Query Performance Prediction for Microblog Search: A Preliminary Study

Maram Hasanain
maram.hasanain@qu.edu.qa

Rana Malhas
rana.malhas@qu.edu.qa

Tamer Elsayed
telsayed@qu.edu.qa

Department of Computer Science and Engineering
Qatar University
Doha, Qatar

ABSTRACT

Microblogging has recently become an integral part of the daily life of millions of people around the world. With a continuous flood of posts, microblogging services (e.g., Twitter) have to *effectively* handle millions of user queries that aim to search and follow recent developments of news or events. While predicting the quality of retrieved documents against search queries was extensively studied in domains such as the Web and news, the different nature of data and search task in microblogs triggers the need for re-visiting the problem in that context. In this work, we re-examined several state-of-the-art query performance predictors in the domain of microblog ad-hoc search using the two most-commonly used tweets collections with three different retrieval models that are used in microblog search.

Our experiments showed that a temporal predictor was generally the best to fit the prediction task in the context of microblog search, indicating the importance of the temporal aspect in this task. The results also highlighted the need to either re-design some of the existing predictors or propose new ones to function effectively with different retrieval models that are used in our tested domain. Finally, our experiments on combining multiple predictors resulted in achieving considerable improvements in prediction quality over individual predictors, which confirmed the results reported in the literature but in different domains.

Categories and Subject Descriptors

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

Keywords

query difficulty; microblog search; temporal retrieval

SoMeRA'14, July 11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-3022-0/14/07 ...$15.00.
http://dx.doi.org/10.1145/2632188.2632210.

1. INTRODUCTION

Microblogging online services have been widely spreading during the past few years. Twitter is certainly one of the most rapidly growing microblogging services. Millions of Twitter users share information, news, and opinions about ongoing events and activities through 140-character long posts called tweets. Given this continuous stream of millions of tweets posted daily[1], users in parallel are issuing lots of search queries to get the latest information about events and news. While some of these queries will get high quality results, others might be difficult to answer by the search system. However, the system can attempt to improve retrieval performance for poorly-performing queries if it can accurately-enough estimate the performance of queries in advance. Query Performance Prediction (QPP) aims to estimate the quality of retrieval for a query given a retrieval model and a collection of documents in absence of relevance information [19, 1]. There is a large body of research work on studying methods of predicting the performance of a query, either by only examining the query terms (i.e., *pre-retrieval* predictors) [6, 18, 5], or by analyzing the retrieved documents as well (i.e., *post-retrieval* predictors) [2, 7, 20, 3, 15]. Most of these studies were conducted on ad-hoc search in the news and Web domains [2, 7, 20, 15].

Microblogs are naturally different from news and Web documents; they are very short, very informal, and of conversational nature, compared to the long, well-formed, non-conversational documents in typical TREC Web and news collections used in previous studies. The temporal aspect is also highly-manifested in microblogs, due to their posting frequency. Moreover, Twitter users for example, tend to be interested in retrieving relevant and *fresh* tweets [9], indicating that *recency* is an important factor in microblog real-time ad-hoc search. These distinguishing features of both the data and the search task in microblogs make it necessary to revisit the problem of query performance prediction in such domain; up to our knowledge, there was no previous studies conducted on this problem in the domain of microblogs.

In this work, we target three main research questions:

1. How well do the existing state-of-the-art predictors perform in the context of microblog search?
2. Will their performance be consistent across different retrieval models, more specifically the temporal ones, that are used in microblog search?

[1] https://blog.twitter.com/2013/celebrating-twitter7

3. Can we improve the prediction quality in the studied domain using a combination of predictors?

Accordingly, our contributions in this work are two-fold:

1. This is the first study of query performance prediction in the context of microblog search. We examined several pre- and post-retrieval predictors using the two most-widely used microblog collections (Tweets2011 [13] and Tweets2013 [11]) with three different retrieval models that are specifically used in such context. Our experiments gave insights on their performance in that domain when used individually and when combined using linear regression.

