ISSN:0975-9646

Filtering Undesired Messages from Online Social Networks: A Content Based Filtering Approach

Dhruv Vashistha
Computer Science and Engineering
SRM University
Kattankulathur, Chennai, Tamil Nadu, India

Sivagami.G.
Assistant Professor
Computer Science and Engineering
SRM University
Kattankulathur, Chennai, Tamil Nadu, India

***Abstract*— This paper proposes a system that implements a content-based message filtering service for Online Social Networks (OSNs). Our system allows OSN users to have a direct control on the messages that are posted on their walls. This is done through a rule-based system, that allows a user to customize the filtering criteria, which is to be applied to their walls, and a Machine Learning based classifier which can automatically produce membership labels for the support of our content-based filtering mechanism.**

I. INTRODUCTION

In the last years, Online Social Networks (OSNs) have become a popular interactive medium to communicate, share and spread a considerable amount of human life information. Daily and continuous communication leads to exchange of several types of content, including text, image, audio and video. 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 within the data and then provide an active support in sophisticated tasks involved in social networking analysis and management. A main part of social network content is composed by short text, a notable example are the messages permanently written by OSN users on particular public/private areas, called in general walls.

The aim of the present work is to propose and a system, called Filtered Wall (FW), able to filter unwanted messages from social network user walls. The idea of the proposed system is the support for content-based user preferences. This is possible with the use of a Machine Learning (ML) text categorization procedure [11] able to automatically assign with each message a set of categories based on its content. We believe that the proposed strategy is a useful service for social networks as in today social networks users have little control on the messages displayed on their walls. For example, 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 setting a set of *filtering rules*. Filtering rules are very flexible in terms of the 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.

The remainder of this paper is organized as follows: in Sect. 2 we describe work closely related to this paper, sect.3 explains the existing system and its drawbacks, Sect. 4 introduces the conceptual architecture of the proposed system. Sect.5 describes the ML-based text classification method used to categorize text contents, whereas Sect. 6 provides details on the content-based filtering system. Finally, Sect. 7 concludes the paper.

II. RELATED WORK

Content-based filtering has been investigated by exploiting ML techniques [2, 9, 10] as well as other strategies [8, 5]. However, the problem of applying content-based filtering on the contents exchanged by users of social networks has received up to now few attentions in the scientific community. One of the few examples in this direction is the work by Boykin and Roychowdhury [3] that proposes an automated anti-spam tool that, exploiting the properties of social networks, can recognize unwanted commercial e-mail, spam and messages associated with people the user knows. However, it is important to note that the strategy just mentioned does not exploit ML content-based techniques.

The advantages of using ML filtering strategies over other engineering approaches are a very good effectiveness, flexibility to changes in the domain and portability in differ applications. However difficulties arise in finding an appropriate set of features by which to represent short, grammatically ill formed sentences and in providing a consistent training set of manually classified text.

Focusing on the OSN domain, interest in access control and privacy protection is quite recent. As far as privacy is concerned, current work is mainly focusing on privacy-preserving data mining techniques, that is, protecting information related to the network, i.e., relationships/nodes, while performing social network analysis?

Work more related to our proposals is those in the field of access control. In this field, many different access control models and related mechanisms have been proposed so far (e.g., [4, 12, 1, 6]), which mainly differ on the expressivity of the access control policy language and on the way

access control is implemented (e.g., centralized vs. decentralized). Most of these models express access control requirements in terms of relationships that the requestor should have with the resource owner. We use a similar idea to identify the users to which a filtering rule applies. However, the overall goal of our proposal is completely different, since we mainly deal with filtering of unwanted contents rather than with access control. As such, one of the key ingredients of our system is the availability of a description for the message contents to be used by the filtering mechanisms' well as by the language to express filtering rules. In contrast, no one of the access control models previously seen to use the content of the resources to enforce access control. We believe that this is the basic change and difference. Moreover, the use of blacklists and their management are not considered by any of these access control models.

III. EXISTING TECHNIQUE:

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 preference is supported and therefore it is not possible to prevent undesired messages, such as political or vulgar ones, no matter of the user who posts them. Providing this service is not only a matter of using previously defined web content mining techniques for a different application, rather it requires to design ad hoc classification strategies. This is because wall messages are constituted by short text for which traditional classification methods have serious limitations since short -texts do not provide sufficient word occurrences.

