International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P): 2349-3968, ISSN (O): 2349-3976
Volume 1 Issue 6, October 2014

USER GROUP ESTIMATION AND RECOMMENDATION MODEL FOR SOCIAL NETWORKS

Ms. S.Rabiyathul Basiriya
II Year M.E (CSE)
Shree Venkateshwara Hi-Tech Engg College, Gobi
rabiyasabu19@gmail.co

Dr. T. Senthil Prakash
Professor & HOD
Shree Venkateshwara Hi-Tech Engg College, Gobi
jtyesp@yahoo.co.in

Ms. P.Sudhaselvanayaki
II Year M.E (CSE)
Shree Venkateshwara Hi-Tech Engg College, Gobi
sudhaselva44@gmail.com

ABSTRACT- *Communities with explicit profiles indicate the interests of community members. User and check-in venue details are used to cluster users with different preferences and interests into different Communities. Multimode multi-attribute edge-centric coclustering model is used to discover overlapping and hierarchical communities. Overlapping communities of users can be recovered by replacing each edge with its vertices in edge clusters. Inter mode and intra mode features are used in group discovery process. User-venue similarity and venue-user similarity are the inter mode features. Three intra mode features are used in the community detection process. They are User Social-Influence Similarity, User Geo-Span Similarity and Venue Temporal Similarity features. K-means based multimode multi-attribute edge clustering (M^2Clustering) algorithm is used for community detection with fixed K value. Hierarchical multimode multi-attribute edge clustering (HM^2Clustering) algorithm is used to detect overlapping communities of LBSNs. The overlapping community detection mechanism is enhanced with recommendation models. The community details are classified with location or regions. The system is enhanced with feature selection and fusion mechanism. The system is improved to provide friends recommendation model with place details.*

Keywords: *overlapping community detection, M^2clustering, HM^2clustering, friends recommendation model*

1 INTRODUCTION

A social network is a social structure made up of a set of social actors and a set of the dyadic ties between these actors. The social network perspective provides a set of methods for analyzing the structure of whole social entities as well as a variety of theories explaining the patterns observed in these structures [4]. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics.

Social networks and the analysis of them is an inherently interdisciplinary academic field which emerged from social psychology, sociology, statistics, and graph theory. Georg Simmel authored early structural theories in sociology emphasizing the dynamics of triads and "web of group affiliations." Jacob Moreno is credited with developing the first sociograms in the 1930s to study interpersonal relationships. These approaches were mathematically formalized in the 1950s and theories and methods of social networks became pervasive in the social and behavioral sciences by the 1980s. Social network analysis is now one of the major paradigms in contemporary sociology, and is also employed in a number of other social and formal sciences. Together with other complex networks, it forms part of the nascent field of network science.

The social network is a theoretical construct useful in the social sciences to study relationships between individuals, groups, organizations, or even entire societies [5]. The term is used to describe a social structure determined by such interactions. The ties through which any given social unit connects represent the convergence of the various social contacts of that unit. This theoretical approach is, necessarily, relational. An axiom of the social network approach to understanding social interaction is that social phenomena should be primarily conceived and investigated through the properties of relations between and within units, instead of the properties of these units themselves. Thus, one common criticism of social network theory is that individual agency is often ignored although this may not be the case in practice. Precisely because many different types of relations, singular or in combination, form these network configurations, network analytics are useful to a broad range of research enterprises. In social science, these fields of study include, but are not limited toanthropology, biology, communicationstudies, econ omics, geography, information science,

organizational studies, social psychology, sociology, and sociolinguistics.

