SCIENCE CHINA
Information Sciences

• HIGHLIGHT •

September 2018, Vol. 61 098105:1–098105:3
https://doi.org/10.1007/s11432-017-9272-y

Who will retweet? A prediction method for social hotspots based on dynamic tensor decomposition

Yunpeng XIAO*, Xiaojuan LI, Shasha YANG & Yanbing LIU

Chongqing Engineering Laboratory of Internet and Information Security,
Chongqing University of Posts and Telecommunications, Chongqing 400065*, China*

Received 29 April 2017/Revised 20 August 2017/Accepted 25 September 2017/Published online 21 May 2018

Citation Xiao Y P, Li X J, Yang S S, et al. Who will retweet? A prediction method for social hotspots based on dynamic tensor decomposition. Sci China Inf Sci, 2018, 61(9): 098105, https://doi.org/10.1007/s11432-017-9272-y

With the development of Web 2.0, the emergence of blogs, forums makes the Internet to maintain substantial amounts of user behavioral data. These data are rich and valuable, and mining useful information from massive data has become a popular research. We can acquire an understanding of the distribution of public behavior based on analysis of user behavior. Therefore, retweet prediction is one of the important research fields in social network analysis, information dissemination and public opinion mining.

In recent years, scholars have conducted extensive research on the information retweeting behavior on online social networks [1]. Although retweeting prediction has achieved significant research results, there are still some challenges. First, inaccurate calculation. There can be a huge number of data in social networks, but useful information is extremely sparse, creating difficulties for the precise calculation of attributes. Second, network dynamics. Online social networks have both static and dynamic properties. In general, nodes and edges of the network change over time, while traditional research has considered networks as static, resulting in what is known as the static problem of dynamic network [2]. Third, topic timeliness. Hotspots in social networks typically undergo evolution in a process of generation, development, and extinction. In hotspots evolution, participation in different stages of the topic is uneven. Regarding uneven data, how to predict user behavior has become the primary difficulty in retweeting prediction research.

Motivated by the above challenges, we propose a prediction method with the advantages of tensor decomposition in feature extraction. We first construct a model of "hot user-alternative user-interactive behavior" based on tensor decomposition in data space and projection to solve the problem of inaccurate calculation caused by data sparsity. Then, Time decay function is introduced to optimize the calculation of user interaction dynamically. Finally, a prediction method that involves user retweeting behavior based on logistic regression is proposed. And it can solve the problem of uneven user interaction regarding hotspots with time slices and discretization.

Proposed method. To solve the above problems, a method of user retweeting prediction is proposed based on user information, behavioral and relational data. The details of this method are introduced in three stages, influence quantification, dynamic modeling, and retweeting prediction, as shown in Figure 1. In the first stage, three factors affecting user behavior are quantified, and two driving mechanisms are defined to represent them. In the second stage, we move to dynamic modeling based on the relationship data to analyze the influence of friends on user interaction. In the third stage, the time slices and discretization are used to deal with the hotspots life cycle, and the user retweeting prediction model is constructed to

* Corresponding author (email: xiaoyp@cqupt.edu.cn)

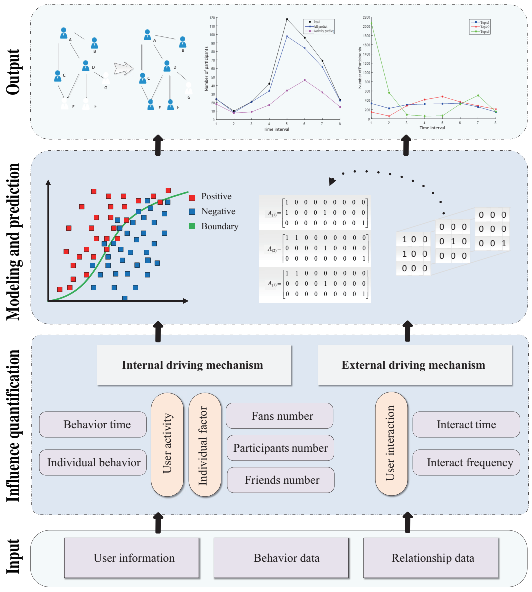

Figure 1 (Color online) Framework for user retweeting prediction.

predict the behavior of the user.

(i) Influence quantification. In social networks, a user will forward the content of hotspots based on the influence of a friend. Sometimes, according to user interests, users will participate in a discussion of hotspots spontaneously. This shows that diverse factors affect user behavior. First, we introduce internal driving mechanism. Whether a user would retweet a topic is usually related to the user own characteristics, such as user activity or interest. The factors related to the user characteristics are defined as factors influenced by internal driving mechanisms. The specific description is given as follows.

