MICROBLOGGING CONTENT PROPAGATION MODELING USING VIRALITY AND SUSCEPTIBILITY ANALYSIS

Dr. G.Anandharaj, Ms. S.R.Sivadharani
1Associate Professor & Head, Department of Computer Science and Applications, Adhiparasakthi College of Arts and Science (Autonomous), Kalavai, Vellore -632506,India ,
2Research Scholar, Department of Computer Science and Applications, Adhiparasakthi College of Arts and Science (Autonomous), Kalavai, Vellore -632506,India,

Abstract : Due to the rise of sites like Facebook, Twitter, and Weibo, political information is more and more frequently encountered in a social context: even stories published by mainstream media sites are often encountered by users after having been shared by others. Clearly, this social context can influence how information is interpreted and re-shared. When a microblogging user adopts some content propagated to her, they can attribute that to three behavioral factors, namely, topic virality, user virality and user susceptibility. Topic virality measures the degree to which a topic attracts propagations by users. User virality and susceptibility refer to the ability of a user to propagate content to other users, and the propensity of a user adopting content propagated to her, respectively. In this work, we study the problem of mining these behavioral factors specific to topics from microblogging content propagation data. It first construct a three dimensional tensor for representing the propagation instances. Then propose a tensor factorization framework to simultaneously derive the three sets of behavioral factors. Based on this framework, it develop a numerical factorization model and another probabilistic factorization variant. The work also develop an efficient algorithm for the models’ parameters learning.

I.  INTRODUCTION

Content propagates among microblogging users through their follow links, from followees to followers. The former are the senders, and the latter are known as the receivers. A receiver may adopt the content exposed to her based on a number of factors, namely the: (a) virality of the sender, (b) susceptibility of the receiver, (c) virality of the content topic and (d) strength of relationships between sender and receiver. User virality refers to the ability of a user in getting others to propagate her content, while user susceptibility refers to the tendency of a user to adopt her followees content. Topic virality refers to the tendency of a topic in getting propagated. Since microblogging has been shown rather an information source than a social networking service. We assume in this paper that most relationships among users in a microblogging site are casual and identical in strength. We therefore focus on modeling the user and content factors that drive content factors that drive content propagation without considering the pairwise relationships among users.

The modeling of the virality and susceptibility factors has many important applications. In advertisement and marketing, companies may hire viral users to propagate positive content about their products, or to the advertisement with viral content so as to maximize their reach. Similarly, politicians may leverage on viral users to disseminate their messages widely or to conduct campaigning. Also one may detect events by tracking those mentioned by non susceptible users and detect rumors based on susceptible users interactions with the content.
Inter-relationship among user virality, user susceptibility and content virality. Prior empirical research have suggested there are inter-dependencies among the three factors. Hence, the measurement of a user’s susceptibility requires the virality of topics of tweets propagated to her and the virality of users propagating the tweets. The same can be said about the measurement of user virality and topic virality. Existing models however measure the three behavioral factors separately. That is, they measure a user’s virality by aggregating propagations on her content without considering the virality of content and susceptibility of the receivers. Again, similar remarks are applicable to existing works that measure users’ susceptibility and topics’ virality. Such simplistic approaches may lead to less accurate modeling results.

Consider the example scenario of propagation in Twitter. In this example, content are tweets \((t_1, ..., t_3)\), and the tweets are propagated from the authors \((u_1, u_2 \text{ and } u_3)\) to their followers \((v_1, v_2, \text{and } v_3)\) when the followers retweet (forward). Without considering the followers’ susceptibility one may conclude that \(u_3\) is more viral in propagating tweets than \(u_1\) since the former gets more retweets. However, \(v_3\) is observed to be much more susceptible than \(v_1\) and \(v_2\) since \(v_3\) retweets all the followees’ tweets. The same is not observed on \(v_1\) and \(v_2\) since \(v_3\) is susceptible to all the followers. Moreover, \(u_3\) receives retweets mostly by \(v_3\) while all \(u_1’s\) tweets are retweeted by all the followers.
II. RELATED WORKS

In this section, we review prior works on analyzing content virality, user virality and susceptibility in online social networks that are closely related to ours. Also, we review works on retweet analysis since retweet is the most common action that generates content propagation in microblogging sites.

User virality and susceptibility. In many works, a user’s virality is simply measured by FanOut, i.e., the average number of friends the user diffuses item(s). Other existing works borrow user influence as a proxy for user virality. On the other hand, prior works has measured a user’s susceptibility by FanIn, i.e., the number or fraction of items the user adopts once she is exposed to them. Early studies of user influence and susceptibility in online social networks focus on examining the existence of these factors and distinguishing them from other related factors.

Content virality. In some previous works, content virality has been simply measured by popularity, and viral coefficient. Popularity can be defined differently including the number of users adopting the item, the number of views, likes, comments, and shares, and the number of downloads and citations. Viral coefficient is defined by the average number of new adopters generated by each existing adopter. For microblogging data, the viral coefficient of a tweet is the same as the retweet count of the tweet. Previous works on analyzing item virality include (a) empirical works on examining effects of different factors on item virality; and (b) works on predicting item virality.

