International Journal of Advances in Computer Science & Its Applications– IJCSIA
Volume 5 : Issue 2 [ISSN : 2250-3765]
Publication Date: 30 October, 2015

Relation and Attribute Fusion to Detect Communities of Online Social Networks

Mohammad H. Nadimi-Shahraki, Mehrafarin Adami

Abstract—**Community detection is a vital research area for online social networks. Since there is not a formal context in users` profiles, a new data source of user`s attributes is extracted from online social networks. Then in this paper a novel algorithm is investigated. Both attributes and relations of users are used in the proposed algorithm; therefor e communities can be detected through users' similar characteristics or users common relationships. The experiments show that the accuracy of the algorithm is comparable to other well-known algorithms; moreover detected communities are self-descripted through the mode of each community members.**

Keywords—**Social networks, Community detection, Influential nodes, Self-descriptive communities.**

I. Introduction

Social network community detection is an important issue for means of effective advertisements, accurate recommender systems and tracking changes. Since social networks consist of several data sources, they can be very informative that yield self-descriptive communities.

Through the community detection process, it is possible to categorize common relations between users and analyse each related part of a network, community, separately in more detail. Changes can also be shown by tracking communities.

First studies on community detection methods focused on link analysis. Since those methods failed to extract semantic of the communities, hybrid methods are favored recently. These methods enhance the efficiency of community detection results [1-4]. Almost all of these methods use Bayesian models, and they combine two algorithms or two data sources. These data sources are often links and context. Using context as a data source has some limitations that are listed below. Handling context as a data source for community detection is a voluminous task that needs algorithms with high computational complexities; therefore, the majority of research studies extract the most frequent words or keywords from the context, and then use these limited words as the content. Furthermore, Contexts are extracted from scientific papers written with correct grammatical instructions and formal vocabularies. In the case of online social networks comments and descriptions usually consist of informal words and phrases, which are particularly used in virtual environments. In addition, users' contexts are written in different languages. Because online social networks have users from all over the world; Hence, Extracting the context is not suitable in order to detect the communities in online social networks. In this study, each node of the social networks is described by both attribute and link (relation) data sources, e.g. demographic information as attributes and relations or even a separate social network as links. Intuitively, in this paper a community is a connected subgraph of nodes, consisting of users who have the most similarity to each of the community members based on both of these sources. Relation data source form communities, when their users are connected to each other and at the same time, they have similar characteristics.

This paper is organized as follows. In section 2, the related works are reviewed. Then, the proposed algorithm is described in section 3. Afterwards in section 4, the real data set from Facebook users and Soybean dataset are described, in order to demonstrate the accuracy of the algorithm. Finally, section 5 concludes the contribution and introduces some future works.

II. Related Work

A. Simple community detection approaches

The simple community approaches are graph partitioning, devise methods, spectral methods, Bayesian models, and clustering [11-17]. These approaches use one data source to construct the graph of social networks. This data source can show explicit links such as friendship relations or implicit relations that are gathered from the content of network. Two major content analyses are topic models and scientific authors' network, called citation graph. For detecting communities, each document is changed to some frequent keywords, and the document-word matrix is constructed for further analysis [5, 6].This preprocessing is a costly step for community detection methods.

Citation graphs are social networks constructed from the references and used keyword in scientific papers. Experimental results show that the accuracy of community detection can be enhanced through this information [18].

In topic model studies, each community has one or more topics that its users are writing about that. Bayesian models are used for community detection algorithms; hence, most

Mohammad H. Nadimi-Shahraki and Mehrafarin Adami are with Faculty of Computer Engineering, Najafabad branch, Islamic Azad University, Najafabad, Esfahan, Iran,.

topic models are generative and vulnerable to words that are irrelevant to the target topic [1]. Furthermore, There are only a handful of scalable Bayesian approaches to community discovery in graph [10] that mostly extend Latent drichlet Allocation [19-20].

B. Combinatorial community detection approach

Neither link information nor content information is sufficient to decide the community membership. Combining the link and content for community detection usually achieves better result. Two improvements of this approach is given following.

Combining different approaches covers drawbacks of every single algorithm. HCDF presented by Henderson et al, uses Latent Dirichlet Allocation on Graphs (LDA-G) as the core Bayesian method for community detection. A key aspect of HCDF is its effectiveness in incorporating hints from a number of other community detection algorithms and producing results that outperform the constituent parts. Furthermore, it produces algorithms that can predict links [10].

