Semantic Topics Modeling Approach for Community Detection

Hassan Abbas Abdelbary
Computer Science Department, Faculty of Computers and Information, Zagazig University
Elzeraa Square, Zagazig, Egypt

Abeer El-Korany
Computer Science Department, Faculty of Computers and Information, Cairo University
5 Dr.Ahmed Zewail Street, Orman, Giza, Egypt

ABSTRACT
Social networks play an increasingly important role in online world as it enables individuals to easily share opinions, experiences and expertise. The capability to extract latent communities based on user interest is becoming vital for a wide variety of applications. However, existing literature on community extraction has largely focused on methods based on the link structure of a given social network. Such link-based methods ignore the content of social interactions, which may be crucial for accurate and meaningful community extraction. In this paper, we present a novel approach for community extraction which naturally incorporates the content published within the social network with its semantic features. Two layer generative Restricted Boltzmann Machines model is applied for community discovery. The model assumes that users within a community communicate based on topics of mutual interest. The proposed model naturally allows users to belong to multiple communities. Through extensive experiments on the Twitter data for scientific papers, we demonstrate that the model is able to extract well-connected and topically meaningful communities.

General Terms
Social Networks, Community Detection, Restricted Boltzmann Machines (RBM).

Keywords
Social network, Community discovery, Machine learning, Restricted Boltzmann Machines, Topic modeling, Semantic similarity.

1. INTRODUCTION
In recent years, social networks have been spread widely specially with the appearance of social web sites like Facebook, Twitter and LinkedIn. Social networks create a pool of users with different interests, from different geographical regions, topics, opinions and feelings. Furthermore, social network demonstrates how the Internet continues to better connect people for various social and professional purposes. Contents broadcasted in those social networks almost cover several topics or interest in the world like marketing, politics, science, sports, movies and other. Both business companies and academics start to get attention to this rich environment which contains different kinds of people, interests and topics to be used for extracting useful knowledge to be helpful in making decisions. Knowledge extracted from social networks can be used to predict human real events like elections, marketing, movies box office, information spreading, stock index prices and other events. Recently, publisher uses social network analysis as quantitative or qualitative indicators of the impact that a single article has had [24] .Discovering communities is considered as one of the valuable production to be extracted from this rich pool of information. Community is a collection of users who share the same interest(s) and interact with each other most likely than other users in the network. Discovering these communities finds its importance in many applications like marketing, elections, stock index and computer science. Community discovery helps to connect people with common interests and encourages people to contribute and share more contents. Furthermore, it gives insights about the dynamics within each community and provides a good indicator about the status of the whole network and its health.

However, discovering common interests shared by users is a fundamental problem in social networks. Two main approaches are used to discover shared interests in social networks. One is user-centric, which focuses on detecting social interests based on the social connections among users; the other is item-centric, which detects common interests based on the common items such as hobbies, behavior, or topics of discussion. In the first approach the network is considered as a graph constructed of nodes and edges where the nodes represent the users and edges represent the relationships among those users. So discovering communities based on links analysis is considered as a graph clustering problem. In the other approach, discovering communities is based on analysis of contents published by users which represents their interests. Content broadcasted among users could be: posts, blogs, emails, tags, or tweets which contain topics that is used to identify communities share the same interest. Another approach integrates the advantage of the previous two approaches was proposed in [3] where people are clustered within a social network based on combined knowledge that decomposed explicit information defined in profiles of users representing their interests and knowledge extracted from dynamic interaction and social behavior overtime.

Different from other approaches, this research aims to find people who share the same interests no matter whether they are connected by a social graph or not. The proposed model assumes that two bloggers discuss the same topic(s) without necessarily being friends could be part of the same community and should have strong tie based on similarity between their published content. Therefore, the proposed model focus on connecting people within the social network based on their topic of interest. It focuses on directly detecting social interests or topics by taking advantage of user posts. The proposed model combines the advantage of generative machine learning with semantic correlation among users in

social networks. Generative model is used for topic modeling by analyzing user posts, extract topics, and semantically cluster users having the same topic (s). Within the proposed model, community discovery is applied using two-layer Restricted Boltzmann Machine (RBM) [6] to automatically identify the topics and their corresponding communities. Given a collection of user' posts (visible variables), the model learns a mixture of topic distribution over posts (latent variables) and hence discovers communities. In order to enhance community coherence, semantic similarity is applied to identify the correlation index between members within the same community using semantics features of discovered topics. *Semantic features* provide consistent correlation between topics within a community using the advantage of ontology.

The rest of this paper is structured as the follows: In section 2, some prior work on community discovery in social networks is reviewed. Section 3 presents the proposed model. Section 4 discusses the architecture of the community detection framework. In Section 5 detail about the datasets and the experimental results is illustrated. In Section 6 we conclude the work and upcoming future work is presented.

2. RELATED WORD

Community discovery has been extensively studied in various research areas such as social network analysis, web community analysis, etc. Several works done in the area of community discovery in social networks based on analysis of either links or content.

