Predicting Query Performance In Microblog Retrieval

Jesus A. Rodriguez Perez, Joemon M. Jose
School of Computing Science
University of Glasgow
Glasgow
(Jesus.RodriguezPerez; Joemon.Jose)@glasgow.ac.uk

ABSTRACT

Query Performance Prediction (QPP) is the estimation of the retrieval success for a query, without explicit knowledge about relevant documents. QPP is especially interesting in the context of Automatic Query Expansion (AQE) based on Pseudo Relevance Feedback (PRF). PRF-based AQE is known to produce unreliable results when the initial set of retrieved documents is poor. Theoretically, a good predictor would allow to selectively apply PRF-based AQE when performance of the initial result set is good enough, thus enhancing the overall robustness of the system. QPP would be of great benefit in the context of microblog retrieval, as AQE was the most widely deployed technique for enhancing retrieval performance at TREC. In this work we study the performance of the state of the art predictors under microblog retrieval conditions as well as introducing our own predictors. Our results show how our proposed predictors outperform the baselines significantly.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Ad-hoc Retrieval; Query Performance Prediction; Query Expansion

1. INTRODUCTION

Most information retrieval systems experience a high variability in retrieval performance across different queries. Whilst many queries are satisfied successfully, the system produces poor results for many others. Since a number of retrieval approaches rely on the initial set of results, it would be highly desirable to predict when retrieval results are not satisfactory enough.

This task is known as query performance prediction (QPP), and has been an active and challenging area of research over the last decade. Multiple predictors have been proposed in the literature with varying degrees of success. These predictors fall mainly into two categories: pre-retrieval and post-retrieval predictors. Pre-retrieval predictors are computed before retrieving any documents, thus relying solely on query term features. On the other hand, post-retrieval predictors rely on features extracted from the retrieved documents. Post-retrieval predictors mainly estimate how well a query is represented by the retrieved documents.

In this work, we study pre and post retrieval predictors for microblog retrieval tasks. Although much work has been done in predicting the performance of queries over web collections, to the best of our knowledge, no work has been done in the context of microblogs. Microblogging platforms such as Twitter have gained momentum over recent years providing a new way of sharing information and broadcasting short messages over a network of users. Microblogs present many differences with respect to web documents both in morphology and content [9]. Mainly, microblogs constitute a time ordered stream of very short documents as they are published. Moreover, microblogs contain community defined tags to refer to certain topics (hashtags), or people (mentions), which we intent to investigate in our QPP study.

The motivation behind studying QPP for microblogs resides in increasing the robustness of existing retrieval approaches. More specifically, QPP can be especially handy for selectively applying pseudo relevance feedback (PRF) based automatic query expansion (AQE) approaches [2]. PRF-based AQE approaches rely on the initially retrieved set of documents. Thus if these documents loosely represent the initial information need, PRF-based approaches most likely result in unexpected behaviour, and unreliable results.

Effective QPP represents an opportunity to estimate the performance of a system for a particular query, based on pre-retrieval and post-retrieval features. In turn, this would allow an IR system to selectively perform AQE, when the circumstances are most propitious, based on estimates given by the predictors.

Our work is driven by two research questions. **(RQ1)** To what extent, can we predict the performance of a retrieval model for microblog corpora?. **(RQ2)** To what extent, the combination of predictors can improve overall prediction performance, in the context of microblogs?.

In this work, we investigate the performance of previously proposed predictors [5], in the context of microblogs. We then show that they fail to perform effectively which prompts the need to develop better predictors. We propose a number of predictors, which take into consideration the

SIGIR'14, July 6–11, 2014, Gold Coast, QLD, Australia.

http://dx.doi.org/10.1145/2600428.2609540.

characteristics of microblogs. Our evaluation findings show how our predictors outperform those found in the literature. Finally we further improve our performance by learning a prediction model, combining our predictors by means of a support vector machine for regression.

2. RELATED BACKGROUND

One of the main works in query performance prediction is that by [3]. In their work, they proposed a predictor based on the Kullback-Leiber divergence between the query's and the collection's language models. This predictor attempts to quantify the "clarity" of the query. In other words, the non-ambiguity of the query which in turn should reflect on how well it represents a particular topic. Their evaluation shows good correlation of their predictor with average precision, using Spearman's ranking correlation tests.

Work by [7] suggests other predictors such as the standard deviation of IDF values within the query. They also defined a simplified version of the "Clarity Score" proposed by [3] namely Simplified Clarity Score (SCS). Finally, they proposed an alternative to SCS called query scope (QS). Their main objective was to investigate pre-retrieval predictors, as post-retrieval predictors are normally computationally more expensive to use.

