Electronic Preprint for Journal of Information Processing Vol.23 No.3

Regular Paper

Assigning Keywords to Automatically Extracted Personal Cliques from Social Networks

Maike Erdmann[1,a)] Tomoya Takeyoshi[2,b)] Kazushi Ikeda[1,c)] Gen Hattori[1,d)] Chihiro Ono[1,e)] Yasuhiro Takishima[1,f)]

Received: August 18, 2014, Accepted: February 4, 2015

Abstract: In Twitter and other microblogging services, users often have large social networks formed around cliques (communities) such as friends, coworkers or former classmates. However, the membership of each user in multiple cliques makes it difficult to process information and interact with other clique members. We address this problem by automatically dividing the social network of a Twitter user into personal cliques and assigning keywords to each clique to identify the common ground of its members. In this way, the user can understand the structure of their social network and interact with the members of each clique independently. Our proposed method improves clique annotation by not only extracting keywords from the tweet history of the clique members, but individually weighting the extracted keywords of each member according to the relevance of their tweets for the clique. The keyword weight is influenced by two factors. The first factor is calculated based on the number of connections of a user within the clique, and the second factor depends on whether the user publishes personal information or information of general interest. We developed the prototype of a Twitter client with clique management functionality and conducted an experiment in which on average 46.96% of the keywords extracted from our proposed method were relevant for the cliques as opposed to 38.31% for the baseline method.

Keywords: news sources, keyword extraction, clique annotation, social networks

1. Introduction

In microblogging services such as Twitter, users can choose whose posts they want to read by "following" other user accounts. Twitter users often have large social networks formed around cliques (communities) such as friends, coworkers or former classmates. However, the membership of each user to multiple cliques makes it difficult to process information and interact with other clique members, thus many Twitter users are overwhelmed with managing their network connections and dealing with information overload.

We address this problem by automatically dividing the social network of a Twitter user into personal cliques and assigning keywords to each clique in order to identify the common ground of its members. We define a clique as a "small group of people, with shared interests or other features in common" as described in the Oxford American English Dictionary [*1]. Unlike the clique definition in graph theory, a member of a clique does not have to be connected to all other members of the clique. However, the members of a clique can be assumed to be closely interconnected. The automatically extracted and annotated cliques help Twitter users to understand the structure of their social network and interact with the members of each clique independently.

Our proposed clique annotation method extracts keywords from the tweet history of the clique members based on term frequency, and individually weights the extracted keywords of each clique member according to the relevance of their tweets for the clique. The keyword weight is influenced by two factors. The first factor is calculated based on the number of connections of a user within the clique, since users with many connections often post information which is more relevant to the clique than users with few connections. The second factor depends on the type of user. Users who post information of general interest are given priority over users who post personal information.

The rest of the paper is structured as follows. First, we present related research in Section 2 and explain our proposed method in Section 3. In Section 4, we introduce the Twitter client with clique management functionality that we have developed, and discuss the results of our experiment. Finally, we draw a conclusion and outline future work in Section 5.

2. Related Work

In this section, we introduce related research for clique extraction and clique annotation, since they form the basis of our proposed method. Clique extraction denotes the process of dividing the social network of a user into small groups of users that have some kind of relationship, e.g., friends, coworkers or former

1 KDDI R&D Laboratories, Fujimino, Saitama 356–8502, Japan
2 KDDI, Chiyoda, Tokyo 102–8460, Japan
a) ma-erdmann@kddilabs.jp
b) to-takeyoshi@kddi.com
c) kz-ikeda@kddilabs.jp
d) gen@kddilabs.jp
e) ono@kddilabs.jp
f) takisima@kddilabs.jp

*1 http://www.oxforddictionaries.com

classmates. Clique annotation denotes the selection of representative keywords for each extracted clique, in order to understand the common ground of the clique members.

2.1 Clique Extraction

In Twitter, users can divide their social network into cliques (called lists) manually, and use tools such as TweetDeck *2 to display tweets of selected cliques. However, the manual creation of cliques is difficult and time-consuming.

The application NodeXL *3 [10] can import the social network of a Twitter user and divide the network into cliques using the Clauset-Newman-Moore (CNM) algorithm [7], [8], a standard algorithm in the field of network analysis. However, it is difficult for a user to understand the composition of their cliques unless they are annotated, e.g., with representative keywords.

2.2 Clique Annotation

In order to annotate the extracted cliques, keywords representing the common ground of the clique members have to be collected from the tweets. This is a challenging task, since information posted on microblogging services tends to be short and informal and is therefore difficult to analyze.

