User Community Reconstruction using Sampled Microblogging Data

Miki Enoki
IBM Research – Tokyo
1623-14 Shimo-tsuruma
Yamato, Kanagawa, Japan
+81 46 215 4998
enomiki@jp.ibm.com

Yohei Ikawa
IBM Research – Tokyo
1623-14 Shimo-tsuruma
Yamato, Kanagawa, Japan
+81 46 215 4998
yikawa@jp.ibm.com

Raymond Rudy
IBM Research – Tokyo
1623-14 Shimo-tsuruma
Yamato, Kanagawa, Japan
+81 46 215 4998
raymond@jp.ibm.com

ABSTRACT
User community recognition in social media services is important to identify hot topics or users' interests and concerns in a timely way when a disaster has occurred. In microblogging services, many short messages are posted every day and some of them represent replies or forwarded messages between users. We extract such conversational messages to link the users as a user network and regard the strongly-connected components in the network as indicators of user communities. However, using all of the microblog data for user community extraction is too costly and requires too much storage space when decomposing strongly-connected components. In contrast, using sampled data may miss some user connections and thus divide one user community into pieces. In this paper, we propose a method for user community reconstruction using the lexical similarity of the messages and the user's link information between separate communities.

Categories and Subject Descriptors
J.4 [Computer Applications]: Social and Behavioral Sciences

Keywords
Microblogging, Twitter, Community reconstruction, Social Media

1. INTRODUCTION
In contrast to other social media services, a microblogging service such as Twitter [1] has special characteristics in that the frequency of users' posts is high and strongly related to real-time topics. Also, there is intense interaction between users through replies and forwarded posts. Many people around the world use microblogging as an important new type of Internet communication tool [6, 7].

After the Great East Japan Earthquake occurred of 2011, cellular and landline phone services in Japan were either unavailable or unreliable for some days, while Internet services were largely unaffected [10]. As a result, many people got real-time information from social media rather than that from mass media. This included such important data as crucial resources for damaged areas and new evacuation sites. In social media services, people became not only information receivers, but also senders propagating new information [17].

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

To find valuable information in a flood of data on social media, classifying it from various perspectives is effective. There have been several classification techniques for Twitter data. Mathioudakis et al. [12] find trending topics by detecting bursty keywords and grouping them into trends based on their clustering. Yamaguchi et al. [19] discovered appropriate topics for a user by using Twitter list function. Java et al. [3] presented their observations of microblogging phenomena by studying the topological and geographical properties of the tweets and tried to find communities in the follower networks.

User community recognition is also valuable for marketing. Many companies see microblogging social media services as ideal places to directly obtain timely opinions from customers and potential customers. They actively look for ways to use these insights for better customer relationships and for product marketing. Guan et al. [20] found a tendency for users who purchased a product to often be in the same community in a social-media friendship network. Therefore, targeting a campaign at potential users who belong to a community that has a significant number of users who have already purchased the product is likely to be effective in stimulating new purchases. The potential users are likely to be influenced through their interactions in the social media with those users who have purchased the product.

In this paper, we propose a method to extract user communities in a microblogging service based on their intense interactions. We do this by building a network in which the nodes correspond to users and the edges correspond to interactions (replies and forwarded messages) among the users, identifying sets of strongly connected nodes in the network, and then confirming the corresponding users of the identified node sets are a community.

Unfortunately, obtaining all of the interactions of the users in a social media system such as Twitter is often impossible due to the computing time and storage required for processing the tremendous amounts of data, not to mention the restrictions on using APIs to collect such data over a network. For these reasons, methods to analyze the social media interactions are severely constrained and the communities are likely to be divided when only some interactions among the members are sampled.

In this paper, we also address the problem of extracting user communities from samples. Our method uses the network structure and the text content of the sampled interaction networks to recognize unobserved interactions.

2. FINDING TWITTER COMMUNITIES

2.1 Twitter

Twitter is relatively mature with over 300 million users (as of June 2011). Twitter users can simply post any message up to 140 characters, which is called a "tweet". All of the normal tweets are by default public. Each tweet has a unique tweet ID and links to such information as the user account name, profile, and the user's location.

A typical Twitter user usually 'follows' the tweets of a limited number of people, but all of the public tweets can be read or cited by navigating through account names (the user IDs), by searching for words, or by using URL links to specific tweets. "To follow" means one user subscribes to all of another user's tweets so that those tweets are included in the first user's personal view. The followed user is added to a list of followed users. The users who follow a user are called "followers" of that user. A "timeline" is a list of a user's own tweets and followed tweets, in chronological order.

Twitter also has two special tweet types for communication among users. These are called replies and retweets.