2.1 Link Analysis Community Discovery

In social networks, communities between individual are constructed by specifying and establishing friendship connections with each other. The link-based methods aim to find communities such that the friendship connections are dense within communities and sparse between them. Several methods use traditional graph partition methods to discover communities. In [18] conditional random fields used to construct a classification system for labeling users relationships in mobile social network and used these labeling to extract communities based on these labels by either weighting those labels or just treat them equally. Link based method used in [8] applied K-means clustering algorithm to find clusters (communities) based on the links among nodes in the network. Genetic algorithm approach is applied in [12] to find communities within social networks by relaying only in the graph structure of the network. Community discovering extended in special kinds of social networks called customer relationship networks in [9] by depending on the graph structure of the network by first identifying all maximal cliques then cliques are merged or extended into new structure by adding vertices to these cliques, these cliques identified by the density of connected component represents a community. In [21] communities were detected by labeling each node which represented a user with his interests based on his/her interaction with other users It select labels from other neighbor users with specified probability, after several iterations each user reach set of labels identifying her/his interest. In [22], dynamic of community change over time is considered. They applied a cut spectral clustering algorithm for detecting communities from social network that is represented by nodes and weighted edges. They applied the clustering algorithm at fixed time slices and view the communities change over time periods which give good intuition about the dynamically of social network. Recently, researchers focused on community mining for heterogeneous networks [2] using graph mining techniques which is already available. Such as MinCut algorithm, Regression based algorithm, Max-Min modularity measure, LM algorithm and SECI model. For example the work of [13] applied Min-cut and Regression for community mining and was applied on several real world datasets such as DBLP, Orkut, Facebook and Enron and was able to detect an acceptable number of hidden communities.

2.2 Content-based Community Discovery

Recently, content based approaches used to detect communities tend to be more realistic because social networks are not just a solid graph with nodes and edges only but also it contain contents published and spread periodically by users. The content basically represents the user interests and thoughts. Most of these approaches for discovering the communities are based on trying to find the joint probability distribution among different random variables, the visible and hidden ones. Visible variables or observable variables represent the content posted or broadcasted by the user and the hidden or latent variables represent the topics, communities or both. In [14,15], generative model was built to discover communities based on the discovered topics, interaction types and the social connections among people. Bayesian model is used for extracting the latent communities from the social network by assuming that community membership is dependent on the topics of interest among users. Another probabilistic model is used in [11] to discover communities from email exchange by considering both topics and link information. Another category of methods to discover communities based on the contents are those which apply g machine learning techniques in topic modeling. Restricted Boltzmann Machine (RBM) [4] was applied in [16,17] to extract the topics from the words then another model is trained to discover the link influence between web pages. The learning is measured by applying the resulted features discovered from the trained model into a supervised learning algorithm on labeled data to classify the links. The advantage of content-based approach is that it allow user to be member in more different community according to its different topics and interest.

3. SEMANTIC TOPIC-BASED CUMMUNITY DISCOVERY MODEL

The proposed model of community discovery integrated the concept of semantic using ontology with topic modeling. Topic modeling is used to detect major topics within a text, while ontology is used to identify the correlation strength between topics within a community. First, two-layer Restricted Boltzmann Machine (RBM) model is applied to extract topics from published contents from users in social network and discover topically related communities. Next, semantic similarity is applied to compute groups of users that are more densely connected to each other than to the rest of the network. The novelty of this model is that it incorporates machine learning technique for topic detection with semantic similarity measurement to tie related discovered topics and forming communities. This section describes the proposed framework which is shown in figure1.

3.1 Generative Topic Modeling

Probabilistic topic modeling has been applied to relate documents and words through variables which represent the main topics inferred from the text itself.