In our previous research [2], we have applied df-idf [9] for extracting keywords from social networks. Li et al. [4] have proposed a similar method based on tf-idf [3] and other linguistic features. Wu et al. [11] use TextRank [6] to annotate the tweets of Twitter users with keywords. Kim et al. [1] annotate manually extracted cliques of Twitter users based on χ^2 feature selection [5].

However, the difference in accuracy of the listed keyword extraction methods does not differs substantially (e.g., in Ref. [4], the precision increased by only 1.27% for the top 5 keywords), and all of them face the problem that user posts contain a lot of irrelevant small talk, causing noise that is difficult to filter out. For this reason, our proposed method applies weights to the extracted keywords in order to filter out keywords from users who post information that is irrelevant for their clique.

2.3 Clique Extraction and Annotation

In Ref. [2], we have attempted to combine clique extraction with clique annotation. However, the majority of the automatically extracted cliques were not useful for the test subjects, thus we added a function to adjust the cliques manually. Besides, many “noisy” keywords were found in the clique annotations. Therefore, we refined the clique annotation method in this research.

3. Proposed Method

In this section, we describe our method for automatically extracting and annotating personal cliques of Twitter users and introduce a clique management function that resolves conflicts between automatic clique extraction and manual clique modification.

3.1 Clique Extraction

In the first step, the 1.5-hop social network of a Twitter user is constructed from their social network and divided into cliques with the help of the CNM algorithm [7], [8]. A 1.5-hop network includes the direct neighbors of a node and the edges among them. The difference between a 1.5-hop network and other n-hop networks is visualized in **Fig. 1**. The dark nodes and edges in the figure represent those that are included in the n-hop network.

The connections among direct neighbors are essential to divide the social network into cliques. The nodes in 2-hop range are excluded, since they would affect the performance of clique extraction in a negative way by causing one clique to be formed around each node in 1-hop range.

The CNM algorithm starts with regarding each node in a network as a separate clique, and iteratively merges cliques that are closely connected. The process stops if no more cliques can be merged without negatively affecting the network modularity score. Since the targeted Twitter user should be located in the center of the social network, we remove them from their originally assigned clique and place them into a new, empty clique. The concept of a social network divided into cliques is visualized in **Fig. 2**.

3.2 Clique Annotation

In this subsection, we first give an overview on the clique annotation process and then explain the keyword extraction process in detail.

In order to annotate an extracted clique, we analyze the tweet history of all clique members. We could simply combine the tweets of all clique members in one document and extract important keywords from that document, but this distorts the results, since some clique members tweet more frequently than others. Instead, we extract keywords for each clique member individually, merge them by adding up their values and rank them by their values, as shown in **Fig. 3**.

The keyword extraction process of Fig. 3 is visualized in **Fig. 4**. We identify important keywords in the tweets of a clique mem-

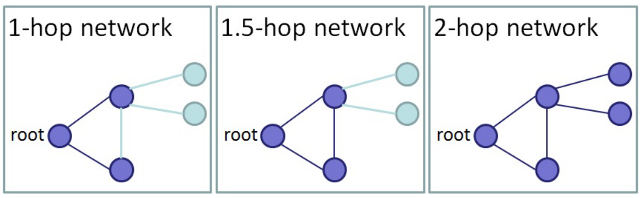

Fig. 1 Visualization of n-hop networks.

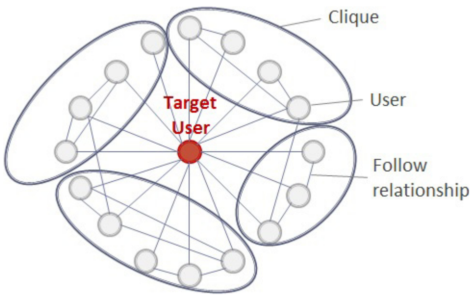

Fig. 2 Clique extraction.

*2 http://www.tweetdeck.com
*3 http://nodexl.codeplex.com

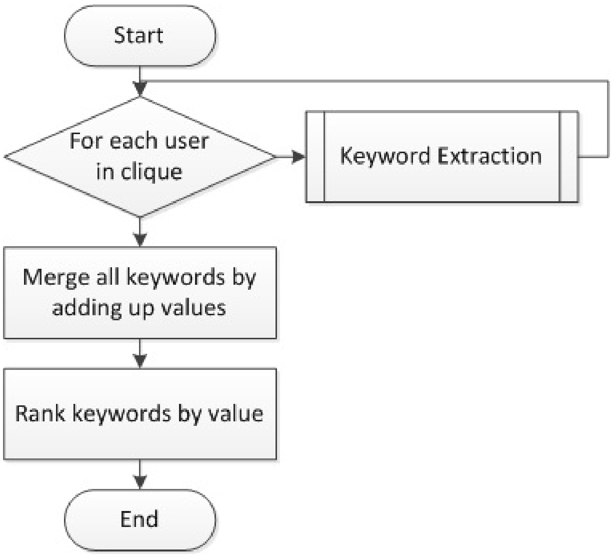

Fig. 3 Clique annotation process.

