Filter Wall: To prevent undesired messages posted on OSN user wall

Amruta Kachole, S.D. Jondhale

ME(computer), amrutakachole@gmail.com

Abstract—This paper is support for content based user preferences. It is possible to the use of a Machine Learning (ML) text categorization procedure able to automatically assign with each message a set of categories based on its content. The proposed approach is a key service for social networks where users have little control on the messages displayed on their walls. For instance, Facebook allows users to state who is allowed to insert messages in their walls (i.e., friends, friends of friends, or defined groups of friends). However, no content-based preferences are supported. For instance, it is not possible to prevent political or vulgar messages. In contrast, by means of the proposed mechanism, a user can specify what contents should not be displayed on his/her wall, by specifying a set of filtering rules. Filtering rules are very flexible in terms of the filtering requirements they can support, in that they allow to specify filtering conditions based on user profiles, user relationships as well as the output of the ML categorization process. In addition, the system provides the support for user-defined blacklist management, that is, list of users that are temporarily prevented to post messages on a user wall.

Keywords—Online social networks, content based filtering, short text classification, Space Vector Model (SVM)

INTRODUCTION

On-line Social Networks (OSNs) have become a popular interactive medium to communicate, share and disseminate a considerable amount of human life information. Daily and continuous communication implies the exchange of several types of content, including free text, image, audio and video data. The huge and dynamic character of these data creates the premise for the employment of web content mining strategies aimed to automatically discover useful information dormant within the data and then provide an active support in complex and sophisticated tasks involved in social networking analysis and management. The main part of social network content is constituted by short text, a notable example are the messages permanently written by OSN users on particular public/private areas, called in general walls.

In this paper an automated filtering system is implemented for Content based filtering that allows OSN users to have a direct control on the messages posted on their walls. This proposed approach can automatically filter unwanted messages from OSN user walls on the basis of content of message. It also proposes a flexible rule-based system that allows users to customize the filtering criteria to be applied to their walls and a Machine Learning-based soft classifier automatically labeling messages in support of content-based filtering. The core components of the proposed system are the Content-Based Messages Filtering (CBMF) and the Short Text Classifier modules. The short text classifier component aims to classify messages according to a set of categories. STC is performed as a hierarchical two level classification process. The first-level classifier performs a binary hard categorization that labels messages as Neutral and Non-neutral. The second-level classifier performs a soft-partition of Non-neutral messages assigning a gradual membership to each of the non-neutral classes. Therefore, ML-based short text classifier extracts metadata from the content of the
(message) In contrast, the Content-Based Messages Filtering component exploits the message categorization provided by the STC module to enforce the FRs specified by the user. BLs can also be used to enhance the filtering process.

**Modules**

Module 1: Vector Presentation

Module 2: Binary Classification

Module 3: Multi Label Classification

Module 4: Filtering Rules Specification

Module 5: Blocking Management

**Module 1: Vector Representation**

In this project, training sets are prepared from the data set, WmSnSec which is available online at http://www.dicom.uninsubria.it/~marco.vanetti/wmsnsec. Training set is divided into neutral and non-neutral training datasets to train the SVM classifier. SVM is trained by extracting the content of messages in the dataset (wmsnsec). This approach follows Vector Space model, according to which a text message \( d_j \) is represented as a vector of binary or real weights \( d_j = \{w_{1j}, w_{2j}, w_{3j} \ldots w_{|T|j}\} \) where \( T \) is the set of terms that occur at least once in at least one document of the collection \( T_r \) and \( w_{kj} \in [0;1] \) represents that how much the term \( k \) contributes to the semantics of the document. Training \((T_r,S_D)\) set are transformed into a form of vector representation

\[
T_rS_D = \{(x_1^r, y_1), \ldots (x_{|T|SD}^r, y_{|T|SD})\}
\]

Test set \((T_e,S_D)\) are transformed into a vector representation as &.

\[
T_eS_D = \{(x_1^e, y_1), \ldots (x_{|T|SD}^e, y_{|T|SD})\}
\]

The term frequency-inverse document frequency is used to calculate the weight of term \( t_k \) in document \( d_j \) as follows

\[
Tf-idf \text{ weighting} = \frac{\# (t_k, d_j) \times \log N}{\# (t_k, N)}
\]

\( \# (t_k, d_j) \) is the term frequency where the number of occurrences of term \( t_k \) in the document \( d_j \).

