ISSN:0975-9646
Mahesh Jadhav et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (4) , 2015, 3611-3614

Customizable Content Based Filtering Method to Filter Texts and Images from Social Network User Wall

Mahesh Jadhav[1], Subhash Pingale[2], Ravindra Takalikar[3]

[1,2,3] *Department of Computer Science and Engineering, Solapur University*
SKN Sinhgad College of Engineering, Pandharpur, Maharashtra,India

***Abstract*— Today online social media are most valuable and essential member of human life. People can online communicate with their family, society and friends to exchange several type of information including text, images, audio & video data. Therefore, there is need of control of user over the contents which are published on his wall. Hence , user is able to avoid undesirable content that is displayed on his wall. But, currently social networks provide this service in very small extent. To provide this service, we plan a user defined filtering rules and machine learning categorization technique in this paper. Filtering rules grant users to personalize the filtering norms that are employed to the contents which are published on their walls. Automatic categorization based on content of messages into proposed categories is possible through machine learning technique. Also, proposed system can restrict the undesired images that are published on online social network (OSN) user private space by using content wise image filtering along with K-Nearest Neighbour (KNN) classifier.**

***Keywords*— Online Social Network (OSN), Machine Learning, K-Nearest Neighbour (KNN) Classifier, Information Filtering.**

I. INTRODUCTION

Nowadays social media are most popular channel of communication that provides easy connection between people over the world .OSNs like Facebook provides easy manner to share, exchange and communicate huge extent of personal life information. Live communication between people is possible through OSNs. Online conversation involves sharing of huge extent of static data (text, image) and constantly varying data (video and audio). Facebook statistics states that 90 parts of contents are created by ordinary user each month and more than 30 billion parts of information are shared each month. Therefore, unwanted messages are likely to post on user private wall. Customizable and automatic information filtering is needed to avoid posting of unneeded messages on user wall. Today OSNs are not in favor to automatically restrict the messages published on their walls. Content wise filtering is not provided by OSNs. e.g.- In Facebook one can state who is allowed to publish the messages on his private area but he cannot state what content should be posted on his private area. Hence, it is impossible to restrict the nonstandard messages such as vulgar, offense, hate etc. independent from the person who publishes them. Most of the social media messages are made up of short length. As short texts do not provide enough word occurrences and standards, traditional classification methods are poor to classify short texts. In past periods, digital image databases are expanded in huge amount. A large no of images are shared and exchanged by people in OSNs. But user cannot decide which images are allowed to publish on his private area. Therefore, to overcome above issues, we plan a system that comprises of filtering wall, machine learning technique, content wise image filtering along with KNN classifier. Undesired messages are restricted from publishing on user private space by filtering wall. Machine learning classifier can autonomously categorize each message into proposed categories depending on content of that message. Section II states related work. Section III states proposed system. Section IV states machine learning techniques.. Section V proposes short text classifier. Section VI states filtering rules and blacklist. Section VII states online setup assistant. Section VIII explains content based image filtering. Section IX concludes the paper. .

II. RELATED WORK

To provide customizable content wise message filtering and content wise image filtering, we must consider following research in area of information filtering and retrieval. Boykin and Roychowdhary [4] provides an antispam technique that can automatically identify unexpected commercial email. This technique utilizes the characteristics of social networks. Still, it does not support customizable content wise filtering of emails. Summary of recommendation methods is presented by Adomavicius and Gediminas Tuzhilin [3] .Following are two main recommendation methods-

Content-wise recommendations [10], here those items are predicted to user that are close to the items favoured in former period. System must rate sufficient no of items to know the user tastes and present the user with reliable recommendations. Hence, users with low rankings cannot get exact recommendations.

Collaborative-based recommendations [17], here those items are forecasted to user that are close to the people with close interests. Huge dataset is required for collaborative filtering. Only popular items are effectively recommended by collaborative approach. Collaborative filtering does not give accurate information based on content.

Yusuf and Othman [2] propose categorization of web content by using feature extractor and machine learning phase. Feature extractors convert the text from its normal form to its semantic form. Machine learning phase classify the content depending on its semantics into out of two possible classes. Hence, machine learning phase enlarge the classification edge. But limitation is that proposed approach is effective for only long form text and not for short texts that are broadly used in OSN. Short length texts are hard to categorize as they have nonstandard terms, less no of occurrences. Hirish and Zelikovitz [16] try to enhance the categorization of short length text by imposing supervised learning approach. To well understand the trending topics posted on Twitter, Kathy Lee and his team[6]classify them into ordinary categories such as sports, politics, technology,

etc. by using text classification method (NaïveBayes classifier) [13] and network classification method(decision tree)[6]. Still, user cannot state how and to which extent filtering out unwanted information. Chich-Chin Lai,Ying-Chaun Chen [12] propose retrieval of image based on user interests to quickly and accurately retrieve the required images from abundant image database. They consider user subjectivity and preferences in the retrieval process. Still, user cannot decide which images are allowed to post on his wall. Macro Vanetti and Elisabetta Binaghi [1] proposes system to prevent undesired contents from social network users private space. But, they propose this system for only text form messages and not for messages in other form like video, image etc. Information access can be personalized according to user defined policies. Policy based personalization approach [5] group every tweet into various categories like news, events etc. So, user can view tweets of his interest. But, here user cannot rule the result of classification process. Therefore, we prefer to provide customizable content wise filtering in online social media.

