Contents lists available at ScienceDirect

Information Sciences

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

Real-time recommendation for microblogs

Xianke Zhou [a], Sai Wu [a,*], Chun Chen [a], Gang Chen [a], Shanshan Ying [b]

[a] College of Computer Science, Zhejiang University, Hangzhou 310027, PR China
[b] School of Computing, National University of Singapore, Computing 1, 13 Computing Drive, Singapore 117417, Singapore

ARTICLE INFO

Article history:
Received 19 April 2013
Received in revised form 19 February 2014
Accepted 29 March 2014
Available online xxxx

Keywords:
Twitter
Microblog
Real-time recommendation

ABSTRACT

Existing microblogging systems, such as Twitter, provide global or local discussion trends, so that users can easily find hot topics to follow on. The discussion trends are discovered by analyzing statistics of the microblogging database globally or regionally. As a consequence, users who select the same region, receive the same set of topics irrespective of their interests. This strategy fails to deliver relevant recommendations to the right users. To address this problem, in this paper, we propose a real-time customized recommendation scheme for microblogging systems. Our approach identifies users' interests via their personal tags and builds tag-user graphs for computing the similarities between microblogs and users. The subsequently submitted microblogs are considered as an input stream that flows through our tag-user graphs and buffered for the interested users. The user's browser can update his/her recommendations in real-time by pulling microblogs from our buffer. A statistics-based pruning approach, called APS (Approximate Pruning Scheme), is applied to reduce the processing cost by effectively avoiding unnecessary comparisons. We evaluate our system with two real datasets, namely the Twitter dataset and the Netease dataset. Our extensive experimental study shows the scalability and efficiency of our approach.

1. Introduction

The popularity of smart phones boosts the use of microblogging systems, where users can update their status and share news snippets using their mobile phones anywhere anytime. Due to its real-time requirement, the microblogging system is challenging the conventional news media. It allows the non-journalists to report their nearby breaking news and hence improves the freshness and timeliness of the news. Therefore, some people call the microblogging systems "the electronic mouth" [19]. As an example, the first wave of reports of 2008 Mumbai terrorist attack was from Twitter (http://twitter.com), neither CNN nor BBC.[1] Conventional news media are aware of the potential of microblogging systems and have started incorporating their services into those systems.[2]

In current microblogging systems, such as Twitter, news are disseminated to users via one or more of the following three services. First, all events of a user's followees are pushed to him/her, which allows the user to track the status of the people he/she follows. But this service limits the sources of news (only from the followees) and is not capable of reflecting breaking news in the system. Second, some discussion trends are retrieved by the system and recommended to all users. As the

* Corresponding author. Tel.: +86 15605813329.
E-mail addresses: xiankz@zju.edu.cn (X. Zhou), wusai@zju.edu.cn (S. Wu), chenc@zju.edu.cn (C. Chen), cg@zju.edu.cn (G. Chen), shanshan@comp.nus.edu.sg (S. Ying).

[1] http://techcrunch.com/2008/11/26/first-hand-accounts-of-terrorist-attacks-in-india-on-twitter/.

[2] http://www.zdnet.com/blog/btl/twitter-and-news-media-a-long-term-relationship-or-just-a-fling/16409.

http://dx.doi.org/10.1016/j.ins.2014.03.121
0020-0255/© 2014 Elsevier Inc. All rights reserved.

recommendation is based on the statistics of the system, the top discussion trends typically reflect the global or regional breaking news, such as Iran election in 2009, Gulf Oil Spill in 2010 and Japanese earthquake in 2011. Besides the global trends, users can also switch between different regional trends, e.g., Twitter has trending topics for about 30 countries. However, regardless of users' interests, the system always recommends the same trends set to users who choose the same region. Even if your city is holding a carnival, you are not likely to be notified in this strategy. Finally, the search engine equips users with the tool to retrieve microblogs via keywords. Most microblogging search engine are fast and return real-time search result; the problem is however, due to the huge volume of microblogs being submitted every second, a user has to browse a large set of results before locating his/her desired ones.

