A Novel Technique for Spam Mail Detection using Dendric Cell Algorithm

Rekha
Mtech student, Department of CSE
Delhi Institute of Technology
Gannaur, Sonepat

Sandeep Negi
Assistant Professor, Department of CSE
Delhi Institute of Technology
Gannaur, Sonepat

ABSTRACT
Today most of the personal and professional communication is done using the electronic media such as E-Mailing, SMS etc. But these all services also suffer from the problem of unwanted messages or the communicating information called Spam. The Spam Message can be an email virus, charity latter, commercial advertisement etc. But it affects the user time, memory and the attention. In this paper, a DCA based improved decision tree approach is suggested to identify the spam emails over the dataset. The work is implemented in integrated weka environment. The obtained result from the system shows the effective identification of spam mails over the dataset.

Keywords
Spam Email, DCA, Decision Tree, Secure Communication

1. INTRODUCTION
Email has now become one of the best ways for advertisements due to which spam emails are generated. Spam emails are the emails that the receiver does not wish to receive. A large number of identical message are sent to several recipients of email. Increasing volume of such spam emails is causing serious problems for internet users, Internet Service Providers, and the whole Internet backbone network. One of the examples of this may be denial of service where spammers send a huge traffic to an email server thus delaying legitimate message to reach intended recipients. Spam emails not only waste resources such as bandwidth, storage and computation power, but may contain fraudulent schemes, bogus offers and scheme. Apart from this, the time and energy of email receivers is wasted who must search for legitimate emails among the spam and take action to dispose the spam. Dealing with spam and classifying it is a very difficult task. Moreover a single model cannot tackle the problem since new spams are constantly evolving and these spams are often actively tailored so that they are not detected adding further impediment to accurate detection.

A spam filter is a program that is used to detect unsolicited and unwanted email and prevent those messages from getting to a user's inbox. Like other types of filtering programs, a spam filter looks for certain criteria on which it bases judgments. For example, the simplest and earliest versions (such as the one available with Microsoft's Hotmail) can be set to watch for particular words in the subject line of messages and to exclude these from the user's inbox. This method is not especially effective; it may omit legitimate messages(called false positives) and passing actual sp am messages.More sophisticated programssuch as Bayesian filte rs orother heuristic filters; attempt to identify spam through suspicious word patterns or word frequency.

Collaborative identification of spam exploits the fact that every spam message is usually sent by an automatic system to many recipients. In general, function "spam/ham" is not a computable function, and an accurate determination can only be based on the evaluation of the collective opinion of the user population. Different approaches provided which use different methods and techniques but none of them provide 100% solution. In this paper a new approach is provided which gives better results.

2. LITERATURE SURVEY
Aleksander Kocz has defined a review on challenges of service-side personalize. In this paper author describes the challenges associated with implementing large-scale personalized spam-filtering service ranging from the need to scale with the user population to the challenge of being constrained by a fixed budget [1]. An enhanced approach of email filtering based on Combining Similarity Graphs is proposed by the author [2]. The author performed the spam filtration by striking a balance between generalizing. The presented approach is effective as the similarity between the users is analyzed and relative decision is taken place. Anirudh Ramachandran presented an approach based on behavior blacklisting for the spam identification. According to this approach the email classification is been suggested based on the sender behavior rather than his identity. The author has defined a fast clustering algorithm for quick change in the message pattern. The ratio analysis is also been performed while identifying the pattern. To perform the blacklisting decision the ip address of the sender is considered [3].

