Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.238 – 242

RESEARCH ARTICLE

Discovering Communities of People from Social Network Using Novel Graph Mining Approach

Darshana P. Patel[1], Sandip Chauhan[2]

[12]Computer Department & Gujarat Technological University, Gujarat, India

[1] darshanapatel.1990@gmail.com; [2] sandymba2006@gmail.com

Abstract— Social networks are often modelled as graphs in which a node denotes a person, and an edge indicates some relationship, e.g., in Facebook, Twitter and LinkedIn. Various methods are existing of clustering for community detection like partitional clustering, hierarchical clustering, spectral methods and Galois lattices. Among of them Galois lattices are fulfilling our problem at some level but in this method a significant number of groups are generated. In this paper, we include improved Galois lattices which reduce limitations of this method.

Keywords— Social networks; Community detection; Clustering; Galois lattices

I. INTRODUCTION

Graph mining is a popular topic in research area. Social networks are often modeled as graphs in which a node denotes a person, and an edge indicates some relationship, e.g., in Facebook, Twitter and LinkedIn. In the actual interconnected world, and the rising of online social networks the graph mining and the community detection become completely up-to-date. Finding groups with high concentrations of relations within the group and low concentration between these groups, which is called community detection.

There are three main parts in graph mining: Mining frequent subgraph patterns, graph indexing and graph similarity search. And there are two main approaches for the subgraph pattern mining: Apriori-Based Approach and Pattern-Growth Approach [1]. Graph pattern matching used for graph clustering, graph indexing.

Some methods are existing of clustering for community detection like partitional clustering, hierarchical clustering, spectral methods and Galois lattices. In partitional clustering you have to add number of clusters as input parameter, and an inappropriate choice may leads too non significant results. In hierarchical clustering two types of algorithm agglomerative algorithms and divisive algorithm exists. Agglomerative algorithms which start with a set of small initial clusters and iteratively merging these clusters into larger ones. In this algorithm vertices of a community may be not correctly classified, and some nodes could be missed even if they have a central role in their cluster. divisive algorithms which split the dataset iteratively or recursively into smaller and smaller clusters. The main problem with this algorithm is it become NP complete problem and when to stop splitting graph. Spectral algorithms constitute a very particular class of techniques. And this particularity is to perform the classification based on eigenvectors of matrix build upon the adjacency (or weight) matrix. spectral clustering has complexity issue because it requires the computation of the eigenvectors of the Laplacian matrix, and if the graph is large, an exact computation require large complexity.

All above methods are non overlapping but Galois lattice is more informative method. A Galois Lattice clusters the data (called objects) in classes (called concepts) using their shared properties. The Galois lattice based clustering is costly and not simple to read but has several advantages in comparison with other clustering methods. The notion of similarity is more strict in Galois lattices than in similarity based clustering. Given a dataset the Galois lattice turns a context into a unique and complete lattice of concepts, while classical clustering produces a result over all possible classes, depending of several choices (methods, parameters,...).

Among of them Galois lattices are fulfilling our problem at some level but in this method a significant number of groups are generated. So, in proposed work we try to solve problem with improved Galois lattices which reduce limitations of this method. To reduce number of groups generated in Galois lattices we put user defined threshold value so if the community have less member than that threshold value at time the new group generation process from that community will be stopped so that number of iteration will be reduced and we can get improved result.

II. LITERATURE REVIEW

In this section, I present the different kind of algorithm are exist for detecting community from social network. Here, we show some famous algorithm from them and also comparison among them.

A. Partitional Clustering

Partitional algorithms try to find a partition of a set of data, with a given number of cluster equal to k. The "best" partition is searched using a distance or dissimilarity measure. **k- means clustering** Find a partition C = (C1, · · · ,Ck) of a set of n objects denoted {x1, · · · , xn}, which minimize the sum-of-square (SSQ) criterion1:

$$g_n(C) = \sum_{i=1}^{k} \sum_{x \in Ci} d(x, ci)$$

where, d(x, ci) measures the dissimilarity between the object x and the centroid of the class Ci. Expression is called SSQ, because initially d(x, ci) was computed in the following way : $\|x - ci\|^2$.[17]

B. Hierarchical Algorithms

Hierarchical clustering algorithms are divided into two types, depending on whether the partition is refined or coarsened during each iteration: 1. Agglomerative algorithm 2. Divisive algorithms.

Agglomerative algorithms which start with a set of small initial clusters and iteratively merging these clusters into larger ones. At the starting point, the n objects to cluster {x1, · · · , xn} are their own classes: {{x1}, ⋯,{xn}} ,then at each stage we merged the two more similar clusters.

