ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(33), DOI: 10.17485/ijst/2015/v8i33/80128, December 2015* ISSN (Online) : 0974-5645

A Content Filtering Scheme in Social Sites

J. Adamkani[1*] and K. Nirmala[2]

[1]Quaid-E-MillathGovt College for Women, Chennai - 600002, Tamil Nadu, India; adam_kani@rediffmail.com
[2]Department of Computer Science, Quaid-E-MillathGovt College for Women, Chennai - 600002, Tamil Nadu, India; nimimca@yahoo.com

Abstract

Objectives: In recent scenario, online social networks such as Face book, Twitter and Google+ have become one of the fastest emerging e-services. There are several issues affected these e-services. Since it is emerging service and reliability to communicate, in social networks privacy is often a key concern by the users. Since millions of people are willing to interact with others, it is also a new attack ground for malware creators. Some users and pages spreading malicious content and sending spam messages by taking advantage on the users' inherent trust in their relationship network**. Methods**:This proposed work handles the most prevalent issues and threats targeting different social networks recently. And finally finds the authentication scheme for those attacks. This proposes a detecting and blocking scheme for social sites using data mining techniques. **Findings**: This system helps to detect suspicious URLs for social network by considering the following parameters, i).Text and keywords appears in the URL. ii). URL descriptions iii). Detection of scam messages which is done in manual script attacks on social sites. **Application/Improvement**: This performs two techniques which are message filtering and MLE (Maximum Likelihood Estimation).

Keywords: Feature Selection, Machine Learning, Security, Social Network

1. Introduction

Social network is a system of social connections, individual relationship and committed website or other application which empowers clients to speak with one another by posting data, remarks, messages, pictures, and so forth. Nowadays, business organizations are utilizing informal communities exceedingly[1]. There are such kinds of approaches to utilize a long range informal communication in a business situation, thus numerous advantages of doing, for example, Inexpensive Marketing, Banner and Text Ad Advertising, Customer Relation Management Tool, Global Exposure, Online Meeting Places[2]. At the point when considering, "use of social network in society or organization", that can be being used for some things, great and awful. A considerable lot of the previously stated ways that they can be utilized are gainful to organizations[3].They can likewise do a great deal more like to enhance the client pictures of the business, to get input on new items and administrations, to unite companions, family and departed school companions, for thought sharing and the making dialogue Online interpersonal organizations are gainful from various perspectives. They uproot a considerable lot of the complexity of the logged off world. Furthermore, they are frequently an extremely fun leisure activity and get-together information from social networks[4].BelowFigure 1 shows[5,6] a sample Social networking

Structure. It is the diagram of email conversations within a company. The colors represent people (in the network it is known as nodes) and the lines between the nodes represent the relationship between the people in a company. Clusters of nodes represent people reporting to important people.

A group (frequently termed as a community, e-group or club) is a component in numerous social network services which permits clients to create, post comment to and read from their own interest and specialty particular discussions, regularly inside of the domain of virtual groups[7]. The propensity of individuals to meet up and

* *Author for correspondence*

form groups is inherent in the structure of society; and the courses/topics in which such group comes to fruition and advance after some time is a topic that goes through expansive parts of sociological examination. Though the sociogram can give a general feeling of the system initially, analysts have added to a mixture of measurements for evaluating essential contrasts in system structure.

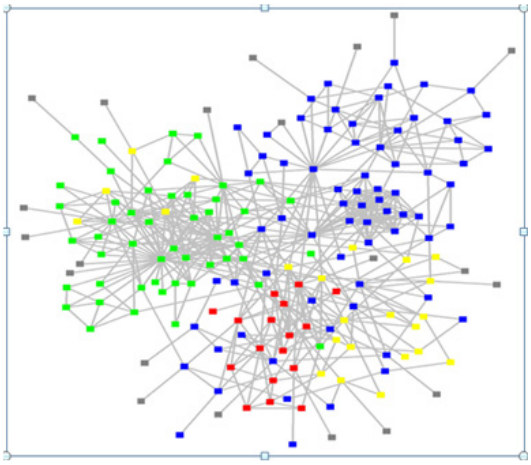

Figure 1. Social networking structure.

