DOI: http://dx.doi.org/10.14236/ewic/FDIA2017.6

Detection of Trending Topic Communities: Bridging Content Creators and Distributors

Lorena Recalde
Web Research Group, DTIC
Universitat Pompeu Fabra
Barcelona, 08018 Spain
lorena.recalde@upf.edu

The rise of a trending topic on Twitter leads to the temporal emergence of a set of users currently interested in that topic. Given the temporary nature of the links between these users (more users as well as more interactions among them appear while the topic evolves), being able to dynamically identify communities of users related to this trending topic would allow for a rapid spread of information. In this paper, we tackle this challenge, by identifying coherent topic-dependent user groups, linking those who generate the content (*creators*) and those who spread this content, *e.g.*, by retweeting/reposting it (*distributors*). This is a novel problem on group-to-group interactions in the context of recommender systems and user modeling. Analysis on real-world Twitter data compare our proposal with a baseline approach that considers the retweeting activity, and validate it with standard metrics. Results show the effectiveness of our approach to identify communities interested in a topic where each includes content creators and content distributors, facilitating users' interactions and the spread of new information.

Trending topics; community detection; content creators; content distributors; Twitter.

1. INTRODUCTION

In the context of Online Social Networks (OSNs), when a noun, a phrase, or a hashtag is used with a high frequency in posts and discussions, it is said to be associated to a *trending topic*. With the rise of a trending topic, a set of users interested in it also emerges. However, multiple points of view might be associated to it (*e.g.*, the `#donaldtrump` hashtag, related to the US president, is used by people with opposing political views). Being able to manage these users and detect communities related to a given trending topic is a problem of central interest in social recommender systems. Indeed, having a community of users who are linked and have the same interests would allow a system to generate suggestions (i) for individual users, by providing recommendations of content related to the trending topic and generated by the other users in the community; or (ii) for the community as a whole, by providing group recommendations with new content related to the trending topic[1]. At the same time, the problem is challenging, since trending topics are characterized by their temporary nature and evolve quickly. Thus, an approach that connects compatible users, and detects communities in this context should run quickly and dynamically adapt to the evolution of the trending topic.

In other words, the increasing popularity of the topic implies new users interested in it (the number of users grows over time), being some of them creators of content and the others its distributors. The existent interactions among them and also, possible relationships with the already observed users in previous periods of time need to be considered. Accordingly, we tackle the problem of defining the significant (social) links that connect content creators and distributors in order to favor the spreading of information in the evolving graph of interest. When the type of links are established, we address the detection of trending topic communities; thus, short-term similar users are brought together.

The scenario in which we propose our research is Twitter. In such platform, follower users are interested in tracking down significant users to follow, whereas the followed (leader) users wish to accumulate a lot of followers. However, to generate significant content it is necessary to obtain interesting, trendy, and relevant information. One way of doing this is to form a "coalition" with other content creators or influencers in the domain. As a result, the influential group is able to share and filter key news before they become widely known,

[1]Once being part of a community, the connected users share the same interest in certain kind of content related to the given trending topic. The like-minded users may be seen as a group, be modeled in such way, and be provided with a specific recommendation.

Proceedings of the 7th Symposium on Future Directions in Information Access 2017

and then potentiate its diffusion through the group of users interested in that topic (distributors or consumers). Accordingly, we present a method to identify groups of topic-dependent "content creators" (*CCs*). Another key element of our proposal is the identification of their matching spreader groups or topic-dependent "content distributors" (*CDs*). After the identification of these two categories of users, both CCs and CDs are linked by our approach in a unique community, which represents the user base for the different forms of recommendation.

Given this real-world application scenario, our contribution is the detection of communities of users who (i) are associated to a given trending topic, (ii) are interested in the same content, (iii) are linked among themselves (*i.e.*, they follow each other), and (iv) can be either identified as content creators or content distributors. We validate our proposal on a real-world dataset extracted from Twitter, by employing standard metrics and by comparing it with a baseline approach that only requires the retweeting activity.

