Discovery of interesting users in Twitter by overlapping propagation paths of retweets

Yusuke Ota
The University of Electro-Communications

Kazutaka Maruyama
The University of Tokyo

Minoru Terada
The University of Electro-Communications

Abstract—In recent years, social networking services have come into wide use to people. Especially, one of micro blog services, Twitter is a significant service. Twitter user gets information by following other users whose tweets match his interest. Retweet is one of Twitter functions which spreads tweets to other users. Using retweets, one can read tweets originated by users who are not followed by him. Our goal is to discover Twitter users who retweet many tweets which match the interest. We focus on the propagation of retweets and build a graph, the Overlap Graph, which contains users who share same retweets. Finally, we validate the users appearing in the graph by checking the frequency and the content of their retweets.

Keywords-Twitter; Retweet; Social network;

I. INTRODUCTION

In the whole world, many social networking services are appearing, which include Facebook1, Twitter2, Myspace3 and so on. To use these web services, a user has social connections to many other users. Twitter is a web service which enable us to collect much information from many users. In Twitter, a user follows others to see their short messages (tweets) on his home timeline. When user A follows user B, A is a follower of B, and B is a followee of A. In Twitter, there is much information about a user, including the profile, tweets, and the social relationship with other users. Users judge whether to follow a user based on this information.

In this paper, we focus on retweet. A user can repost other users’ tweets which he likes by using retweet. There are similar functions in other social networking services, such as the Reblog function in Tumbler. A user who retweets often can be seen as a hub because he collects and distributes interesting tweets, even if he does not make tweets by himself. So, we think such hub user who shares common interest with us is a target user who collects and distributes interesting tweets, even if he does not make tweets by himself. Our goal is to discover Twitter users who retweet many tweets which match the interest. We focus on the propagation of retweets and build a graph, the Overlap Graph, which contains users who share same retweets. Finally, we validate the users appearing in the graph by checking the frequency and the content of their retweets.

In this research, we consider that retweet has various powers which are as follows:

1) Filtering by user: Good tweets are retweeted by many users, which means that retweet is a filtering by users. A user consciously judges whether specific tweet is worth retweeting since retweet is an action that a user actively do. Therfore, only tweet that a user wants to tell his followers is selected. In this aspect, we consider retweets are more valuable than normal tweet.

2) Same interest user group: Boyd et al. [1]find users retweet because they agree to the tweet and want to spread content of the tweet to other users. For that reason, retweeted users are interested in content of retweet. Thus, it can be said that retweeted users are a group of users sharing common interest about content of retweet. Also, retweets can be utilized for a barometer of user’s interest.

Rest of the paper is organized as follows: In section II, we describe the power of retweet. In section III, we present some researches about mining Twitter. In section IV, we show our idea to find target users. In section V, we present the way to estimate propagation path of a retweet and to construct overlap graph. In section VI, we present the result of verification that users appearing in the overlap graph satisfy the purpose of our research by three measures. In section VII, we discuss informal retweets, visualizing graph, and automatic categorization of tweets. In section VIII, we present the conclusion of our research and future works.

II. RETWEET

Retweet is one of Twitter functions that a user shows other user’s tweet to his followers. There are two kinds of retweets, one is provided by Twitter, which is called formal retweet, and the other is informal retweet, which is a usual tweet that includes someone’s tweet with comments.

A. Propagation of Retweet

A user may retweets followee’s tweets (and retweets) which are displayed on his timeline. And, his followers see this retweet, some of them retweet. As a result, tweets are propagated to users who are not direct followers of the original user. Retweet propagates from one user to others, resulting a tree-shaped path of users. We define some terms about retweet propagation paths of this imaginary tree.

- Parent User: User who is retweeted by child user
- Child User: User who retweeted from parent user
- Retweet Root User: User who tweets retweet

An imaginary retweet propagation tree represents how a certain retweet are propagated from retweet root user to many users.

B. Potential of retweet

In this research, we consider that retweet has various powers which are as follows:

1) Filtering by user: Good tweets are retweeted by many users, which means that retweet is a filtering by users. A user consciously judges whether specific tweet is worth retweeting since retweet is an action that a user actively do. Therfore, only tweet that a user wants to tell his followers is selected. In this aspect, we consider retweets are more valuable than normal tweet.