2 RELATED WORK

In this section, we briefly review the related work that can be classified into three categories. The first category contains the research on understanding the collective user behaviors based on LBSNs. Scellato et al. [7], [8] analyzed the social, geographic and geo-social properties of four social networks. Noulas et al [9] investigated the user checkin dynamics and the presence of spatio-temporal patterns in Foursquare. Cheng et al. [10] studied the mobility patterns of Foursquare users and revealed the factors affecting people's mobility. Vasconcelos et al. [11] analyzed how Foursquare users exploited three features to uncover different behavior profiles. Only two studies aimed at uncovering group profiles in LBSNs. Li et al. [12] proposed two different clustering approaches to identify user behavior patterns on BrightKite. Noulas et al. [3] used a spectral clustering algorithm to group Foursquare users based on the categories of venues they had checked in, aiming at identifying communities and characterizing the type of activity in each region of a city. Although the aforementioned studies offer important insights into properties of user interactions in LBSNs, none of them worked on overlapping community detection using network links and node attributes. Our work aims to fill in this gap by discovering and profiling communities in an overlapping manner.

The second category involves the work on community detection that is a classical task in complex network analysis [2]. In order to detect communities from a network of nodes, one typically chooses an objective function based on the intuition that a cluster is a set of nodes with better internal connectivity than external connectivity, and then applies approximate or heuristic algorithms to extract node clusters by optimizing the objective function. In general, community detection can be classified into two categories: overlapping and non-overlapping approaches. Some popular methods are modularity maximization, Girvan Newman algorithm, Louvain algorithm, clique percolation, link communities [6], etc. As users in LBSNs have rather weak and sparse relations [1], one cannot naively apply community detection based solely on the network links and expect to generate interpretable communities.

The third category focuses on community detection by considering both links and node attributes, which are the closest to our work. Several existing works on attributed graph clustering fall into this category. The main idea is to design a distance/similarity measure for vertex pairs that combines both structural and attribute information of the nodes. Based on this measure, standard clustering algorithms such as k-medoids and spectral clustering are then applied to cluster the nodes. For instance, a weighted adjacency matrix is used as the similarity measure, where the weight of each edge is defined as the number of attribute values shared by the two end nodes. The authors applied graph clustering algorithms on the constructed adjacency matrix to perform clustering. The state-of-the-art distance-based approach is the SA-cluster that defined a unified distance measure to combine structural and attribute similarities. Attribute nodes and edges are added to the original graph to connect nodes that share attribute values, and a neighborhood random walk model is used to measure the node closeness on the augmented graph. Afterward, a clustering algorithm SA-cluster is proposed based on the k-medoids method.

The last category attempted to optimize two contradictory objective functions and intended to identify disjoint communities; thus, the communities detected were not optimal and had no clear semantic meanings. In this paper, we propose to leverage both the structure links between users and venues, and their attributes to discover the overlapping community structure. Specifically, we formulate the overlapping community detection problem into a multimode multi-attribute edge clustering issue, viewing both intermode links and intramode attributes as unified features for clustering. With this novel representation, users and venues together with their attributes are grouped in a natural way, where the detected communities have explicit semantic meanings that can be interpreted as community profiles.

3 DISCOVERING COMMUNITIES IN LBSNs

With the wide adoption of GPS-enabled smartphones, location-based social networks (LBSNs) have been experiencing increasing popularity, attracting millions of users. In LBSNs, users can explore places, write reviews, upload photos, and share locations and experiences with others. The soaring popularity of LBSNs has created opportunities for understanding collective user behaviors on a large scale, which are capable of enabling many applications, such as direct marketing, trend analysis, group search, and tracking.

One fundamental issue in social network analysis is the detection of user communities. A community is typically thought of as a group of users with more and/or better interactions amongst its members than between its members and the remainder of the network. Unlike social networks that provide explicit groups for users to subscribe to or join, the notion of community in LBSNs is not well defined. In order to

capitalize on the huge number of potential users, quality community detection and profiling approaches are needed. It has been well understood that people in a real social network are naturally characterized by multiple community memberships. For example, a person usually belongs to several social groups such as family, friends, and colleges; a researcher may be active in several areas. Thus, it is more reasonable to cluster users into overlapping communities rather than disjoint ones.