A content based analysis is suggested by Jose Marea in 2006 to perform the SMS spam filtration. The work was performed on English as well as on Spanish. In this work, the message was analyzed for specific words and the patterns. For the filtration process the Bayesian filtration is performed by the user [4]. Another work on SMS spam filtration was done by Gorden V. Cormack in 2007. The author defined a feature based analysis under the supporting hypothesis to identify the spam SMS. The featured extraction from the message improves the analysis process and the efficiency of the complete system was improved upto a new level [5]. Clustering is also the process to reduced the actual dataset and to process on message parts instead of complete message based system. The author defined two cluster oriented approaches for spam filteration. In very first approach, labels are assigned to these clusters and the training and classification is performed based on medoids analysis. In the second approach, clustering on email message is performed separately so that the filtration of messages will be performed [6].

Another clustering based semi-supervised approach for filtration was proposed by John S. Whissell. The author present two spam filtering approaches for this scenario, both of which start with a clustering of training email. Our first approach uses the true labels of the medoids of each cluster to train a spam filter; our second approach functions similar to the first, except that the true label of each cluster's medoid is used as the label of every email within the cluster, giving a much larger set of labels for training, while still only requiring only a few labels [7]. One more work on SMS spamming is performed by Kuldeep Yadav for the mobile based system. The work included an intelligent approach for the analysis. The author has used the Bayesian network learning to identify the spammer and then perform the blacklisting mechanism to block the id of the user. The keyword and the pattern based analysis is been presented by the researcher. The researcher has implemented the work in real environment [8]. Miklos Erdely in year 2009 defined an approach for Web spam filtration. The author defined the analysis process based on the spam altering needs, opportunities and blockers for Internet archives via analyzing several crawl snapshots and the difficulty of migrating filter models across different crawls [9]. A spam filtration approach for Smart Senders was proposed by Pattaraporn Klangpraphant. In this present work various protection methods and software have been implemented. However, miscellaneous problems caused by spams still remain. Therefore, this paper proposed a novel method, called EMAS, to certify delivered mails. The benefits obtained from this system do not only solve the spam problem, but also untie the indirect effects from spams that no other methods have been missed [10].

3. EXISTING APPROACHES

The design has been articulated bearing in mind the complex nature of processes required. The practical approaches have reminded us that the reporting of spam can be at different levels and sources and as outlined earlier from the client to Host or peer networks or global reports etc. now there cannot be one universal filter that can impede the approach of these spam in to the host network or one specific filter for one unique host in a network or a collection of clients. Different Approaches of spam filtration are defined here under

A. Digest Comparision

The Architectural design seeks to incorporate these factors while processing for spam. The first level filter starts with the Digest comparison. Here the process to create a Digest from the incoming email message is prompted and this digest is used in an ingenious way to compare for similarities with known and reported spam email messages. The system seeks to be optimistic in that it gives a measure of flexibility in which to report spam. The flexibility is not too stringent nor too lenient.

B. User Spam Report

This method again uses three different approaches here. It checks the spam email address filter by which just the email addresses are checked to detect spam. This is again simple and a logical first step. Then the space count filter is activated if that fails. Here those email messages that uses spaces in between strings to masquerade as normal emails, are brought in to intense scrutiny to again detect spam. This is a complex procedure and the system tries to eliminate spaces intra and inter strings in a logical manner to detect spam. This is also a powerful filter and uses a previously loaded database of spam words. These words are then processed and matched for spam detections. This process has a high probability of Yielding spam if the keywords match.

C. Address Book Matcher

This is a combination of spam sources to verify the incoming addresses with the stored repository of addresses. This process also yields good results if the spam was already reported for that particular address. If spam was detected in this stage then again it is filtered and the process grinds to a halt. So as the system tries its best to detect spam at each of the above mentioned stages either independently or in unison, it also is flexible to not reporting spam of ambiguous emails. This is very important else most of the emails that would even partially ascribe to one of the above filters would be reported as spam. The system continuously seeks to update its spam dictionaries. Some of the advantages that this system offers is that even if the spam passes through undetected in one of the filters it can always be filtered out in the other.

D. Pattern Matching

Special cases when spammer includes the space between characters of a particular string to avoid being detected as spam by dictionary checking in those cases. Eliminate the intra string spaces and form the word and compare with spam Dictionary which some unwanted words and giving individual percentage to all the words.