Habitually utilized measurements incorporate on-screen character degree centrality (the degree to which performers send or get direct ties), between centrality (the degree to which on-screen characters have ties with other people who are not specifically joined), closeness centrality (the degree to which on-screen characters are straightforwardly or in a roundabout way associated with whatever remains of the on-screen characters in the system), correspondence (the degree to which there are common ties in the middle of on-screen characters), and transitivity (the degree to which on-screen characters who are associated with each other are additionally joined with the same different on-screen characters)[8].

In social networks security is often a key concern by the users. Since a huge number of individuals are willing to communicate with others, it is additionally another assault for malware creators. They are taking so as to spread malicious code and sending spam messages by advantage of the clients' inalienable trust in their relationship system. Security change in informal organization locales (SNS) is a testing errand, where the social destinations experience the ill effects of a few sorts of assaults, for example, Identity theft, stalking, Unintentional fame, unauthorized access and so on[9,10].

How to solve this issue using data mining? : scam messages are not directly posted, instead along with the comments and message it will be included. So, finding those scam ,malicious content and URL by using data mining is the major task. The proposed method uses data mining techniques, which finds the malicious contents and manual script attacks and performs authentication scheme for those attacks[2-5].A significant number of the data mining algorithms[11] used to recognize spam and patterns of abuse on SNSs are composed with the assumption that the information and the classifier are free. Be that as it may, on account of spam, extortion and different malicious content, users will regularly alter their behavior to evade detection, prompting degraded classifier execution and the need to re-train classifiers much of the time[14]. Naive Bayes classifier to recognize and reclassify information considering the ideal optimal technique that an adversary could choose reverse engineering framework uses a classifier to determine whether an adversary can efficiently learn enough about a classifier to effectively defeat it[12].

For example, training data is converted to a series of feature that consist of a set of values for attributes. These features construct the input of supervised machine learning algorithm. After training, a classification model is applied to distinguish whether the specific user belongs to normal user or spammer. Because spammers and non-spammers have different social behavior, through analyzing content feature and user behavior, it is capable to distinguish abnormal behavior from legitimate ones.

Many data mining and machine learning researchers have worked on spam detection and filtering, commonly treating it as a basic text classification problem[15-18]. The

Table 1. Comparative study of existing data extraction tools with FFS

Techniques	Spam message detection	Scam message filtering	Score range analysis	Server side filtering	Active learning	Client side filtering
FFS_MLE	available	possible	possible	possible	possible	possible
TKNN	available	possible	possible	possible	X	X
KNN	available	X	X	X	X	X

problem is popular enough that it has been the subject of a Data Mining Cup contest as well as numerous class projects. Also combination of methods can be used, like combination of Naïve Bayes and memory based classifiers [19,20]. Bayesian analysis has been very popular, but researchers have also used SVMs, decision trees, memory and case-based reasoning, rule learning, Artificial Neural Networks , K-Nearest Neighbor, Instance-based learning, Statistical Data Compression Models, Latent semantic indexing, [21,22] memory based classifiers and even genetic programming. Other methods such as Markov Model have been used in character-level as opposed[23] to common word-level methods[24]. But all these evolutions have been on static datasets. A rule-base model could be designed to define a structure for spams and hams[25,26]. These rules should be updated by a human expert; this is a time-consuming and inaccurate method. Bayesian Classifier, Naïve Bayes and K-Nearest Neighbor (KNN)[27,28] were the mostly used. Naïve Bayes is most commonly used[29]. approaches by spam filters as a static text classification[30-32]. Lazy learning techniques like case-based reasoning have been used for such dynamically changing contexts.

