Applying to Twitter Networks of a Community Extraction Method Using Intersection Graph and Semantic Analysis

Toshiya Kuramochi, Naoki Okada, Kyohei Tanikawa,
Yoshinori Hijikata, and Shogo Nishida

Graduate School of Engineering Science, Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka, Japan
kuramochi@nishilab.sys.es.osaka-u.ac.jp,
{hijikata,nishida}@sys.es.osaka-u.ac.jp

Abstract. Many researchers have studied about complex networks such as the World Wide Web, social networks and the protein interaction network. One hot topic in this area is community detection. For example, in the WWW, the community shows a set of web pages about a certain topic. The community structure is unquestionably a key characteristic of complex networks. We have proposed the novel community extracting method. The method considers the overlaps between communities using the idea of the intersection graph. Additionally, we address the problem of edge inhomogeneity by weiting edges using content information. Finally, we conduct clustering based on modularity. In this paper, we evaluate our method through applying to real microblog networks.

Keywords: complex network, community extraction, intersection graph, hierarchical clustering, text mining, microblog network.

1 Introduction

Many researchers, having studied complex networks such as the World Wide Web, SNS networks, and the protein interaction network, have reported scale-free characteristics, the small-world effect, the property of high-clustering coefficient, and so on [1,2]. Recently, the community structure in complex networks is gaining increased attention from many researchers. The community structure shows the appearance of densely connected groups of nodes, with only sparse connections among groups. Many community detection methods have been proposed based on this definition [11]. They are applied to various complex networks. Communities in SNS networks shows a set of people with the same background or hobby. Communities in the WWW show sets of web pages related to a certain topic [4] and those in the protein interaction network show sets of proteins having the same function [6].

For community detection, researchers have started to show interest in whether overlaps between communities can be extracted [15,17,23]. The overlaps signify that one node belongs to several communities. For example, the Apple Inc. page

M. Kurosu (Ed.): Human-Computer Interaction, Part III, HCII 2013, LNCS 8006, pp. 314–323, 2013.

is categorizable among computer category pages and audio category pages. It is important that a community detection method be able to assign a node not only to one community but also to several communities. Weights of all edges in complex networks are assumed to be the same in many community detection methods [19]. However, edges are rarely homogeneous in real networks. For example, various human connections such as those of businesses, hobbies, and organizations exist in SNS networks. It is important that the weights of these edges are not be treated as identical. They should be set individually. Many researchers conduct hierarchical clustering methods to complex networks for community detection. Most hierarchical clustering methods require advance input [3]: the number of clusters. However, the number of real communities is often unknown in real networks. Therefore, it is important that the number of proper clusters be decided automatically.

We have proposed a novel community detection method that can solve these three problems [8]. Our proposed method can extract overlaps between communities using the idea of the intersection graph. When there exist several subgraphs individual members are connected densely, we can make a new graph where the above subgraphs are converted to new nodes and edges are created when two subgraphs have common elements. Graph created by the above process is called intersection graph [10]. We also determine the weights of the edges in the intersection graph using two types of information: the degree of overlaps of the members and the similarity of content information such as text information and attribute information which appear in the network. Moreover, we use the hierarchical clustering method based on modularity proposed by Newman [11,13,14]. This method does not necessitate manual input of the number of clusters. In this paper, we evaluate general versatility of our method through applying to microblog networks.

2 Related Works

The problem of community detection in complex networks has been examined in various areas such as those of computer science and medical science [2].

Some researchers have attempted to extract communities in complex networks including the overlaps between communities. The overlaps mean that one node belongs to several communities. Everett et al. found them using the idea of the intersection graph [3]. Palla et al. found them by detecting cliques whose size was k and merging the cliques that shared $k-1$ nodes [15]. Fuzzy clustering, the method considers the notion of fuzziness and can assign one node to several communities is often used to extract overlaps of communities [17,23].

Our study weights edges in complex networks for dealing with edge inhomogeneity. Weighting edges in a network (usually a document network or hyperlink network) is popular in the area of the information retrieval. Some researchers improved the effectiveness of link analysis using content information. Jiang and Conrath measured the similarity between words using link information and semantic information of words [7]. Hung et al. improved the HITS algorithm by analyzing anchor text [5].

Many researchers use hierarchical clustering methods for detecting communities. The methods need input, which is the number of clusters preliminarily. Newman and Borgatti reported modularity as an indicator of how well the clusters are formed [12]. Newman proposed some clustering methods based on modularity [11,13,14]. These methods do not obviates manual input.

