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. 3, Issue. 7, July 2014, pg.974 – 979

RESEARCH ARTICLE

A System to Customize Content Based Messages Filtering for On-Line Social Networks

Miss. Vidya Alone[1], Mrs. R.B.Talmale[2]

[1]Final Year M. Tech CSE, Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur

[2]HOD of CSE Department, Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur

Email: vidya.alone@gmail.com, roshanikambe@rediffmail.com

Abstract* — Today On-line Social Networks (OSNs) is one of the most popular interactive medium to share and communicate among the internet user. But there is one fundamental issue is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now OSNs provide little support to this requirement. These papers give comprehensive review on various techniques of flexible rule-based system, and Machine Learning based soft classifier. Whereas flexible rule-based system helps to customize the filtering criteria. Machine Learning based soft classifier for automatically labeling messages in support of content-based filtering.

***Index Terms* - On-line Social Networks, Information Filtering, Short Text Classification, Policy-based Personalization**

I. INTRODUCTION

On-line Social Networks (OSNs) are today one of the most popular interactive medium to communicate, share information. in the form of text, image, audio and video data etc. Today every day and every month a huge amount of data is shared so, it require active support in complex and difficult task involved in OSN management such as information filtering and access control. Information filtering concerns textual documents and web content (e.g., [5], [1]). Main aim is to provide ability to user to automatically control the messages written on their own wall by avoiding unwanted messages. This wall messages very short so do not provide sufficient word occurrences so traditional classification methods fail here. Filtered Wall (FW), specify Filtering Rules can support a variety of different filtering criteria that can be combined and customized according to the user needs to filter unwanted messages from OSN user walls. Machine Learning (ML) is a text categorization techniques it automatically categories text based on its content. Neural learning is the efficient solutions in text classification [4]. Radial Basis Function Networks (RBFN) is use for short text classification, managing noisy data and fuzzy classes. User-defined Blacklists (BLs), that is, lists of users that are temporarily prevented to post any kind of messages on a user wall. The rest of the paper is organized as follows. Section II surveys related work. Section III filtered wall architecture section IV Text classification techniques to categorize text contents.
Section V illustrates FRs and BLs. Finally, section VI concludes the paper.

II. LITERATURE REVIEW

Nicholas J. Belkin and W. Bruce Croft has been discussed relationship between information filtering and information retrieval and they come to the conclude that both are two sides of the same coin [11]. The previous recommended systems use social filtering methods that base recommendations on other users' preferences. By contrast R. J. Mooney and L. Roy describe a content-based book recommending system that utilizes information extraction and a machine-learning algorithm for text categorization. This way they improve access to relevant products and information [1]. The assignment of natural language texts categorization is an important component in many information organization and management tasks. S. Dumais, J. Platt, D. Heckerman, and M. Sahami compare the effectiveness of five different automatic learning algorithms for text categorization in terms of learning speed, real-time classification speed, and classification accuracy. and they conclude that Linear Support Vector Machines (SVMs) are most accurate classifier, fastest to train, and quick to evaluate. They used SVMs for categorizing email messages and Web pages .They also hope to extend their work by including the additional structural information about documents, as well as knowledge-based features for classification accuracy and automatically classify items into hierarchical category structures [12].The widely use text representation techniques for text retrieval are phrase indexing and clustering. D. D. Lewis studied the properties of phrasal and clustered indexing, to isolation from query interpretation issues. He worked on same number of features for each category and there was no automated feature selection.[13].R. E. Schapire and Y. Singer, describe in detail an implementation, called BoosTexter, of the new boosting algorithms for text categorization tasks. and also they compare its performance with a number of other text-categorization algorithms on a variety of tasks.[14]. Neural network allows us to model the higher order interaction between document terms and to simultaneously predict multiple topics using shared hidden futures. E. D. Wiener, J. O. Pedersen, and A. S. Weigend, presents an application of nonlinear neural network to topic spotting. Whereas topic spotting is the problem of identifying which of the set of predefined topics are present in a natural language document [15]. Sarah Zelikovitz, Haym Hirsh describe a method for improving the classification of short text strings using a combination of labeled and unlabeled but related longer documents [16].

