International Journal of Advanced Trends in Computer Applications
www.ijatca.com

A FRAMEWORK TO FILTER UNWANTED MESSAGES FROM OSNs USER WALL

Kanika Sharma
Department of Computer Science and Engineering, PEC University of Technology, Chandigarh, India
kanikapisces@gmail.com
Prof. Manavjeet Kaur
Department of Computer Science and Engineering, PEC University of Technology, Chandigarh, India
manvjeet_1980@yahoo.com

Abstract: *Social networking sites plays a significant role in today's society, it is now one of the daily activities in everyone's regular life. With the help of smart phones, its use has increased drastically. At present, online Social Networks does not provide its users the capability to control the messages posted on their own confidential space/private wall, to avoid the unwanted content being displayed. To fill this gap, in the present paper, we suggest a system allowing OSN users to have a direct control on the messages posted onto their wall. This is achieved through a flexible rule-based system, that allows users to specify the filtering criteria to be applied to their walls, and with the help of Machine Learning based soft classifier the short text messages are classified into different categories and can be filtered as desired by the users.*

Keywords: *Online social networks, Content based filtering and Filtering System.*

I. INTRODUCTION

Online Social Networks (OSNs) plays a significant role in today's society; it is now one of the daily activities in everyone's regular life. With the help of smart phones, its use has increased drastically. OSNs are today's one of the most popular medium amongst the people of all age groups to share and stay connected with the social world. Daily and continuous communications imply the exchange of several types of content, including text, image, audio, and video data. Facebook and Twitter are replacing email and search engines as users' primary choices to the Internet. Communication on these sites involves exchange of various types of content including text as well as multimedia data. A social networking site generally include blogs, private messaging, chat facility and file, photo sharing functions and other ways to share text and multimedia data. Users of the online networking sites can share their feelings and ideas in terms of wall messages too. In OSN, a wall is a section of the user profile where others can post messages or send images to its wall owner. This wall is a public space so others can view what has been written on the wall. Therefore, in OSNs, there is possibility of posting bad or undesirable messages on wall which is visible to others too. To provide solution to this problem, wall messages should be classified and the unwanted messages should be filtered out as required by the wall owner.

As in today's OSN, there is a very high chance of posting unwanted content on public/private areas, generally called as walls. Existing OSNs provides very less support to prevent unwanted messages on user walls. For example, Facebook allows users to manage access for who is allowed to post messages onto their walls (i.e., friends, friends of friends, or defined groups of friends). However, no content-based preferences are supported and therefore it is not possible to prevent undesired messages, such as vulgar, offensive or political ones, no matter of the user who posts them. This is because wall messages are constituted by short text for which traditional classification methods have serious limitations since short texts do not provide sufficient word occurrences.

The aim is therefore to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls with the help of Machine Learning (ML) text

categorization techniques to automatically assign with each short text message a set of categories based on its content. A hierarchical two level classification strategy is used with Neural Networks, Support Vector Machine and Latent Dirichlet Allocation. In the first level, the ICA categorizes short messages as Neutral and Non-Neutral. Then, in the second stage, Non-Neutral messages are classified with the help of the machine learning techniques to identify the category to which the message belongs.

Framework also has the facility to provide direct control to the users for managing the content being posted onto their walls with the help of filtering rules (FR). Filtering rules allow users to state constraints on message creators like by imposing conditions on their profile's attributes or exploiting information on their social graph. There is review done for malicious behaviors of OSN users, and discussed several solutions to detect misbehaving users. Thus additional feature can be provided as Black List (BL) where based on the user's specification the system will be able to determine the users to be inserted in the BL list. Based on the relative frequency that let the system be able to detect those users whose messages continue to fail the filtering rules will be blacklisted. Additional features to enhance the learning of the classification system like Key term identification, Querying Microsoft Word Thesaurus/Word Net or using Google Sets. It can be used to give users the ability to automatically control the messages written onto their own walls, by filtering out unwanted messages.

COMMON FILTERING TECHNIQUES

The common content filtering techniques are:

- Content based filtering
- Collaborative filtering
- Policy based filtering

A. Content-based filtering

Content Filtering (also known as information filtering) is blocking undesirable or unwanted content over the network i.e. a Content Filter helps to decide which content is acceptable for viewing and access through a given system.

