International Journal of Innovative Research in Advanced Engineering (IJIRAE)
Volume 1 Issue 1 (April 2014)

An analysis on Filter for Spam Mail

Anand G Sharma
PG-Scholar/IT
Institute of Engineering and Technology, Alwar (RAJ)
Alwar, India
sharma.anand2008@gmail.com

Prof. Vedant Rastogi
Associate Professor/IT
Institute of Engineering and Technology, Alwar (RAJ)
Alwar,India
vedantnoki@gmail.com

Abstract— Electronic-mail is widely used most suitable method of transferring messages electronically from one person to another, rising from and going to any part of the world. Main features of Electronic mail is its speed, dependability, well-equipped storage options and a large number of added services make it highly well-liked among people from all sectors of business and society. But being popular it also has negative side too. Electronics mails are preferred media for a large number of attacks over the internet.. A number of the most popular attacks over the internet include spams. Some methods are essentially in detection of spam related mails but they have higher false positives. A number of filters such as Checksum-based filters, Bayesian filters, machine learning based and memory-based filters are usually used in order to recognize spams. As spammers constantly try to find a way to avoid existing filters, a new filters need to be developed to catch spam. This paper proposes to find an resourceful spam mail filtering method using user profile base ontology. Ontologies permit for machine-understandable semantics of data. It is main to interchange information with each other for more efficient spam filtering. Thus, it is essential to build ontology and a framework for capable email filtering. Using ontology that is particularly designed to filter spam, bunch of useless bulk email could be filtered out on the system. We propose a user profile-based spam filter that classifies email based on the likelihood that User profile within it have been included in spam or valid email.

Keywords— Electronic mail, spam, ontology, filter, RDF

I. INTRODUCTION

As the number of Internet user's increases Electronic mail has been a capable and well-liked communication system. Electronic mail is most convenient method of transmitting messages electronically from one person to other person and going to anywhere in the world. Electronic mail connects each and every location around the world, eliminating geographical limitations and bridging people closer. On the technical side, it includes a number of protocols such as SMTP, TCP/IP, POP, and so on, for transferring messages from one mailbox to another mailbox. Electronic mail can be appropriately defined as a method for interchanging digital information from one author to one or more recipients and it has become the standard media of communication in various areas. It provides many eye-catching features by its virtue such as simple, fast and free access, global acceptance, support for instant messaging protocols, support for file attachments, etc. [3] Therefore, managing email became an important and growing problem for individuals and organizations because it is prone to mistreatment. The blind posting of needless email messages is said to be spam is an example of the misuse. Spam is normally defined as sending of unwanted bulk email - that is, email that was not asked for by multiple recipients. A another common definition of a spam is limited to unwanted commercial email, a definition that does not consider non-commercial solicitation such as political or religious pitches, even if unwanted, as spam. Email was by far the most common form of spamming on the internet. According to the data estimated by Ferris Research [1], spam accounts for 15% to 20% of email at U.S.-based corporate organizations. Half of users are receiving 10 or more spam emails per day while some of them are receiving up to several hundred unwanted emails. International Data Group [2] expected that global email traffic surges to 60 billion messages daily. It involves sending identical or nearly identical unwanted messages to a large number of recipients. Unlike valid commercial email, spam is generally sent without the explicit permission of the recipients, and frequently contains various tricks to bypass email filters.

II. LITERATURE SURVEY

1. MEANING OF AN ONTOLOGY

Ontology is an open specification of a conceptualization. Ontologies can be taxonomic hierarchies of classes, class definitions, but need not be limited to these forms. Also, ontologies are not limited to conservative definitions. To specify a conceptualization one needs to state principle that constraint the possible interpretations for the defined terms. Ontologies play a key role in capturing domain knowledge and providing a common understanding. Generally, ontologies consist of classified class hierarchy, domain knowledge base, and relationships between classes and instances. An ontology has different relationships depending on the schema or classification builder, and it has different restrictions depending on the language used. Also, the domain, range, and cardinality are different based on ontology builder. Ontologies permit for machine-understandable semantics of data, and make the effective search, interchange, and addition of knowledge for business-to-business and business to- consumer (B2C) e-commerce. By using semantic data, the usability of e-technology can be facilitated. There are several languages such as extensible markup

language (XML), resource description framework (RDF), RDF schema (RDFS), DAML+OIL, and OWL. Many tools have been developed for implementing metadata of ontologies using these languages. However, current tools have problems with interoperation and association.