In this paper, we design a real-time customized recommendation scheme for the microblogging system. Considering the system as a streaming system, tweets (to simplify the presentation, we use *tweet* to denote all types of *microblogs*) are inserted into the system continuously. Each user in the system is associated with a set of weighted tags to quantify his/her interests. Based on the users' interests, the recommendation service streams the tweets to the corresponding buffers. The server periodically pushes the recommended tweets to each online user. In this manner, the recommendation service provides customized news for each online user. It enables the user to filter the information and select a few interested tweets among thousands of candidates. Note that the recommended tweets are not used to replace the global trends. Instead, global trends still show the most popular topics in the whole network, while our recommendation service provides real-time news feed based on the users' interests. It targets at individual users and is a good supplement to the trend-based solution.

To illustrate the importance of real-time customized recommendation, Fig. 1 compares the difference between the global discussion trend and tweets recommended by our technique. In this example, we use tweets published on June 26th, 2009, the second day when the pop-star Michael Jackson was found dead, as well as the next day of the 2009 NBA Draft. The Global Trend from Twitter, on the leftmost column in Fig. 1, perfectly captured what were people discussing that day with keywords such as "michael", "dead", "pop", and "die". As a fan of NBA, however, a user may keen to see the news and people's discussion about the NBA Draft, which was sadly submerged by the pop-star's departure in the Global Trend. The three columns on the right present four top-ranked real-time customized recommendations for every 20 s, for a user tagged with "kobe bryant", "lakers la" and "NBA". Apparently, our customized recommendation fits more to user's interests and servers as a perfect complement of the Global Trend. Besides, the real-time mechanism catches the fresh news specially for each user, speeding up the broadcasting of breaking fresh news to the targeted users.

Fig. 2 shows how our service provides customized recommendation results for different users. We randomly pick three users from three different categories: entertainment, sports and computing. For each user, we show his/her personal tags and top-ranked real-time recommended tweets. We can see that the topics of the recommended tweets are very different from the global trends, as expected. The first user still gets the news about Michael Jackson, as he is interested in the entertainment news. The other two users are notified with the NBA news and IT news instead, capturing their personal interests well.

For a large-scale microblogging system however, providing customized recommendation for each user is rather expensive. Conventional recommendation approaches, such as collaborative filtering [3,6,7] suffering from scalability and sparsity, cannot be adopted in such circumstance. The challenges of providing real-time customized recommendation are twofold: the quality and the efficiency. Given a tweet t, we need to measure its relevance to a user u. For this purpose, a set of tags are created to identify users' interests. Some microblogging systems, such as Sina (http://t.sina.com.cn) and Netease (http://t.163.com),[3] allow the users to create their personal tags, which can be directly utilized by our approach. For other systems (e.g., Twitter and Tumblr), tags are automatically inferred based on user's tweets published. It is obvious that for different users, they show different interestingness on the same tag, e.g., one user prefers sports to music, while another chooses the opposite. To indicate the importance of a tag for a specific user u, a weight matrix M is created based on the social relationships of u to figure out the weights.

Popular microblogging systems support millions of users and handle thousands of tweets every second on average. Computing the similarity between every user-tweet pair incurs extremely high overheads to the recommendation service. To reduce the cost, in this paper, we propose the APS, *Approximate Pruning Scheme*. The APS exploits the statistics of tags and tweets to prune trivial tweets for each user without buffering it. The intuition is to avoid unnecessary computations since only a small number of high-relevant tweets have the opportunity to be sent to the potential readers. The APS also groups users with similar interests. Instead of keeping the recommendation buffer for each online user, APS computes the similarity between a tweet and a group of users. The grouping strategy further alleviates system cost while maintaining high accuracy. By adopting the Approximate Pruning Scheme, we can handle more than ten thousand of tweets per second.

In summary, the contributions of this paper are as follows:

1. A weighted-tag retrieval approach upon social relationships. We adopt the conventional IR approach to identify the most expressive keywords from a user's tweets, which are referred to as his/her tags. Besides, to distinguish the importance of tags, for each user, we create a tag weight vector, which exploits his/her social relationships to other users (via the following and followed links).

[3] Sina (NASDAQ:SINA) and Netease (NASDAQ:NTES) are the largest internet companies in China, which are currently ranked the 17th and 31th most visited sites in Alexa Traffic Rank (http://www.alexa.com/topsites) respectively. For more details of Netease, please refer to http://ir.netease.com/phoenix.zhtml?c=122303&p=irol-IRHome.

Please cite this article in press as: X. Zhou et al., Real-time recommendation for microblogs, Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.03.121