Existing Technique explanation:
Inductive Learning Algorithms

Step 1: Find Similar
Our Find Similar method is a variant of Rocchio's method for relevance feedback which are a popular method for expanding user queries on the basis of relevance judgments'. In Rocchio's formulation, the weight assigned to a term is a combination of its weight in an original query, and judged relevant and irrelevant documents.

Step 2: Decision Trees
The decision trees were grown by recursive greedy splitting, and splits were chosen using the Bayesian posterior probability of model structure. A class probability rather than a binary decision is retained at each node.

Step 3: Naive Bayes
A naive-Bayes classifier is constructed by using the training data to estimate the probability of each category given the document feature values of a new instance. We use Bayes theorem to estimate the probabilities:

$$P(C = c_k \mid x) = \frac{P(x \mid C = c_k)P(C = c_k)}{P(x)}$$

Step 4: Bayes Nets
It is allows for a limited form of dependence between feature variables, thus relaxing the very restrictive assumptions of the Naïve Bayes classifier. We used a 2-dependenceBayesian classifier that allows the probability of each feature *xi* to be directly influenced by the appearance/non-appearance of at most two other features.

Step 5: Support Vector Machines
Support Vector Machines have only recently been gaining popularity in the learning community. In its simplest linear form, an SVM is a hyper plane that separates a set of positive examples from a set of negative examples with maximum margin.

Some Drawbacks:
It is machine based classifier used for this system. It also has only the classifier so after compare to contents, message will display the public wall and user cannot handle the spam messages directly.

.
We believe that inductive learning methods like the ones we have described can be used to support flexible, dynamic, and personalized information access and management in a wide Variety of tasks. Linear SVMs are particularly promising since they are both very accurate and fast.

IV. PROPOSED ARCHITECTURE

The aim of this paper is to develop a method that allows OSN users to easily filter undesired messages, according to content based criteria. In particular, we are interested in defining a language-independent system providing a flexible and customizable way to filter and then control incoming messages. Before the architecture of the proposed system, we briefly introduce the basic model underlying OSNs. In general, the standard way to model a social network is as directed graph, where each node corresponds to a network user and edges denote relationships between two different users. In particular each edge is labelled by the *type* of the established relationship (e.g., friend of, colleague of, parent of) and, possibly, the corresponding *trust* level, which represents how much a given user, considers trust worthy with respect to that specific kind of relationship the user with whom he/she is establishing it. Therefore, there exists a direct relationship of a given type and trust value between two users, if there is an edge connecting them having the labels mentioned. Moreover, two users are in an indirect relationship of a given type if there is a path of more than one edge connecting them, such that all the edges in the path have labels [7].
In general, the architecture in support of OSN services is a three-tier structure. The first layer commonly aims to provide the basic OSN functionalities (i.e., profile and relationship management). Additionally, some OSNs provide an additional layer allowing the support of external Social Network Applications (SNA). Finally, the supported SNA may require an additional layer for

their needed graphical user interfaces (GUIs). According to this reference layered architecture, the proposed system has to be placed in the second and third layers (Figure 1), as it can be considered as a SNA. In particular, users interact with the system by means of a GUI setting up their filtering rules, according to which messages have to be filtered out. Moreover, the GUI provides users with a FW that is a wall where only messages that are authorized according to their filtering rules are published.