2. DEVELOPMENT OF ONTOLOGY

Ontology development tools can be applied to all phases of the ontology lifecycle together with the formation, implementation and maintenance of ontologies. An ontology can be used to maintain various types of knowledge management including knowledge recovery, sharing and storage. In one of the most accepted definitions, ontology is the specification of shared knowledge. For a knowledge management system, ontology can be defined as the categorization of knowledge. Ontologies are different from conventional keyword-based search engines in that they are metadata, capable to afford the search engine with the functionality of semantic match. Ontologies are able to search more efficiently than usual methods. Basically, ontology consists of hierarchical explanation of main concepts in a domain and the descriptions of the properties of each concept. Traditionally, ontologies are built by both skilled knowledge engineers and domain specialists who may not be well-known with computer software. Ontology creation is a time-consuming task. Its tools require users to be trained in knowledge representation and guess logic. XML is not suited to describe machine understandable documents and interrelationships of resources in ontology [7]. Ontology tools have to tolerate more expressive power and scalability with a large knowledge base, matching and reasoning in querying. Also, they need to support the use of high-level language, visualization, modularity.

3.IDENTIFYING SPAM MAIL

A algorithm was developed to decrease the feature space without sacrificing remarkable correctness, but the usefulness was based on the excellence of the training dataset, demonstrated that the feasibility of the approach to find the best learning algorithm and the metadata to be used, which is a very significant contribution in email categorization using Rainbow system[8]. A graph based mining approach for email categorization structures/patterns can be extracted from a pre-classified email folder and the same can be used effectively for categorizing incoming email messages [7].Which are spam-specific features in their work, could improve the categorization results. A good performance was obtained by removing the classification error by finding temporal relations in an email sequence in the form of temporal sequence patterns and embedding the discovered information into content-based learning methods. It shows the work on spam filtering using feature selection based on heuristics. It introduced a method to help various classifiers to increase the mining of classified profiles. Upon receiving a document, the technique helps to create dynamic classified profiles with respect to the document, and accordingly helps to make proper filtering and categorization decisions.

III. PROPOSED WORK

The training datasets are the set of emails that gives us a categorization result. The test data is actually the email that will run through our system which we test to see if classified correctly as spam or not. This will be an ongoing test process and so, the test data is not finite because of the learning procedure, and will sometimes merge with the training data. To do that, the training datasets should be modified as a compatible input format. To query the test email, an ontology should be created based on the categorization result. To create ontology, an ontology language is required. MDF will be use to create an ontology. The categorization result of MDF format will be inputted to Jena, and inputted MDF will deploy through Jena, finally, an ontology will be created. An ontology generated in the form of MDF data model is the base on which the incoming mail is checked for its legitimacy. Depending upon the assertions that we can conclude from the outputs of Jena, the email can be defined as spam or valid. Customized ontology filter would be different with each other depending on each user's background, preference, etc. That means one email might be spam for person A, but not for person B. The ontology evolves periodically and adaptively. The input to the system is mainly the training datasets and then the test email. The test email is the first set of emails that the system will classify and learn and after a certain time, the system will take a variety of emails as input to be filtered as spam or valid. The ontology technique enables us to decide the way different headers and the data inside the email are linked based upon the word frequencies of each words or characters in the datasets. The mapping also enables us to obtain assertions about the legitimacy and non-legitimacy of the emails. The next part is using this ontology to decide whether a new email is spam or valid. Queries using the obtained ontology will process again through Jena. The output obtained after querying will be the decision that the new email is spam or valid [11].