E. Space Count Filter

The spammers in order to confuse the system resort to inserting spaces in-between letters of strings when they actually need not be there. They do this in order to separate or divide that string which would otherwise be reported as spam in to innocuous words that can escape from the spam filter. This system recognizes this and hence makes an attempt to filter such spams too. This process scans the words in the message and then attempts to use a special process whereby it eliminates the spaces and then matches it with spam dictionary. Now even when this is being done the system also uses the novel method of counting the number of spaces in a message as a means of identifying spam.

F. Rule Based Filtration

Rule-based filters assign a spam "score" to each email based on whether the email contains features typical of spam messages, such as fake SMTP components, keywords, HTML formatting like fancy fonts and background colors. A major problem with rule-based scores is that since their semantics is not well-defined, it is difficult to aggregate them and to establish a threshold that can actually limit the number of false positives. Also, experience has shown that spammers quickly learn feature-based rules and freely investigate ways to overcome them. The filters used by Rule based filtration approach are

i. *Preferred List*

This list maintains the preferred list of e-mail for each client separately. This list is compared for granting access to the client's inbox .if the client's preferred list submitted to his service provider does not contain the email id of the inward email, it is filtered.

ii. Master Spam Report

This is a comprehensive report that contains the list of spams reported across geographic and domains .the two very important sources are *Source 1:* From clients of server who report spam. This can either be intra network or internetwork. *Source 2:* From Global spam report by other server also called an Altert. The illustration depicts some of the spam reporting that the system recognizes.

G. Gateway Filration

In this approach, all inbound email is routed through a filtering gateway before being delivered to the mail server. Gateway services work well with web based and mobile access to email, and may increase robustness since they queue emails if the client network or server is off-line. On the other hand, the gateway itself is a single point of failure and may be difficult to manage in presence of multiple mail servers within an organization. A correct approach to spam filtering should not mandate any of the above choices. P2P architectures can provide high flexibility, because they smoothly adapt themselves to the underlying network and emerging application architectures.

H. Fingerprinting

Comparing the fingerprint values can detect spam. Fingerprint is a vector of digest value that is unique for a e-mails, whereby specific spam e-mails are identified and a unique "fingerprint" is developed. It is calculated based on a fingerprint algorithm using substrings. This fingerprint size is smaller than the email size. This fingerprint's are smaller than the email messages. Using Finger print vector can compare this fingerprint values.

4. PROPOSED APPROACH

Email Spamming is one of the most critical problem that includes the unwanted message communication over the web. In this work, an effective Dendric Cell Algorithm is suggested to identify the spam mails over the dataset. The work is defined as an application of data mining where the spam information is available in the form of statistical data driven from external sources. This data contains different information associated with emailing system. The work will be implemented with the integration of weka environment. This work will present an intelligent DCA approach to perform spam mail detection. This theory is based on danger cell theory in which death cell detection is performed under different constraints. The work will be defined as the fragmented molecules to identify the cell distribution and bad cell detection. It is expected that the work will provide better recognition rate.

A) DCA

In this presented work DCA and DBT based approach is presented to predict the spam mail over the dataset. The presented work is a probabilistic model in which different kind of Outliers over the dataset will be identified and intrusion detection will be performed.

