Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Discovering and Profiling Overlapping Communities in Location Based Social Networks using Feature Description

Manju Joseph
Department of Computer Science & Engg
Mangalam College of Engineering,
Ettumanoor, India

Seena George
Department of Computer Science & Engg
Mangalam College of Engineering,
Ettumanoor, India

Abstract— With the soaring popularity of location–based social networks (LBSNs), such as Twitter, Foursquare and Face book Places, huge digital footprints of user's locations, profiles and online social connections are available to service providers, which are capable of providing lots of applications such as, trend analysis, direct marketing, group search, tracking etc. LBSNs usually have no explicit community structure as that of social networks like Face book, Flicker etc. Quality community detection and profiling approaches are needed for capitalizing the potential users. Meanwhile, the diversity of people's interests and behaviours when using LBSNs suggests that their community structures overlap. In this paper, based on the user check-in traces at venues and user/venue attributes, mainly the profile attributes or the feature attributes, we come out with a novel multimode multi-attribute edge-centric co-clustering framework to discover the overlapping and hierarchical communities of LBSNs users. The proposed framework is able to group like-minded users from different social perspectives, different contexts like travelling, walking, stationary etc and discover communities with explicit profiles indicating the interests of community members. The efficacy of our approach is validated by intensive empirical evaluations using the collected Twitter dataset.

Keywords— community profiling, hierarchical clustering, location-based social networks, feature description, overlapping community detection.

I. INTRODUCTION

The wide espousal of GPS-enabled smart phones has upshot the increasing popularity of location based social networks (LBSNs) through which the users can explore and share locations and experiences with others, upload photos etc. The towering hip of LBSNs has created opportunities for discernment of collective user behaviours on a large scale, thereby providing many applications like trend analysis, tracking, group search and direct marketing. Detection of user communities is a vital matter in social network analysis. A community is a group of users who are more similar with users within the group than those outside the group [2], [3]. Quality community detection and profiling approaches are needed to furnish on the huge number of potential users in LBSNs since the perception of community is not well defined in it. A real world example for polymorphism can be a particular person who acts as a family member, a friend, a faculty etc. If we are incorporating the impression of community structure into this, we can categorize this as a person belonging to several groups like a family, friends, colleagues etc. So, it is more reasonable to cluster users into overlapping communities rather than disjoint users.

We present an example of the user-venue check-in network in figure 1 with five users and four venues. In this, users and venues exemplify two types of nodes and each check-ins as an edge between a user node and a venue node. On performing edge clustering to this attributed bipartite network, we obtain two overlapping communities: Group 1 (Anna, Geo) and Group 2 (Geo, Anju, Elsa, Abin). By implicitly using the venue mode to characterize the user mode (intermode)

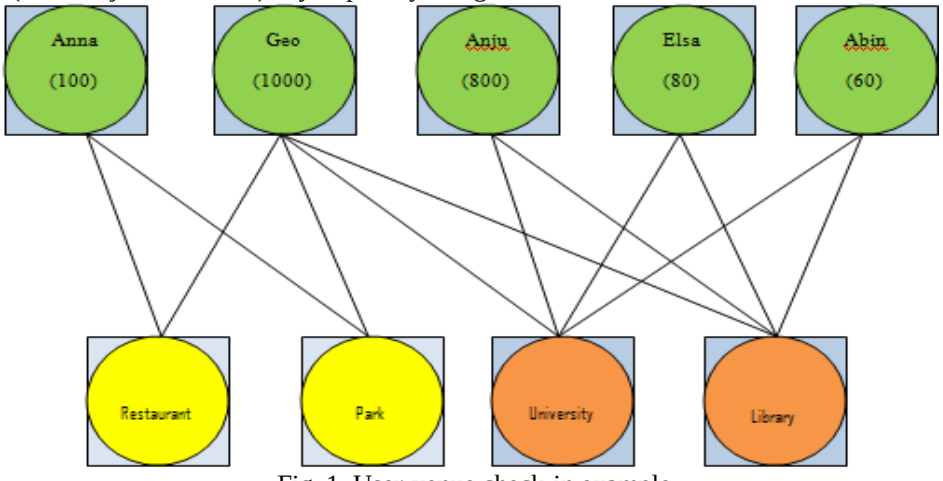

Fig. 1. User-venue check-in example

we can consider Group 1 as a family community and Group 2 as a colleague community. However, if we consider both the check-in network (intermode features) as well as the attributes of users and venues (intramode features), we conclude as three overlapping communities: Group 1 (Anna, Geo), Group 2 (Geo, Anju) and Group 3 (Elsa, Abin). In this scenario, even though Geo, Anju, Elsa and Abin have similar check-in patterns, they are again grouped into two separate communities. Since Geo and Anju travel frequently whose radius of gyration, r_g (given below their names in the node) are 1000 km and 800 km, while Elsa and Abin mainly stay locally whose r_g are 80 km and 60 km respectively. So we can obviously label Group 1 as a family community, Group 2 as a research staff community and Group 3 as a community of teaching staffs.

Existing community detection approaches are based on structural features (eg. links) but in social network, this information is sparse and weak. So, we cannot make the interpretations completely by using this structural information alone. LBSNs provide rich information about 'user' and 'venue' through check-in and clustering is performed using this elucidation. Clustering is performed on edges rather than on nodes.

Indeed, it is more plausible to exploit both structural information (intermode) as well as the node attributes (intramode) to cluster users, as we can deduce communities with richer and interpretable information. Classical coclustering is a method to perform this type of community partitioning, the identified communities are disjointed that contradicts the social settings. Edge has been proposed for overlapping community clustering but it does not consider intramode features. Overlapping community detection problem in LBSN is formulated as a coclustering issue, considering both user-venue check-in network and attributes of users and venues equally. We consider both community detection and profiling in a single integrated framework and acquire communities containing user and venue particulars simultaneously.

II. RELATED WORKS

The related works can be categorized into three:

A. Category 1

This contains the research on comprehension of the collective user behaviours based on LBSNs. Only two works addressed at husking group profiles in LBSNs. Li *et al.* [4] contemplated two different clustering methods to ascertain user behaviour patterns on Bright Kite. Noulas *et al.* [5] exploited a spectral clustering algorithm to sort out Foursquare users based on the genre of venues they had checked in, striving at identifying communities and typifying the activities in each part of a city. All these works offer important notion into characteristics of user interactions in LBSNs, but none of them considered overlapping community detection through network links and attributes of nodes.

B. Category 2

This group involves the work on community detection which is a classical problem in complex network analysis. For detecting communities from a network of nodes, one makes use of an objective function followed by an approximate or heuristic algorithm that optimizes the objective function to extract node clusters. Some popular methods for community detection include, modularity maximization, Girvan Newman algorithm [2], link communities [6] etc. One cannot simply apply community detection depending absolutely on the network links and wait for the generation of interpretable communities in the case of LBSNs as it has a weak and sparse relation.

C. Category 3

This category is similar to our work, which focuses on both links and node attributes for community detection. The key idea is to design a distance/similarity measure for vertex pairs that aggregates both structural and attribute information of nodes. Depending on this measure, standard clustering algorithms like *k-medoids* and spectral clustering are applied to cluster nodes. The state-of-the-art distance-based approach is the SA-cluster [7] defines a unified distance measure to merge structural and attribute similarities.

In this paper, we tried to purchase the structural links between users and venues together with their attributes to key out the overlapping community structure. Generally, we formulate the overlapping community detection hitch into a multimode, multi-attribute edge clustering issue, considering both intermode links and intramode attributes as aggregated features for clustering. With this novel representation, users and venues along with their attributes are shunt in a natural way, where the detected communities have explicit semantic meanings that can be interpreted as community profiles. An edge is generated between two vertices only if some particular feature attributes are satisfied by the origin vertex, a user or a venue. So, the graph generated is similar to those condition controlled automata transitions ie, a transition from state A to B will occur only if some particular attributes are satisfied by state A.

III. PROBLEM STATEMENT

A community is defined as a cluster of edges (check-ins) with user and venue as the two modes. We have, U = (u1, u2, ..., um) , represents the user set and V = (v1, v2, ... , vn) , represents the venue category set. A community C_i ($1 \leq i \leq k$) is a subset of users and venue categories , k is the count of communities. Check-in network between users and venue categories form a matrix M , where each entry $M_{ij} \in [0,\infty)$ accounts to the number of check-ins that u_i has performed over v_j. So each user can be denoted by a vector of venue categories and vice-versa. Also, users and venues have independent attributes as (a_{i1} , $a_{i2, ..., a_{ix}}$), and ($b_{j1, b_{j2, ...}}$, b_{jy}) respectively. An example for user or venue attribute includes,

a particular number of followers or followings in Twitter, and a venue category may have a certain operating time. This indicates that the two modes, user and venue has intermode as well as intramode representations.

Based on the above said notations, overlapping community detection problem in LBSNs can be represented as a multimode multi-attribute edge-centric coclustering problem as below:

1. A check-in matrix $M(|U|\times|V|)$,
   where $|U|$ → the numbers of users and
   $|V|$ → the numbers of venue categories.
2. A user attributes matrix $M(|U|\times|A|)$,
   where $|A|$ → the number of user attributes.
3. A venue category attributes matrix $M(|V|\times|B|)$,
   where $|B|$ → the number of venue category attributes.
4. The number of communities k,
   (optional, based on the clustering algorithm.)

Output:

1. k overlapping communities with both users and venue categories.

IV. MULTIMODE MULTI-ATTRIBUTE EDGE CLUSTERING FRAMEWORK

The steps involved in community discovering and profiling framework are:

1. Features are picked based on the characteristics of composed LBSNs dataset, followed by feature normalization and fusion.
2. Overlapping community structure is spotted by using the recommended edge clustering algorithm.
3. The detected communities are united with the user/venue metadata to retrieve the community profiles to transcribe the social and semantic meanings of communities.

The above-said framework can be depicted as follows:

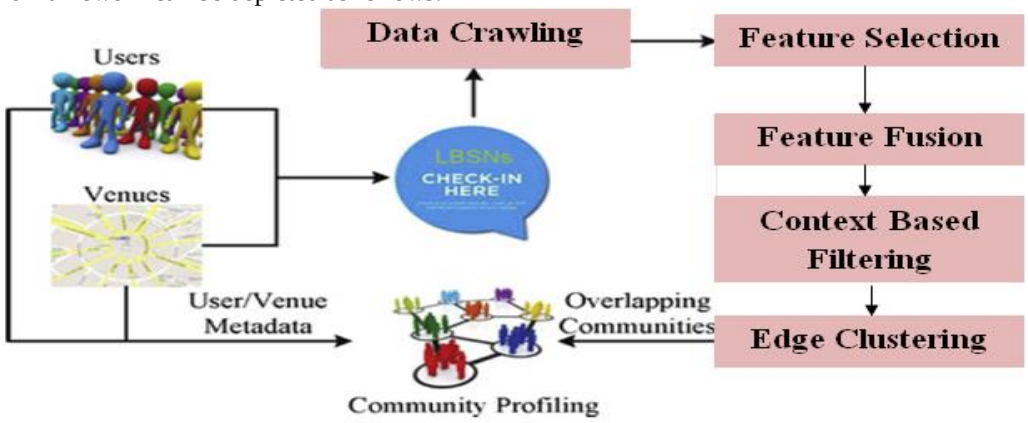

Fig 2 : Community discovering and profiling framework

In LBSN community, the users within a group have much more similarities than that for users outside the group. Communities that congeries similar users and venues together is identified by maximizing intracluster similarity. In a user-venue check-in network, each edge is related with a user vertex and a venue vertex. In edge-centric view, each edge can be considered as an instance with its two vertices as its features. So, for a pair of edges, similarity between corresponding pair of user and venue vertices are as follows:

$$sim_{edge}(e_i, e_j) = F(sim_u(u_i, u_j), sim_v(v_i, v_j))$$

where,

$sim_u(u_i, u_j)$ → similarity between two users

$sim_v(v_i, v_j)$ → similarity between two venues

F → function to combine similarities

Commonly adopted F are average and multiplication. Here, we make use of multiplication.

Each community has a certain number of edges and the similarity between an edge e_i and a community C_j is given as,

$$sim(ei, Cj) = \frac{1}{|Cj|}\sum_{ec \in Cj} simedge(ei, ec)$$

Where, |Cj| → number of edges within community Cj

V. FEATURE DESCRIPTION

The feature description includes both intermode as well as intramode features.

A. Intermode features

1. User-venue similarity
2. Venue-user similarity

B. Intramode features

1. User social-influence similarity
2. User geo-span similarity
3. Venue temporal similarity
4. User profile attribute similarity

Since we make use of different similarity features and different calculation methods, the similarity values estimated will be in different value ranges. So, we normalize each and every similarity to [0,1] using cosine normalization.

‘

VI. CLUSTERING ALGORITHM

Two step hierarchical clustering algorithm is employed. We make use of Multimode Multi-attribute Feature based edge clustering algorithm (M^2F) and a hierarchical H M^2F algorithm.

A. M^2F algorithm

Step 1: Randomly select k edges.
Step 2: Construct the initial centroids based on the k edges.
Step 3: Initialize an event list with the attributes of users.
Step 4: Calculate similarity of each edge, similarity it lost in last rearrangement and current similarity.
Step 5: Centroid is updated as that, for each edge e_i, which is most similar to itself and this similarity is denoted as $maxsim_i$.
Step 6: At the end of every iterations, current value of Objective function, Obj_{cur} is calculated to compare with its previous value Obj_{pre}.
Step 7: Iteration completes only if, $| Obj_{cur} - Obj_{pre} | > Є$, where Є is a pre-defined threshold.

B. H M^2F *algorithm*

Step 1: Invoke M^2F algorithm
Step 2: Obtained groups are agglomerated using average – linkage hierarchical algorithm.
Step 3: All edges now belong to a single cluster and, the history of clustering process is stored in a dendogram.

A two step hierarchical clustering approach is adopted rather than the classical hierarchical clustering since the later is quite time consuming when processing large datasets.

VII. PERFORMANCE EVALUATION

To evaluate the performance of the proposed framework, we worked over 1000s of Twitter accounts. Dataset has been preprocessed by removing invalid venues. Only those users who has at least a check-in a week are considered (active users) which means, inactive users together with their check-ins are excluded. Finally, users who used agent software conducting remote and large scale automatic check-ins (with a checkin speed faster than than 1200 km/h, which is the common airplane speed) are defined as sudden move users and check-ins from these users are eliminated as well. After the above data cleansing, we retrieve the dataset for the three targeted cities as follows. We first calculate the home location of all the active users, and then a set of users for each city are selected based on the distance between their home locations and the geometric center of the corresponding city. Based on the dataset we conducted experiments to evaluate the quality of the detected communities when using different algorithms and different feature sets.

Our experiments compared the depth of clustering criteria variations with respect to the database size for both the algorithms. In the previous work [1], the experiments were not focused on the profile information of the user like, gender, age etc. Here, in this work, we do consider the basic information of a user ie, the finest and crucial information about a Twitter user. We also worked on the fluctuations of accuracy with respect to the variations in the number of followers. For this, we considered thousands of Twitter accounts holding a wide range of follower count.

Through our experiments it has been found that, by considering an event controlled ie, by taking into account the profile information of user, the accuracy of overlapping community detection has been increased as illustrated by figure 3.

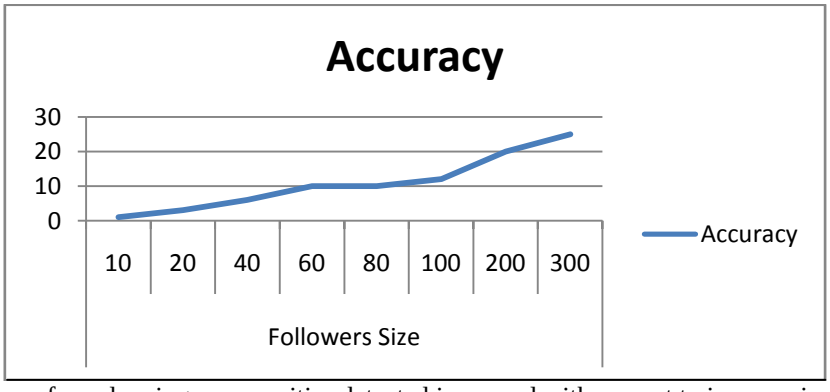

Fig 3: Accuracy of overlapping communities detected increased with respect to increase in follower size.

