Towards Personalized Search for Tweets

Akshay Choche and Lakshmish Ramaswamy

Department of Computer Science,
The University of Georgia,
Athens Georgia 30602
{choche,laks}@cs.uga.edu

Abstract. Powerful search capabilities are fundamentally important for micro-blog-based information systems such as Twitter. While recently there has been some works aimed at enhancing the scalability of micro-blog search, very few existing techniques incorporate personalization into their search and ranking processes. This paper argues that since Twitter is a social network (SN)-based micro-blog system, it is essential to personalize search results taking into account the social relationships among various users. In this paper, we outline a scalable and personalized tweet search framework that takes into account the search parameters, the distances of the follower relationships, and the temporal aspects of the tweets when ranking the search results.

Keywords: Micro-blogs, Search, Social Network, Personalization.

1 Introduction

Powerful search techniques are indispensable to micro-blog-based cooperative information systems such as Twitter, as they seek to establish themselves within the highly competitive online social media space [5]. However, searching on SN-based microblog platforms is radically different than searching content on the World Wide Web, and it poses several unique challenges. First and foremost, micro-blogs (henceforth referred to as *tweets*) being very short pieces of text (maximum of 140 characters), heavyweight text analytic techniques are not very effective. Second, as Twitter is most commonly used for sharing updates about current events and issues, freshness and timelines of search results become critically important. Third, since Twitter also incorporates SN features, it is important to take SN-relationships into account during search and ranking processes. Unfortunately, very few existing works comprehensively address the above challenges [1] [2] [4]. To the best of our knowledge, none of the existing works incorporate *personalization* into their search and ranking strategies. Thus, the social network aspect of Twitter is ignored during search.

In this paper, we argue that it is important to personalize search results based on SN relationships. The follower relationship in Twitter is often an indicator of commonality of interests and opinions. Thus, it is natural for a user issuing a query to expect higher rankings to matching tweets from users that she

Y.T. Demey and H. Panetto (Eds.): OTM 2013 Workshops, LNCS 8186, pp. 722–725, 2013.
© Springer-Verlag Berlin Heidelberg 2013

is transitively following even when those tweets are not the most recent ones. However, personalization also raises new scalability issues not only because it requires additional analytics, but also because it reduces the efficacy of caching. In this paper, we outline a scalable and personalized search technique for tweets. Our technique combines three major tweet relevance factors, namely, degree of syntactic match, distance of the follower relationships, and the temporal recency of tweets into a unique tweet ranking metric.

2 Personalized Tweet Search Scheme

Shortness of tweets (limited to 140 characters), makes it necessary to use additional contextual information during search and ranking. The follower relationships in Twitter often embody commonalities in interests and opinions. When a user U_1 chooses to follow another user U_2, it inherently signifies that U_1 is interested in the updates and opinions being posted by U_2. Applying this logic transitively, we hypothesize that U_1 will be much more satisfied with the search results if matching tweets (tweets containing one or more search query words) from users that she is directly or indirectly following are ranked high. A second benefit of using SN-relationships is in terms of *search query disambiguation*. For example, the search query "Stephen King" can either refer to the famous author or the famous soccer player. It may be possible to discern the querier's intent by checking if she predominantly follows literary personalities or sport commentators.

2.1 Technique Overview

Our search technique operates in two phases. First, we use a simple keyword-based filter to retrieve tweets that are relevant to a given query. This phase is intentionally kept simple so as to filter-out large fractions of obviously-irrelevant tweets. The tweets that are deemed relevant are processed by the second phase which generates personalized search results by using our novel personalized tweet ranking algorithm. Adopting a two phase scheme vastly reduces the overheads of personalization [3].

Personalized Tweet Ranking: One of the main challenges in ranking tweets is that several factors influence the relative importance of tweets. In our work, we have identified three main factors that influence the relative significance of tweets to a given query. The first is the strength of the SN-relationship between the tweet's author and the user issuing the query. We call this the *personalization factor*. The second is the *temporal significance factor*, which capture the temporal relevance of a tweet to the query. The third is the *author influence* factor, which measures social influence of a tweet's author. Each of these factors have very distinct significance, and we combine these factors into a personalized tweet ranking function as follows.

$$CRS(Tw_k) = w_1 * PS(Tw_k) + w_2 * TS(Tw_k) + (1 - w_1 - w_2) * AIS(Tw_k) \quad (1)$$