Divisive Algorithm which split the dataset iteratively or recursively into smaller and smaller clusters. The algorithm given in [7] splits the network into clusters by removing, step after step, edges with the higher betweenness value. In the algorithm, the two following steps are repeated:

1. compute the edge betweenness for all edges of the running graph,
2. remove the edge with the largest value (which gives the new running graph).

Several community detection's algorithms are based on the minimization of the number of edges which link the clusters. These algorithms are based on the minimization for each cluster of the cut size between the cluster and the outside of this latter. The *Kernighan-Lin* algorithm is one the first algorithm in this way[8].

C. Spectral methods

The classification based on eigenvectors of matrix build upon the adjacency (or weight) matrix[11]. Suppose that G is an undirected, weighted graph, with positive symmetric weights matrix W, $w_{i,j} = w_{j,i} \geq 0$. Moreover we need to define the degree matrix D: D = W · I. where I is the identity matrix. D is such that we found the degrees deg(vi) on the diagonal. Now, we are able to define the spectral clustering: the Laplacian matrix : L = D −W. Then, given a Laplacian L and a number of cluster k, the general algorithm for spectral clustering is the following:

1. compute the eigenvalues and sort them such
2. compute the last k eigenvectors $u_{n-k} \cdots, u_n$,
3. form matrix $U \in R^{nxk}$ with $u_{n-k} \cdots, u_n$ as columns, and matrix $Y = U^t$,
4. cluster the points yi using the k-means into clusters A1, · · · ,Ak,
5. build the communities C1, · · · ,Ck such Ci = {vj |yj 2 Ai}.

D. Galois Lattice

All the previous clustering methods are conceived to build non overlapping clusters, and are non exhaustive but there exists more informative methods. One of these interesting methods, is the Formal Concept Analysis (FCA) and the Galois Lattices [9] [10], which can be used for the conceptual clustering [11][12]. A Galois Lattice clusters the data (called objects) in classes (called concepts) using their shared properties. The concepts found using Galois Lattices can be communities of people sharing connections, but also shared informations or opinions in an online network.

We give definition of Galois lattices. Let us suppose that we have a set of objects O, and a set of possible attributes A for these objects. The possession of a property a $\in$A by an object o $\in$ O is fulfilled when there is a relation I between them: aIo. Relations I between O and A are captured in a binary incidence matrix. The triplet K = (O,A, I) is called a formal context or simply a context. Finally the Galois lattice is the set of concepts L with the following partial order $\leq$:

(X1, Y1) $\leq$ (X2, Y2) , X1 $\subseteq$ X2(or Y1 $\supseteq$ Y2).
The Galois lattice is denoted T = (L, $\leq$) . To find a concept is not difficult:
1. pick a set of obects X,
2. compute Y = f(X),
3. compute X′ = g(Y),
4. (X′, Y) is a formal concept.

The dual approach can be taken starting with a set of attributes. A more difficult task is, given a formal context table, to generate all the existing concepts and build the Hasse diagram of the Galois lattice. Wedetail here one of the simplest methods to build a Galois lattice, the Bordat's algorithm [16], which build L and its Hasse diagram. Let us, firstly, define the cover of a concept C = (X, Y), denoted Ç,

$$Ç = \{C' | C' \leq C \text{ and } \forall C'' : C' \leq C'' \leq C\}$$

The algorithm starts with the set of concept given by the single objects (o, f(o)) and then find all their children nodes which are added to L and linked to their parent. This child generation process is iteratively repeated for every new pair:

– L < −{(o, f(o))|o $\in$ O)},

– for each concept C $\in$ L:
– build C,

– for each C′ $\in$ C:

• if C′ $\notin$ L, then L < −L $\cup$ {C′},
• add the edge between C and C′.
– end for.
– end for.

Several other algorithms were proposed and the reader can take a look at [14] for a review and a comparison of performances. The first use of Galois lattices in social network data analysis is owed to Freeman in [15], where he use the belonging to a clique as attribute.

E. Comparison

Partial clustering method is cheapest and easy method but not give efficient result. The hierarchical methods are more scalable than Partial methods. The spectral clustering continue with a more greedy but very efficient technique. The Galois lattices which are costly methods, but with very rich results. Galois lattice based clustering is then costly, and moreover not simple to read but, but has several advantages in comparison with similarity clustering [16].The notion of similarity is more strict in Galois lattices than in similarity based clustering: two objects in a Galois lattice are similar if they share identical properties, while they are alike to a degree quantified by a proximity measure in the other case.