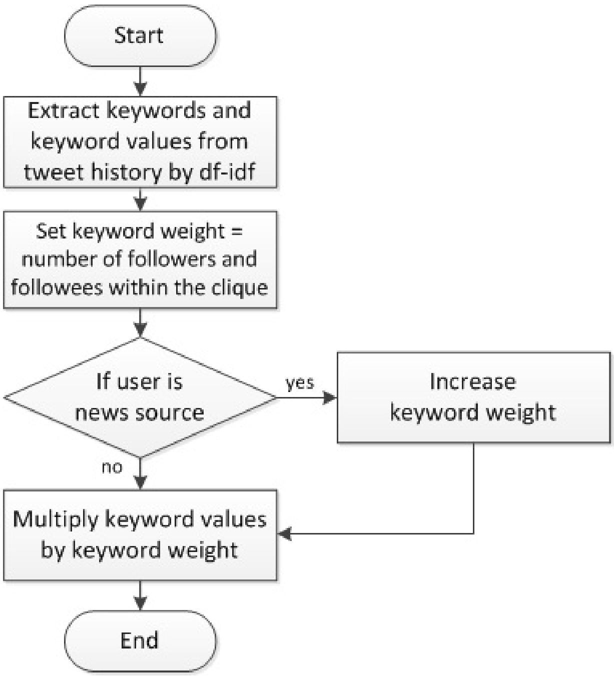

Fig. 4 Keyword extraction process.

ber by using df-idf [9]. As opposed to tf-idf, df-idf calculates the document frequency, i.e., it measures how often a term occurs in a local document set in comparison to a global document set, ignoring how often a term occurs in a single document. In our case, the local document set consists of all tweets of a user, with one tweet being one document. For the global document set, we collect the tweets of randomly selected Twitter accounts.

Whereas the difference of tf-idf and df-idf does not appear critical, a preliminary experiment with keywords extracted from the tweets of five test subjects have indicated that df-idf is more suitable for our application (increase of precision by about 2%), since the term frequency tf is usually not higher than 1 in short texts such as tweets.

In order to keep the keyword extraction method simple, we decided not to use other linguistic features, TextRank or χ^2 feature

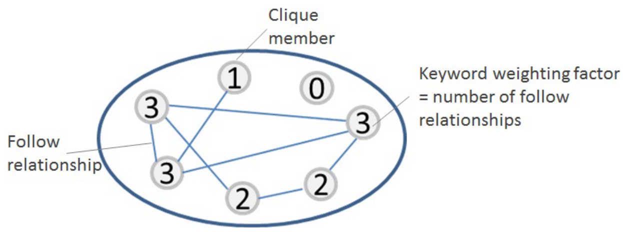

Fig. 5 Weighting of clique members.

Tweet examples of news sources
- Accel Leads $52.5M Round In Cloud-Based Data Storage And Backup Company Code 42 http://tcrn.ch/zk2TAW
- Crowds cheer as pro-democracy leader Aung San Suu Kyi applies to stand for election to #Burma's parliament. More soon http://bbc.in/AAdIv4
- Watch @drosennhl's top season games on ?#NHLNetwork US?. We've got the 1/7/12 VAN/BOS @ 4PM ET on 7/14.
- Kate Hudson shares her no-fuss skincare routine & fave products + more http://wwwear.it/Mede3d

Tweet examples of average users
- can't sleep, too exited.
- wow i had to see that didnt i ?
- Back on twitter after a long time .... Hmm
- waking up with a migraine <<<

Fig. 6 Features used for classifier.

selection. As shown in the respective papers, the difference in accuracy is not critically high. However, it is possible to replace the keyword extraction algorithm without losing the benefit of our proposed keyword weighting method described in the following.