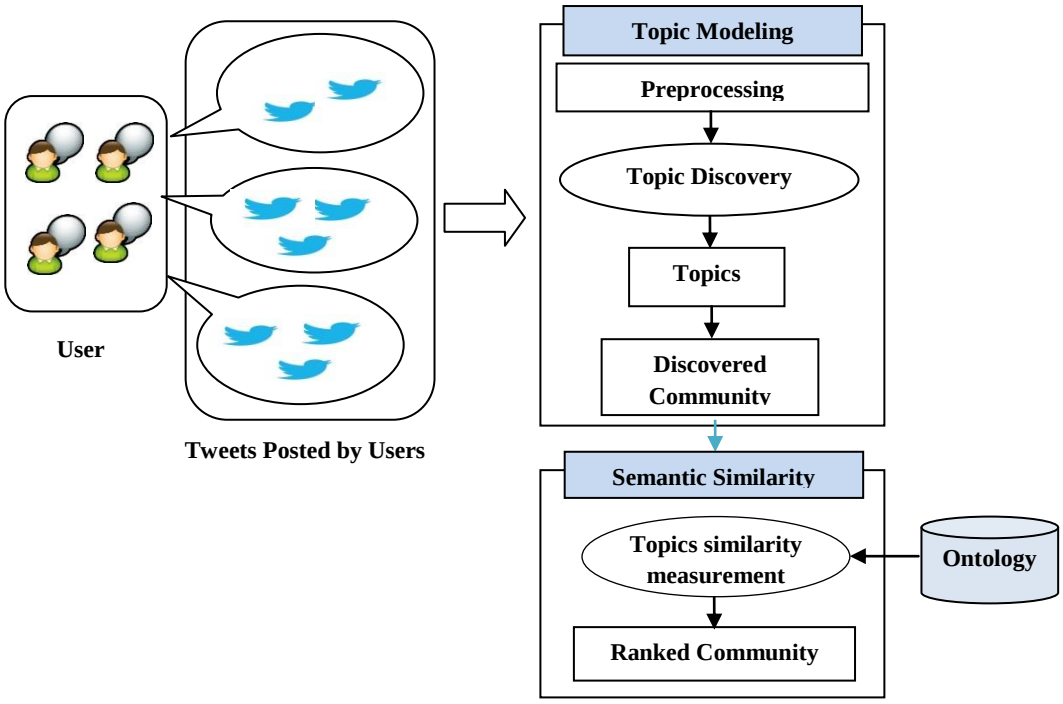

Fig 1 : The proposed framework for community discovery

In the context of this work, the concept of topic modeling is applied on posts published by users within a social network. Previous work which applied latent semantic analysis to link nodes to certain topics in the network construction [1]. Other work denoted in [23], applied Latent Dirichlet Allocation (LDA) based web crawling framework to discover different topics from Dark Web forum cites. Unlike previous work, multi-level Restricted Boltzmann Machine (RBM) is used to learn topics from text corpus collected from posts and then, cluster topics at the community level as shown in figure2. The proposed model assumes that communities are formed when users with similar interests aggregate together. Thus, it contains two levels for modeling variables: the first one is for modeling topic distribution over posts published by users, and the other is for modeling community of users over discovered topics. The first level represents a set of hidden variables to model a (topic) distribution over visible variables (words of posts) and the second level to model community distribution over topics. Accordingly, one community can correspond to multiple topics and multiple communities can share the same topic.

3.2 Semantic Community Detection

In order to determine the relevance between members inside discovered community, correlation between extracted topics within a community is measured using semantic similarity. Ontology-based semantic similarity is applied in order to measure the closeness between discovered topics which represent the member's interest inside each community. Measuring of semantic similarity between topics is naturally based upon a precise understanding of how the topics space is structured. WordNet term collection is used for the semantic grounding of the topics similarity measures.

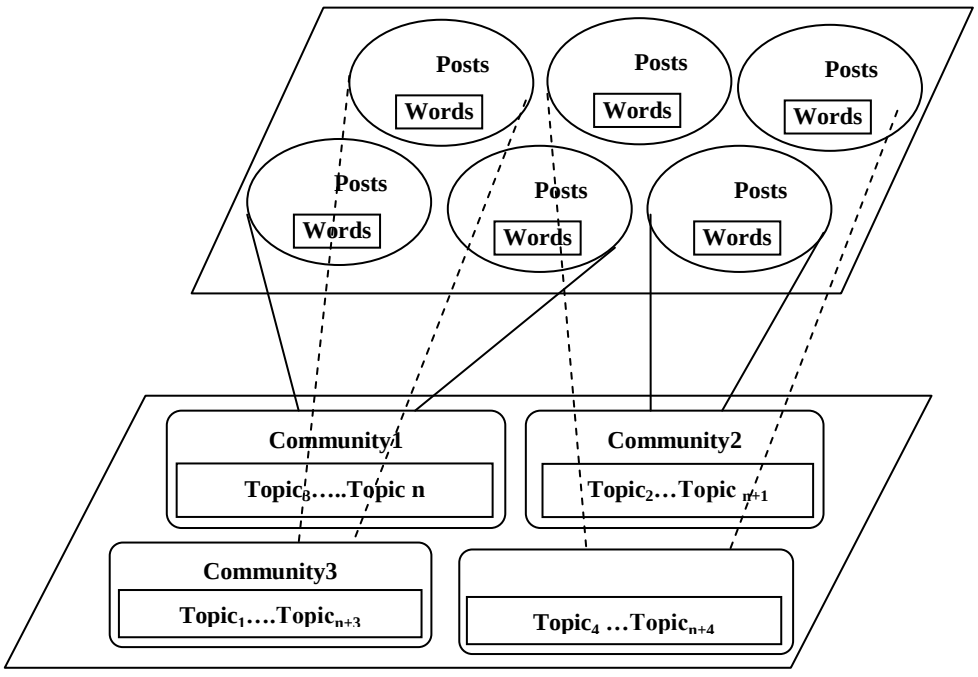

Fig 2: The mapping from words to communities which contains topics discovered from words

4. COMMUNITY DETECTION ARCHETICTURE

The community detection framework is decomposed of three phases: topic discovery, community detection, and semantic similarity measurement. Before topic discovery phase takes place, the system preprocesses the content that represents the posts of users as shown in figure3. Community discovery process is represented such that each topic has a multinomial distribution over words represent the users' posts and each user has a multinomial distribution over topics.

4.1 Preprocessing

During this step, all posts are parsed in order to clear all special characters, numbers, dates, stop words and single characters. This yields to construct a vocabulary that represents the set of words that have been used by all the users of the social network within a specific time period. The final step of the preprocessing is to convert the posts of each user from characters representation to vector of binary representation, where each word is represented by a binary variable indicating its presence or absence in the vocabulary we build. This is achieved by iterating over all posts and for each one if the current word appears in the vocabulary replace it by one and zero otherwise. The output of that phase is a set of binary vectors each represents a posts of each user.

4.2 Topic Discovery

The proposed model uses binary vectors for each user represent her/his published contents as an input to a stack of RBM in order to produce mixture of topic distribution over collected posts and next use these topics to relate users within a community. During this phase, the RBM model is trained over the binary data generated from the preprocessing step. RBM is an energy based models for unsupervised learning that has been successfully applied to problems involving high dimensional data such as images [5] and text [10,15,19]. It consists of two layers; one visible layer and one hidden layer. Visible layer which is clamped with the observed data (posts) and the hidden layer that is used for modeling the probability distribution of variables in visible units. The model is trained over collected terms until it reaches an acceptable error percentage and the top words for each hidden units are selected.

