POPULARITY PREDICTION BASED ON INTERACTIONS OF ONLINE CONTENTS

Qingchao Kong[1], Wenji Mao[1], Chunyang Liu[2]

[1]State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2]CNCERT/CC, Beijing 100029, China
{qingchao.kong, wenji.mao}@ia.ac.cn, lcy@isc.org.cn

Abstract: The interactive behavior of Web users often makes some online contents more popular than others. Thus the popularity of online contents can help us understand public interest and attention behind user interactions. Modeling and predicting the popularity of online contents is an important research topic and can facilitate many practical applications in different domains. Previous work on popularity modeling and prediction usually treat each online content separately, and neglect the interaction information between online contents, represented as interaction relations. In this paper, we explore the interaction relations between online contents, specifically competition and cooperation relations, for popularity prediction. We first define the interaction relations between different online contents and propose a method for the calculation of interaction information. We then apply the non-negative matrix factorization (NMF) technique to get a low dimensional representation of interaction features for online contents, which are used by classifiers for popularity prediction. We finally evaluate the proposed approach using two datasets from SinaWeibo (i.e., original tweets and topic hashtags). The experimental results show that interaction features alone can yield relatively good performance, and by incorporating interaction features into traditional feature based methods, our method can further improve popularity prediction results.

Keywords: popularity prediction; interactions of online contents; user interactions; social media analytics

1 Introduction

Social media platforms have become extremely popular in recent years. The online contents generated by Web users are valuable resources for acquiring, understanding and assessing public attitudes, especially those with high popularity attracting user's attention. Meanwhile, there is increasing research interest on modeling and predicting popularity of online contents [1]. Popularity prediction has become an important research topic to understand how user interacts with online contents and how information diffuses via social networks. It can also facilitate many real-world applications in business, security and public management, etc.

Popularity prediction is a challenging task for the following reasons. Firstly, it is hard to model and explore how various factors contribute to the popularity of online contents. Secondly, as the popularity of different online contents co-evolve over time and complex interactions between them make popularity prediction even harder. For the first challenge, previous research work has designed many types of features [2–6] that affect the volume of popularity, and proposed different models [7–10] that describe the dynamic evolution of popularity. However, there is few research dealing with the second challenge.

In this paper, we propose to predict the popularity of online contents based on interactions between them. There are two main types of interaction relations between online contents: competition (i.e., one online content refrains the other with respect to popularity growth) and cooperation (i.e., one online content facilitates the other with respect to popularity growth) [11–13]. The interactions of two online contents are defined using the users who once commented, retweeted or saw them. In this way, the interactions of two online contents can be represented as an interaction matrix and we propose a method to efficiently calculate the interaction matrix. We apply the non-negative matrix factorization (NMF) technique to factorize the interaction matrix to get a low dimensional representation of interaction features. The interaction features are then incorporated into typical feature based methods for popularity prediction. Finally, we use two datasets (i.e., original tweets and topic hashtags) to test the performance of our interaction based popularity prediction method.

Our work has made several contributions. We make the first attempt to model the interactions between online contents for popularity prediction and propose a method to compute the degrees of interaction relations, represented as an interaction matrix. To extract information from the interaction matrix, we apply the NMF technique to get a low dimensional representation of the interaction features. We collect two datasets from SinaWeibo to evaluate the effectiveness of our proposed popularity prediction method.

978-1-5090-1256-5/16/$31.00 ©2016 IEEE

2 Related work

2.1 Popularity measures

The popularity of online contents reflects the degree of user attention distributed over them, which is usually measured by some quantitative criteria. For example, the popularity of tweets can be measured by the number of retweets[3, 8] and the popularity of YouTube videos can be measured by the number of views [6, 14]. However, there are other ways to measure the popularity of online contents by combining different criteria. For example, the popularity of videos can be measured using the ratio of the view count and upload time [15], and the popularity of Digg shares can be measured by the difference between positive and negative votes [16]. In this paper, we measure the popularity of tweets using the number of retweets and the popularity of hashtags using the number of tweets containing that hashtag.

