Identifying Community For Important Intensions In Complex Data Structure On The Online Social Networks

Gowthami U .[1], Laura Juliet P.[2]

[1] *Research Scholar, Department of Computer Science, Vellalar College for Women Tamilnadu, India*
[2] *Assistant Professor, Department of Computer Applications, Vellalar College for Women Tamilnadu, India*

---***---

***Abstract** - In Social Media, large multidimensional data's exploring and gaining the interest in mining the community based on the information.Here many research focus on community discover based on keywords and entities through affinity calculation. Unfortunately, the Community discover process-mining approaches do not take into account hidden aspect of intentions behind the data sharing in user activities, recognizing and detecting the hot topics in the network about public opinion on the focus of the community discovery. By using HMM a community is constructed starting with a seed consisting of one or more items of the entities believed to be participating in a viable community. Given the seed item, we iteratively adjoin new items by evaluating the affinity between the items to build a community in the network. As there are multiple interactions among the items from different dimensions/entities in a multidimensional network, the main challenge is how to evaluate the affinity between the two items in the same type of entity (from the same dimension/entity) or in different types of entities (from different dimensions/entities)..Inter behavior and intra behavior of user is obtained .Multimodal perspective is used to avoid the co clustering issues. Data Redundancy is eliminated among the communities. Multiple intentions are taken for clustering.*

***Key Words:** social network,* Spectral Clustering,hidden markov model.

1. INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Data mining involves integration of techniques from multiple disciplines such as database technology, statistics, machine learning, neural networks, information retrieval, etc. Data mining is the process of discovering meaningful patterns and relationships that lie hidden within very large databases. Data mining is a part of a process called knowledge discovery in databases (KDD). This process consists basically of steps that are performed before carrying out data mining, such as data selection, data cleaning, pre-processing, and data transformation.

There are many other terms carrying a similar or slightly different meaning to data mining such as knowledge mining from databases, knowledge extraction, Data/pattern analysis, Data archaeology and Data dredging. A standard definition for data mining is the non-trivial extraction of implicit, previously unknown, and potentially useful knowledge from data.

Hidden Markov Model

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be presented as the simplest dynamic Bayesian network. Markov models, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly. It is stochastic process that satisfies the Markov property. A Markov process can be thought of as 'memory less': loosely speaking, a process satisfies the Markov property if one can make predictions for the future of the process based solely on its present state just as well as one could knowing the process's full history. i.e., conditional on the present state of the system, its future and past are independent.

2. RELATED WORK

To better understand of community discover process, it is useful to review and examine the existing research works in literature. Therefore, recent approaches and methodologies used for community discover have been discussed.

Dinesh Babu Jayagopi and Daniel Gatica-Perez [2010] [5] have proposed the automatic discovery of group conversational behaviour is a relevant problem in social computing. We analyse an approach to address this problem by defining a novel group descriptor called bag of group-nonverbal-patterns (NVPs) defined on brief observations of group interaction, and by using principled probabilistic topic models to discover topics. The proposed bag of group NVPs allows fusion of individual cues and facilitates the eventual comparison of groups of varying sizes. The use of topic models helps to cluster group interactions and to quantify how different they are from each other in a formal probabilistic sense. Results of behavioural topics discovered on the Augmented Multi-Party Interaction (AMI) meeting corpus are shown to be meaningful using human annotation with multiple observers. Proposed method facilitates "group behaviour-based" retrieval of group conversational segments without the need of any previous labelling

Vasanthan Raghavan,Aram Galstyan, and Alexander G.Tartakovsky1[2013] [4] have focus on this work is to analyse probabilistic models for temporal activity of users in social networks (e.g., posting and tweeting) by incorporating the social network influence as perceived by the user. Although prior work in this area has developed sophisticated models for user activity, these models either ignore social network influence completely or incorporate it in an implicit manner. We overcome the no transparency of the network in the model at the individual scale by proposed a coupled hidden Markov model (HMM), where each user's activity evolves according to a Markov chain with a hidden state that is influenced by the collective activity of the friends of the user. We utilize generalized Baum-Welch and Viterbi algorithms for parameter learning and state estimation for the proposed framework. We then validate the proposed model using a significant corpus of user activity on Twitter.