After extracting the keywords of a user and calculating their df-idf scores, the keyword values are adjusted based on the characteristics of the user. We realized that keywords extracted from central clique members, i.e., users with many followers and followees within the clique, tend to represent common tweet topics of a clique more accurately. Keywords from marginal clique members tend to be less representative. For that reason, we decided to multiply the keyword weight by the number of connections a user has within the clique, as illustrated in **Fig. 5**. For instance, if a clique member is connected to two other clique members, their weight is set to 2. If a clique member has no clique connections at all, i.e., it is connected only to the target user, their weight is set to 0, which means that all keywords are ignored.

By analyzing the number of connections within the clique, we can emphasize the keywords extracted from central clique members, since they are expected to represent the common ground more adequately than marginal clique members.

Apart from that, we distinguish "news sources," users who are posting mainly news and information of general interest, in each clique from "average users," users who are posting mainly personal information. 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. A few representative example tweets are listed in **Fig. 6**.

In our previous research [2], we distinguished news sources from average users based on the number of tweets and the ratio of followers and followees, since news sources tend to tweet much more frequently than average users and most news sources are followed by more users than they follow themselves. A fol-

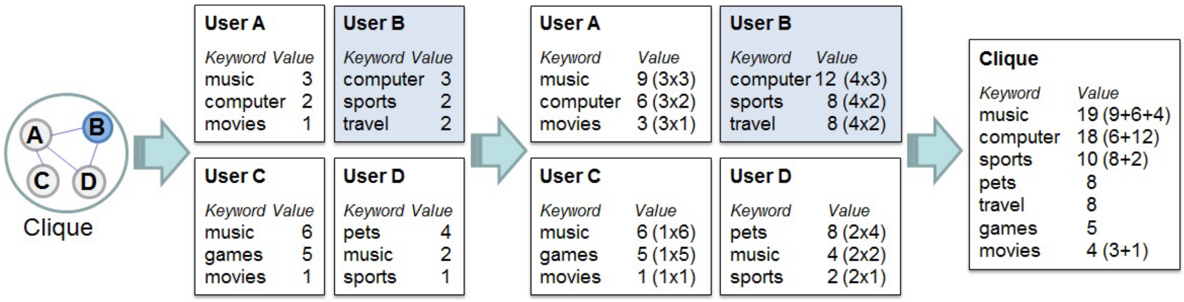

Fig. 7 Merging of keywords from different clique members.

lowee is somebody who is being followed by the user, whereas a follower is somebody who themselves is following the user.

This time, we conducted a small experiment in which 1,500 Twitter accounts were manually categorized into news sources and average users. When analyzing the characteristics of the two types of accounts, we noticed that news sources usually have more followers and followees than average users and post long tweets containing many hyperlinks and hashtags. Moreover, the profiles of news sources often contain keywords such as "official," "news," "corporation," "information" or "introduction." Based on those observations, we created a simple heuristic consisting of eight criteria. A user is categorized as a news source if at least half of the following eight criteria are fulfilled:

(1) Number of followers / number of followees ≥ 5
(2) Number of followers ≥ 100
(3) Number of followees ≥ 65
(4) Number of tweets ≥ 500
(5) Avg. tweet length ≥ 70 characters
(6) Avg. number of hyperlinks in tweets ≥ 0.25
(7) Avg. number of hashtags in tweets ≥ 0.05
(8) More than one characteristic keyword in user profile

In our experiment, the news source discrimination algorithm based on eight criteria was able to detect news sources with a precision of 83.5% and a recall of 54.4%. The previous method using two criteria only achieved a precision of 60.7% and a recall of 28.6%.

Figure 7 illustrates how keywords are extracted and ranked in the following steps.

- Extract keywords for all clique members (users A-D) separately and calculate the df-idf values.
- Weight extracted keywords based on the number of connections of the users within the clique (e.g., keywords of user A are multiplied by 3).
- Double the weight of all keywords extracted from news sources (user B).
- Merge the keywords and ranke them by their values in order to annotate the clique using the keywords with the highest values.

3.3 Clique Management

It can be expected that in automatically created cliques, some of the clique members will be misplaced, making it necessary to modify the clique structure manually. However, when the social network of a Twitter user changes and automatic clique extraction has to be reconducted, the manual modifications should not be discarded. Therefore, we propose two simple conflict resolution rules exemplified in **Fig. 8**.