Our work here overcomes this shortcoming by showing how the behavioral factors can be jointly derived from the data traces of user-information content interactions in the propagation process. However, the former does not measure the factors specific to topics, while the latter does not model virality of topics. Aggregating the factors across topics over simplifies the problem and would result in less than optimal models. Also, without topics’ virality, dealing with future content items requires more side information which is not always available. Our work, on the other hand, aims to model topic virality.
specific virality and susceptibility factors, which would can be easily used to predict propagation of future content.

**Retweet analysis:**
Existing works on retweet analysis include: (a) empirical works on studying the effects of different factors on tweets’ retweetability, and (b) works on modeling retweet actions. In the former category, researchers have examined the correlation between retweetability with authority features, social and emotional features and content and linguistic features. Most of works in the latter category formulated the retweet modeling problem as a tweet recommendation task in which retweets are considered as positive user feedback. While these works are reported to achieve high performance, they suffer from a few shortcomings. First, they use features that require a large dataset covering user activities over a long time period (e.g., users’ tweets, retweets, and interactions) or even no longer available system features of Twitter (e.g., the retweet traces of tweets). In contrast, our models only requires the retweet data and considers new topic and user factors. Second, they can only perform in-matrix recommendation: only tweets in the training dataset can be recommended to users. Hence, they cannot be applied to predict retweets for the future tweets like our models.

**III. EMPIRICAL FINDINGS ABOUT CONTENT PROPAGATION**
We conduct an empirical analysis of content propagation on a large dataset collected from Twitter. The methodology used to derive content propagation behaviour and topics will be presented. The study will show that virality and susceptibility contributing to content propagation should be modeled at topic level. In microblogging, retweet is the most common form of content propagation. We therefore use retweet to define propagation in the remaining part of this section. That is, each original tweet is considered as a content item, and we say user v is exposed to m if (a) v follows m’s author, and (b) v receives and reads m. Lastly, m is said to be propagated from its author u to v if (i) v follows u and (ii) v retweets m. We do not consider in this work the subsequent retweets of m by v’s followers and by followers of the followers, since: (1) only less than 5% of retweets are subsequent retweets, and (2), as aforementioned, Twitter no longer provides subsequent retweets’ trace.

**Methodology**
Both content propagation and content topics are usually not observable when the microblogging data are crawled. We have therefore devise the methodological steps to infer them as described below.

**Determining user-tweet exposure.** In Twitter, the latest tweets posted by a user’s followees always appear at the top of her timeline. Hence, many tweets may have been missed by the user who does not monitor the timeline closely, and such tweets would never be retweeted. As Twitter API does not reveals the tweets seen by users, we define a time window in which the received tweets will be read. We know that every retweet by a user v comes with a corresponding tweet m that v must have read. We first count the number of other tweets v receives within the duration from the time v receives m to the time v retweets m. Based on this count we estimate N the number of tweets a user may read on her timeline whenever she performs a retweet. We found that N follows a long tail distribution. For more than 90% of the times, N is not larger than 200. We therefore determine that a user v receives and actually reads through the tweet m, i.e., v’s expose to m, if and only if m is among last 200 tweets posted by v’s followees up to the time v makes a retweet. Otherwise, v is considered not exposed to the tweet m.
**Topic discovery.** We applied TwitterLDA model to automatically identify the topics of every original tweet. This step is conducted for every time window, independently from each other. We first remove all retweets and non-informative tweets, e.g., tweets generated by third party applications like Foursquare or Instagram. We then remove from remaining tweets all stop words, slang word, and non-English phrases. Next, we iteratively filter away words, tweets, and users such that: each word must appear in at least remaining tweets, each tweet contains at least 3 remaining words, and each user has at least 20 remaining tweets. These minimum thresholds are designed to ensure that for each user, tweet, and word, we have enough observations to learn the latent topics accurately.

Based on the learnt topics and topic distributions of users, we compute the topic distribution of every remaining tweet m with author u as follows.

\[
D_{m,k} \propto \theta_{u,k} \cdot \prod_{w \in m} \phi_{k,w}
\]

where \(D_{m,k}\) and \(\theta_{u,k}\) is the probability of topic k of tweet m and user u respectively; and \(\phi_{k,w}\) is probability of word w given topic k.

Due to the filtering steps above, many tweets are filtered away, and there is only 15% of tweets are topically modelled by the TwitterLDA model. We therefore expanded the set of modeled tweets as follows. First, we include in the set all the tweets of filtered away users that contain at least 3 remaining words. Then, we compute the topic distribution of each of these tweets using their (remaining) words and the learnt topics, assuming the tweet’s author u (who is filtered away) has uniform topic distribution (i.e., \(\theta_{u,k} = 1/K\)).

Moreover, as each tweet is a short document, we are not interested in tweets that cover many topics. Instead, we only consider tweets having some dominating topics. To do this, we filter away tweets whose sum of top K topic probabilities is less than 0.95.

