ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol.3 , No.5, Pages : 439- 442 (2014)
Special Issue of ICACSSE 2014 - Held on October 10, 2014 in St.Ann's College of Engineering & Technology, Chirala, Andhra Pradesh

An Efficient Filtering Technique over Undesirable Data of OSN User Walls

[1]**Rakesh Singampalli**, [2]**K.Vinaykumar**
[1]Final M.TechStudent ,[2]Assistant professor
[1,2]Dept of Computer Science and Engineering
[1,2]Pydah College of Engineering, Visakhapatnam, AP, India

Abstract: Prevention of unwanted and unauthorized data over Internet has become popular mechanism in online social networks .Authorization and control given to the users to prevent those spams either messages or images. In this paper we are proposing an efficient spam filtering mechanism with Naïve Bayesian classification by forwarding the testing samples over training samples of the dataset. Online social networks should be optimized with content based filtering. Wehave proposed an enhanced filtering mechanism by using a machine learning technique based on a content filtering.

INTRODUCTION

The personalization mechanism is a concept used in different contexts. Personalization limits toadapt a service in a specific situation and to perceive individualgoals by providing services to the users with a best product or service they really require and it can be used at its best. Later we used personalization for telecommunication services and assume it as the passageand a user can abstract incomingcommunication requests based on personal preferences. For best concernslet us assume two completely different samples of communication types andtheir personalization. By discussing about communication, we think people talking togetherbased on some communication device like phones on a telephone network or audio applications on web. Obviously we may wish to personalize the way of communication iswell handled [1][2].

Some other way the communication can be perceived much ingeneral as some discussion carried out in people orwith a computerized system. People can start avirtual office on a network of distributed systems and each participant may personalize the way his or her discussion is carried out [3]. The overall advantages are:

(1) It is able to use personal configuration of devices and register services and

(2) It is able to continue collaborative interactions independently from individual technical properties or registrations.

There is a rapid growth in development and usage of social-networking. Technology always has both advantages and disadvantages. Filtering of unauthorized content is a complex task because, identification may disturbs the genuine user even though they are not malicious, so we need an efficient spam filtering mechanism to identify spam in both content and images. Various cluster based mechanisms available to cluster the spam data but it cannot analyze behavior of the new sample which is not available in the current set and accuracy of the cluster based approaches are low than classification approach. In classification approach input message component isforwarded towards training dataset which consists of existing spam information of both messages and images [4].

Information filtering system removes duplicate information from information stream using automated methods to present data to end user. Its main aim is to maintain the information overburden and increasing semantic signal distortion rate [5]. To execute this, user profile is compared to reference features. These features may initiate from the information item or the user's social habitat. Whereas in data transmission signal filters are used to defend syntax distortion on the binary level and the methods engaged in information filtering act on the semantic level.

The phases of machine methods build on the similarrules as those are used for information abstraction process. An application can be found in the field of email spam filters. Therefore it is not only the information corruption and that is necessary for some types of filters and also maliciously releases the pseudoinformation [6].

There are some Recommender systems like active information filtering systems that are involved to present the user information, items such as film, television, music etc.The user is curious in thesesystems which are append the information items from the information passed towards the user and opposes by deleting information items from the information passed towards the user. The Recommender systems use filtering methods or a combination of the filtering and data-based filtering methods.

Filtering methods of this style include many tools that help people identify the most information. So within the limited time you can commit to read or listen or view and it is reallyproposed in the much interesting and valuable files aside from the most inconsequential. These filters are used to maintain and structure information in a real and understandable passage. In addition to set of messages on the mail addressed. These filters are necessary in the results of the searching browsers on the web. The methods of filtering isincreasing for getting web files and much efficiently [7][8].

RELATED WORK

Classification approaches analyzes the training samples efficiently by using existing data samples along with their decision classes, various traditional classification algorithms available to classify the input data sample, but every classification technique has their own merits and demerits and they are depend on context. Classification algorithms like SVM, KNN, ID3, C4.5, Bayes theorem and Naïve Bayesian classification.

A. Naive Bayesian classification

Considering different module classifier process of prioryprobability and class conditional probability,Naive Bayesian classification algorithm is one type of module classifier. It is very simpleprobability classifier and it is based on applying the much known Bayes theorem with strong (naive) independentbases/ assumptions. The main advantage of classifier isthat it requires only a few training data to consider the features such as means & variances of the variables which is necessary for classification. By evaluating and identifying the dependency among different attributes. The Naive Bayes is very simple for implementation and calculation. So it is used for pre-processing operation.

