Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 5, Issue. 2, February 2016, pg.281 – 285

A Survey Paper on OSN User Wall to Filter Unwanted Message

Pauras Bangar[1], Lalita Randive[2]

[1]Department of Computer Engineering, Maharashtra Institute of Technology, Aurangabad
Dr.Babasaheb Ambedkar Marathwada University, India

[2]Asst.Professor, Department of Computer Engineering, Maharashtra Institute of Technology, Aurangabad
Dr.Babasaheb Ambedkar Marathwada University, India
[1] bangarpauras@gmail.com; [2] l.randive@gmail.com

Abstract— Nowadays, As web is growing rapidly, one of it's use is social communication. OSN are the most effective way of interactive medium to share, communicate and exchange vital information about human life in text, audio, video format. Current OSNs provide very little support to prevent unwanted messages on user walls. In OSNs, information filtering can be used for a different function. In OSNs there is the possibility of posting or commenting other vulgar, sexual, offensive posts on particular's public/private regions, called as general walls. Information filtering can therefore be used to allow users the ability to automatically filter the messages posted on their own walls, by filtering out unwanted messages. For instance Facebook allows who can post on Wall (i.e., friends, defined groups of friends or friends of friends).But no OSN allows , content based filtering to prevent unwanted message from posting on wall. For instance sexual, offensive, no matter of the user who post them. To propose and experimentally evaluate text based automated system called Filtered Wall, able to filter undesired message from OSN user wall

Keywords— Online social networks, Information filtering, Short text classifier, Policy based personalization, User wall.

I. INTRODUCTION

In recent years Online Social Networks(OSNs)became the most popular interactive channel to communicate, share and disseminate a considerable amount of human life information about Entertainment, Medical Science, History etc. On Daily basis and continuous communications imply the sharing of different types of content, including images, text data, audios and video data. According to Facebook, they have1.5 billion active monthly users, generates 4million post every minutes. Instagram with 300 million monthly users. Also 1 average user creates 90 pieces of substance every month, while more than 30 billion quantity of substance (web address, news information, study notes, blog posts, various albums, etc.) are distributed every month around globe. They are ready to give a dynamic support in complex and sophisticated function involved in OSN administration area, for example such information filtering on web. Information filtering has been significantly searched for what concerns textual documents and, more recently, web content [1], [2].The vast and dynamic behaviour of this information creates the scope for the employment of web analysis mining strategies which aimed to automatically discover usefulness of knowledge contained by the raw data.

In OSNs, information filtering can also be to provide users the ability to automatically control the messages posted on their individual private/public walls, by filtering out unwanted message. We believe that this is a key OSN service that has not been provided.

The aim of present work is to propose and experimentally evaluate an automated system, called Filtering Wall (FW), able to filter undesired messages from OSN user walls. We use Machine Learning (ML) text categorization techniques [3] to automatically assign with each short text message a set of categories based on its content. We developed robust short text classifier (STC) are used in the extraction and selection of a set of characterizing and discriminate aspects of the message[4].

In this paper, section II gives brief information about related work already performed for OSN user wall. In section III provides various content based machine learning methods so far used for text classification purpose.

II. RELATED WORK

A. CONTENT BASED FILTERING:

In content-based filtering, each user is assumed to operate independently. As consequence of it a content-based filtering system selects information items based on the correlation between the content of the document and the user priority as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences[5],[1]. Documents processed in content-based filtering are likely in text format in nature and this makes content-based filtering very close to text classification approach. The activity of filtering can be formulated, as a case of single label text classification, binary classification, partitioning incoming documents into relevant and no relevant classes. Most of the complex filtering methods include multilabel text categorization automatically labelling messages into predefined categories.

Content-based filtering is mainly based on the use of the Machine Learning paradigm according to which a classifier is automatically developed by learning from a set of preclassified data set. A remarkable different and related work has recently came which differ for the adopted feature extraction methods, model learning, and collection of sample data.

B. COLLABORATIVE FILTERING:

Collaborative filtering system selects information item based on user's preferences, actions and predicts what users will like based on his similarities to other users. Items are rated on the basis of user likes and dislikes [6]. Collaborative filtering involves collaboration of multiple agents while filtering information. Collaborative filtering system mostly needs huge amount of dataset.Amazon.com uses item to item collaborative filtering for its Recommendation system developed for it intent users. The collaborative approach is well suitable for popular items but in effective content information is not much good in result as opposite to content based approach for filtering which is more suitable for unpopular items in documents and effective content information is easily available for filtering purpose.