2. We highlight the need for query performance predictors that consider the temporal nature of the microblog data and the corresponding search task.

The remainder of the paper is organized as follows. We first describe the query performance predictors we used in our study in Section 2. Experimental setup is then presented and results are discussed in Section 3, followed by the conclusion and some guidelines for future work in Section 4.

2. PERFORMANCE PREDICTORS

To study query performance prediction in the context of microblog search, we experimented with several existing pre- and post-retrieval predictors. In this section, we describe those predictors we used in our study. We note that the studied predictors aim to predict the Average Precision (AP) for a given query [1].

2.1 Pre-retrieval Predictors

For pre-retrieval prediction, we basically experimented with two main categories of predictors. The first category is based on the inverse document frequency (IDF) of query terms. Under this category, we considered the maximum (MaxIDF), sum (SumIDF), average (AvgIDF), standard deviation (DevIDF), and variance (VarIDF) of IDF values of query terms. The other category is based on a score for collection query similarity (SCQ) [1] defined as follows:

$$SCQ(w) = (1 + \log(cf_{w,C})) * IDF(w) \quad (1)$$

where $cf_{w,C}$ is the collection frequency of query term w in the document collection C. Under that category, we considered the maximum (MaxSCQ), sum (SumSCQ), and average (AvgSCQ) of SCQ values of query terms. Additionally, we experimented with the simplified clarity score (SCS) [6] that estimates the divergence between the query language model based on the query terms and the collection language model.

2.2 Post-retrieval Predictors

Post-retrieval predictors require a list R of k retrieved documents in response to a given query Q, in order to predict the performance of Q [1]. We selected four predictors based on their reported high prediction quality when experimented with different types of collections [20, 3, 15]. We have also examined the performance of a fifth predictor [7] that emphasizes the *temporal* aspect of the data, which is important in the search task in the microblog domain. For each of the predictors presented next, k is a free parameter.

- **Clarity (CLR):** CLR is one of the very first proposed predictors [2]. The prediction is based on estimating the coherence of the list R with respect to the collection of documents C using the KL-divergence [17] between the query language model induced by R and the collection language model. The query language model is represented as follows:

$$P(w|Q) = \sum_{D \in R} P(w|D)P(D|Q) \quad (2)$$

where $P(w|D)$ is estimated using the maximum likelihood estimate (MLE) as follows: $P(w|D) = \frac{tf_{w,D}}{|D|}$, where $tf_{w,D}$ is the term frequency of w in D. $P(D|Q)$ is computed as the normalized (over all documents in R) query likelihood of D, assuming uniform prior probabilities for the documents as shown next:

$$P(D|Q) = \frac{\prod_{w \in Q} P(w|D)}{\sum_{D' \in R} \prod_{w \in Q} P(w|D')} \quad (3)$$

Finally, the clarity score is computed using KL-divergence as follows:

$$\text{CLR}(Q) = \sum_{w \in V} P(w|Q) \log \frac{P(w|Q)}{P(w|C)} \quad (4)$$

where V is the vocabulary set and $P(w|C)$ is estimated by MLE over C.

- **Normalized Query Commitment (NQC):** NQC [15] measures the amount of query drift in the results list R; that is, the commitment of documents in R to aspects related to Q. NQC is computed as follows:

$$\text{NQC}(Q) = \frac{\sigma_R}{|Score(C)|} \quad (5)$$

where σ_R is the standard deviation of retrieval scores of documents in R, and $Score(C)$ is the retrieval score of the collection when viewed as one very long document.

- **Normalized Standard Deviation (NSD):** With a similar intuition to NQC, NSD [3] is computed as the standard deviation of document retrieval scores, but normalized by the square root of the query length instead of the collection score. It also differs from NQC, when computing the standard deviation, in considering only top documents in R with retrieval scores $\geq x\%$ of the score of the top-ranked document. The predicted value is computed as follows:

$$\text{NSD}(Q) = \frac{\sigma_{x\%}}{\sqrt{|Q|}} \quad (6)$$

where $\sigma_{x\%}$ is the standard deviation of retrieval scores of documents matching the $x\%$ cut-off criterion. $x\%$ is a free parameter.

