

Filtering Information for Short Text Using OSN

S. Venkata Lakshmi, K. Hema

^{1,2}Assistant Professor, Dept. of Computer Science & Engg., KMMITS college, Tirupati, India

Abstract

Online social networks (OSN) are used to share the information among the different kind of people. There is a major task of online social network is information filtering. An online social network provides the little support for allowing sharing the information on the user walls. Using, machine learning algorithms text classification is to be done. Text categorization is applied to the set of pre classified documents. Only the unwanted messages will be blocked not the user. To avoid this issue, BL (Black List) mechanism is proposed Machine Learning Text Categorization is also used to categorize the short text messages.

Keywords

Online social networks, information filtering, text classification, machine learning

I. Introduction

Online social networking sites like Orkut, YouTube, and Face book are among the most popular sites on the Internet. Users of these sites forms a social network, which provides a powerful means of sharing, organizing, and finding useful information. Unlike web content, the Online social networks (OSN) are organized around more number of users joins the network, shares their information and create the links to communicate with other users. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users. In online social networks information filtering can be used for avoiding the undesired messages sharing or commenting on the user Walls. Various machine learning methods are used for classification. The development of social networking among the people, it allows users to customize their own privacy settings. Our main aim is to survey the OSN for filtering the undesired messages.

Information and communication technology plays a significant role in today's networked society. It has affected the online interaction between users, who are aware of security applications and their implications on personal privacy. There is a need to develop more security mechanisms for different communication technologies, particularly online social networks. OSNs provide very little support to prevent unwanted messages on user walls. With the lack of classification or filtering tools, the user receives all messages posted by the users he follows. In most cases, the users receive a noisy stream of updates. In this paper, an information Filtering system is introduced.

The aim of the present work is therefore to propose and experimentally evaluate an automated system, called Filtered Wall (FW), able to filter unwanted messages from OSN user walls. We exploit Machine Learning (ML) text categorization techniques to automatically assign with each short text message a set of categories based on its content. The major efforts in building a robust short text classifier are concentrated in the extraction and selection of a set of characterizing and discriminant features.

Besides categorization capabilities, the system offers a powerful rule layer utilizing a flexible language to specify Filtering Rules (FRs), by which users can state what substances, should not be showed on their walls. FRs can maintain a variety of different filtering criteria that can be combined and customized according to the user requirements. In particular, FRs utilizes user profiles, user relations as well as the production of the ML categorization process to state the filtering criteria to be forced. Additionally, the system gives the support for user-defined Blacklists (BLs), that is, lists of

users that are temporarily prevented to post any kind of messages on a user wall. Main dissimilar includes a different semantics for filtering rules to best fit the measured domain, an OSA to aid users in FR specification, the extension of the set features considered in the classification process, a more deep performance evaluation plan and an update of the prototype implementation to reflect the changes made to the classification methods.

II. Related Work

The main goal of information filtering system is to filter unwanted content from input data before its presentation to the end user. It takes into account user profile and compares it with referred characteristics or properties. Recommender systems have become popular in recent years.

Filtering Based Contents

Information filtering systems are designed to classify a stream of dynamically generated information dispatched asynchronously by an information producer and present to the user those information that are likely to satisfy his/her requirements. Focusing on the OSN domain, interest in access control and privacy protection is relatively recent. As future as confidentiality is disturbed, current work is essentially focusing on privacy-preserving data mining methods, that is, protecting data associated to the network, i.e., relations/nodes, while performing social network study. Effort more associated to our schemes is those in the field of access control. In this field, various dissimilar access control models and associated mechanisms have been proposed so far which essentially differ on the expressivity of the access control policy language and on the way access control is enforced (e.g., centralized vs. decentralized). The majority of these models convey access control requirements in terms of relationships that the requestor should have with the resource holder. We use a related idea to classify the users to which a filtering rule applies. Though, the general purpose of our suggestion is absolutely different, while we effectively agreement with filtering of unwanted substances rather than with access control. For itself, one of the key elements of our scheme is the availability of an explanation for the message contents to be exploited by the filtering mechanism as well as by the language to express filtering rules.