Most of the existing community detection approaches are based on structural features, but the structural information of online social networks is often sparse and weak; thus, it is difficult to detect interpretable overlapping communities by considering only network structural information. Fortunately, LBSNs provide rich information about the user and venue through check-ins, which makes it possible to cluster users with different preferences and interests into different communities. Specifically, the observation that a check-in on LBSNs reflects a certain aspect of the user's preferences or interests enlightens us to cluster edges instead of nodes, as the detected clusters of check-ins will naturally assign users into overlapping communities with connections to venues. Once edge clusters are obtained, overlapping communities of users can be recovered by replacing each edge with its vertices, i.e., a user is involved in a community as long as any of her check-ins falls into the community. In such a way, the obtained communities are usually highly overlapped.

We present an example of the user-venue check-in network, which consists of five users and four venues. In such a network, users and venues are represented as two types of nodes, and each check-in is represented as an edge between a user node and a venue node. For this attributed bipartite network, since both users and venues have their own attributes, if we perform edge clustering to group users based solely on network structure, we can get two overlapping communities: Group 1 (Mary, Tom) and Group 2 (Tom, David, Bob, Eva). By implicitly using the venue mode to characterize the user mode, we can interpret Group 1 as a family community and Group 2 as a colleague community. If we consider not only the check-in network but also the attributes of users and venues, we can get three overlapping communities: Group 1 (Mary, Tom), Group 2 (Tom, David), and Group 3 (Bob, Eva). In this case, even though Tom, David, Bob, and Eva have similar check-in patterns, they are further grouped into two separate communities. Since Tom and David travel frequently whose radius of gyration (i.e., rg) are 1000 km and 800 km, while Bob and Eva mainly stay locally whose rg are 80 km and 60 km, respectively. Here, we probably can label Group 1 as a family community, Group 2 as a research staff community, and Group 3 as a teaching staff community.

Apparently, it is more reasonable to exploit both the structural information (intermode) and the node attributes (intramode) to cluster users, as we can naturally obtain communities with richer and interpretable information, even though it is a highly challenging task. While classical coclustering is one way to conduct this kind of community partitioning, the identified communities are disjointed, which contradicts with the actual social setting. Edge clustering has been proposed to detect communities in an overlapping manner, but it did not take intramode features into consideration.

From the perspective of service providers, it is equally important to identify communities with similar interests and understand what each community is interested in. In contrast to existing community detection approaches that seldom address the profiling of detected communities, we intend to take community profiling into account when designing the community detection framework. We believe that it's crucial to characterize communities in a semantic manner to effectively support real-world applications. Due to the limitation of available node information, not much work has been done on community profiling. The rich user and venue metadata available in LBSNs, especially the hierarchical structure of venue categories, provides us the possibility to semantically characterize the identified communities. In this paper, we aim to make the following two contributions.

1) We formulate the overlapping community detection problem in LBSNs as a coclustering issue that considers both the user-venue check-in network and the attributes of users and venues. Specifically, we detect overlapping communities from an edge-centric perspective, where each edge is viewed as a link between two modes, i.e., a user mode vertex and a venue mode vertex. While existing multimode clustering methods mainly concern the intermode features, we adopt both intermode and intramode features for clustering. By introducing different attributes of users and venues as intramode features, we show that various perspectives of social communities can be revealed.

2) We consider both community detection and profiling in one unified framework and obtain communities containing user and venue information simultaneously. In such a way, each community explicitly shows who is interested in where with what attributes, which is very useful in enabling real applications. In the meantime, we analyze and compare the detected user community profiles in London, Los Angeles and New York, with interesting findings.