In content-based filtering each user is assumed to operate independently. As a result, a content-based filtering system selects information items based on the correlation between the content of the items. For example OSNs such as Facebook uses content based filtering policy. In that by checking users profile attributes like education, work area, hobbies etc. suggested friend request may send.

The activity of filtering can be modeled, in fact, as a case of single label, binary classification, partitioning incoming documents into relevant and non relevant categories. More complex filtering systems include multi-label text categorization automatically labeling messages into partial thematic categories.

B. Collaborative filtering

In collaborative filtering, information is selected on the basis of user's preferences, actions, predicts, likes, and dislikes. Match all this information with other users to find out similar items. Large dataset is required for collaborative filtering system. According to user's likes and dislikes items are rated.

C. Policy-based filtering

In policy based filtering system, users filtering ability is represented to filter wall messages according to filtering criteria of the user. For example, associating a set of categories with each tweet describing its content on Twitter. The user can then view only certain type of tweets based on his/her interests. In policy-based filtering, the communication policy can be defined between two communicating parties.

II. FILTERED WALL ARCHITECTURE

Three Tier architecture is used in OSN services. The three layers are:

- Graphical User Interface (GUI)
- Social Network Application (SNA)
- Social Network Manager (SNM)

Graphical User Interface (GUI)

The system graphical user interface composed of interface to insert user credentials to login into system as well as new user registration. The Filtered wall interface consist of components to post a message on user wall which on submission go through social network application layer and social network manager layer before being published on user wall.
The second layer comprises Content Based Message Filtering (CMBF) and Short Text Classifier. This is very important layer for the message categorization according

ISSN: 2395-3519

to its CBMF filters. Also Blacklist is maintained for the user who sends frequently bad words in message.

Social Network Manager (SNM)

The Social Network Manager layer provides the essential OSN functionalities (i.e., profile and relationship administration). It also maintains all the data regarding to the user profile. The social network manager layer extract data from user social profile and provide it to the social network application layer to impose filtering rules.

III. SHORT TEXT CLASSIFICATION

Traditional techniques like Bag-Of-Words work well with the documents which are typically large and are rich with content as the word occurrence is high and though the order is lost, word frequency is enough to capture the semantics of the document. Alternate approaches like TF-IDF help to counter some loop holes in the Bag-Of-Words approach by weighing the terms.
Short text is characterized by shortness in the text length, and sparseness in the terms presented, which results in difficulty in managing and analyzing them based on the bag- of-words representation. It has a wide range of extension, such as mobile short messages, instant messages, news titles, online chat record, blog comments, news comments, etc. And its main characteristic is that the text length is very short, no longer than 200 characters. As mobile messages which we commonly used are no more than 70 characters, news titles are less than 30. Instant messaging (IM) software also limits its length, such as Windows Live Messenger of Microsoft allows the longest message 400 characters.
However, when dealing with shorter text messages, traditional techniques will not perform that well as they would have performed on larger texts. Since these techniques rely on word frequency and short texts do not provide sufficient word occurrences, also they offer no sufficient knowledge about the text itself.

There are other approaches like integrating short text messages with Web search engines like Google, Bing to extract more information about the short text. With the help of statistics on the engine results for each pair of short text, similarity score is determined. However, these techniques require additional entity disambiguation approaches. For example, “bat” and “bird” are highly related. But, when thesaurus search or web search is performed, more hits may be related to the game “cricket” than the bird “bat”. Hence, there is a need to get explicit feedback from the user to direct the searching and text inflation process. It is not feasible to perform semantic similarity search on every pair of short text messages as it is time consuming and not suitable for real-time applications.

IV. NEURAL NETWORK

Neural frameworks are made out of simple elements which work in parallel. A neural framework can be arranged to perform a particular function by changing the estimations of the weights between elements. Network function is determined by the connections between elements. There is activation functions used to produce relevant output.