- **Weighted Information Gain (WIG):** WIG [20] measures the difference between the average retrieval score of documents in R and the collection retrieval score. In this study, we adopted a reduced version that is based on query likelihood model [19], and thus computed as follows:

$$\text{WIG}(Q) = \frac{1}{k} \frac{1}{\sqrt{|Q|}} \sum_{D \in R} (Score_{QL}(D) - Score_{QL}(C)) \quad (7)$$

where $Score_{QL}(D)$ is the query log-likelihood score of D calculated as follows:

$$Score_{QL}(D) = \sum_{w \in Q} \log P(w|D) \quad (8)$$

and $Score_{QL}(C)$ is a query log-likelihood score computed by considering the collection as one very long document as follows:

$$Score_{QL}(C) = \sum_{w \in Q} \log P(w|C) \quad (9)$$

- **Temporal Clarity (t-CLR):** t-CLR [7] is a variant of the clarity predictor that measures the KL-divergence between the *temporal profile* of the query (represented by $P(t|Q)$) and the *temporal profile* of the collection (represented by $P(t|C)$), as follows:

$$t\text{-CLR}(Q) = \sum_{t \in T} P(t|Q) \log \frac{P(t|Q)}{P(t|C)} \quad (10)$$

where $P(t|C)$ is estimated as a uniform distribution over all timestamps in C, and $P(t|Q)$ is estimated by first computing $\tilde{P}(t|Q)$ as follows:

$$\tilde{P}(t|Q) = \sum_{D \in R} P(t|D)P(D|Q) \quad (11)$$

where $P(t|D)$ is 1 for documents posted within the timestamp t, and 0 otherwise, and $P(D|Q)$ is estimated as discussed earlier. t is measured in units of h hours and h is a free parameter. $P(t|Q)$ is then computed by smoothing $\tilde{P}(t|Q)$ with the collection temporal model as follows:

$$P(t|Q) = \lambda \tilde{P}(t|Q) + (1 - \lambda)P(t|C) \quad (12)$$

where the smoothing factor λ is another free parameter for this predictor.

3. EXPERIMENTAL EVALUATION

In this section, we describe our experimental setup, followed by the results of the individual and combined predictors experiments respectively.

3.1 Experimental Setup

We conducted our experiments with two TREC tweets collections: Tweets2011 [13] and Tweets2013 [11]. Along with Tweets2011 collection, we used a merged set of the queries provided by TREC-2011 and TREC-2012 microblog tracks. As for Tweets2013, we used the queries provided by TREC-2013 microblog track. Both collections were accessible remotely through a search API provided by the microblog track organizers [11], who also made the collection statistics for both available. The queries distributed in microblog tracks are short (3.10 and 3.28 words on average for Tweets2011 and Tweets2013 respectively), resembling title-only queries in typical TREC collections. Table 1 below briefly presents both test collections.

Collection	Tweets (Time)	Queries	Source
Tweets2011	16M (16 days)	108	TREC'11-12
Tweets2013	243M (2 months)	60	TREC'13

Table 1: Tweets test collections used in our experiments.

3.1.1 Evaluation Measures

As commonly-used in similar studies, we used Pearson's r and Kendall's-τ correlations to measure the quality of each predictor [1]. The quality of prediction is estimated by correlating the actual Average Precision values (at cut-off 1000) of a set of queries, with their corresponding Average Precision values estimated by a predictor. In our experiments, Kendall's-τ correlation results showed very similar relative rankings of predictors to the ones resulting from Pearson's; thus we only report Pearson's r correlation results here. Higher values of Pearson's coefficient indicate better prediction quality.