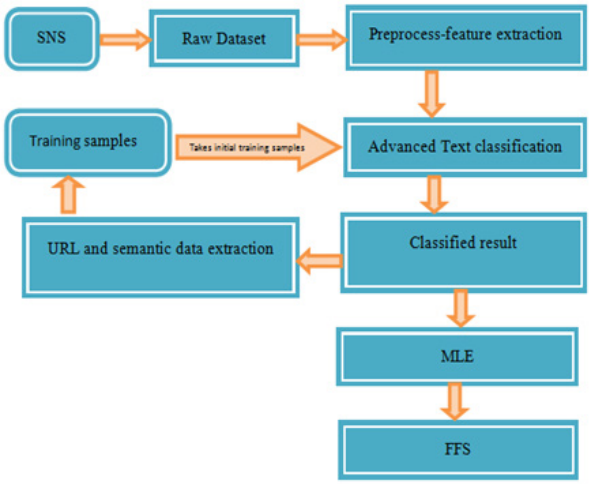

Figure 2. Overall process of scam detection in SNS.

2. Problem Defination

Security improvement in SNS is a challenging task, where the social site experience the ill effects of a few sorts of assaults. The proposed framework handles the accompanying issues. scam messages, which will do by the manual script assaults over web, SNS even on portable. Normally scammers send an instant message that seems to come straightforwardly from the authorized service provide. In the event that the client taps the connection or message which has been sent by the attackers, then the user system will be adulterated. Beneficiaries of this message ought not to tap the connection or content, which unites with a site that could contaminate the client's gadget with malware, scramble your information or introduce spyware. The user additionally shouldn't answer or reacts to those messages. In existing, the trick sifting plans have not completely coordinated with the SNS. Some centralized authorities will gather and verify the scam messages and they will answer to the client.

3. Proposed Framework

Due to the tremendous technological developments, security and privacy issues are become very tough to handle. For example, the popular online social networks such as facebook, twitter, google + and n number of sites gives protection on user data with respective privacy policies and settings. But still the malwares, irrelevant content posting and malicious content spreading problems are not fully restricted. With the help of data mining techniques, the proposed work handles the above problem in the social sites. Our system aims to provide security analysis against the following issues in SNS".

- Malicious data insertion via SNS.
- Manual script attacks in SNS

Malicious and unwanted data management in SNS is very tedious due to its abnormal data size, because every day millions of users uses the above mentioned popular social network sites. Verification and filtering malicious data from the huge data needs more computation time. And the finding of manual script attacks also similar to the above point. So the proposed work introduces a new filtering technique named as FFS (Filter-Forward-synchronize), which performs an effective filtering and data and Machine learning approaches to perform the effective filtering process below figure 2 shows that overall process of scam detection in SNS.

4. Methodology

This work focuses on the problem of automatically extracting and summarization of data records that are encoded in the query result pages generated by web databases.

4.1 Data extraction

Here, the data description extraction about the URL process is given.Scam URL content Identification initially identifies with all possible data description, which included in the training samples. If the data is not appeared in the training samples, the extraction process begins and it extracts necessary keyword and its descriptions. It also identifies Scam URL content with similar data records in the web pages. This step mines every scam URL data in the training samples that contains similar data records. Instead of mining data records directly, the system first tries to find in the test samples.

4.2 Pre processing and Feature Extraction

Before classification and filtering, pre-processing steps should be applied on message in the SNS. The following steps are:

4.3 Content Extraction

When a new message received in the SNS, the textual content will be extracted and stored as the dataset, for every data from the dataset, the content separation has been done. The message contains several contents, so initially we need to eliminate some irrelevant contents such as the header, footer and images if any. And the next process is finding and eliminating unwanted and malicious content from the message by matching the blacklist, which has been included in the training samples. When the sender ID or some group is found in a classifier list, it is simply classified as malicious.

4.4 Removing Unwanted HTML Tags

In some cases, the data can't be easily classified because; the scam messages may change regularly. In certain situations, the proposed system finds the scam message and URL by extracting its content and verifying it before delivering it to the receiver.

