Classification of Twitter Users Based on Following Relations

Takuya Yamashita, Haruhiko Sato, Satoshi Oyama, Masahito Kurihara

***Abstract*—Recently, Twitter which provides both social networking and micro-blogging services has become the focus increasing attention. Twitter has a function called follow which allows a user to subscribe to another user's information transmissions called tweets. Following is assumed to be based on a common interest or a shared attribute. In the following set (the set of Twitter accounts a user is following), if there are many different commonalities, we can consider that users are following each other based on a common attribute and that users in the set can be clustered on the basis of nature of the commonality.**

***Index Terms*—twitter, social networks, clustering, data analysis, social media.**

I. INTRODUCTION

Recently Twitter [1] which provides both social networking services (SNSs) and micro-blogging service has attracted increasing attention. The microblogging service allows users to share short passages (less than 140 characters) called tweets. The SNS service is provided by the follow function, enabling a user to subscribe to, or follow, other Twitter users' information transmissions(tweets).

Following is typically based on some commonality, such as Friends, common interests (Figure 1). In the following set, i.e. the set of Twitter accounts a user is following, if there are many different commonalities, we can consider that users are following each other based on a common attribute and that users in the set can be clustered on the basis of the nature of the commonality. For example, users who are fans of a particular football team may choose to follow each other.

A. Purpose

This study aims to investigate the assumption that users who are followed by a common user based on a common attribute are following each other and to analyze the sets of followers. In the future, we want to determine the human relationship between the following set and the user.

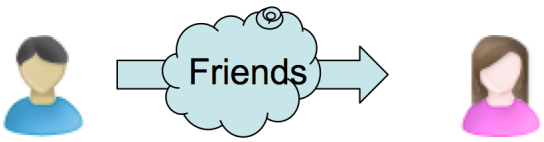

Fig. 1. Example of Follow

T. Yamashita, H. Sato, S. Oyama and M. Kurihara are with the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan 060-0814. E-mail:t_yamashita@complex.ist.hokudai.ac.jp, haru@complex.ist.hokudai.ac.jp, oyama@ist.hokudai.ac.jp, kurihara@ist.hokuda.ac.jp

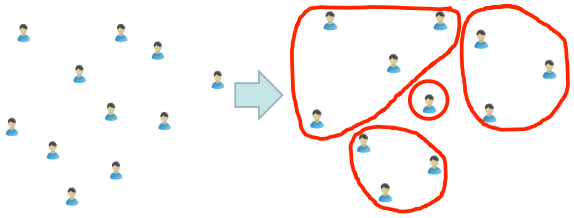

Fig. 2. Clusters of followers

1) Organizing Information: In the official Twitter interface, the list of followers is displayed in chronological order; the most recently selected follower is displayed first. However, a classification of the following set can reveal commonalities and could allow users to be displayed on the basis of those commonalities. Twitter has a list function that can be used to classify users by commonality. However, they are curated lists that must be set up and managed by users and as such are not useful for automatically determining commonalities among followers.

2) User Recommendation: In this study, we clusters users based on commonalities among followers and can recommend that cluster members follow each other's Twitter feeds.

B. Related Work

Recently, Twitter has been the subject of considerable research. For example, Kwak et al. used various data, including a large amount of user data, to explore aspects of Twitter, such as trending topics and retweets [2]. Other studies have reported the relation between microblogging service and social graphs [3][4], and have investigated how and why we use Twitter [5]. In addition, there has been considerable research into particular aspects of tweets, such as retweets [6], hyperlinks [7], use of hashtags [8], analysis of tweets with tags using Bayesian filtering [9].

Research has also been specifically conducted on Twitter users; for example, users' interest words [10] and user attributes in relation to the list function [11].

II. PROPOSAL METHOD

In this study, we treat following relations among users as a directed weighted graph. We define a user as a node and a following relation as an edge. From this, we assume that a subgraph of users based on a common relation is dense and we present a classification of the graph's density.

The system's input is a set of Twitter users and the output is the clustering of these users. Initially, the system gets input about a user's following relations and a following

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

set using the TwitterAPI [12]. We construct an adjacency matrix using following relations and a following set as a directed weighted graph. Using the adjacency matrix, we construct a similarity matrix which represents the strength of the users' connections. Finally, we perform clustering using the similarity matrix and the number of the clusters.

A. Constructing the Adjacency Matrix

An adjacency matrix, in which the rows and columns represent users, represents the existence of follow relation using 1, 0. If user i follows user j, the ith row and jth column are 1. If user i does not follow user j, the ith row and jth column are 0. The matrix element which represents the target user is 0.

