Topic-STG: Extending the Session-based Temporal Graph Approach for Personalized Tweet Recommendation

Jianjun Yu
Computer Network Information Center, CAS
yujj@cnic.ac.cn

Yi Shen
University of Chinese Academy of Sciences
shenyi@cnic.ac.cn

Zhenglu Yang
Institute of Industrial Science, The University of Tokyo, Japan
yangzl@tkl.iis.u-tokyo.ac.jp

ABSTRACT

Micro-blogging is experiencing fantastic success in the world-wide. However, during its rapid development, it has encountered the problem of information overload, which has troubled many users. In this paper, we mainly focus on the task of tweet recommendation to address this problem. We extend the session-based temporal graph (STG) approach as Topic-STG for tweet recommendation which comprehensively considers three types of features in Twitter: the textual information, the time factor, and the users' behavior. The experimental results conducted on a real dataset demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering; H.2.8 [**Database Applications**]: Data Mining

Keywords

recommendation; graph; topic model; time-sensitive.

1. INTRODUCTION

Twitter and other micro-blogging systems not only act as the role of social relation between people, but also as important sources for people to obtain useful information. Currently, there are more than 200 million messages generated on Twitter each day. This scale of data benefits to users while floods them with huge volumes of noise, and thus puts them at risk of information overload. Currently, researchers mainly employ the content features, user relationships as well as interactions to build user interest model for tweet recommendation. However, most of them have not taken into account the effect of the time factor. We have observed that users' behaviors on Twitter are determined by both long-term and short-term interests, a similar scenario described in [1]. In other words, users' interest may change with time. Moreover, a considerable part of tweets are time-sensitive (e.g. a tweet about "WWW 2014"). As a result, the temporal tweet recommendation is necessary when modeling user interest and improving the precision of recommendation.

Xiang et al [2] proposed a novel recommendation approach named session-based temporal graph (STG). As shown at the bottom part of Figure 1, STG is a bipartite graph with three types of nodes: user-node, item-node, and session-node. The session-node is a combined node of a user and a specific time bin, i.e, $S1(U_1, t_1)$. The data for temporal recommendation is converted into the following form: <user, item> and <session, item>, where the former represents the long-term interests and the latter indicates the short-term interest. Every edge has a weight in STG. Finally, the authors transformed the recommendation problem into calculating the sum of the scores of the shortest paths between corresponding nodes.

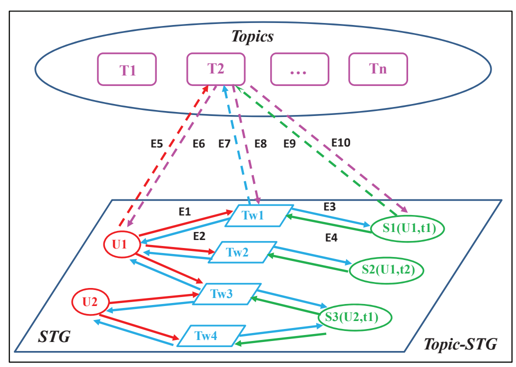

Figure 1: An example of Topic-STG.

Although STG-based approach has been proved to be significantly effective on most kinds of datasets, it is still not suitable to the tweet-recommendation task for two reasons: firstly, the original STG approach neglects the textual information of tweets, which presents rich sentiment information to predict user preference in Twitter; secondly, STG only recommends those "previous visited" tweets, however, our task is to recommend those potential tweets that a user may be interested in. As a result, we propose an extended approach called Topic-STG, which can bridge STG and textual information via adding "topic-node" into the existing bipartite graph.

2. TOPIC-STG RECOMMENDATION

As the same with STG approach, once a user U_1 operates on a tweet Tw_1 (i.e, "retweet", "comment", or "favorite"),

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.
WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577328.

Table 1: The weight of edges in Topic-STG

id	edge type	weight [*]
E1	user → tweet	$W_U^{tw} = \frac{1}{N(U,Tw)}$
E2	tweet → user	$W_{Tw}^{U} = \frac{W_U^{Tw}}{\sum_{U'} W_U^{Tw} + \sum_{S'} W_{S(U,t)}^{Tw}}$
E3	tweet → session	$W_{S(U,t)}^{Tw} = \frac{W_{S(U,t)}^{Tw}}{\sum_{u'} W_u^{Tw} + \sum_{S'} W_{S(U,t)}^{Tw}}$
E4	session → tweet	$W_{Tw}^{S(U,t)} = \frac{1}{N(S(U,t),Tw)}$
E5	user → topic	$W_U^T = P(T\|U)$
E6	topic → user	$W_T^U = \frac{W_U^T}{\sum_{U'} W_U^T + \sum_{S'} W_{S(U,t)}^T}$
E7	tweet → topic	$W_{Tw}^T = P(T\|Tw)$
E8	topic → tweet	$W_T^{Tw} = \frac{P(T\|Tw)}{\sum_{tw'} P(T\|Tw)'}$
E9	session → topic	$W_{S(U,t)}^T = P(T\|S(U,t))$
E10	topic → session	$W_T^{S(U,t)} = \frac{W_{S(U,t)}^T}{\sum_{U'} W_U^T + \sum_{S'} W_{S(U,t)}^T}$

[*] $N(U,Tw)$ is the number of tweets Tw operated by U, $N(S(U,t),Tw)$ is the number of Tw operated by U at time t. $P(T|U)$ is U's preference on topic T , $P(T|S(U,t))$ is U's preference on T at t, $P(T|Tw)$ is Tw's probability distribution on T. W_{Tw}^T, W_U^T, $W_{S(U,t)}^T$ are normalized weight of corresponding topic-related in-degree edges. W_U^{Tw} and $W_{S(U,t)}^{Tw}$ are normalized weight of corresponding operation-related in-degree edges.

two pairs of edges (E1-E4) will be created. The difference is that the topic-node will be automatically generated by LDA (Latent Dirichlet Allocation). We train the latent topics to infer tweets' topic distribution, long-term topic distribution and short-term topic distribution of users, and the correspondingly topic-related edges will be created, i.e, E5-E10.