1. ARCHITECTURE OF SPAM FILTER

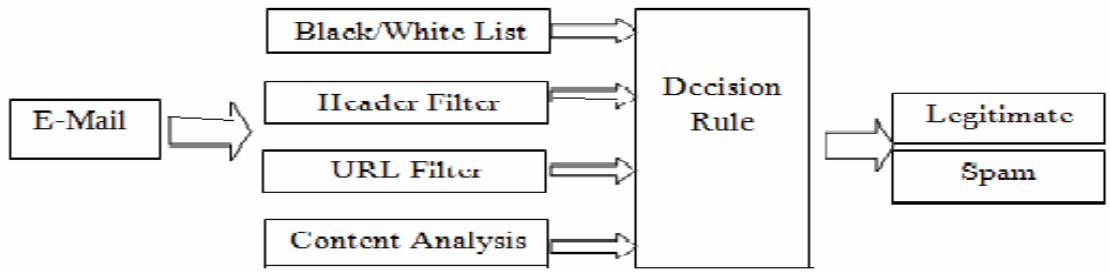

Figure 1 *Schematic representation of spam filters.*

Figure 1 shows our framework to filter spam. The training dataset is the set of email that gives us a categorization result. The test data is actually the email will run through our system which we test to see if classified correctly as spam or not. This will be an ongoing test process and so, the test data is not finite because of the learning procedure, the test data will sometimes merge with the training data. The training dataset was used as input to categorization algorithm. To do that, the training dataset should be modified as a compatible to query the test email in Jena, an ontology should be created based on the categorization result. To create ontology, an ontology language was required. MDF was used to create ontology. The categorization result in the form of MDF file format was inputted to Jena, and inputted RDF was deployed through Jena, finally, ontology was created. Ontology generated in the form of MDF data model is the base on which the incoming mail is checked for its legitimacy. Depending upon the assertions that we can conclude from the outputs of Jena, the email can be defined as spam or otherwise. The email is actually the email in the format that Jena will take in (i.e. in a CSV format) and will run through the ontology that will result in spam or not spam. The input to the system mainly is the training dataset and then the test email. The test email is the first set of emails that the system will classify and learn and after a certain time, the system will take a variety of emails as input to be filtered as a spam or not. The training dataset which we used, which had categorized values for features on the basis of which the decision tree will classify, first be given to get the same. The categorization results need to be converted to ontology. This ontology enabled us to decide the way different headers and the data inside the email are linked based upon the word frequencies of each words or characters in the dataset. The mapping also enabled us to obtain assertions about the legitimacy and no legitimacy of the emails. The next part was using this ontology to decide whether a new email is a spam or not. This required querying of the obtained ontology which was again done through Jena. The output obtained after querying was the decision that the new email is a spam or not [15].

2 SPAM IDENTIFICATION METHODS

The several different methods to identify incoming messages as spam are, White list/Blacklist, verifying URL , Checking Mail header, Keyword checking.

A. ***WHITE LIST/BLACKLISTS***

White list is a list, which contains all addresses from which we always wish to functional domains. white list management tool that eliminates the need for administrators to manually input approved addresses on the white list and ensures that mail from particular senders or domains are never flagged as spam. The number of records can be configured. When an overflow occurs, obsolete records are overwritten. A blacklist works similarly to competitive alternatives: this is a list of addresses from which we never want to receive mail [15].

B. ***KEYWORD CHECKING***

Another method widely used in filtering spam. It works by scanning both email subject and body. Using "conditions" i.e. combinations of keywords is a good solution to enhance filtering efficiency. We can specify combinations of words and update the list that must appear in the spam email. All messages that include these words will be blocked [9].

C. ***CHECKING MAIL HEADER***

This technique consists of a set of rules that, if a mail header matches, triggers the mail server to return messages that have blank "From" field, that lists a lot of addresses in the "To" from the same source, that have too many digits in email addresses (a fairly popular method of generating false addresses). It also enables to return messages by matching the language code stated in the header.

D. ***VERIFYING URL***

Verifying URL email can be classified on base of URL. Blocks all the mails from the list of addresses from which user never want to receive mail.