C. POLICY BASED FILTERING APPROCH:

In a classification method we classify short text messages in order to avoid large amount of users of micro blogging services by raw data. Some OSN provider uses relationship among the list of friends for providing filtering services, so that user can view only interested post, such as Golbeck and Kuter. Such systems do not provide a filtering policy layer by which the user can exploit the result of such classification methods to decide how and up to which extent filtering out unwanted information is carry out. In opposite of this, our filtering policy allows us to set the Filtering Rules as according to a variety of criteria, that consider the results of the classification process as well as the relationships of the wall owner with other OSN users who want to share something as well as information on the users private profile with others.

The only social networking service provider is MyWOT,3 a social networking service provider which gives its user the ability to: 1) rate OSN user with respect to four criteria:, child safety ,vendor reliability, privacy trustworthiness , and 2) specify preferences for determining whether the browser should block access to a given OSN user, or should simply give a warning message on the basis of the specified rating for the first time.

Content filtering can be considered as an extension of access control, since it can be used both to protect objects from unauthorized OSN users, and users from inappropriate access.[1] In the field of OSNs, the majority of access control models proposed to use topology-based access control mechanism, according to which access control requirements are expressed in terms of relationships that the requested user should have access with the wall owner. We use similar methods to identify the OSN users to which a Filtering Rules are applied. However, our filtering policy methods extend the models proposed for access control policy specification in OSNs to meet the demands of the extended requirements of the filtering domain.

III. CONTENT BASED FILTERING TECHNIQUES: RELATED WORK

Content-based filtering is mainly based on the use of the Machine Learning techniques, by which a short text classifier classifies incoming message automatically to predefined set of classes. These classes are created by using learning data.

A. Bag of word Approach:

The items extraction procedure relates text form document into a correct representation of its content and is uniformly applied to training data set . Experiments proved that Bag of Words (BoW) approaches gives good performance and satisfied result in some domain and also in more sophisticated text representation that may have good semantics but lower statistical quality of text data [7].

B. Radial Basis Function Neural Network.

Radial basis function neural network (RBFN) is can be used for a wide range of application primarily because it is able imprecise any regular function and its training is more faster than that of a multi-layer system. This faster learning speed comes from the realization that RBFN has two layers of weights and each of these layer can be determined sequentially for filtering purpose.

RBFNs has one abstract layer of processing units with local, restricted activities of the domain: a Gaussian function is commonly used, but any other function can be used for the same purpose. They were introduced as a neural network evolution of exact interpolation [8], and are demonstrated to have the universal approximation property The first level classifier is then restructured as a regular RBFN. In the second level of the classification process we use a modified standard of RBFN.

C. Machine Learning Approach:

A Machine learning approach learns from training data and develops classifiers for the categorization of new dataset. The main task of text classification is to assign a predefined classes with each text available in document. Text classification is done on the basis of large collection of training data set. The machine learning, based methods learn how to classify the categories of incoming text data on the basis of features available from the set of training data set. Below are the key methods which are commonly used for text classification.

1. Neural network classifiers
2. Support vector machines
3. Decision tree
4. K-Nearest Neighbors
5. Random Forest.

1. Support Vector Machine:

The support vector machine classifiers analyse data and recognize pattern in it. They are based on supervised learning approach and are able to perform nonlinear classification in addition to linear classification. The support vector machine classifier is good for large amount of unlabeled text data set and small amount of labelled text data set [6]. The high dimensional input space, irrespective of features in documents, creates document vectors and linearly separates text data which makes support vector machine suitable for text categorization purpose [6].

SVMs are very universal learners for mining purpose . In their basic form, SVMs learn linear threshold functions. One good property of SVMs is that their way to learn and can be independent of the dimensionality of the feature s in documents. [9]

2. Decision Tree:

Decision trees classification are used for a hierarchical decomposition of the data set . It finds the predicate or a condition depending on attribute values of text data set. Class labels are in the leaf node and are used for classification purpose. In order to reduce the over fitting of data is required in decision tree. This classifier model requires repetitive training methods and is more sensitive to training data set.[6]

When decision tree is used for text classification purpose it is consist tree in internal node and are label by term, branches separating from them are labelled by testing on the weights, and leaf nodes are represents corresponding class labels for classification. Tree can classify the text document by running the query structure from root to until it reaches to specified certain leaf node , which shows the goal for the classification of the document that is class label. Most of training data will not fit in memory for decision tree construction and it becomes inefficient due to swapping of training data. [10]

3. **Neural network classifiers:**

Neural network classifiers method consists of neurons which are arranged in different layers that converts an input vector into output vector space. The likely operated neural network is multilayer feed forward network in which a unit always gives its output to all the units of the next layer and there is no feedback to the previous layer from them. Radial basis function network is an artificial neural network where radial basis function as an activation function for text mining purpose. The output of this network is a linear combination of input to the radial basis functions and available neuron parameters. It is very robust to different outliers [5] and therefore more suitable in text filtering.