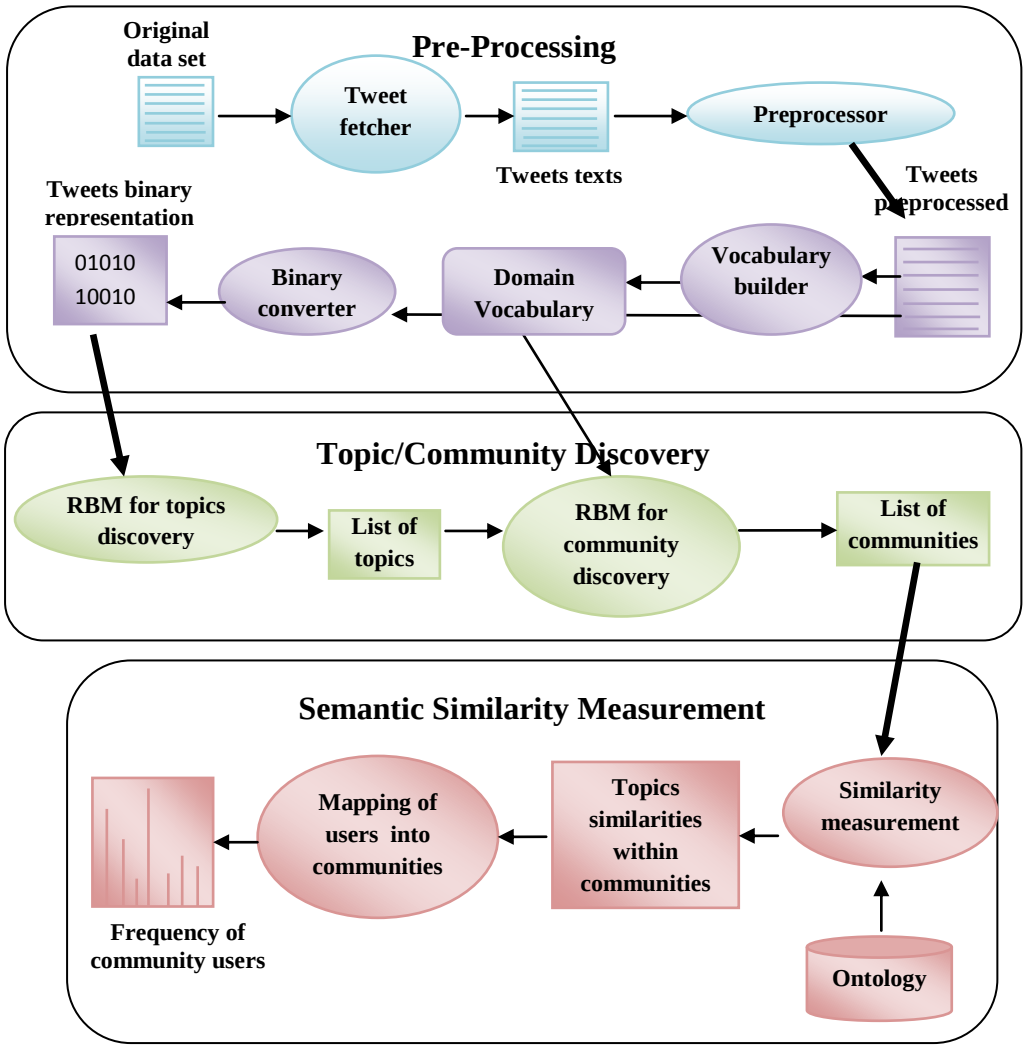

Fig 3: Overall process of community discovery

RBMs are usually trained using the contrastive divergence (CD) learning procedure [4]. This requires a certain amount of practical experience to decide how to set the values of numerical meta-parameters such as the learning rate, initial values of the weights, the number of hidden units and the size of each mini-batch. An energy function E(v,h) is used to promote competition between units which has the following form:

$$E(v,h) = -\sum_i v_j b_i + \sum_k h_k b_k + \sum_{i<j} v_i v_j w_{ij} + \sum_{i,k} v_j h_k w_{ik} + \sum_{k,l} h_k h_l w_{kl}$$

The weight update rule is defined as a function of gradient which is calculated by finding how the log probability of one training vector changed with respect to the change of weight. In order to find this gradient a derivative of log probability p(v) with respect to weight, the following learning rule is used for updating the weight:

$$\frac{\partial \log p(v)}{\partial w_{ij}} = < s_i s_j >_v - < s_i s_j >_{model}$$

Where:

$< s_i s_j >_v$: is the expected value of product of states when visible units is observed.

$< s_i s_j >_{model}$: is the expected value of all possible configurations for the whole model.

During each iteration, when we approximate the second term of the gradient by just one sample per one visible data vector this is called CD-1 and if we get more than one sample it become CD-n. Using more than one iteration to calculate the reconstruction depends on the computational. Another important measure of training process is the cross entropy reconstruction error which calculated by the following equation:

$$cere(x) = \sum_{i=1}^{n} (x * \log_2 r + (1 - x) * \log_2 (1 - r))$$

where x is the visible data and r is the reconstruction generated by the model.

Next, each hidden unit will present a set of ranked topics, top five of each unit is selected to be used by the next phase for communities discovery.

4.3 Community Detection

The previous phase produces a set of ranked topics which are then used to train the second level of RBM model in order to discover communities. This phase aims to estimate the probability distribution of communities over those topics. To figure most topic(s) discussed in each community, hidden units are ranked and only top 10 (weights) that connects these units and the visible units is used. After that, topics representing each community are gathered to form discovered communities.

