Detecting Like-minded Communities with Common Interests on Twitter

Kwan Hui Lim
School of Computer Science and Software Engineering
The University of Western Australia
Crawley, WA 6009, Australia
email: kwanhui@graduate.uwa.edu.au

ABSTRACT

The popularity and prevalence of online social networks (OSN) have made them efficient platforms for advertising and marketing campaigns. One important problem in target advertising and viral marketing on OSNs is the efficient identification of communities with common interests. Most current approaches use only topological links to first detect all communities, and then determine the interests of these communities. These approaches are both computationally intensive and may result in communities without a common interest. As topological links do not translate to actual user interactions, many of these detected communities may include members who rarely communicate with each other. This problem affects the application of targeted advertising and viral marketing to OSNs, since such campaigns require communities to be highly connected and interactive for the rapid diffusion of product/service information. We propose approaches to detect highly interactive and connected communities that share common interests on Twitter. Our approaches utilize Wikipedia to first identify and classify popular celebrities into various interest categories, before detecting communities based on linkages and communication patterns among followers of these celebrities. We also study the topological characteristics and interaction behaviour of these communities, showing them to be cohesive, connected and highly interactive about their common interests.

Keywords: Twitter, Online Social Networks, Community Detection

1. INTRODUCTION

With the rapid growth and proliferation of online social networking sites, many companies have embraced social media as a new outlet for their targeted advertising and viral marketing efforts. For example, the Twitter social network comprises 500 million users who produces 2,200 tweets per second. This large user base and high user activity provide tremendous opportunities for these companies to effectively reach out to a large audience group (of potential consumers) on such OSNs. In turn, this audience group may further propagate information about the products/services provided by these companies.

One important problem in the application of targeted advertising and viral marketing to OSNs is the efficient identification of like-minded communities with common interests in large social networks [3, 5]. Such communities should ideally comprise users of the right demographics who are also well-connected among themselves. The identification of the right demographic group is important to ensure the right product-audience matching [3] and the connectedness of this group facilitates the subsequent word-of-mouth advertising [5].

For the purpose of detecting like-minded communities comprising users with common interest, most of the current approaches involve first detecting all communities, followed by determining the interests of these communities [4, 7]. These approaches involve a lengthy and intensive process of detecting communities for the entire social network, which is both large and growing daily. In addition to the lengthy and intensive community detection process, many of the detected communities may not share the interest we are looking for.

Our proposed methods allow us to first select users with common interests and then identify communities comprising like-minded individuals with common interests on Twitter (and other OSNs). These methods differs from existing ones that first detect all communities, followed by identifying the topics they are interested in [4, 7]. Also, our methods do not unnecessarily detect communities that do not share any specific interest. Instead, our methods allow for the efficient detection of only communities sharing a common interest and can be applied to targeted advertising and viral marketing. In addition, our methods are able to detect communities at different levels of interest. As far as we are aware, there has been no prior study on the detection of communities with common interest on Twitter.

Our main contributions are as follows:

- A selection algorithm for identifying users with common interest based on their following of celebrities.
- A topological-based approach for detecting social network communities that share common interest, termed as the Common Interest Community Detection (CICD) method.

This paper is a short (partial) summary of my M.Sc. thesis and comprises the main ideas presented in [8, 9, 10].

18th School of Computer Science & Software Engineering Research Conference, Fairbridge, Pinjarra, Western Australia, 7–8 November 2012.

- An interaction-based approach for detecting highly interactive communities that frequently communicate about their common interests, termed as the Highly Interactive Community Detection (HICD) method.

2. METHODOLOGY

In this section, we first introduce the notations and definitions used, then describe our proposed selection algorithm for identifying users with common interest. Next, we present two approaches for detecting like-minded communities.

2.1 Notations and Definitions

We model the Twitter social network as a directed graph, $G = (U, L)$ where U refers to the set of users/nodes and L refers to the set of links/edges.[1] A followership link $(i, j) \in L$ indicates that user $i \in U$ is a follower of user $j \in U$, while a friendship link $Fr_{i,j}$ indicates $(i, j) = (j, i)$. We classify a Twitter user as a celebrity if he/she has more than 10,000 followers. This choice of 10,000 followers is supported by an extensive research on Twitter which shows that most users have less than 10,000 followers and those who do are mostly real-life celebrities [6]. As required, the definition of a celebrity can also be made more or less stringent by respectively increasing or decreasing the required number of followers.