3.1.2 Training and Testing Setup

To evaluate the quality of the predictors, we adopted a train-test approach proposed by Shtok et al. [15]. To tune the parameters of each predictor and measure its performance, we randomly split a query set into two subsets: a *training* (i.e., *tuning*) subset with 75% of queries and a *testing* subset with the remaining 25%. We tuned the free parameters of the predictors (by optimizing Pearson's coefficient) over the training subset, and then tested the optimized predictors over the testing subset. To avoid having a biased evaluation, we repeated this (split-tune-test) process 120 times and measured the final quality of each predictor by averaging the correlation values over the 120 splits[2]. Two-tailed paired t-test, with a significance level $\alpha = 0.05$, is used to determine statistically-significant differences in quality of the predictors [14].

While it is theoretically possible to follow a different evaluation approach in which we tune the predictors' parameters on one of the tweets collections and run the predictors on the other, we chose not to follow it due to the large difference in collection sizes and thus relevance sets.

3.1.3 Retrieval Models

To examine the robustness of predictors across different retrieval approaches, we measured the quality of predictions with three different retrieval models that were used earlier for ad-hoc search task in microblogs. The first is the Query Likelihood (QL) model [12] that is typically used in related QPP studies [2, 7, 15]. The other two adopt temporal models that consider the temporal nature of the data and the task: Time-based Exponential Priors (t-EXP) [10] and Time-based Query Relevance Modeling (t-QRM) [8]. For each retrieval model, tweets were ranked based on their retrieval scores, and the actual AP (at cut-off 1000) of retrieval for a query was computed given this ranking and the corresponding relevance judgments provided by TREC.

t-EXP has shown good retrieval performance for recency queries [4]. The model simply extends the QL model using an exponential decay factor as a document prior as follows:

$$P(D|Q) \propto P(Q|D) \cdot r \cdot e^{-r.t_d} \quad (13)$$

where $P(Q|D)$ is the query likelihood of the document D, r is the decay rate parameter, and t_d is the time difference in

[2]We tried different number of splits and different train/test distributions, but found the reported setting to produce the best results.

days between the posting time of D and the posting time of Q. In our experiments, we set r to 0.05.

t-QRM [8] is a variant of the typical query relevance modeling approach [12] in which the relevance model of the query is temporal and computed as follows:

$$P(w|Q) = \sum_{t \in T} P(w|t, Q)P(t|Q) \quad (14)$$

$$P(w|t, Q) = \sum_{D \in t} P(w|D)P(D|t, Q) \quad (15)$$

where t is a timestamp in unit of days, $P(t|Q)$ is estimated as the normalized sum of retrieval scores of documents in R posted within t, and $P(D|t, Q)$ is assumed to be uniform over all documents posted within t. In our experiments, we set number of tweets and terms considered in the model to 25 and 5 respectively.

The normalized retrieval score of a document, computed by a retrieval model, in the ranked list R was used to estimate $P(D|Q)$ in CLR and t-CLR for the corresponding retrieval model. The retrieval score of a document, computed by a retrieval model, was used instead of the log-likelihood score in WIG for the corresponding retrieval model. Collection score in NQC and WIG was always computed using the typical query likelihood model.

3.2 Individual Predictors

In this section, we discuss the evaluation results of the individual predictors with the three retrieval models. For each retrieval model, we followed the train-test approach discussed in section 3.1.2 to tune the free parameters and evaluate the predictors. Tables 2 and 3 show the prediction quality (measured by Pearson's r correlation) over Tweets2011 and Tweets2013 collections respectively. We only report the prediction quality of the best-performing three pre-retrieval predictors, i.e., SumIDF, DevIDF and SumSCQ (as SumIDF and DevIDF were the best performing over Tweets2011, and SumIDF and SumSCQ were the best over Twees2013), while we report the quality of all post-retrieval ones.

Predictor	QL	t-EXP	t-QRM
SumIDF	0.3332	0.3112	0.3540
DevIDF	0.3115	0.3393	0.2819
SumSCQ	0.2553	0.2404	0.2762
NQC	0.3542	0.3537	**0.4465**
NSD	0.3679	0.3470	(0.4161)
WIG	0.3852	0.4148	-0.1020
CLR	(0.5107)	(0.4818)	0.2646
***t*-CLR**	**0.5340**	**0.5225**	0.4155

Table 2: Pearson's correlation using Tweets2011. Best result per retrieval model is boldfaced and second-best is surrounded by parentheses.

