THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Machine Learning Approach to Filter Undesirable Messages from Online Social Networks

Vanaja U.
M.Tech (Computer Science and Engineering)
S.V. Engineering College for Women, Tirupati, India
M. Parthasaradhi
Assistant Professor, Computer Science and Engineering
S.V. Engineering College for Women, Tirupati, India

Abstract:
Now a day's one of the major problem in online social networks is filtering undesirable messages posted on user walls. OSN (Online Social Network) users have the ability to post and exchange information to their friends. But, theusers do not control the undesirable messages posted on their walls. In this paper we proposed the dominant approach to this problem is based on machine learning technique. It is ageneral inductive process automatically builds a classifier by learning from a set of preclassified documents and the characteristics of the categories.

Keywords: *Online Social Networks, Information filtering, text representation*

1. Introduction

Today social networks are part in modern life. Social network is an interaction among people in which they create, share or communicate and exchange several types of information as text, audio and video data. OSNs provide very little support to prevent undesirable messages posted on user walls. For example, Facebook established in 2004, became the largest social networking site in the world. It allow users to state who is allowed to insert messages in their walls i.e., the list present in friends, friends of friends, or defined group of friends. But no content based preferences are supported and therefore it is not possible to prevent unwanted messages, such as violence, sex or vulgar ones, no matter of the user who posts them. So, to control this type of activity and to prevent unwanted messages posted on user walls or private space, we can implement filtered wall(FW) able to filter unwanted messages posted on user wall's and Machine Learning (ML) text categorization techniques, it automatically assign each short text message to a set of categories based on its content.The system also provides a rule layer utilizing flexible language to specify Filtering Rules (FRs), by which users can state what content should not be displayed on their walls. FRs can support different filtering criteria according to the user needs. More specifically FRs utilizes user profiles, user relationships as well as the Machine Learning categorization process to state the filtering criteria to be required. In addition, the systemprovides the support for user-defined Black Lists (BLs), that is, lists of users that are temporarily prevented to post anykind of messages on a user wall.

2. Related Work

Information filtering systems are designed to examine a stream of dynamically examined documents and display only those are relevant to a user's interest, it must operate over relatively long scales and the ability to observe model adapt to their persistence, variation and interaction of interests are important. However our work has relationships with the Content-based filtering and Policy-based filtering. We, survey in both the fields.

2.1. Content-based filtering

Content based filtering system selects items based on the correlation between the content of the items and user's preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences.Content-based filtering is mainly based on the use of the ML paradigm according to which a classifier is automatically induced by learning from a set of preclassified examples. A content-based filteringselects information based on the content of the items and user preferences. The filtering can be modeled as label, binary form and partitioning the documents in to relevant and nonrelevant and automatically label the messages in to partial thematic categories.

2.2. Policy-based personalization content:

A classification method is proposed to categorize short text messages in order to avoid tremendous users of micro blogging services by raw data. For example, Twitter associates a set of categories with each tweet describing its content. The user can view only certain types of tweets based on their interests. In contrast,one application is proposedby Golbeck and Kuter called Film Trust exploits OSN trust relationships and provenance information to personalize access to the website. In such systems do not provide a filtering policy layer by which the user can utilize the result of the classification process to decide how and what to which extent filtering out unwanted information. Our filtering policy allows the setting of FRs according to a variety of criteria that do not consider only results of the classification process but also the relationships of the wall owner with other OSN users as well as user information on the user profile. Black List (BL) mechanism is another filtering procedure.

3. Filtered Wall Architecture

The architecture to filter unwanted messages consists of three layers. So, it is called three tier architecture. The first tier is Social Network Manager (SNM), second tier is Social Network Applications (SNA) and final tier is Graphical User Interface (GUI)

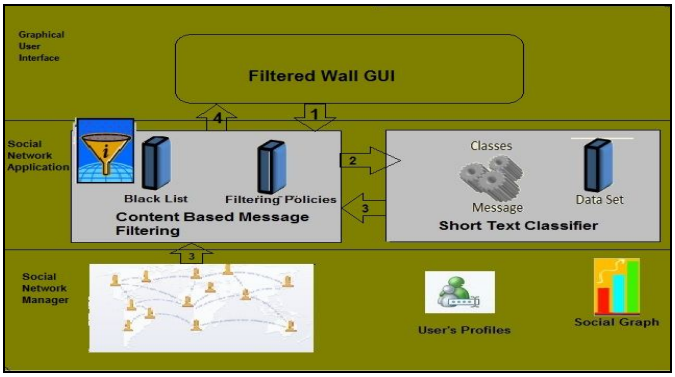

Fig 1: Filtered Wall Conceptual Architecture

3.1. 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).

3.2. 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 CBMF filters. Also Black list is maintained for the user who sends frequently bad words in message.

