*International Journal of Engineering Science Invention*
*ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726*
*www.ijesi.org ||Volume 5 Issue 3|| March 2016 || PP.10-16*

# Multi-Mode Conceptual Clustering Algorithm Based Social Group Identification For Collaborative Web Search Using Web Log Data Sets

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:*** *The problem of web search time complexity and accuracy has been visited in many research papers, and the authors discussed many approaches to improve the search performance. Still the approaches does not produce any noticeable improvement and struggles with more time complexity as well. To overcome the issues identified, an efficient multi mode conceptual clustering algorithm has been discussed in this paper, which identifies the similar interested user groups by clustering their search context according to different conceptual queries. Identified user groups are shared with the related conceptual queries and their results to reduce the time complexity. The multi mode conceptual clustering, performs grouping of search queries and users according to number of users and their search pattern. The concept of search is identified by using Natural language processing methods and the web logs produced by the default web search engines. The author designed a dedicated web interface to collect the web log about the user search and the same data has been used to cluster the social groups according to number of conceptual queries. The search results has been shared between the users of identified social groups which reduces the search time complexity and improves the efficiency of web search in better manner.*

***Index Terms:*** *Social Networks, Multi Mode Networks, collaborative web search, Web Log data set, Conceptual clustering.*

## I. INTRODUCTION

The modern days more impact of internet where the people likes to explore many things through the web but the size of web resource is growing in day by day factor. For any particular detail, we can see enormous number of web pages available in the world. So that a web user cannot visit all the web pages related to any concept. For example if a web user likes to visit web pages related to "Data Mining" , then the number of pages gives details about the query has no limit and the user cannot view all the pages. In most cases, the user visits only few web pages and misses more informatics web pages in most times. So, in order to reduce the search time and to maintain the list of informatics web pages, the web log is necessary and using that the web user can be provided with more informatics web pages. The web log data set may maintain number of information according to the mechanism employed to retrieve the web page or to index the web page.

The social network is a growing multi mode network where the users are grouped in a name and shares many information between them. Even the user does not knew them personally but they become friends and shares many information. The users are grouped under a name and a single user may present in more than one group. In general case, they discuss about many things and chat about anything and the modern social networks keep track of what they are discussing. It is customary that the user in the social network search about many things and to find a more informatics web page every user spends more time in the web. This increases the search time and reduces the search efficiency.

Collaborative web search is one, where the search history of any user and the set of web pages visited about any similar query can be shared between them. For example, a user has performed search about a concept C, and visited N number of web pages. Among them, the user has found set of web pages are considered as more informatics and marked by the user by performing many actions or by spending more time in the web page. Such web pages are logged into the web log by the framework and used to produce more efficient result to another user, who comes with the same query.

Similar to that, the users of the social network chats about many topics and shares many information. From their chat we can identify, what topic they are interested about and can identify the interest of any user. Similarly, we can identify the interest of all the users of the social network. The problem here is, the social

network user may have more than one interest because, a single user is participated in different groups and has more than one interest. Because of a single user has links with more than one group, it can be considered as multi mode network. Each group in the social network has unique concept of interest and because of a single user has participated in multiple groups , every user has more than one concept of interest. By identifying the multiple modes of every user, they can be grouped or clustered to form a group using multi mode clustering algorithm.

Once the user could be clustered using multi mode network then their search history can be shared or used to produce more effective search result, which is called as collaborative web search and thus it reduces the time complexity of the web search.

## II. RELATED WORKS:

There are number of approaches has been discussed earlier to improve the performance of web search using web log data. Here we discuss set of few methods around the problem.

Predicting Friendship Links in Social Networks Using a Topic Modeling Approach [1], propose a topic modeling approach to the problem of predicting new friendships based on interests and existing friendships. Specifically, we use Latent Dirichlet Allocation (LDA) to model user interests and, thus, we create an implicit interest ontology. We construct features for the link prediction problem based on the resulting topic distributions. Experimental results on several LiveJournal data sets of varying sizes show the usefulness of the LDA features for predicting friendships.

Social Network Analysis with Content and Graphs [2], analyzes the Social network, which undergone a renaissance with the ubiquity and quantity of content from social media, web pages, and sensors. This content is a rich data source for constructing and analyzing social networks, but its enormity and unstructured nature also present multiple challenges. Work at Lincoln Laboratory is addressing the problems in constructing networks from unstructured data, analyzing the community structure of a network, and inferring information from networks. Graph analytics have proven to be valuable tools in solving these challenges. Through the use of these tools, Laboratory researchers have achieved promising results on real-world data.