Andrew Mehler and Steven Skiena[2009] [1] has proposed a Network communities refer to groups of vertices within which their connecting links are dense but between which they are sparse. A network community mining problem (or NCMP for short) is concerned with the problem of finding all such communities from a given network. A wide variety of applications can be formulated as NCMPs, ranging from social and/or biological network analysis to web mining and searching. Network communities and their properties based on the dynamics of a stochastic model. Relationship between the hierarchical community structure of a network and the local mixing properties of such a stochastic model has been established with the large-deviation theory. Topological information regarding to the community structures hidden in networks can be inferred from their spectral signatures. Based on the above-mentioned relationship, this work proposes a general framework for characterizing, analyzing, and mining network communities.

Lei Tang, Huan Liu, Jianping Zhang[2012] [2] explore a multimode network consists of heterogeneous types of actors with various interactions occurring between them. Identifying communities in a multimode network can help understand the structural properties of the network, address the data shortage and unbalanced problems, and assist tasks like targeted marketing and finding influential actors within or between groups. In general, a network and its group structure often evolve unevenly. In a dynamic multimode network, both group membership and interactions can evolve, posing a challenging problem of identifying these evolving communities. In this literature, we try to address this problem by employing the temporal information to analyze a multimode network. A temporally regularized framework and its convergence property are carefully studied.

Guanfeng LIU,Yan Wang,Mehmet A. Orgun,Ee Peng LIM[2013] [3] presents a Online Social networks which provided the infrastructure for a number of emerging applications in recent years, e.g., for the recommendation of service providers or the recommendation of files as services. In these applications, trust is one of the most important factors in decision making by a service consumer, requiring the evaluation of the trustworthiness of a service provider along the social trust paths from a service consumer to the service provider. However, there are usually many social trust paths between two participants who are unknown to one another. In addition, some social information, such as social relationships between participants and the recommendation roles of participants, has significant influence on trust evaluation but has been neglected in existing studies of online social networks. Furthermore, it is a challenging problem to search the optimal social trust path that can yield the most trustworthy evaluation result and satisfy a service consumer's trust evaluation criteria based on social information.

Overview of Existing system:

In the exsisting scenario a framework (MultiComm) to identify a seed-based community in a multi-dimensional network by evaluating the affinity between two items in the same type of entity or different types of entities (different dimensions) from the network. To calculate the probabilities of visiting each item in each dimension, and compare their values to generate communities from a set of seed items. In order to evaluate a high quality of generated communities and to study a local modularity measure of a community in a multi-dimensional network. Experiments based on synthetic and real-world data sets suggest that the framework is able to find a community effectively. Experimental results have also shown that the performance of the algorithm is better in accuracy than the other testing algorithms in finding communities in multi-dimensional networks. On the other hand, in social networks, user actions are constantly changing and co-evolving. here,it is required to adapt the proposed model to be time-varying. Hence, to overcome all these issues the Hidden markov model(HMM) algorithm is utilized in the current work.

3. PROPOSED METHOD

A Hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be presented as the simplest dynamic Bayesian network. Markov models , the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

It is stochastic process that satisfies the Markov property. A Markov process can be thought of as 'memory less': loosely speaking, a process satisfies the Markov property if one can make predictions for the future of the process based solely on its present state just as well as one could knowing the process's full history. i.e., conditional on the present state of the system, its future and past are independent.
Our HMM representation is defined in terms of the following vector and matrix quantities:

[P1]i = Pr[x1 = i]
[P2;1]ij = Pr[x2 = i; x1 = j]
[P3;x;1]ij = Pr[x3 = i; x2 = x; x1 = j] 8x 2 [n];

where P1 2 Rn is a vector, and P2;1 2 Rn_n and the P3;x;1 2 Rn_n are matrices. These are the marginal probabilities of observation singletons, pairs, and triples.

4. Methodology

The modules in the current work are listed below:

- Establish the online social network.
- Synthesis data preprocessing.
- Partitioning of the network using Spectral Methods.
- Classify the training data.
- Affinity value is calculated.
- Eliminate the overlapping Profiles using learning algorithm.
- Establishing a Complex data Analysis technique using Hidden Markov model.
- Analyses the result
-

4.1 Establishing Online Social Network or Data Pre-processing of OSN dataset

Here, we either build the online Social network to construct the profile structure to yield the profile data with multidimensional fields which is considered to the synthesis dataset. Another way to gather is multidimensional data for a community discovery process is through real dataset from the Online Social network . Data Pre-processing involves stemming and Stop word removal process.