Our proposed method [8] considers the overlaps between communities using the idea of the intersection graph. Furthermore, we address the problem of edge's inhomogeneity by weighting edges using the degree of overlaps and the similarity of content information between sets (nodes of the intersection graph). Finally, we conduct a clustering method based on modularity, which does not necessitate manual input of the number of clusters. No study deal with all the above problems for detecting communities.

3 Proposed Method

In this section, we explain our proposed method [8]. The input of our proposed method is a graph of $G = (V, E)$, where V stands for the set of nodes and E signifies the set of edges. Additionally, content information is given to the nodes. We apply the following four steps to this graph.

Step 1. Enumeration of Dense Subgraphs: This method enumerates dense subgraphs (generally, they are called cliques) from an input graph of $G = (V, E)$.
Step 2. Conversion to the Intersection Graph: This method regards each subgraph enumerated in Step 1 as one new node and converts the input graph G to the intersection graph of $G' = (V', E')$.
Step 3. Calculation of the Weights of Edges: This method calculates the weights of edges E' in the intersection graph G' using the degree of overlaps and the similarity of content information between nodes V' (dense subgraphs) in G'.
Step 4. Clustering Based on Modularity: This method divides nodes V' into clusters using a clustering method based on modularity.

We applied the method of Everett et al. [3] to **Step 1** and **Step 2**. In **Step 1**, their method enumerates maximal cliques as dense subgraphs in an input graph of $G = (V, E)$. A clique is a subgraph in which an edge exists between any two nodes. Next, the method converts the input graph G into the intersection graph $G' = (V', E')$ in **Step 2**. Our method regards each dense subgraph enumerated in **Step 1** as one special node and makes the intersection graph $G' = (V', E')$ from the input graph $G = (V, E)$. When several sets (dense subgraphs) S_i $(i = 1, \cdots, n)$ are enumerated, our method generates a special node v'_i for each set S_i. If a common element exists in two arbitrary nodes v'_i and v'_j, then a special edge is put between them. The intersection graph is a new graph composed of special nodes and special edges [10]. When the method puts a special edge between special nodes, we can set the threshold of the number of common elements between the subgraphs corresponding to these special nodes. Finally, the method of Everett et al. conducts hierarchical clustering for the intersection graph. Our method address the edge inhomogeneity (in **Step 3**) and automatically detection of extracting communities number (in **Step 4**).

In **Step 3**, the proposed method calculate weights of edges in the intersection graph generated in **Step 2**. We use the degree of overlaps and the content information similarity between each subgraphs. There exist many types of measurement for presenting the degree of overlaps between X and Y ($d(X, Y)$) such as co-occurrence frequency, mutual information, Dice coefficient, Simpson coefficient and Jaccard coefficient [9,16]. For example, Jaccard coefficient is defined as:

$$d(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}. \tag{1}$$

The proposed method uses vector space model [18] to calculate the similarity of the content information between two arbitrary sets X and Y. The method regards each set as one vector and calculates the *tf·idf* score for the keyword in the texts in the set. This *tf·idf* score becomes the element of the vector. Finally, the method calculates the similarity $sim(X, Y)$ between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ corresponding to two sets X and Y using cosine similarity:

$$sim(X, Y) = \cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\| \boldsymbol{x} \| \| \boldsymbol{y} \|}. \tag{2}$$

Then, the proposed method calculates the weights $w(i, j)$ for the special edge between special nodes v'_i and v'_j (corresponding to set X and Y) using the degree of overlaps of sets $d(X, Y)$ and the similarity of content information $sim(X, Y)$. We can use several types of calculation function. In this work, we use function emphasizing the similarity of content information:

$$w(i, j) = w(X, Y) = \frac{d(X, Y)}{1 + \epsilon - sim(X, Y)}. \tag{3}$$

Here, ϵ $(0 < \epsilon < 1)$ is a constant used to keep the denominator from being 0.

Finally, in **Step 4**, the proposed method conducts clustering for community detection in the intersection graph. When a method extracts several clusters in a network, we must evaluate the currently detected clusters. Modularity is a broadly accepted indicator for evaluation. The indicator is simple and intuitive. Therefore, we adopt a clustering method based on the modularity that is proposed by Newman et al.[11,13,14]. When k clusters are given and P_k is defined as the sets of these clusters, the module function $Q(P_k)$ is the following.

$$Q(P_k) = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}(e) - |e^2| \tag{4}$$

$$\begin{cases} e_{ij} = \frac{1}{2m} \sum_{vw} A_{vw} \delta(c_v, i) \delta(c_w, j) \\ a_i = \frac{1}{2m} \sum_v k_v \delta(c_v, i) \end{cases}$$

4 Applying to Microblog Networks

We apply our method to real social networks. We select Twitter for this study which is the most popular microblog searvice. Twitter users post *tweets* (short

messages) and have conversations with other users through their tweets. If tweets have the word "@username," they are *mentions*—tweets for certain users. The word "*#hashtag*" in tweets means these tweets concern certain topics. Therefore, we can get the large amount of content information. In the Twitter network, all people can *follow* anyone without approvals, then the network contains links represent the unilateral interest. Links between users may represent relationships in the real world (university friend etc.) or that of interest (hobby friend etc.).