2.2 Popularity prediction methods

Recent work in popularity prediction falls into two major categories depending on the prediction methods, that is, feature-based discriminative methods and generative model based methods. For the feature-based discriminative methods [2–6], it is important to design and test different kinds of features that affect the popularity and then apply classification or regression algorithms depending on specific problem definition. For example, the SH model [2] assumes there is a linear correlation between the (log-transferred) early stage and future popularity. To predict the size of Facebook photo cascades, Cheng *et al.* [5] propose five types features, including content features, original poster features, resharer features, structural features and temporal features. Compared to the above static factor based method which do not consider dynamic popularity evolution information, some work [14, 17] propose to use dynamic factors to predict the volume of popularity and the popularity trend. Feature-based discriminative methods are very popular in popularity prediction, because it is convenient to incorporate various factors that influence popularity, and can produce good prediction results if the features are well-designed. However, this method offers little understanding of how the popularity evolves.

Generative model based methods [7–9] usually predict popularity by describing the generative process of the dynamic popularity evolution. To predict the number of votes of Digg shares, Lerman and Hogg [7] propose a stochastic model of the user browsing behavior. Generative methods usually need a large number of training samples to get good prediction results and suffer from a lack of flexibility of adding more influencing factors of popularity growth.

2.3 Interactions of online contents

Previous work treat each online content separately and models the popularity itself, while neglecting the fact that the popularity evolutions of multiple online contents co-exist. There is currently no previous work which uses the interaction information to predict popularity. Recent research on the interactions of online contents have focused on tweets [18], hashtags [12, 19], Amazon online products [11] and Douban book & movie [13]. To model the conditional probability of the user retweeting behavior, Myers and Leskovec [18] use the interaction of tweets as an offset with respect to the prior retweet probability. Using “entropy” to measure the user attention, Weng *et al.* [19] have found that user attentions are limited and memes are competitive against each other based on their agent-based model.

In this paper, we first define and calculate the interaction information between different online contents and represent the interaction relations as an interaction matrix. We then propose to apply the NMF technique to extract interaction features from the interaction matrix and build classification models using the interaction features for popularity prediction. We finally evaluate our proposed method using two SinaWeibo datasets.

3 Proposed method

Given an online content, such as a tweet or a hashtag, denote its popularity at time t as $C(t)$. Instead of predicting the exact volume of popularity, we divide the popularity volume into several intervals defined using thresholds, such as (0, θ_1], (θ_1, θ_2], …. Let t_r be the prediction time after publication and t_t be the target time ($t_t > t_r$). The popularity prediction problem can be defined as a classification task: given the relevant information at time t_r as input, predict which interval $C(t_t)$ will belong to.

3.1 Calculation of interaction matrix

To calculate the interactions of two online contents, we adapt the method proposed in [13], which determines the interactions of two online products based on global adoption probability (GAP). Compared to other methods of interaction calculation [11, 12], the GAP-based method has more interpretability for online contents in social media.

For our problem, the GAP of two online contents is defined as a 4-element tuple: $P = (p_{A|\varnothing}, p_{A|B}, p_{B|\varnothing}, p_{B|A}) \in [0, 1]^4$, where A and B are any two online contents. Here $p_{A|\varnothing}$ is the probability that a user retweets A if the user saw someone retweet A in her timeline (i.e., the tweets posted by a user’s followees) and the user has not retweeted B (B is a retweet in the user’s timeline); $p_{A|B}$ is the probability that a user retweets A if she has retweeted B. $p_{B|\varnothing}$ and $p_{B|A}$ follow similar definitions as $p_{A|\varnothing}$ and $p_{A|B}$. Based on the above definition, $p_{A|\varnothing}$ can be calculated as follows:

$$p_{A|\varnothing} = \frac{\left| R_A \setminus R_{B \prec_{retweet} A} \right|}{\left| I_A \setminus R_{B \prec_{see} A} \right|} \qquad (1)$$

where R_A is the set of users who have retweeted A, I_A is the set of users who have saw A ($R_A \subseteq I_A$) in the user's timeline, $R_{B \prec_{retweet} A}$ is the set of users who have retweeted B before retweeting A and $R_{B \prec_{see} A}$ is the set of users who have retweeted B before seeing A. Similarly, $p_{A|B}$ can be calculated as follows:

$$p_{A|B} = \frac{\left|R_{B \prec_{retweet} A}\right|}{\left|R_{B \prec_{see} A}\right|} \tag{2}$$