4.2 Partitioning of the network using Spectral Methods

The Social data network make use of eigenvectors and matrix representations of the network. We show that with certain choices of free parameters appearing in these spectral algorithms within the spectral approximations used here, there is no difference between the modularity and inference-based community detection methods, or between either graph partitioning.

- Apply Spectral similarity-based clustering to nodes.

- Vertex similarity is defined in terms of the similarity of their neighborhood.

- Structural equivalence: two nodes are structurally equivalent if they are connecting to the same set of actors.

- Optimal solution: top eigenvectors with the smallest Eigen values.

- Structural equivalence is too restricting for practical use.

4.3 Establishing a MultiComm community discovery model based on Affinity calculation

Data is partitioned as tensors .In this probability Distribution is used to calculate maximum likelihood between the data affinity. We consider a random walker choose randomly among the available interactions between the items in different dimensions, and makes a choice with probability and going back to a set of items in the current community. A community is constructed starting with a seed consisting of one or more items of the entities believed to be participating in a viable community. Given the seed item, we iteratively adjoin new items by evaluating the affinity between the items to build a community in the network. Based on their probability values, we can determine the candidate items in different dimensions that are closely related to the current items in the community.

Algorithm to Discovery the Community based Multi dimensional data

Input: Data Source – User Details and Activity formed in terms of Multidimensional data

Process:

- ➢ Classify the user details and activity based on the different constraints.
- ➢ Constraints have modeled as learning algorithm.
- ➢ Classify the Training data into class based on the attributes of the Dataset.
- ➢ Classify the attribute based on Domain Knowledge and Value types.
    - Value types = {Single Value Attribute, Two Value Attribute …. Multi- Value attributes}
    - Domain Knowledge = {Personal info, Employment, Lifestyle, Sports, Entertainments, cuisines}
- ➢ Inference of the Data through Application is carried out.
- ➢ Application analysis is carried out for behaviors.
- ➢ Application Gain is calculated based on the attributes.
    - Behaviors = Trained Data of the learning Algorithm.
    - Behaviors = No. of. Similarities between the attribute and data source.
- ➢ Affinity value is calculated based probability tensor.
- ➢ Probability is denoted P
    - $P= (1-\alpha)\, W_p + \alpha\, e_I$ -------> [1]
    - Above equation is a steady state probability.
    - Where W is weighted matrix associated graph details of the User profiles.
    - Sensitivity is classified based on the application characteristic sand behavior valuate.
    - Largest value in the W and α leads to the maximum Support to form the Community group.
- ➢ Learning Algorithm extracts data based on the application category.

Output: Secured disclosure of the information

In this algorithm, the computations require several iterations, through the collection to adjust approximate probability values of items of the entities in the multidimensional network to more closely reflect their theoretical true values. When communities vary in different subsets of dimensions, we can make use of affinity to identify which dimension of its corresponding item with the highest probability joins in the community.

4.4 Eliminate the overlapping Profiles

Due to the characteristic of various similarity features, different calculation methods might be used which lead to the different value ranges. Therefore, the absolute values of different features must be normalized. Classical co clustering is one way to conduct this kind of community partitioning. Clauset defined a measure of community structure for a graph. The idea is that a good community should have a sharp boundary, i.e., it will have few connections from its boundary to the other portion of the network, while having a greater proportion of connections from the boundary back into the community. Here we extend this idea to define a local modularity of a community in a multi-dimensional network. The identified communities are disjointed, which contradicts with the actual social setting. Edge clustering has been proposed to detect communities in an overlapping manner.

Normalized Mutual Information

- Entropy: the information contained in a distribution.
- Mutual Information: the shared information between two distributions.