With the new-added edges and their weights (expressed in Table 1), we are able to recommend candidate tweets to a user U at a timestamp t based on the following steps:

1. Select the corresponding user-node and session-node as the source nodes.

2. Use breadth-first strategy (BFS) to search the shortest paths between the source nodes and the candidate tweet-nodes. There exist 10 types of shortest paths comparing with STG approach: P1: U→ I → U→ I; P2: U→ I → S→ I; P3: S→ I → U→ I; P4: S→ I → S→ I; P5: U→ I → T→ I; P6: S→ I → T→ I; P7: U→ T → S→ I; P8: S→ T → U→ I; P9: U→ T → U→ I; P10: S→ T → S→ I, where P1-P4 are the same as the STG approach, and the other paths are new-added ones. Obviously, P5, P6 are calculated by the content-based recommendation approach, and P7-P10 are calculated by the user-based collaborative filtering approach.

3. The score of each path $\phi(P) = \prod_{v_k \in p} w(v_k, v_{k+1})\gamma(v_0)$, where $\gamma(v_0)$ is an indicator for the influence of different starting points. $\gamma(v_0) = \beta$ if $v_0 = v_U$, and $\gamma(v_0) = 1 - \beta$ if $v_0 = v_S$.

4. The preference of U on Tw is computed as $p_U^{Tw} = \sum_{P \in P(U,Tw)} \phi(P)$, where $P(U,Tw)$ represents the set of the shortest paths between U and Tw.

5. Finally, we choose the top-N tweets with the highest p_U^{Tw} as our recommendation results.

The complexity of the BFS is $O(e \cdot v)$, where e and v are the number of edges and vertexes respectively. Actually, we can utilize some pruning strategies to improve the performance.

3. EXPERIMENTAL EVALUATION

We crawled a dataset from Twitter which includes 14,023 users and 271,695 tweets from 20/04/2013 to 22/05/2013. Those inactive accounts and spammers are filtered to get a dense dataset with 8,923 users and 37,226 tweets. For each user in the dataset, the latest tweet operated by the user is selected as the test data, the other tweets are used as the training data. Another 110,361 randomly selected tweets are collected to train an LDA topic model with 100 topics. As for the parameters in Topic-STG, we set the size of the time window as one day and $\beta = 0.4$ through heuristic learning approach. We employ the Hit Ratio (HR) metric [2] to evaluate the performance of our method, which indicates how many tweets appears in the recommendation list $R(U,t)$ for each user U at time t. Moreover, two baselines are compared to evaluate the effectiveness of the Topic-STG approach [2]. One is the original STG approach, and the other is the approach introduced in [3], which mainly utilizes textual feature to recommend tweets.

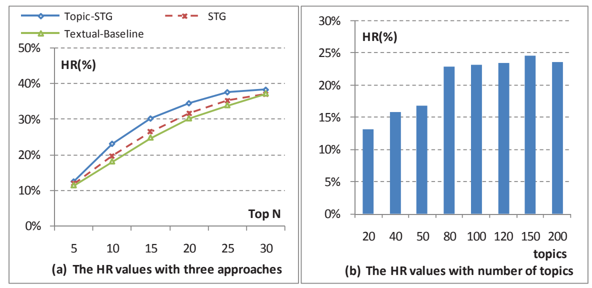

Figure 2: The HR with different approaches and topics.

As illustrated in Figure 2(a). HR is positively correlated with the length N of the recommendation list. The HR values of three approaches are very close when N is set to 30. Moreover, the Topic-STG outperforms the baselines when $N \in [10, 25]$. Also the recommended tweets would get better rank positions in the list generated by Topic-STG. The number of topics would also influence HR. As shown in Figure 2(b), the performance of Topic-STG tends to be stable when $Num(Topic) \in [80, 150]$, which means we can set $Num(Topic)$ to an arbitrary value in [80,150]. However, we should note that the complexity of Topic-STG will increase as we add more topics, so we decide to set $Num(Topic) = 80$. In summary, the new-added topic-nodes play an important role to optimize the recommendation results. In the future, we aim to present a parallel version of Topic-STG to improve its performance.

4. ACKNOWLEDGE

This research was supported by NSFC Grant No.61202408 and CAS 125 Informatization Project XXH12503.

5. REFERENCES

[1] K. Chen, T. Chen, and et al. Collaborative personalized tweet recommendation. In *SIGIR'12 Conference Proceedings*, pages 661–670, 2012.

[2] X. Liang, Y. Quan, and et al. Temporal recommendation on graphs via long-and short-term preference fusion. In *KDD'10 Conference Proceedings*, pages 723–732, 2010.

[3] P. Marco, S. Fabrizio, and et al. Making your interests follow you on twitter. In *CIKM'12 Conference Proceedings*, pages 165–174, 2012.