In distinguish no one of the access control models previously cited exploit the content of the resources to enforce access control. We consider that this is an essential difference. Furthermore, the concept of blacklists and their administration are not believed by any of these access control models. The application of content-based filtering on messages posted on OSN user walls poses

additional challenges given the short length of these messages other than the wide range of topics that can be discussed. Short text categorization has acknowledged up to now few attentions in the scientific community.

Using rule base engine components, filtering concept is applied to the Online Social Network user wall. Latest effort highlights complexities in significant robust aspects, effectively due to the fact that the explanation of the short text is brief, with various misspellings, nonstandard conditions, and noise. Zelikovitz and Hirsh attempt to improve the classification of short text strings developing a semi-supervised learning strategy based on a combination of labeled training data plus a secondary corpus of unlabeled but related longer essays.

It is a type of information filtering system that predicts the preference that user might give to an item or to the social element. It takes into account user interest and recommends an item. Recommender systems works in one of three ways:

1. Content based filtering
2. Collaborative filtering
3. Policy based filtering

A. Content-based filtering

Content based filtering selects an item based on user interest. It uses items previously preferred by the user and then suggests the best matched item. Each user acts independently in content based system. This kind of system chooses item depending on relation between item content and user recommendations against collaborative system that selects item based on relation between people with similar preferences. The content based system creates a content based profile of a user based on rated items of a user. Items features are weighted based on features preferred by the user and recommendations are given by the system accordingly. In content based filtering, the main issue is whether the system is able to learn from user's actions related to a particular content source and use them for other content types. Text classification is similar to content based filtering as documents processed in such type of system are mostly textual. In online social network user's social profile has to be taken into account and this makes content based filtering system difficult to apply in OSN domain as a standalone system.

B. Collaborative filtering

In collaborative filtering information will be 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. Twitter is the best example for policy based filtering. In that communication policy can be defines between two communicating parties.

Architecture of Filtered Wall

The architecture of OSN services is a three-tier structure of three layers (Figure 1). These three layers are

1. Social Network Manager (SNM)
2. Social Network Application (SNA)
3. Graphical User Interface (GUI)

1. Social Network Manager (SNM)

The initial layer is 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. After maintaining and administrating all users data will provide for second layer for applying Filtering Rules (FR) and Black lists (BL).

2. Social Network Application (SNA)

In second layer Content Based Message Filtering (CMBF) and Short Text Classifier is composed. This is very important layer for the message categorization according to its CMBF filters. Also Black list is maintained for the user who sends frequently bad words in message.

3. Graphical User Interface (GUI)

Third layer provides Graphical User Interface to the user who wants to post his messages as a input. In this layer Filtering Rules (FR) are used to filter the unwanted messages and provide Black list (BL) for the user who are temporally prevented to publish messages on user's wall.

In general, the architecture in support of OSN services is a three-tier configuration. The initial layer generally aims to offer the essential OSN functionalities (i.e., profile and relationship administration). In addition, some OSNs offer an extra layer allowing the support of external Social Network Applications (SNA) 1. Finally, the supported SNA may require an additional layer for their needed graphical user interfaces (GUIs). According to this orientation layered structural plan, the planned system has to be positioned in the second and third layers (Figure 1), as it can be considered as a SNA. Particularly, users cooperate with the system by means of a GUI setting up their filtering laws, along with which messages have to be filtered out. In addition, the GUI offers users with a FW that is a wall where only messages that are authorized according to their filtering rules are published. The core components of the proposed system are the Content-Based Messages Filtering (CMBF) and the Short Text Classifier elements. The latter element aims to categorize messages according to a set of categories. In compare, the first element exploits the message categorization offered by the STC module to implement the FRs specified by the user. As graphically illustrated in Fig. 1, the path pursued by a message, it can be summarized as follows:

1. After entering the private wall of one of his/her associates, the user attempts to post a message, which is captured by FW.
2. A ML-based text classifier extracts metadata from the content of the message.
3. FW uses metadata provided by the classifier, mutually with data extorted from the social graph and users' profiles, to implement the filtering and BL rules.
4. Depending on the result of the previous step, the message will be available or filtered by FW.