An interesting observation drawn from Table 2 is that SumIDF, a pre-retrieval predictor, had comparable performance to NQC and NSD with QL and t-EXP models. It also outperformed CLR and WIG with the t-QRM model. This strong performance of SumIDF compared to post-retrieval predictors is in line with findings of Shtok et al. [15] over ClueWeb09 Web collection.

Table 2 also shows that t-CLR outperformed all other predictors for both QL and t-EXP retrieval models, indicating that temporal predictors might be more effective in predicting query performance for microblog search. A possible reason behind this is that temporal predictors consider the temporal nature of the tweets, queries, and task. Compared to CLR (which is the second-best predictor for QL and t-EXP), t-CLR has statistically significantly higher prediction quality with t-EXP and t-QRM, while difference was not significant with QL. Moreover, for t-QRM model, the improvement of performance of NQC and NSD (first and second best predictors) was not statistically significant compared to t-CLR.

Predictor	QL	t-EXP	t-QRM
SumIDF	0.1789	0.2481	0.2230
DevIDF	0.0111	0.0348	-0.0017
SumSCQ	0.1868	0.2227	0.2469
NQC	0.2450	0.1785	**0.5446**
NSD	0.3063	0.3274	0.2476
WIG	0.3284	(0.3651)	0.0121
CLR	**0.3800**	**0.3887**	0.1575
***t*-CLR**	(0.3470)	0.3431	(0.3787)

Table 3: Pearson's correlation using Tweets2013. Best result per retrieval model is boldfaced and second-best is surrounded by parentheses.

As Table 3 shows, SumSCQ had a better performance compared to WIG and CLR with t-QRM. It also had a better quality compared to NQC with t-EXP. Similar to Tweets2011, SumIDF continues to exhibit comparable performance to some post-retrieval predictors over Tweets2013, including NSD with t-QRM. It also showed better performance than WIG and CLR with t-QRM, and NQC with t-EXP. This robustness of SumIDF across different retrieval models and tweets collection, and given the very short length of queries, shows that SumIDF can be a somewhat helpful indicator of query performance in our studied domain. Furthermore, SumIDF can be efficiently-computed compared to post-retrieval predictors that require an initial search step; a feature that might be desirable in the real-time setting of the ad-hoc search task in microblogs.

Table 3 also indicates that CLR outperformed all other predictors with QL and t-EXP models. It also shows that t-CLR is the second-best performing predictor with QL and t-QRM models. It has a prediction quality that is statistically significantly higher compared to CLR ($p < 0.01$) with t-QRM. It exhibits a relatively good performance for both QL and t-EXP models, with non-significant difference compared to CLR with QL and a significant one with t-EXP. Moreover, the higher performance of WIG (second-best with t-EXP) was not statistically significant compared to t-CLR with t-EXP. That supports our argument above that temporal predictors are potentially good-fit for microblog search.

Considering both sets of results, we noticed that the quality of predictors is generally lower over Tweets2013 compared to Tweets2011. We argue that the small size of the query set (60 queries), and thus training set in particular (45 queries only), in Tweets2013 is the main reason behind that, which hindered better-tuning of the predictors parameters.

We also noticed that NSD and t-CLR are generally robust across different retrieval models in both collections. On the contrary, prediction quality of CLR had a *slight* drop with t-EXP (over Tweets2011) and a *severe* drop with t-QRM (over

both collections), compared to its performance with QL. A possible justification for that is the inconsistency in computing both the query and collection models. Recall that the collection model is computed in all cases using the typical query likelihood model. However, the query model used for t-EXP is generated using query-likelihood model with a non-uniform prior for document scoring (i.e., relatively closer to the estimated collection model), while the one used for t-QRM follows the temporal relevance model (i.e., very different from the the estimated collection model), which justifies the severe drop in quality in the latter case compared to the former. Following the same argument, we noticed a similar severe drop in quality of WIG with t-QRM model over both collections. We further investigated that by trying a variation of WIG that did not consider the collection score. The variation was adopted from Shtok et al. study [15]. Surprisingly, the variation still resulted in a very low prediction quality with the t-QRM. This indicates that some of the existing state-of-the-art predictors (with their current design) did not exhibit a consistent performance across the different retrieval models that are used in microblog search; which answers our second research question and highlights the need for further investigation.

