IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3, May 2014.
www.ijiset.com
ISSN 2348 - 7968

Filter for Spamming Email by Using Ontology

Anand G Sharma[1], Mr. Vedant Rastogi[2]
[1]M-Tech Scholar, Department of Information Technology
[2]Associate Professor, Department of Information Technology
Institute of Engineering and Technology, Alwar (RAJ), ALWAR,INDIA

Abstract: As the number of users of email increases it tends to public irritation over spam. Email has become one of the easiest, fast and most economical form of communication. During the past few years as the number of email users increases it results in the dramatic increase of spam emails. As spammers always try to uncover a way to avoid existing filters, new filters require to be developed to catch spam. Ontologies allow for machine-understandable semantics of data. It is essential to share information with each other for more efficient spam filtering. Thus, it is necessary to construct ontology and a framework for effective email filtering. Using ontology that is specifically designed for filtering spam, bunch of unsolicited bulk email could be filtered out on the system. Similar to other filters, the ontology evolves with the user requests. Hence the ontology would be modified for the user. This project proposes to find an efficient method for filtering of spam email using user profile based ontology. We propose a user profile-based spam filter that classifies email based on the possibility that user profile within it have been included in legal or in spam email. If a exacting User profile is more frequently seen in spam, the system classifies the host message accordingly.

1. INTRODUCTION

Email has been an efficient way of information exchange as the number of email users increases, it is became a main growing problem for individuals and organizations to manage email, because it tends to misuse. Spam is commonly defined as sending of unwanted bulk email - that is, email which is not called by multiple recipients. A additional common definition of a spam is restricted to unwanted commercial email, a definition that does not consider non-commercial solicitation such as political or religious pitches, even if unsolicited, as spam. Email is the most common form of spamming on the internet. According to the data estimated by Ferris Research [6], 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 hundreds unsolicited emails. International Data Group [8] expected that global email traffic surges to 60 billion messages daily.

It involves sending similar or nearly identical unsolicited messages to many of recipients. Unlike legitimate commercial email, spam is generally sent without the prior permission of the recipients, and frequently contains various ways to deny email filters. We constructed a framework for efficient email filtering using ontology. Ontologies allow for machine-understandable semantics of data, so it can be used in any system [12]. It is important to share the information with each other for more effective spam filtering. Thus, it is necessary to build ontology and a framework for efficient email filtering. Using ontology that is specially designed to filter spam, bunch of unsolicited bulk email could be filtered out on the system. This project proposes to find an efficient spam email filtering method using ontology. We used Waikato Environment for Knowledge Analysis (Weka) explorer, and Jena to make ontology based on sample dataset. Emails can be classified using different methods. Different people or email agents may maintain their own rules for email specification. The problem of spam filtering is somewhat old one and there are already different methods approaches to the problem that have been implemented. The problem is more particular to areas like artificial intelligence and machine learning. Some methods had various trade-offs, difference performance metrics and different classification efficiencies.

1.1. Characteristics and Definition of Spam

There exist various definitions for spam or junk mail. Explaining and how it differs from legitimate mail (also called non-spam, genuine mail or ham). The shortest among the popular definitions characterizes spam as \unsolicited bulk email" [8]. Sometimes the word commercial is added, but this extension is debatable. The TREC Spam Track relies on a similar definition: spam is unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user"[5] . Another widely accepted definition states that Internet spam is one or more unsolicited messages, sent or posted as part of a larger collection of messages, all having substantially identical content" [18]. Direct Marketing Association proposed to use the word “spam" only for messages with such kinds of content, such as pornography, but this idea met no enthusiasm, being considered an attempt to legalize other kinds of spam [7]. As we can see, the common point is that spam is unsolicited; according to a widely cited formula spam is about consent, not content. It is necessary to mention that the notion of being unsolicited is hard to capture. In fact, despite the wide agreement on this type of definitions the filters have to rely on content and ways of delivery of messages to recognize spam from legitimate mail.