In Partitional clustering you have to add number of clusters as input parameter, and an inappropriate choice may leads too non significant results. In hierarchical clustering two types of algorithm agglomerative algorithms and divisive algorithm exists. Agglomerative algorithms which start with a set of small initial clusters and iteratively merging these clusters into larger ones. In this algorithm vertices of a community may be not correctly classified, and some nodes could be missed even if they have a central role in their cluster. divisive algorithms which split the dataset iteratively or recursively into smaller and smaller clusters. The main problem with this algorithm is it become NP complete problem and when to stop splitting graph. Spectral algorithms constitute a very particular class of techniques. And this particularity is to perform the classification based on eigenvectors of matrix build upon the adjacency (or weight) matrix. spectral clustering has complexity issue because it requires the computation of the eigenvectors of the Laplacian matrix, and if the graph is large, an exact computation require large complexity.

All above methods re non overlapping but Galois lattice is more informative method. A Galois Lattice clusters the data (called objects) in classes (called concepts) using their shared properties. The Galois lattice based clustering is costly and not simple to read but has several advantages in comparison with other clustering methods. The notion of similarity is more strict in Galois lattices than in similarity based clustering. Given a dataset the Galois lattice turns a context into a unique and complete lattice of concepts, while classical clustering produces a result over all possible classes, depending of several choices (methods, parameters,...).

F. Motivation from Literature

Among of all method for detecting community from social network, Galois lattices are fulfilling our problem at some level but in this method a significant number of groups are generated . So, in proposed work we try to solve problem with improved Galois lattices which reduce limitations of this method. To reduce number of groups generated in Galois lattices we put user defined threshold value so if the community have less member than that threshold value at time the new group generation process from that community will be stopped so that number of iteration will be reduced and we can get improved result.

III. PROPOSED APPROACH

In this section, method used for proposed work is discussed in detail. In proposed work to reduce number of groups generated in Galois lattices we put user defined threshold value so if the community have less member than that threshold value at time the new group generation process will be stopped.

The steps for proposed work are as follows:

1. Read database of objects and their attributes.
2. Set threshold value (th). (This is user defined threshold value.)
3. each object compute total number of attribute t.
4. if t > th then go to step 5 else remove object go to step 3.