**Topics of tweets and retweets at network level**

To compare the likelihood of getting retweeted across topics, in each time window and for each topic k, we derive the relative popularities of topic k among the set of all original tweets and the bag of retweets in the time window. The former is called generating popularity of the topic k, denoted by \(G_k\). and the later is called propagating popularity, denoted by \(P_k\). The two popularities are defined based as follows.

\[
G_k = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} D_{m,k} \quad (2)
\]

\[
P_k = \frac{1}{\sum_{m \in \mathcal{M}} p_m \sum_{m \in \mathcal{M}} [p_m \cdot D_{m,k}]} \quad (3)
\]

where, in each time window, \(\mathcal{M}\) is the set of all content items, and \(p_m\) is number of time m is propagated successfully. Since we use tweets and retweets to define content and propagation respectively, \(\mathcal{M}\) is the set of original tweets while \(p_m\) is number of m’s retweets.
To examine the difference between the two popularities of topics, we use their Pearson rank correlation coefficient PRCC, defined as below

$$PRCC = \frac{\sum_{k=1}^{K} (r_{G}(k) - \bar{r})(r_{P}(k) - \bar{r})}{\sqrt{\sum_{k=1}^{K} (r_{G}(k) - \bar{r})^2} \sqrt{\sum_{k=1}^{K} (r_{P}(k) - \bar{r})^2}}$$

Where \( r_{G}(k) \) is the rank of generating popularity of topic \( k \) (i.e., the rank of \( G_k \) in \( G_1, \ldots, G_k \)), \( r_{P}(k) \) is the case of propagating popularity of topic \( k \).

**4.3.5 Topics of tweets and retweets at individual level**

In each time window, to compare the likelihood of user \( u \) getting retweeted for different topics, we compare the relative popularities of each topic \( k \) in the set of tweets posted by \( u \), and in the bag-of-retweets that \( u \) got. The former is called sender-specific generating popularity of \( u \) for topic \( k \), while the latter one is called sender-specific propagating popularity of \( u \) for topic \( k \). The two popularities are denoted by

$$G_{u,k} = \frac{1}{|\mathcal{M}_u|} \sum_{m \in \mathcal{M}_u} D_{m,k}$$

$$P_{u,k} = \frac{1}{\sum_{m \in \mathcal{M}_u} p_m} \sum_{m \in \mathcal{M}_u} [p_m \cdot D_{m,k}]$$

Similarly, we compute Pearson rank correlation coefficients between \( G_{u,k} \) and \( P_{u,k} \) for each user \( u \), and between \( P_{u1,k} \) and \( P_{u2,k} \) for each pair of different users \( u1 \) and \( u2 \).
On the receiver side. Similarly, in each time window, to compare the likelihood of retweeting by user \( v \) for different topics, we compute the relative popularities of each topic \( k \) in the set of tweets \( v \) (exposed to and in the set of tweets \( v \) retweeted. The former popularity is called receiver-specific exposing popularity of user \( v \) for the topic \( k \), and the latter is called receiver-specific adopting popularity of user \( v \) for topic \( k \). The two popularities are denoted by \( E_{v,k} \) and \( A_{v,k} \) respectively, and are defined below.

\[
E_{v,k} = \frac{1}{|\mathcal{M}_v^e|} \sum_{m \in \mathcal{M}_v^e} D_{m,k} \\
A_{v,k} = \frac{1}{|\mathcal{M}_v^a|} \sum_{m \in \mathcal{M}_v^a} D_{m,k}
\]

Where \( \mathcal{M}_v^e \) and \( \mathcal{M}_v^a \) are the set of content items \( v \) has exposed to and the set of content items \( v \) has adopted due to propagation, respectively. In this section, \( \mathcal{M}_v^e \) is consist of original tweets \( v \) has received and read, while \( \mathcal{M}_v^a \) is the set of retweets by \( v \).

IV. DATA FLOW DIAGRAM

[Diagram of user interactions and data flow]

Admin

User

Check

Login

Registration

View User Details

View User Profile Details

View User Share Image Details

New Tweet

View Tweets

Identify Viral User

Search Friends

Identify Susceptible User

Follow Friends

Share Tweets
V. SCREEN SHOTS

Default

![Default Screen Shot]

Admin login

![Admin Login Screen Shot]
Admin View User Details:

All Tweet:
Total Tweet:

![Image of Micro Blogging page]

User Profile:

![Image of User Profile page]
Follower:

User Search Details:
Create Tweet:

View Tweet:
VI. CONCLUSION

We study user and content factors underlying content propagation in microblogging. Motivated by an empirical study showing that different topics have different likelihood of getting propagated at both network and individual levels, we propose to model these factors to topic level. We develop V2S, a tensor factorization-based framework and its associated models, to learn topic-specific user virality and susceptibility, and topic virality from content propagation data. Our experiments on a large Twitter dataset show that the proposed V2S-based models outperform baseline models significantly in propagation prediction. Our experiments on synthetic databases also show that our proposed models outperform all the other baseline methods in learning the topic-specific factors.

VII. FUTURE ENHANCEMENT

In the future, we want to relax the assumption on the tie identical strength by incorporating heterogeneous pair-wise influence among users in modeling the propagation. We would also like to incorporate more fine-grained factors affecting the propagation. These factors include users’ positions in the network, linguistic features in content, and emotion factors of users.

VIII. REFERENCES