Ontology-Based Link Prediction in the LiveJournal Social Network [3], is a social network journal service with focus on user interactions. As for many other online social networks, predicting potential friendships in the LiveJournal network is a problem of great practical interest. Previous work has shown that graph features extracted from the graph associated with the network are good predictors for friendship links. However, contrary to the intuition, user data (e.g., interests shared by two users) does not always improve the predictions obtained with graph features alone. This could be due to the fact that features constructed from a large number of user declared interests cannot capture the implicit semantic of the interests. To test this hypothesis, we use a clustering approach to build an interest ontology, and explore the ability of the ontology to improve the performance of learning algorithms at predicting friendship links, when interest-based features are used alone or in combination with graph-based features.

Social Network depicts the relationship like friendship, common interests etc. among various individuals. Social Network Analysis deals with analysis of these social relationships. Link prediction algorithms are used to predict these social relationships. Given a social network graph in which a node represents a user and an edge represents the relationship between the users, link prediction algorithm predicts the possible new relationships that can be created in the future. Comparison Analysis of Link Prediction Algorithms in Social Network [4], compares these link prediction algorithms on the basis of performance metrics like accuracy, precision, specificity and sensitivity.

Toward Predicting Collective Behavior via Social Dimension Extraction [5], examine how we can predict online behaviors of users in a network, given the behavior information of some actors in the network. Many social media tasks can be connected to the problem of collective behavior prediction. Since connections in a social network represent various kinds of relations, a social-learning framework based on social dimensions is introduced. This framework suggests extracting social dimensions that represent the latentaliations associated with actors, and then applying supervised learning to determine which dimensions are informative for behavior prediction.

Scalable Learning of Collective Behavior Based on Sparse Social Dimensions [6], propose an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other nonscalable methods.

Relational Learning via Latent Social Dimensions [8], 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 [10], 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 [11], 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.

All the above discussed approaches has the problem of identifying the social group interest in efficient manner and using the web log analysis efficiently to reduce the web search time.

## III. PROPOSED MODEL:

The multi mode conceptual clustering algorithm based collaborative web search model has the following functional components namely : Preprocessing, , Multi Mode Graph Generation, Multi mode concept clustering, Interest Prediction, Result Generation.

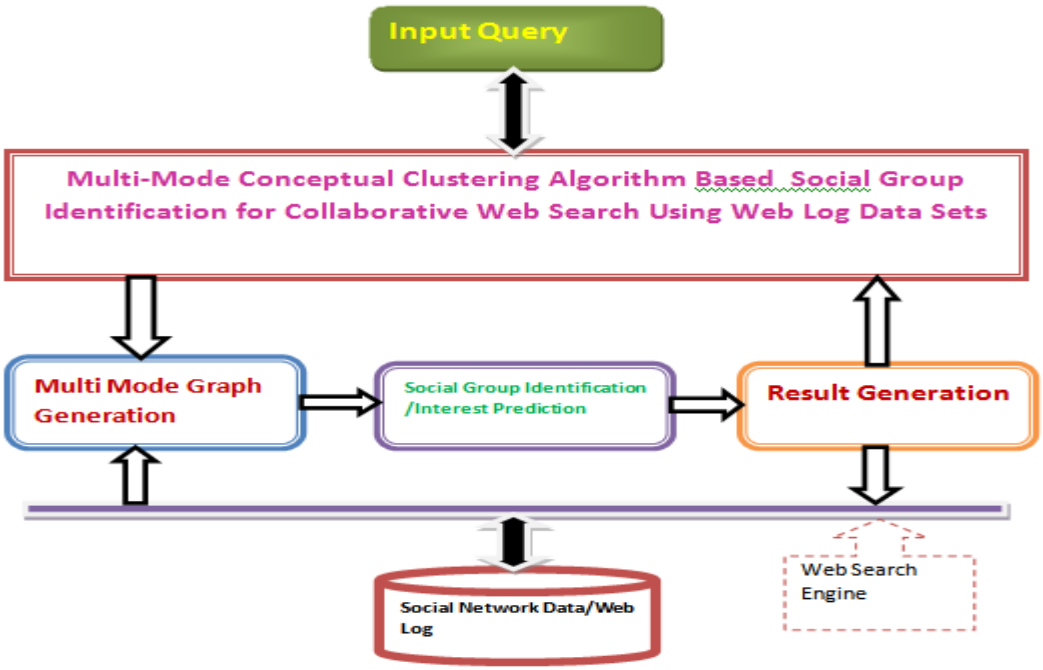



**Figure2: Architecture of Multi Mode Conceptual Clustering**

The Figure 2, shows the architecture of the proposed approach and its functional components. We discuss each of the functional module in detail in this section.

