International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 5, ISSN No. 2455-2143, Pages 43-47
Published Online March – April 2016 in IJEAST (http://www.ijeast.com)

A SURVEY ON VARIOUS CLASSIFIERS DETECTING GRATUITOUS EMAIL SPAMMING

Garima Jain

***Abstract*— Email becomes the major source of communication these days. Most humans on the earth use email for their personal or professional use. Email is an effective, faster and cheaper way of communication. The importance and usage for the email is growing day by day. It provides a way to easily transfer information globally with the help of internet. Due to it the email spamming is increasing day by day. According to the investigation, it is reported that a user receives more spam or irrelevant mails than ham or relevant mails. Spam is an unwanted, junk, unsolicited bulk message which is used to spreading virus, Trojans, malicious code, advertisement or to gain profit on negligible cost. Spam is a major problem that attacks the existence of electronic mails. So, it is very important to distinguish ham emails from spam emails, many methods have been proposed for classification of email as spam or ham emails. Spam filters are the programs which detect unwanted, unsolicited, junk emails such as spam emails, and prevent them to getting to the users inbox. The filter classification techniques are categorized into two either based on machine learning technique or based on non-machine learning techniques. Machine learning techniques, such as Naïve Bayes, Support Vector Machine, Adaboost, and decision tree etc. whereas non- machine learning techniques, such as black/white list, signatures, mail header checking etc. in this paper we review these techniques for classifying emails into spam or ham**

Keywords— *Ham, Spam, Email Spamming, Spam Filter, Email Spam*

I. INTRODUCTION

Data mining is the process of mining or extracting knowledge from large databases. Data mining is also known as "Knowledge Discovery Process" or "Knowledge mining". There are many other terms which define data mining such as knowledge extraction, knowledge mining from large amounts of data, data analysis. Data mining is applicable on various kinds of data repositories such as data warehouses, relational databases, transactional databases, data streams, flat files and World Wide Web. Data mining is an essential step in the process of discovery of relevant knowledge.

The process of knowledge discovery or knowledge extraction is an iterative process [1] and it contains the following steps:

1. Data cleaning involves cleaning of noisy data. It removes the noise and inconsistent, irrelevant data from databases.
2. Data integration where data from different multiple data sources are combined together and collected in one data store.
3. Data selection where the data which is relevant to the task under analysis are selected and retrieved from the database.
4. Data transformation where the data are transformed into the forms appropriate for the mining process by performing the summary or aggregation operations.
5. Data mining, the process where methods are applied to mine or extract important data patterns.
6. Pattern evaluation considers identifying the interesting patterns which represent knowledge based on the interestingness measures.
7. Knowledge presentation where the knowledge representation methods are used to present the mined and extracted knowledge to the database user.

Steps 1 to 4 are the steps which are used for pre-processing the data, where the data is processed prior to the mining so that and inconsistency, irrelevant or noisy data is removed from the database. This pre-processed data is passed to the data mining algorithms and techniques which produces an output in some forms of patterns. Data mining step interact with the user or a knowledge base. The patterns which are interesting and true are presented to the database user and can be stored as the new knowledge in the knowledge base. Data mining is the essential and most important step in knowledge discovery process because it mines the hidden patterns from the database which is important for the data evaluation and various data analysis tasks.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 5, ISSN No. 2455-2143, Pages 43-47
Published Online March – April 2016 in IJEAST (http://www.ijeast.com)

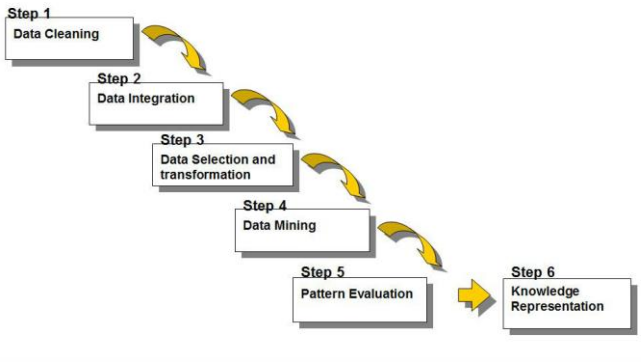

Fig 1. Knowledge Discovery

Email becomes the major source of communication these days. Most humans on the earth use email for their personal or professional use. Email is an effective, faster and cheaper way of communication. It is expected that the total number of worldwide email accounts is increased from 3.3 billion email accounts in 2012 to over 4.3 billion by the end of year 2016[email statistic report 2012] . Now days, almost every second user in the earth has an email account. The importance and usage for the email is growing day by day. It provides a way to easily transfer information globally with the help of internet.