There is a growing scientific literature pointing to the characteristics of the spam phenomenon. In general, spam is used to advertise different kinds of goods and services, and the percentage of advertisements dedicated to a particular kind of goods or services changes over time [11]. Quite often spam serves the needs of online frauds. A special case of spamming activity is phishing, namely hunting for sensitive information (passwords, credit card numbers, etc.) by imitating official requests from trusted authorities, such as banks, service providers. Another type of malicious spam content are viruses. Sometimes a massive spam attack can be used also to upset the work of a mail server [10]. To sum up, the sender of a spam message pursues one of the following tasks: to advertise some goods, services, or ideas, to cheat users out of their private information, to deliver malicious software, or to cause a temporary crash of a mail server. Characteristics of spam traffic are different from those of legitimate mail traffic, in particular legitimate mail is concentrated on diurnal periods, while spam arrival rate is stable over time. Spammers

ISSN 2348 - 7968

usually mask their identity in different ways when sending spam, but they often do not when they are harvesting email addresses on websites, so recognition of harvesting activities can help to identify spammers [15].

1.2. Email can result in exposure to the following

1.2.1. Loss of confidential or sensitive data

Sensitive information regarding proprietary technology, trade secrets, corporate strategy or financial data can easily find its way into an outbound email and to a opponent, whether through intentional efforts by an end-user or through unintentional means. (Example: The user unintentionally hits "Reply All" rather than "Reply.")

1.2.2. Proliferation of spam

Unsolicited email or junk mail, known as spam, consumes valuable disk space, diverts employees attention and increases network congestion. The volume of spam has continually increased over the past several years, and that trend is expected to continue thereby straining an organization's network resources.

1.2.3. Loss of employee productivity

End-users often consider corporate email to be a personal communication tool, and spend a large amount of time sending and receiving nonwork- related information, such as jokes, electronic greeting cards, audio and video files and chain letters. This impacts the productivity of the organization.

1.2.4. Legal liability

Because many end-users treat email as casual conversation, emails often contain off-color, offensive or inappropriate material. The transmission of a single offensive email can initiate a long and expensive litigation process for an organization and also severely damage its reputation.

1.2.5. Negative impact on corporate image

Organizations must realize that outbound email can be regarded as an official communication (i.e., sent on 'electronic letterhead') and, when used inappropriately by end-users, can deliver a severe blow to an organization's reputation.

1.2.6. Exposure to virus threats

The widespread use of the Internet and rapid spread of complex viruses via email have created security issues for organizations of all sizes. Infected emails can be broadcast to entire corporate networks through gateways and mail servers, thereby halting mission-critical business processes.

1.2.7. Degradation of network performance

Large volumes of emails containing attachments (many not work-related) can severely impact the performance of an organization's Internet connection, as well as obstruct transmission of true business-critical traffic.

2. RELATED WORK &LITERATURE SURVEY

2.1. Understanding of an Ontology

An ontology is an explicit specification of a conceptualization. Ontologies can be taxonomic hierarchies of classes, class definition, or subsumption relation, but need not be restricted to these forms. Also, ontologies are not incomplete to conventional definitions. To denote a conceptualization one needs to shape axioms that constrain the possible interpretations for the defined terms [4]. Ontologies occupy a key role in capturing area knowledge and providing a common understanding. Generally, ontologies contain of taxonomy, domain knowledge base, class hierarchy and relationships between classes and objects. Ontology has different relationships depending on the schema or taxonomy builder, and it has different limitations depending on the language used. Also, the domain, range, and cardinality are different based on ontology builder. Ontologies allow for machine-understandable semantics of data, and facilitate the search, exchange, and integration of knowledge for business-to-business (B2B) and businessto- consumer (B2C) e-commerce. By using semantic data, the usability of e-technology can be facilitated. There are several languages like 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 collaboration.

2.2. Filtering of Spam

A algorithm was developed to reduce the feature space without sacrificing remarkable classification accuracy, but the effectiveness was based on the quality of the training dataset [17] 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 classification using Rainbow system[14]. A graph based mining approach for email classification that structures/patterns can be extracted from a pre-classified email folder and the same can be used effectively for classifying incoming email messages [1].