### 3.1 Multi Mode Graph Generation:

**T**his is the very first stage of the proposed approach, it read the input query from the user and read the web log and input search query. From the web log, the method identifies the set of unique users Ui, and for each user from the log the method identifies the set of all conversation made with others. For each user Uk, a graph is being generated, we identify the conversation made and if any conversation exists with particular user, then a link is generated. Simiarly N number of graphs are generated and the nodes are initialized with the node id and the conversation details. Generated graph will be used in the further stages to produce the result.

Procedure:
Input: Weblog Wl
Output: Graph Set Gs.
Start
Read Web log wl.
Idetify unique user Set Us.
Us = $\sum_{k=0}^{size\,(Wl)}$ Wl(Ui)₡Us
For each user Ui from Us
Initialize Graph Gi.
Collect set of all logs from Wl.
Logs Ul = $\sum_{k=0}^{size\,(Wl)}$ Wl(Ui) == Ui
For each log $l_i$ from Ul
Create Node Ni.
Initialize Ni with User id and Text conversation.
Add the node to the graph Gi.
Gi = ∑(Nodes €Gi)+Ni
End.
End
Stop

The above discussed algorithm generates multi mode social graph and it has been generated for each of the user present in the social network. The nodes are assigned with the concern user id and the conversation text.

**3.2 Social Group Identification:**

The method maintains set of taxonomy which has many categories in the taxonomy which is used to identify the social groups. From the input graph set, for each graph, the method extract the conversation text. Extracted conversation text is split into words and applied with stop word removal, stemming and tagging process. Finally a small set of terms will be selected as top words and verified with the taxonomy of concepts. For each of the concept the method computes the conceptual similarity measure. The method maintains various groups and for each user with the conversation logs, the conceptual similarity is computed. Finally with all the conceptual measures of each user, the method compute the cumulative weight for ech of the concept and choose a top valued concept as the interest of the user. This will be performed for each of the user present in the social group.

Procedure
Input: Taxonomy Tm, Graph Set Gs.
Output: Social group Sg, Cumulative Conceptual Similarity CCS.
Start
Initialize Social group Sg.
For each graph Gi from Gs
Collect all the nodes and conversation text.
Conversation Text Set CTS = $\sum_{k=0}^{size\,(Gi)}$ ∑ConversationText(Ni)
For each conversation Ci from CTS
Ts = Termset of (Ci)
Perform Stop word removal.
Perform stemming and tagging.
Ts =Identify pure terms.
For each concept Cp
Compute conceptual similarity measure CSM.
CSM = $\frac{\sum \text{Terms (Ts)} € \text{Termset (Taxonomy (Cp))}}{size\,(Termset\ (cp))} \times 100$
End.
End
Compute cumulative conceptual similarity CCS.
CCS = $\frac{\sum \text{CSM}}{Total\ number\ of\ conversation} \times 100$
Choose the maximum similarity concept.
Class Cl = Max(Css).
Add user Ui the semantic group Cl.
End
End
Stop

**a. Result Generation:**

At this stage, the method conceptual similarity measure using the input text query and a single concept is selected and the same input query will be passed to the web search engine and receive the results of the search engine. With the results of search engine,the method identifies the popular results viewed by the other users of the group and rank them based on click stream to generate the final result to the web user.

Procedure:
Input:Cumulative Conceptual Similarity set CCS, Input Query Q, Web log Wl.
Output: Final Recommendation.
Start
Compute Conceptual Similarity measure on query q.
Csm = CSM(q).
Identify the concept of interest.
Interest In = Max(Csm).
Identify the similar interested group from SG.
Identify the set of results viewed under the concept In.
Page viewed Pv = $\sum_{i=1}^{size\ (Wl)} Sg(Pi©In)$
Pv = Pv+ ResultFromSearchEngine(q).
Rank Results according to click stream.
Add to final recommendation and return.
End.