Denote the interaction matrix as **M**, whose shape is $N \times N$ and N is the total number of online contents. Based on the above definitions, each entry of **M** can be calculated as:

$$\mathbf{M}_{A,B} = p_{A|B} - p_{A|\varnothing} \tag{3}$$

B is cooperative towards A if $\mathbf{M}_{A,B} > 0$, while B is competitive towards A if $\mathbf{M}_{A,B} < 0$. If $\mathbf{M}_{A,B} = 0$, there is no interaction between A and B. Obviously, $\mathbf{M}_{A,B}$ is in range [0, 1] and $|\mathbf{M}_{A,B}|$ can measure the strength of the competition or cooperation interaction. We can also see that: 1) P is not specific to any particular user; 2) the interaction between any two online contents is asymmetric: for instance, if A is cooperative towards B, it does not follow that B is cooperative towards A.

We propose a method to calculate the above interaction matrix. As the size of **M** can be huge when there are a lot of online contents and thus calculating each entry of **M** one by one will be very time-consuming. To efficiently calculate the interaction matrix, we count the number of users in each user set, such as $R_{B \prec_{retweet} A}$, in Eq. (1) and Eq. (2) by checking each user in the whole user set. For each user u, instead of considering every pair of online contents, we only calculate the interaction relations of online contents which are related to u (i.e., u has seen or retweeted).

Taking the original tweets for example, we next describe the calculation of interaction matrix in detail. First we give some notations. We define four matrices of size $N \times N$: $\mathbf{S}^{top}$, $\mathbf{S}^{bottom}$, $\mathbf{T}^{top}$ and $\mathbf{T}^{bottom}$. Let $\mathbf{S}^{top}_{A,B}$ and $\mathbf{S}^{bottom}_{A,B}$ record the number of users in $R_A \setminus R_{B \prec_{retweet} A}$ and $I_A \setminus R_{B \prec_{see} A}$, $\mathbf{T}^{top}_{A,B}$ and $\mathbf{T}^{bottom}_{A,B}$ record the number of users in $R_{B \prec_{retweet} A}$ and $R_{B \prec_{see} A}$. Thus according to Eq. (3), we can calculate each entry of the interaction matrix as follows:

$$\mathbf{M}_{A,B} = \frac{\mathbf{T}^{top}_{A,B}}{\mathbf{T}^{bottom}_{A,B}} - \frac{\mathbf{S}^{top}_{A,B}}{\mathbf{S}^{bottom}_{A,B}} \tag{4}$$

Now we show how to calculate $\mathbf{S}^{top}_{A,B}$, $\mathbf{S}^{bottom}_{A,B}$, $\mathbf{T}^{top}_{A,B}$ and $\mathbf{T}^{bottom}_{A,B}$, respectively. We construct two retweet sets for user u and her followees: denote Q_u as the retweet set posted by user u and Q^u_F as the retweet set posted by u's followees. For any two retweets i and j ($i \in Q_u$ and $j \in Q^u_F$) with A and B being the corresponding original tweets for i and j, if the publication time of j is earlier than i, than increase $\mathbf{T}^{bottom}_{A,B}$ by 1, otherwise increase $\mathbf{S}^{bottom}_{A,B}$ by 1. Similarly, for any two retweets i and j ($i, j \in Q^u_F$) with A and B being the corresponding original tweets for i and j, if the publication time of j is earlier than i, than increase $\mathbf{T}^{top}_{A,B}$ by 1, otherwise increase $\mathbf{S}^{top}_{A,B}$ by 1.