4.4 Semantic Similarity Measurement of Community

Users in microbloging systems belong to multiple communities. Therefore, it is significant to measure the closeness between members within a community. Semantic similarity is applied in order to measure the strength of the tie between users according to the similarity between extracted topics shared by community members. Thus, we calculate similarity between two terms given the underlying domain ontology .Several methods have been applied in order to calculate semantic similarities between concepts of ontology such as methods based on semantic distance between concepts, methods based on information contents of terms, methods based on features of terms, and methods based on the hierarchical structure of ontology. Here, we consider one of the methods used to calculate the semantic similarity between two concepts which is calculated as a function of distance between concepts in a hierarchical structure of the underlying ontology. According to [7] which have obtained the best correlation of Semantic Distance between Terms compared with the average scores obtained by the humans. It considered Semantic Distance between Terms w1 and w2 as a function of shortest path between two words. The shortest path is calculated according to [20] by considering to position relation of Topic1 and Topic2 to their nearest common ancestor Topic. Here topic was the node with fewest is-a relationship as their ancestor node which appeared at the lowest position on the ontology hierarchy. This method consider semantic neighborhood of entity classes within their own ontologies. The following mathematical formula for calculating similarity between T1 and T2 is denoted as

$$Sim(T1, T2) = \frac{2H}{D1+D2+2H} \qquad \text{Equation1}$$

Where D1 and D2 were, respectively, the shortest paths from T1 and T2 to Topic, and H the shortest path from T to the root.

5. EXPIREMENTS

For validating the correctness of the proposed model, it was applied on Twitter blogs that discus scholar papers1. As the largest one of the microblogging service, Twitter's user base has grown, and it has attracted attention from corporations and others interested in customer behavior and service. Scientific community focuses on tracking, collecting, and measuring the spread of scholarly content using Twitter. Therefore, the dataset used in our experiment is a collection of tweets that mentioned a scientific articles that have been assigned at least one recognized scholarly identifier such as (DOIs), or (PMIDs) between 1st and 31st of July 2011. The dataset is formatted in records and each record contains user ID, tweet ID, and URL for each tweet.

5.1 Experiment Setup

A preprocessing step is done before using the dataset in the stack of framework. The tweets dataset is prepared to be in the appropriate format to be used. During this phase, the posts of each user are collected from corresponding URL and returned in html file that represent each tweet. The next step is parsing each html file to get the tweet text part. After extracting all the texts for each tweet, redundant, non-English text, all special characters, numbers, dates, stop words and single characters are removed. All the text converted to lower case and we consider only tweets that are of more than three words longer. A vocabulary that represents all words used in the tweets is constructed, and we also remove the words in the vocabulary which has frequency less than three. The final step in the preprocessing phase is to convert the tweets from characters representation into binary form as explained in section 4.1.We did several experiments on this dataset on different configurations. The total number of tweets is 54868 records. After the preprocessing steps the total number of posts reduced to 29217 cases. We run the experiments on a server machine with 16GB Ram, Intel Xeon CPU with two

altmetric.com

processors 2.40 Ghz for each. And 64-bit Windows Server 2008 R2 Operating system.

5.2 Illustrative Example for the Trained RBM

The first phase focused on training the first layer of RBM for topic discovery using the Contrastive divergence algorithm CD-1 with stochastic gradient descent on mini-batches. Meta parameters such as the number of hidden units, learning rate, and the size of each mini-batch is adjusted according to [24] to 50 hidden units and several learning rate between (0.01, 0.001, 0.0001) with no momentum or weights decay were used. Once a model is trained, we define a topic for a hidden unit by considering only the top 5 words with the highest connections to that unit. After 10-30 epochs, we compute the average free energy of a representative subset of the training data and compare it with the average free energy of validation set. We monitor the performance of learning by calculating the cross entropy reconstruction error between the clamped data and the reconstructed data from the model. After training using different learning rate, the validation error is calculated and the best one is only considered. The best model was with learning rate 0.001,50 hidden units and 30 iterations using CD-1 according to figure 4 and figure 5 that shows the histogram of the weights for the best model.

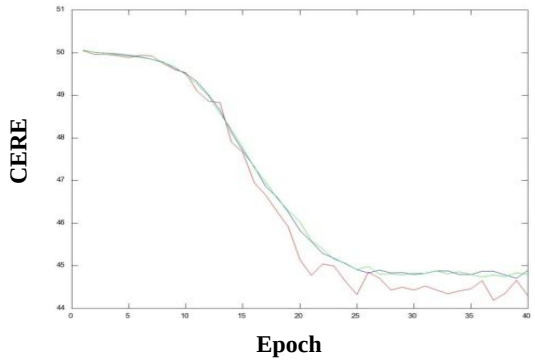

Fig 4: Learning curves (red curve for training error, blue curve for validation error and green curve for testing error

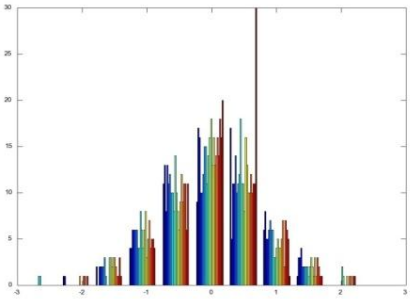

Fig 5: The final weights after finishing the training

5.2.1 RBM Topic Detection

The top 5 ranked weights for each hidden unit are captured and the items found in each one of are listed in table 1. Those topics are then used as the training set for the next RBM for detecting communities.

