Information Processing and Management 53 (2017) 1320–1341

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Query performance prediction for microblog search

Maram Hasanain*, Tamer Elsayed

Department of Computer Science and Engineering, College of Engineering, Qatar University, Qatar

ARTICLE INFO

Article history:
Received 22 January 2017
Revised 3 June 2017
Accepted 14 August 2017

Keywords:
Algorithms
Experimentation
Performance
Query difficulty
Temporal retrieval

ABSTRACT

Query performance prediction (QPP) is the task of estimating the effectiveness of a retrieval system given a search query in the absence of any feedback from the searcher. The task has been proven to be very challenging, and thus it attracted a lot of research attention in domains like news and Web retrieval. However, search in microblogs poses new challenges for the task due to the more prevalent temporality in microblogs and the different types of information needs in such domain. In this work, we aim at studying QPP for microblog search. We conducted large-scale experiments, testing 37 state-of-the-art predictors using several types of retrieval models usually used in microblog search. Moreover, we propose a set of predictors that exhibit statistically-significant improvements over the state-of-the-art predictors with the maximum percentage of improvement reaching 55% over all studied retrieval settings. Further experimental explorations show that using expanded queries in predicting the performance of query expansion models gives much better prediction quality than using the original queries, and that the examined predictors were generally much more effective over temporal queries compared to non-temporal ones; both phenomena have never been studied in the context of microblog search before. As microblog search is considered a major step in several retrieval tasks in the domain (such as timeline generation, summarization, and question answering), improving QPP for microblog search has a high potential to help improve the effectiveness of those closely-related tasks.

1. Introduction

In recent years, users have turned to microblogging services to share information that is as personal as their daily life details, up to the most general topics discussed world-wide. Twitter is indeed one of the fastest growing microblogging services. In 2013, Twitter reported that more than half a billion tweets are posted daily.[1] Twitter users usually share information, news, and opinions about ongoing events via *tweets*, where a tweet is the unit of information sharing on Twitter, with a maximum length of 140 characters that possibly contain hashtags,[2] mentions,[3] and URLs.

With the large number of tweets posted daily, a large number of search queries are being issued by Twitter users, who expect relevant tweets to their queries. Golovchinsky and Efron (2010) surveyed 23 people on their usage pattern of Twitter

* Corresponding author.
E-mail addresses: maram.hasanain@qu.edu.qa (M. Hasanain), telsayed@qu.edu.qa (T. Elsayed).

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

[2] A hashtag is constructed using the # symbol followed by one or more words and is usually added to a tweet to reflect the topic of the tweet.

[3] A mention is represented by the @ symbol followed by a Twitter username and is a means of *tagging* other Twitter users in one's tweet.

http://dx.doi.org/10.1016/j.ipm.2017.08.002
0306-4573/

search interface and found that more than half of them search in Twitter at least once a week. Moreover, Teevan, Ramage, and Morris (2011) collected search queries issued to Twitter through Bing Toolbar and found that during two weeks in November 2009, more than 126K queries were issued to Twitter by users from the United Stated alone. In a more recent work, Elsweiler and Harvey (2015) asked 68 Twitter users about a recent search task they have done and 45% of the respondents stated that they do such search task at least once a week. Conducting an analysis over tweeting and search behavior on a much larger scale, Lin and Mishne (2012) inspected the full Twitter stream during October 2011 and found that Twitter search serves more than 2B queries and API requests a day.

In a typical microblog ad-hoc search system, a user poses a query reflecting her information need and the system responds with a set of retrieved microblogs usually arranged in a form of a ranked result list. In that list, tweets are usually ranked in a descending order by their potential relevance (that is usually determined by a retrieval score assigned by the system and hidden from the user). The result list includes supposedly relevant microblogs to that information need. Some users' queries might be handled *effectively* by the system, i.e., the system will manage to retrieve relevant microblogs and rank them high in the result list of those queries. However, other queries can be difficult for the system to answer, resulting in a poor quality of results and thus poor user experience.

The microblog search system can attempt to improve retrieval effectiveness for poorly-performing queries. Yet, for the system to do so, it should be able to accurately *estimate* (or *predict*) how satisfied the user will be with the retrieved results prior to presenting the result list to her, and specifically in the general situation where the system lacks user-provided relevance information. The process of doing so is called **Query Performance Prediction (QPP)**.