3.2 Extraction of interaction features for popularity prediction

To extract interaction features from the interaction matrix **M**, we apply the widely used NMF technique to get a low dimensional representation for each online content. Specifically, **M** is first transformed to be non-negative, such as $(\mathbf{M}_{i,j}+1)/2$. Through NMF, **M** is factorized into two matrices **U** and **V**, whose shapes are $N \times r$ and $r \times N$ respectively, where r is the target number of dimensions. Thus the columns of **V** are the interaction features of each online content.

We incorporate the acquired interaction features into typical feature based methods for popularity prediction. We first design multiple features, including the numeric features (e.g., number of retweets and users) and structural features (e.g., average path length, density and max depth of the retweet tree, average path length, density and mean degree of the author reply network). We then propose two different ways to incorporate interaction features. One way is to extend the features of the above methods using interaction features (denoted as "feature-extend"). The other is to use a balancing parameter α to combine the prediction results of two methods (denoted as "combination"), that is, the result of the method only using interaction features (i.e., $score_{interaction}$), and the result of the typical feature based popularity prediction method (i.e., $score_{typical}$):

$$score = \alpha \times score_{interaction} + (1-\alpha) \times score_{typical} \tag{5}$$

We apply various classifiers, such logistic regression and SVM, in the above methods.

4 Experiments

4.1 Datasets

To evaluate our proposed method, we construct two datasets of tweets and hashtags using a publicly available SinaWeibo dataset [20]. We first find all original tweets along with their retweets with text content, user information and publication time, then extract hashtags from all the tweets, for example, "#Xiaomi#" and "#Wenzhou Train Crash#", resulting about 230,000 original tweets and 80,000 hashtags. To ensure enough information when performing prediction, we restrict that there are at least 50 retweets before prediction time. After the above filtering, there are 5000+ original tweets and

3000+ hashtags, each with millions of retweets in our dataset.

4.2 Popularity prediction results with interaction features

In the first experiment, we construct classifiers to evaluate the effectiveness of the interaction feature based method for popularity prediction. For the two datasets, the average lifecycles for the tweet and hashtag dataset are 8 hours and 9 days, respectively. For the tweet dataset, the prediction time is set to 1 hour and the target time is 3 hours. For the hashtag dataset, the prediction time is set to 3 days and the target time is 7 days.

We divide the popularity value at target time into three intervals defined by two threshold parameters θ_1 and θ_2, that is, (0, θ_1], (θ_1, θ_2], (θ_2, ∞]. At target times, the distributions of the popularity values for the two datasets are shown in Figure 1. For the tweet dataset, θ_1 and θ_2 are set to 70 and 140 respectively. For the hashtag dataset, θ_1 and θ_2 are 110 and 300 respectively. Another consideration of choosing thresholds is to ensure that the number of samples in each interval is approximately equal.

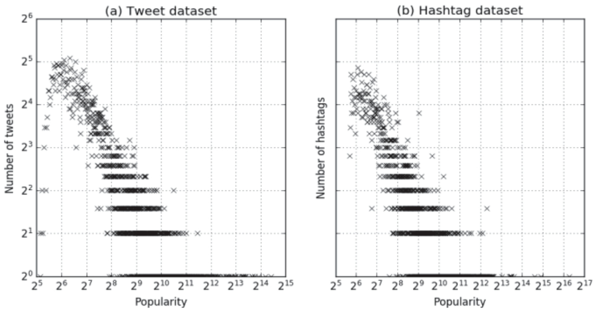

Figure 1 Popularity distributions for the tweet and hashtag dataset

We choose five classifiers for evaluation, logistic regression (LR), SVM with RBF kernel, decision tree (DT), kNN (k = 5) and random forest (RF). For each dataset, we apply 5-fold cross validation and report the average F1 score as the evaluation metric. See Table I for the prediction results.