\( \log N/ \# (t_k, N) \) is the inverse document frequency i.e., document frequency the number of documents have the term \( t_k \) among all the documents \( N \).
Module 2: First level Binary classification

In this project, SVM perform two level of classification to filter the unwanted short text based on its content. Let m1 be the first level classifier used to classify the messages into two types such as Neutral and Non neutral messages.

Module 3: Multi Label Classification

In this module, the messages which are labeled as non neutral messages are given as an input to the second classifier M2 that performs multi-label classification. In order to perform classification, the classifier M2 is trained using multi-class training set as follows. The performance of the model M2 is then evaluated using the test set TeS2.
Module 4: Filtering Rule Specification

Besides classification facilities, this project provides a powerful rule layer exploiting a flexible language to specify Filtering Rules (FRs), by which users can state what content should not be displayed on their walls. FRs supports the specification of content-based filtering using variety of different filtering criteria in order to combine and customize the user needs. More precisely, FRs exploit user profiles, user relationships as well as the output of the ML categorization process to state the filtering criteria to be enforced. FRs should allow users to state constraints on message creators. This implies to state conditions on type, depth, and trust values of the relationship(s) creators should be involved in order to apply them the specified rules. A Filtering Rule FR is represented as a tuple as follows

\[ \text{FR} = (\text{author, creatorSpec, contentSpec, action}) \]

\( \rightarrow \) author is the user who specifies the rule;

\( \rightarrow \) creatorSpec is a creator specification,

\( \rightarrow \) contentSpec is a Boolean expression that expresses constraints in the form \((c, ml)\) where \(C\) is a class of the first or second level and \(ml\) is the minimum membership level threshold required for class \(C\) to make the constraint satisfied.

\( \rightarrow \) action \(\in\) \{block; notify\} denotes the action to be performed by the system on the messages matching contentSpec and created by users identified by creatorSpec.

Module 5: Block List Management

Block List management is used to avoid messages from undesired creators, independent from their contents. To achieve this, user specifies the information through a set of rules called as BL rules. These rules are directly managed by the system, which determine who are the users will be inserted into the Block List and decide when user retention in the BL is finished. A BL rule is a
tuple \{author, creatorSpec, creatorBehavior, T\}, where creatorBehavior consists of two components RFBlocked and minBanned. RFBlocked \((RF, \text{mode}, \text{window})\) is defined such that

\[
\text{Relative Frequency} = \frac{\text{bMessages}}{\text{tMessages}},
\]

\(\rightarrow \text{tMessages}\) is the total number of messages that each OSN user identified by creatorSpec has tried to publish in the author wall (\text{mode} = \text{myWall}) or in all the OSN walls (\text{mode} = \text{SN}); whereas \#bMessages is the number of messages among those in \#tMessages that have been blocked;

\(\rightarrow \text{window}\) is the time interval of creation of those messages that have to be considered for RF computation;

\(\rightarrow \text{minBanned} \equiv (\text{min}, \text{mode}, \text{window})\), where \text{min} is the minimum number of times in the time interval specified in \text{window} that OSN users identified by creatorSpec have to be inserted into the BL due to BL rules specified by author wall (\text{mode} = \text{myWall}) or all OSN users (\text{mode} = \text{SN}) in order to satisfy the constraint.

\(\rightarrow T\) denotes the time period the users identified by creatorSpec and creatorBehavior have to be banned from author wall.

CONCLUSION
In OSN environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to the difficulties in traditional classification approaches, thereby requiring intensive investigation. This project presented a system to filter undesired messages from OSN walls. This project improved the quality of classification using short text classifier. It provides high privacy and flexibility to manage OSN walls. In this project, we have investigated the privacy problem of user by flexible filtering rule specification and designed a group of hierarchical level classification to assign the metadata for each of the message posted by the user. This project creatively specified flexible rule specification to directly control the messages posted on their walls. It concretely accomplishes the automated filtering in a highly flexible way.

REFERENCES: