ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

WEB SEARCHING IN DATA MINING

Mr. Manimaran.P[1], Ms. Preethi.M[2], Ms. Naresh Kumar.C[3], Ms. Poovizhi.M[4]

[1]Assistant Professor, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
[2,3,4]Students, Department of CSE, K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu, India
Email: manimaran7@yahoo.com, *preethibe9555@gmail.com,naresh.cse95@gmail.com, poovizhim1994@gmail.com*

Abstract

Social networks are constructed to make communication over the group of people. Location Based Social Networks (LBSN) uses the spatial and temporal information for analytic purpose. Profile location and checkin location are analyzed in the LBSN. User and venue relationship are compared with proximity properties. Twofold Search Algorithm (TSA) and Aggregate Index Search (AIS) schemes are applied to discover the friends with social and spatial proximity measures.

Friendship recommendation is an important task in the social networks. The recommendation should by release with spatio temporal factors and similar interest factors. Newsfeeds post by the members are also important criteria for recommendation process. The hierarchical clustering schemes are applied to group up the network structured data values. Edge cluster based recommendation scheme is designed to perform community discovery and friendship recommendation process. The community discovery and recommendation tasks are carried out with the location and newsfeeds details. Hybrid recommendation model suggests suitable friends associated with the user profiles and newsfeeds.

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. 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. 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 to anthropology, biology, economics, geography, organizational studies, social psychology, sociology and sociolinguistics.

The task of classification is to assign an object to a class from a given set of classes based on the attribute values of this object. In spatial classification the attribute values of neighboring objects are also considered. The relevant attributes are extracted by comparing the attribute values of the target objects with the attribute values of their nearest neighbors. The determination of relevant attributes is based on the concepts of the nearest hit the nearest neighbor belonging to the same class and the nearest miss the nearest neighbor belonging to a different class. In the construction of the decision tree, the neighbors of target objects are not considered individually. Instead, so-called buffers are created around the target objects and the nonspatial attribute values are aggregated over all objects contained in the buffer. For instance, in the case of shopping malls a buffer may represent the area where its customers live or work. The size of the buffer yielding the maximum information gain is chosen and this size is applied to compute the aggregates for all relevant attributes. Whereas the neighborhood relations cannot directly express the nearest neighbor, it would be possible to extend the set of neighborhood relation accordingly.

2. Related Work

Detecting communities in a social network is still an open problem in social network analysis. In literature, many community detection methods have been proposed. According to [1], these approaches can be divided into four categories: node-centric, group-centric, network-centric and hierarchic centric. Some popular methods are modularity maximization, Givan-Newman algorithm, Louvain algorithm, clique percolation, link communities [8]. [2] and [9] provide a throughout review of the topic. These methods ignore the attributes of the nodes. Below are some studies that incorporate node attributes in the clustering process. Steinhae user et al. [10] proposed an edge weighting method NAS (Node Attribute Similarity) that takes into account node attributes. A community detection method is then proposed based on random walks. The complexity of the algorithm is O(n2logn) or O(n) where n is the number of nodes. Zhou et al. [11] defined a unified distance measure to combine structural and attribute similarities. Attribute nodes and edges are added to the original graph to connect nodes which share attribute values [7]. A neighborhood random walk model is used to measure the node closeness on the augmented graph. A clustering algorithm SA-Cluster is proposed, following the K-Medoids method. The time complexity of the algorithm is O(n3).

Coupling relationship and content information in social network for community discovery is an emerging research area because current methods do not focus on social graphs or they are not efficient for large-scale datasets.

3. Social and Spatial Proximity Analysis on Social Networks

The emergence of social networks (SNs) brings a new era in the organization and browsing of online information. Manufacturers and service providers are becoming increasingly interested in exploiting popular SNs to promote their products and services. Recently, Microsoft's search engine has integrated social information from Facebook to return webpages that are popular among the friends of users [3]. Studies like have investigated the influence between users of SNs and quantified the probability of a user performing an action after his/her friend(s) did. Current text search systems have also incorporated social influence into query processing by taking into account friend relationships for the ranking of documents/objects [4].

Location-based services are an indispensable feature in SNs. This fact becomes increasingly prominent as the number of users who access SN applications on mobile devices is growing steadily. The most popular SN, Facebook, includes a set of location-based features, while others are explicitly based on the management of user locations. Motivated by this trend, we investigate the integration of social and spatial information in a single query. Consider a service like badoo.com, where a user u1 who is looking for company to have lunch or watch a movie, may browse the profiles of nearby users and invite them to join him/her. Existing systems apply a traditional k-nearest neighbor query [5], potentially with some binary conditions, to provide u1 with the profiles of users in the vicinity. While recommended users are indeed near u1 geographically, his/her true preferences of companions would be better captured if SN information was also taken into account. Assume, for example, that the users' euclidean coordinates and social connections are and 1b respectively. The closest user to u1 in the spatial domain is u5. U4 might be a better match because he locates only slightly farther but is "closer" in the social network. Conversely, the closest user socially (u2) may be too far spatially. Therefore, to provide meaningful recommendations, both social proximity and spatial proximity should be incorporated into the search [6]. In this paper we propose and study the social and spatial ranking query (SSRQ). SSRQ reports the top-k users in the SN based on a ranking function that incorporates social and spatial distance from the query user. Our key contributions are:

- We conduct the first study on a joint search by social and spatial user proximity.
- We propose a suite of processing methodologies, including a highly scalable and robust approach that relies on indexing and social summaries.
- We equip the latter with sophisticated optimizations, based on computation sharing, intermediate result caching and an accuracy-enhancing strategy that complements social summaries in proximity estimation.
- We use real SN data to experimentally evaluate our algorithms.