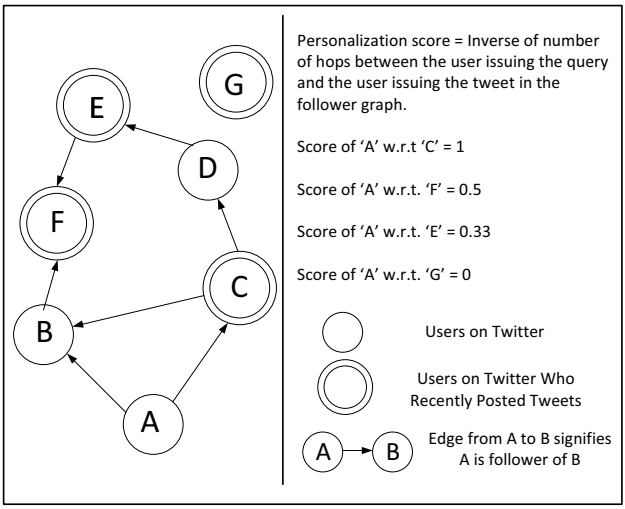

Fig. 1. Personalization using relationship property

In this function, $CRS(Tw_k)$ indicates the cumulative ranking score for tweet TW_k, $PS(Tw_k)$ denotes the tweet's personalization score, $TS(Tw_k)$ denotes its temporal score and $AIS(Tw_k)$ indicates its author influence score. w_1 and w_2 are system-defined weight parameters such that $w_1 \geq 0$, $w_2 \geq 0$ and $w_1 + w_2 \leq 1$. The values of w_1 and w_2 determine the relative importance of the different factors. We now discuss our mechanisms for quantifying each of these factors.

Personalization Score (PS)*:* This score measures the *social affinity* of the user issuing the query to the tweet's author. To compute the social affinity, we model the follower relationship as an unweighted directed graph in which vertices correspond to users and edges correspond to the follower relationship between users. We use the inverse of the length (number of hops) of the shortest path to quantify the personalization score. Our choice is influenced by computational efficiency and stability of the parameter over time. Figure 1 illustrates the personalization score computation on a hypothetical follower graph. Personalization score of A with respect to C is 1 because A is a direct follower of C, whereas the personalization score of A with respect to G is zero because there is A is not a direct or indirect follower of G.

Temporal Score (TS)*:* The temporal score measures the temporal closeness of a tweet to a query. The rationale is to provide higher preference to more recent tweets. Suppose a tweet Tw_i was posted at time T_x. This tweet's temporal score with respect to query Q_j issued at time T_y is quantified as $\frac{1}{T_y - T_x + 1}$.

Author Influence Score (AIS)*:* The author influence score measures the influence of a tweet's author on the Twitter user community. The rationale is to give higher ratings to tweets from more popular Twitter users. The author influence score of a tweet is defined as the ratio of the number of followers of the tweet's author to the total number of currently active Twitter users.

Ranking Algorithm: At a conceptual-level, the ranking algorithm is quite simple – the CRS of each matching tweet is calculated and the tweets are ranked in the decreasing order of their CRS values. However, this naive implementation does not scale well because personalization score computation will quickly become a bottleneck. Thus we need smarter ways to generate the ranking. Our system incorporates an approximation strategy to accelerate the ranking process. Our algorithm proceeds as follows. First, we select only tweets whose temporal score exceeds a certain threshold μ. Computing this list is computationally scalable because we are essentially selecting tweets that are more recent than a certain time-point. We compute the PS and the AIS values only for the tweets whose TS value exceeds μ. Once these values are available, we generate the ranked list. Since PS values (which are the most expensive) are computed only for a limited number of tweets, our algorithm is much more scalable and efficient. Our experiments show that the above approximation strategy is very effective in minimizing personalization overheads [3].

3 Summary

In this paper, we outlined a scalable approach for incorporating personalization into search and ranking of tweets. Our approach is based upon a unique tweet ranking metric for generating personalized search results. This ranking metric takes into consideration three major tweet relevance factors, namely, the syntactical similarities between the tweet and the query, the distance of follower relationships, and the temporal recency of the tweet. We also presented a two-phased approximation strategy for enhancing the efficiency and scalability of personalized search.

References

1. Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S., Lin, J.: Earlybird: Real-time search at Twitter. In: ICDE (2012)
2. Chen, C., Li, F., Ooi, B.C., Wu, S.: TI: An Efficient Indexing Mechanism for Real-time Search on Tweets. In: ACM-SIGMOD (2011)
3. Choche, A., Ramaswamy, L.: REPLETE: A Realtime Personalized Search Engine for Tweets. Technical Report, Dept. of Computer Science, The University of Georgia (2013)
4. Dong, A., Zhang, R., Kolari, P., Bai, J., Diaz, F., Chang, Y., Zheng, Z., Zha, H.: Time is of the essence: improving recency ranking using twitter data. In: ACM-WWW 2010 (2010)
5. Teevan, J., Ramage, D., Morris, M.R.: # TwitterSearch: A Comparison of Microblog Search and Web Search. In: ACM-WSDM 2011 (2011)