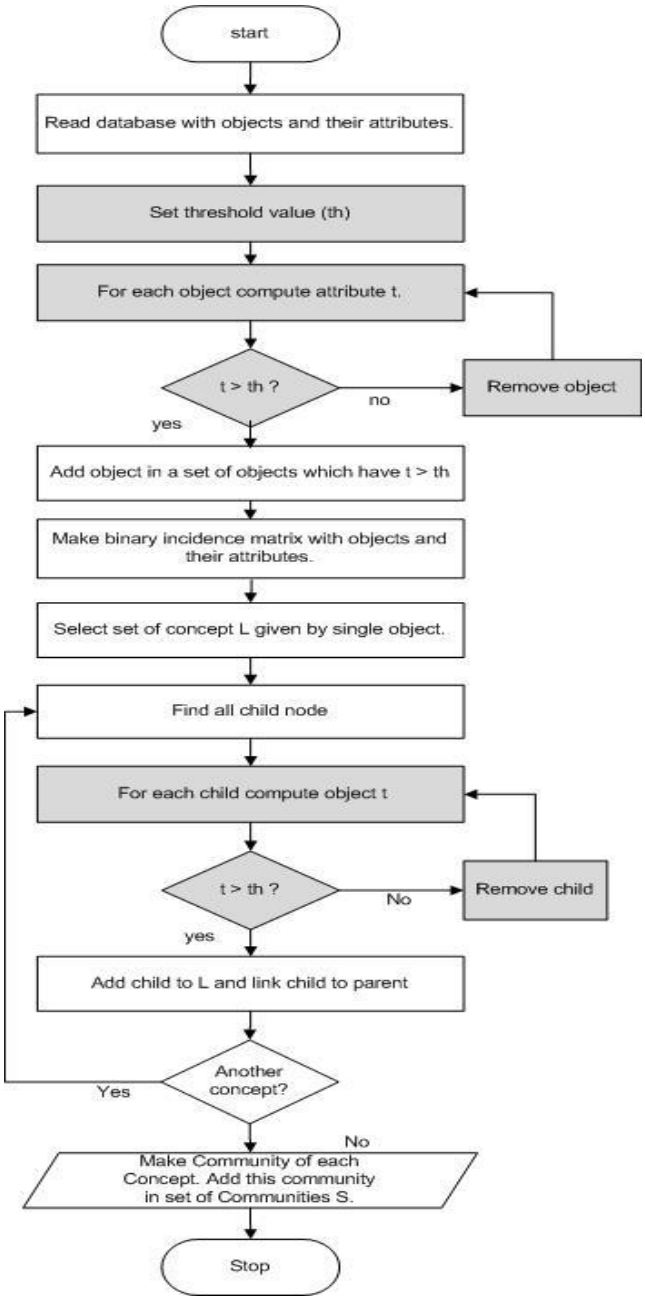

Fig.1 flow to reduce iteration in Galois Lattices

5. add object in a set of objects which have t > th.
6. Make binary incidence matrix with objects and their attributes.
7. select set of concept given by single object.
8. for each concept C ∈ L.
9. Find all child node of $\underline{C} \in C$
10. For each child compute total number of object t
11. if t > th ? then go to step 12 else remove child go to step 9.
12. Add child to L and link child to parent
13. end for
14. end for
15. Make Community of each Concept. Add this community in set of Communities S.
16. End

IV. CONCLUSIONS

To reduce number of groups generated in Galois lattices we put user defined threshold value so if the community have less member than that threshold value at time the new group generation process from that community will be stopped so that number of iteration will be reduced and we can get improved result. We get communities in the form of cluster. This communities are more scalable cause it have number of objects greater than threshold value.

REFERENCES

[1] Jiachuan Shi, Graph Pattern Mining: A survey.
[2] K.Lakshmi1 and Dr. T. Meyyappan2, FREQUENT SUBGRAPH MINING ALGORITHMS A SURVEY AND FRAMEWORK FOR CLASSIFICATION,CS&IT,189.
[3] Divya Prakash ,Subu Surendran , Detection and Analysis of Hidden Activities in Social Networks, IJCA, *Volume 77 – No.16, September 2013*
[4] Harsh J. Patel1, Rakesh Prajapati2, Prof. Mahesh Panchal3, Dr. Monal J. Patel4,A Survey of Graph Pattern Mining Algorithm and Techniques, *IJAIEM,* Volume 2, Issue 1, January 2013.
[5] M. E. J. Newman, The structure and function of complex networks, SIAM REVIEW 45 (2003) 167-256.
[6] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264-323.
[7] Fragkiskos D. Malliarosa, Michalis Vazirgiannisa,b, Clustering and Community Detection in Directed Networks:A Survey.
[8] Etienne Cuvelier and Marie-Aude Aufaure, Graph Mining and Communities Detection, Springer -Verlag Berline Heidelberg 2012 ,pp 117-138.
[9] Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E 69(2) (Feb 2004) 026113
[10] Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99 (2002) 7821–7826
[11] Kernighan, B., Lin, S.: An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal 49(2) (1970) 291308
[12] Barbut, M., Monjardet, B.: Ordre et classification, Algebre et combinatoire, Tome 2. Hachette (1970)
[13] Birkhoff, G.: Lattice Theory. Volume 25. American Mathematical Society, New York (1940) Carpineto, C., Romano, G.: Galois: An order-theoretic approach to conceptual clustering. In:Proc. Of the 10th Conference on Machine Learning, Amherst, MA, Kaufmann.(1993) pp.33– 40.
[14] Wille, R.: Line diagrams of hierarchical concept systems. Int. Classif. 11 (1984) 77–86
[15] Bordat, J.: Calcul pratique du treillis de galois dune correspondance. Math´ematique, Informatique Sciences Humaines 24(94) (1986) 31–47
[16] Bordat, J.: Calcul pratique du treillis de galois dune correspondance. Math´ematique, Informatique et Sciences Humaines 24(94) (1986) 31–47
[17] Bordat, J.: Calcul pratique du treillis de galois dune correspondance. Math´ematique, Informatique et Sciences Humaines 24(94) (1986) 31–47
[18] Kuznetsov, S.O., Obedkov, S.A.: Comparing performance of algorithms for generating concept lattices. In: Concept Lattices-based Theory, Methods and Tools for Knowledge Discovery in Databases (CLKDD'01), Stanford, July 30, 2001. (2001)
[19] Freeman, L.: Cliques, galois lattices, and the structure of human social groups. Social Networks 18 (1996) 173–187
[20] MacQueen, J.: Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab.Univ. Calif. 1965/66, 1, 281-297 (1967). (1967)
[21] Neubauer Nicolas and Obermayer Klaus. Towards Community Detection in k-Partite k-Uniform Hypergraphs. In *Proceedings NIPS 2009* . . . .
[22] Michel Planti_e, Michel Crampes, Survey on Social Community Detection, HAL Id: hal-00804234, Sub0mitted on 25 Mar 2013.