The purpose of this evaluation is to verify three questions:

- **Whether our method achieves better results than the conventional method:** We compare our method with the conventional method proposed by Everett et al. [3]. The conventional method converts an input graph into the intersection graph and conducts a simple hierarchical clustering.
- **Whether it is efficient to use content information for weighting edges:** Our method uses not only information about the degree of overlaps of sets but also the content information. We compare the method using both kinds of information with the method using only the degree of overlaps of sets. We confirm the effectiveness of the content information.
- **Whether the kinds of content information affect the results:** Twitter networks have several types of content information. We examine whether the results change according to the kind of content information.

4.1 Dataset

We make a dataset for the evaluation inviting test subjects who give true relationships between them and each member in the extracted communities. We followed users from a test subject to two in the radius (from the test subject up to the friends of the test subject's friends). Our experiment assume a situation extracting the communities of the real world or strong interest. There are some celebrities followed by million users in the Twitter network. A test subject may be not able to answer the proper relation between stranger users connected via such celebrities. We think it is important that all test subjects can answer the all relations between users in the dataset. Therefore, we set the threshold of the number of follow users and that of followers. Then, we removed such hub users (celebrities) in advance. Thresholds are set as 400 from prior study. Additionally, we collected profile texts and tweets (contain @usernames and #hashtags) as content information. The test subjects are 9 users who are all university students.

4.2 Implementation

Parameter Settings of the Proposed Method. We adopt the maximal clique as the dense subgraph in Step 1. We can use various sizes of the maximal clique (the clique threshold). If the clique threshold is 5, then the method uses only the maximal cliques that comprise more than four nodes. We set 3, 4 and 5 as the clique threshold. In Step 2, our method converts the original graph to the

Table 1. Statistical information of the dataset

method	content information
Prf	nouns in profile texts
Prf-H	nouns and #hashtags in profile texts
Prf-N	nouns and @usernames in profile texts
Prf-H+N	nouns, #hashtags and @usernames in profile texts
Twt	nouns in tweets
Twt-H	nouns and #hashtags in tweets
Twt-N	nouns and @usernames in tweets
Twt-H+N	nouns, #hashtags and @usernames in tweets

intersection graph. Here, it creates a special edge between two dense subgraphs (two special nodes) when they have common elements. We can set the threshold of the number of common elements (the overlap threshold) in this step. We set 2, 3 and 4 as the overlap threshold (In our prior work [8], we have found that the performance of the proposed method is very low when the overlap threshold is 1 under the influence of clustering method based on modularity). We conduct nine threshold conditions (clique threshold, overlap threshold) = $(3, 2)$, $(4, 2)$, $(4, 3)$, $(5, 2)$, $(5, 3)$ and $(5, 4)$.

In Step 3, we selected the Jaccard coefficient (eq. (1)) as the degree of overlaps of sets. We also selected profile texts and tweets as the content information. Our method extracts nouns, #hashtag and @username as keywords by conducting morphological analysis of the content information. The method calculates $tf \cdot idf$ scores for all keywords within one maximal clique (corresponding to a special node). The maximal clique can represent one vector. The similarity between maximal cliques is calculated using eq. (2). Finally, the weights between maximal cliques are calculated using eq. (3). We set $\epsilon = 0.1$ in this experiment. We use a greedy approach that repeatedly merges a pair of nodes to maximize the increment of the modularity [13].

Implementation of Community Extraction Method. As described in section 4.1, we use content information of two types: profile texts and tweets. We respectively designate the cases using profile texts and tweets as *Prf*, *Prf-H*, *Prf-N*, *Prf-H+N*, *Twt*, *Twt-H*, *Twt-N* and *Twt-H+N* (see Table 1). Hereinafter, we regard *Twt* as a representative example of these cases using the content information. We implemented the case using only the degree of the overlaps between sets (Jaccard coefficient) as the weights of edges in Step 3 to examine the contribution of content information analysis. We designate the case *NonCA* (without content analysis).