We also experimented that as the database size increases, the depth of clustering criteria has increased as compared with the older scenario [1], as in figure 4

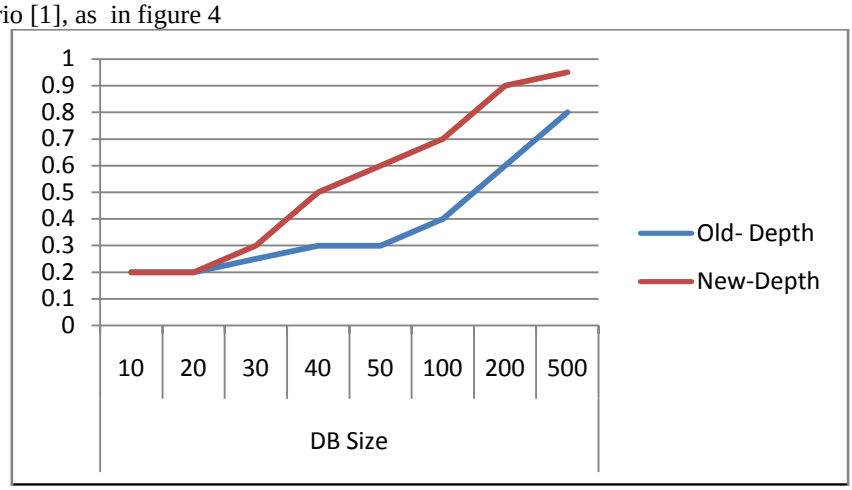

Fig 4: Depth of clustering criteria increases with increase in database size.

VIII. CONCLUSION AND FUTURE SCOPE

By leveraging the user-venue check-in network and user/venue attributes, a multimode multi-attribute edge-centric coclustering framework is proposed to detect overlapping communities for LBSNs users. Experimental results showed that the proposed framework was able to discover high quality overlapping communities from different perspectives and at multiple granularities, which can be used to facilitate different applications, such as group advertising and marketing. The preliminary study suggested several interesting problems that were worth further exploring. Providing a framework to guide the selection and fusion of different features is one direction to work on. The proposed community detection framework can also help the study of friend recommendation mechanisms. These recommendations can be provided based on the exact location of the user considering whether the user is mobile or stationary. More feature attributes can be considered to achieve maximum accuracy.

ACKNOWLEDGMENT

The authors would like to thank all of their friends and faculty members for their discussions and suggestions.

REFERENCES

[1] Zhu Wang, Daqing Zhang, Xingshe Zhou, Dingqi Yang, Zhiyong Yu and Zhiwen Yu "*Discovering and Profiling Communities in Location- Based Social Networks*". IEEE transactions on Systems, Man and Cybernetics.

[2] M. E. J. Newman and M. Girvan, "*Finding and evaluating community structure in networks*," *Phys. Rev. E*, vol. 69, no. 2, pp. 26 113–26 127, 2004.

[3] S. Fortunato, "*Community detection in graphs*," *Phys. Rep.*, vol. 486, nos. 3–5, pp. 75– 174, 2010.

[4] M. A. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida, "*Tips, dones and todos: Uncovering user profiles in Foursquare,"* in *Proc. WSDM*, 2012, pp. 653–662.

[5] N. Li and G. Chen, "*Analysis of a location based social network*," in *Proc. CSE*, 2009, pp. 263–270.

[6] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek, "*Uncovering the overlapping community structure of complex networks in nature and society*," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.

[7] K. Steinhaeuser and N. V. Chawla, "*Community detection in a large real-world social network*," in *Social Computing, Behavioral Modeling, and Prediction*, H. Liu, J. J. Salerno, and M. J. Young, Eds. New York, NY, USA: Springer US, 2008, pp. 168–175.