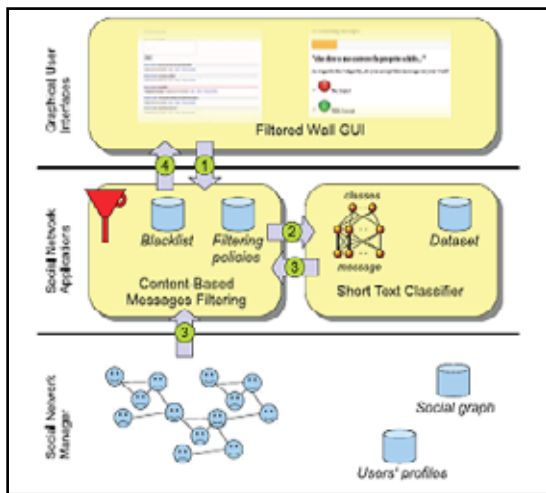


Fig. 1: Architecture of Filtered wall

III. Filtering Rules and Blacklist Management

A. Filtering Rules

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. Action
4. Content Spec

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

Content spec on the messages matching content Spec and created by users identified by creator Spec.

B. Blacklists

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.

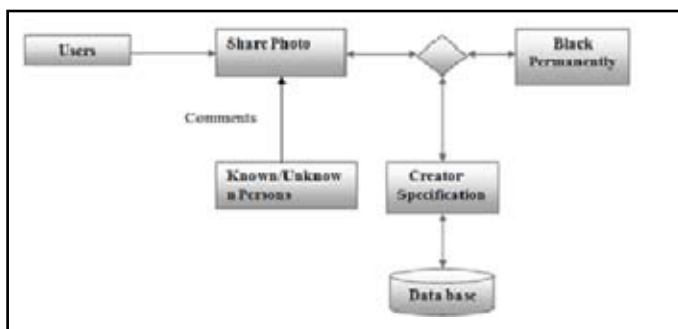


Fig. 2: Blacklist System

By applying the BL rule, owner can identify which user should be blocked based on the relationship in OSN and the user's profile. The user may have bad opinion about the users can be banned for an uncertain time period. We have two information based on bad

attitude of user. Two principles are stated. First one is within a given time period user will be inserted in BL for numerous times, he /she must be worthy for staying in BL for another sometime. This principle will be applied to user who inserted in BL atleast once. Relative Frequency is used to find out the system, who messages continue to fail the FR. Two measures can be calculated globally and locally, which will consider only the message in local and in global it will consider all the OSN users walls.

IV. Machine Learning Techniques

A Machine learning approach learns from training data and creates classifiers for the classification of new dataset. The main task of text classification is to assign a predefined category with each text. Text classification is accomplished on the basis of endogenous collection of data. The machine learning, based classifier learns how to classify the categories of incoming data on the basis of features extracted from the set of training data. Below are the key methods which are commonly used for text classification.

1. Naive Bayes classifier
2. Neural network classifiers
3. Support vector machines
4. Decision tree

A. Naive Bayes classifier

Naive Bayes classifier is a probabilistic classifier based on Bayes theorem with independence assumption. Given a class variable, it assumes the presence or absence of specific feature is unrelated to the presence or absence of any other feature. For instance a fruit is considered to be cherry if it is red, round and small in size. Bayes classifier considers each of these features independently to the probability that the fruit is cherry regardless of the presence or absence of any other feature. The main advantage of this classifier is that it requires a small amount of training data to estimate the parameters required for classification.