Reply

A "reply" is a tweet sent in response to an earlier tweet. The reply includes the earlier tweet ID. This type of tweet typically includes "@account_name" followed by the new message. All of the users following either the replying user or the replied-to user can read this tweet.

ReTweet (RT)

When a tweet is "retweeted" by a user, it is broadcast to that user's followers. For an official RT with Twitter's retweet command, the tweet is simply reposted without changing the original message An unofficial RT is created by retweet users using the "RT @account_name original message" notation, possibly followed by additional text.

2.2 Identification of Strongly Connected Components from User Networks

A notable feature of Twitter beyond the large volume of message data is the streaming structure of the conversations based on the relationships between tweets. Of the various types of tweets, about 47% are conversational [14]. A later tweet may be a reply to an earlier tweet using the reply mechanism. Another kind of link is when a tweet cites another tweet within its text using the RT mechanisms. Subsequences of these kinds of tweets can produce conversational streams involving multiple users.

We can extract a user community from these conversational streams by building a network whose nodes correspond to users and whose edges correspond to direct interactions (Replies or RTs) among users. In this network, we identify sets of strongly connected nodes and define the corresponding users of the identified node sets as forming communities. The strongly connected components form a maximal subset of nodes containing a directed path from each node to all of the others in the subset [5]. Since the users in the community have histories of Reply or RT message among each other, they are regarded as relatively close relationships similar to friendships. In addition, we can determine the topic of each community by analyzing the tweets that the members exchanged. For example, we can extract the most frequent words for each community.

2.3 Sampling Twitter Data

When we analyze user community from Twitter, we often face the difficulties in using coarsely-sampled data. For example, calculating the networks based on 65 million tweets per day [8] would be an extremely time-consuming job. In addition, continuously storing a thick tweet stream would require too much storage space and CPU resources. Also, the no-fee public Twitter API [2] can only collect about 1% of the data, so communities are likely to be divided due to interactions among the members that are lost in the sampling process.

3. ASSESSING USER COMMUNITIES FROM SAMPLED INTERACTION DATA

In this section, we describe a promising method to restore user communities in Twitter from communities that are observed in the sampled tweets. The method utilizes both the textual content of the tweets and the network structure of the user friendships in the sampled data to restore complete communities that are related to specific topics or activities.

3.1 Similar Textual Contents in Tweets

Our hypothesis is that users of a particular community tend to exchange tweets that are related to the topics or activities of the community. Therefore, it should be possible to identify the members of a community from the similarities of the keywords and phrases that appear in their tweets. For example, technical keywords that describe lenses, manufacturers' names, types of cameras, and so on are expected to be observed more frequently within a community of DSLR (Digital Single-Lens Reflex) camera enthusiasts. To test this hypothesis, we parsed the tweets of users within the community to identify the keywords whose frequencies are then used as elements of the feature vector of the community. Since keywords that appear in many communities are not helpful for characterization, we assign tf-idf scores as the weights of the keywords in the feature vectors so that the keywords that are unique to a community are emphasized.

With this procedure, we can construct a feature vector for each observed community and compute the similarity between two communities by using the cosine similarity score of their feature vectors. If the similarity score is above a given threshold, the two communities should be merged to form a larger community.

3.2 Similarities of Twitter Follow Relationships

Other than the "Reply" and "RT" relationships, Twitter users are also connected by the semi-static "Follow" relationships. The users of a particular community are also likely to have some actual relationship in the real world. If the relationship is reflected in their community, they are likely to follow (or be followed by) other similar users. Thus, by comparing the lists of followers of the users of a divided community, we can merge communities that share more than some threshold number of users to merge the underlying divided community.

3.3 Proposed Method

In Sections 3.1 and 3.2, we discussed similarity metrics that can be used to recognize divided communities. Relying only on the metric of Section 3.1 might link accidentally similar interests without true relationships, so they might be mistakenly grouped into a community without actual friendship relationships.

Alternatively, if only the metric of Section 3.2 were used, then communities that share some friendship relationships could be recognized but different interests might be blurred together, since they are exchanging tweets on various topics. Spurious communities are less valuable for extracting focused insights. Therefore, we propose using the combined similarity metrics from Section 3.1 and 3.2 to find meaningful communities.

4. EXPERIMENTAL RESULTS

In this section, we evaluate the effectiveness of user community extraction for tweets after the Great East Japan Earthquake of 2011. Then we measured the accuracy rate of our community reconstruction using the sampled Twitter data.