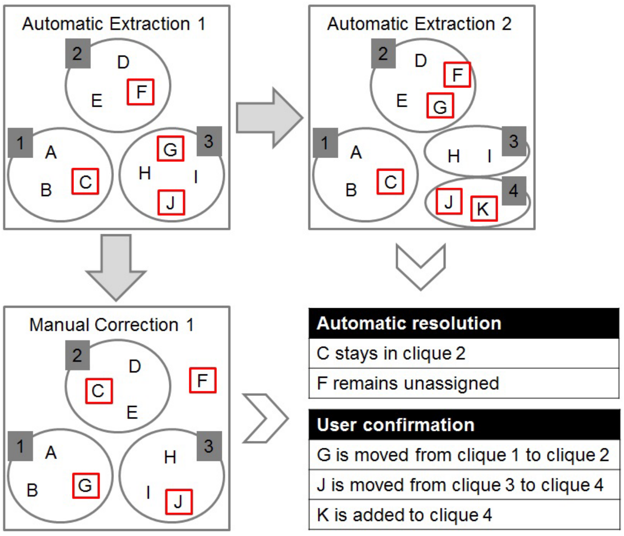

Fig. 8 Clique management.

Automatic resolution

If a clique member has been moved manually, it cannot be placed back into the original clique through automatic clique extraction. Examples in Fig. 8 are the clique members C and F.

User confirmation

If a clique member has been assigned to a clique that is different from both the old automatically assigned clique and the manually assigned clique, the user should confirm whether the clique member should be transferred to the new clique or stay in the manually assigned clique. In Fig. 8, the user needs to decide whether clique member G should be moved to Clique 2, whether J is moved to Clique 4 and whether K is newly added to Clique 4.

In order to judge whether the new and previously assigned cliques are the same clique, the degree of co-occurrence of Twitter accounts is calculated using the Jaccard Index, and the cliques with the highest scores are matched.

4. Experiment

In this section, we introduce the Twitter client with clique management function that we have developed, and discuss an experiment in which we extracted and annotated cliques from the social networks of 40 Twitter users.

4.1 Twitter Client

We developed the prototype of a Twitter client that automati-

Fig. 9 Twitter client.

Table 1 Clique extraction statistics.

	Cliques	Very useful	Useful	Not so useful	Not useful	Unrelated users	Missing users
before modification	7.75	27.8%	39.6%	13.54%	16.93%	18.31	11.72
after modification	6.95	52.37%	31.61%	6.15%	3.41%	2.17	0.45

cally divides the 1.5-hop range social network of a Twitter user into cliques and assigns keywords to each clique based on the method proposed in Section 3. If the user is not satisfied with the results, they can manually adjust the clique structure by adding new cliques, merging ord deleting existing cliques or moving individual users into a different clique. It is also possible to assign a clique name manually. When the user is satisfied with the clique structure, they can use it to filter the timeline by showing only the tweets sent from members of the selected clique. Apart from that, the user can send direct messages to all members of the selected clique at once. The application runs on iOS and Android devices.

Four screenshots of the clique management function of the implemented Twitter client are shown in **Fig. 9**. The screenshot on the left side shows the tweet timeline in which tweets of all cliques are displayed with their corresponding clique names. The second screenshot shows the clique management main screen where each clique is displayed with its top ranked keywords and icons of selected clique members. The third screenshot displays the information of a selected clique. In the example, the clique name was manually changed from “Clique 2” to “Bandai.” Beneath the clique title, a list of keywords and icons of some clique members are displayed. The fourth screenshot shows the user modification screen, where selected users can be deleted or moved to different cliques.

4.2 Clique Extraction

For 20 Twitter users, we extracted the 1.5-hop social network and collected the complete tweet history (most tweets were written in Japanese) of all Twitter accounts in the network. The 20 test subjects were then given the option to further divide the cliques into subcliques if they regarded the original cliques as too coarse-grained.

Table 1 shows statistics of the clique extraction experiment. On average 7.75 cliques and subcliques were created per test subject. Then, the test subjects were asked to add, merge or delete cliques as well as move individual clique members in order to improve the extracted clique structure. After the manual modification of the clique structure by the test subjects, on average 6.95 of cliques remained per test subject, of which 52,37% were labeled as “very useful” and 31.61% as “useful” by the test subjects. Besides, the number of unrelated users accidentally placed in a clique was reduced to 2.17 and the number of users missing from a clique was reduced to 0.45 per clique.

Table 2 Examples of extracted cliques.

Type	Examples
affiliation-based cliques	coworkers, former classmates, family, academic relationships, friends
interest-based cliques	news, IT, soccer, music, data mining, fashion, baseball, fun
other cliques	unknown, miscellaneous, inactive accounts

The test subjects added 1.05 cliques, merged 1.75 cliques, deleted 0.35 cliques and moved 40.05 clique members on average. Considering that the test subjects had on average 151 followers and 155 followees in their personal social network, our approach can be considered significantly less time consuming than the entirely manual construction of cliques.