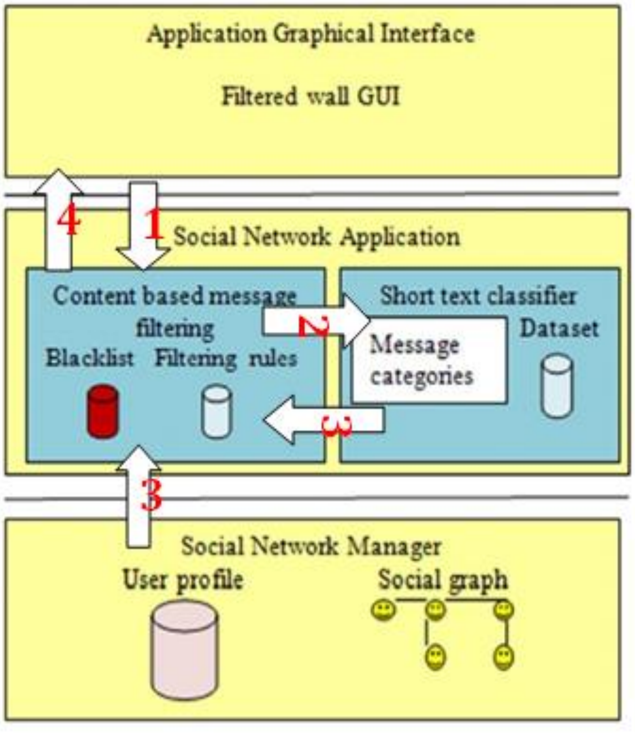

Fig. 1 Conceptual Architecture of Filtered Wall

III. FILTERED WALL ARCHITECTURE

The architecture of OSN services is a three-tier structure. It consists of three layers (Fig 1) [7]. First layer is Social Network Manager (SNM), second layer is Social Network Application (SNA) third and last layer is Graphical User Interface (GUI). Profile and relationship management is the main task of SNM layer. It maintains the data related to user profile and provides the data to the second layer for applying filtering rules (FR) and blacklists (BL). SNA layer composed of short text classifier and Content Base Message Filtering (CBMF). SNA layer most important layer. Because here the classifier categorizes each message according to its content and CBMF filters the message according to filtering criteria and blacklist provided by the user. Third layer consist of graphical user interface by which user provide his input and is able to see published wall messages. Moreover GUI also provides user the facility to apply filtering rules for his wall messages and helps to provide list of BL user who are temporally prevented to publish messages on user's wall. The GUI also consists of Filtered Wall (FW) where the user is able to see his desirable messages. As per the filtered wall architecture, when the user tries to post a message on a private wall of his or her contact it is first intercepted by the filtered wall. Then a short text classifier categories a message according to its content and CBMF applies FR and BL as per the data provided by the third layer. Based on the result of above step only messages that are authorized according to their filtering rules are published OSN user walls.

IV. AUTOMATIC TEXT CLASSIFICATION

Today, text classification is a necessity due to the very large amount of text documents that we have to deal with daily. Text Classification is the task to classify documents into predefined classes. Text Classification is also called text categorization, document classification, and document categorization. There are two approaches for classification manual classification and automatic classification. Relevant technologies for Text Classification are

1. Text Clustering
2. Information Retrieval (IR)
3. Information Filtering
4. Information Extraction (IE)
5. Text Classification.

Text clustering that Create clusters of documents without any external information. Information Retrieval (IR) that retrieve a set of documents relevant to a query. Information Filtering Filter out irrelevant documents through interactions. Information Extraction (IE) extracts fragments of information, e.g., person names, dates, and places, in documents. In text Classification there is no query, no interactions, no external information, it only decide topics of documents. Text Classification use in various fields like E-mail spam filtering, categorize newspaper articles and newswires into topics, organize Web pages into hierarchical categories, sort journals and abstracts by subject categories (e.g., MEDLINE, etc.), assigning international clinical codes to patient clinical records etc. Established techniques used for text classification work well on datasets with large documents such as newswires corpora [6], but suffer when the documents in the corpus are short. There are various representation techniques and neural learning strategy that can be use combine to semantically categorize short texts. For classifying text or short texts a hierarchical strategy is use [5]. A different sets of features has been consider for text categorization ,there are three different issues pertaining namely document representation, classifier construction, and classifier evaluation. Text categorization is the activity of labeling natural language texts with thematic categories from a predefined set [4]. However the most appropriate feature set and feature representation. are, Bag of Words(BoW), Document properties (DP) and Contextual Features (CF).[7][3]

A. BAG-of-Words Model (BOW)

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval (IR). In this model, a text (such as a sentence or a document) is represented as the bag (multi set) of its words, disregarding grammar and even word order but keeping multiplicity. Recently, the bag-of-words model has also been used for computer vision. The bag-of-words model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

B. Documents Properties (DP)

Document properties as a uniform basis for interaction. Document properties express high-level features of documents that are meaningful to users and usable by systems. Document properties are directly associated with documents, rather than with document storage locations. This means that documents will retain properties even when moved from one place to another, and that property assignment can have a fine granularity.

