International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Web User Profile Inference for User Group Interest Prediction on Social Networks using Domain Ontology

M. Mohamed Iqbal Mansur[1], Dr. C. Kavitha[2], Dr. K. Thangadurai[3]

[1]Assistant Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Karur - 639 005, Tamilnadu, India.

[2]Assistant Professor, Department of Computer Science, Thiruvalluvar Govt Arts College, Rasipuram, Tamil Nadu, India.

[3]Assistant Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Karur - 639 005, Tamilnadu, India.

Abstract: ***Web inference techniques have become more sophisticated and which can be used for many real world applications like market strategies, business intelligence and etc… In social networks the interest of a single user represents the interest of the whole group to which he belongs to. There exist various approaches to find out the interest of a group, but suffers with the accuracy of clustering users into groups and predicting their interest. We propose a new methodology to predict the interest groups on social networks using domain ontology and their profile details. Each user would be communicating with others in the group or network about some topic and we consider that a single user has N number of interest or topic of conversation. Our ultimate aim is to group similar interested user's and their interest to infer some valuable knowledge from clustered results. Each user conversation and logs are retrieved to get set of conversation they have with others. From each log the topic of conversation and interest is identified using domain ontology. According to the classes available in the domain ontology the user's are grouped to form a cluster. The generated clusters are validated by computing overlap measure and the process is iterated till the overlap measure becomes low. The domain ontology has number of classes and each class has different labels which represent the properties and values of domain attributes. The proposed method has produced efficient clusters and the prediction accuracy is also higher.***

Keywords: Multimode Networks, Social Networks, Temporal Data, Community Evolution, Data Mining.

1. Introduction

The usage of social network has been increased due to the development of information technology, where the users can share any type of information and their opinions and others could view and write comments on the information. The social networks can be adapted to generate any business intelligence or towards any field as a supporting factor. For example marketing company could use the business intelligence generated from the interactions and actors from the social network in order to increase the market of a product. Just like a voting can be conducted on the forum of social network to which the other actors will write comments as an interaction. Using the votes we can come to conclusion about the contest.

The social networks like face book, twitter, YouTube are considered in this paper. An real time example of social network is you tube, where a single user posts a video and others in the network simply views the video , writing comment on that and uploading another video as a reply for that. What happens here is, a set of user groups writing comments on particular topic, and uploading videos which shows their interest on the topic or subject. The ultimate aim of this paper is to identify and predict the user interest from the web logs generated from earlier conversation patterns.

For any user from social network, he belongs to a group and shares informations within the group. Also, the interest of the group are same for some extent and mining information from such user group is more important in many market strategies. The interest of single user implicates the interest of the whole group and we consider such a factor to mine information to support business intelligence and other usages. The inference achieved from such behavior patterns of social user could be used for market analysis and other functionalities.

Once the community of any actor is identified then the ultimate interest of the actor could be identified. We focus on identifying the interest of the actor using the inference model based on social ontology. The social ontology has set of classes and relations with some attributes. A single attribute or class represent an interest and the user can be identified as he is interested in particular topic if the behavior pattern contains the social relation. For example, from figure1, the users A,B,C has conversation about different topics like Data Mining, Image Processing, Natural language processing (NLP). It shows that the users A, B and C have conversation with each other but the topic differs. The user A has conversation in all the three topics with B and C, but B has conversation in NLB with A and C has conversation in NLP with B and NLP , Multimedia with A.

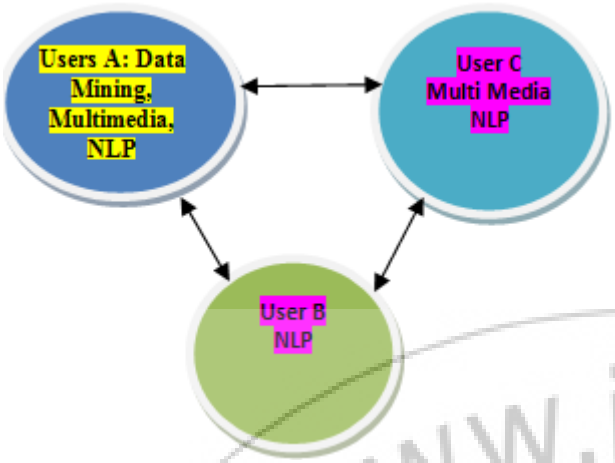

Figure1: Simple interaction within users of social group

From figure 3, out scope is to identify the user interest from the social data, which could be used for any market analysis or some other usage. We design such a framework to infer useful knowledge from social data and our approach converts the social data into social behavior pattern and computes weight according to the occurrence of relations. We discuss the proposed method in detail at later stage in this paper.