To predict query difficulty [8] proposed a query coherence score (QC-1) which attempts to quantify how related are the query terms to the retrieved set of documents as well as measuring the differences between the language used in the retrieved set and the query, with respect to the collection. They found that their approach correlates well with average precision, using Spearmans's rank correlation test. Furthermore they also suggested two other versions of this score however their performance was poorer than their simpler first version. For their evaluation they used a number of retrieval models, including BM25 and TFIDF to retrieve documents from the TREC Robust track collection.

Work by [10] proposed a series of pre-retrieval performance predictors. One of the most succesful was SCQ. The aim of SCQ is to compute a similarity score between the queries and the collection. Moreover, they also proposed a variability measure relying on the standard deviations of TFIDF scores for the query terms. Furthermore, they also proposed a predictor using both previous approaches together. Their evaluation showed how the combined predictor outperformed all previous approaches. It is important to note that their joint approach is slightly better than their simple SCQ, only when the linear interpolation gives most of the weight to SCQ, albeit being much more complex, thus computationally much more expensive. In this work, we will evaluate the performance of SCQ in our particular context.

A short but comprehensive survey of performance predictors was produced by [6]. This study helped on deciding which predictors to include in our study in the context of microblogs, as it showed results for many state of the art predictors across multiple collections.

Previous evaluations. The "de facto" evaluation procedure in previous work has been the statistical correlations between the predictors and the evaluation metric results for a given system. More often than not, the evaluation metric used was Average Precision (AP). The better the predictor estimates the performance of the system in terms of an evaluation metric, the higher the correlation scores.

The most used correlation metrics are Kendall-Tau (K.Tau) and Spearman's (SP.Rho) rank correlation coefficients. The SP.Rho correlation coefficient is a measure of statistical dependence between two variables, by which it is estimated how well their relation is represented by a monotonic function (I.e.: grows/decreases always in the same direction). SP.Rho uses Pearson's correlation coefficient in such a way that is much less sensitive to outliers. Kendall-Tau's correlation coefficient is slightly different in that, it does not rely on the values of the variables themselves, but it rather measures the similarity in the ordering of the data provided when ranked by each of the variables.

State of the art prediction. The correlation coefficients obtained for AP in web collections vary wildly. The Kendall-tau coefficients, with respect to AP, for the best performing pre-retrieval predictors range from 0.30 to 0.49 depending on the collection [1]. On the other hand, the Kendall-tau coefficients for post-retrieval predictors are generally higher.

It is important to note the high variability in terms or predicting performance, with respect to the collection. The collections used in the literature include "TREC Vol. 4+5"; "WT10g" and "GOV2", where it is often the case for a particular predictor to be the best for a particular collection and the worst for another.

Selective Query Expansion. One of the main applications of QPP is selective Query Expansion [1]. It refers to selectively applying automatic query expansion (AQE) whenever predicted performance is above a certain threshold. This serves as a warranty for PRF-based AQE approaches, as they rely on the top N retrieved documents to perform optimally.

3. PREDICTORS

In this section, we describe the state of the art predictors we will be considering in our evaluation, including our proposed predictors. Subsequently, we introduce the evaluation approach followed to benchmark and compare their performances.

Baseline Predictors: As a starting point we selected those predictors performing best from the survey by [6].

The **QueryTermIdf** predictor utilizes the IDF values of query terms for making estimations of retrieval performance. The intuition is that the higher the IDF value the more specific a term is, thus score variations across terms may indicate drifting concepts, negatively affecting the performance. We derive different predictors considering the mean, median, standard deviation (Std), max, min and diff($max - min$) IDF scores from each query [6]. Moreover, **Simplified Clarity Score (SCS)** proposed by [7], attempts to model the clarity of a query, i.e. how well it targets a particular topic based on the collections metrics. An homologous predictor to SCS is **Query Scope (QS)**, which was also proposed by [7].

Another predictor is **Similarity of Collection w/ Query (SCQ)** which was proposed by [10] to compute the similarity between the collection and the query at hand. In a similar context, **Term Weight Variability (VAR)** was proposed. This predictor measures the variability of weights for a term across the collection. [10] hypothesises that the higher standard deviation for a term, the more discriminative it is.

Moreover, the work by [1] introduced four post_retrieval predictors namely **NQC**, **WIG**, **QF** and **Clarity**. NQC measures the normalized standard deviation of the top scores. The intuition behind this predictor is that relevant documents are assumed to have a much higher score than that of the mean score. Similarly WIG measures the divergence of retrieval of the top-ranked results scores from that of the documents in the corpus. QF and Clarity are predictors that take into account the actual content of the documents. QF measures the divergence between the original top results for the query and the results that would be obtained for a query constructed from the top results. Finally, Clarity measures the KL divergence between a (language) model induced from the result list and the corpus model.