2. RELATED WORK

2.1. OSN Analysis to Discover User's Interests

It has been shown that friends share some similar interests (3). Therefore, recommender systems might make suggestions for the target user based on her/his friends' preferences. Thus, *social recommender systems* have emerged with the aim of modeling the user's preferences by using the information s/he and their friends have published in OSNs. For instance, the study done in (8) demonstrated that friends of the target user provided more useful and better recommendations than recommender systems. Ma *et al.* (7) also modeled the preferences of the user in a social recommender system under the idea that some of the user's friends might have different interests. The premise is that people tend to look for their friends recommendations; hence, this work establishes the difference between trust relationships and social friendships. In our paper we also consider the exploration of users' connections in the Social Web. However, our approach differs from (8) and (7), since the item recommendation for the user may be not only based on his/her direct friends, but also on a community to which the user belongs and which is related to a topic of interest.

2.2. Social Entity Recommendation on Twitter

In (4), the authors present a framework that merges a traditional collaborative ranking approach with Twitter features such as content information and social relations data, so the model can generate better personalized tweet recommendations. In (1), the authors make a proposal to solve the news feed filtering problem in OSNs by presenting a method that automatically reorganizes the feeds and filters out irrelevant posts. The authors in (5) propose a "users to follow" recommender, implemented by using data from Twitter. The details about algorithms and strategies used in their recommender system are presented in http://twittomender.ucd.ie.

Other social entities to recommend to Twitter users are hashtags (#). The hashtags give some relevant meaning and structure to the users' posts as a folksonomy. In (6), a method that recommends hashtags is presented. It is based on finding similar user-tweet pairs to the target user-tweet pair, so the hashtags used by the neighbors may be recommended. Compared to the state of the art, our approach may also be used to generate recommendations of news feeds, users to follow, hashtags, and other social entities. However, the novelty of our method is to employ a trending topic of interest to a set of users; consequently, the recommendations that can be generated are topic-dependent and are different for users who are content creators and for those who are distributors.

3. APPROACH

This section provides the details of our approach, named $TreToC$ (which stands for "*Tre*nding *To*pic *C*ommunities"), able to identify content creators and content distributors, as well as detect topic dependent communities. The approach works in three steps described next.

3.1. Identification of CCs

Users with a certain number of followers, whose tweets are quickly propagated because of their content, and who are experts or somehow represent a specific domain, may be considered creators of significant content. Given a trending topic $h \in H$, we collect the set of tweets T_h that contain h and consider the set of users U_h associated to these tweets (*i.e.*, those that either tweeted or retweeted content in T_h). Out of all the collected tweets, let T'_h denote the set of tweets that do not represent retweets (*i.e.*, those tweets that contain original content).

Every tweet $t \in T'_h$ is created by users who promote the content amplification over the social network. However, it is essential that the content is considered as interesting by other users, who retweeted a given tweet $t \in T'_h$ at least once. For this reason, we build a set $\hat{T'_h} \subseteq T'_h$, which contains these tweets:

$$\hat{T'_h} = \{t \in T'_h : retweets(t) > 0\}$$

where $retweets()$ returns the number of times a given tweet was retweeted.

Given the previously defined set, we designate as $CCs \subseteq U_h$ the collection of *content creators*, who favor the content generation:

$$CCs = \{u \in U_h : \exists t \in \hat{T}'_h \; s.t. \; author(t) = u\}$$

where $author()$ returns the author of a given tweet.

3.2. Identification of CDs

If a user retweets certain content as it is, s/he is showing an agreement with it. Therefore, the fact that a user *retweets* the tweets of another user is an important source of information to identify the content distributors. Consider that every user $u \in CCs$ posts a tweet $t \in \hat{T}'_h$. Let R_t be the set of tweets that represent a retweet of t:

$$R_t = \{t' \in T_h \setminus \hat{T}'_h : rt(t', t) = true\}$$

where $rt()$ returns true if a tweet t' is originated by a tweet t (*i.e.*, if it is a retweet of t).

We define as *content distributors* (CDs) the set of users who retweet content in $\hat{T}'_h$ and act as propagators:

$$CDs = \{u \in U_h : \exists t' \in \cup_{t \in \hat{T}'_h} R_t \; s.t. \; author(t') = u\}$$

3.3. Detection of Trending Topic Communities

Given the sets of users CCs and CDs, the first goal is to find an effective way to link them. Indeed, in order to allow a rapid spread of information, users should follow each other. Moreover, we have to ensure that an explicit connection between a CC and her/his CDs is present. In order to detect the communities related to a trending topic $h \in H$, it is first necessary to build a graph $G = (V, E)$ that represents the mentioned connections. The set V of vertices is represented as the union of the two sets of users identified in the previous two steps, $V = CCs \cup CDs$.

In order to build the set E of edges that represent the connections among the users, we consider three types of relationships. The first is the following relationship between two topic-dependent content creators:

$$F_C = \{(u_x, u_y) : follow(u_x, u_y) = true, u_x, u_y \in CCs\}$$

where $follow()$ returns true if the first user follows the second.

The second type of connection we consider is the following relationship between two topic-dependent content distributors:

$$F_D = \{(u_x, u_y) : follow(u_x, u_y) = true, u_x, u_y \in CDs\}$$

In the third type of connection we link a CC to a CD only if the CD retweeted content generated by the CC:

$$Ret = \{(u_x, u_y) : \exists (t', t) \in T_h \; s.t. \; rt(t', t) = true \wedge author(t') = u_x \; \wedge \; author(t) = u_y\}$$

Finally, the set E of edges in the graph is represented as:

$$E = F_C \cup F_D \cup Ret$$

At this point, the Louvain method (2) is applied to detect topic-dependent communities of interest in the graph G. The choice of employing Louvain was made since it can easily handle networks with millions of nodes in a very short time. This characteristic of the algorithm fits with our need to detect communities that rapidly evolve and are characterized by a temporary nature.

4. ANALYTICAL SETUP

To validate our proposal, four sets of analysis were performed:

1. **Characterization of the trending topics.** Given a trending topic, we analyze the number of content creators and distributors that characterize it.
2. **Analysis of the disconnected users.** We analyze the percentage of disconnected users from the graph (which would not be involved in the community detection and thus would not benefit of the information spreading).
3. **Analysis of the cohesion among the users.** For each community, we evaluate its quality by measuring the cohesion between the users in it, using standard metrics such as modularity, ratio between the number of communities and the number of users, and density.
4. **Analysis of the community structure.** For each community, we analyze its composition, by measuring the ratio of content creators and distributors in it, and their clustering coefficient.

In order to verify the choices made in our approach we compare our proposal with a baseline approach named *Retweeting-Based Communities* (RBC). In the RBC method, the set of edges in the graph connects two users only if one retweeted the other. The dataset contains 368 trending topics with 67,607 tweets, 15,918 unique creators, and 36,890 unique distributors.

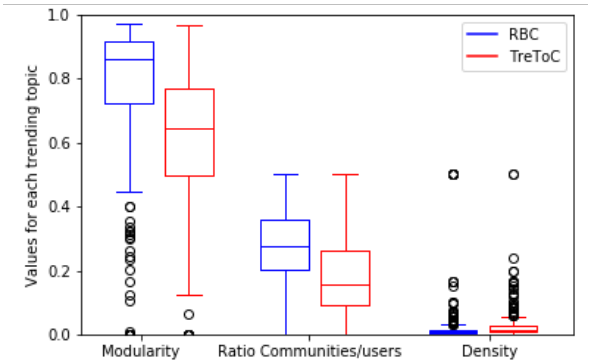

Figure 1: *Cohesion among the users.*

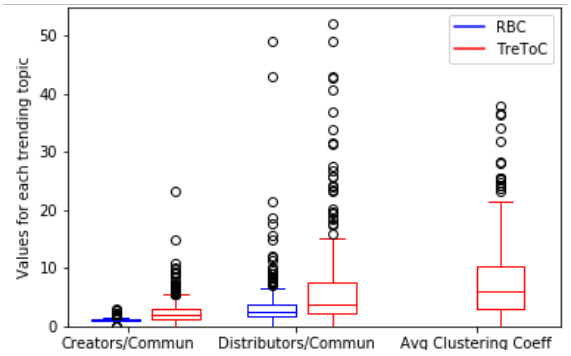

Figure 2: *Community structure.*

5. RESULTS AND DISCUSSION

When working with trending topics, their *characterization* showed that the data related to tweets presents a normal distribution (very few outliers) for which the average number of tweets per trending topic is 55.51, while the number of creators is 46.20. In contrast, with respect to the content propagation, the distribution is skewed to the right, showing that few trending topics reached a high incidence of retweets/distributors. The median value for the retweets per trending topic is 84.5 and the median number of distributors is 69. The analysis of the *disconnected users* showed that to detect communities that are related to a trending topic and involve most of the users, it is necessary to link the users both with the "following" and "who-retweeted-who" relationships. Indeed, the retweeting relationship alone (RBC) leaves around 24% of the users out of the graph, while the TreToC method does not consider around 15%. The analysis of the *cohesion among the users* (Figure 1) showed that the communities we created are large (the number of communities is very low if compared to the number of users), that the users in a community are well connected (density is high) and that the communities themselves are connected (modularity is not high); this means that the evolution of a trending topic over time would allow a user to be moved from one community to another, to better fit with her/his current interests and the evolution of the trending topic itself. The analysis (Figure 2), which studied the *structure of the communities* showed us that each community contains both a proper number of content creators and distributors (this would allow the distributors to get in touch with diverse content, generated by their content creators counterpart); moreover, the clustering coefficient confirmed that the nodes in the communities tend to cluster well together (the values are high), thus enabling the desired spread of information.

6. FUTURE WORK

In the context of recommendations, we propose for future work to generate suggestions of social items for groups of Twitter users by leveraging information about their corresponding topic-based creator groups. Diverse recommendations could be made by suggesting content from one community to another.

Acknowledgement

The author would like to thank ThoughtWorks for the funding provided to cover the registration fee for ESSIR 2017.

REFERENCES

[1] S. Berkovsky and J. Freyne. Personalised network activity feeds: Finding needles in the haystacks. In M. Atzmueller, A. Chin, C. Scholz, and C. Trattner, editors, *Mining, Modeling, and Recommending 'Things' in Social Media*, volume 8940 of *Lecture Notes in Computer Science*, pages 21–34. Springer International Publishing, 2015.

[2] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[3] P. Bonhard and M. A. Sasse. 'Knowing me, knowing you' – using profiles and social networking to improve recommender systems. *BT Technology Journal*, 24(3):84–98, July 2006.

[4] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu. Collaborative personalized tweet recommendation. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, pages 661–670, New York, NY, USA, 2012. ACM.

[5] J. Hannon, M. Bennett, and B. Smyth. Recommending Twitter users to follow using content and collaborative filtering approaches. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 199–206, New York, NY, USA, 2010. ACM.

[6] S. M. Kywe, T. A. Hoang, E. P. Lim, and F. Zhu. On recommending hashtags in Twitter networks. In *Proceedings of the 4th International Conference on Social Informatics*, SocInfo'12, pages 337–350, Berlin, Heidelberg, 2012. Springer-Verlag.

[7] H. Ma, D. Zhou, C. Liu, M. Lyu, and I. King. Recommender systems with social regularization. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 287–296, New York, USA, 2011.

[8] R. R. Sinha and K. Swearingen. Comparing Recommendations Made by Online Systems and Friends. In *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries*, 2001.