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p1(x)p2(y)}\right)$$

- Normalized Mutual Information (between 0 and 1).
- Consider a partition as a distribution (probability of one node falling into one community), we can compute the matching

between the clustering result and the ground truth.

$$NMI(\pi^a,\pi^b)=\sum_{h=1}^{k(a)}\sum_{l=1}^{k(b)} nh,l\log(\frac{n.nh.1}{nh^{(a)}\cdot nl^{(b)}})/$$
$$\sqrt{(\sum_{h=1}^{k(a)} n_h^{(a)}}\log\frac{n_h^a}{n})(\sqrt{(\sum_{l=1}^{k(b)} n_l^{(b)}}\log\frac{n_l^b}{n})$$

4.5 Establishing a Complex data Analysis technique using Hidden Markov model

We assume that the intentional process models underlying user activities by using Intention mining techniques can discover the important information and entities as a community to gather the large members and exploit information's. The aim of this paper is to propose the use of probabilistic models to evaluate the most likely intentions behind traces of unobserved activities and mixture information's in the complex data structures, namely Hidden Markov Models (HMMs).

Algorithm

LEARNHMM(m;N):

Inputs: m - number of states, N - sample size
Returns: HMM model parameterized by fbb1;bb1; b Bx 8x 2 [n]g

1. Independently sample N observation triples (x1; x2; x3) from the HMM to form empirical estimates
b P1; b P2;1; b P3;x;1 8x 2 [n] of P1; P2;1; P3;x;1 8x 2 [n].

2. Compute the SVD of b P2;1, and let bU be the matrix of left singular vectors corresponding to the m
largest singular values.

3. Compute model parameters:
(a) bb1 = bU> b P1,

(b) bb1 = (b P>bU)+P1,

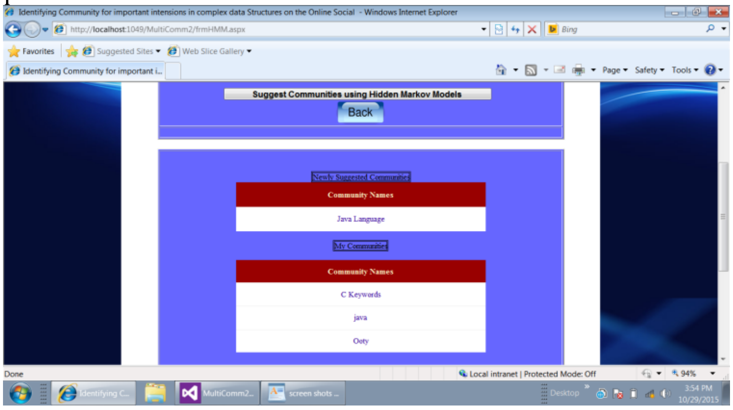
(c) b Bx = bU> b P3;x;1(bU> b P2;1)+ 8x 2 [n].

The random sampling, the running time of the learning algorithm is dominated by the SVD computation of an n_n matrix. The time required for computing joint probability calculations is O(tm2) for length t sequences—same as if one used the ordinary HMM parameters (O and T). For conditional probabilities, we require some extra work (proportional to n) to compute the normalization factor. However, our analysis shows that this normalization factor is always close to 1, so it can be safely omitted in many applications.
This algorithm is used to identify hidden activities of the user.so by using this algorithm we can able to find the community of the user. Finally concluded the proposed work yields superior performance rather than existing work.

5. RESULTS AND DISCUSSION

A framework (HMM) is proposed to determine communities in a multi-dimensional network based on probability distribution of each dimension/entity computed from the network described in Both theoretical and experimental results have demonstrated that the proposed algorithm is efficient and effective. Performance of the proposed System is determined through the following parameters.

Precision, Recall, F Measure and NMI are calculated in terms of the ground-truth community. We construct one "ground-truth" community and avoid profile overlapping, and then check how HMM can recover this community and after that community prediction is done.


Identifying Community for important intensions in complex data Structures on the Online Social - Windows Internet Explorer

http://localhost:1049/MultiComm2/frmHMM.aspx

Suggest Communities using Hidden Markov Models

Back

Newly Suggested Communities

Community Names
Java Language

My Communities

Community Names
C Keywords
java
Ooty

Figure 5.1 Newly Suggested Groups

Community is obtained based on the HMM through the probabilistic models with latent variables hidden with less proportion in the comments and stories. Filtering is carried using the latent analysis over the hidden data's. Process outline is depicted. In the existing system the community prediction is done and the quality of the community is predicted. In the proposed system based the likes the new community will be suggested to the user based on the HMM.