Overall, the test subjects were satisfied with the clique extraction results, because the users who could not successfully be categorized into cliques were often not considered worth categorization, as we have found out in a previous experiment [2].

Table 2 shows some typical examples of extracted cliques. Overall, about half of the cliques extracted in the experiment were “interest-based cliques,” consisting of users who have a hobby or other interest in common, such as “baseball.” The rest were “affiliation-based cliques,” consisting of users affiliated, in the past or present, with the same organization, such as a “company,” as well as a few “other cliques” that did not fit in any of the two categories.

4.3 Clique Annotation

After extracting cliques from the social network of each test subject, we annotated the successfully extracted cliques with keywords representing the common ground of the clique members.

For the global document set, which is needed for calculating the df-idf scores, we collected more than 5 million tweets.

Then, we extracted two different sets of keywords:

- NS2: News sources identified using 2 criteria (baseline)
- NS8: News sources identified using 8 criteria (proposal)

In both methods, the keywords were weighted according to the number of clique connections. In addition, the weight of news sources was doubled. We did not evaluate clique annotation using no keyword weights or keyword weights based on only the clique connections, since our previous experiment has already demonstrated that the performance of those two methods is low [2].

The cliques together with the top 10 extracted keywords of each extraction method were presented to the test subjects and they were asked to categorize each keyword into "very useful," "useful," "not so useful" and "not useful," where "very useful" was assigned to keywords that were highly relevant for the clique. The label "not useful" was assigned to keywords for which no relevance at all could be recognized.

Tables 3 and **4** show example keywords (translated from Japanese) extracted by the proposed method (NS8) for a clique of machine learning and text mining related Twitter accounts (interest-based clique) and a clique of accounts of former university friends (affiliation-based clique). For the machine learning related clique, some of the keywords, such as "machine learning" or "language processing," are helpful understand the common ground of the clique members. On the other hand, too general terms, such as "poster introduction," and too specific terms, such as "SAS user meeting," were also extracted. Sometimes, terms were extracted that were unknown to the test subject, such as "CCC." For the clique of former university friends, many of the extracted keywords, such as "MSRA (Microsoft Research Asia), "SOM (self organizing map)" and "cat cafe" represent only the interests of a few members of the clique, thus were too specific to be useful for annotation.

Tables 5 and **6** give an overview of the experimental results for the top 10 and top 5 keywords respectively. The best keywords were extracted for interest-based cliques, such as "baseball" or "fashion." This was not surprising, since users in such cliques tweet about similar topics. For the proposed method using eight criteria to indicate whether an account is a news source, 51.69% of the top 10 keywords and 53.37% of the top 5 keywords were considered "suitable" by the test subject, a significantly larger percentage than for the baseline method using only two criteria.

For affiliation-based cliques, the performance of both clique annotation methods was significantly lower than for the interest-based cliques, since people who e.g., once belonged to the same high school usually have few common interests, which is also reflected in their tweets.

Overall, our proposed method extracted 46.96% of suitable keywords for the top 10 and 45.33% for the top 5 keywords, clearly better results than the baseline. This improvement (p value < 0.01) is considered statistically significant.

We also asked the test subjects to state their motivation for labeling keywords as "not so useful" and "useful" and found out that the keywords categorized as "not so useful" were often too general or too specific terms, whereas the keywords categorized as "not useful" were often completely unrelated or unknown terms.

Table 3 Clique annotation example (machine learning related).

Keyword	Very useful	Useful	Not so useful	Not useful
SAS user meeting			◦	
machine learning	◦			
poster introduction			◦	
mining		◦		
SPM		◦		
data scientist		◦		
language processing		◦		
CCC				◦
research jokes		◦		
HPC		◦		

Table 4 Clique annotation example (former university friends).

Keyword	Very useful	Useful	Not so useful	Not useful
seminars, etc.			◦	
MSRA		◦		
SOM		◦		
cat cafe			◦	
WIRE			◦	
clustering		◦		
certain documents				◦
overseas guest				◦
manuscript check		◦		
SDK			◦	

4.4 Clique Categories

It seems promising to distinguish interest-based and affiliation-based cliques automatically, in order to develop clique annotation methods tailored to the specific characteristics of each type. Since affiliation-based cliques cannot be represented adequately with keywords extracted from tweets, other means of clique annotation, e.g., using profile or location information of clique mem-

Table 5 Clique annotation statistics (Top 10).