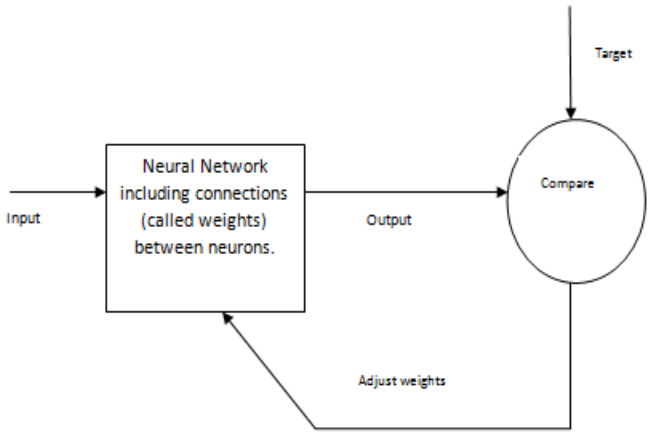

Figure 1: Neural network

Training can be either supervised or unsupervised. In supervised training, network adjusts by endeavouring to predict results for known delineations. System learns by comparing the differences in the results and its expectations for the known inputs and adjusts its weights accordingly. In unsupervised training, no yield or result is exhibited as a part of training. With the delta rule, as with diverse sorts of back spread, "learning" is an overseen procedure that happens with each cycle or epoch (i.e. each time the framework is given another data outline) through a forward incitation stream of yields, and the retrogressive slip causing of weight changes. Essentially, when a neural framework is at first given a case, it makes a subjective "assessment" in admiration to what it might be. It then sees how far its answer was from the certifiable one and makes a fitting acclimation to its affiliation weights. Inside every hidden layer node is a sigmoid activation function which delights framework activity and helps it to be stable in nature.

FEED FORWARD BACK PROPAGATION NEURAL NETWORK

This neural network architecture is very popular, because it can be applied to many different tasks. To understand this neural network architecture, we must examine how it is trained and how it processes a pattern. The first term, "feed forward" describes how this neural network processes and recalls patterns. In a feed forward neural network, neurons are only connected forward. Each layer of the neural network contains connections to the next layer (for example, from the input to the hidden layer), but there are no connections backwards. This differs from the Hopfield neural network that was examined to be fully connected, and its connections are both forward and backward. The term "back propagation" describes how this type of neural network is trained. Back propagation is a form of supervised training. When using a supervised training method, the network must be provided with both sample inputs and anticipated outputs. The anticipated outputs are compared against the actual outputs for given input. Using the anticipated outputs, the back propagation training algorithm then takes a calculated error and adjusts the weights of the various layers backwards from the output layer to the input layer

The back propagation and feed forward algorithms are often used together; however, this is by no means a requirement. It would be quite permissible to create a neural network that uses the feed forward algorithm to determine its output and does not use the back propagation training algorithm. Similarly, if you choose to create a neural network that uses back propagation training methods, you are not necessarily limited to a feed forward algorithm to determine the output of the neural network.

V. SUPPORT VECTOR MACHINE (SVM)

SVMs are very widespread apprentice. Support Vector Machines (SVM's) are a relatively new learning method used for binary categorization. The essential idea is to find a hyper plane which separates the d-dimensional data perfectly into its two categories. However, since example data is frequently not linearly separable, SVM introduces the notion of a "kernel induced feature space" which casts the data into a higher dimensional space where the data is divisible. Typically, casting into such a gap would cause problems computationally, and with over appropriate. The key near used in SVM's is that the higher-dimensional space doesn't need to be dealt with directly (as it turns elsewhere, only the formula for the dot product in that space is needed), which eliminates the above concerns. In addition, the VC-dimension (a measure of a system's likelihood to perform well on unseen data) of SVM's can be explicitly calculated, unlike other learning types like neural networks, for which there is no measure. Overall, SVM's are intuitive, theoretically well founded, and have shown to be nearly successful. SVM's have also been absolute to solve complex tasks (where the system is trained to output a numerical value, rather than "yes" "no" classification). In their fundamental form, SVMs study linear threshold function. Support vector machines are based on the Structural Risk Minimization theory from computational knowledge hypothesis. SVM are independent of the dimensionality of the feature space. Characteristics of SVM:

- High dimensional input space
- Document vectors are sparse
- Few irrelevant features
- Mainly text classification problems are linear