2) Same interest user group: Boyd et al. [1]find users retweet because they agree to the tweet and want to spread content of the tweet to other users. For that reason, retweeted users are interested in content of retweet. Thus, it can be said that retweeted users are a group of users sharing common interest about content of retweet. Also, retweets can be utilized for a barometer of user’s interest.

1http://www.facebook.com/

2http://twitter.com/

3http://www.myspace.com/
3) **Relationship between users:** Using retweets, a user can find users who are ‘distant’ in the sense of follow relations, because tweet is spread to many users and shared by being retweeted. They could be found by tracing follow relations. It can be said that finding user to suit by using information of retweet may be more efficient than random search.

4) **Discovery of active user:** Users who are retweeting tend to use Twitter positively in a bid to tell other user about tweet which is interesting. Users who are only to tweet thinks that they want to post about them to other user, but users who are retweeting quite often are interested in other user’s tweet, and they are trying to spread that with devotion. In our research, the latter user is important because they provide to a lot of information.

## III. RELATED WORKS

In this section, we show some research to recommend users to follow. There are various types of users whom a user wants to follow, such as acquaintance in real world (friend, colleague, or family), user who tweets about interested field, and widely known user. So, there are many ways to recommend users to follow. There may be two types of approaches to recommend candidates to follow; content-based approaches which analyzes content of user’s tweet, and social graph based approaches which analyzes relationships of follow.

Hannon, et al. [4] propose that Twitter users can be usefully modeled by the tweets and relationships of their Twitter social graphs. They explore a variety of different recommendation strategies including content-based techniques and collaborative filtering style approaches. Collaborative filtering is a recommending technique which filters for information or pattern by using many users’ interest information. They use collaborative filtering on social connection in Twitter, such as follower and followee. Social graph based approach can explore users who are ‘near’ in the sense of follow relations, but it is difficult to find users who are ‘distant’ in the sense of follow relations.

On the other hands, in order to use content-based approach, it is necessary to analyze exactly the content of tweets. Michelson, et al. [5] discover user’s topics of interest by leveraging Wikipedia as a knowledge base and classifying tweets. Wikipedia includes information of relationship between items and provides encyclopedic knowledge. In this way, several studies have been made on method of tweet categorization using a Wikipedia as knowledge base. Also, Yoshida, et al. [8] propose an algorithm to solve one type of name disambiguation problems. Assuming that tweets are retrieved by the organization name query, the problem is to decide whether each organization name found in each tweet represents the target organization or not (such as “Apple PC” for the former and “Apple Pie” for the latter for the query “Apple”). Their algorithm which is based on the intuition that organization names can be classified into “organization-like name” and “general-word-like names” solves this problem. Because tweet is complex and obscurity, categorizing tweet is hard. However, our approach which uses retweet can find users who are interested in a content of retweet without analyzing retweet and being dependent of any specific language.

## IV. PROPOSED METHOD

In order to find users who are likely to give interesting tweets, we propose two graphical representations of user network based on the propagation of retweets. In section II, we mentioned that users who retweeted a certain tweet have common interest about it. From this insight, we can extract users who share common retweets and add them as followees of a user because they will have similar interest to that of the user. Additionally, those users are ‘active’ in retweeting and are good candidates to follow.

The social relationship is important in following users. We construct a tree which represents the propagation path of a retweet, which we call the **Retweet Propagation Path Tree**. Users are represented as nodes in the tree, the original author of the retweet (the root user) is the root node. An edge in the tree represents a single step of retweet, which is usually a follow relation in Twitter. Using this tree, we can see the social relationship of users; the distance between users is shown as the number of edges, nodes in between are the chain of acquaintances. These information is one of the key informations when a user decides whether to follow someone.