B. Neural Network Classifier

Neural network classifiers consist of neurons arranged in layers converting an input vector into output. The most commonly used neural network is multilayer feed forward network in which a unit feeds its output to all the units of the next layer but there is no feedback to the previous layer. Radial basis function network is an artificial neural network which uses radial basis function as an activation function. The output of this network is a linear combination of radial basis functions of the inputs and neuron parameters. It is robust to outliers and therefore more suitable in this context.

C. Support Vector Machines

The support vector machine classifiers analyze data and recognize pattern in it. They are based on supervised learning model and are able to perform nonlinear classification in addition to linear classification. The support vector machine classifier is suitable for large amount of unlabeled data and small amount of labeled data. The high dimensional input space, irrelevant features, sparse document vectors and linearly separable text classification makes support vector machine classifier suitable for text categorization.

D. Decision Trees

Decision trees classifiers are used for a hierarchical decomposition of the data space. It determines the predicate or a condition

depending on attribute value. Class labels in the leaf node are used for classification. In order to reduce the over fitting data pruning is required in decision tree. This classifier requires iterative training procedure and is oversensitive to training data .

V. Short Text Classifier

A hierarchical two level classification is advantageous to short text classification as per the suggestion .The first level of a classifier labels the message into neutral and non-neutral. In second level non neutral messages are estimated into one or more of the conceived categories.

A. Text Representation

Text representation of a given document is important task strongly affecting the performance of classification process. It is done by extracting features for a given document. The investigation from suggest three types of features important for text representation. They are Bag of Words, Document properties (Dp) and Contextual Features (CF).The first two types of features are entirely derived from the information contained within the text of the message whereas contextual features are exogenous. Text representation using endogenous. In Bag of Words representation terms are identified with words. It is also important to use Feature which is extracted from outside the message content but related to message itself. A contextual feature is introduced in that characterize the environment where the user is posting. According to .It determines the semantics of message .Vector space Model is the model of text representation by which a text document is represented as a vector of binary or real weights. These three features are experimentally evaluated for short text classification in for their appropriateness.

B. Machine Learning-based Classification

As short text classification is hierarchical two level tasks and it should be robust to outliers hence RBFN is used for short text classification. An RBFN model is chosen as per the experimental evaluation in among the other classifiers.

VI. Conclusion

Existing system is used to filter undesired messages from OSNs wall using customizable filtering rules (FR) enhancing through Black lists (BLs). In present system (www.winow.in), we are more focus on an investigation of two interdependent tasks in depth. This system approach decides when user should be inserted into a black list.

The system developed GUI and a set of tools which make BLs and FRs specifications more simple and easy. Investigation tools may be able to automatically recommend trust value of the user. The primary work of this system is to find out trust values used for OSN access control. In this system we will provide only core set of functionalities which are available in current OSNs like Facebook, Orkut, Twitter,etc. In existing OSNs have some difficulties in understanding to the average users regarding privacy settings. But this problem will be overcome in present OSNs system.

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] 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.
- [5] A.Adomavicius and G.Tuzhilin, "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the -Art and Possible Extensions, " *IEEE Trans. Knowledge and Data Eng.*,vol.17, no.6,pp.734-749, June 2005
- [6] B.Sriram, D.Fuhry, E.Demir, H.ferhatatosmanoglu, and M.Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," *Proc.33rd Int'l ACM SIGIT Conf. Research and Development in InformationRetrieval(SIGIR '10)*, pp.841-842,2010.
- [7] V.Bobicev and M.Sokolova, "An Effective and Robust Method for Short Text Classification," *Proc.23rd Nat'l Conf. Artificial Intelligence (AAAI)*, D.Fox and C.P.Gomes, eds., pp.1444-1445,2008.
- [8] 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.
- [9] J.Nin, B.Carminati, E.Ferrari, and V.Torra, "Computing Reputation for Collaborative Private Networks," *Proc.33rd Ann. IEEE Int'l computer Software and Applications Conf., Vol.1*, pp. 246-253, 2009.
- [10] R.J.Mooney and L.Roy, "Content-Based Book recommending Using Learning for Text Categorization", 2000.