We implemented the method Everett et al. proposed [3] as a baseline method. The method comprises three steps. The first two steps of the method are the same as the first two steps (Step 1 and Step 2) of our method. Unlike our method, it uses a simple hierarchical clustering method in the third step. In detail, the method searches for a pair of special nodes that have maximal Jaccard

coefficient and merges the pair, repeatedly. We set the number of output clusters as the number of clusters including the test subject becomes the number of true communities (provided by each test subjects). We designate this case as *Everett's method*.

4.3 Evaluation Method of Extracted Clusters

Clusters of two kinds are extracted by a community extraction method: a cluster that includes the test subject and a cluster that does not contain the test subject. It is difficult for test subjects to judge the connection of members in the latter clusters. Therefore, we specifically addressed only those clusters containing the test subject. To measure the accuracy of the extracted clusters, we adopt the following evaluation process.

Step 1. A test subject enumerates all relation names for each person in the dataset (number of relation names is regarded as the number of true communities). The test subject can see the profile texts and latest 200 tweets of each person.
Step 2. We assume that the relation in the extracted cluster corresponds to each relation name. We calculate the precision, recall and$F$-measure per relation name for the cluster (For the calculation, we consider that the relation name is the name of correct relation for the cluster). The precision, recall and$F$-measure of a relation name N are calculated as follows.

$$\text{Precision}(N) = \frac{\#\text{ people whose relation name is } N \text{ in the extracted cluster}}{\#\text{ people in the extracted cluster}}$$

$$\text{Recall}(N) = \frac{\#\text{ people whose relation name is } N \text{ in the extracted cluster}}{\#\text{ people whose relation name is N in the dataset}}$$

$$F\text{-measure}(N) = \frac{2 \cdot \text{Precision}(N) \cdot \text{Recall}(N)}{\text{Precision}(N) + \text{Recall}(N)}$$

Step 3. We use the highest F-measure calculated in Step 2 among all relation names as the F-measure of the extracted cluster. We regard the relation name $\bar{N}$ that marked the highest F-measure as the relation of the cluster. We also use the precision and recall of a relation name $\bar{N}$ as the precision and recall of the clusters.
Step 4. We calculate the average values of the precision, recall and F-measure of all clusters. These values are regarded as the evaluation value of one test subject.

5 Evaluation of Extracted Clusters

We conducted an experiment of community extraction for clarifying three questions mentioned in Section 4. We show the average values of the precision, recall and F-measure in all 9 test subjects in the conventional method and our method in Figure 1-3.

We compare our method (*NonCA* and *Twt*) with the conventional method (*Everett's method*). In precision (Figure 1), the results of our method tend to become

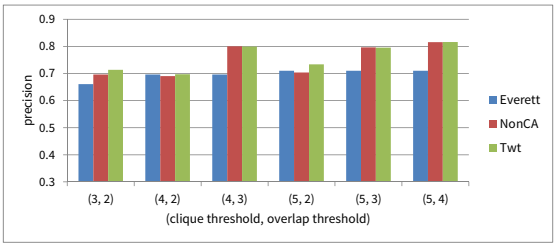

Fig. 1. Comparing the precision of Everett's method and our method

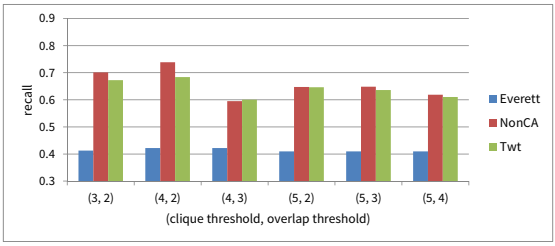

Fig. 2. Comparing the recall of Everett's method and our method

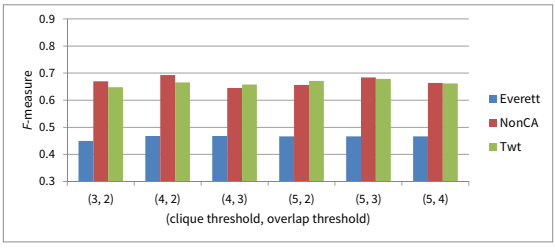

Fig. 3. Comparing the F-measure of Everett's method and our method

better when both thresholds become are larger. This is because users connecting strongly each other tend to survive while making the intersection graph. The proposed method overcomes the conventional method in the case when the overlap threshold become larger. In recall (Figure 2), the results of our method become better when both thresholds become smaller. Our method shows much higher recall than the conventional method in all conditions. At last, in F-measure (Figure 3), our method overcomes the conventional method in all conditions. Overall, we found our method brings a better result than the conventional method.

We compare *NonCA* and *Twt* to find out the effectiveness of content information analysis. In many threshold conditions, the precisions of *Twt* are greater than those of *NonCA* (Figure 1). Other hand, in recall, *NonCA* tends to overcome *Twt* (Figure 2). We cannot determine which method is better for extracting

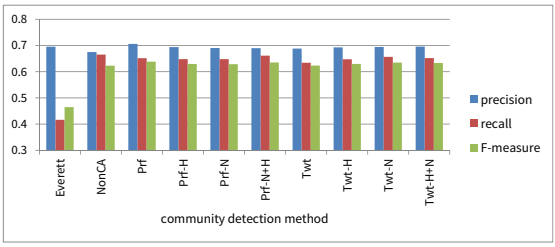

Fig. 4. Evaluation in condition $(4, 2)$ using content information of various kinds

communities because the difference in the F-measure is small (Figure 3). We think the reason is that text information in Twitter dataset contains many colloquial words and coined words. We conduct a simple morphological analysis that may be not enough effective to accurate extracting nouns from tweets.

Finally, we examine whether the results change according to the type of content information. We evalutate the cases using nouns, #hashtags, and @usernames in profile text and tweets as the content information in the threshold condition $(4, 2)$. We present results of *Everett's method*, *NonCA*, *Prf*, *Prf-H*, *Prf-N*, *Prf-H+N*, *Twt*, *Twt-H*, *Twt-N* and *Twt-H+N* in Figure 4. The results of cases using content information (all method without *Everett's method* and *NonCA*) are mutually similar. As we explained above, the influence to community detection of content information is not enough strong in our method (because of a simply morphological analysis). Additionally, many users in this experiment do not post tweets containing #hashtags or @usernames much. The volume of these information is not enough to influence the performance of our method.

6 Conclusion and Future Work

As described in this paper, we evaluated the proposed method for community detection [8] through applying to the Twitter dataset. We demonstrated the superiority of the proposed method compared to the conventional method. Although we compared the case using content information with the case not using that, we cannot show the advantage of proposed method. We think that is because the morphological analysis in our method is too simply to applying coined words and colloquial words in tweets and our experimental dataset do not contain much #hashtags and @usernames. As future workCwe will improve the morphological analysis and apply our method to other test subjects.

References

1. Albert, R., Barabasi, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
2. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex Networks: Structure and Dynamics. Phys. Rep. 424(4–5), 175–308 (2006)

3. Everett, M.G., Borgatti, S.P.: Analyzing Clique Overlap. Connections 21(1), 49–61 (1998)
4. Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.: Self-Organization of the Web and Identification of Communities. IEEE Computer 35(3), 66–71 (2002)
5. Hung, B.Q., Otsubo, M., Hijikata, Y., Nishida, S.: HITS Algorithm Improvement using Semantic Text Portion. WIAS 8(2), 149–164 (2010)
6. Huss, M., Holme, P.: Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks. IET Systems Biology 1(5), 280–285 (2006)
7. Jiang, J.J., Conrath, D.W.: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In: Proc. ROCLING 1997, pp. 19–33 (1997)
8. Kuramochi, T., Okada, N., Tanikawa, K., Hijikata, Y., Nishida, S.: Community Extracting Using Intersection Graph and Content Analysis in Complex Network. In: Proc. WI 2012, pp. 222–229 (2012)
9. Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (2002)
10. McKee, T.A., McMorris, F.R.: Topics in Intersection Graph Theory, vol. 2. SIAM, Discrete Mathematics and Applications (1999)
11. Newman, M.E.J.: Detecting community structure in networks. Eur. Phys. J. B 38(2), 321–330 (2004)
12. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2) (2004)
13. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69(6) (2004)
14. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E 74 (2006)
15. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043), 814–818 (2005)
16. Rasmussen, E.: Clustering Algorithms, Information Retrieval: Data Structures and Algorithms. In: Frakes, W.B., Baeza-Yates, R. (eds.), pp. 419–442. Prentice-Hall (1992)
17. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E 74(1), 16110 (2006)
18. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications ACM 18(11), 613–620 (1975)
19. Scott, J.: Social Network Analysis: A Handbook, 2nd edn. Sage Publications (2000)
20. Scripps, J., Tan, P.-N., Esfahanian, A.-H.: Node Roles and Community Structure in Networks. In: WebKDD/SNA-KDD 2007, pp.26–35 (2007)
21. Tasgin, M., Bingol, H.: Community Detection in Complex Networks using Genetic Algorithm. In: Proc. ECCS (2007)
22. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
23. Zhang, S., Wang, R., Zhang, X.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A 374(1), 483–490 (2007)