Proposed Predictors: The first two predictors we defined are to measure the coverage of terms over the documents retrieved. The **CoveredQueryTerms (QTCov)** predictor measures how well the query is being represented by the documents in the result list. For each document, we divide the number of query terms found in the document by the total number of query terms, which gives a normalized value between 1 and 0. (1 being a document that completely fits the query). Similarly we defined **TopTermsCoverage (TTCov)** which measures the coverage of the top N terms in the result list. Intuitively, the more times these terms appear the more likely documents are to revolve around a particular topic.

Another predictor is **TimeCohesion (TimeCH)**, which taps into the distribution of retrieved tweets over time. We assume that the closer the documents appear with respect to time, the more likely they refer to the same event. To compute it, we take the differences between retrieved document timestamps. Differences are taken only between contiguous documents in the rank.

Also, exploiting microblog specific features we defined the **Http** predictor. This predictor measures how common is to find a Url in the retrieved set of documents. To this end we count the number of documents with Urls and divide by the total number of documents retrieved. (I.e the rate of Urls). This predictor relies in previous findings which suggest that the presence of Urls indicates the presence of relevant documents [4].

Finally, we defined **HashTagCount** as the rate of documents with hashtags in the retrieved results set. Hashtags are important in the context of Twitter as they refer to particular topics. Thus the presence of similar hashtags often indicates that users are dealing with the same topic.

3.1 Evaluation

Datasets. In this evaluation we utilize the Tweet2011, 2012 and 2013 collections with a total of contains 170 topics. The collections have been merged together to produce enough evidence for learned predictor models.

Retrieval Model Used. We utilized the DFRee for producing the runs, as it provides competitive performance. (P@10 = 0.527 & P@30= 0.409).

Predictor's Correlation. In the literature, evaluations have mainly taken into account either K.Tau or SP.Rho as correlation measures with respect to average precision (AP).

AP correlations			
Predictor	**K.Tau**	**SP.Rho**	**Pearson**
post_TTCov_Mean	0.302 **	0.447 **	0.403
post_TTCov_Median	0.253 **	0.312 **	0.274
post_TTCov_upper	**0.356 ** **	**0.463 ** **	**0.434**
post_TTCov_Lower	0.178	0.218 **	0.197
post_TimeCH_Lower	-0.202 **	-0.300 **	-0.273
post_TimeCH_Median	-0.200 **	-0.291 **	-0.310
post_TimeCH_Upper	-0.122 *	-0.188 *	-0.231
post_TimeCH_Mean	-0.197 **	-0.288 **	-0.281
pre_SCQ_Sum	0.094	0.138	0.254
pre_QueryTermIdf_Diff	0.140 **	0.209 **	0.200

Table 1: Correlations of DFRee runs with AP (**$p < 0.01$ & *$p < 0.05$)

In this work, we also pay attention to Pearson's correlation coefficient.

In microblog retrieval, it is most important to optimise performance for the first retrieved documents due to its real-time nature. It has been agreed in the literature that a user will not look further than the first 30 documents, thus AP might not be appropriate for this task. Moreover, to help in selectively applying PRF-based AQE, we focus on the very top retrieved documents, thus we also study prediction in terms of P@10.

4. RESULTS AND DISCUSSION

In this section we introduce and discuss the results we obtained during the evaluation of the above mentioned predictors. Tables 1 and 2 show the correlation coefficients in terms of K.Tau, SP.Rho and Pearson for a subset of predictors. Since it was not possible to show all the predictors in this paper, we have chosen to include only those achieving a Pearson coefficient higher than 0.19. The predictors are prefixed with either "pre_" or "post_" to indicate whether they are pre-retrieval or post-retrieval predictors. Furthermore, the suffixes: Mean, Median, Std, Max, Min, Lower and Upper; denote mean, median, Standard Deviation, maximum, minimum, lower percentile and upper percentile, of the predictor values respectively. Moreover, Sum refers to the Sum of all predictor values, whereas Diff is the difference between Max and Min.

Table 1 shows the correlations coefficients in terms of AP. In the survey done by [6] the maximum correlation achieved using K.Tau ranged from 0.30 to 0.49 depending on the collection.

State of the art predictors such as VAR, SCS, NQC or WIG (Described in Section 3) performed poorly in the context of microblogs, as their K.Tau coefficient values ranged between 0 and 0.16, thus are not shown in Tables 1 or 2. This demonstrates how challenging performance prediction is in the context of microblog retrieval, and the need for tailored predictors to this new task. On the other hand, the predictors we proposed achieved a much better correlation than that obtained by the state of the art in this context. **post_TTCov_upper** is one of such predictors, achieving a K.Tau coefficient of 0.356, being the best correlation with respect to AP. This predictor takes the upper percentile of the rate at which top terms appear in the retrieved set of documents.