Method	Clique Type	Suitable Keyword	Unsuitable Keyword
NS2 (baseline)	affiliation	33.00%	67.00%
	interest	42.24%	57.76%
	all	**38.31%**	**61.69%**
NS8 (proposal)	affiliation	42.35%	57.65%
	interest	51.69%	48.31%
	all	**46.96%**	**53.03%**

Table 6 Clique annotation statistics (Top 5).

Method	Clique Type	Suitable Keyword	Unsuitable Keyword
NS2 (baseline)	affiliation	37.13%	62.87%
	interest	43.15%	56.85%
	all	**41.29%**	**58.71%**
NS8 (proposal)	affiliation	44.81%	55.19%
	interest	53.37%	46.63%
	all	**45.33%**	**54.69%**

bers, should be explored as well.

In order to distinguish interest-based and affiliation-based cliques automatically, we can analyze the absolute values of keywords extracted from tweets. Since users in affiliation-based cliques do not share many common interests, the overlap of extracted keywords is low, resulting in smaller keyword values. In a small experiment, we have confirmed this assumption. By setting a simple threshold based on the keyword value of the top ranked keyword, we were able to distinguish the two types of cliques with an accuracy of 75%. With a more sophisticated approach, it should be possible to distinguish interest-based and affiliation-based cliques with high accuracy.

4.5 Overall User Evaluation

In order to further evaluate the usefulness of our proposed clique annotation and evaluation method, we asked 40 experienced Twitter users, including the 20 test subjects in the first part of the experiment, to fill out a questionnaire about the usefulness of our implemented Twitter client. The 40 test subjects were between 20 and 50 years old and the majority (75%) of them was male. Unfortunately, three of the test subjects were not able to use the Twitter client because their personal social networks were exceptionally large (more than 2,000 followers and followees) causing the software to malfunction. The remaining 37 Twitter users responded to the questionnaire as follows.

All test subjects were interested in using automatically created lists. 32 test subjects stated that they want to use interest-based cliques and 29 users want to use affiliation-based cliques. To our surprise, some of the test subjects want to use automatically created cliques for other purposes as well, such as separating news sources from average users, private from business accounts or bots from human accounts. The majority of test subjects want to use cliques for sorting tweets in the timeline, whereas only 8 test subjects are interested in the option of sending direct messages to the members of a selected clique.

5. Conclusion

In this paper, we proposed a method for automatically extracting and annotating personal cliques from social networks to help users of microblogging services such as Twitter understand the structure of their social network and interact with the members of each clique independently.

Our proposed clique annotation method counts the number of followers and followees of users within the same clique, since keywords extracted from central clique members tend to represent common tweet topics of a clique more accurately than keywords of marginal clique members. The proposed method also distinguishes users who publish information of general interest (news sources) from users who publish private information (average users), since news sources tend to post less “noise” than average users. Based on eight criteria, including e.g., the number of followers and followees, the ratio of followers to followees, the number of tweets and the average tweet length, news sources can be identified accurately. In that way, we can weight keywords extracted from each clique member individually according to their relevance for the clique.

We developed a prototype of a Twitter client with clique management functionality. After this, we conducted an experiment in which we extracted and annotated cliques from the personal social networks of 20 Twitter users to show that weighting the keywords extracted from central clique members and emphasizing news sources identified using eight characteristics of news sources helps improve the accuracy of clique annotation.

In order to further improve the accuracy of clique annotation, we want to experiment with other ways to emphasize keywords that can identify the common ground of the clique members. Besides, we will develop a method for distinguishing interest-based and affiliation-based cliques automatically.

Our goal is to enable microbloggers to create and manage lists of followers and followees automatically, and use that information to keep track of their social network connections. Using the proposed method as a foundation, applications for sender-side or client-side information filtering can be implemented in order to help users deal with information overload. Clique annotation also helps users to understand the topics of interest of others in his social network and join their communication successfully.

Our proposed clique extraction and annotation algorithm is mainly targeting microblogging services such as Twitter. However, it can also be applied to other social networking services (SNS) in which user relationships can be expressed as a link structure. It can be expected that in all SNS, some users are central clique members whose keywords represent the common ground of the clique members more accurately than marginal clique members. However, news source discrimination is meaningful only for SNS in which unidirectional relationships among users are common. Apart from that, typical SNS such as Facebook provide much more information about their users than microblogging services, including detailed user profiles and status updates without length restriction, which should be exploited as well.