B. Support Vector Machines

The positive and negative training datasets are notcommon for another classification algorithms and it is required by SVM. These training datasets are needed for the SVM is to seekfor the decision space and these are separates the positive fromthe negative data in the 'n' dimensional space andthen the similar is called as the hyper plane. It supports vectors that are document representatives and which are similar to the decisionsurface. The goal of SVM is to identify the best classification function in order to differentiate between objects of two classes in the training dataset in a twoclasseslearning operation [10].

ARCHITECTURE

C. K-Nearest Neighbors (KNN)

It is one of the various classifiers and KNN classifier is a casebased learning algorithm which is based on a distance or similarity for various pairs of case study such as the Euclidean distance. It is tested for many applications because of its efficiency and non-parametric and it is easy to implementation features. Under this method the classification time is very long and it is difficult to find out the optimal result value of K. The best substitute of k to be chosen and depends upon the data. The effect of distortion on the classification is decreased by the long valuesof k but it makes limitations between classes few distinct. Byvarious heuristic techniques usage a best 'k' can be chosen. To achieve the above explained drawbackmodify previous KNN with various K values fordifferent classes other than fixed value for all classes [11][12].

PROPOSED WORK

In our proposed spam detection mechanism, we classify the testing samples of the data component and message sample with existing log of data samples,over experimental classification can be simple message or images, both can be filtered based on the meta information of the samples.

Effective spam detection in OSN, is achieved by the use Machine learning technique. Content based features can be mined from the messages. The machine learning categorization can be used to classify the messages based on its contents. People involving in online social network are interested in posting their views and ideas mostly in the form of text. And the users in OSN environment are communicating through short messages. Content based features are extracted from the short messages posted from the user walls.

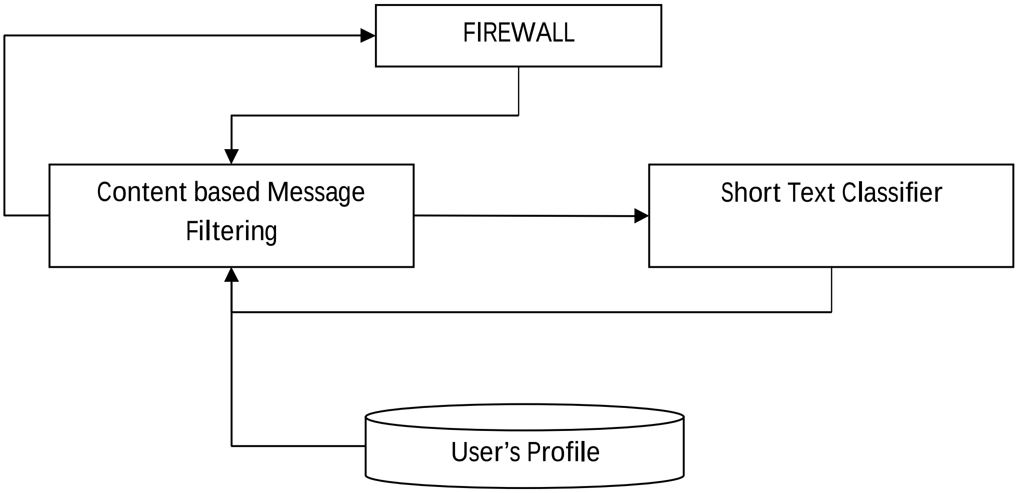

Fig 1:- Proposed Architecture

Our proposed architecture allows the spam filtering mechanism based on machine learning approach, to classify Meta data of the testing sample or message sample with existing training data samples,here we are proposing content based filtering mechanism initially it computes the prior probability follows by attribute based probability to compute posterior probability of the sample the following algorithm shows classification of Naïve Bayesian classification algorithm as follows

Naïve Bayesian Classification

Sample space: set of agent

H= Hypothesis that X is a node

P (H/X_i) is our confidence that X_i is an incoming node

P(H) is Prior Probability of H and it is probability that any given training sample is an agent regardless of its anomaly or not anomaly behavior

P(H/X) is a conditional probability and P(H) is independent of X

Estimating probabilities

P(H), P(X_i) and P(X_i/H) may be estimated from given training and testing data samples

$P(H|X_i)=P(X_i|H)*P(H)/P(X_i)$.

Steps Involved

1.Each training data sample is of attribute type

X= (x_j) j =1(1….n), where x_j is the values of X for attribute A_j

2.Suppose there are m decision classes C_j, j=1(1…m).

$P(C_i|X) > P(C_j|X)$ for 1<= j <= m, j>i

i.eclassifier assigns X to decision class C_j having highest posterior probability conditioned on testing sample X

The decision class for which $P(C_j|X)$ is maximum is knownas maximum posterior hypothesis of the sample.