International Journal On Engineering Technology and Sciences – IJETS™
ISSN (P): 2349-3968, ISSN (O): 2349-3976
Volume 1 Issue 6, October 2014

4 PROBLEM STATEMENT

Communities with explicit profiles indicate the interests of community members. User and check-in venue details are used to cluster users with different preferences and interests into different Communities. Multimode multi-attribute edge-centric coclustering model is used to discover overlapping and hierarchical communities. Overlapping communities of users can be recovered by replacing each edge with its vertices in edge clusters. Intermode and intramode features are used in group discovery process. The intermode feature describes the structure similarity between a pair of edges based on the check-in relationships. User-venue similarity and venue-user similarity are the intermode features. The intramode feature depicts attributes similarity where each attribute corresponds to a certain social aspect of users or venues. Three intramode features are used in the community detection process. They are User Social-Influence Similarity, User Geo-Span Similarity and Venue Temporal Similarity features. K-means based multimode multi-attribute edge clustering (M^2Clustering) algorithm is used for community detection with fixed K value. Hierarchical multimode multi-attribute edge clustering (*HM^2Clustering)* algorithm is used to detect overlapping communities of LBSNs. The following problems are identified from the existing system.

- Feature selection and fusion process is not optimized
- Recommendation mechanism is not provided
- Community preferences are not analyzed
- Geographical region analysis is not performed

5 MULTIMODE MULTI-ATTRIBUTE EDGE CLUSTERING

We define a community in LBSNs as a group of users who are more similar with users within the group than users outside the group. Therefore, communities that aggregate similar users and venues together should be detected by maximizing intracluster similarity. This objective function is formulated as

$$\text{Odj}=\arg\max_{c} \sum_{j=1}^{k} \sum_{e_c \in C_j} sim\left(e_c, C_j\right) \quad (1)$$

where k is the number of communities, C = {C_1,C_2, . . . ,C_k} is the detected community set, e_c denotes an edge of community C_j and sim(e_c,C_j) is the similarity between e_c and C_j.

With the above objective function, the key issue is to characterize the similarity between an edge and a community. To this end, we first introduce the definition of edge similarity. In a user-venue check-in network, each edge is associated with a user vertex and a venue vertex. By taking an edge-centric view, each edge can be treated as an instance with its two vertices as features. In other words, the similarity between a pair of edges can be defined as the similarity between the corresponding pair of user vertices and venue vertices as $sim_{edge}(e_i, e_j)$ = F($sim_u(u_i, u_j)$, $sim_v(v_i, v_j)$) (2)

where sim (u_i, u_j) is the similarity between two users, $sim_v(v_i, v_j)$ is the similarity between two venues, and F represents the function used to combine these two similarities. The formalism of F depends on the characteristics of the expected communities and the targeted applications. Considering the similarity trade-off between user mode and venue mode, two widely used formalisms of F are average (i.e., $(sim_u + sim_v)/2$) and multiplication (i.e., $\sqrt{sim_u \times sim_v}$. In this paper, we adopt the second notion to ensure that a pair of edges are of high similarity if and only if they are of high similarity in both user-mode and venue-mode.

Each community contains a set of edges, based on (2), the similarity between an edge e_i and a community C_j is defined as

$$\text{sim}_{ei,Cj} = \frac{1}{|C_j|} \sum_{e_c \in Cj} sim_{edge}(e_i, e_c) \quad (3)$$

where $|C_j|$ refers to the number of edges within community C_j.

The edge similarity is defined based on two mode similarities. We compute the mode similarities by taking into account both intermode and intramode features.

6 USER GROUP ESTIMATION AND RECOMMENDATION SCHEME

The overlapping community detection mechanism is enhanced with recommendation models. The community details are classified with location or regions. The system is enhanced with feature selection and fusion mechanism. The system is improved to provide friends recommendation model with place details. The system integrates the location, time and textual data values for the community detection process. Community classification is performed with sub class details. Customized friendship recommendation process is provided in the system. The system is divided into six major

modules. They are social network data analysis, similarity estimation, newsfeeds analysis, clustering process, community identification and recommendation process.