References

[1] Kim, I.-C.M.D., Jo, Y. and Oh, A.: Analysis of twitter lists as a potential source for discovering latent characteristics of users, *ACM Conference on Human Factors in Computing Systems* (*CHI*) *Workshop on Microblogging* (2010).

[2] Erdmann, M., Takeyoshi, T., Hattori, G. and Ono, C.: Extraction and annotation of personal cliques from social networks, *International Symposium on Applications and the Internet* (*SAINT*), pp.172–177 (2012).

[3] Hulth, A.: Improved automatic keyword extraction given more linguistic knowledge, *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.216–223 (2003).

[4] Li, Z., Zhou, D., Juan, Y.-F. and Han, J.: Keyword extraction for social snippets, *World Wide Web Conference* (*WWW*), pp.1143–1144 (2010).

[5] Liu, H. and Setiono, R.: Chi2: Feature selection and discretization of numeric attributes, *IEEE International Conference on Tools with Artificial Intelligence*, pp.338–391 (1995).

[6] Mihalcea, R. and Tarau, P.: Textrank: Brining order into texts, *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), pp.404–411 (2004).

[7] Newman, M.E.J. and Girvan, M.: Finding and evaluating community structure in networks, *Physical Review E*, Vol.69, No.2 (2004).

[8] Newman, M.E.J. and Girvan, M.: Finding community structure in very large networks, *Physical Review E*, Vol.70, No.6 (2004).

[9] Shinzato, K. and Torisawa, K.: Acquiring hyponymy relations from web documents, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*HLT-NAACL*), pp.73–80 (2004).

[10] Smith, M.A., Shneiderman, B., Milic-Frayling, N., Rodrigues, E.M.,

Barash, V., Dunne, C., Capone, T., Perer, A. and Gleave, E.: Analyzing (social media) networks with nodexl, *International Conference on Communities and Technologies* (*C&T*), pp.255–264 (2009).

[11] Zhang, B., Wu, W. and Ostendorf, M.: Automatic generation of personalized annotation tags for twitter users, *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (*HLT-NAACL*), pp.689–692 (2010).

Maike Erdmann received her B.Sc. in Computer Science from CvO University Oldenburg, Germany in 2006, and M.E. and Ph.D. in Information Science and Technology from Osaka University, Japan in 2008 and 2011 respectively. Since 2011, she is working in the Intelligent Media Processing Group of KDDI R&D Laboratories Inc., Japan. Her research interests include social media analysis and natural language processing.

Tomoya Takeyoshi received his B.E. in Information Engineering and M.S. in Computer Science from Hokkaido University, Japan in 2005 and 2007 respectively. Since joining KDDI R&D Laboratories, Inc. in 2007, he has been working on Web data mining technology and social media analysis. He is currently an assistant manager in the Advertisement Department of KDDI Corporation, Japan.

Kazushi Ikeda received his B.E., M.E. and Ph.D. in Information Science from Osaka University, in 2006, 2008 and 2013 respectively. Since 2008, he has been working in the Intelligent Media Processing Laboratory in KDDI R&D Laboratories, Inc., Japan. His research interests include natural language processing, machine learning and social media analysis. He is currently a visiting researcher at Stanford University.

Gen Hattori received his B.E. and M.E. degrees of Electrical and Electronics Engineering from Kobe University, Japan, in 1996 and 1998 respectively. He received the Ph.D. in Information Science from Osaka University, Japan, in 2009. Since joining KDD R&D Laboratories Inc., Japan in 1998, he has been working on network management systems, intelligent transportation systems, software mobile agent systems and Web text analysis. He is currently an R&D manager in the Intelligent Media Processing Laboratory of KDDI R&D Laboratories Inc. He received the Young Engineers Award of IEICE in 2003.

Chihiro Ono received his B.E. in Electrical Engineering, M.S. in Computer Science and Ph.D. in Science for Open and Environmental Systems from Keio University, Japan, in 1992, 1994 and 2009 respectively. Since joining KDD in 1994, he has been working on database systems, software agent technologies, personalization technologies and social media analysis. From 1999 to 2000, he was a visiting researcher at Stanford University. He is currently a senior manager in the Data Mining Application Laboratory of KDDI R&D Laboratories, Inc., Japan. He received the Best Paper Award for Young Researchers of the National Convention of IPSJ in 1996.

Yasuhiro Takishima received his B.S. and M.S. in Electrical Engineering and the Ph.D. in Electronic Engineering from the University of Tokyo, Japan, in 1986, 1988, and 1997 respectively. He has been working at KDDI R&D Laboratories Inc., Japan, since 1988. His current interests include human communication, signal processing and their appications.