Spam is an unwanted, junk, unsolicited bulk message which is used to spreading virus, Trojans, malicious code, advertisement or to gain profit on negligible cost. Spams are of many types based on the way of transmission i.e. email spam, social networking spam, web spam, blog or review platform spam, instant message spam, text message spam and comment spam. Spam message can contain text, image, video and also voice data. Spam can be sent via web, fax, telephonic, sms (text messages).

The email spamming is increasing day by day because of effective, fast and cheap way of exchanging information with each other. According to the investigation, it is reported that a user receives more spam or irrelevant mails than ham or relevant mails. About 120 billion of spam mails are sent per day and the cost of sending is approximately zero. According to a spam report of Symantec, the spam rate for December, 2015 was 53.1 percent. Spam not only wastes user time, energy, consumes resources, storage, computation power, bandwidth but also irritates the user with unwanted messages. For example, if you received 100 emails today. Then about approximately 70 emails are spam and only about 30 emails are ham. So, it takes time to identify the ham or important emails from it, which irritated the user. Email user receives hundreds of spam emails per day with a new address or identity and new content which are automatically generated by robot software.

Email is a spam email if it meets the following criteria:

1. Unsolicited email: - The email which is not requested by recipient.
2. Bulk mailing/mass mailing: - The email which is sent to large group of people.
3. Nameless emails: - The email in which the address and identity of the sender are hidden.

Spam emails cost billions of dollars per year to the internet service provider because of the loss of bandwidth. Spam emails causes serious problem for intended user, internet service provider and an entire internet backbone network. One of the examples to explain it, may be denial of service where the spammers send bulk emails to the server thus delaying relevant email to reach the intended recipient. Spam is a major problem that attacks the existence of electronic mails. So, it is very important to distinguish ham emails from spam emails, many methods have been proposed for classification of email as spam or ham emails.

Spam filters are the programs which detect unwanted, unsolicited, junk emails such as spam emails, and prevent them to getting to the users inbox. The filter classification techniques are categorized into two parts:

1. Based on machine learning technique.
2. Based on non-machine learning techniques.

Machine learning techniques, such as naïve Bayes, support vector machine, neural network, and decision tree etc. whereas non- machine learning techniques, such as heuristics, black/white list, signatures, Mail heading checking etc.

It is found that classification based on machine learning success ratio is very high as compared to classification based on non-machine learning.

The email is classified into spam or ham by extracting features from an email. Therefore the email classifications are based on two feature selection.

1. Header based features
2. Content based features

Both the set of features to detect spam emails have their own pros and cons. Header features can easily bypassed by the spammers.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 5, ISSN No. 2455-2143, Pages 43-47
Published Online March – April 2016 in IJEAST (http://www.ijeast.com)

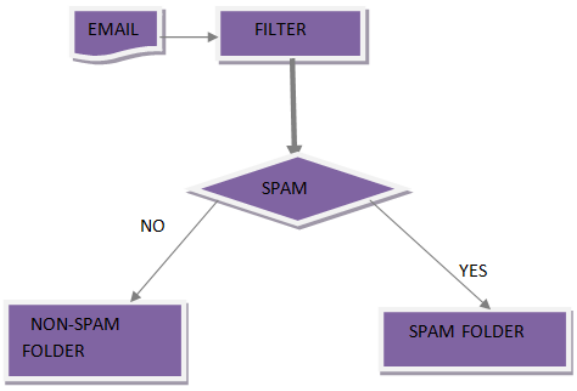

Fig 2. Flow chart of Spam filters

This paper is organized as follow section 2 presents related work, section 3 comprised of comparison of techniques, section 4 presents conclusion.