Table 1. List of topics associated with hidden units, ND (not defined) topics for hidden unit

HU Index	Topic	HU Index	Topic
1	Addiction	26	ND
2	Medical	27	Biophysics
3	ND	28	ND
4	Pathology	29	Biology
5	Biology	30	Psychology
6	ND	31	Economy
7	Electrical Engineering	32	ND
8	Investigation	33	Politics
9	Biology	34	Nervous System
10	Science	35	Safety
11	Medical	36	ND
12	ND	37	Mapping
13	Pharmacology	38	Medical
14	Pharmacology	39	Chemistry
15	Management	40	Medical
16	Military	41	ND
17	Genomics	42	Social Networks
18	Biology	43	Medical
19	Genomics	44	Geography
20	Physics	45	Astronomy
21	Experiment	46	Bioinformatics
22	Diseases	47	Biology
23	Geography	48	Technology
24	ND	49	Neuroscience
25	Pathology	50	Chemistry

5.2.2 RBM Community Discovery

The second phase focuses on training the RBM using 50 visible layers equal to number of topics discovered from first layer according to table1. Each tweet is converted to binary form by mapping it to topics appears in table1. We start by assuming that the hidden units are equal to 10 units and monitor the results. Next, we regularly increase the number of hidden units until the error rate improves, which actually happen with 20 hidden units. We use mini batches approach in learning the model by updating the weights at each mini batch. CD-10 is used for training with the new reduced dimension. CD-10 is more desirable when RBM is used to model the joint probability. We run the training between 10-50 epochs. Figure 6 shows the training error, validation error, testing error and figure 7 shows histogram of weights. The top 10 ranked topics from each hidden units are listed in table 2.

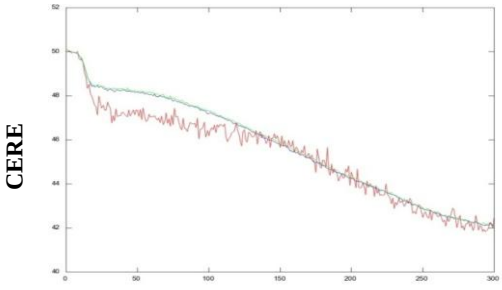

Fig 6: Learning curves (red curve for training error, blue curve for validation error and green curve for testing error

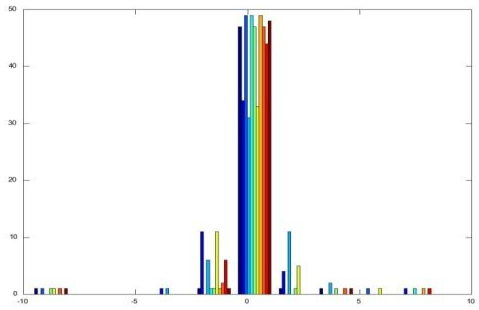

Fig 7: The final weights after finishing the training

5.2.3 Semantic-Based Communities Detection

In order to measure the correlation coefficient between members of discovered community, we use semantic of ontology to indicate the relationship among discovered topics representing each community. Thus, we start by applying HIT algorithm to identify the key topic in each community by creating adjacency matrix and calculate the authority value. Then, similarity between each topic and key topic is measured using equation1. According to table3 which summarize each community, key topics, and similarity among topics, the same community may have diverse topics which are true especially in scientific conference where a paper may occupy a multi-disciplinary work.

6. CONCLUSION AND FUTURE WORK

This paper presents a novel approach that integrated semantic feature of ontology with unsupervised learning in order to discover communities of micoblog users. The proposed approach applies two-level restricted Boltzmann machine (RBM) model to identify topics from contents published by users which represent user interest, then use the discovered topics to discover communities of users. Semantic relation between topics of each community is used to measure the closeness between members within discovered communities. As future work, we plan to include several semantic relations that would enhance community discovery process such as link influence and trust relationship. Furthermore, other knowledge could be added about user interest and could be extracted from other social network such as facebook, or FOAF.

Table 2. List of communities and topics associated with them and ND means not defined topic

Community Number	$Topic_1$	$Topic_2$	$Topic_3$	$Topic_4$	$Topic_5$	$Topic_6$	$Topic_7$	$Topic_8$	$Topic_9$	$Topic_{10}$
1	Mapping	Safety	Investigation	Electrical Engineering	Chemistry	Social Networks	Medical	ND	Geography	Psychology
2	Genomics	Astronomy	ND	Chemistry	Mapping	ND	ND	Geography	Biology	ND
3	Chemistry	Biophysics	Biology	ND	ND	Social Networks	Biology	Pharmacology	Electrical Engineering	Bioinformatics
4	Medical	ND	Medical	Biology	ND	Geography	Biology	Genomics	Pharmacology	Pathology
5	ND	Technology	Experiment	Nervous System	ND	Addiction	Psychology	Economy	Medical	Diseases
6	Investigation	ND	Physics	ND	Medical	ND	Pharmacology	Medical	Military	ND
7	ND	Experiment	ND	Mapping	Management	Addiction	Medical	Biology	ND	Medical
8	Medical	Biophysics	ND	Nervous System	Physics	ND	Genomics	Pathology	Electrical Engineering	Pathology
9	ND	Science	Technology	ND	ND	Genomics	Nervous System,	Politics	Biology	Chemistry
10	Medical	Neuroscience	ND	Nervous System	ND	Medical	Electrical Engineering	Biology	Astronomy	Management
11	Geography	Economy	Biology	ND	Social Networks	Management	Investigation	Medical	Pathology	ND
12	Economy	Biophysics	Pathology	ND	Science	Neuroscience	Biology	Safety	ND	Management
13	Biology	Diseases	Biology	Biophysics	Geography	Medical	Pharmacology	Addiction	ND	Astronomy
14	ND	Investigation	ND	Biophysics	Medical	Diseases	Geography	Biology	Pathology	Science
15	Psychology	Astronomy	Medical	Genomics	Pharmacology	Safety	Biology	ND	Biology	Experiment
16	Genomics	ND	Physics	Biology	ND	Politics	ND	Genomics	Investigation	Diseases
17	Astronomy	Geography	Addiction	ND	Pathology	Geography	Diseases	Military	Pharmacology	ND
18	Astronomy	Science	Nervous System	Chemistry	Mapping	ND	ND	Biology	ND	Pharmacology
19	Medical	Physics	Economy	Genomics	Military	ND	Technology	Social Networks	Pathology	Biology
20	Experiment	Physics	Science	Medical	Medical	Geography	Politics	Bioinformatics	Military	Medical