This step identifies tags, which helps to identify the URL and paragraphs in the web content. It places tags in classes depending on their semantic annotations.

4.5 Replacing all Sequences of Whitespace Characters (Tabs, Spaces and Newline Characters) by a Single Space

This process has been done using simple regular expression concepts. A simple pattern matching functionality can effectively identify these types of characters.To extract the texts and replacing the spaces, newline we use the following,

In order to obtain all words that are used in a given input with eliminating tabs and other keywords, this replacing process is required, i.e. a message will split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of all different words have obtained by merging all messages of a dataset.

4.6 Eliminating "Stop Words"

After concatenating the words, stop word elimination process will begin. Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text contents are prepositions, articles and pro-nouns, etc. that does not give the meaning of the documents. These words are treated as stop words. Example for stop words: the, in, a, an, with, etc. Stop words are removed from documents because those words are not measured as keywords in text mining applications.

4.7 Stemming

This method is used to identify the root/stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word "connect" The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space. the proposed framework used porter stemmer algorithm.

4.8 Porter Stemmer Algorithm

The preprocessing process includes the stemming process, which eliminates unnecessary keys. All stemming algorithms can be roughly classified as affix removing, statistical and mixed. Affix removal stemmers apply set of transformation rules to each word, trying to cut off known prefixes or suffixes.

Porter stemmer utilizes suffix stripping techniques rather than prefix methods. The porter stemmer Algorithm dates from 1980.

Step 1: Gets rid of plurals and -ed or -ing suffixes

Step 2: Turns terminal y to i when there is another vowel in the stem

Step 3: Maps double suffixes to single ones:-ization, -ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final –e

The above steps represent the process and elimination of porter stemmer algorithm.. The importance of the stemmer algorithm is, it reduces the difficulties of data classification when the training data's are insufficient. This effectively eliminates the suffix words such as 'ed', 'ing' etc.,

4.9 Frequency Counting using TFIDF

Term Frequency–Inverse Document Frequency Tf–IDF is a numerical statistic which reveals that a word is how important to a document in a collection. The Tf–IDF is often used as a weighting factor in information retrieval and text mining. The value of Tf–IDF increases proportionally to the number of times a word appears in the document, but is counteracting by the frequency of the word in the corpus. This can help to control the fact that some words are generally more common than others. Tf–IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification. Tf–IDF is the product of two statistics which are termed frequency and inverse document frequency. To further distinguish them, the number of times each term occurs in each document is counted and sums them all together.

Table 2. Performance analysis

Scam messages	Datasets			
	Email Based		Face Book	
	TKNN	FFS	TKNN	FFS
Filtered Scam message	500	330	350	340
Correctly Filtred Scam messages	480	325	330	335
Record Level Precision (%)	96	98.4	94.2	98.5
Record Level Recall (%)	90.2	95.4	96.3	99.7
Page Level Precision (%)	80.5	91	93.8	95.1

4.10 Scam Filtering and Data Synchronizing

The scam issues, which are malicious activity or policy violations on the user's data. Then there should be an attempt to stop an filter attempt. This module is for avoiding scam by analyzing the data. The person of a company will not be allowed to send any message to the restricted user without the administrator permission. If the person receives review with more than a threshold the data, which contains scam will be blocked.

When filtering process begins, the system detects the scam by using the following attributes.

A suspicious email address

Generic salutations rather than using a name.

Alarmist messages or urgent requests to download or install something. The scammer is trying to create a sense of urgency so you'll respond without thinking.

Grammatical errors and misspellings, which are used to break through phishing filters.

Requests to verify or update your account, stop payment on a charge, and the like.

4.11URL and Semantic Data Extraction

The FFS performs data extraction process, when there are no enough training samples available on the dataset. Usually, detection of scam messages is not always successful with training samples, because the scam messages will be vary regularly.