Furthermore, we construct another graph to recognize relationship among users on several retweets. We select a user (the **center user**) and collect his retweets. Overlapping propagation trees for these retweets, we construct a directed graph, the **Overlap Graph**. Overlap graph contains users who appear more than once on retweet propagation path trees, each retweet root users, and users who appear on retweet propagation paths. From the graph we can find users who often retweet tweets that matches the center user’s interest, with the indication how those users are related to the user. The number of occurrence of a user in the overlap graph indicates the similarity to the center user.

## V. IMPLEMENTATION

### A. Data set

The public Twitter API\(^4\) enables us to get various information about tweets. We used the Twitter4J\(^5\), a JAVA wrapper for TwitterAPI, to get the following information:

- tweets
- retweets, and
- followers.

For a tweet, we can get the author of the tweet, time when user tweets, whether the tweet was retweeted and the number of users who retweeted it. We can pick up only tweets that have been retweeted, which we call retweets. For a retweet, we can get a list of users who retweeted it and time when they retweeted. We collected the latest two hundred retweets a user retweeted, and inferred propagation paths for them using the relationship between the followers of the user.

\(^4\)https://dev.twitter.com/
\(^5\)http://twitter4j.org/en/index.html
B. Retweet propagation path tree

For formal retweets, Twitter API enables us to retrieve the list of users who retweeted a tweet, but it does not provide the order of retweet and we cannot know who retweets a tweet just before a certain user. So, we need to infer the retweet propagation path based on information which is time when users retweet and users’ followers. Although informal retweet keeps information of propagation path in the tweet, it is difficult to make good use because of the limitation of the length of a tweet (we discuss informal retweet on section VII-A). It is necessary to presume propagation path of formal retweet.

In this research, we presume propagation path of retweet by time when user retweets. There are two ways in which a user finds a tweet to be retweeted:

- it appears in his timeline,
- it appears at other sources (e.g. from search result).

A timeline of a user contains his followees’ tweets, and a user usually retweets to followee’s tweet. When a user retweets a tweet retweeted by his followee in his timeline, it is clear that his followee’s retweet occurs before his own. However, if there are some followees who retweet earlier than one’s own, it is unable to decide parent user uniquely. In this time, assuming that a user retweets when they find first, we decide that a parent user is the earliest user who retweeted the tweet in the candidates.

For the second way, the user’s timeline does not contain the parent tweet. We consider that a user retweeted directly from the retweet root user.

C. Overlap Graph

Method of overlap graph construction is simple. First, pick up users who appear more than one times on retweet propagation path trees. Second, add the paths from them to the root user on each tree to overlap graph. Figure 1 shows an example of construction of overlap graph. This overlap graph is made by three retweet propagation path trees (A, B, C).

D. Visualization of Graphs

Figure 2 and Figure 3 are visualization examples of retweet propagation path tree and overlap graph respectively.

1) Figure2: Retweet propagation path tree (Figure2) displays that retweets spread along the direction of the arrow from retweet root user (triangular node). The depth of color of the nodes color indicates the time of retweet, deeper color for later retweets. The dotted edge indicates that parent user was not found properly (in the second case in the sectionV-B). We can see the overall flow of the retweet and the relationship between users from this tree.

2) Figure3: Element in the overlap graph are as below:
   - Square node: the center user
   - Triangular node: retweet root user
   - Node color: deeper color for a user which appears more frequently
   - Direction of edge: direction of retweet transmission

- Thickness of edge: the number of passage of path

Overlap graph contains users who share more than one retweets (nodes with color), retweet root users (triangular node), and users who exist in retweet path from retweet root user (white color node). Just like retweet propagation path tree, retweets spread along the direction of the arrow from retweet root user. Figure 3 is an example of a overlap graph of fifty retweet propagation path trees.

3) Effectiveness of overlap graph: In Figure 3, there are users (like user A) whose node color is deep and who connects
the center user through many hops. Without the overlap graph, the center user cannot find them although they retweet just as him. So we call them potential same interested user. Overlap graph is suitable for finding these users visually because they are hard to be found from information of a single retweet.

VI. VERIFICATION

In this section, we verify that users who appear in an overlap graph actually retweet about the center user’s interest, based on the frequency and contents of retweets. We evaluate as below:

- the number of retweets in a certain period,
- the ratio of good retweets to total retweets in the period, and
- similarity of categories of retweet subjects between the center user and the potential user.