Performance Analysis Of The Community Discovery Frame Work

Technique	Existing (Multicomm)	Proposed(HMM)
Precision	0.92	0.94
Recall	0.90	0.90
F Measure	0.91	0.92
NMI	0.11	0.65

Table 5.1 Performance Analysis

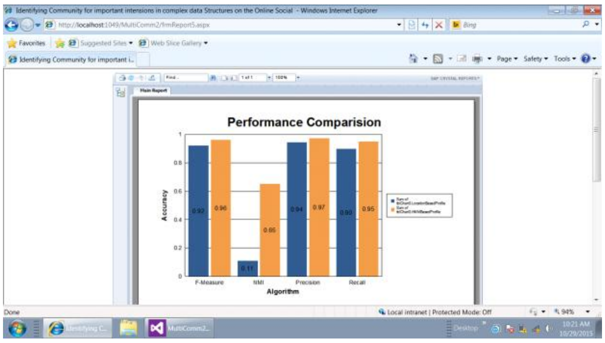

Identifying Community for important intensions in complex data Structures on the Online Social - Windows Internet Explorer
http://localhost:1049/MultiComm2/frmReport5.aspx
Favorites
Suggested Sites
Web Slice Gallery
Page
Safety
Tools
Main Report
Performance Comparision
Accuracy
1
0.8
0.6
0.4
0.2
0
0.92
0.96
0.11
0.65
0.94
0.97
0.90
0.95
F-Measure
NMI
Precision
Recall
Algorithm
Done
Local intranet | Protected Mode: Off
94%
10:21 AM
10/29/2015

Figure 5.2 Performance Analysis

This above table 5.1 and fig 5.2 illustrate that proposed technique has some advantages over the Multicomm algorithms that is one with no direct interaction between the same entities; the second is that the interactions are duplicated. Local modularity changes with respect to the number of items joined in the community on two generated multi-dimensional networks. As each of these two multi-dimensional networks is represented by multiple tensors, here the local modularity refers to the average value of local modularity's corresponding to these tensors.

Consistency score defines the magnitude of associatively between entities in entity-set data. Here, we study the effect of consistency threshold in the formation of communities, in terms of F-measure with respect to task of entity recommendation. As the threshold is increased, the number of edges in the co-occurrence consistency graph would decrease, resulting in drop in density of graph as well as decrease in size and number of communities discovered from the graph.

5. Conclusion and Future Work

A framework (HMM) is proposed to determine communities in a multi-dimensional network based on probability distribution of each dimension/entity computed from the network. The Proposed algorithm is better in accuracy than the other testing algorithms in finding communities. Inter behavior and intra behaviors of user are obtained. Multimodal perspective is used to avoid the co clustering issues. Data Redundancy is eliminated among the communities. Multiple intentions are taken for clustering. Both theoretical and experimental results have demonstrated that the proposed algorithm is efficient and effective. On the other hand, in social networks, user actions are constantly changing and co-evolving.

FUTURE WORK

This research work can be enhanced in the future with the following scopes:

- Comparing the two CNM (Clasuet Newman Moore)-based policy management model enhancements (Assisted Friend Grouping and Example Friend Selection) in terms of policy definition, openness, and their human effects.
- To develop a prototype that leverages user privacy sentiment for the mass customization of a privacy management model.

REFERENCES

1. **Andrew Mehler And Steven Skiena**" Expanding Network Communities from Representative Examples" P1: vlmacmb126a-07 acm-transaction March 26, 2009.
2. **Lei Tang, Huan Liu, Jianping Zhang**" Identifying Evolving Groups in Dynamic Multimode Networks" ieee transactions on knowledge and data engineering, vol. 24, no. 1, january 2012.
3. **Guanfeng LIU,Yan WANG,Mehmet A. Orgun,Ee Peng LIM**" Finding the Optimal Social Trust Path for the Selection of Trustworthy Service Providers in Complex Social Networks" ieee transactions on services computing, vol. 6, no. 2, april-june 2013
4. **Vasanthan Raghavan, Greg Ver Steeg, Aram Galstyan, and Alexander G. Tartakovsky** "Modeling Temporal Activity Patterns in Dynamic Social Networks" arXiv:1305.1980v1 [physics.soc-ph] 9 May 2013.
5. **Dinesh Babu Jayagopi and Daniel Gatica-Perez**" Mining Group Nonverbal Conversational Patterns Using Probabilistic Topic Models" ieee transactions on multimedia, vol. 12, no. 8, december 2010.