3.3. 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 Fig. 1, the path followed by a message flow is summarized as follows:

- The user tries to post an undesired message on user'swall, which is intercepted by FW.
- The metadata is extracted bya ML-based text classifier from the content of the message.
- FW uses metadata provided by the classifier, together with data extracted from theuser profiles and social graph, to enforce the filtering and BL rules.
- The message will be filtered by FW based on the result of the previous step,

4. Short Text Classifier

Theclassifier which is used in previous paper is used to classify the text which contain large amount of data, but it endure when the amount of document is little. To overcome this problem, short text classifier is used. Aim of the short text classifier is to recognize and eradicate the neutral sentences and categorize the non-neutral sentences in step by step, not in single step. This classifier will be used in hierarchical strategy. The first level task will be act as classified with neutral and non-neutral labels. The second level act as a non-neutral, it will develop gradual membership. These grades will be used as succeeding phases for filtering process. Short text classifier includes text representation, machine learning based classification.

4.1. Text Representation

Representing the text of a document is critical, which will affect the classification performance. Many features are there for representation of text, but we judge three types of features. BOW, Document properties (DP) and contextual features. BOW and Document properties are already used in twitter for information filtering, are endogenous that is, text which is entirely

derived from the information within the text message. Endogenous knowledge is well applicable in representation of text. It is genuine to use also exogenous knowledge in operational settings. Exogenous knowledge is termed as any source of information from outside the message but directly or indirectly communicate to the message itself. CF modeling is introduced, its feature is to understand the semantics of message. DP features are heuristically

- Correct words: It calculates as the percentage of correct words in the message.
- Bad words: They are computed similarly to thecorrect words feature, where the collections of "bad words" are represented as set K for the domain language.
- Capital words: It expresses the amount of wordsmostly written with capital letters, calculated as thepercentage of words within the message, havingmore than half of the characters in capital case based on character count. Therationale behind this choice lies in the fact that withthis definition we intend to characterize the willingnessof the author's message to use capital lettersexcluding accidental use or the use of correctgrammarrules. For example, "No" is not uppercasebecause the number of characters count should be strictly less than the capital characters.
- Punctuations characters: It is calculated as the percentageof the punctuation characters over the totalnumber of characters in the message. For example,the value of the feature for the document "Hai!!!What're u doing?" is 5=22.
- Exclamation marks: It is calculated as the percentageof exclamation marks over the total number ofpunctuation characters in the message. Referring tothe aforementioned document, the value is 3=5.
- Question marks: The question marks are calculated as number of punctuation marks are present in the message.

5. Machine Learning-Based Classification

A Short text categorization is a hierarchical two level ML based classification process. In the first-level text categorization technique performs a binary classification that labels message as Neutral and Non-neutral. By considering the first-level filtering task facilitates the subsequent second-level task in which finer-grained classification is performed. The second-level text categorization performs a partitioningof Non-neutral messages and those messages are assignedto non-neutral classes. The different types of multiclass ML models are well suited for text classification. We choose theRBFN model for the experimented competitive behavior with respect to other state-of-the-art classifiers. RBFNs have a single invisible layer of processing units with local, confined activation domain. A Gaussian function is commonly used for Radial Basis Functional Networks (RBFNs), but any other locally tunable function can be used. They were introduced as a neural network evolution of exact interpolation, and are demonstrated to have the universal approximation property. As outlined in, RBFN main advantages are that classification function is nonlinear and classification includes hard decision on the output values, the model may produce trust values and it may be robust to occupants; drawbacks are thepotential overtraining sensitivity, andpotential sensitivity to input parameters. Then the regular RBFN is structured by the first-level classifier. In second level of classification stage, we introduce modification of the standard use of RBFN. In regular use of classification finding the output values became hard decision. So, according to the winner-take-all rule, a given input pattern is assigned with the class corresponding to the best output neuron which has the highest value. In our approach, we consider result of the classification task as all values of the output neurons and we consider them as gradual estimation of multimembership to classes.

5.1. Filtering Rules

In OSNs everyday life, the same message may have different meanings and relevance based on who writes on it. Filtering rules provide constraints on message creators.

Rule1 (Creator Specification): A creator specification denotes a set of OSN users. It has one of the following forms

- A set of attribute constraints of the form*avOP an*, where *av* is a profile attribute value and OP is a comparison operator, *an* is a user profile attribute name, *OP* is compatible with *an*'s domain.
- A set of relationship constraints of the form(*rt,m, maxTrust,minDepth*) denoting with a relationship of type *rt*all OSN users participating with user *m*, and trust value less than or equal to *maxTrust*, having a depth greater than or equal to *minDepth*.
- Rule2: A Filtering Rule FR is a tuple (author, creatorSpec, contentSpec, action) where,
- The user who specifies the rule is author
- CreatorSpec is a creatorSpecification, according to Rule1.
- ContentSpec is a Boolean expression defined on content constraints of 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.
- Action belongs (block, notify) denotes the action to be performed by the system on the messages matching
- Any filtering rule can apply to same user, a message is published only if it is not blocked any of the filtering rules that apply to the message creator.