II. RELATED WORK

Bo Yu a,*, Zong-ben Xu b*(2008) performed "A comparative study for content-based dynamic spam classification using four machine learning algorithms". This paper uses the following techniques Naıve Bayesian; Neural network; Support vector machine; Relevance vector machine :
it states that NN classifier is more sensitive to the change of training set because the parameters of NN model must be decided upon network size and training algorithm. The accuracy of SVM and RVM classifier is higher than NB classifier. Hence, the RVM classification is more suitable to the SVM classification in terms of applications that require low complexity [1].

Tiago A.AlmeidaandAkeboYamakami(2010)performed" Content-Based Spam Filtering" ,using Support Vector Machines. However, there are several forms of Naive Bayes filters. They have conducted empirical experiments using well known, large and public databases. The results state that linear SVM, Boolean NB and Basic NB are the best choice for automatic filtering spams. However, SVM acquired the best average performance for all analyzed databases presenting an accuracy rate higher than 90% for all tested corpus [2].

Loredana Firte Camelia Lemnaru Rodica Potolea(2010)" Spam Detection Filter using KNN Algorithm and Resampling". This paper proposed approached for a spam detection filter. The Messages that are classified with the kNN algorithm based on a set of features extracted from the email's properties and content. [3]

RasimM. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova(2011)" Classification of Textual E-Mail Spam Using DataMining Techniques" In this paper, the problem of clustering of spam messages collection is formalized. The criterion function is a max of similarity between messages in form of clusters, which is defined by k-nearest neighbor algorithm. Genetic algorithm including penalty function for solving clustering problem is introduced in this paper [4].

Rushdi Shams and Robert E. Mercer (2013) performed a work "Classification spam emails using text and readability features". They reported a novel spam classification method that uses features, based on email content language and readability combined with the previously used content based task features. The features are extracted from four benchmark datasets such as CSDMC2010, Spam Assassin, Ling Spam, and Enron-spam. They explain all these features. Features are divided three categories i.e. traditional features, test features, and readability features. The proposed method is able to classify emails in any language because the features are language independent. They use five well-known machine learning algorithms to introduce spam classifier: Random Forest (RF), Bagging, Adaboostm 1, support vector machine (SVM), Naïve Bayes (NB). They evaluate the classifier performances and concluded that Bagging performs the best out of five. At last they compare their proposed method to that of many state-to-art anti-spam filters and concluded that their proposed method can be a good means to classify spam emails. [5]

Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora (2014) performed a work "Text and Image Based Spam Email Classification Using KNN, Naïve Bayes and Reverse DBSCAN Algorithm" The objective of their work is to detect text as well as spam emails. For this purpose they use Naïve Bayes, K- Nearest Neighbor and a new proposed method Reverse DBSCAN (Density-based spatial clustering of application with noise). They use enron cropus dataset of text as well as image. They extract words from image by using Google's open source library called, Tasseract. They use pre-processing of data. They show that preprocessing gives 50 percent better accuracy results with all the three algorithms than without using pre-processing. They concluded that naïve bayes with pre-processing gives the best accuracy among other algorithms. [6]

Masurah Mohamad and Ali Selamat (2015) performed a work "An Evaluation on the Efficiency of Hybrid Feature Selection in Spam Email Classification". They present a hybrid feature selection method, namely The Hybrid Feature Selection, in which they integrate the rough set theory and term frequency inverse document frequency (TF-IDF) to increase the efficiency result in email filters. They explain Feature Selection Methods such as Information Gain (IG), Gini Index, $X^{2\text{-Statistic}}$, Fuzzy Adaptive Particle Swarm Optimization (FAPSO) and Term Frequency Inverse Document Frequency (TF-IDF) and Machine Learning Approaches such as Naïve Bayes and Rough set theory. They use header section and spam behaviours which are non-content based keywords. They use dataset comprises of text messages and images.