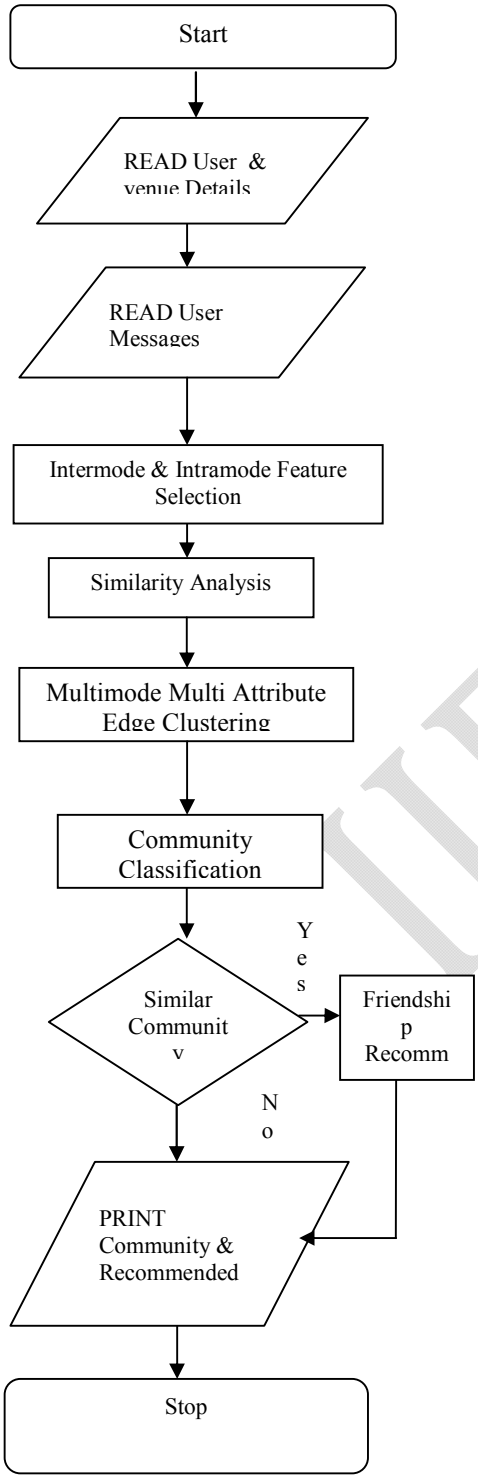

Fig. No: 6.1. User Group Estimation and Recommendation Scheme

6.1 Social Network Data Analysis

User profile, check in details and newsfeeds are collected from social networks. Foursquare and Twitter social network data values are used in the system. User check-in venue and time details are maintained in location data values. User submitted messages are maintained under newsfeeds data collection

.

6.2 Similarity Estimation

Similarity estimation is performed with inter mode and intra mode features. User-Venue relations are analyzed in intermode features. Social influences, geospan and temporal relationships are analyzed using intra mode features. Similarity values are used in the clustering process.

6.3 Newsfeeds Analysis

User messages are analyzed in newsfeeds analysis. Feature selection is performed on textual data values. Noisy data are eliminated from the messages. Term level relationships are analyzed in the similarity estimation process.

Social network data analysis module is used to manage user profile and location details. Inter mode and Intra mode feature analysis is used to estimate the similarity value. Newsfeeds analysis is used to extract features from the textual data values. Clustering process is used group up the similar data values. User groups are identified under the community identification module. Friendship recommendation is carried out under the recommendation process module.

6.4 Clustering Process

Feature integration and separation mechanism are used in the clustering process. Hierarchical multimode multi-attribute edge clustering (*HM*^{2}*Clustering)* algorithm is used to group up the similar users. Location based clustering is performed with user-venue relationships. Hybrid clustering model integrates the user-venue details with message details.

6.5 Community Identification

User groups are identified in the community identification process. Community identification is performed with location and message details. Community identification process is improved with preferences details. Region details are used for the community classification process

.

6.6 Recommendation Process

The recommendation process is used to suggest friends. Region and interest factors are used in the recommendation process. Recommendation

process is customized with user needs. Score based recommendation scheme is used in the system.