These results are very promising, but our main focus is to enable selective PRF-based AQE, thus we present correlations in terms of P@10 in Table 2. As it can be observed, amongst the top performing predictors we find those rely-

P@10 correlations			
Predictor	**K.Tau**	**SP.Rho**	**Pearson**
post_http	0.163 **	0.206 **	0.213
post_QTCov_mean	0.291 **	0.382 **	0.375
post_QTCov_median	0.305 **	0.382 **	0.373
post_QTCov_upper	0.325 **	0.404 **	0.392
post_QTCov_lower	0.266 **	0.336 **	0.312
post_TTCov_mean	0.301 **	0.416 **	0.429
post_TTCov_median	**0.365 ** **	**0.456 ** **	**0.441**
post_TTCov_upper	0.264 **	0.355 **	0.374
post_TTCov_lower	0.253 *	0.303 **	0.298
post_TimeCH_lower	-0.212 **	-0.286 **	-0.236
post_TimeCH_median	-0.145 **	-0.199 *	-0.239
post_TimeCH_mean	-0.170 **	-0.233 **	-0.212
post_TimeCH_diff	0.192 **	0.269 **	0.198

Table 2: Correlations of DFRee runs with P@10 (**$p < 0.01$ & *$p < 0.05$)

ing on microblog specific features, namely **post_TimeCH** which measures how close in time are the retrieved tweets and **post_http** measuring the presence of URL's in documents. Additionally, the correlations achieved by these predictors with respect to P@10, are generally higher than that achieved for MAP, with **post_TTCov_Median** being the best performing predictor.

An interesting observation regarding **post_TTCov_upper** and **post_TTCov_Median** is that they may be referring to the same documents, as with AP a larger set of documents is considered compared to P@10.

Combining Predictors. Since our main objective is predicting performance with AQE in mind, we focus on P@10. To combine the predictors for P@10, we used a Support Vector Machine for regression. To avoid biasing we performed a ten-fold cross-validation. The learned prediction model is defined as follows:

```
P@10 =  0.3028 * TTCov_upper
 + 0.3494 * QTCov_median + 0.3701 * QTCov_upper
 - 0.4745 * twids_median - 0.2641 * TTCov_mean
 + 0.5014 * twids_mean + 0.3394 * TTCov_median
 + 0.2318 * TTCov_lower + 0.3122 * twids_diff
 + 0.2429 * http - 0.1651 * QTCov_lower
 - 0.2745
```

The correlation coefficients obtained for this model, are 0.412 (+12.88%), 0.559(+22.59%), and 0.539 (+22.22%), for K.Tau, SP.Rho and Pearson respectively.

Finally, the predictors proposed in this work outperform those in the literature, within this particular context. However, as in previous attempts, it is uncertain that it will be enough for enabling effective selective AQE. Nonetheless, these predictors may represent an important step towards that much sought after objective, within the context of microblogs.

5. CONCLUSIONS

In this work, we studied the performance of the state of the art predictors in the context of microblogs. The most sought after benefit from predicting query performance is increasing the robustness of PRF-based AQE approaches. To this end we paid special attention to the prediction in terms of the top retrieved documents, specifically Precision@10 (P@10).

Our evaluation suggests that predictors in the literature perform poorly in the context of microblogs, thus we need to come up with predictors that are better fit for purpose. To this end, we defined a number of predictors relying on microblog features and characteristics. We benchmarked their performance and showed that most of them outperform those in the literature, with TTCov being the most correlated with MAP and P@10.

Finally, we used support vector machines for regression to learn a prediction model based on the best performing predictors. The resulting model further increased performance by a +22% in terms of the Pearson correlation coefficient, and +12.88% for K.Tau.

Future work will put these findings to a practical application for selective approaches to PRF-AQE, or in the selection of a baseline model to optimize a system's overall performance given the conditions of a particular query. Furthermore, we will study the performance of other predictors which will consider more microblog specific features.

6. ACKNOWLEDGMENT

This research is partially supported by the EU funded project LiMoSINe (288024).

7. REFERENCES

[1] D. Carmel and E. Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

[2] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.

[3] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM, 2002.

[4] D. F. Gurini and F. Gasparetti. Trec microblog 2012 track: Real-time algorithm for microblog ranking systems. 2012.

[5] C. Hauff. Predicting the effectiveness of queries and retrieval systems. University of Twente, 2010.

[6] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1419–1420. ACM, 2008.

[7] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval*, pages 43–54. Springer, 2004.

[8] J. He, M. Larson, and M. De Rijke. Using coherence-based measures to predict query difficulty. In *Advances in Information Retrieval*, pages 689–694. Springer, 2008.

[9] J. Teevan, D. Ramage, and M. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

[10] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval*, pages 52–64. Springer, 2008.