C. Contextual Features (CF)

The A contextualized strategy might allow IR systems to learn and predict what information a searcher needs, learn how and when information should be displayed, present results relating them to previous information and to the tasks the user has been engaged in and decide who else should get the new information [10].

D. *Vector Space Model* (*VSM*)

The Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. The VSM representation of the text that characterizes the environment where messages are posted There are some strength of SVM: Training is relatively easy i.e. no local optimal, unlike in neural networks, It scales relatively well to high dimensional data, Tradeoff between classifier complexity and error can be controlled explicitly. By performing logistic regression (Sigmoid) on the SVM output of a set of data can map SVM output to probabilities. But still there is one weakness ie we need to choose a “good” kernel function.

Summary: Steps for Classification

1. Prepare the pattern matrix
2. Select the kernel function to use
3. Select the parameter of the kernel function and the value of *C*
4. You can use the values suggested by the SVM software, or you can set apart a validation set to determine the values of the parameter
5. Execute the training algorithm and obtain the a_i
6. Unseen data can be classified using the a_i and the support vectors

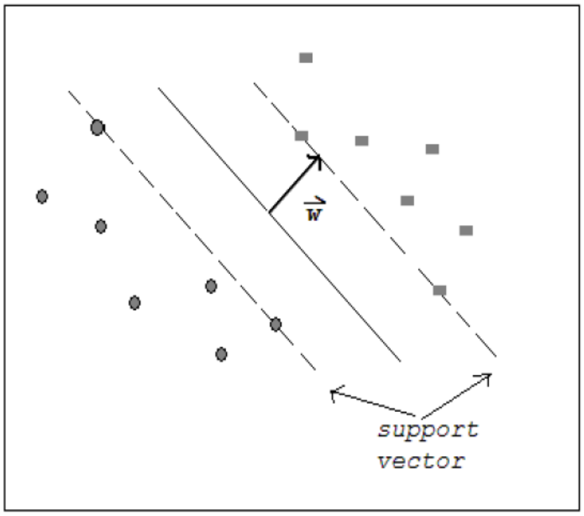

Fig. 2 SVM classification plane.

E. *Machine Learning-Based Classification*

There are two approaches that you can take that is rule based approach and machine learning-based approach. Rule based approach write a set of rules that classify documents. Machine learning-based approach using a set of sample documents that are classified into the classes (training data) automatically create classifiers based on the training data. Machine Learning is another way of getting computers to classify documents. Machine learning is normally not rule based. Instead, it is normally statistically based. It’s the ability of a machine to improve its performance based on previous results so, machine learning document classification is “the ability of a machine to improve its document classification performance based on previous results of document classification.

Examples of ML algorithms.

F. *Naïve Bayes*

This method computes the probability that a document is about a particular topic, T, using a) the words of the document to be classified and b) the estimated probability of each of these words as they appeared in the set of training documents for the topic, T – like the example previously given.

G. *Neural Networks*

During training, a neural network looks at the patterns of features (e.g. words, phrases, or N-grams) that appear in a document of the training set and attempts to produce classifications for the document. If its attempt doesn’t match the set of desired classifications, it adjusts the weights of the connections between neurons. It repeats this process until the attempted classifications match the desired classifications.

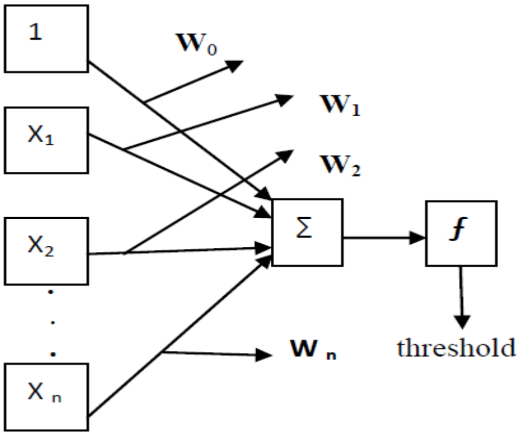

Fig. 3. Neural network Architecture.

H. Instance Based

The Saves documents of the training set and compares new documents to be classified with the saved documents. The document to be classified gets tagged with the highest scoring classifications. One way to do this is to implement a search engine using the documents of the training set as the document collection. A document to be classified becomes a query/search. A classification, C, is picked if a large number of its training set documents are at the top of the returned answer set. There are varieties of multi-class ML models well-suited for text classification. Whereas RBFN model is the best model. [2].

I. Radial Basis Function Network (RBFN)

The In the field of mathematical modelling, a RBFN is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. Radial basis function networks have many uses, including function approximation, time series prediction, classification, and system control. RBFN main advantages are that classification function is non-linear, the model may produce confidence values and it may be robust to outliers; drawbacks of RBFN are the potential sensitivity to input parameters, and potential overtraining sensitivity [2].