4.12 Maximum Likelihood Estimation Model

After successful preprocessing, the scam messages will be filtered according to the pattern matching concept, but in scam filtering on SNS, there is a need of huge training samples. The learning is based on the principle of maximum likelihood estimation, this also based on some refinement of this principle such as maximum a posteriori probability (MAP). This finds conditional word occurrence probabilities for scam message classification.

MLE works on labeled training data. A learning algorithm takes the data produces a classifier, which is a function that takes in new unlabeled test samples and outputs predictions about the scam message labels of those test samples based on the patterns in the training scam data. The MLE find useful patterns in the data: for example, in the context of message classification and filtering they can find a useful partition of documents in naturally occurring classes.

The framework estimates the parameters of the model by using a standard method that is MLE (maximum Likelihood estimation). Consider a set of m points that are

generated from a one dimensional Gaussian distribution model. Consider that the points are generated separately; the probability of these points is just the product of their individual probabilities. The following is the calculation used for probability calculation.

Table 3. Comparison table

Metrics	KNN	TKNN	Proposed FFS
Filtering Accuracy	89	90	94
Detection time	73.71	70.68	49.69
Classification Delay	3702.09	2984.06	2108.08
Number of iterations	64.52	57.81	48.21

Since the probability would be very small in size, the proposed framework based MLE typically work with the log probability. While μ and μ are unknown, it can choose the parameters for which the data is most probable. i.e, choose the μ and μ that maximize the above equation (1). This approach is known as Maximum Likelihood principle and the process of applying this principle to estimate the parameters of a statistical distribution from the data is known as MLE (Maximum likelihood estimation). The values of the parameters for which the log likelihood derived from the probability of equation (1) is maximum, are the parameter values for the underlying Gaussian Distribution.

A: fix a multinomial distribution with parameter vector φ of length V

B: for each word in the document draw a word w according to

Above, step A sets up the probability distributions that are then used in step B to produce the observed training data. The generative process above gives the following global probability distributionGiven:

- A data collection D (e.g., a set of messages from SNS).
- A pair wise similarity measure m defined over D (e.g., word co-occurrence probabilities derived from training set).
- A probability score L for describing elements of D
- A feedback in L of each object in D. Output all subsets c of D such thatc is highly cohesive with respect to m (e.g., the average pair wise similarity of objects in c exceeds some threshold).

c corresponds to a concept describable in L.

Probability distributions that are then used in step B to produce the observed training data.

Using the above, the scam related message will be find and eliminated form SNS.

5. Performance Analysis

The performance of the data extraction methods is compared in three different ways. General data set evaluation presents the performance on the first three data sets, which exhibit a variety of properties and have been used in previous work by others. The other two evaluations focus on specific properties of the data filtering techniques. FSS compares the performance for effective scam filtering on SNS.

This experiment tests on two data sets email scams and Face book, twitter scam messages and email based dataset contains 50-60 messages. The website tests on 3 domains mentioned above. More than 5 results of these messages that contain scam message contents. For each SNS, 10 messages are collected manually.

Two common measures, Recall and Precision, to evaluate the performance of this approach. Recall is the percentage of the number of scam messages that have been correctly filtered over the total number of data records on a SNS. Precision is the percentage of the number of scam messages that have been correctly filtered over the total number of data records that have been extracted.

Pr = Cc / Ce

Rr = Cc / Cr

Where, Cc is the count of correctly extracted scam messages and total messages ,

Ce is the count of filtered Scam messages , and

Cr is the actual count of Scam messages in the training samples.

The number of Scam messages in different SNS page varies from a few to hundreds. Consequently, pages with many Scam messages will dominate the record level metrics. To use a page-level metric, namely page- level precision defined as,

Pp = Cp / Na

Where, cp is the count of correctly filtered messages, which means that all the Scam messages in the pages are correctly filtered and summarized to the user, Na is the count of all the pages from which Scam messages are filtered. To assume that each input page contains at least two Scam messages and data extraction is performed on all input pages.