A. Disadvantages of Existing System

- Existing system do not provide content wise filtering of information. Hence, publishing of undesired messages on user private space cannot be avoided.
- Mostly wall messages are composed of short length texts. As short length texts are rarely occurred and they bear unusual terms, traditional classification methods cannot effectively classify them.

III. PROPOSED SYSTEM

Proposed system architecture is shown in Fig 1. Proposed system filtered out undesired contents in following way-

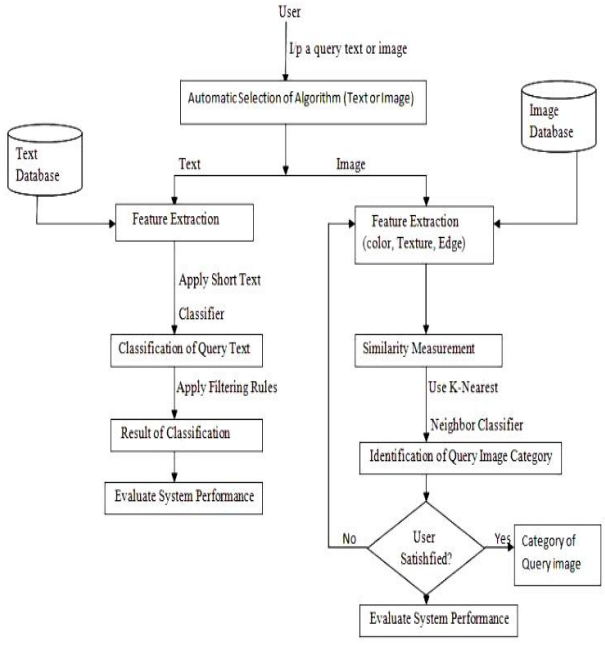

Fig. 1 Proposed system architecture

Firstly it identifies content being published is either image or text. If it is text then first move is to withdraw features from it.

Second move is to categorize that text based on its features by using radial basis algorithm [8].

Then apply filtering rules [1] to that classified text to impose the conditions on it. Finally result of categorization is obtained i.e. text is appeared or blocked on user private space.If content is image then features are extracted from that image by using various feature descriptors [12]. Similarity of query image is measured and its class is identified by using KNN classifier. If user is unsatisfied with identified category then again process starts from feature extraction of image.

IV. MACHINE LEARNING TECHNIQUES

Machine learning technique is efficient and accurate method for content classification. It categorizes the incoming content rest on the previous results of training document classification. It can automatically classify the documents without human interruption. It performs consistent classification. Following are machine learning arts-

A. Naive Bayes

This classifier figure out the possibility that a document represents particular subject [17].Advantage is that it demands a low extent of labelled data for determining the parameters needed for classification.

B. Support Vector Machines

These machines recognize the framework of data by analysing it [17]. They use framework of data for its classification. They use machine learning strategy which is supervised. As they provide top prediction exactness and robustness, it is suitable for text categorization.

C. Neural Nets

These nets are comprises of large processing segments i.e. neuron elements. They are ordered in three layers i.e. input layer, hidden layer and output layer. They map input variable to output variable. This mapping is controlled by various adjustable parameters. Radial basis functions (RBFs) are neural nets. They are robust to outliers [8]. Hence, we select RBFs classifier for classification.

V. SHORT TEXT CLASSIFIER

A. Text Representation

Execution of accurate categorization of text is greatly depend on appropriate set of patterns that are drawn from text. Document properties, bag of words , contextual patterns are three pattern types. Foremost two pattern types [10] are endogenous patterns as they are gained from the message content itself. Contextual patterns are exogenous patterns as they are gained from the content external to the message anatomy. e.g. environment where the messages are published. Bag of words represent the terms in text file in words form. Text file is formatted as real or binary weights vector as per vector space text representation model [15].

B. Preferred Algorithm for Short Text Classification

We choose RBF network algorithm [8] for short text classification.RBF network performs hierarchical

categorization which is two level task. First level task is binary hard categorization that categorizes the content into out of two possible classes i.e. fair or unfair class. Second level task is soft categorization that categorizes the unfair content by calculating its membership value into each of unfair classes. We choose RBF network algorithm as it is robust to outliers and its classification function is non linear.

VI. Filtering Rules And Blacklist

Filtering rules impose the restrictions on the users who create the messages. These users are selected by utilizing their profile attributes. Filtering rule has ingredients namely author, creator specification, content specification, action. Author is person who creates the rule. Creator specification is nothing but group of OSN users. Content specification specifies conditions on the content in Boolean form. Action is in two forms i.e. notify or block action that is applied on the contents which matched with content specification and creator specification [1].

Blacklist consists of users to be prohibited independent of their message content. User can interpret blacklist rule to specify the users to be prohibited and the period for which they are banned from posting of messages on user wall.