We are given l training examples $(x_i; y_i)$, i = 1,……..., l , where each examples has d inputs ($x_i \in R^d$), and a class label with one of two values ($y_i \in \{-1, 1\}$). Now, all hyper planes in R^d are parameterized by a vector (w), and a constant (b), expressed in the equation $w \cdot x + b = 0$ (Recall that w is in fact the vector orthogonal to the hyperplane.) Given such a hyper plane (w,b) that separates the data, this gives the function $f(x) = sign(w \cdot x + b)$ which correctly classifies the training data (and hopefully other "testing" data it has not seen yet). However, a known hyper plane represented by (w,b) is equally expressed by all pairs $\{\lambda w, \lambda b\}$ for $\lambda \in R^+$. So we describe the canonical hyper plane to be that which separates the data from the hyper plane by a "distance" of at least 1. That is, we consider those that satisfy:

$$xi \cdot w + b \geq +1 \text{ when } y_i = +1$$
$$xi \cdot w + b \leq -1 \text{ when } y_i = -1$$

or more compactly:

$$y_i (x_i \cdot w + b) \geq 1 \; \forall i$$

VI. LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation is a way of automatically discovering topics that these sentences contain. It considers each document as mixture of topics in which same words may exists in documents of other topics but with different probabilities. LDA is a hierarchical Bayesian model where document of a collection is

modeled as a finite mixture of underlying topics and topics are modeled as infinite mixture over an underlying set of topic probabilities. The base for LDA is the premise that words contain strong semantic information about the document. Therefore, it is reasonable to assume that documents on roughly similar topics will use the same group of words. Latent topics are thus discovered by identifying groups of words in the corpus that frequently occur together within documents. Each document is characterized by its own topic weight vector which indicates the amount of contribution of each of the K topics in that document using Dirichlet prior distribution. Then LDA uses Bayesian rule to determine the posterior distribution of latent topic variables based on the words in the document.

VII. RESULTS

In this work, three classifiers i.e. Neural Networks, SVM and LDA are used. These are used for classifying the short text messages posted on the OSNs user walls to categories like abusive/vulgar, politics and spam. The experiments are carried out on the sample messages taken from OSNs user walls. The aim of the work is to experimentally evaluate the performances of these three classifiers numerically to find the classifier for filtering the OSNs user wall short text messages The figure below shows the training section.

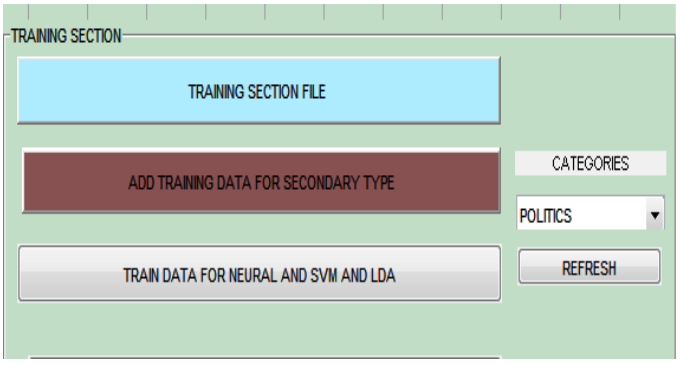

Figure 2: Training Section

This section deals with the training section .The neural network defines the targets and SVM defines the group. The training data has three types of categories that are- Politics, Abusive and Spam. The application is trained for the respective categories. Option is provided for the users to sign-up or login to the application and post messages to other users. Users can log-in and process the messages posted to their wall to check the classification type of that message.