From Bayes Theorem

3. P(X_i) is constant and Only need be maximized.

if class initial probabilities not known prior then we can assume all decision classes to be more equally likely decision classes

Otherwise maximize the samples

$P(C_i)$ = Si/S

4. Naïve assumptionfor attribute independence

$P(X|C_j) = P(x_1,…..,x_m|C) = PP(x_k|C)$

5. To classify an unknown testing sample X_i, compute each decision class C_i and Sample X is assigned to the class

iff ($Prob(X_i|C_i)P(C_i)> P(X_i|C_j) P(C_j)$).

Short Text Classifier:

Short Words	Long Words
Abt	About
Lemme	Let me
Tmrw	Tomorrow
Yup	Yes

Fig2: Short text classifier example dataset

The main task of the proposed system are the Content-Based Messages Filtering (CBMF) and Short Text Classifier .It additionally supports the classification of message based on the category set. Also it focuses on the following:

1.Filtered wall (FW) is used to intercept the message posted on the private walls of the user.

2.From the content of the message, Meta data are extracted based on the Machine learning (ML).

3.The extracted meta data are used by Filtered wall (FW) based on the classification and users' profile.

4.Based on the result obtained, the messages are filtered by Filtered wall (FW). Here we develop a relationship based filtering .To filter the posts automatically based on the relation.

The spamicity of a word

Recent statisticsshow that the current probability of any message being spam is 80%, at the very least:

$Pr(S)=0.8; Pr(H)=0.2$

However, most Bayesian spam detection software makes the assumption that there is no a priori reason for any incoming message to be spam rather than ham, and considers both cases to have equal probabilities of 50%:

$Pr(S)=0.5; Pr(H)=0.5$

The filters that use this hypothesis are said to be "not biased", meaning that they have no prejudice regarding the incoming email. This assumption permits simplifying the general formula to:

$$Pr(S|W)=\frac{Pr\ (W|S)}{Pr(W|S)+Pr\ (W|H)}$$

This is functionally equivalent to asking, "what percentage of occurrences of the word "replica" appear in spam messages?"

This quantity is called "spamicity" (or "spaminess") of the word "replica", and can be computed. The numberPr(W|S) used in this formula is approximated to the frequency of messages containing "replica" in the messages identified as spam during the learning phase. Similarly,Pr(W|H) is approximated to the frequency of messages containing "replica" in the messages identified as ham during the learning phase. For these approximations to make sense, the set of learned messages needs to be big and representative enough. It is also advisable that the learned set of messages conforms to the 50% hypothesis about repartition between spam and ham, i.e. that the datasets of spam and ham are of same size. Of course, determining whether a message is spam or ham based only on the presence of the word "replica" is error-prone, which is why Bayesian spam software tries to consider several words and combine their spam cities to determine a message's overall probability of being spam.

CONCLUSION

We are concluding our research work with efficient spam detection system with content based filtering mechanism in a machine learning approach., testing samples of the message content or image data sample can be forwarded to training data set and compute posterior probability and spamicity of a word in case of textual information. Our experimental results show more accurate results than the traditional approaches.

REFERENCES

[1] A. Adomavicius, G.andTuzhilin, "Toward the next generation ofrecommender systems: A survey of the state-of-the-art and possibleextensions," IEEE Transaction on Knowledge and Data Engineering,vol. 17, no. 6, pp. 734–749, 2005.

[2] M. Chau and H. Chen, "A machine learning approach to web pagefiltering using content and structure analysis," Decision SupportSystems, vol. 44, no. 2, pp. 482–494, 2008.

[3] R.J. Mooney and L. Roy,"Content-based book recommending usinglearning for text categorization," in Proceedings of the Fifth ACMConference on Digital Libraries. New York: ACM Press, 2000, pp. 195–204.

[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," in Proceedings of ECML/PKDD Workshop on Privacy and Security issues in DataMining and Machine Learning (PSDML 2010), 2010.

[6] N. J. Belkin and W. B. Croft, "Information filtering and informationretrieval: Two sides of the same coin?" Communications of the ACM, vol. 35, no. 12, pp. 29–38, 1992.

[7] P. J. Denning, "Electronic junk," Communications of the 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," Communications ofthe ACM, vol. 35, no. 12, pp. 51–60, 1992.

[9] P. S. Jacobs and L. F. Rau, "Scisor: Extracting information from onlinenews," Communications of the ACM, vol. 33, no. 11, pp. 88–97, 1990.

[10] S. Pollock, "A rule-based message filtering system," ACM Transactionson Office Information Systems, vol. 6, no. 3, pp. 232–254, 1988.

[11] P. E. Baclace, "Competitive agents for information filtering," Communicationsof the 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," in Proceedings of6th IEEE Conference on Artificial Intelligence Applications (CAIA-90). IEEE Computer Society Press, Los Alamitos, US, 1990, pp. 320–326.