As per theoretical comparison and proof from the current experiment setup, the comparison study has developed. The proposed FFS shows better results, as a well known data record extraction system.

As per Table 2, this approach has much better experimental results than existing approach TKNN, and in almost every domain this approach significantly outperforms TKNN. The precision and recall of this approach are both high across all datasets, approaching 100%. This approach can also extract HTML related content for the appropriate URL.

To evaluate the performance of the proposed schemes, execution time and storage are the main measurement of performance evaluation. Without loss of generality, this defines processing delay and classification delay for scam message classification. Processing delay indicates the execution time for classification to produce frequent items and corresponding interest before page load. Scam message detection delay is also evaluated by measuring time spent on processing time on classification frequent items and interest in the proposed schemes. Another criterion is cost evaluation. Cost evaluation involves storage and computation aspects.

The performance of this proposed work FFS using MLE Scheme is compared with two existing approaches KNN and TKNN . The table 3 shows the performance comparison of the proposed method with other existing approaches based on the four different metrics classification delay, time, processing delay, number of iterations.

6. Conclusion

The paper provides a novel spam and scam detection approaches on SNS datasets. In the literature, the authors were only concentrated on the spam filtering, the proposed system finds the unwanted and malicious messages and URL and filters before delivering the data using MLE. New approach named as FFS has been proposed to match the co-occurrences of scam messages, which able to find dynamic scam messages changes but also simple, accurate and fast. Experiments show this new FFS have better results than TKNN. The result is also useful for other types of datasets which may have fraudulent intents, like securities and SNS frauds.

7. References

1. Trusov M, Bucklin RE, Pauwels K. Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. Journal of Marketing. 2009; 73(5):90–102.
2. Jensen D, Neville J. Data mining in social networks. na.2003.
3. Zuber M. A Survey of data mining techniques for social network analysis. International Journal of Research in Computer Engineering and Electronics. 2014; 3(6).
4. Aggarwal C. An introduction to social network data analytics. Springer US, 2011.
5. Asur S, Huberman B. Predicting the future with social network. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT) 2010IEEE. 2010; 1.
6. Au Yeung CM, Iwata T. Strength of social influence in trust networks in product review sites. Proceedings of the fourth ACM International Conference on Web Search and Data Mining, ACM; 2011. p. 495–504.
7. Bekkerman R, McCallum A. Disambiguating web appearances of people in a social network. In Proceedings of the 14th International Conference on World Wide Web, ACM; 2005. p. 463–70.
8. Castellanos M, Dayal M, Hsu M, Ghosh R, Dekhil M. U LCI: A social channel analysis platform for live customer intelligence. Proceedings of the 2011 International Conference on Management of Data. 2011.
9. Chelmis C, Prasanna VK. Social networking analysis: A state of the art and the effect of semantics. Privacy, security, risk and trust (passat), 2011 IEEEThird International Conference on and 2011 IEEEThird International Conference on Social Computing (socialcom). IEEE, 2011.
10. Barabasi A-L, Jeong H, Neda Z, Ravasz E, SchubertA, Vicsek T. Evolution of the Social Network of Scientific Collaborations. Physica A: Statistical Mechanics and its Applications. 2002; 311(3):590–614.
11. Yang B, Cheung W, Liu J. Community Mining from Signed Social Networks, Knowledge and Data Engineering, IEEE Transactions. 2007; 19(10):1333–48.
12. Bonchi F, Castillo C, Gionis A, Jaimes A. JaimesA. Social Network Analysis and Mining for Business Applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):22.
13. Chen J, Fagnan J, Goebel R Rabbany R, Sangi F, Takaffoli M, Verbeek E, Zaiane O. Meerkat: Community Mining with Dynamic Social Networks. Data Mining Workshops (ICDMW), IEEE International Conference. 2010; p. 1377–80
14. Gundecha P, Liu H. Mining Social Media: A Brief Introduction. Tutorials in Operations Research. 2012; 1(4).
15. FawcettT. In vivo' spam filtering: a challenge problem for data mining. KDD Explorations. 2003 Dec; 5(2).
16. Airoldi E, Malin B. Data mining challenges for electronic safety: The case of fraudulent intent detection in e-mails.Proceedings of the Privacy and Security Aspects of Data Mining Workshop, in conjunction with the 4th IEEE Internation Conference on Data Mining. Brighton, England, 2004 Nov. p. 57–66.
17. Tretyakov K. Machine learning techniques in spam filtering. Institute of Computer Science, University of Tartu Data Mining Problem-oriented Seminar, MTAT, 2004; 3:60–79.