Table I Prediction results using interaction features only

Classifiers	LR	SVM	DT	kNN	RF
tweet	**0.59**	0.47	0.48	0.54	0.36
hashtag	0.48	**0.53**	0.46	0.45	0.37

As shown in Table I, using only the interaction features, the F1 score can be as high as 0.59 for the tweet dataset and 0.53 for the hashtag dataset. Almost all classifiers achieve relatively good prediction results, especially LR and SVM.

The target number of dimensions r is an important parameter for NMF and we next evaluate the influence of r on the prediction results. In this experiment, we use the LR classifier and report the prediction performance with varying r. See Figure 2 for the prediction results.

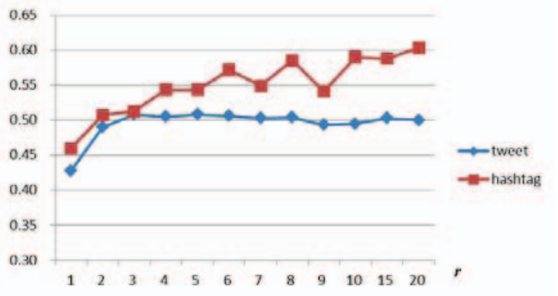

Figure 2 Prediction Results with Varying r

In Figure 2, as r increases, the F1 score increases for both datasets, however, the increase rate of F1 score becomes lower. This is expected because larger r means more interaction information is retained after NMF. At the same time, the computation time of NMF is also rapidly increasing with larger r, so there is a tradeoff between high F1 score and low NMF computation time.

4.3 Popularity prediction results incorporating interaction features

Our second experiment compares the prediction results of typical feature based methods (i.e., baselines) and two methods incorporating interaction features (i.e., “feature-extend” and “combination”). As the static factor based (denoted as “SF”) and dynamic factor based (denoted as “DF”) methods are two main approaches for feature based popularity prediction, we choose them as the baseline methods. Specifically, for the SF method, we use the features proposed in section 0 at prediction time. For the DF method, we record the values of each feature once every fixed time interval to form the feature vector, and the final feature vector is obtained by joining the feature vectors of all proposed features. We use logistic regression as the classifier as it performs best in the first experiment. The balancing parameter α is set to 0.5. The target number of dimensions r is 5 for the tweet dataset and 20 for the hashtag dataset. Other parameter settings, evaluation metric and experimental process are the same as the first experiment in section 0. Table II show the prediction results after incorporating interaction features.

Table II Prediction results when incorporating interaction features

Dataset	tweet		hashtag	
Method	SF	DF	SF	DF
baseline	0.52	0.31	0.51	0.37
feature-extend	0.48	0.32	0.52	0.36
combination	**0.64**	**0.48**	**0.69**	**0.61**

As shown in Table II, the SF method generally performs better than the DF method in all the cases. On possible reason is that, historic popularity dynamics bring more noise for this specific prediction task. Both the SF and DF methods perform worse than the method only using interaction features. We can also see that the combination method performs better than the feature-extend method in all cases. The combination method greatly improves the prediction results compared to the baseline methods and the method only using interaction features. This indicates that the baseline methods and interaction feature based method are good at predicting different types of online contents, and the combination method can make more meaningful prediction decisions by taking advantages of these two methods.

5 Conclusions

In this paper, we model the interactions of online contents in social media and exploit the interaction information for popularity prediction problem. Based on the interaction relations between online contents, we develop the interaction feature based method and incorporate it into typical popularity prediction methods. We first define the interaction relations of online contents and propose a method to calculate the interaction matrix. We then apply the NMF technique to extract interaction features from the interaction matrix, and develop two popularity prediction methods combining interaction features. The experimental results show that using only the interaction features can achieve relatively good performance. Further experiment shows that the performance of previous typical feature based popularity prediction methods can be improved by incorporating interaction features.

Acknowledgements

This work is supported by NSFC grants #61175040 and #U1435221, and grant #2015A069.