5.2. Blacklists

The Blacklist mechanism is to block messages from unwanted creators, independent from their contents. BLs are directly managed by the system. The flexibility of information given to the system through a set of rules called BL rules. The owners of the walls specify BL rules regulating who has to be banned from their walls and for how long. Similar to FRs, by using BL rules the owner of the wall can able to identify users to be blocked according to their profiles as well as they may have bad opinion of this person. This banning can be elected for an uncertain time period or for a particular time window andbanning criteria may also done based on user's behavior in the OSN. Based on user's badbehaviorwe have focused on two main measures. The first measure is related to that if within a given time interval a user has been inserted into a BL for several times, say greater than a

given threshold, he/she deserve to stay in the BL and another principle as his/her behavior is not improved. These measures work for the users who are already inserted in the considered BL at least one time. The another one is to catch new bad behaviors, we use the Relative Frequency (RF) to detect those users whose messages not filter in FRs.The measures computed either locally or globally, that is, by considering only the message or by considering all OSN user walls and BLs.
The BL rule is defined as follows:
Rule 3 (BL rule): A BL rule is a tuple (*author, creatorBehaviour, creatorSpec, T*), where

- The OSN user who specifies the rule is *author*
- *creatorBehaviour* consists of two components. They are *minBanned, RFBlocked*
- *creatorSpec is defined according to Rule1*.

6. Conclusion

In this paper, we describe our work to provide unwanted message filtering for social networks. We have presented a system to filter undesired messages from OSN walls. The system exploits a Machine Learning soft classifier to enhance customizable content-based FRs. Moreover, any type of message is filtered through FRs and management of BLs. we would like to remark that the system proposed in this paper represents just the core set of functionalities needed to provide a sophisticated tool for OSN message filtering. In particular, future plans survey an investigation on two interdependent tasks. The first concerns extraction and collection of contextual features and second task is learning phase i.e. the collection of preclassified data may not be represented in longer time.

7. References

1. 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.
2. 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.
3. 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.
4. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.
5. M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, "Content-Based Filtering in On-Line Social Networks," Proc. ECML/PKDD Workshop Privacy and Security Issues in Data Mining and Machine Learning (PSDML '10), 2010.
6. N.J. Belkin and W.B. Croft, "Information Filtering and Information Retrieval: Two Sides of the Same Coin?" Comm. ACM, vol. 35, no. 12, pp. 29-38, 1992.
7. P.J. Denning, "Electronic Junk," Comm. ACM, vol. 25, no. 3, pp. 163-165, 1982.
8. P.W. Foltz and S.T. Dumais, "Personalized Information Delivery: An Analysis of Information Filtering Methods," Comm. ACM, vol. 35, no. 12, pp. 51-60, 1992.
9. P.S. Jacobs and L.F. Rau, "Scisor: Extracting Information from On-Line News," Comm. ACM, vol. 33, no. 11, pp. 88-97, 1990.
10. S. Pollock, "A Rule-Based Message Filtering System," ACM Trans. Office Information Systems, vol. 6, no. 3, pp. 232-254, 1988.
11. P.E. Baclace, "Competitive Agents for Information Filtering," Comm. ACM, vol. 35, no. 12, p. 50, 1992.
12. P.J. Hayes, P.M. Andersen, I.B. Nirenburg, and L.M. Schmandt, "Tcs: A Shell for Content-Based Text Categorization," Proc. Sixth IEEE Conf. Artificial Intelligence Applications (CAIA '90), pp. 320-326, 1990.
13. G. Amati and F. Crestani, "Probabilistic Learning for Selective Dissemination of Information," Information Processing and Management, vol. 35, no. 5, pp. 633-654, 1999.
14. M.J. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," Machine Learning, vol. 27, no. 3, pp. 313-331, 1997.
15. Y. Zhang and J. Callan, "Maximum Likelihood Estimation for Filtering Thresholds," Proc. 24th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 294-302, 2001.
16. R.E. Schapire and Y. Singer, "Boostexter: A Boosting-Based System for Text Categorization," Machine Learning, vol. 39, nos. 2/3, pp. 135-168, 2000.
17. 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.
18. 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.
19. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short Text Classification in Twitter to Improve Information Filtering," Proc. 33rd Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '10), pp. 841-842,2010.
20. J. Moody and C. Darken, "Fast Learning in Networks of Locally- Tuned Processing Units," Neural Computation, vol. 1, no. 2, pp. 281-294, 1989