VII. Online Setup Assistant

In OSA process [1], collection of short length messages is given to the user. For every message, user will take the stand to receive or dismiss the message. From these decisions, OSA calculates the minimum membership level threshold needed for class.

VIII. Content Based Image Filtering

For content based image filtering, it is necessary to extract the visual features such as colour, texture, edge etc. of image. Following feature extractors are used to extract the visual features of image-

A. Hue, Saturation, Value (HSV) Descriptor

HSV [12] colour descriptor is used to withdraw colour features of image as it consider the human interpretation of colours.

B. Grey Level Co-Occurrence Matrix (GLCM)

Texture is an important feature of image. It is used to withdraw surface characteristics of an object. GLCM [12] method is effective and more accurate method to extract texture of image.

C. Edge Histogram Descriptor (EHD)

Edge is an vital attribute of image to describe the content of image. Human eyesight can easily interpret the image from its edge characteristics. EHD [12] can appropriately visualize the edge distribution of an image.

D. K-Nearest Neighbour (KNN) Classifier

Image is categorized into proposed categories of images in database by withdrawing the feature of that image.KNN classifier can appropriately perform this classification. KNN classifier selects k-nearest images that are similar with the query image. Then it applies majority voting to obtain category of the query image.

IX. Conclusions And Future Scope

In this survey, we outline a system that can restrict undesired texts and images from private space of user of social network. We studied different text representation techniques, feature extraction techniques and machine learning techniques required for classification of OSN content. Short length text categorization algorithm i.e. RBF network classifier can automatically categorize the messages into group of categories. Proposed system grants user to establish flexible filtering rules that are applied on categorized messages to restrict the undesired messages that are published on user private space. Presented system can block undesired images that are published on user private space by applying various feature extractors and KNN classifier on images. So, user has command on the messages that are published on his private area. As underlying domain is persistently changing, a group of precategorized data used for training intend is inadequate for longer time. Future work is to implement blacklist.

X. Acknowledgment

I am profoundly grateful to Prof. Subhash Pingale and Prof. Ravindra Takalikar for their expert guidance and continuous encouragement throughout to see that this paper rights its target since its commencement to its completion. At last I must express my sincere heartfelt gratitude to all people who helped me directly or indirectly during this course of work.

References

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

[2] L.Yusuf and M.Othman, “Web Classification using Extraction and Machine Learning Techniques,” *IEEE Trans. DataMining*, 2010.

[3] A.Adomavicius and G.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.

[4] P.Boykin and V.Roychowdhury, “Leveraging social networks to fight spam,”*IEEE*,2005.

[5] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu,and M.Demirbas, “Short Text Classification in Twitter to Improve Information Filtering,” *Proc. 33rd Intl ACMSIGIR Conf. Research and Development in Information Retrieval (SIGIR 10)*, pp. 841-842,2010.

[6] Kathy Lee, Diana Palsetia, Ramanathan Narayanan, Md. Mostofa Ali Patwary, Ankit Agrawal, and Alok Choudhary,”Twitter Trending Topic Classification,” *11th IEEE International Conference on Data Mining Workshops*,2011.

[7] P. E. Baclace, “Competitive agents for information filtering,” *Communications of the ACM*,” vol. 35, no. 12, p. 50, 1992.

[8] Oliver Buchtala, Manuel Klimek, and Bernhard Sick,” Evolutionary Optimization of Radial Basis Function Classifiers for Data Mining Applications,” *IEEE transactions on systems, man, and cybernetics—part b: cybernetics*, vol. 35, no. 5, October 2005.

[9] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[10] 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.

[11] M.J.D. Powell, “Radial Basis Functions for Multivariable Interpolation: A Review,” *Algorithms for Approximation*, pp. 143-167, Clarendon Press, 1987.

[12] Chich-Chin Lai, Ying-Chaun Chen, "A User Oriented Image Retrieval System Based on Interactive Genetic Algorithm," *IEEE Transaction on Instrumentation and Measurement*,Vol.60,No.10,2011.

[13] Zhou Faguo, Zhang Fan, Yang Bingru and Yu Xingang, "Research on Short Text Classification Algorithm Based on Statistics and Rules," IEEE, *Third International Symposium on Electronic Commerce and Security*,2010.

[14] D.D. Lewis, Y. Yang, T.G. Rose, and F. Li, "Rcv1: A New Benchmark Collection for Text Categorization Research," *J. Machine Learning Research*, vol. 5, pp. 361-397, 2004.

[15] C.D. Manning, P. Raghavan, and H. Schu¨ tze, *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008.

[16] S. Zelikovitz and H. Hirsh, "Improving Short Text Classification Using Unlabeled Background Knowledge," *Proc. 17th Int'l Conf. Machine Learning (ICML '00)*, P. Langley, ed., pp. 1183-1190, 2000.

[17] K. Nirmala, S. Satheesh kumar, "A Survey on Text Categorization in Online Social Networks," *in Proceedings of International Journal of Emerging Technology and Advanced Engineering*, Volume 3, Issue 9, September 2013.