4.1 Experimental Data

For our experiments, we collected posted tweets at the default access level of the Twitter Streaming API [2] from 03/16/2011 to 04/03/2011. The Twitter Streaming API has three levels with decreasing volumes of data, firehose, gardenhose, and spritzer. The firehose level returns all of the public tweets, but this is only available to the privileged users. The spritzer level is the default level that allows anyone to access streaming tweets without special permission, but the returned messages are only a 1% sample for all of the public tweets. In this paper, we regard the sampled data as all of the tweet data for our experiment to measure the accuracy rate of reconstruction.

In the collected data, we filtered using disaster related keywords (in Japanese) such as "disaster", "disaster-affected", "earthquake", "nuclear power", "buyout", "brownout", etc. The filtered tweets were posted by about 300,000 users.

4.2 Evaluation Methodology

In our experiments, we identified strongly connected nodes with more than three users as a user community. Four may seem to be a small minimum for a user community, but the filtered user communities were already small in our experimental data as collected with the default-level API. With more data, such as the firehose level, we could increase the minimum size.

To evaluate the accuracy of our community reconstruction, we created two sampled datasets by randomly deleting 20% and 50 % of the experimental data. The complete set of all of the communities in our experimental data is regarded as the correct set of communities. When a community lacks more than 50% of users of the actual members of the original community, we call it a 'divided community'. If a merged community includes more than 80% of the members of the original community (found in the complete data), then we call it a 'correctly reconstructed community'. For comparison, we measured the accuracy rate of community reconstruction by using three metrics: (1) Using only the similarities of textual contents (Section 3.1), (2) Using only the similarities from the Twitter follow relationships (Section 3.2), (3) Our proposed method (Section 3.3).

4.3 Experimental Result

4.3.1 Community Extraction

We extracted the strongly connected nodes from our experimental data to evaluate the existing communities. Our data contained 203 communities. We reviewed the tweets of each community and found that each community had close relationships among the members and some communities had a local topic of interest. For example, members of the community shown in Figure 1 seem to live in Fukuoka prefecture (in western Japan) and exchanged messages about a volunteer event for East Japan held in Fukuoka. Another community's members were interested in pet shelters for the disaster area. They seemed to have pets and were concerned about the animals in the disaster area.

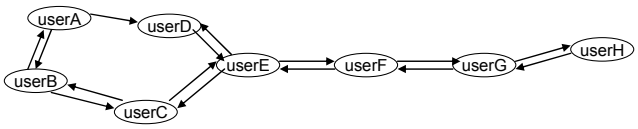

Figure1. Extracted Community.

4.3.2 Community Reconstruction

In the next experiment, we evaluated the accuracy of community reconstruction when the data was sampled. Figure 2 shows the results of community reconstruction using 80% of the data. The "Correct" is for the correctly reconstructed communities relative to the divided communities. The "False-positive" represents the ratio of incorrectly reconstructed communities to the divided communities. The results show that our proposed method had the highest accuracy while using only the textual similarities had the lowest. The result using only the similarities of the Twitter follow relationships was the second most accurate, but the rate of false-positives was quite high. The results using 50% of the data (shown in Figure 3) show even larger differences.

As a result, we found that the metric of textual similarities is not very effective in reconstructing divided communities extracted from a large network, but if we use the textual similarity metrics after selecting candidate communities for reconstruction by using the similarities of the Twitter follow relationships, this is effective in avoiding spurious communities.

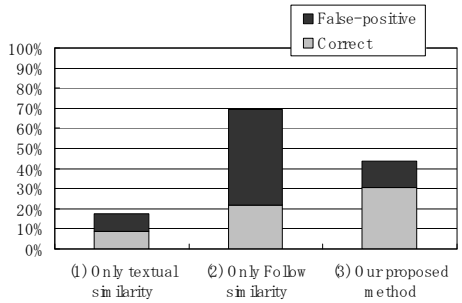

Figure 2. Accuracy of Community Reconstruction using 20% Reduced Twitter Data.

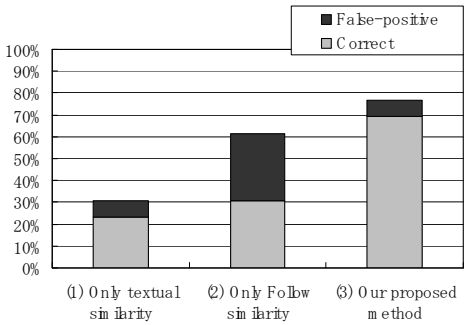

Figure3. Accuracy of Community Reconstruction using 50% Reduced Twitter Data.