(1) User activity. The activity of alternative user v_j is defined as $\mathrm{activity}(v_j) = \alpha \cdot \mathrm{origNum}(v_j) + \beta \cdot \mathrm{retwNum}(v_j)$, where $\mathrm{origNum}(v_j)$ and $\mathrm{retwNum}(v_j)$ are the original and the retweeting numbers, respectively, of alternative users of the latest month before the topic was launched. α, β are adjustable parameters with $\alpha, \beta \in [0, 1]$. And hot users are the users who participate in hotspots time period t, alternative user v_j is one of the fans of hot users in hotspots time period t.

(2) User individual factors. Whether a user would retweet information on a topic is related to his or her inherent attributes. We refer to these inherent attributes as user individual factors. The inherent attributes of alternative user v_j include $\mathrm{CountofFans}(v_j)$: the number of fans of alternative user v_j, $\mathrm{CountofIdol}(v_j)$: the number of friends of alternative user v_j, and $\mathrm{CountofHU}(v_j)$: the number of participant users followed by alternative user v_j.

Second, external driving mechanism is introduced. Since a user participates in the discussion of a topic based on the influence of friends, and the influence of friends can be quantified by the interaction between users, we refer to the factors that affect user behavior as user interaction. As user interaction is quantified based on user relationships, it is defined as the factors influenced by external driving mechanism. The interaction between the hot user and the alternative user is defined as

$$\begin{aligned} &\mathrm{strengthInteract}(u_i, v_j) \\ &= I_{ij} \sum_{k=1}^{K} \sum_{b=1}^{3} \mathrm{interact}(\mathrm{blog}_{kb}), \end{aligned}$$

where

$$I_{ij} = \begin{cases} 1, & u_i \text{ is friend of } v_j, \\ 0, & \text{others}, \end{cases}$$

is an indicator function, blog_{kb} indicates that alternative user v_j has retweeted or commented his friend's (u_i) k-th microblog, K is the total number of microblogs of hot user u_i and

$$\mathrm{interact}(\mathrm{blog}_{kb}) = \begin{cases} 1, & \text{alternative user } v_j \text{ parti-} \\ & \text{cipated in hot user } u_i\text{' s} \\ & k\text{-th blog based behavior} \\ & b\ (\mathrm{retweet}, \mathrm{comment}), \\ 0, & \text{others}. \end{cases}$$

(ii) Dynamic modeling. In social networks, users interact with each other through social relations. However, the intensity of interaction between users will change with the growth of users life experience or changes in interests. That is, interaction between users in social networks has aging characteristics. These characteristics result in the interaction between users having the form of an exponential decline. To quantify the influence of interaction on user behavior dynamically, an exponential decay function is introduced to optimize the calculation of user interaction:

$$\begin{aligned} &\mathrm{strengthInteract}(u_i, v_j) \\ &= I_{ij} \sum_{k=1}^{K} \sum_{b=1}^{3} \mathrm{interact}(\mathrm{blog}_{kb}) \\ &= I_{ij} \sum_{k=1}^{K} \sum_{b=1}^{3} I_{kb} \mathrm{e}^{(-(1+\kappa \cdot (t-t_k)))}, \end{aligned}$$

where

$$I_{kb} = \begin{cases} 1, & \text{alternative users participate in the hot} \\ & \text{user's } k\text{-th blog based behavior } b, \\ 0, & \text{others}, \end{cases}$$

t is the current time of the hotspots, t_k is the time that the hot user published the k-th microblog, and κ is the adjustable parameter.

The interactions between users is calculated based on explicit relationships, but the data sparsity problem of user interaction behavior is not solved. We construct the 3-order tensor $A \in R^{I\times J\times K}$ of "hot user–alternative user–interactive behavior", and we take advantage of the characteristics of tensor decomposition in data space and projection to resolve the data sparsity of user interaction behavior, where I is the dimension of hot users, J is the dimension of alternative users, and K is the dimension of interaction between users. The high-order singular value decomposition (HOSVD) [3] model is used to factorize the user interaction tensor. The 3-order user interaction tensor A is decomposed into three factor matrices $U^{(1)}$, $U^{(2)}$, $U^{(3)}$, and a core tensor S, and the following calculations are required.

(1) The three matrix unfolding operations are defined as follows: $A_{(1)} \in R^{I\times(JK)}$, $A_{(2)} \in R^{J\times(KI)}$, $A_{(3)} \in R^{K\times(IJ)}$, where $A_{(1)}$, $A_{(2)}$ and $A_{(3)}$ are called the 1-mode, 2-mode, and 3-mode matrix unfolding of A, respectively.

(2) Singular value decomposition (SVD) is applied to each $A_{(n)}$, $1 \leqslant n \leqslant 3$. The SVD decomposition of the 1-mode matrix unfolding is as follows: $A_{(1)} = U_{I\times I}\cdot\Sigma_{I\times JK}\cdot V^{\mathrm{T}}_{JK\times JK}$.