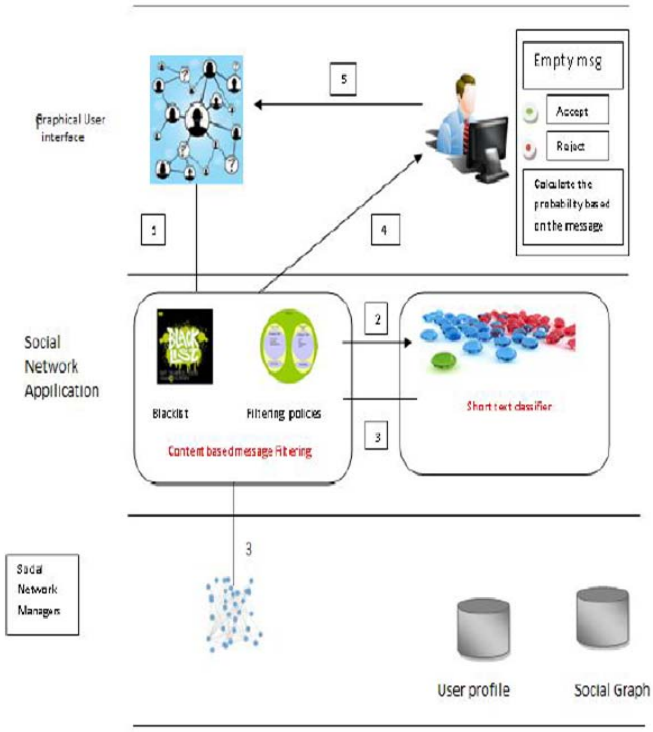

Figure.1

The core components of the proposed system are the Content-Based Messages Filtering (CBMF) and the Short Text Classifier (STC) modules. The latter component aims to classify messages according to a set of categories. The strategy underlying this module is described in Sect. 4. In contrast, the first component exploits the message categorization provided by the STC module to enforce the filtering rules specified by the user. Note that, in order to improve the filtering actions, the system makes use of a blacklist (BL) mechanism. By exploiting BLs, the system can prevent messages from undesired users. More precisely, as discussed in later Sect., the system is able to detect who are the users to be inserted in the BL according to the specified user preferences, so to block all their messages and for how long they should be kept in the BL.

V. SHORT TEXT CLASSIFIER

Established techniques used for text classification work well on data sets with large documents such as newswires corpora but suffer when the documents in the corpus are short. In this context, critical aspects are the definition of a set of characterizing and discriminant features allowing the representation of underlying concepts and the collection of a complete and consistent set of supervised examples.

Our study is aimed at designing and evaluating various representation techniques in combination with a neural learning strategy to semantically categorize short texts. From a ML point of view, we approach the task by defining a hierarchical two-level strategy assuming that it is better to identify and eliminate “neutral” sentences, then classify “nonneutral” sentences by the class of interest instead of doing everything in one step. This choice is motivated by related work showing advantages in classifying text and/or short texts using a hierarchical strategy [1]. The first-level task is conceived as a hard classification in which short texts are labelled with crisp Neutral and Nonneutral labels. The second-level soft classifier acts on the crisp set of nonneutral short texts and, for each of them, it “simply” produces estimated appropriateness or “gradual membership” for each of the conceived classes, without taking any “hard” decision on any of them. Such a list of grades is then used by the subsequent phases of the filtering process.

VI. FILTERING RULES AND BLACKLIST MANAGEMENT

In this section, we introduce the rule layer adopted for filtering unwanted messages. We start by describing FRs, and then we illustrate the use of BLs. In what follows, we model a social network as a directed graph, where each node corresponds to a network user and edges denote relationships between two different users. In particular, each edge is labelled by the type of the established relationship (e.g., friend of, colleague of, parent of) and, possibly, the corresponding trust level, which represents how much a given user, considers trustworthy with respect to that specific kind of relationship the user with whom he/she is establishing the relationship. Without loss of generality, we suppose that trust levels are rational numbers in the range 0; 1Š. Therefore, there exists a direct relationship of a given type label named and trust values (assume as X) between two users, if there is an edge connecting them having the labels and values. Moreover, two users are in an indirect relationship of a given type if there is a path of more than one edge connecting them, such that all the edges in the path have label as same. In this paper, we do not address the problem of trust computation for indirect relationships, since many algorithms have been proposed in the literature that can be used in our scenario as well. Such algorithms mainly differ on the criteria to select the paths on which trust computation should be based, when many paths of the same type exist between two users.

i) Filtering Rules

In defining the language for FRs specification, we consider three main issues that, in our opinion, should affect a message filtering decision. First of all, in OSNs like in everyday life, the same message may have different meanings and relevance based on who writes it. As a consequence, FRs should allow users to state constraints on message creators. Creators on which a FR applies can be selected on the basis of several different criteria; one of the most relevant is by imposing conditions on their profile’s attributes. In such a way it is, for instance, possible to define rules applying only to young creators or to creators with a given religious/political view

ii) Blacklists

We make use of a BL mechanism to avoid messages from undesired creators. BL is managed directly by the system, which according to our strategy is able to:

(1) Detect who are the users to be inserted in the BL,

(2) Block all their messages, and

(3) Decide when user's retention in the BL is finished.