The interest of a user in a category cat, Int_{cat} is inferred by the number of celebrities (of category cat) that the user follows. Although Int_{cat} represents the interest level of a user in a category, this metric is subjective due to the celebrities selected. The accuracy of Int_{cat} is dependent on the correct classification of celebrities into their respective categories, which is subjective as some celebrities loosely belong to multiple categories (e.g. a singer that has starred in some movies). We minimize this subjective judgment by using information on Wikipedia[2] to classify these celebrities into their respective categories. On the Wikipedia page of a celebrity, there is an "occupation" field which we use to determine the categories this celebrity belong to. Thus, this process minimizes the chances of classifying celebrities into the wrong category.

Twitter users communicate directly to other users by posting a tweet containing @username of the other user, along with the actual message. This direct communication is also called the @mentioning of other users. We define $M_{i,j}$ as a tweet posted by user i that contains a @mention of user j (i.e. the @mentioning process). Next, we also model the communication intensity $I_{i,j}$ of user i to j as the number of @mentions user i makes of user j. Table 1 lists a summary of all our notations and definitions.

While we describe our notations and definitions on the basis of Twitter, these notations and definitions can also be applied to other OSNs. Followership and friendship links respectively correspond to uni-directional and bi-directional links (or their appropriate representation) on other OSNs. As such, the definitions of a celebrity and Int_{cat} remains unchanged (as earlier described). Similarly on other OSNs, $M_{i,j}$ represents the private messages (or wall messages) that user i sends (or posts) to user j, and $I_{i,j}$ is the frequency of this messaging process.

[1]The terms nodes and users are used interchangeably but they refer to the same thing. Similarly, the terms edges and links are used in the same way.

[2]http://en.wikipedia.org/

Table 1: Notations and Definitions

Notations	Definitions
(i, j)	A followership (uni-directional) link from user i to user j
$Fr_{i,j}$	A friendship (bi-directional) link between users i and j
Celebrity	A user with more than 10,000 followers
Int_{cat}	The interest level of a user in category cat
$M_{i,j}$	A tweet containing a @mention of user j by user i
$I_{i,j}$	The number of times user i @mentions user j

2.2 Overall Framework

The overall framework for detecting like-minded communities involves first identifying users with common interest, followed by detecting communities among this set of users based on either topological or interaction links. The type of links used for detecting like-minded communities divides our framework into two different methods, namely the CICD method using topological links and the HICD method using interaction links. Both the CICD and HICD methods are based on the set of users with common interest which are selected based on their followings of representative celebrities. We shall first begin by describing our proposed selection approach for identifying users with common interest.

2.2.1 Identifying Users with Common Interest

The first step in selecting users with common interest is to identify a set of k celebrities $c_1, c_2, ..., c_k$, that belongs to a common interest category. As mentioned in Section 2.1, the classification of celebrities into their respective interest categories could be subjective. As such, we utilize information on Wikipedia to ensure the accuracy of our classification of these celebrities.

Our next step is to retrieve the set of Twitter users who follow all celebrities in a given category. Given a pre-identified set of k celebrities $c_1, c_2, ..., c_k$, we next identify all the followership links for each individual celebrity in this set. Consider celebrity $c_j, 1 \leqslant j \leqslant k$, and all the followership links for this celebrity $\bigcup_i link(i, c_j)$. We construct the set:

$$\mathcal{P} = \bigcap (\bigcup_i link(i, c_j)), for\ 1 \leqslant j \leqslant k$$

$\mathcal{P}$ is the set of fans who follow all the k celebrities in the set $\bigcup c_j, for\ 1 \leqslant j \leqslant k$. The criteria for constructing Set $\mathcal{P}$ can also be relaxed such that we select users who follow x out of k celebrities, where $0 < x < k$. In doing so, the value of x would determine the interest level Int_{cat} of the resulting Set $\mathcal{P}$. As the value of Int_{cat} is inversely correlated to the size of Set $\mathcal{P}$, a user has the flexibility to construct a larger Set $\mathcal{P}$ with the trade-off in a lower level of interest.

After constructing Set $\mathcal{P}$ (comprising users with common interest), we next determine the common links among these

users, where such links can be either topological or interaction links. Topological links correspond to followership or friendship links while interaction links correspond to communication (@mentioning) links. The choice of link type results in two different approaches to detecting like-minded communities, namely the topological-based CICD method and the interaction-based HICD method.

2.2.2 Topological Approach: Common Interest Community Detection

Our first proposed approach, the CICD method aims to detect like-minded communities comprising users with common interest, using only topological links. Topological links can be either followership or friendship links as introduced in Table 1. In this approach, we consider only friendship links (among Set $\mathcal{P}$) for community detection as friendship links are stronger and more reflective of real-life interactions. Using this set of friendship links (which corresponds to an undirected graph), we try to detect communities among the members of Set $\mathcal{P}$ next using the Clique Percolation Method (CPM) developed by Palla et al. [11].

The CPM defines a community as one with a series of adjacent k-cliques, where a k-clique comprises k nodes that are interconnected. We first identify all k-cliques in the network and connect them if they are adjacent. Two k-cliques are adjacent if they share $(k-1)$ common nodes. This procedure of connecting k-cliques continues iteratively until no adjacent k-cliques can be found. The result is a series of communities formed based on the k-cliques and adjacency criteria. For our experiments, we use CPM with a k-value of 3 as this produces the best results in detecting communities compared to other k-values.

Similarly, we also detect communities among the members of $\mathcal{P}$ next using the Infomap algorithm developed by Rosvall and Bergstrom [12]. Infomap approaches community detection as a coding or compression problem where the network graph can be compressed to retain its key structures. These key structures represent communities or clusters that are found within the network graph. Infomap uses random walks on the network graph to analyze information flow where the random walker is more likely to traverse within a cluster of nodes belonging to the same community.

Using both CPM and Infomap show that our proposed method produces results that are independent of the chosen community detection algorithm and their unique characteristics. CPM was chosen due to its ability to detect overlapping communities (which reflects real-life social communities) while Infomap was selected due to its superior performance compared to other algorithms [2].

2.2.3 Interaction Approach: Highly Interactive Community Detection

Our previous approach, the CICD method considers only topological information (such as follower/following links) but not user activity (such as communication/tweeting patterns and frequency). In a community where its users share common interests and are well-connected, the tweeting frequency and content of tweets are other factors that determine the speed of information diffusion. Many studies also support this observation, noting that only a small subset of users (among those connected by topological links) frequently interact with each other [1, 13]. Thus, it is necessary to consider user activity (communication/tweeting patterns and frequency) in addition to topological information for community detection, especially for advertising and marketing purposes. We now present the HICD method for identifying communities where its members not only share common interests but actively and frequently communicate about the common interests.

Our proposed HICD method detects a highly interactive community using the communication (tweeting) pattern and frequency among the users. This approach involves identifying community members based on their frequency of direct communication with other users in the community. As previously defined in Table 1, $I_{i,j}$ is the number of times user i @mentions user j in his/her tweets. Using this definition of $I_{i,j}$, we next build a list of weighted edges between two users i and j as a tuple $(i, j, I_{i,j})$ where $i, j \in \mathcal{P}$, and user j could be either an ordinary user or celebrity. Using a pre-determined intensity threshold T, we remove all tuples $(i, j, I_{i,j})$ if $I_{i,j} < T$ or $I_{j,i} < T$. In short, we are building a new set of users $\mathcal{Q}$ comprising only edges that exceed the threshold T.

Similar to the CICD method, we now detect communities among this set $\mathcal{Q}$ of users using CPM and Infomap. These stringent requirements for constructing Set $\mathcal{Q}$ ensures that the resulting Com_{HICD} is well-connected, cohesive and communicate frequently about their common interest.

3. EXPERIMENTS AND RESULTS

As the CICD and HICD methods have been extensively evaluated in our previous papers (listed in [8, 9, 10]), we shall give a broad overview of the experiments conducted and a summary of the results. Pre-print versions of the above-mentioned papers are also available at the following website, https://sites.google.com/site/limkwanhui/publications.

The datasets used for our experiments include a Twitter dataset collected by Kwak et al. [6] in June 2009 and another Twitter dataset that we collected from Nov 2011 to Jan 2012. The dataset by Kwak et al. comprises 41.7 million Twitter users and 1.47 billion links, while our dataset consists of 17,941 Twitter users and 1.9 million tweets.

We evaluated the proposed CICD and HICD methods using both topological and interaction measures. The topological measures include community size, average clustering coefficient, average path length and average degree, while interaction measures include the frequency and content of tweets among the communities, specifically the usage of @mentions, #hashtags, URLs and keywords.