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 5, ISSN No. 2455-2143, Pages 43-47
Published Online March – April 2016 in IJEAST (http://www.ijeast.com)

Then they explain their proposed spam filtering framework. In their experimental work they show that rough set theory and TF-IDF were able to work together in order to generate concise and more accurate results. [7]

Izzat Alsmadi and Ikdam Alhami (2015) performed a work "Clustering and Classification of Email Contents". In this they explain various research papers based on spam detection, ontology classification on email content and other research goals. They use the data set of general statistic about the email from Google report provided for Gmail account user. They classify the dataset based on two methods. 1) Classification based on WordNet class 2) Clustering and Classification evaluation. For clustering they use K-Means algorithm and for classification they use support vector machine. Three SVM models are evaluated such as 1. Top 100 words- VS- email before removing stop words, 2. Top 100 words-VS- email after removing stop words, 3. N Gram terms -VS- email. They concluded that the True Positive(TP) rate is shown to be very high in each case but the False Positive (FP) rate is shown to be best in case of NGram based clustering and classification .[8]

Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi, Ms.P.Lakshmisurya (2011) performed a work "Spam Classification based on Supervised Learning using Machine Learning Techniques". [9]

Megha Rathi and Vikas Pareek (2013) performed a work "Spam Email Detection through Data Mining-A Comparative Performance Analysis". [10]

Savita Pundalik Teli and Santosh Kumar Biradar (2014) performed a work "Effective Email Classification for Spam and Non-spam" [11]

Rekha and Sandeep Negi (2014) performed a work "A Review on Different Spam Detection Approaches" [12]

III. SPAM DETECTION TECHNIQUES

There are various spam detection techniques. Out of which some are machine learning and some are non- machine learning. Some of them are defined below:

A. Machine Learning Techniques

- **AdaBoost Classifier**: - A machine learning algorithm proposed by Freund and Robert Schapiro. It is a Meta algorithm which can be used in aggregation with some other learning algorithms to improve the performance of AdaBoost algorithm. AdaBoost classifier uses Confidence based label sampling that works with the concept of active learning. Classifier is trained by the variance and obtains a scoring function which is used to classify the mail as spam or ham. The labelled data is used to train the data. The trained classifier generated the required functions which classify the message as spam. This algorithm improves training process.
- **Naïve Bayes**: - A machine learning algorithm, Naive Bayes classifier is based on Baye's theorem of conditioned probability. It is used to recognize an email to be spam or ham. Conditioned Probability is given as

$$P(H/X) = P(X/H)\, P(H) / (P(X).$$

Where H denotes hypothesis, X is some evidences, P (H/X) is the probability of given evidence (X) holds by the hypothesis (H). P (X/H) is probability of X conditioned on H. P (H) – prior probability of H, independent on X. There are particularly significant words used in spam emails and ham emails. These words have probability of occurring in both emails. In advance the filters don't know these probabilities; we must train the filter to build them up. After training the word probabilities are used to compute the probability that an email have that belong to either spam or ham emails. Each particular word or only the most interesting words contribute to email's spam probability. Then, the emails spam probability is computed for every word in the emails. If this total probability exceed over certain threshold then the filters will mark that emails as spam.

- **Support Vector Machine**: - it specifies data for Classification and regression analysis. An SVM model is a represents as points in space, mapped so that separate categories are notified by a clear gap as clear as possible. Then examples are fed and then mapped into that same space and predicted to belong to a category based on which side they fall on. SVMs can efficiently perform a non-linear classification using what is called the ***kernel trick***, mapping of inputs into high-dimensional spaces.
- **Knn method**: - the k-Nearest Neighbour algorithm (or k-NN for short) is a non-parametric method used for classification and regression. It is a type of instance-based learning, or lazy learning, where the function is only estimated locally and all computation is deferred until classification. The k-NN algorithm is the simplest form of a machine learning algorithms. Both for classification and regression, they are being assign weight to the neighbours, so that the nearest neighbours can contribute more to the average than the more distant ones. For example, a common weighting method is in giving each neighbour a weight of 1/d, where d is the distance to the neighbour.
- **Relevance vector machine**:- In mathematics, a Relevance Vector Machine (RVM) is a machine learning technique that uses Bayesian inference to obtain parsimonious solutions for regression and probabilistic classification.[1] The RVM has an identical functional form to the support vector machine

International Journal of Engineering Applied Sciences and Technology, 2016
Vol. 1, Issue 5, ISSN No. 2455-2143, Pages 43-47
Published Online March – April 2016 in IJEAST (http://www.ijeast.com)

$$k(x,x') = \sum_{j=1}^{N} \frac{1}{a^j} \varphi(x,x_j)\varphi(x',x_j)$$

where φ is the kernel function (usually Gaussian), α_j's as the variances of the prior on the weight vector $w \sim N(0, \alpha^{-1} I)$.