A good retweet is a retweet whose category is included in the interest of the center user.

A. Dataset

We verify the three highest users (A, B, C) in the number of occurrences in the overlap graph and the five lowest users (V, W, X, Y, Z) who are chosen randomly from the users who appear only once in the graph. We call them verification users.

In order to verify them we collected the latest 100 retweets of the center user and verification users’ tweets for three days, which are distinct from the tweets used for the construction of the overlap graph.

B. Categorizing retweet

We categorize retweets according to their subjects. First, the center user (one of the authors of this paper) categorized his retweets by himself. Each retweet of the center user is put into a category if it already exists, or a new category is added to include the tweet. The example of category is “programming”, “anime”, “study” and so on. The resulting center user’s categories are marked as $CAT_1 \sim CAT_n$. We add another category $CAT_{n+1}$, which includes tweets which failed to be categorized in $CAT_1 \sim CAT_n$. Then, the center user also categorized verification users’ retweets which were extracted from tweets for three days into those categories.

Using the result, we made a n+1 dimensional vector ($C_{user}$) for each user.

$$C_{user} = (C_1, C_2, \ldots, C_n, C_{n+1})$$

We call this vector CategoryFeatureVector. $C_i$ indicates the number of retweets which is categorized as $CAT_i$. $C_{n+1}$ also includes the number of verification users’ tweets which are not retweeted for three days.
C. Number of retweets

The first metric is the number of retweets of the verification users. We counted the number of the verification users’ retweets for three days.

D. Ratio of good retweets

The second metric is the ratio of good retweets to all retweets. As defined in previous section, a good retweet is a tweet that matches the interest of the center user. A high score in this metric means the verification user’s retweets are mostly included in the interest of the center user.

E. Similarity of retweet categories

The third metric is a similarity of retweet categories. We calculated a cosine similarity between the center user and each verification user from their category feature vectors. While the ratio of good retweets indicates the degree of the matching of some categories, the cosine similarity shows the distance of interest of users. A cosine similarity between user $A$ and user $B$ is defined as below.

$$\frac{C_{\text{user}A} \cdot C_{\text{user}B}}{|C_{\text{user}A}| \times |C_{\text{user}B}|}$$

F. Result of verification

<table>
<thead>
<tr>
<th>Category</th>
<th>Center user</th>
<th>$A$</th>
<th>$V$</th>
<th>$Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of Electro-Communications</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Club</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Game</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Anime</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Programming</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Pokemon</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Categorization Failed</td>
<td>0</td>
<td>3</td>
<td>12</td>
<td></td>
</tr>
</tbody>
</table>

Table I shows the results of the categorization about the center user and two typical verification users, $A$ and $V$. $A$ and $V$ do retweeting about four and three categories respectively, and the numbers of each retweet are one to three. From that point, they seem equivalent, but are still different. In the categories about which the center user often retweets, such as ‘Twitter’ and ‘Club’, there is few retweet by $V$. In addition, in the category of categorization failed, the number of retweet by $A$ is fewer than $V$. These facts imply that $A$ rather than $V$ is suitable for a user to be followed by the center user. The difference among the verification users become clear by using other measure.

Table II summarizes all the results of the verification. In this verification, we regard the ratio of good retweets as the most import metric because we search users who often retweet about a center user’s interest. Table II shows that the highest users have higher scores in the ratio of good retweet and value of cosine similarity than ones of the lowest users. Thus it is clear that they often retweet about center user’s interest.

Z seems to be a potential user in terms of the ratio of good retweet the cosine similarity, $Z$ is still one of the lowest users because the number of his good retweet is only one. So, it is proper to judge comprehensively by three metrics.

Two of the lowest users $V$ and $Y$ retweet many times, but the categories of those retweets are not included in those of the center user. They may tweet about topics in various categories and might be good candidates as hub users for topics in wide areas. Because our goal is to find a personalized hub user who matches the specific interest of the center user, we regard those users as inappropriate ones. From these results, we can conclude that the highest users are hub users who retweet more often and these retweets match the center user’s interest.