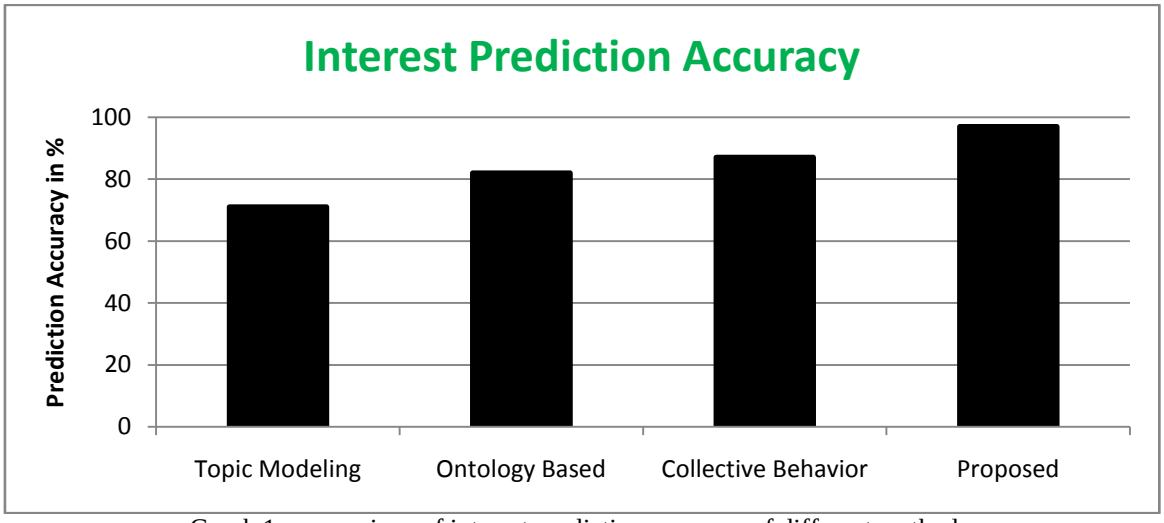
The above discussed algorithm shows how the result beign produced based on the conceptual clustering based approach.

## IV. RESULTS AND DISCUSSION:

The proposed approach has been implemented and tested for its efficiency using different data sets and a real time implementation which is developed in advanced java. The proposed approach has been evaluated for its efficiency using different data sets. The enron data set has been used to evaluate the performance of clustering and interest identification.

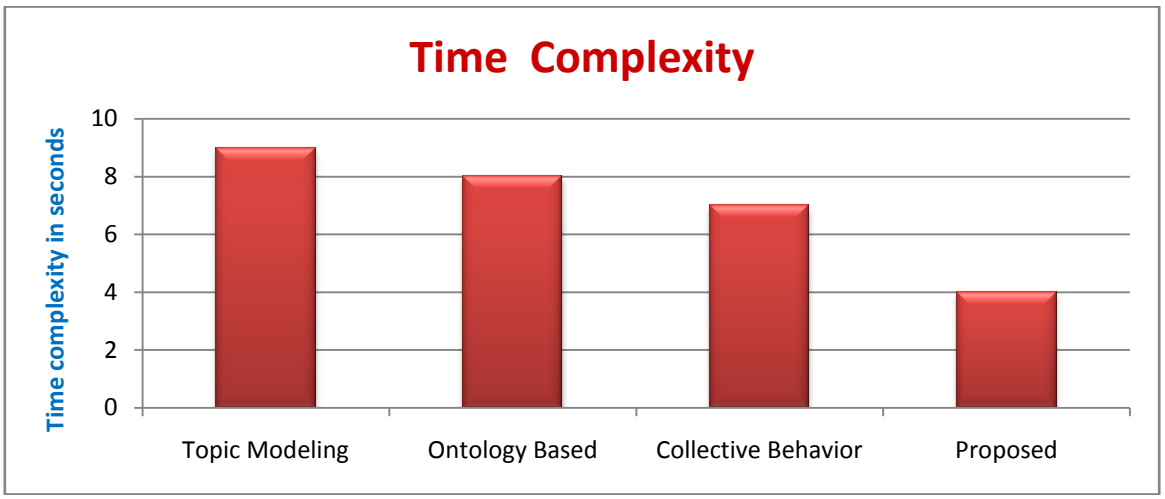
| Data Set | Number of Users | Number of Messages |
|---|---|---|
| Enron | 2359 | 32,789 |

Table1: shows the data set used.