TABLE I

COMPARISON OF DIFFERENT SPAM DETECTION TECHNIQUES

Technique	Advantages	Disadvantages
Naïve bayes	Best choice for automatic filtering spams.	Based on 'naive' Bayesian filtering, which assumes events are occurred mutually exclusively.
Support vector machine	Accuracy of this classifier is high.	Higher algorithmic complexity.
Relevance vector machine	RVM classification is more suitable to the SVM classification in terms of applications that require low complexity.	Rvm is slower as compared to other machine learning algorithms.
Knn method	The criterion function is a maximization of similarity between messages in clusters, which is defined by k-nearest neighbour algorithm.	Need to determine the value of k always. Computation cost is high for determining distance vectors.
AdaBoost	powerful classifier that works well on both basic and more complex recognition problems	AdaBoost could be sensitive to noisy data

IV. Conclusions

The impact of ensemble hybrid feature ranking method is analyzed on the benchmark classifier, Naïve Bayes.
As we have noticed that naïve classifier is the best far so on using this with "Swarm" hybrid ensemble feature ranking method, the proposed swarm intelligence algorithm can be used to solve intrusion detection as classification problems.

.

V. References

[1] Yu, Bo and Xu, Zong-ben,"A comparative study for content-based dynamic spam classification using four machine learning algorithms", Elsevier Knowledge-Based Systems,2008.

[2] Almeida, Tiago A and Yamakami, Akebo,"Content-based spam filtering", IEEE Neural Networks (IJCNN),International Joint Conference, pp.1-7,jul,2010.

[3] Firte, Loredana and Lemnaru, Camelia and Potolea, Rodica," Spam detection filter using KNN algorithm and resampling ",IEEE, 6th International Conference on Intelligent Computer Communication and Processing ,pp.27—33,Romania, Aug. 26-28, 2010.

[4] Rasim, MA and Ramiz, MA and Saadat, AN," Classification of Textual E-mail spam using Data Mining Techniques", the Journal of Applied Computational Intelligence and Soft Computing, JAN 2011.

[5] Rushdi Shams and Robert E. Mercer, "Classification spam emails using text and readability features," *IEEE 13th International Conference on Data Mining*, 2013.

[6] Anirudh Harisinghaney, Aman Dixit, Saurabh Gupta, and Anuja Arora , "Text and image based spam email classification using KNN, Naïve Bayes and reverse DBSCAN Algorithm, " *ICROIT 2014, India,* Feb 6-8 2014.

[7] Masurah Mohamad and Ali Selamat, "An evaluation on the efficiency of hybrid feature selection in spam email classification," *IEEE International Conference on Computer Communication, and Control Technology (14CT 2015)*, April. 2015.

[8] Izzat Alsmadi and Ikdam Alhami, "Clustering and Classification of email contents*," Journal of King Saud University-Computer and Information Sciences*, vol. 27, pp. 46-57, Jan. 2015.

[9] Ms.D.Karthika Renuka, Dr.T.Hamsapriya, Mr.M.Raja Chakkaravarthi, Ms.P.Lakshmisurya, "Spam Classification based on Supervised Learning using Machine Learning Techniques," *IEEE*, 2011.

[10] Megha Rathi and Vikas Pareek, "Spam Email Detection through Data Mining-A Comparative Performance Analysis," *I.J. Modern Education and Computer Science, vol. 12, pp. 31-39,* 2013.

[11] Savita Pundalik Teli and Santosh Kumar Biradar, "Effective Email Classification for Spam and Non-spam," *International Journal of Advanced Research in Computer and software Engineering, vol. 4,* June 6, 2014.

[12] Rekha and Sandeep Negi, "A Review on Different Spam Detection Approaches," *International Journal of Engineering Trends and Technology (IJETT), Vol. 1,* May 6, 2014.