VII. DISCUSSION

A. Informal retweet

In this paper, we build overlap graph based on formal retweets only and ignore informal retweets. Informal retweet has a good point: a user who retweets a tweet can explicitly indicate a source followee of the tweet in his retweet, for example, “RT @someuser: _INCLUDED”. But the limitation of the length of a tweet in Twitter (140 characters) forces the followers to trim the information added by their followees. Furthermore, it is difficult to collect all information of retweet propagation paths about one of retweet because informal retweet does not have information of all retweeting users. But, informal retweet tell us user’s opinion about content of retweet because user add comment about this. So, informal retweet could be useful for tweet content classification or relationship analysis between a user and the contents of his retweets. In this research, we use formal retweet because we focus on retweet propagation paths.

B. Visualizing graph

![Fig. 4. Focusing overlap graph.](image-url)

We discuss visualizing overlap graph here. The list of users who share retweets could be found by applying OR operation...
to the sets of retweet users for each retweet. By visualizing the graph, we can show two important things:

First, visualized graph makes the relationship of users intelligible. A list of retweet users is insufficient to grasp the relation to the center user. Visualizing retweet paths enables users to recognize stronger relationships between their followees.

Secondly, we can understand user’s behavior on twitter. Figure 4 shows a portion of overlap graph. There are several types of users on Figure 4: a user (like user B) who receives many retweets from many users, a user (like user C) who spreads retweets to many users. These unusual users are found by visualizing retweet path. It is likely that the former user actively retweets and the latter user has many followers by being retweeted. Thus we can comprehend much information about user by visualizing overlap graph.

C. Automatic categorization of tweet

In our verification, a center user categorized his retweet by himself. Automatic categorization of tweets is required to process more retweets. However, standard natural language processing is hard because sentences of a tweet is too short. It is necessary to analyze utilizing Twitter.

Categorizing tweet is also hard because the sentences of a tweet is shorter than normal articles and include slang words. In other hands, they have many metadata, which includes the information of the location user tweets. We might be able to categorize by considering not only content of tweet but also the state of timeline and some users who relate to tweet.

VIII. CONCLUSION

The purpose of this study is finding a user who propagates retweets interesting for another increases the benefit of Twitter. For this reasons, we propose retweet overlap graph to find such user. The graph is based on some retweet propagation path trees. We verify that users on overlap graph match the purpose of this research in three measures: (1) the number of retweet, (2) ratio of good retweet, (3) cosine similarity. The user who has the highest score in an overlap graph, also has high scores in the measures. Therefore, overlap graph represents users to be followed.

In this paper, we made a overlap graph based on the latest fifty retweets of a center user, so the potential user found in a graph implicitly includes all the topics of the retweets. The center user could pick up a part of the retweets to find other potential users who include the specified topics.

Therefore, user may want to find user who is interested in a portion of his interest. We think that it can find them by using overlap graph which is built from retweets that center user selects voluntarily.

Furthermore, we will create a retweet viewer based on collaborative filtering or retweet recommendation system could be developed because of the similarity between retweet and filtering.

REFERENCES


<table>
<thead>
<tr>
<th>User</th>
<th>Number of retweet</th>
<th>Number of good retweet</th>
<th>Ratio of good retweet</th>
<th>Cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>11</td>
<td>8</td>
<td>0.7273</td>
<td>0.0154</td>
</tr>
<tr>
<td>B</td>
<td>11</td>
<td>10</td>
<td>0.9090</td>
<td>0.0123</td>
</tr>
<tr>
<td>C</td>
<td>26</td>
<td>20</td>
<td>0.7692</td>
<td>0.0175</td>
</tr>
<tr>
<td>V</td>
<td>18</td>
<td>6</td>
<td>0.3333</td>
<td>0.0092</td>
</tr>
<tr>
<td>W</td>
<td>5</td>
<td>1</td>
<td>0.2000</td>
<td>0.0003</td>
</tr>
<tr>
<td>X</td>
<td>2</td>
<td>1</td>
<td>0.5000</td>
<td>0.0542</td>
</tr>
<tr>
<td>Y</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Z</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0386</td>
</tr>
</tbody>
</table>