In this prediction model, a weighted analysis is been performed on the dataset attribute and based on the dendric cell algorithm the initial training of data is performed. Respective to this training the dataset over the nodes is constructed. The probabilistic relationship between the attributes is identified. The dataset is defined with a set of attributes called X=(X1, X2…Xn). Each attribute is defined with some discrete value represented by Val(X). When the training algorithm is applied on this attribute set, some weighted value is identified for each attribute. Based the weighted values, the relationship between the attributes is identified. This relationship and probabilistic weighted values collective helped to generate a graph over the attribute set. The dendric cell algorithm also deals with the attributes using the conditional probability analysis. When one cell attribute is compared with the outside values conditionally. The cell attributes are considered as the independent attributes and outside cell attributes are dependent attributes. Based on this relationship, the conditional probability is estimated for the dataset. The conditional probability between two attributes I and j is given by P (Xi/Xj). Once all the attributes are defined with probabilistic attributes, then the particular instance value for all attributes is given by

Product(P(Xi/Xj), P(Xj/Xi)) where i >=1 and i<=n, j>=1 and j<=n

Some examples of probabilistic decision is listed as under

P(X1=Outlier)=0.2 P(X1 = not)=0.8

The dataset based construction over the attribute set is performed using dendric cell algorithm. This cell algorithm defined the attribute sets in two categories, the attribute present within the cell and outside the cell. The inside cell attributes are considered as the independent variables and the attributes defined outside the cell are external variables or dependent variable. The decision about the cell and non cell attributes depends on the weightage value assigned to these attributes. To construct the graph from the attribute analysis, the work is given as

1) Identify the probabilistic value for each attribute set based on value analysis on all instances.
2) Based on these probabilistic value assign the weightage to the attributes
3) Identify the high weighted attributes and present them as the cell members
4) Attribute with low weightage are identified as outside cell attribute
5) Identify the conditional value of attribute based on dendric cell algorithm
6) Identify the bound between the cell attributes so that the dataset will be constructed.

Once all the attributes and the probabilistic and conditional probabilistic value is obtained the next work is to perform the learning. The learning is here defined as

1) Parametric learning in which each attribute is learning respective to the weightage.
2) Perform the value based learning on the input attribute values respective to other values

Once the learning process is completed, the next work is performing the classification process using Dendric Cell Algorithm. The analysis of this work will be done using goal based analysis. In this algorithm the certainty and uncertainty analysis is performed based on the weighted graph analysis derived from the Dendric cell process. In this algorithmic process, the graph representation is given in the form of probabilistic weighted attribute. This value is defined based on deterministic analysis and the probabilistic table. Based on this probabilistic analysis the classification of attack will be done. The parent child relationship between the attributes and the attribute values is been performed. Based on which each instance classification value is identified. Now this obtained value is classified using belief theory. This belief theory is the deterministic analysis of probabilistic and conditional probabilistic values. The relational analysis between the attributes will be done so that the analysis between the attributes will be done to identify the attribute. The analysis is been performed on the differential constraints so that different dataset instance will be identified respective to the difference values.

5. RESULTS

The presented work is about to identify the spam emails over the email dataset defined with email characteristics. The work is implemented in integrated weka environment. The authentication of this work depends on the approach applied to perform the dataset encoding. The dataset considered in this work is described in table 1.

Table 1: Dataset Properties

Property	Value
Name of Dataset	SpamBase
URL	http://www.cs.uu.nl/docs /vakken/dm/spambase.arff
Type	Arff
Number of Instances	4601
Number of Attributes	58
Class Attributes	1

The complete dataset is divided in training and testing sets and later on the DCA approach is applied to identify the spam emails over the dataset. The results driven from the approach is shown in figures.