V. Filtering Rules And Blacklist Management

A. Filtering Rules

The same message on OSNs may have different meanings and relevance based on who writes it. It is necessary to apply constraints on messages. Constraints can be selected on several different criteria's. User can state what contents should be blocked or displayed on filtered wall by means of Filtering rules. Filtering rules are specified on the basis of user profile as well as user social relationship.FR is dependent on following factors.

1. Author
2. Creator Spec
3. Content Spec
4. Action

An author is a person who defines the rules. Creator Spec denotes the set of OSN user and Content Spec is a Boolean expression defined on content. Action denotes the action to be performed by the system on the messages matching content Spec and created by users identified by creator Spec [9].

B. Online Setup Assistant for FRS Thresholds

There are problem of setting thresholds to filter rule. Online Setup Assistant (OSA) procedure allows user to select a set of messages from dataset of messages. Such messages are selected by certain amount of non-neutral messages taken from a fraction of the dataset and not belonging to the training/test sets, are classified by the ML in order to have, for each message, the second level class membership value. On the basis of this the user tells the system the decision to accept or reject the message. The collection and processing of user decisions on an sufficient set of messages circulated over all the classes allows to compute customized thresholds representing the user attitude in accepting or rejecting certain contents.

VI. BLACKLIST

BL users are those users whose messages are prevented independent from their contents. BL rules enable the wall owner to determine users to be blocked on the basis of their profiles and relationship with wall owner. This banning can be done for a specified period or forever according wall owner's desire. Like FR, BL is also dependent on author, creator specification and creator behavior. For denoting users' bad behavior considered two main measures. The first is, if any user has already into BL and again inserted many times in it for given time interval, say greater than a given threshold, he/she might deserve to stay in the BL. Second is to consider Relative Frequency (RF), it detect those users whose messages continue to fail the Fr^s.

VII. CONCLUSION

In this paper, we have presented a system to filter unwanted messages from OSN walls. Text classification is the most important task in machine learning and data mining. Our work surveys on the different kinds of text categorization techniques ,machine learning methods like decision trees, naïve bayes, support vector machines, neural networks and Instance Based, Radial Basis Function Network .Using, these techniques classification is efficiently done. The flexibility of the system in terms of filtering options is enhanced through the management of BLs. Using machine learning, filtering rules; black list management should be effectively implemented.

REFERENCES

[1] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in Proceedings of the Fifth ACM Conference on Digital Libraries. New York: ACM Press, 2000, pp.195–204.

[2] R. Prashant Tomer1, "On Line Social Network Content And Image Filtering Classifications," ISSN 2319-5991 Vol. 2, No. 4, © 2013 IJERST. November 2013

[3] M. Carullo, E. Binaghi, I. Gallo, and N. Lamberti, "Clustering of short commercial documents for the web," in Proceedings of 19th International Conference on Pattern Recognition (ICPR 2008), 2008.

[4] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1–47, 2002.

[5] 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.

[6] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," Journal of Machine Learning Research, 2004.

[7] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-based filtering in on-line social networks," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML 2010),2010.

[8] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, Moreno Carullo,"A System to Filter Unwanted Messages from OSN User Walls," IEEE Transaction on Knowledge and Data Engineering, vol. 25, 2013.

[9] DIK L. LEE "Document Ranking and the Vector-Space Model" Hong Kong University of Science and Technology HUEI CHUANG, Information Dimensions KENT SEAMONS, Transarc.

[10] Carla Teixeira Lopes "Context Features and their use in Information Retrieval" Doctoral Program in Informatics Engineering of Faculdade de Engenharia da Universidade. 2007.

[11] N. J. Belkin and W. B. Croft, "Information filtering and information retrieval: Two sides of the same coin?" Communications of the ACM, vol. 35, no. 12, pp. 29–38, 1992.

[12] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in Proceedings of Seventh International Conference on Information and Knowledge Management (CIKM98), 1998, pp. 148–155.

[13] D. D. Lewis, "An evaluation of phrasal and clustered representations on a text categorization task," in Proceedings of 15th ACM International Conference on Research and Development in Information Retrieval (SIGIR-92), N. J. Belkin, P. Ingwersen, and A. M. Pejtersen, Eds. ACM Press, New York, US, 1992, pp. 37–50.

[14] R. E. Schapire and Y. Singer, "Boostexter: a boosting-based system for text categorization," Machine Learning, vol. 39, no. 2/3, pp. 135– 168, 2000.

[15] E. D. Wiener, J. O. Pedersen, and A. S. Weigend, "A neural network approach to topic spotting," in Proceedings of 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR-95), Las Vegas, US, 1995, pp. 317–332.

[16] 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.