A-033

Extraction and Annotation of Personal Cliques from Social Networks

マイケ・エルドマン† 武吉 朋也† 帆足 啓一郎† 小野 智弘†
Maike Erdmann Tomoya Takeyoshi Keiichiro Hoashi Chihiro Ono

1. Introduction

In microblogging systems such as Twitter, users can choose whose posts they want to read by "following" other people's accounts. Twitter users often have large social networks, including connections to e.g. family members, friends and coworkers, thus many of them are overwhelmed with managing their network connections and dealing with information overload.

We want to address this problem by dividing the social network of a Twitter user into personal cliques and annotating each clique with keywords representing common interests of its members. Our proposed method categorizes Twitter accounts into news sources and average users and weights keywords extracted from news sources higher in order to improve accuracy of clique annotation.

Through clique extraction and annotation, we want to enable Twitter users to create and manage lists of followers and followees without effort in order to use these lists to control information flow.

2. Related Work

Some tools (e.g. TweetDeck [1]) are available that allow the manual categorization of a user's followees into cliques, but this can be difficult and time-consuming for large networks. Other tools (e.g. MentionMap [2], Twitter Browser [3]) automatically extract and visualize the social network of a Twitter user, but extract only selected "followers" and "followees" (people a user is following) and don't have the functionality to extract cliques.

An application named NodeXL [4] can automatically import the social network of a Twitter user and divide the network into cliques using the Clauset-Newman-Moore (CNM) algorithm [5,6], a standard algorithm in the field of network analysis. NodeXL can also visualize graphs with e.g. the Harel-Koren Fast Multiscale algorithm [7].

Li et al. [8] have proposed a method for extracting keywords from social networking services based on term frequency–inverse document frequency (tf-idf). Unfortunately, the usage of tf-idf for keyword extraction faces the problem that, as Li et al. themselves state, information posted on microblogging services tends to be "short and informal" and is therefore difficult to analyze. Besides, user posts contain a lot of smalltalk, causing noise that is difficult to filter out with tf-idf.

3. Proposed Method

In this paper, we propose a method for automatically extracting and annotating personal cliques from Twitter. An overview of the system is shown in Figure 1.

In the first step, the social network of a Twitter user is constructed by extracting all followees, and divided into cliques with the help of the CNM algorithm.

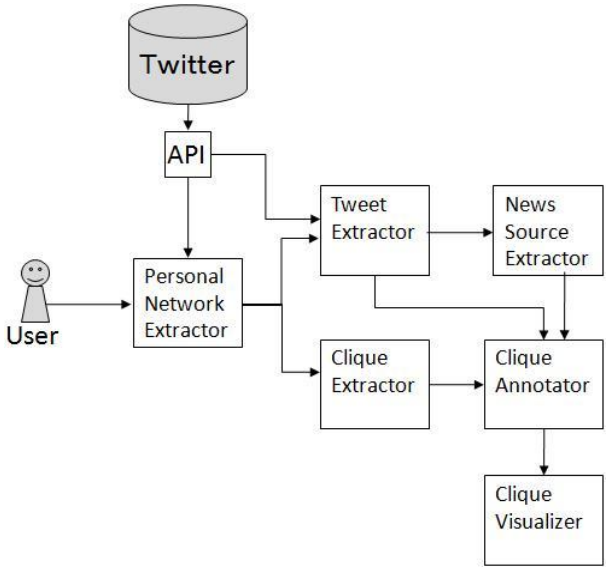

Fig. 1 System Overview

In the next step, we collect tweets of each member in a clique and identify keywords by calculating tf-idf. We distinguish "news sources" (users posting mainly news and information of general interest) in each clique from "average users" (users posting mainly smalltalk and social information). Using a simple heuristic, a user is categorized as news source if one of the following two criteria is fulfilled:

(1) number of followers / number of followees ≧ 2
(2) number of tweets ≧ 10,000

We weight keywords extracted from news sources higher than keywords extracted from average users, since news sources tend to post less "noise" than average users.

In the last step, labels are assigned to each clique based on the extracted keywords, enabling automatic annotation of each clique according to common interests of its users. Then, the annotated cliques are visualized in a graph structure using the Harel-Koren Fast Multiscale algorithm. For the visualization, we place the target user into an artificially created empty clique and visualize only edges between cliques, weighted by the number of individual user connections.

4. Experiment

In order to evaluate the proposed method, we extracted the 1.5-hop followee network for five test subject and collected the 200 latest tweets for each Twitter account in the personal network of the test subjects. We asked the test subjects to give feedback on the clique extraction results, and further divided

† KDDI 研究所　KDDI R&D Laboratories

cliques into sub-cliques in cases where test subjects regarded the original cliques as too large.