7 CONCLUSION

Location based Social Networks (LBSNs) manages the users with their access location details. Community structure identification and profiling operations are carried out in LBSN. Multimode multi-attribute edge-centric coclustering framework is used to discover overlapping and hierarchical communities. The community detection scheme is improved with friend recommendation scheme. Location based recommendation scheme is adopted by the system. The system produces high clustering accuracy levels. Clustering is performed without predefined cluster count. Community discovery accuracy level is improved by the system.

REFERENCES

[1] L. Tang and H Liu, "Community detection and mining in social media," Synthesis Lectures Data Mining Knowledge Discovery, 2010.
[2] S. Fortunato, "Community detection in graphs," Phys. Rep., vol. 486, nos. 3–5, 2010.
[3] A. Noulas, S. Scellato and M. Pontil, "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks," in Proc. ICWSM, 2011.
[4] Eunice E. Santos and John Korah, "Infusing Social Networks With Culture" IEEE Transactions On Systems, Man, And Cybernetics: Systems, January 2014
[5] Michele Nitti and Luigi Atzori, "Trustworthiness Management in the Social Internet of Things" IEEE Transactions On Knowledge And Data Engineering, May 2014
[6] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," Nature, pp. 761–764, 2010.
[7] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora, "Distance matters: Geo-social metrics for online social networks," in Proc. WOSN, 2010.
[8] S. Scellato and C. Mascolo, "Socio-spatial properties of online location-based social networks," in Proc. ICWSM, 2011.
[9] A. Noulas, C. Mascolo, and M. Pontil, "An empirical study of geographic user activity patterns in Foursquare," in Proc. ICWSM, 2011.
[10] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, "Exploring millions of footprints in location sharing services," in Proc. ICWSM, 2011.
[11] M. A. Vasconcelos and V. Almeida, "Tips, dones and todos: Uncovering user profiles in Foursquare," in Proc. WSDM, 2012.
[12] N. Li and G. Chen, "Analysis of a location-based social network," in Proc. CSE, 2009.

AUTHORS BIOGRAPHY

Ms.S.Rabiyathul Basiriya pursuing M.E(CSE) degree in Shree Venkateshwara Hi-Tech Engineering College, Erode, India in 2014 and B.E(CSE) degree from Sri Ramakrishna Institute of Technology, Coimbatore, India in 2013. She has published 1 National conferences, 4 workshops. She is a Member of Computer Society of India(CSI). Her research interests include Data Mining, Networks.

Dr.T.Senthil Prakash received the Ph.D. degree from the PRIST University, Thanjavur, India in 2013 and M.E(CSE) degree from Vinayaka Mission's University, Salem , India in 2007 and M.Phil.,MCA.,B.Sc(CS) degrees from Bharathiyar University, Coimbatore India, in 2000,2003 and 2006 respectively, all in Computer Science and Engineering. He is a Member in ISTE New Delhi, India, IAENG, Hong Kong..IACSIT, Singapore SDIWC, USA. He has the experience in Teaching of 10+Years and in Industry 2 Years. Now He is currently working as a Professor and Head of the Department of Computer Science and Engineering in Shree Venkateshwara Hi-Tech Engineering College, Gobi, Tamil Nadu, and India. His research interests include Data Mining, Data Bases, Artificial Intelligence, Software Engineering etc.,He has published several papers in 17 International Journals, 43 International and National Conferences.

Ms. P. Sudhaselvanayaki received the B.E (CSE) degree from the RVS College of Engineering and Technology, Coimbatore, India in 2009-2013 and pursuing ME (CSE) degree in Shree Venkateshwara Hi-Tech Engineering College, Erode, India in 2013-2015, all in Computer Science and Engineering. She is a Member of Computer Society of India (CSI). Her research interests include Data Bases, Data Mining, and Image Processing. She published 1 National Conferences, 4 Workshops.