Complementary information used to propose new improved community detection methods. Influential nodes seem to be the centroids of communities. By this consideration, [21] proposed a new method that formed communities by influential nodes called leaders, and assigned other nodes to these leaders as followers. This complementary information is extracted from link sources but yields more accurate communities.

III. Hybrid Community Detection Algorithm

Almost all the previous hybrid methods use context as content of social network. Context handling needs special models such as Bayesian models, but not all of them are scalable. Also preprocessing task is costly as related topic should be extracted. Furthermore, in some cases like online social networks, the context is not written in a proper grammatical and formal way; therefore, quality of community detection results will be decreased.

In this paper the proposed algorithm utilizes two data sources of relations and attributes to detect communities. Therefore, the algorithm is able to detect communities of a social network without mentioned concerns. Using this algorithm, each node would be a member of a community if it has similar characteristics with others, or if it has common neighbors with other nodes of its community.

This section is followed by describing how the nodes in each community are scored based on their relations or attributes, and then the way of fusing these two data sources is described. Finally the proposed algorithm shows how to detect communities.

A. Relations and the neighborhood score

Relationship data source is modeled by adjacency matrix; each element in the matrix is computed as follows [22]:

$$m_{ij} = 1 \quad iff \;\; relation\,(x_i, x_j) = True \tag{1}$$

The neighbors of each node are compared with all the cores` neighbors, then that node will be assigned to the community that has the maximum score of common neighbors with its core. The score of neighborhood is computed as follows, where N is the number of users in their community.

$$score_{i,j}(neighborhood) = \frac{|common\_neighbors(node_i, core_j)|}{N} \tag{2}$$

B. Users' attributes and the similarity score

Attribute data source can be represented as an n*m user - attribute matrix, where n is the number of users or nodes, and m represents the number of attributes that each user has. The proposed algorithm uses categorical variables as attribute data source. Such as k-modes algorithm, each node should be assigned to a cluster/community with the most similar core. The dissimilarity between two nodes are computed by simple mismatching [23]. Simple mismatching between two nodes of x_i, x_j with d attributes is computed as follows:

$$D(x_i, x_j) = \sum_{l=1}^{d} \delta(x_{il}, x_{jl}) \tag{3}$$

$$\delta(x_{il}, x_{jl}) = \begin{cases} 0 & x_{il} \neq x_{jl} \\ 1 & x_{il} = x_{jl} \end{cases}$$

Recently, a new dissimilarity measurement parameter was introduced by Cao et.al that enhanced the accuracy of k-modes clustering algorithm [24]. This parameter for p attributes is computed as follows:

$$NDis_p(z_i, x_i) = \sum_{a \in P} NDis_a(z_i, x_i) \tag{4}$$

$$NDis_a(z_i, x_i) = 1 - Sim_a(z_i, x_i) \times m_a$$

$$m_a = \frac{|x_i | f(x_i, a) \equiv f(z_i, a), x_i \epsilon c_i|}{|c_i|}$$

$$f(x, a) \equiv f(y, a) = \begin{cases} 1 & \text{if } f(x, a) \neq f(y, a) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Finally, the similarity score is computed as follows:

$$score_{i,j}(similarity) = 1 - dissimilarity(node_i, core_j) \tag{6}$$

C. Detecting communities based on fusing two sources

The proposed algorithms need k representative nodes as cores of communities. Each node in the data set will be assigned to one of these cores which has the maximum score by that core. First communities are formed by attribute source. Next, an iterative process happens until communities undergo no change, refine the communities and update the best core of each community. The proposed algorithm, like other hill climbing algorithms such as k-modes, is sensitive to its initial nodes that the algorithm will be started by them. Moreover, the number of communities is another input parameter for this algorithm.

To detect communities based on similar attributes and relations, two scores should be considered. The way of fusing these scores is summation. If one of these sources has greater influence on the social network, the related score can be multiplied by an instant number as the weight of that source. In this study, we do not assign weight to sources. Before summation of two scores, scores should be normalized. By min-max normalization, both of the scores will be in the same range of values. Finally, total score will be:

$$\text{total score}_{i,j} = \text{normalized score}_{i,j}(\text{neighborhood}) + \text{normalized score}_{i,j}(\text{similarity}) \quad (7)$$

Algorithm 1 highlights the major steps of the hybrid community detection algorithm. The major steps are forming the communities and updating the cores. First assignments form first communities based on the similarity between nodes and cores, but for the iterative section, both the similarity and neighborhood scores are considered. Experiments show that first formation will be refined by iterative section. That is why it only used the similarity score of attributes for the first discovering of communities. Iterative section considers both scores. Common neighbors between each node and the core should be computed through all of the native nodes. If the core or the considered node - has common neighbors with other communities, these neighbors should be ignored as they are not members of that community. Also the similarity will be compared to the nodes of each core`s community.