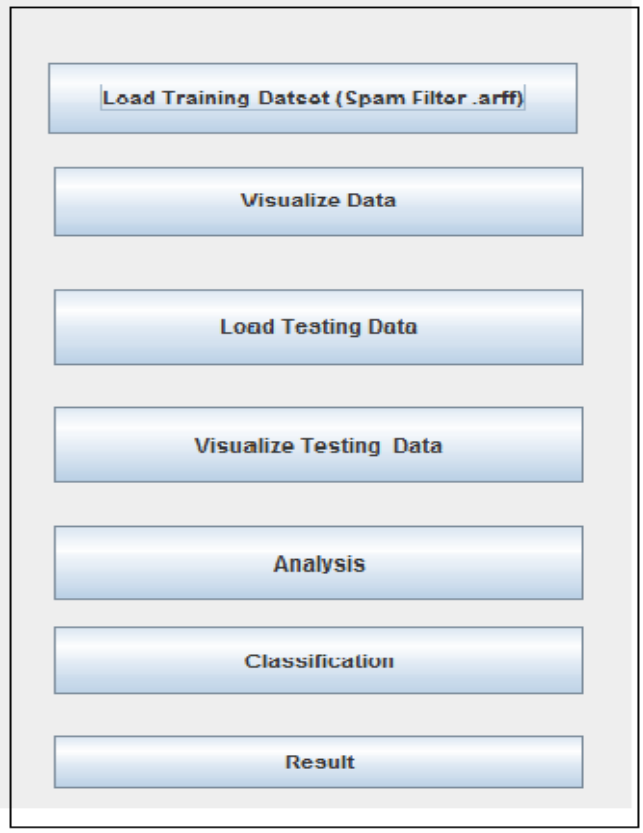

Figure 1: GUI designed for Approach

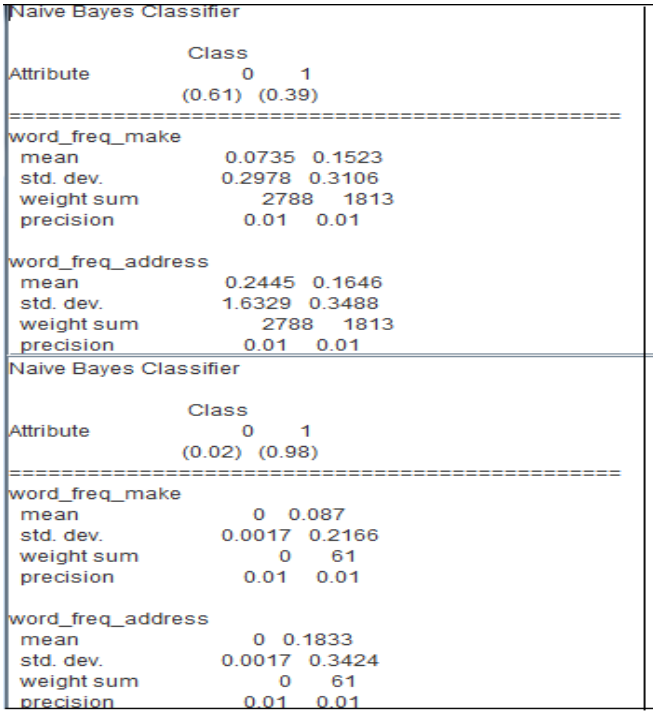

Figure 2: Analysis of same dataset with Naïve Bayes classifier Approach

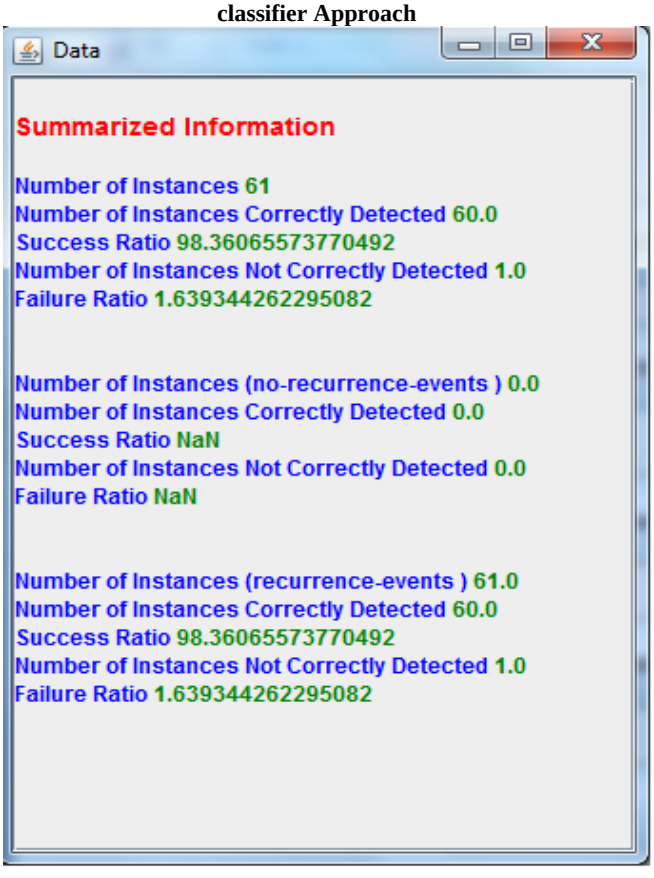

Figure 3: Result obtained using the dataset with Approach

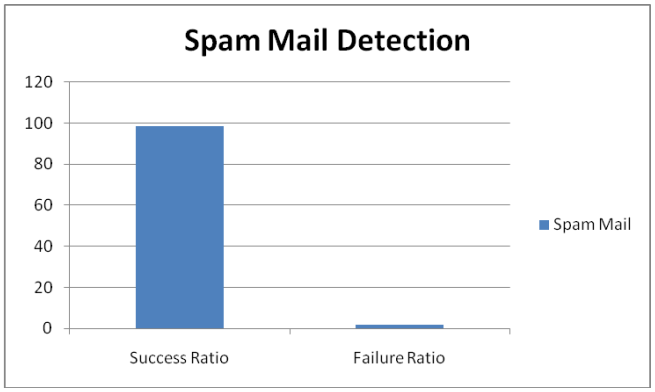

Figure 4: Graph showing the success and failure ratio of used Approach

Here figure 1 is showing the results obtained from the work for a testing set of 62 records. The result shows that the out of 62, 61 records are identified successfully and 1 is recognized as wrong result. The success ratio obtained in this work is about 98.36%. The result shows the presented work is effective enough to identify the spam emails over the dataset.

6. CONCLUSION

The aim of this paper is to explore different spam detection methods and classify them as such. The paper has defined an effective DCA based approach for spam mail detection over the dataset. The work is applied on a dataset of 62 records and about 98.3% accuracy is achieved from the work. There is very much scope for identifying mail as spam emails or legitimate mails for text as well as multimedia messages.

7. REFERENCES

[1] Aleksander Kojcz, "The challenges of service-side personalized spam ltering: scalability and beyond", Proceedings of the 1st international conference on Scalable information systems, May 2006