Query Performance Prediction is the problem of predicting retrieval performance for a query given: (a) the retrieval model used to answer the query and (b) the collection of documents from which the retrieval model retrieves documents. Prediction is usually performed in the absence of relevance information (Carmel & Yom-Tov, 2010; Zhou, 2007). We will be referring to the approach that performs prediction given these variables by a *predictor* that will be used to compute a *predicted quality* of the results returned by the retrieval model given the query.

A large body of research work has studied QPP. Some studies focused on QPP methods ranging from those that only examine the query terms (i.e., pre-retrieval predictors) (Hauff, Azzopardi, & Hiemstra, 2009; He & Ounis, 2004; Zhao, Scholer, & Tsegay, 2008) to methods heavily relying on analyzing retrieved documents for a query (i.e., post-retrieval predictors) (Cronen-Townsend, Zhou, & Croft, 2002; Cummins, Jose, & O'Riordan, 2011; Shtok, Kurland, Carmel, Raiber, & Markovits, 2012; Zhou & Croft, 2007). Others have recently proposed frameworks designed to theoretically categorize and evaluate predictors (Makarenkov, Shapira, & Rokach, 2015) or to better explain and integrate them (Shtok, Kurland, & Carmel, 2016).

Most of those studies were conducted on the ad-hoc search task in the context of typical TREC Web and news collections (e.g., WT10g (Hawking, 2000) and GOV2 (Clarke, Craswell, & Soboroff, 2004)). Web and news documents are generally long with an average document length in many Web and news collections exceeding 265 words (Hauff et al., 2009). However, tweets are much shorter with a maximum length of 140 characters. Tweets are also naturally different from those documents as they tend to be conversational and more temporal with a very short lifespan (Alonso, Baeza-yates, Strtgen, & Gertz, 2011; Boyd, Golder, & Lotan, 2010; Elsweiler & Harvey, 2015). Moreover, the task of searching tweets is naturally different from searching the Web, as the information needs are very different (e.g., following recent events) than the typical types of web search (e.g., transactional or navigational) (Oeldorf-Hirsch, Hecht, Morris, Teevan, & Gergle, 2014; Teevan et al., 2011). The distinct features of both the *microblog search task* (Teevan et al., 2011) and the *microblog data* trigger the need to revisit the problem of query performance prediction in such domain.

1.1. Research questions

Few recent research studies have tackled this problem in the context of microblog search (Hasanain, Malhas, & Elsayed, 2014; Rodriguez Perez & Jose, 2014), which raises several research questions.

First, we are not sure (or we cannot anticipate) if the current state-of-the-art predictors will perform as good on microblog search as they do on the well-studied Web and news search. Moreover, earlier work on studying state-of-the-art predictors in microblog search is just preliminary with very little analysis carried to understand this problem within this domain (Hasanain et al., 2014; Rodriguez Perez & Jose, 2014). In an effort to understand how existing predictors will perform with microblog retrieval settings, we experiment with 37 state-of-the-art predictors covering predictors that are very different from each other. We test QPP over the two commonly-used tweet collections: Tweets11 (Ounis, Macdonald, Lin, & Soboroff, 2011) and Tweets13 (Lin & Efron, 2013).

Second, whether state-of-the-art predictors proved to be effective or not, can we further improve the prediction quality for microblog search by proposing new predictors? We propose few simple yet highly-effective changes to some existing microblog-specific predictors that significantly improve prediction quality.

Third, it is crucial to measure the quality of QPP for microblog search across different retrieval models that are proved effective for that task. Existing QPP studies in the domain have either studied few predictors with several retrieval models (Hasanain et al., 2014) or only experimented with a single retrieval model (Rodriguez Perez & Jose, 2014); we experiment with four retrieval models covering a variety of types typically used in this domain (e.g., temporal, and query expansion models).

Finally, with the temporal nature of data and search task, it is interesting to study the effect of the temporality of queries (i.e., temporal or non-temporal) on the prediction quality. For that, we have categorized the queries into tempo-