Both tables 2 and 3 show notable improvement in performance of WIG over t-EXP. NQC also exhibited a notable improvement in performance with t-QRM; in fact, it was the best performing predictor with t-QRM in both collections. We are still investigating the reasons behind such behavior.

Finally, considering the results over both collections with QL model, Pearson's r correlation values of the best performing predictors in our study generally lie in a range that is comparable to the correlation values reported in related studies in the news and Web domains [3, 15]. This indicates that some of the existing state-of-the-art predictors exhibit a relatively good performance in the domain of microblogs as well, suggesting an answer to our first research question. However, the fact that t-CLR specifically was always among the best performing predictors, consistently with all retrieval models and across both collections, indicates that more attention should be given to that kind of predictors in the microblog domain. More investigation on such predictors is left as future work.

3.3 Combined Predictors

As combining predictors exhibited noticeable improvements in prediction quality in previous studies [2, 7], we conducted preliminary experiments on that in the context of microblog search using linear regression. We carried our experiments with all of the predictors and retrieval models used in our earlier experiments, but considering Tweets2011 only[3]. Due to the need for tuning the predictors parameters in addition to learning the regression model, we adopted a different experimental setup than the one described in Section 3.1. We adopted a 40-60 split approach of the query set, where 40% of the queries were used for tuning the predictors free parameters, and the remaining 60% were used for both training and testing the linear regression model using 10-fold cross-validation; this process was repeated 120 times. Again, Pearson's correlation coefficient, averaged over the 120 trials, was used as a measure of the prediction quality of individual predictors and any combined set of them.

For feature selection, we adopted a 2-step greedy-like approach that is a variant of the Greedy Stepwise approach [16] for searching the space of predictors. It first ranks the predictors descendingly based on their individual prediction quality, then incrementally combines the predictors, by adding one at a time, and then computes the average correlation of each new set. The algorithm then proceeds by adopting a leave-one-predictor-out approach on the optimal combined set to track the predictor(s) that improve prediction quality when eliminated. This leave-one-predictor-out process is repeated until no further improvement (by eliminating predictors) is achieved.

Applying our approach with the three retrieval models yielded a different optimal combined set of predictors with each. Tables 4 and 5 show the average correlation scores achieved by top-performing individual predictors[4] and by each optimal set, respectively. As expected, combining predictors did exhibit a noticeable improvement in prediction quality. The optimal combined sets of predictors achieved 21.6%, 27.8%, and 46.5% improvement in prediction quality over the best-performing individual predictor with QL, t-EXP, and t-QRM models respectively. This clearly confirms a positive answer for our third research question.

Predictor	QL	t-EXP	t-QRM
SumIDF	0.2012	0.1662	0.2393
SCS	-0.2542	-0.2704	-0.2711
NQC	0.2480	0.2464	**0.3575**
NSD	0.2573	0.2021	0.3330
WIG	0.3201	0.3515	-0.1309
CLR	**0.4518**	(0.4140)	0.1379
***t*-CLR**	(0.4251)	**0.4207**	(0.3415)

Table 4: Pearson's correlation achieved by individual predictors using Tweets2011 averaged over 40-60 splits. Best correlation per retrieval model is boldfaced and second-best is surrounded by parentheses.

Model	Optimal Set of Predictors	Pearson's
QL	{t-CLR, CLR, WIG, SCS}	0.5496
***t*-EXP**	{t-CLR, WIG, SCS}	0.5375
***t*-QRM**	{t-CLR, NQC, NSD, SumIDF}	0.5238

Table 5: Pearson's correlation achieved by optimal combined predictor sets using Tweets2011.