18. Bratko A, Filipic B. Spam filtering using character-level markov models: Experiments for the TREC 2005 spam track. Text Retrieval Conference, 2005.
19. Cournane A, Hunt R. An analysis of the tools used for the generation and prevention of spam. Computers and Security. 2004; 23(2):154–66.
20. Bratko A, Cormack GV, Filipic B, Lynam TR, Zupan B. Spam filtering using statistical data compression models. Journal of Machine Learning Research 7. 2006Dec; 2699–720.
21. Androutsopoulos I, KoutsiasJ, KonstantinosV, Chandrinos V, Paliouras G, Spyropoulos C. An evaluation of naive Bayesian anti-spam filtering.In: PotamiasG, MoustakisV, van SomerenM,editors. Proceedings of the ECML 2000 Workshop on Machine Learning in the New Information Age. 2000; 9–17.
22. Leonard D. E-mail threats increase sharply. IDG News Service. 2002 Dec12.
23. Androutsopoulos I, KoutsiasJ, PaliourasG, KarkaletsisV, SakkisG, SpyropoulosC, StamatopoulosP.Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. 4th PKDD workshop on machine learning and textual information access. 2000.
24. Androutsopoulos I, PaliourasG, MichelakisE. Learning to filter unsolicited commercial email. Tech rpt 2004/2, NCSR Demokritos, 2004.
25. DruckerHD, WuD, VapnikV.Support vector machines for spam categorization. IEEE Transactions On Neural Networks. 1999; 10(5): 1048–54.
26. GeeKR.Using latent semantic indexing to filter spam. Proceedings of the 2003 ACM Symposium on Applied Computing (SAC), ACM. 2003; 460–4.
27. Delany SJ, Cunningham P, Tsymbal A, Coyle L. A case-based technique for tracking concept drift in spam filtering. Knowledge-Based Systems. 2005; 18(4-5):187–95,
28. PantelP, LinD. SpamCop: A spam classification and organization program. Learning for Text Categorization—Papers from the AAAI Workshop, Madison Wisconsin.(AAAI Technical Report WS-98-05). 1998; 95–98.
29. Sahami M, Dumais S, Heckerman D, Horvitz E. A bayesian approach to filtering junk email. AAAI-98 Workshop on Learning for Text Categorization. Madison, Wisconsin. (AAAI Technical Report WS-98-05). 1998; 55–62,
30. Carpinter J, Hunt R. Tightening the net: A review of current and next generation spam filtering tools. Computers and Security. 2006; 25(8):566–78
31. ProvostF, FawcettT,KohaviR. The case against accuracy estimation for comparing induction algorithms. In:ShavlikJ, editor. Proceedings of ICML- 98, San Francisco, CA. Morgan Kaufmann.1998; 445–3
32. Kolcz A, AlspectorJ.SVM-based filtering of e-mail spam with content-specific misclassification costs. Proceedings of TextDM'2001, IEEE ICDM-2001 Workshop on Text Mining, San Jose CA, 2001.
33. Khongbantabam SD. A New Feature Selection Algorithm for efficient spam filtering using Adaboost and Hashing Technique. Indian Journal of Science and Technology. 2015; 8(13):65753.