IV. APPLICATION

1. Effective email client

2. Private White list and blacklist. User can maintain their own lists of secured email senders (white list) and suspicious email addresses, from which you are used to receive spam (blacklist). User can update list whenever want to update.
3. Certain spam solutions make the choice about spam emails not only based on the sender's mail address, they also analyze the subject lines and the message content and remove the spam mail before entering to inbox
4. Elimination of virus from emails
5. Provide the user only required content.

V. CONCLUSION

In this paper, we present Spam or unwanted Electronic mail has become a major problem for organizations and private users. From the study we found that, many of the filtering techniques has been proposed for sorting the mails are based on Blacklist, White list, Bayesian filter methods and there is no method can maintain to offer an ideal solution. Our approach is relay on applying profile base categorization techniques by means of Ontology tools. There is lot of scale for research in classifying text messages as well as multimedia messages. A personalized ontology filter was evolved based on specific user's background. Hence, as expected, better sorting for spam can be achieved using a customized ontology filter which is scalable. Text oriented email datasets are adapted, but the same idea can be applicable to other categorization work.

REFERENCES

[1] Ferris Research. Spam Control: Problems & Opportunities, 2003.
[2] Lourdes Araujo and Juan Martinez-Romo, Web Spam Detection: New Categorization Features Based on Qualified Link Analysis and Language Models. Proceeding of IEEE Transactions on information forensics and security,vol.5, no. 3, September.
[3] Take, Aura Kumar & Tapas, Shashikala, (2010) "Knowledge Base Compound Approach towards Spam Detection", *Recent Trends in Network Security and Applications*, Communications in Computer and Information Science, Vol 89, Part 2, pp 490-499, DOI=10.1007/978-3-642-14478-3_49, Published by Springer Berlin, Heidelberg
[4] Giorgio Fumera, Ignazio Pillai, Fabio Roli, Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. Proceedind of Journal of Machine Learning Research 7 (2006) 2699-2720.
[5] Youn, S, and McLeod, D. Ontology Development Tools for Ontology-Based Knowledge Management.Encyclopedia of E-Commerce, E-Government and Mobile Commerce, Idea Group Inc, 2006.
[6] Seongwook Youn, Dennis McLeod, Spam Email Categorization using an Adaptive. Proceedind of Journal of Software, vol.2, no.3, September.
[7] Shankar, S., and Karypis, G. Weight adjustment schemes for a centroid based classifier. Computer ScienceTechnical Report TR00-035, 2000.
[8] Yang, J., Chalasani, V., and Park, S. Intelligent Email Categorization Based on Textual Information and Metadata. IEICE TRANS. INF. & SYST., VOL. E82, NO.1 JANUARY 1999.
[9] Rathore, Shashikant, Jassi, Palvi & Agarwal, Basant, (2011) "A New Probability based Analysis for Recognition of Unwanted Electronic mails", *International Journal of Computer Applications (IJCA)*,Vol. 28, Issue 4, Published by Foundation of Computer Science, New York, USA.
[10] Luo, Yan, (2010) "Workload characterization of Spam Electronic mail Filtering Systems", *International Journal of Network Security & Its Applications (IJNSA)*, Vol. 2, No.1.
[11] Yang, J., Chalasani, V., and Park, S. Intelligent Email Categorization Based on Textual Information and Metadata. IEICE TRANS. INF. & SYST., VOL. E82, NO.1 JANUARY 1999.
[12] Sakkis, G., Androutsopoulos, I., & Paliouras, G., (2003) "A Memory based Approach to Antispam Filtering", *Information Retrieval*, Vol 6, pp 49-73.
[13] Saraubon, Kobkiat & Limthanmaphon, Benchaphon, (2009) "Fast Effective Botnet Spam Detection", *Fourth International Conference on Computer Sciences and Convergence Information Technologies (ICCIT)*, pp 1066-1070.
[14]Pundt, H., and Bishr, Y. Domain ontologies for data sharing: An example from environmental monitoring using field GIS. Computer & Geosciences, 28, 98-102, 1999.
[15] Jindal, Nitin & Liu, Bing, (2007) "Analyzing and Detecting Review Spam", *Seventh IEEEInternational Conference on Data Mining (ICDM)*, Published by IEEE.