Considering two data sources of attributes and relations for community detection make updating step more complex in comparison to one source used algorithms, such as k-modes. The next subsection focuses on selection of new cores after the new formation of each community. The core of a community is the most similar node to all the other nodes in its community. Since the attributes are categorical variables, the mode of each community is a good choice to show that core. Mode of a cluster or a community is a vector of attributes that can be not a node`s attributes in the data set. That is why we do not use the mode to present the core. On the other hand, the node with the maximum influence in the community can also be an appropriate choice for selecting of cores. Influence can be measured by degree centrality, which is one of the best and simplest measurements [25]. It is computed using the following formula where N is the number of nodes and $deg(n,c)$ is the degree of node n in community of c:

$$DC(n) = \frac{deg(n,c)}{N-1} \quad (8)$$

Algorithm 1 The proposed algorithm

```
INPUT:  dataset (attributes, adjacency matrix), k
Selecting  cores
// first assignments
for all nodes n $\epsilon$ dataset do
   assign n to the community which has minimum dissimilarity with its real centroid(core) and make first communities
end for
for all communities do
   Update the core of community // Algorithm 2
end for
// iterative section
Repeat
   for all nodes n $\epsilon$ dataset do
      assign n to the community which has maximum total score with its core
   end for
   for all communities do
      Update the core of community// Algorithm 2
   end for
until there is no change in communities
```

IV. EXPERIMENTS

The proposed algorithms are coded in Matlab 7.10.0 programming language. Other related results and illustrations are using NodeXL 10.0.1.229. Consistently, two popular datasets Soybean and Facebook are conducted. The accuracy of proposed algorithm is compared with two well-known algorithms; k-modes for attributes, and Newman-Girvan algorithm [15] to test the link data source. Finally, the accuracy of hybrid sources is computed. The dissimilarity parameter for Soybean data set is the simple mismatching parameter [12]. Table 2 summarizes the number of members in each community and the accuracy of the proposed algorithm and Newman-Girvan algorithm.

Since the ground truth of Facebook data set is not available, the attribute data source efficiency is just compared through the Soybean data set. Therefore, to reveal the accuracy of hybrid sources, the Soybean communities were compared to the kind of disease that each instance has. This study consisted of two set of initial nodes to start the algorithm. Similar to other hill climbing algorithms, initial nodes have effects on the accuracy. The experiment set was run for 10% of dataset nodes and the average results of accuracy on Soybean data set are summarized in table 3. In this test, number of links is not constant. For the first experiment, the adjacency matrix is null, then it has just 61 edges and the other experiments have full adjacency matrix. The last experiment did not select the first cores randomly; instead each core was an instant with one of the fourth kind of diseases. The increasing value of the accuracy shows that this source is really essential in this algorithm. As each core is a node with real influential score, accuracy for the link-based method is comparable to other methods; however, the core is not really the mode of that community; hence, we expect that the effect of links is greater, especially for Soybean data set.

Table 2. Facebook members of each community and algorithm accuracy.

	1st community	2nd community	3rd community	Accuracy
Newman-Girvan	318	220	65	$\frac{601}{603} = 0.996$
Proposed algorithm	318	216	69	$\frac{600}{603} = 0.995$
Friends of seeds	320	219	73	-

Table 3. Accuracy vs. link source completeness.

Number of edges	First core selection method	Accuracy
0	Randomly	0.357
61	Randomly	0.6
271	Randomly	0.71
271	Specified nodes	0.893

V. Conclusion and future work

The previous hybrid community detection methods were proposed based on the context and link sources. Since context processing is a voluminous task and online social networks contain informal vocabularies, a new source is introduced in this paper. Users have profiles with their information that yields to detect self-descriptive communities besides the link data source. Hence, the proposed algorithm is investigated to detect communities based on these two data sources. Experiments show the accuracy of the proposed algorithm; the results are comparable to well-known algorithms. The next extension is specifying a solution to extract the best first cores, and estimating the number of communities in the social network.