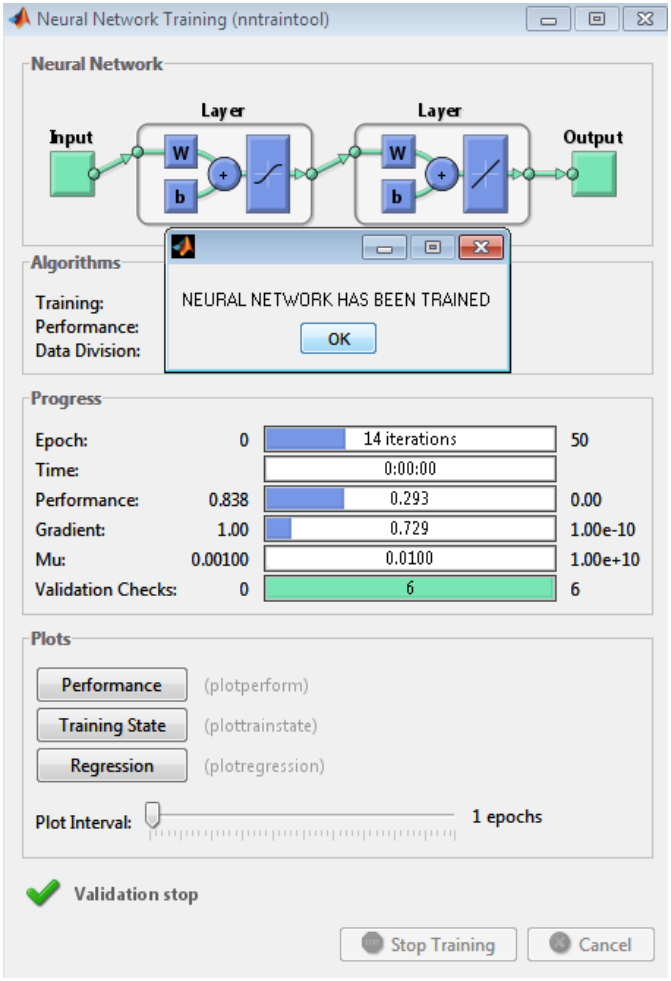

Figure 3: Neural network train tool window

The above figure shows the neural network toolbox window which displays the parameters like number of hidden neurons, number of iterations. The neural architecture deals with input layer, hidden layers and output layer which deal with the synaptic weights. The connection also deals with the activation function which processes the information from hidden layer to the output layer in number of epochs with magnitude of weights and validation checks. Newff (training set, target, hidden neurons) method is used for initializing the neural network.

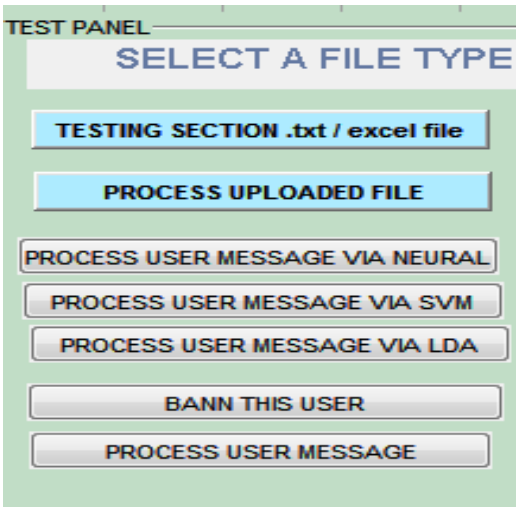

Figure 4: Testing section

This section shows the testing panel.

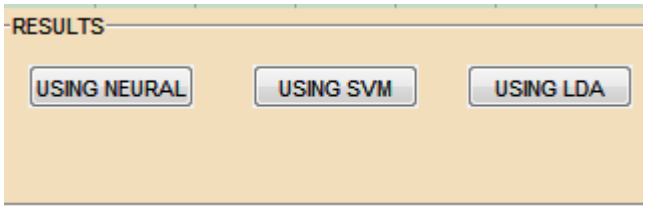

Figure 5: Result UI control

Table 1: SVM Values

FRR	FAR	ACCURACY	F.MEASURE
6.3667	3.7535	76.27	4.7227
12.7373	1.168	78.16	2.1397
10.8321	4.125	73.43	5.9747
7.1955	2.261	78.91	3.4408
6.2193	4.213	80.001	5.0232
8.2003	4.279	82.83	5.6235
5.8316	5.231	72.63	5.5149

Table 2: Neural Network values

FRR	FAR	ACCURACY	F.MEASURE
0.27	0.683	88.28	0.38701
0.29033	0.45	83.48	0.35295
1.1983	3.621	93.6713	1.80069
1.4711	3.7471	77.9687	2.11274
2.0228	4.2703	78.001	2.74522
1.319	2.8231	94.061	1.79796
1.5216	3.398	94.821	2.10196

Table 3: LDA values

FRR	FAR	ACCURACY	F.MEASURE
4.3216	2.7536	70.27	3.3638
10.6361	1.176	78.16	2.1178
5.821	5.216	73.46	5.5019
7.1906	3.279	79.81	4.5040
6.2831	4.209	81.86	5.0410
5.2093	5.216	70.83	5.2126
8.2836	3.221	72.16	4.6384