B. Constructing the Similarity Matrix

The similarity matrix represents the strength of relationships among users. The rows and columns of this symmetrical matrix are users. Elements in this matrix range from 0 to 1, where 1 represents the strongest relationship and 0 represents the weakest relationship. If user i and user j are following each other, the ith row and jth column are 1. If only user i or user j is following the other user the ith row and jth column are 0.5. If no following relation exists between the ith and jth users, the ith row and jth column are 0. The number of user relationships represented in the matrix is directly proportional to the size of the matrix.

C. Spectral Clustering

In this study, we chose a spectral method for clustering. The inputs for the spectral method are the similarity matrix and the number of clusters, and the output is the clustering result. Herein, we employ a function which dimidiates the following relation set by density, i.e. most and least dense. To solve this evaluation function, we take advantage of the fact that this evaluation function's optimum solution corresponds to a certain eigenvalue solution. In addition, we use clustering to recursively solve this evaluation function.

We use Min-MaxCut($MCut$) as an evaluation function. Figure 3 shows the evaluation function when a graph dimidiates to A and B. $sim(a, b)$ represents similarity between nodes a and b.

$$W(A, B) = \sum_{a \in A, b \in B} sim(a, b)$$
$$W(A) = W(A, A)$$

Using the above, we write the MCut evaluation function as follows.

$$MCut = \frac{W(A, B)}{W(A)} + \frac{W(A, B)}{W(B)}$$

Typically spectral clustering is performed by the following processes. First, the eigenvalue problem of M shown below is solved.

$$M = I - D^{-1/2} W D^{-1/2}$$

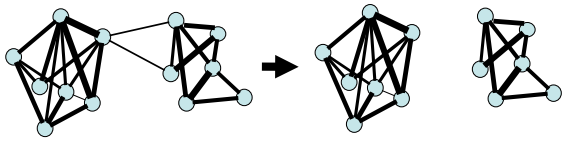

Fig. 3. An Example of A Step of Clustering

Here, W is a similarity matrix, $D = diag(We)$, and $e = (1, 1, \ldots, 1)$. Then, the eigenvector u_2 which corresponds to the second smallest eigenvalue of M is solved. Second, using $\hat{q} = D^{-1/2} u_2$, we get $\hat{q}$. Then, the elements of $\hat{q}$ are sorted and divided into two clusters using a threshold value. The elements that are larger than the threshold are divided to cluster one and those that are smaller than the threshold are divided to cluster two. In the proposed method, we can substitute this process by dividing $\hat{q}$'s elements based on whether they are positive or negative. From investigation the results of clustering and the evaluation function, we can obtain a proper division position.

III. EXPERIMENT

The experiment was performed by five participants as testers.

A. Purpose

The purpose of this experiments was to determine if testers' ideal classification of their following sets could be obtained. The result would enable us to assess the validity of our assumptions.

B. Experiment

First, the testers provided an ideal classification of their following sets. Next, the system outputs a range of possible clustering results. Then, by comparing the testers' ideal analyses with the clustering results, the system outputs the clustering result with the best comparative evaluation result.

We used $RandIndex$ value to evaluate the clustering results. We assume that A is the system's output and B is the tester's ideal. We tagged each pair of users to denote whether they were in a same cluster for each result. $RandIndex$ is the ratio of the tagged matches of A and B and is calculated by the following formula.

$$Rand\ Index = \frac{number\ of\ pairs\ tags\ match}{number\ of\ pairs\ of\ users\ in\ a\ result}$$

IV. RESULTS

Table I shows the experimental results. Accuracy of classification was approximate for most attributes such as same common interest or friends. In particular clustering for following and followed sets of limited size was accurately analyzed. Users with an extremely large number of followed or following users, such as entertainers or artists, were referred to as authority users.

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

TABLE I
RESULT

	tester 1	tester 2	tester 3	tester 4	tester 5
Size of following set	13	55	208	47	226
RandIndex value (system's output)	1.00	0.90	0.78	0.96	0.73
Number of clusters (system's output)	8	29	20	13	14
RandIndex value (number of clusters in ideal classification)	1.00	0.89	0.40	0.96	0.72
Number of clusters in ideal classification	8	9	10	13	16

A. Problem 1

One identified problem was that authority users were not distributed in a manner that correlated with the tester's ideal classification; therefore the system output a greater number of clusters than expected. Tester 2 is good example of this problem. In this tester's ideal classification, authority users were attributed to one or two clusters ('bot', 'entertainer', etc); however, the cluster's users are not actually following each other. This did not satisfy our supposition that users follow each other based on a commonality. For example, an artist and a comedian were grouped in a tester's ideal classification; however, they were not actually following each other.