To make the system able to automatically perform these tasks, the BL mechanism has to be instructed with some rules, hereafter BL rules. In particular, these rules aim to specify (*a)* how the BL mechanism has to identify users to be banned and *(b)* for how long the banned users have to be retained in the BL, i.e., the retention time. Before going into the details of BL rules specification, it is important to note that according to our system design, these rules are not defined by the Social Network manager, which implies that these rules are not meant as general high level directives to be applied to the whole community. Rather, we decide to let the users themselves, i.e., the wall's owners to specify BL rules regulating who has to be banned from their walls. As such, the wall owner is able to clearly state how the system has to detect users to be banned and for how long the banned users have to be retained in the BL. Note that, according to this strategy, a user might be banned from a wall, by, at the same time, being able to post in other walls.

In defining the language of BL rule specification we have mainly considered the issue of how to identify users to be banned. We are aware that several strategies would be possible, which might deserve to be considered in our scenario. Among these, in this paper we have considered two main directions, postponing as future work a more exhaustive analysis of other possible strategies. In particular, our BL rules make the wall owner able to identify users to be blocked according to their profiles as well as their relationships. By means of this specification, wall owners are able to ban from their walls, for example, users they do not know directly (i.e., with which they have only indirect relationships), or users that are friend of a given person as they may have a bad opinion of this person. This banning can be adopted for an undetermined time period or for a specific time window. Moreover, banning criteria take in consideration also users' behaviour in the OSN. More precisely, among possible information denoting users' bad behaviour we have focused on two main measures. The first is related to the principle that if within a given time interval a user has been inserted into the BL for several times, say greater than a given threshold, he/she might deserve to stay in the BL for another while, as his/her behaviour is not improved. This principle works for those users that have been already inserted in the BL at least one time.

VII. CONCLUSIONS

We have presented system direct control to the user block unwanted messages on their social network wall. The systems using the machine language soft classifier to label the contents are Neutral and Nonneutral. And then applying the Filter Rule based on the creators. Moreover, the flexibility of the system in terms of filtering options is enhanced through the management of BLs.

REFERENCES

[1] Ali, B., Villegas, W., Maheswaran, M.: A trust based approach for protecting user data in social networks. In: Proceedings of the 2007 conference of the center for advanced studies on Collaborative research. pp. 288–293. ACM, New York, NY, USA (2007)

[2] Amati, G., Crestani, F.: Probabilistic learning for selective dissemination of information. Information Processing and Management 35(5), 633–654 (1999)

[3] Boykin, P.O., Roychowdhury, V.P.: Leveraging social networks to fight spam. IEEE Computer Magazine 38, 61–67 (2005)

[4] Carminati, B., Ferrari, E., Perego, A.: Enforcing access control in web based social networks. ACM Trans. Inf. Syst. Secur. 13(1), 1–38 (2009)

[5] Churcharoenkrung, N., Kim, Y.S., Kang, B.H.: Dynamic web content filtering based on user's knowledge. International Conference on Information Technology: Coding and Computing 1, 184–188 (2005)

[6] Fong, P.W.L., Anwar, M.M., Zhao, Z.: A privacy preservation model for facebook style social network systems. In: Proceedings of 14th European Symposium on Research in Computer Security (ESORICS). pp. 303–320 (2009)

[7] Golbeck, J.A.: Computing and Applying Trust in Web-based Social Networks. Ph.D. thesis, PhD thesis, Graduate School of the University of Maryland, College Park (2005)

[8] Hanani, U., Shapira, B., Shoval, P.: Information filtering: Overview of issues, research and systems. User Modeling and User-Adapted Interaction 11, 203–259 (2001)

[9] Kim, Y.H., Hahn, S.Y., Zhang, B.T.: Text filtering by boosting naïve bayes classifiers. In: SI- GIR '00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval. pp. 168–175. ACM, New York, NY, USA (2000)

[10] Pérez-Alcázar, J.d.J., Calderón-Benavides, M.L., González-Caro, C.N.: Towards an information filtering system in the web integrating collaborative and content based techniques. In: LA-WEB '03: Proceedings of the First Conference on Latin American Web Congress. p.222. IEEE Computer Society, Washington, DC, USA (2003)

[11] Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys 34(1), 1–47 (2002)

[12] Tootoonchian, A., Gollu, K.K., Saroiu, S., Ganjali, Y., Wolman, A.: Lockr: social access control for web 2.0. In: WOSP '08: Proceedings of the first workshop on Online social networks. pp. 43–48. ACM, New York, NY, USA (2008)