2. Related Works

The grouping problem is discussed by different researchers and we discuss few of them here. Estimation and Prediction for Stochastic Block structures [1] ,proposes a posteriori block modeling for graph. The model assumes that the vertices of the graph are partitioned into two unknown blocks and that the probability of an edge between two vertices depends only on the blocks to which they belong. Statistical procedures are derived for estimating the probabilities of edges and for predicting the block structure from observations of the edge pattern only. ML estimators can be computed using the EM algorithm, but this strategy is practical only for small graphs. A Bayesian estimator, based on the Gibbs sampling, is proposed. This estimator is practical also for large graphs. When ML estimators are used, the block structure can be predicted based on predictive likelihood. When Gibbs sampling is used, the block structure can be predicted from posterior predictive probabilities.

Orthogonal Nonnegative Matrix t-Factorizations for Clustering [2] , provide a new approach of evaluating the quality of clustering on words using class aggregate distribution and multi-peak distribution. We also provide an overview of various NMF extensions and examine their relationships.

Community Evolution in Dynamic Multi-Mode Networks [3], address the problem of identifying group membership and interactions by employing the temporal information to analyze a multi-mode network. A temporally-regularized framework and its convergence property are carefully studied. We show that the algorithm can be interpreted as an iterative latent semantic analysis process, which allows for extensions to handle networks with actor attributes and within-mode interactions. Experiments on both synthetic data and realworld networks demonstrate the efficacy of our approach and suggest its generality in capturing evolving groups in networks with heterogeneous entities and complex relationships.

Relational Learning via Latent Social Dimensions [4], propose to extract latent social dimensions based on network information first, and then utilize them as features for discriminative learning. These social dimensions describe different affiliations of social actors hidden in the network, and the subsequent discriminative learning can automatically determine which affiliations are better aligned with the class labels. Such a scheme is preferred when multiple diverse relations are associated with the same network. We conduct extensive experiments on social media data (one from a real-world blog site and the other from a popular content sharing site).

Toward Collective Behavior Prediction via Social Dimension Extraction [6], present an innovative algorithm that deviates from the traditional two-step approach to analyze community evolutions. In the traditional approach, communities are first detected for each time slice, and then compared to determine correspondences. This approach is inappropriate in applications with noisy data. The FacetNet for analyzing communities and their evolutions through a robust unified.

A Bayesian Approach Toward Finding Communities and Their Evolutions in Dynamic Social Networks [7], propose a dynamic stochastic block model for finding communities and their evolutions in a dynamic social network. The proposed model captures the evolution of communities by explicitly modeling the transition of community memberships for individual nodes in the network. Unlike many existing approaches for modeling social networks that estimate parameters by their most likely values (i.e., point estimation), in this study, a Bayesian treatment for parameter estimation that computes the posterior distributions for all the unknown parameters is employed. This Bayesian treatment allows us to capture the uncertainty in parameter values and therefore is more robust to data noise than point estimation. In addition, an efficient algorithm is developed for Bayesian inference to handle large sparse social networks.

Constant-Factor Approximation Algorithms for Identifying Dynamic Communities [8], propose a method to generate the structure of network based on maximum weight bipartite matching. We use a similar idea to design an approximation algorithm for the general case where some individuals are possibly unobserved at times, and to show that the approximation factor increases twofold but remains a constant regardless of the input size. This is the first algorithm for inferring communities in dynamic networks with a provable approximation guarantee.

The above discussed methods are focused on identifying the groups and interactions as static or dynamic but suitable for smaller attributes and actors. Also in most of the methods the weight for the actor or interaction is assigned in a static manner which reduces the performance of the group identification. We propose a new method to solve the

problem of identifying dynamic groups and interactions with dynamic weight allocation process using social graphs.

3. Proposed Model

The proposed model has four stages namely: Preprocessing, , Social Graph Construction, clustering, and Interest Identification.

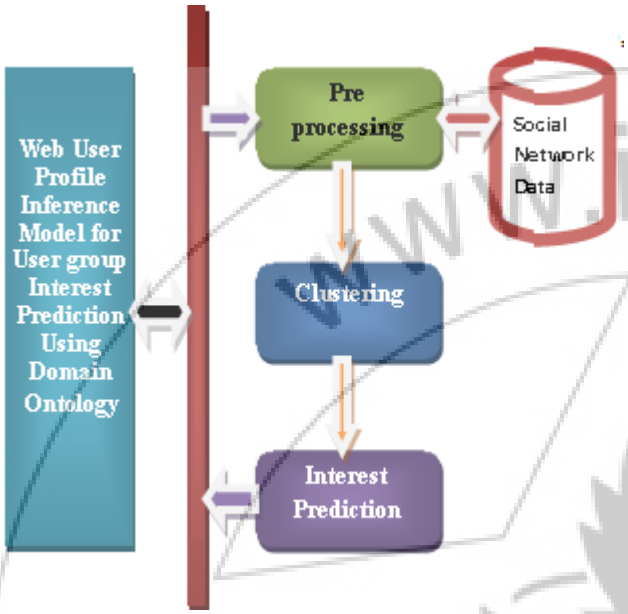

Figure 2: Architecture of Web User Profile Inference Model

3.1 Preprocessing

At the preprocessing phase, the input data set "Enron", is converted into the processing form of the proposed model. Each conversation from data set D_s is read and textual information from each document D_i is retrieved and generates a term set T_s. From the term set T_s , unnecessary words are removed as stop words and verbs also identified using standford part of speech tagger and removed from the term set. With the remaining terms in the term set T_s, stemming process is performed to get pure nouns from the terms in the term set. The selected pure nouns are used to calculate other measures.

Algorithm

Step1: for each document D_i from Document Set D_s.
Extract text content from Document D_i.
Split text into Paragraphs.
Split paragraph into statements.
Remove punctuation marks.
Split text into individual terms and collect as term set T_s.
Remove stop words from the term set T_s.
For each term in the term set T_s.
Use pos tagger to identify verb/noun.
If verb
Remove from Term set T_s.
Else
Perform stemming process.
End.
Step 2. Return Terms Set T_s.

3.2 Clustering

The clustering operation is performed using social ontology. The social ontology has many relations and classes., the user converstion messages has top terms and class. We identify the classes from the term set Ts and compute how many relations it has. From identified relations, semantic class measure and semantic depthness measure is computed. Using both the measure we compute the semantic weight measure, which represent the interest of the message.

Algorithm

Step1**:** start
Step2: read semantic ontology set O, read data set Ds, initialize term set Ts, Weight matrix W.
Step3: for document D_i from Ds
Ts = preprocess Di.
For each category Ct from ontology set O.
Compute semantic weight measure Sw.
Sw = CSM+SDM.
CSM=no. of class names present in term set TS/
Total number of classes in O.
SDM = No of sub class names present in Ts/size of O.
W = W+{ct,SDM, En}.
end
End.
Step4: stop.

3.3 Interest Identification

The interest of the user is identified using the semantic weight computed in the previous phase of the proposed method. In the previous step we have computed semantic weight for each entity of the mode. There may be number of entity of an user with different interests. We have to identify the persistent interest , so that the group evolution could be performed effectively. We identify the interest of the user at each time frame with the computed semantic weight and selects the interest which is occurring at all the user patterns.

Algorithm:

Step1: start
Step2: read computed semantic weight SW, initialize interest set Is..
Step3: for each message from data set *Ds*
Identify interest with more weight.
Is =Is+ $Sw(A) \times T_w$.
End.
End.
Step4: Group similar user interest and users.
Step5: Identify set of interest occurs at all user pattern.
Step4: stop.

4. Results And Discussion

We have used following data set for the evaluation of the proposed approach. The table 1, shows the data set used and proposed solution has been implemented and evaluated using different data sets. We used both DBLP and ENRON data sets which are openly available.

Table1: shows the data set used.

Data Set	Number of Users	Number of Messages
Enron	2359	32,789
DBLP	343, 103	491726
Blog Catalog	32,700	4,52,000

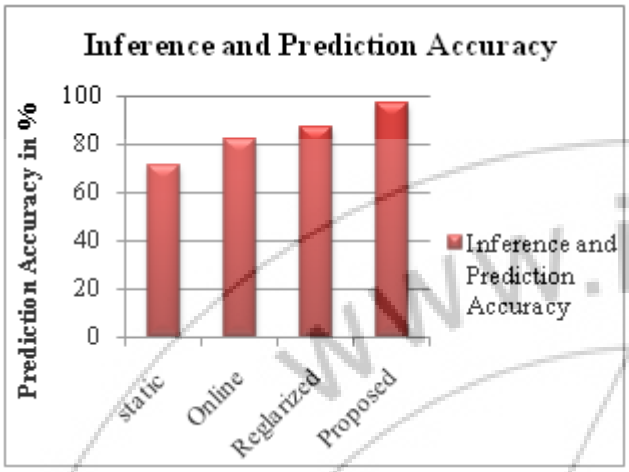

Figure 3: comparison of prediction accuracy of different methods.

The figure 3 shows the prediction accuracy produced by different methods like static, online, regularized cluster and semantic ontology based web inference model which is the proposed method. It shows that the proposed model has produce higher rate of prediction accuracy.

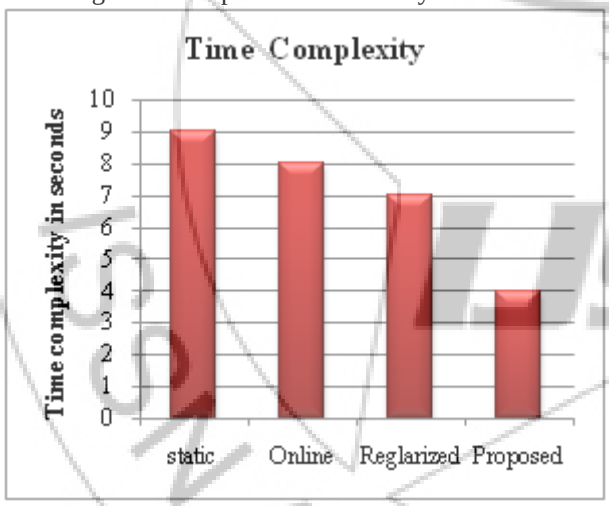

Figure 4: shows the time complexity of different approaches.

The figure 4, shows the time complexity produced by various approaches, while using Enron data set, and it shows that the proposed model has produced less time complexity than others.

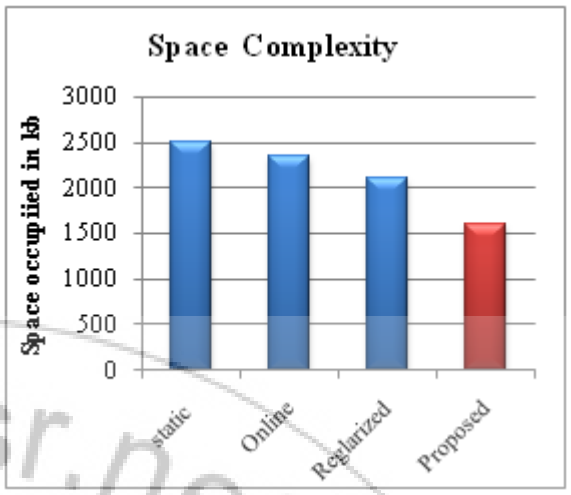

Figure 5: shows the space occupied by different algorithms on Enron data set.

The figure 5: shows the space occupied by different algorithms we compared to evaluate the proposed method. It shows that the proposed method has used only less memory where as other has taken more memory.

5. Conclusion

The proposed web inference model for user interest prediction has been focused on multimode networks. Initially we performed preprocessing, and then we have generated the clusters using which interest prediction is performed. The proposed method has produced efficient results and less time complexity value.

References

[1] K. Nowicki and T.A.B. Snijders, "Estimation and Prediction for Stochastic Blockstructures," J. Am. Statistical Assoc., vol. 96, no. 455, pp. 1077-1087, 2001.

[2] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal Nonnegative Matrix t-Factorizations for Clustering," Proc. 12th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '06), pp. 126-135, 2006.

[3] L. Tang, H. Liu, J. Zhang, and Z. Nazeri, "Community Evolution in Dynamic Multi-Mode Networks," Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '08), pp. 677-685, 2008.

[4] L. Tang and H. Liu, "Relational Learning via Latent Social Dimensions," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 817-826, 2009.

[5] L. Tang and H. Liu, "Toward Collective Behavior Prediction via Social Dimension Extraction," IEEE Intelligent Systems, vol. 25, no. 4, pp. 19-25, July-Aug. 2010.

[6] Y.-R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B.L. Tseng, "Facetnet: A Framework for Analyzing Communities and Their Evolutions in Dynamic Networks," Proc. 17th Int'l Conf. World Wide Web (WWW '08), pp. 685-694, 2008.

[7] T. Yang, Y. Chi, S. Zhu, Y. Gao, and R. Jin, "A Bayesian Approach Toward Finding Communities and

Their Evolutions in Dynamic Social Networks," Proc. SIAM Int'l Conf. Data Mining, 2009.
[8] C. Tantipathananandh and T. Berger-Wolf, "Constant-Factor Approximation Algorithms for Identifying Dynamic Communities," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), pp. 827-836, 2009.
[9] J. Sun, C. Faloutsos, S. Papadimitriou, and P.S. Yu, "Graphscope: Parameter-Free Mining of Large Time-Evolving Graphs," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 687-696, 2007.
[10] T. Xu, Z.M. Zhang, P.S. Yu, and B. Long, "Evolutionary Clustering by Hierarchical Dirichlet Process with Hidden Markov State," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM '08), pp. 658-667, 2008.
[11] A. Java, A. Joshi, and T. Finin, "Approximating the Community Structure of the Long Tail," Proc. Second Int'l Conf. Weblogs and Social Media (ICWSM '08), Mar. 2008.
[12] L. Tang and H. Liu, "Scalable Learning of Collective Behavior Based on Sparse Social Dimensions," Proc. 18th ACM Conf. Information and Knowledge Management (CIKM '09), pp. 1107-1116,2009.
[13] Daqing Zhang, Bin Guo, Zhiwen Yu, "Social and Community Intelligence," Computer, IEEE computer Society Digital Library. IEEE Computer Society,2011.
[14] Efthimios Bothos, Dimitris Apostolou, Gregoris Mentzas, "Using Social Media to Predict Future Events with Agent-based Markets," IEEE Intelligent Systems, 11 Oct. 2010. IEEE computer Society Digital Library. IEEE Computer Society,
[15] Maria Luisa Damiani, Claudio Silvestri, Elisa Bertino, "Fine-grained cloaking of sensitive positions in location sharing applications," IEEE Pervasive Computing, 15 Mar. 2011. IEEE computer Society Digital Library. IEEE Computer Society,
[16] Carmen Ruiz Vicente, Dario Freni, Claudio Bettini, Christian S. Jensen, "Location-Related Privacy in Geo-Social Networks," IEEE Internet Computing, 17 Feb. 2011. IEEE computer Society Digital Library. IEEE Computer Society
[17] Jaehong Park, Ravi Sandhu, Yuan Cheng, "User-Activity-Centric Framework for Access Control in Online Social Networks," IEEE Internet Computing, 28 Feb. 2011. IEEE computer Society Digital Library. IEEE Computer Society.