Table 3. List of communities and weights for each associated topic according to key topics listed in last column

Community	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Key Topic
1	Mapping	Safety	Investigation	Electrical Engineering	Chemistry	Social Networks	Medical	ND	Geography	Psychology	Electrical Engineering
HITS(authority)	0.28245	0.88015	0.22200	1	0.99514	0.52085	0.94514	0	0.52085	0.22200	
Similarities	0.25	0.25	0.25	0	0.333	0.2	0.333	0	0.333	0.333	
2	Genomics	Astronomy	ND	Chemistry	Mapping	ND	ND	Geography	Biology	ND	Chemistry
HITS(authority)	0.85463	0	0	1	0	0	0	0.46081	0.85463	0	
Similarities	0.2	0.25	0	0	0.2	0	0	0.25	0.25	0	
3	Chemistry	Biophysics	Biology	ND	ND	Social Networks	Biology	Pharmacology	Electrical Engineering	Bioinformatics	Chemistry
HITS(authority)	1	0.83836	0.95447	0	0	0	0.95447	0.64174	0.39205	0.83836	
Similarities	0	0.2	0.25	0	0	0.2	0.25	0.2	0.25	0.2	
4	Medical	ND	Medical	Biology	ND	Geography	Biology	Genomics	Pharmacology	Pathology	Pharmacology
HITS(authority)	0.75000	0	0.75000	0.75000	0	0	0.75000	0.50000	1	0.50000	
Similarities	0.333	0	0.333	0.25	0	0.333	0.25	0.2	0	0.25	
5	ND	Technology	Experiment	Nervous System	ND	Addiction	Psychology	Economy	Medical	Diseases	Diseases
HITS(authority)	0	0	0.79622	0.79622	0	0.58287	0.26794	0	0.73205	1	
Similarities	0	0.25	0.25	0.25	0	0.333	0.25	0.333	0.333	0	
6	Investigation	ND	Physics	ND	Medical	ND	Pharmacology	Medical	Military	ND	Medical
HITS(authority)	0	0	0.99622	0	1	0	0.99622	1	0	0	
Similarities	0.333	0	0.5	0	0.5	0	0.333	0	0.5	0	
7	ND	Experiment	ND	Mapping	Management	Addiction	Medical	Biology	ND	Medical	Medical
HITS(authority)	0	0.33467	0	0.75723	0	0	1	0.88395	0	1	
Similarities	0	0.333	0	0.333	0.5	0.333	0	0.333	0	0	
8	Medical	Biophysics	ND	Nervous System	Physics	ND	Genomics	Pathology	Electrical Engineering	Pathology	Pathology
HITS(authority)	0.72701	0.50718	0	0.99377	0.93377	0	0.27512	0.78670	0.72701	1	
Similarities	0.333	0.25	0	0.25	0.333	0	0.25	0	0.25	0	
9	ND	Science	Technology	ND	ND	Genomics	Nervous System	Politics	Biology	Chemistry	Science
HITS(authority)	0	1	0.44504	0	0	0.80193	0	0.44504	0.80193	0.90193	
Similarities	0	0	0.5	0	0	0.25	0.5	0.5	0.333	0.333	
10	Medical	Neuroscience	ND	Nervous System	ND	Medical	Electrical Engineering	Biology	Astronomy	Management	Nervous System
HITS(authority)	1	0.56155	0	1	0	0.96159	0.56155	0.99155	0	0	
Similarities	0.333	0.2	0	0	0	0.333	0.25	0.333	0.333	0.333	
11	Geography	Economy	Biology	ND	Social Networks	Management	Investigation	Medical	Pathology	ND	Geography
HITS(authority)	1	0.71308	0.25430	0	0.87129	0.38727	0.48401	0.25430	0	0	
Similarities	0	0.5	0.5	0	0.25	0.5	0.333	0.5	0.333	0	
12	Economy	Biophysics	Pathology	ND	Science	Neuroscience	Biology	Safety	ND	Management	Biology
HITS(authority)	0	0.81411	0.33721	0	0.81411	0	1	0.60009	0	0.15133	
Similarities	0.333	0.2	0.333	0	0.333	0.25	0	0.25	0	0.333	
13	Biology	Diseases	Biology	Biophysics	Geography	Medical	Pharmacology	Addiction	ND	Astronomy	Diseases
HITS(authority)	0.84238	1	0.84238	0.61315	0	0.85142	0.91917	0.27399	0	0	
Similarities	0.333	0	0.333	0.25	0.333	0.333	0.25	0.2	0	0.333	
14	ND	Investigation	ND	Biophysics	Medical	Diseases	Geography	Biology	Pathology	Science	Biology
HITS(authority)	0	0	0	0.61803	0.91803	0.61803	0	1	0	0	
Similarities	0	0.333	0	0.2	0.333	0.25	0.333	0.25	0.333	0.333	
15	Psychology	Astronomy	Medical	Genomics	Pharmacology	Safety	Biology	ND	Biology	Experiment	Medical
HITS(authority)	0.20500	0.48060	1	0.36777	0.50374	0.20500	0.56893	0	0.44392	0.59995	
Similarities	0.5	0.5	0	0.333	0.333	0.25	0.5	0	0.5	0.333	
16	Genomics	ND	Physics	Biology	ND	Politics	ND	Genomics	Investigation	Diseases	Genomics
HITS(authority)	1	0	0	0.98077	0	0	0	0.78077	0	0.78077	
Similarities	0	0	0.333	0.333	0	0.333	0	0	0.25	0.333	
17	Astronomy	Geography	Addiction	ND	Pathology	Geography	Diseases	Military	Pharmacology	ND	Addiction
HITS(authority)	0	4.94065	1	0	0.46081	4.94065	0.85463	0	0.85463	0	
Similarities	0.5	0.5	0	0	0.5	0.5	0.333	0.5	0.25	0	
18	Astronomy	Science	Nervous System	Chemistry	Mapping	ND	ND	Biology	ND	Pharmacology	Chemistry
HITS(authority)	0	0.74277	0.47697	1	0	0	0	0.68263	0	0.56304	
Similarities	0.333	0.333	0.2	0	0.333	0	0	0.25	0	0.333	
19	Medical	Physics	Economy	Genomics	Military	ND	Technology	Social Networks	Pathology	Biology	Biology
HITS(authority)	0.95921	0	0.28549	0.94160	0.28549	0	0.14013	0.15921	0.46348	1	
Similarities	0.5	0.5	0.5	0.333	0.5	0	0.333	0.25	0.5	0	
20	Experiment	Physics	Science	Medical	Medical	Geography	Politics	Bioinformatics	Military	Medical	Medical
HITS(authority)	0.94920	0.74131	0.74131	1	0.79210	0	0	0.42868	0	1	
Similarities	0.333	0.333	0.5	0	0	0.5	0.333	0.25	0.5	0	