References

[1] A. Tatar, M. D. de Amorim, S. Fdida, and P. Antoniadis, “A survey on predicting the popularity of web content,” *J. Internet Serv. Appl.*, vol. 5, no. 1, pp. 1–20, Dec. 2014.

[2] G. Szabo and B. A. Huberman, “Predicting the popularity of online content,” *Commun ACM*, vol. 53, no. 8, pp. 80–88, Aug. 2010.

[3] L. Hong, O. Dan, and B. D. Davison, “Predicting popular messages in Twitter,” in *Proceedings of the 20th International Conference Companion on World Wide Web*, New York, NY, USA, 2011, pp. 57–58.

[4] Z. Ma, A. Sun, and G. Cong, “On predicting the popularity of newly emerging hashtags in Twitter,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. 7, pp. 1399–1410, Jul. 2013.

[5] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, “Can cascades be predicted?,” in *Proceedings of the 23rd International Conference on World Wide Web*, New York, NY, USA, 2014, pp. 925–936.

[6] D. Vallet, S. Berkovsky, S. Ardon, A. Mahanti, and M. A. Kafaar, “Characterizing and predicting viral-and-popular video content,” in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, New York, NY, USA, 2015, pp. 1591–1600.

[7] K. Lerman and T. Hogg, “Using a model of social dynamics to predict popularity of news,” in *Proceedings of the 19th International Conference on World Wide Web*, New York, NY, USA, 2010, pp. 621–630.

[8] T. Zaman, E. B. Fox, and E. T. Bradlow, “A bayesian approach for predicting the popularity of tweets,” *Ann. Appl. Stat.*, vol. 8, no. 3, pp. 1583–1611, Sep. 2014.

[9] J. G. Lee, S. Moon, and K. Salamatian, “An approach to model and predict the popularity of online contents with explanatory factors,” in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, 2010, vol. 1, pp. 623–630.

[10] Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, and J. Leskovec, “SEISMIC: A self-exciting point process model for predicting tweet popularity,” in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2015, pp. 1513–1522.

[11] M. Gomez Rodriguez, J. Leskovec, and A. Krause, “Inferring networks of diffusion and influence,” in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2010, pp. 1019–1028.

[12] M. Coscia, “Competition and success in the meme pool: a case study on quickmeme.com,” in *Proceedings of the 7th International Conference on Weblogs and Social Media*, 2013.

[13] W. Lu, W. Chen, and L. V. S. Lakshmanan, “From competition to complementarity: comparative influence diffusion and maximization,” *Proc. VLDB Endow.*, vol. 9, no. 2, pp. 60–71, Oct. 2015.

[14] H. Pinto, J. M. Almeida, and M. A. Gonçalves, “Using early view patterns to predict the popularity of youtube videos,” in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2013, pp. 365–374.

[15] A. Khosla, A. Sarma, and R. Hamid, “What makes an image popular?,” in *Proceedings of the 23rd International Conference on World Wide Web companion*, Seoul, Korea, 2014.

[16] P. Yin, P. Luo, M. Wang, and W.-C. Lee, “A straw shows which way the wind blows: ranking potentially popular items from early votes,” in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, New York, NY, USA, 2012, pp. 623–632.

[17] Q. Kong, W. Mao, D. Zeng, and L. Wang, “Predicting popularity of forum threads for public events security,” in *Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference*, Hague, Netherlands, 2014, pp. 99–106.

[18] S. A. Myers and J. Leskovec, “Clash of the contagions: cooperation and competition in information diffusion,” in *Proceedings of 12th IEEE International Conference on Data Mining*, 2012, pp. 539–548.

[19] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, “Competition among memes in a world with limited attention,” *Sci. Rep.*, vol. 2, no. 335, Mar. 2012.

[20] J. Zhang, B. Liu, J. Tang, T. Chen, and J. Li, “Social influence locality for modeling retweeting behaviors,” in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, Beijing, China, 2013, pp. 2761–2767.