[2] Anirban Dasgupta, "Enhanced Email Spam Filtering through Combining Similarity Graphs", Proceedings of the fourth ACM international conference on Web search and data mining, February 2011

[3] Anirudh Ramachandran, "Filtering Spam with Behavioral Blacklisting", Proceedings of the 14th ACM conference on Computer and communications security, October 2007

[4] Carlos Laorden, "Enhancing Scalability in Anomaly-based Email Spam Filtering", Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, September 2011

[5] Gordon V. Cormack, "Feature Engineering for Mobile (SMS) Spam Filtering", SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands. ACM, 2007

[6] John S. Whissell, "Clustering for Semi-Supervised Spam Filtering", Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, September 2011

[7] José María Gómez Hidalgo, "Content Based SMS Spam Filtering", Proceedings of the 2006 ACM symposium on Document engineering, October 2006

[8] Kuldeep Yadav, "SMSAssassin: Crowdsourcing Driven Mobile-based System for SMS Spam Filtering", Proceedings of the 12th Workshop on Mobile Computing Systems and Application ACM, March 2011

[9] Miklós Erdélyi, "Web Spam Filtering in Internet Archives", Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web, April 2009

[10] Pattaraporn Klangpraphant, "E-Mail Authentication System: A Spam Filtering for Smart Senders", Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, November 2009