4. Community Discovery in Social Network

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 and Group 2. 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, Group 2 and Group 3. In this case, even though Tom, David, Bob and Eva have similar check-in patterns, they are 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. 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 and Group 2. 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, Group 2 and Group 3. 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 are 1000 km and 800 km, while Bob and Eva mainly stay locally whose rg are 80 km and 60 km, respectively. 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.

It is more reasonable to exploit both the structural information and the node attributes 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. 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. Each community explicitly 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.

5. Hierarchical Data Clustering Concepts

Clustering is the task of assigning a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, information retrieval and bioinformatics.

Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix. Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion to use. Popular choices are known as single-linkage clustering, complete linkage clustering or UPGMA (Unweighted Pair Group Method with Arithmetic Mean). Furthermore, hierarchical clustering can be computed agglomerative or divisive.

While these methods are fairly easy to understand, the results are not always easy to use, as they will not produce a unique partitioning of the data set, but a hierarchy the user still needs to choose appropriate clusters from. The methods are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge. In the general case, the complexity is $o\left(n^3\right)$, which makes them too slow for large data sets. For some special cases, optimal efficient methods are known: SLINK for single-linkage and CLINK for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did provide inspiration for many later methods such as density based clustering.

6. Hierarchical Cluster based Recommendation Scheme

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, Proximity Analysis, Newsfeeds Analysis, Clustering Process, Community Identification and Recommendation 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 proximity measure. Newsfeeds analysis is used to extract features from the textual data values. Clustering process is applied 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.

6.1. 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. Proximity Analysis

Proximity estimation is performed with inter mode and intra mode features. User-Venue relations are analyzed in inters mode features. Social influences, geospan and temporal relationships are analyzed using intra mode features. Proximity measures are used in the index and query 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.

6.4. Clustering Process

Feature integration and separation mechanism are used in the clustering process. Edge 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 adapted to suggest friends. Region and interest factors are considered in the recommendation process. Recommendation process is customized with user needs. Score based recommendation scheme is used in the system.

7. Conclusion

Social network analysis models are build to recommend the suitable friends for the users. User profile and location factors are normally used for the recommendation process. The social and spatial proximity measures are used to find out the friends. User newsfeeds are also integrated with the system for the community discovery and recommendation process. Location based recommendation scheme is proposed to find out the friends associated with the user location and community levels. High clustering accuracy levels are achieved by the edge clustering scheme. Clustering is performed without predefined cluster count under hierarchical clustering environment. Community discovery accuracy level is improved by the system. Indexing operations are adapted to upgrade the clustering process.

References

[1] L. Tang and H. Liu, Community Detection and Mining in Social Media (Synthesis Lectures on Data Mining and Knowledge Discovery). Morgan Claypool, 2010.

[2] S. Fortunato, “Community detection in graphs,” Physics Reports 486, 75-174, 2009.

[3] M. Helft. (2011, May). Bing taps facebook data for fight with google. [Online]. Available: http://bits.blogs.nytimes.com/2011/05/16/

[4] P. Yin, W.-C. Lee and K. C. K. Lee, “On top-k social web search,” in Proc. 19th ACM Int. Conf. Inform. Knowl. Manage., 2010.

[5] D. Papadias, M. L. Yiu, N. Mamoulis and Y. Tao, “Nearest neighbor queries in network databases,” in Encyclopedia GIS, New York, NY, USA: Springer, 2008, pp. 772–776.

[6] Kyriakos Mouratidis, Jing Li, Yu Tang and Nikos Mamoulis, “Joint Search by Social and Spatial Proximity”, IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 3, March 2015

[7] The Anh Dang, Emmanuel Viennet, “Community Detection based on Structural and Attribute Similarities”, ICDS 2012: The Sixth International Conference on Digital Society, IARA, 2012.

[8] Y.-Y. Ahn, J. P. Bagrow and S. Lehmann, “Link communities reveal multiscale complexity in networks,” Nature, vol. 466, no. 7307, pp. 761–764, Jun. 2010.

[9] J. Leskovec, K. J. Lang and M. W. Mahoney, “Empirical comparison of algorithms for network community detection,” CoRR, vol. abs/1004.3539, 2010.

[10] K. Steinhaeuser and N. V. Chawla, “Identifying and evaluating community structure in complex networks,” Pattern Recognition Letters, Nov. 2009.

[11] Y. Zhou, H. Cheng and J. X. Yu, “Graph clustering based on structural/attribute similarities,” Proc. VLDB Endow., vol. 2, pp. 718–729, August 2009.