7. REFERENCES

[1] Bradford, R. (2006). Application of latent semantic \indexing in generating graphs of terrorist networks. Intelligence and Security Informatics, 674-675.

[2] Devi R, R. (2013). A Perspective Analysis of Hidden Community Mining Methods in Large Scale Social Networks. International Journal of Computer Applications, 75(3), 7-12.

[3] El-Korany, A. (2012). Society in Hand: Toward Community Service through Social Network. International Journal of Computer Applications, 51(8), 15-22.

[4] Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7), 1527-1554.

[5] Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. Paper presented at the Proceedings of the 24th international conference on Machine learning.

[6] Larochelle, H., Mandel, M., Pascanu, R., & Bengio, Y. (2012). Learning Algorithms for the Classification Restricted Boltzmann Machine. The Journal of Machine Learning Research, 13, 643-669.

[7] Li, Y., Bandar, Z. A., & McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. Knowledge and Data Engineering, IEEE Transactions on, 15(4), 871-882.

[8] Lin, C.-Y., Koh, J.-L., & Chen, A. (2010). A better strategy of discovering link-pattern based communities by classical clustering methods. Advances in Knowledge Discovery and Data Mining, 56-67.

[9] Liu, J., Liu, F., Zhou, J., & He, C. (2009). Irregular Community Discovery for Social CRM in Cloud Computing. Cloud Computing, 497-509.

[10] Mnih, A., & Hinton, G. (2007). Three new graphical models for statistical language modelling. Paper presented at the Proceedings of the 24th international conference on Machine learning.

[11] Pathak, N., DeLong, C., Banerjee, A., & Erickson, K. (2008). Social topic models for community extraction. Paper presented at the The 2nd SNA-KDD Workshop.

[12] Pizzuti, C. (2008). GA-Net: A genetic algorithm for community detection in social networks. Parallel Problem Solving from Nature–PPSN X, 1081-1090.

[13]Prakash, D., &Surendran, S. (2013). Detection and Analysis of Hidden Activities in Social Networks. International Journal of Computer Applications,77(16), 34-38.

[14] Sachan, M., Contractor, D., Faruquie, T. A., & Subramaniam, L. V. (2012). Using content and interactions for discovering communities in social networks. Paper presented at the Proceedings of the 21st international conference on World Wide Web.

[15] Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. Paper presented at the ACM international conference proceeding series.

[16] Tang, J., & Zhang, J. (2009). A discriminative approach to Topic-Based citation recommendation. Advances in Knowledge Discovery and Data Mining, 572-579.

[17] Tang, J., Zhang, J., Yu, J. X., Yang, Z., Cai, K., Ma, R., . . . Su, Z. (2009). Topic distributions over links on web. Paper presented at the Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on.

[18] Wan, H.-Y., Lin, Y.-F., Wu, Z.-H., & Huang, H.-K. (2012). Discovering Typed Communities in Mobile Social Networks. Journal of Computer Science and Technology, 27(3), 480-491.

[19] Welling, M., Rosen-Zvi, M., & Hinton, G. (2005). Exponential family harmoniums with an application to information retrieval. Advances in neural information processing systems, 17, 1481-1488.

[20] Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. Paper presented at the Proceedings of the 32nd annual meeting on Association for Computational Linguistics.

[21] Xie, J., & Szymanski, B. (2012). Towards linear time overlapping community detection in social networks. Advances in Knowledge Discovery and Data Mining, 25-36.

[22] Xu, K., Kliger, M., & Hero, A. (2011). Tracking communities in dynamic social networks. Social Computing, Behavioral-Cultural Modeling and Prediction, 219-226.

[23] Yang, L., Liu, F., Kizza, J. M., & Ege, R. K. (2009). Discovering topics from dark websites. Paper presented at the Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium on.

[24] Yu, S., & Kak, S. (2012). A Survey of Prediction Using Social Media. arXiv preprint arXiv:1203.1647.