In neural fuzzy network method, for ranking and selection of HTML pages from the resulted web pages provided by external search engines and we attempt to solve the problem of describing user multi-interests in filtering text to improve the accuracy of current search facilities. The proposed system advice to the relevant pages to users according to the multi-interests of users in pages. For this, the system uses of a User Modeling system to acquire and store and restore the user's interests and non-interests about the pages. An advanced neural fuzzy network is applied to provide adaptive information filtering over the text document. The good features of the this system are embedding a user model in the neural fuzzy network to process the user's mixed interests of web page filtering and it is replacement of the traditional cosine measure method by a parameterized nonlinear map, so that it becomes straight forward to process the multi-interests web page filtering[11].

4. **K-Nearest Neighbors:**

K-NN classifier method is a case based learning algorithm that is uses a distances or similarity function for pairs of observations it may use Euclidean distance or Cosine similarity measures. Because of its effectiveness and non-parametric and easy to implement. The classification time is very long and it is difficult to find optimal value of k for classification. The best choice of k depends upon the data set and Mostly larger values of k which reduce the effect of noise on the classification methods, but we make boundaries between classes which are less different among all. A good k can be selected by various heuristic procedures.

A major drawback of the similarity measure used in k-NN is that it uses all features extracted from document in computing distances. In many document data sets, only smaller number of the total vocabulary may be useful in categorizing documents and classifying it. A possible approach to overcome this problem is to learn weights for different features of document in data set. [12] IT also propose the Weight Adjusted k-Nearest Neighbour (WAKNN) classification algorithm which is based on the k-NN classification paradigm by using unsupervised learning. With the help of KNN one can improve the performance and accuracy of text classification from large training data set and one can also increase the satisfactory combination of KNN with another method.

5. **Random Forest:**

Random forest is used for mining high dimensional data set with multiple classes who are used for text classification of data .A best feature is weighting method and tree selection method are used and synergistically served as base for making random forest framework best used to categorize text documents data with different topics. With the help of new feature weighting method for feature space sampling and tree selection methods, we can effectively reduce sub feature space size and increases classification performance without increasing error bounds in data set.

Breiman proposed a forest construction procedures and it is to randomly select a subspace of features set at each node of tree to grow branches of a decision trees. Also we use bagging approach to generate training data subsets from large amount of for building individual trees. Lastly we combine all individual trees created in model to prepare random forests model for classification of the text data.

IV. CONCLUSIONS

In this way we have studied the existing approaches for Short text filtering in OSN user wall as well as many techniques Text classification. There were many techniques implemented to filter unwanted messages from posting on user's private or public wall. So, as we have mentioned different techniques to that can be used in OSN user wall for text filtering purpose.

Current Text filtering methods having some limits, so the researchers are currently working on this area and would propose a system going under in reduce the time complexity and give better results in respect of quality, also proposed system will be language independent. The future work will be concentrate on the improved technique for filtering unwanted message from OSN user wall.

REFERENCES

[1] A. Adomavicius, G.and Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transaction on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[2] 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.

[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] 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.

[6] S. Zelikovitz and H. Hirsh, "Improving short text classification using unlabeled background knowledge," in Proceedings of 17th International Conference on Machine Learning (ICML-00), P. Langley, Ed. Stanford, US: Morgan Kaufmann Publishers, San Francisco, US, 2000, pp. 1183–1190.

[7] C. Apte, F. Damerau, S. M. Weiss, D. Sholom, and M. Weiss,"Automated learning of decision rules for text categorization," Transactions on Information Systems, vol. 12, no. 3, pp. 233–251, 1994..

[8] M. J. D. Powell, "Radial basis functions for multivariable interpolation: a review," pp. 143–167, 1987.

[9] Text Categorization with Support Vector Machines: Learning with Many Relevant Features Thorsten Joachims Universit at Dortmund Informatik LS8, Baroper Str. 301 44221 Dortmund, Germany.

[10] D. E. Johnson F. J. Oles T. Zhang T. Goetz, "A decision-tree-based symbolic rule induction system for text Categorization", by IBM SYSTEMS JOURNAL, VOL 41, NO 3, 2002.

[11] NEURAL NETWORKS-BASED MULTI-INTEREST INFORMATION FILTERING Dai Xuewu1, Vic Grout2, Tang Haokun3 and Li Jianguo1.

[12] Fang Lu Qingyuan Bai, "A Refined Weighted K-Nearest Neighbours Algorithm for Text Categorization", IEEE 2010.