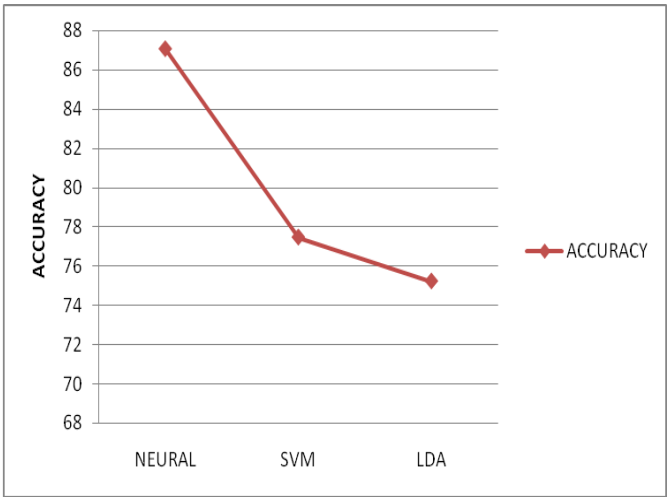

Figure 5: Performance comparison

The above figure shows the performance metric in terms of accuracy of neural network, Support Vector Machine and LDA classifiers which shows that neural network classifier provides better accuracy than other two classifiers for the proposed system.

VIII. CONCLUSION

In this paper, we have proposed a framework to filter undesired messages from OSN walls and experimentally evaluated the most suitable machine learning text classifier for short text messages. The framework develops a machine learning soft classifier for classifying the messages posted on the OSNs user walls into different categories. Additional features are included in the framework like filtering rules, automated blacklisting, additional features to enhance the learning of the classification system like Key term identification, Querying Microsoft Word Thesaurus/Word Net or using Google Sets. Based on the results computed on the experiments conducted, it is concluded that Neural Network is more suitable and provide better accuracy for classification of the short text messages posted on the OSNs user walls.

IX. FUTURE SCOPE

With the extensive use of social networking sites for sharing videos, images etc. other than the textual information, the present work can be extended to analyze the techniques well suited for content filtering of the multimedia files to allow user to have control preventing the unwanted multimedia content being posted.

REFERENCES

1. M. Chau and H. Chen, “A Machine Learning Approach to Web Page Filtering Using Content and Structure Analysis,” Decision Support Systems, vol. 44, no. 2, pp. 482-494, 2008.
2. R.J. Mooney and L. Roy, “Content-Based Book Recommending Using Learning for Text Categorization,” Proc. Fifth ACM Conf. Digital Libraries, pp. 195-204, 2000.
3. F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
4. M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, “Content-Based Filtering in On-Line Social Networks,” Proc. ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML ’10), 2010.
5. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short Text Classification in Twitter to Improve Information Filtering,” Proc. 33rd Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’10), pp. 841-842, 201
6. S. Pollock, “A rule-based message filtering system,” ACM Transactions on Office Information Systems, vol. 6, no. 3, pp. 232–254,1988
7. N.J. Belkin and W.B. Croft, “Information Filtering and Information Retrieval: Two Sides of the Same Coin?” Comm. ACM, vol. 35, no. 12, pp. 29-38, 1992.
8. Sarah Jane Delany," SMS spam filtering: Methods and Data ", Sarah Jane Delany, Mark Buckley, Derek Greene (2012) SMS Spam Filtering: Methods and Data, Expert Systems with Applications 39 (10), p9899-9908, Elsevier.
9. Tarek M Mahmoud," SMS Spam Filtering Technique Based on Artificial Immune System ", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814 www.IJCSI.org.
10. Chen M, Jin X, Shen D., “Short text classification improved by learning multi-granularity topics”. In Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three. AAAI Press, 2011:1776-1781.
11. J. Golbeck, “Combining Provenance with Trust in Social Networks for Semantic Web Content Filtering,” Proc. Int’l Conf. Provenance and Annotation of Data, L. Moreau and I. Foster, eds., pp. 101-108, 2006.
12. K. Strater and H. Richter, “Examining Privacy and Disclosure in a Social Networking Community,” Proc. Third Symp. Usable Privacy and Security (SOUPS ’07), pp. 157-158, 2007.