References

[1] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: a discriminative approach," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 927-936.

[2] K. Yang, "Combining Text-and Link-based Retrieval Methods for Web IR," in *TREC*, 2001.

[3] Y. Wang and M. Kitsuregawa, "On combining link and contents information for web page clustering," in *Database and expert systems applications*, 2002, pp. 902-913.

[4] J. Li and O. R. Zaïane, "Combining usage, content, and structure data to improve web site recommendation," in *E-Commerce and Web Technologies*, ed: Springer, 2004, pp. 305-315.

[5] S. Zhu, K. Yu, Y. Chi, and Y. Gong, "Combining content and link for classification using matrix factorization," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 487-494.

[6] N. F. Chikhi, B. Rothenburger, and N. Aussenac-Gilles, "Combining link and content information for scientific topics discovery," in *Tools with Artificial Intelligence, 2008. ICTAI'08. 20th IEEE International Conference on*, 2008, pp. 211-214.

[7] F. Moser, R. Ge, and M. Ester, "Joint cluster analysis of attribute and relationship data withouta-priori specification of the number of clusters," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 510-519.

[8] C. Wang, Z.-y. Guan, C. Chen, J.-j. Bu, J.-f. Wang, and H.-z. Lin, "On-line topical importance estimation: an effective focused crawling algorithm combining link and content analysis," *Journal of Zhejiang University SCIENCE A*, vol. 10, pp. 1114-1124, 2009.

[9] Y.-M. Li and C.-W. Chen, "A synthetical approach for blog recommendation: Combining trust, social relation, and semantic analysis," *Expert Systems with Applications*, vol. 36, pp. 6536-6547, 2009.

[10] K. Henderson, T. Eliassi-Rad, S. Papadimitriou, and C. Faloutsos, "HCDF: A Hybrid Community Discovery Framework," in *SDM*, 2010, pp. 754-765.

[11] H.-W. Shen and X.-Q. Cheng, "Spectral methods for the detection of network community structure: a comparative analysis," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2010, p. P10020, 2010.

[12] M. E. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577-8582, 2006.

[13] B. Kernighan, Lin, S, " An efficient heuristic procedure for partitioning graphs," *Bell system technical journal*, vol. 49, pp. 291-307, 1970.

[14] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering e-communities," in *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 173-182.

[15] C. C. Aggarwal, *An introduction to social network data analytics*: Springer, 2011.

[16] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, pp. 75-174, 2010.

[17] S. Parthasarathy, Y. Ruan, and V. Satuluri, "Community discovery in social networks: Applications, methods and emerging trends," in *Social Network Data Analytics*, ed: Springer, 2011, pp. 79-113.

[18] D. Greene and P. Cunningham, "Multi-view clustering for mining heterogeneous social network data," presented at the 31st European Conference on Information Retrieval 2009.

[19] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50-57.

[20] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Advances in neural information processing systems*, 2008, pp. 897-904.

[21] R. R. Khorasgani, J. Chen, and O. R. Zaïane, "Top leaders community detection approach in information networks," in *Proceedings of the 4th Workshop on Social Network Mining and Analysis*, 2010.

[22] S. Zhou, A. Zhou, W. Jin, Y. Fan, and W. Qian, "FDBSCAN: a fast DBSCAN algorithm," *RUAN JIAN XUE BAO*, vol. 11, pp. 735-744, 2000.

[23] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, pp. 283-304, 1998.

[24] F. Cao, J. Liang, D. Li, L. Bai, and C. Dang, "A dissimilarity measure for the< i> k</i>-Modes clustering algorithm," *Knowledge-Based Systems*, vol. 26, pp. 120-127, 2012.

[25] J. Sun and J. Tang, "A survey of models and algorithms for social influence analysis," in *Social Network Data Analytics*, ed: Springer, 2011, pp. 177-214.

[26] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7821-7826, 2002.

About Authors:

Mohammad H. Nadimi was born in Iran. He received his Ph.D in computer science from University Putra of Malaysia (UPM) in 2010. Currently, he is a full time assistant professor at the faculty of computer engineering of Islamic Azad University of Najafabad (IAUN). His research interests include data mining, web mining, social network mining and recommender systems.

Mehrafarin Adami received her master of computer software engineering in 2013 from faculty of computer engineering, Islamic Azad University of Najafabad under supervisory of Dr. Nadimi. She is currently continuing her research on social network and community detection methods.