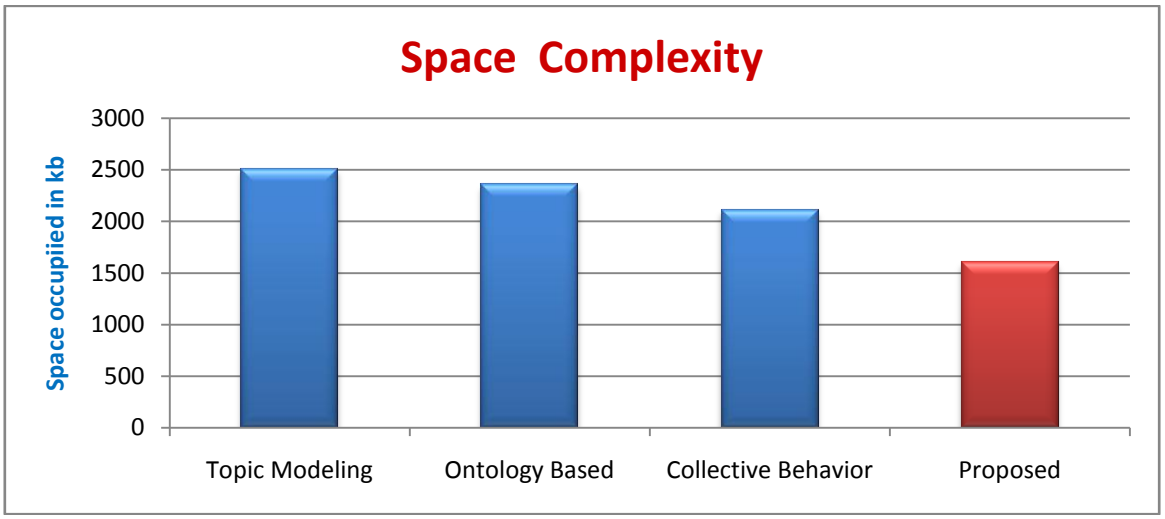

Graph 1: comparison of interest prediction accuracy of different methods.

The Graph 3 shows the comparison of interest prediction accuracy produced by different methods and it shows clearly that the proposed method has produced higher accuracy than other methods.




Graph 4: shows the time complexity of different approaches.

The graph 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.




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

The Graph 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.

## V. CONCLUSION:

We proposed a multi mode conceptual clustering algorithm to predict the user interest in a collaborative web search model to improve the websearch efficiency. The method identifies the social groups according to the conversation the user has with other users and using the conversation text the method computes the conceptual similarity measure for each the conversation. Using the computed measure, the approach computes the cumulative conceptual similarity measure and selects a concept as interest of the user and groups similar interested users. Similarly, the input query is processed and the concept of the query is computed and based on that the web links visited by same interested group user are retrieved. The retrieved links and with the search result from the web search engine is ranked based on click stream and returned to the user. The method produces efficient results and accurate clustering of social groups. The proposed method identifies the user interest in efficient manner and the results are shared between the members of social group and reduces the search time complexity.

## REFERENCES


[1] Rohit Parimi, Doina Caragea, Predicting Friendship Links in Social Networks Using a Topic Modeling Approach, Springer, Advances in Knowledge Discovery and Data Mining, Volume 6635, 2011, pp 75-86

[2] William M. Campbell, Charlie K. Dagli, and Clifford J. Weinstein ,Social Network Analysis with Content and Graphs, LINCOLN LABORATORY JOURNAL  VOLUME 20, NUMBER 1, 2013.

[3] Doina Caragea, Vikas Bahirwani, Waleed Aljandal and William H. Hsu, Ontology-Based Link Prediction in the LiveJournal Social Network, Advancement of Artificial Intelligence, 2009.

[4] Sahil Gupta, Shalini Pandey and K.k.shukla. Article: Comparison Analysis of Link Prediction Algorithms in Social Network.*International Journal of Computer Applications* 111(16):27-29, February 2015.

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

[6] 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.

[7] Daqing Zhang, Bin Guo, Zhiwen Yu, "Social and Community Intelligence," Computer, IEEE computer Society Digital Library. IEEE Computer Society,2011.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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.