The c_1 ($c_1 \prec \min\{I, JK\}$) most important features is used to reconstruct the matrix $\hat{A}_{(1)}$: $\hat{A}_{(1)} = U_{I\times c_1}\cdot\Sigma_{c_1\times c_1}\cdot V^{\mathrm{T}}_{c_1\times JK} = U^{(1)}\cdot\Sigma^{(1)}\cdot V^{\mathrm{T}(1)}$.

Similarly, $\hat{A}_{(2)} = U_{I\times c_2}\cdot\Sigma_{c_2\times c_2}\cdot V^{\mathrm{T}}_{c_2\times JK} = U^{(2)}\cdot\Sigma^{(2)}\cdot V^{\mathrm{T}(2)}$, and $\hat{A}_{(3)} = U_{I\times c_3}\cdot\Sigma_{c_3\times c_3}\cdot V^{\mathrm{T}}_{c_3\times JK} = U^{(3)}\cdot\Sigma^{(3)}\cdot V^{\mathrm{T}(3)}$ can be obtained, where c_2 ($c_2 \prec \min\{J, KI\}$) and c_3 ($c_3 \prec \min\{K, IJ\}$) are the numbers of important features of matrix $A_{(2)}$ and $A_{(3)}$, respectively.

(3) The core tensor is calculated based on the left singular matrix obtained in step (2). $S = A\times_1 U^{(1)\mathrm{T}}\times_2 U^{(2)\mathrm{T}}\times_3 U^{(3)\mathrm{T}}$.

(4) The user interaction tensor is reconstructed based on the core tensor obtained in step (3). $\hat{A} = S\times_1 U^{(1)}\times_2 U^{(2)}\times_3 U^{(3)}$.

The element $\{u, v, b, \mathrm{strength}\}$ of $\hat{A}$ indicates that based on the behavior b, the interaction between hot user u and alternative user v is strength. We can calculate the interaction of hot and alternative users using the results of approximate tensor. $\mathrm{friendInteract}(u, v_j) = \Sigma_{i=1}^{N}\mathrm{strengthInteract}(u_i, v_j)$. The N is the dimension of hot users.

(iii) Retweeting prediction. In addition to the influencing factors of the different driving mechanisms mentioned above, a user retweeting prediction model is proposed based on logistic regression. The function is defined as: $P(r|x) = \frac{1}{1+\mathrm{e}^{-\theta^{\mathrm{T}}x}} = \frac{1}{1+\mathrm{e}^{-(\theta_0+\theta_1x_1+\theta_2x_2+\theta_3x_3+\theta_4x_4+\theta_5x_5)}}$, where x_i refers to influencing factors mentioned above. The parameters θ_i can be updated using gradient descent. It is assumed that when the value of $P(r|x)$ is greater than the threshold ε, the alternative user will retweet the hotspots information in the next period; otherwise, the alternative user will not retweet the topic information.

Experiment. An experiment is carried out with a real dataset, Tencent microblog, to verify the effectiveness of our method. We use Hawkes process [4], learning to rank [5], collaborative filtering [6], and factor graph [7] for comparison with our method. Considering different evaluation metrics (precision, recall and F1-measure), the method we proposed has better predict effect.

Conclusion. We divide users into hot and alternative users based on interactive data of social networks. We try to predict the behavior of alternative users based on basic information, relational data, and previous behavioral data. Then, time decay function is introduced to calculate the interaction strength between users using interaction data. The effects of user interaction are mined by using a model of tensor decomposition. Finally, the behavior of alternative users can be predicted based on logistic regression combined with the factors mentioned.

Acknowledgements This work was supported by National Basic Research Program of China (Grant No. 2013CB329606), National Natural Science Foundation of China (Grant No. 61772098), and Chongqing Science and Technology Commission Project (Grant No. cstc2017jcyjAX0099).

References

1 Li Y, Chen Y H, Liu T. Survey on predicting information propagation in microblogs. J Softw. 2016, 27: 247–263

2 Li C, Feng B Q, Li Y M, et al. Role-based structural evolution and prediction in dynamic networks. J Softw, 2017, 28: 663–675

3 Symeonidis P. ClustHOSVD: item recommendation by combining semantically enhanced tag clustering with tensor HOSVD. IEEE Trans Syst Man Cybern Syst, 2016, 46: 1240–1251

4 Kobayashi R, Lambiotte R. TiDeH: time-dependent Hawkes process for predicting retweet dynamics. In: Proceedings of the 10th International AAAI Conference on WEB and Social Media (ICWSM-16), Cologne, 2016. 191–200

5 Ma H, Qian W, Xia F, et al. Towards modeling popularity of microblogs. Front Comput Sci, 2013, 7: 171–184

6 Pan Y, Cong F, Chen K L, et al. Diffusion-aware personalized social update recommendation. In: Proceedings of the 7th ACM Conference on Recommender Systems, Hong Kong, 2013. 69–76

7 Bian J W, Yang Y, Chua T S. Predicting trending messages and diffusion participants in microblogging network. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, 2014. 537–546