † and ‡ denotes a significant increase over the average and best set of queries respectively, using a Wilcoxon test at the 0.05 level on the topics.

Table 3. MAP for Each Set of Topics Using Predictors to Select Queries

Collection	AP	FBIS	FT	WSJ
Avg. Qry per Topic	0.1698	0.2183	0.2294	0.2394
Best **Set** of User Qrys	0.1846	0.2325	0.2482	0.2669
idf_{avg}	0.1759	0.2353	0.2485	0.2371
$nscq$	0.1794	0.2338	0.2336	0.2427
$clarity$	0.1785	0.2311	0.2506†	0.2383
σ_{100}	0.1846 †	0.2463 †	0.2388†	0.2682†
ncq	0.1808	0.2483 †	0.2632†	0.2613†
$\sigma_{50\%}$	0.1881 †	0.2403 †	0.2511†	0.2679 ‡
$n(\sigma_{50\%})$	**0.1940**‡	**0.2523** †	**0.2623**†	**0.2859** ‡

4 Conclusion

In this paper, we have shown that user queries lie in the top 5% of queries that a user could extract from an information need. We have shown that post retrieval predictors outperform pre-retrieval for actual user queries. Furthermore, we have shown that post retrieval predictors can be used to effectively choose between high performing queries. This has applications to systems that aim to automatically choose between queries of the same topic (e.g. collaborative IR systems).

References

1. Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R. Carvalho. Exploring reductions for long web queries. In *SIGIR*, pages 571–578, 2010.
2. David Carmel and Elad Yom-Tov. *Estimating the Query Difficulty for Information Retrieval*. Morgan and Claypool Publishers, 1st edition, 2010.
3. Stephen Cronen-Townsend, Yun Zhou, and W. Bruce Croft. Predicting query performance. In *SIGIR*, pages 299–306, 2002.
4. Ronan Cummins, Mounia Lalmas, and Joemon Jose. The limits of retrieval effectiveness. In *ECIR 2011*. ACM, 2011.
5. Ronan Cummins, Colm O'Riordan, and Joemon Jose. Improved query performance prediction using standard deviation. In *SIGIR 2011*. ACM, 2011.
6. Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. A survey of pre-retrieval query performance predictors. In *CIKM 2008*, pages 1419–1420, New York, NY, USA, 2008. ACM.
7. Ben He and Iadh Ounis. Inferring query performance using pre-retrieval predictors. In *SPIRE*, pages 43–54, 2004.
8. Giridhar Kumaran and James Allan. Selective user interaction. In *CIKM 2007*, pages 923–926, New York, NY, USA, 2007. ACM.
9. Giridhar Kumaran and Vitor R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR*, pages 564–571, 2009.
10. Joaquín Pérez-Iglesias and Lourdes Araujo. Standard deviation as a query hardness estimator. In *SPIRE*, pages 207–212, 2010.
11. Anna Shtok, Oren Kurland, and David Carmel. Predicting query performance by query-drift estimation. In *ICTIR*, pages 305–312, 2009.
12. Ying Zhao, Falk Scholer, and Yohannes Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR*, pages 52–64, 2008.