After that, we manually categorized about 1,500 Twitter accounts into news sources and average users and used that data to create the heuristic for news source discrimination.

For each user and each clique, we extracted three different sets of keywords:

(1) Baseline: Keywords from news sources are not weighted
(2) News sources de-emphasized: Keywords from news sources are ignored
(3) News sources emphasized: Keywords from news sources are weighted twice as high as average users

The cliques as well as the top 10 extracted keywords of each extraction method were presented to all test subjects and they were asked to categorize each keyword into “good”, “neutral” and “bad”, where “good” had to be assigned to keywords being relevant to the interests of members of a clique and “bad” was assigned to irrelevant keywords.

Table 1 shows example keywords extracted for a clique of NLP researchers. Some of the keywords, such as “言語処理” and “機械学習” adequately represent common interests of members of that clique. On the other hand, terms such as “計画停電” were extracted, which are recent tweet topics but do not represent the general interests of that clique. We want to avoid the extraction of temporary trending topics by extracting the whole history of tweets of each user. Another problem is that some meaningless terms such as “とのこ” were extracted as a result of parsing errors. We need to improve the text parser to be able to extract keywords from tweets, which consists of mostly short and informally written text.

Table 2 gives an overview of the experimental results. The best results were achieved for the method emphasizing news sources by weighting keywords extracted from them twice as high as keywords extracted from average users. 36.52% of all extracted keywords were assigned the label “good” and 70.31% were assigned “good” or “neutral”.

However, since the results of our proposed method do not differ significantly from those of the baseline method, we evaluated the results more thoroughly and realized that for some cliques, news sources represented common interests among others whereas for others, news sources represented only the interests of single users. For instance, the news source “googleresearch” represents interests of a clique of IT research related users, whereas the news source “officialtepco” does not represent the interests of a clique of users sharing the same hobby. Therefore, instead of emphasizing all news sources equally, we need to pay more attention to determining whether a news source represents common interests of a clique and emphasize it accordingly.

5. Conclusion

In this paper, we proposed a method for automatically extracting and annotating personal cliques from social networks. Our method distinguishes news sources from average users in order to emphasize keywords extracted from news sources. In a preliminary experiment, we showed that this discrimination helps improve the accuracy of clique annotation.

Table 1: Examples of Extracted Keywords

Keyword	“good”	“neutral”	“bad”
言語処理	○		
とのこ			○
機械学習	○		
招待講演		○	
コーパス	○		
計画停電			○
機械学習入門	○		
言語処理学会	○		
拡散希望			○
クラスタリング	○		

Table 2: Comparison of Keyword Extraction Methods

Extraction Method	“good”	“good” or “neutral”
Baseline	35.62%	67.12%
News source de-emphasized	28.14%	57.41%
News source emphasized	36.52%	70.31%

In order to improve our proposed method, we want to find a way to determine whether a news source represents common interests of a clique and de-emphasize news sources representing interests of only single users. We also want to improve keyword extraction by adjusting the text parser to the characteristics of tweet texts. Finally, we will conduct a larger experiment with more test subjects and analyze more tweets per user.

Our goal is to enables microbloggers to create and manage lists of followers and followees automatically and use that information to keep track of their social network connections. Clique annotation also helps users to understand the common interests of others in his social network and join their communication successfully.

Using the proposed method as a foundation, applications for sender-side or client-side information filtering can also be implemented in order to help users deal with information overload. Besides, new followees can be suggested to a user if they are present in one of his cliques.

References

[1] TweetDeck, http://www.tweetdeck.com/
[2] MentionMap, http://apps.asterisq.com/mentionmap/
[3] Twitter Browser, http://www.neuroproductions.be/twitter_friends _network _browser/
[4] M. Smith, N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec and C. Dunne, "NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010", http://nodexl.codeplex.com/ from the Social Media Research Foundation, http://www.smrfoundation.org (2010)
[5] M. E. J. Newman and M. Girvan, “Finding and evaluating community structure in networks”, Physical Review E, Vol. 69, No. 2 (2004)
[6] A. Clauset, M. E. J. Newman and C. Moore , “Finding community structure in very large networks”, Physical Review E, Vol. 70, No. 6 (2004)
[7] D. Harel and Y. Koren, “A Fast Multi-scale Method for Drawing Large Graphs”, Proceedings of the 8th International Symposium on Graph Drawing, pages 183-196 (2001)
[8] Z. Li, D. Zhou, Y. Juan and J. Han, “Keyword extraction for social snippets”, Proceedings of the 19th international conference on World Wide Web, pages 1143-1144 (2010)