Based on the topological measures, the results show that our proposed methods detect communities that are larger, more cohesive and only comprise users that share a common interest. Also, we observed how the topological structure of the detected communities become more connected and cohesive with deepening interest in a given category, as indicated by an increasing clustering coefficient and decreasing path length. Similarly, the communities become more connected

and cohesive as users specialize in their interest (e.g. from the general Music category to the specialized Country Music category). In addition, the interaction measures show the effectiveness of our HICD method in detecting communities that are highly interactive about their common interest, based on the content of their tweets (#hashtags and @mentions), their frequent tweets and high number of followers/followings.

4. CONCLUSION

In this paper, we first introduced a selection algorithm for identifying users with common interest. The interests of users are inferred based on the celebrities they follow and we first identify a set of celebrities that represents a certain interest category. Using this set of celebrities, we retrieve the followers of these celebrities and select users who follow all celebrities in this set. This set of selected users are deemed to have a common interest based on their following of these representative celebrities. In addition, we also proposed the CICD and HICD methods for detecting like-minded communities (using the common links among this set of users with common interest). Both methods aim to directly detect like-minded communities comprising users with common interest, without the need to first detect all communities followed by selecting the communities with common interests.

Our two proposed methods differ mainly in the usage of links for community detection. The CICD method detects communities using only topological information such as explicit bi-directional links. These bi-directional links are reflected in Twitter as a pair of users with mutual follower/following links, which are more representative of real-life social relationships. On the other hand, the HICD method uses implicit link information that is derived from communication links. These communication links are based on users @mentioning each other and result in communities that are more interactive, especially about the common interest. Due to this different usage of links, the communities detected by the CICD and HICD methods may overlap but are unlikely to be a subset of one another.

There are distinct advantages and disadvantages to both methods. The CICD method is able to detect like-minded communities using a single snapshot of the topological structure of the OSN (i.e. the topological links among users). However, using only topological links for detecting communities may not necessarily correspond to communities that are highly interactive. On the other hand, the HICD method is able to detect such highly interactive communities using communication (@mentioning) links among these users. However, such communication links cannot be retrieved in a single snapshot (unlike topological links) and instead have to be periodically retrieved at specific time intervals. Thus, the trade-off between the CICD and HICD methods are with the ease of links retrieval and the interactivity of the detected communities.

5. ACKNOWLEDGMENTS

Kwan Hui Lim was supported by the Australian Government, University of Western Australia (UWA) and School of Computer Science and Software Engineering (CSSE) under the International Postgraduate Research Scholarship, Australian Postgraduate Award, UWA CSSE Ad-hoc Top-up Scholarship and UWA Safety Net Top-Up Scholarship.

6. REFERENCES

[1] Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proc. of IMC'08*, pages 57–70, Oct 2008.

[2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

[3] Ganesh Iyer, David Soberman, and J. Miguel Villas-Boas. The targeting of advertising. *Marketing Science*, 24(3):461–476, 2005.

[4] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proc. of WebKDD/SNA-KDD '07*, pages 56–65, Aug 2007.

[5] Andreas M. Kaplan and Michael Haenlein. Two hearts in three-quarter time: How to waltz the social media/viral marketing dance. *Business Horizons*, 54:253–263, 2011.

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proc. of WWW '10*, pages 591–600, Apr 2010.

[7] Daifeng Li, Bing He, Ying Ding, Jie Tang, Cassidy Sugimoto, Zheng Qin, Erjia Yan, Juanzi Li, and Tianxi Dong. Community-based topic modeling for social tagging. In *Proc. of CIKM '10*, pages 1565–1568, Oct 2010.

[8] Kwan Hui Lim and Amitava Datta. Finding Twitter communities with common interests using following links of celebrities. In *Proc. of MSM '12*, pages 25–32, Jun 2012.

[9] Kwan Hui Lim and Amitava Datta. Following the follower: Detecting communities with common interests on Twitter. In *Proc. of HT '12*, pages 317–318, Jun 2012.

[10] Kwan Hui Lim and Amitava Datta. Tweets beget propinquity: Detecting highly interactive communities on twitter using tweeting links. In *Proc. of WI '12*, page to appear, Dec 2012.

[11] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, Jun 2005.

[12] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123, 2008.

[13] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proc. of EuroSys'09*, pages 205–218, Apr 2009.