It is worth noting that not only has t-CLR persisted in the three optimal sets as shown in Table 5, it has also achieved the best or second best prediction quality with the t-EXP, t-QRM and QL retrieval models, as shown in Table 4.

We also notice that the impact of leaving out t-CLR of the optimal set was exceptionally the highest among other predictors across the three retrieval models. Eliminating t-CLR from the three optimal sets reduced the correlation by 11.4%, 48.3% and 11.7% with QL, t-EXP, and t-QRM retrieval models respectively. This ascertains the predictive

[3]The number of queries used with Tweets2013 was too small to support both tuning of predictors parameters and learing the regression model.

[4]using the 40-60 splits in contrast to the 75-25 splits used in the earlier experiments in Section 3.2.

quality of the temporal predictor t-CLR in the context of microblog search and supports our earlier observations about it as well.

4. CONCLUSION AND FUTURE WORK

In this work, we studied query performance prediction in the context of microblog ad-hoc search. We examined several pre- and post-retrieval predictors over two different microblog collections and with three retrieval models. Results showed that, in general, the temporal predictor (t-CLR) demonstrated robust prediction quality over both collections and across the three retrieval models. It was also the best-performing predictor over one of the collections with two retrieval models and the second-best over the other with two retrieval models as well, which indicates that temporal predictors might be more suitable for the nature of the data and the task. The study also shed some light on the performance of the existing state-of-the-art predictors and showed inconsistency in prediction quality of some of them across different retrieval models used in the context of microblog search. Furthermore, experiments with combination of predictors showed that combining predictors improved prediction quality compared to individual predictors.

The conducted study opened up several directions for future work. First, there is a need to re-design some of the existing predictors to fit properly with the state-of-the-art retrieval models in the context of microblog search. Second, the study showed also the need for performance predictors that explicitly consider the temporal aspect of the task and the data. The new predictors might also leverage some specific features of the data, e.g., retweets and hashtags. Third, as the small regression experiments showed promising results, more extensive experiments on combining predictors are needed. Fourth, in this study, predictors were used to estimate Average Precision as the standard measure used in QPP literature; evaluating predictors against other measures usually used in the microblog search domain (e.g., precision at 30) is worth exploring. Finally, leveraging performance predictors in retrieval models (e.g., query expansion) is definitely one of the potential future directions.

5. ACKNOWLEDGMENTS

This work was made possible by NPRP grant# NPRP 6-1377-1-257 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

6. 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, Jan. 2010.

[2] 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*, 2002.

[3] R. Cummins, J. Jose, and C. O'Riordan. Improved query performance prediction using standard deviation. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[4] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[5] C. Hauff, L. Azzopardi, and D. Hiemstra. The combination and evaluation of query performance prediction methods. In *Advances in Information Retrieval*, number 5478 in LNCS, pages 301–312. 2009.

[6] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In *String Processing and Information Retrieval*, number 3246 in LNCS, pages 43–54. 2004.

[7] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14:1–14:31, July 2007.

[8] M. Keikha, S. Gerani, and F. Crestani. Time-based relevance models. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

[10] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, 2003.

[11] J. Lin and M. Efron. Overview of the TREC-2013 Microblog Track. 2013.

[12] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, Cambridge, United Kingdom, 2008.

[13] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the TREC-2011 Microblog Track. 2011.

[14] F. Raiber and O. Kurland. Using document-quality measures to predict web-search effectiveness. In *Advances in Information Retrieval*, number 7814 in LNCS, pages 134–145. Jan. 2013.

[15] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Trans. Inf. Syst.*, 30(2):11:1–11:35, 2012.

[16] I. H. Witten, E. Frank, and M. A. Hall. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, third edition, 2011.

[17] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 2001.

[18] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *Advances in Information Retrieval*, volume 4956 of *LNCS*, pages 52–64. 2008.

[19] Y. Zhou. *Retrieval performance prediction and document quality*. PhD thesis, University of Massachusetts Amherst, 2007.

[20] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.