B. Problem 2

Another problem is that authority users were mixed in a normal user's cluster. It is assumed that this problem developed while constructing the similarity matrix. The similarity matrix was constructed to equally treat the weight of following; however, this is not the case. For example, compare a following from a normal user to a normal user with that from an entertainer to a normal user. Generally, the entertainer is following an extremely small number of users and is not following an extremely large number of users. Therefore, under these conditions, the weight of a following will be different between a normal user and an entertainer. In future, we must examine how to construct a similarity matrix which considers the number of both follower and following users.

V. CONCLUSION

We proposed method to analyze and cluster sets of Twitter users based on a commonality or a shared attribute. We found the system was able to recommend clustering similar to a user's ideal classification. We found that the results, except for the case of authority users, support the supposition that users which are followed by the same user follow and that the following could be attributed to an identifiable commonality.

VI. FUTURE WORK

In future, we expect to narrow the number of analyzed clusters. In this paper, we regarded the output of the system as data with the highest evaluation value compared to tester's ideal classification. In addition, we will work to improve the accuracy of clustering considering the construct of the similarity matrix. We also intend to treat different classes of Twitter users, such as authority users, separately. Such users, who have very large number of followers, should be independently clustered. To complement the work, we will also implement a system for effective labeling of analyzed clusters, which will facilitate further identification of cluster attributes.

REFERENCES

[1] "Twitter." [Online]. Available: http://twitter.com
[2] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772751
[3] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: understanding microblogging usage and communities," in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ser. WebKDD/SNA-KDD '07. New York, NY, USA: ACM, 2007, pp. 56–65. [Online]. Available: http://doi.acm.org/10.1145/1348549.1348556
[4] B. A. Huberman, "Social networks that matter: Twitter under the microscope," *JEL*, vol. 1, p. 9, 2008. [Online]. Available: http://dx.doi.org/10.2139/ssrn.1313405
[5] D. Zhao and M. B. Rosson, "How and why people twitter: the role that micro-blogging plays in informal communication at work," in *Proceedings of the ACM 2009 international conference on Supporting group work*, ser. GROUP '09. New York, NY, USA: ACM, 2009, pp. 243–252. [Online]. Available: http://doi.acm.org/10.1145/1531674.1531710
[6] D. Boyd, S. Golder, and G. Lotan, "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter," in *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, vol. 0. Los Alamitos, CA, USA: IEEE, Jan. 2010, pp. 1–10. [Online]. Available: http://dx.doi.org/10.1109/HICSS.2010.412
[7] T. I. Mitsuo YOSHIDA and M. YAMAMOTO, "Analysis of tweets including urls on twitter." [Online]. Available: https://www.tulips.tsukuba.ac.jp/dspace/handle/2241/111107
[8] H. Takenaka, K. Komiya, and K. Yoshiyuki, "Hashtag classification of tweets in twitter using bayesian filtering," *IPSJ SIG Technical Report. IPSJ SIGFI Report.*, vol. 2011, no. 1, pp. 1–6, 2011-03-21. [Online]. Available: http://ci.nii.ac.jp/naid/110008583688/
[9] Y. Kuroki, T. Ohishi, R. Hasegawa, H. Fujita, M. Koshimura, and Y. Tashiro, "Twitter 発言の時系列解析に基づくハッシュタグの内容説明," *Proceedings of the 73th Annual Convention IPS Japan*, vol. 2011, no. 1, pp. 695–697, 2011-03-02. [Online]. Available: http://ci.nii.ac.jp/naid/110008600988/
[10] J. Saito and T. Yukawa, "Interest extraction and user recommendation based on social bookmark," *IPSJ SIG Technical Report. IPSJ SIGFI Report.*, vol. 2011, no. 2, pp. 1–8, 2011-03-21. [Online]. Available: http://ci.nii.ac.jp/naid/110008583689/
[11] T. Okugawa, T. Ohishi, R. Hasegawa, H. Fujita, M. Koshimura, and K. Kurakado, "Twitter のリスト機能を用いたユーザの特徴抽出," *Proceedings of the 73th Annual Convention IPS Japan*, vol. 2011, no. 1, pp. 687–689, 2011-03-02. [Online]. Available: http://ci.nii.ac.jp/naid/110008600984/
[12] "Twitter api." [Online]. Available: http://apiwiki.twitter.com/Twitter-API-Documentation

ISBN: 978-988-19251-8-3
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)