5. RELATED WORK

There have been many methods originated in the graph theory to find clusters of users and extract communities in the social network. In this paper, we use a linear-time graph algorithm to extract strongly connected components of a directed graph which was firstly proposed by Tarjan [15]. Other graph algorithms such as enumerating pseudo cliques, as in Uno [18], and Haraguchi and Okubo [11] , and discovering large dense subgraphs, as in Gibson et al. [4], might be used instead to extract community candidates. However, we suspect that the number of pseudo cliques and the size of dense subgraphs in the sampled data is not large. Given a sufficient number of network samples, it would be interesting to apply graph-theoretic approaches as in Kumar et al. [16] to locate emerging communities and see how semantic information obtained from the contents of the microblog data are related to them.

Newman et al. [13] advocated modularity as an optimization criterion. Clustered networks with high modularity have dense connections between the nodes within the modules but sparse connections between nodes in different modules. Flake et al. [9] used the hypothesis that the number of links between nodes in dense networks may large and extracted clustered networks with a max-flow min-cut theorem. These approaches were applied to undirected networks, while our focus is on directed networks, because we want to find user communities with close relationships similar to friendships.

6. CONCLUTIONS

User community recognition in social media services is important to identify hot topics or users' interests and concerns in a timely way when a disaster has occurred. After the Great East Japan Earthquake, many people obtained real-time information from social media rather than that from mass media.

In this paper, we proposed a method to extract user communities in a microblogging service based on their intense interactions by building networks where the nodes correspond to users and the edges correspond to interactions, and then identifying sets of strongly connected nodes in the network. Our experimental results showed that each extracted community had close relationships among the members and some of the communities had local topics of interest.

We also addressed the problem of extracting user communities from sampled data. Our method uses the network structure and the text content of the sampled interaction networks to recognize unobserved interactions. As a result, we found that using our method produced the highest accuracy compared with using only textual similarities or using only the similarities of the Twitter follow relationships.

7. REFERENCES

[1] http://twitter.com/

[2] https://dev.twitter.com/

[3] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM,2007.

[4] D. Gibson, R. Kumar and A. Tomkins. Discovering Large Dense Subgraphs in Massive Graphs, VLDB, pp. 721–732, 2005.

[5] E. Nuutila and E. Soisalon-Soininen. On finding the strongly connected components in a directed graph, Information Processing Letters, Vol.49, No.1, pp. 9-14, 1994.

[6] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing user behavior in online social networks. In Proc. of IMC ’09, pages 49-62, New York, NY, USA, 2009. ACM.

[7] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger. Understanding online social network usage from a network perspective. In IMC ’09.

[8] Garrett, Sean. "Big Goals, Big Game, Big Records".http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html. Retrieved February 7, 2011.

[9] G. W. Flake, S. Lawrence and C. L. Giles. Efficient Identification of Web Communities, In Proc. KDD, 2000.

[10] "Japan's phone networks remain severely disrupted", Computerworld. 12 March 2011. Archived from the original on 18 April 2011. http://www.computerworld.com/s/article/9214261/Japan_s_phone_networks_remain_severely_disrupted.

[11] M. Haraguchi, Y. Okubo. A Method for Clustering of Web Pages with Pseudo-Clique Search, Lecture Notes in Artificial Intelligence 3847, pp. 59–78, 2006.

[12] M. Mathioudakis and N. Koudas. TwitterMonitor: trend detection over the twitter stream, Proc. of ACM SIGMOD, pp. 1155-1158,2010.

[13] Newman, M. E. J. Fast algorithm for detecting community structure in networks, Physical Review E, Vol. 69, No. 066133,2004.

[14] Pear Analytics (2009) Twitter Study – August, 2009. http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf

[15] R. E. Tarjan. Depth-first search and linear graph algorithms, SIAM Journal on Computing 1 (2): 146–160, 1972.

[16] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins. Trawling the Web for emerging cyber communities, In Proceedings of the 8th international conference on World Wide Web (WWW), pp. 1481–1493, 1999.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors, In Proceedings of the 19th international conference on World Wide Web (WWW), pp.851–860,2010.

[18] T. Uno. An efficient algorithm for solving pseudo clique enumeration problem, Algorithmica (56-1): 3 - 16, 2010.

[19] Yuto Yamaguchi, Toshiyuki Amagasa, and Hiroyuki Kitagawa. Tag-based User Topic Discovery using Twitter Lists, pp. 13-20, The International Conference on Advances in Social Network Analysis and Mining (ASONAM), July 25-27, 2011.

[20] Z.Guan, J. Wu, Q. Zhang, A. Singh, and X. Yan. Assessing and Ranking Structural Correlations in Graphs, Proc. of ACM SIGMOD, pp. 937-948, 2011.