The spam-specific features in their work, could improve the classification results. A good performance can be obtained by reducing the classification error by finding secular relations in an email sequence in the form of secular sequence patterns and putting the discovered information into content-based learning methods. Spam filter technique helps to a variety of classifiers to improve the mining of category profiles. Upon receiving a document, the technique helps to generate dynamic category profiles with respect to the document, and accordingly helps to make proper filtering and classification decisions.

2.3. Development of an Ontology

Developing Ontology tools can be used to all stages of the ontology lifecycle including the creation, population, implementation and maintenance of ontologies. An ontology can be used to support various types of knowledge management including knowledge retrieval, storage and sharing[10]. In one of the most popular definitions, an ontology is "The specification of shared knowledge". An ontology can be regarded as the classification of knowledge for a knowledge management system. Ontologies are differ

ISSN 2348 - 7968

from traditional keyword-based search engines in that they are metadata, able to provide the search engine with the functionaliy of semantic match. Ontologies are capable to search more efficiently than traditional methods. Normally, an ontology consists of hierarchical description of important concepts in a domain and the descriptions of the properties of each concept. Traditionally, ontologies are built by both highly trained knowledge engineers and domain specialists who may not be familiar with computer software. Ontology construction is a time-consuming and laborious task. Ontology tools also require users to be trained in knowledge representation and predicate logic. XML is not suited to describe machine understandable documents and interrelationships of resources in an ontology [5]. Ontology tools have to support more expressive power and scalability with a large knowledge base, and reasoning in querying and matching. Also, they need to support the use of high-level language, modularity, visualization.

3. ANALYSIS OF PROBLEM

Spam officially call as unsolicited bulk email (UBE) or unsolicited commercial email (UCE). is rapidly becoming a major problem on the Internet. At the end of 2002, as much as 40% of all email traffic consisted of spam and recent reports estimate that this amount has risen to more than 50%.

To hold this increasing load of junk email, several spam filtering techniques exist to automatically sort incoming email as spam, and to discard email classified as such. In this project we are going to examine the effectiveness of these spam filtering techniques and technologies.

For users, receiving spam is quite irritation and costs money. In a recent survey of the European Community 4, it was conventional that the cost for receiving spam for an average Internet user is in the order of 30 euro a year. But the cost of spam goes well beyond the total costs of all recipients. Each ISP pays for each email message received, because it must be stored in a mail box and it takes up a certain amount of bandwidth. The total cost has been projected in the order of 10 billion euro a year.

An another problem with spam is the impact it has on the Internet backbone. Spam sent over the Internet backbone tends delay for all Internet users. Furthermore, because most spammers use mailing lists that not having updated addresses on them, many messages are rejected also said to be bounced. This mandates the operator of the intended destination to send a return response, killing even more bandwidth.

Bulk mailers use several different techniques to send their spam. Often, bulk mailers maltreatment the SMTP protocol or use badly configured MTAs for hide their tracks.

There are at least three basically different techniques to catch spammers. First, bulk mailers can be prohibited to send spam by limiting access to mail servers. Second technique is make spamming less profitable, for example by taxing a cost on every email message sent. A third technique aims to detect and remove all spam once it is sent by using diverse filtering techniques that use the particular characteristics of spam to recognize it.

Our study of counter measures against spam targeted on filtering techniques. We are concerned in measuring the correctness level of these filters in practice. Some of the filtering Techniques not only look at the content of each message, but also think the email traffic at large. To faithfully examine such spam filters, we will build a simulator to generate practical email traffic and test the filters with it. While running at server level, the filter may use the information about connections from the server. On the other hand while using at user level, the filter will use a trained or customized as a function of user specific characteristics. We have found that filters based on genetic algorithms, perform best at ISP level and naive Bayesian filters perform best at user level.

4. PROPOSED WORK

The training datasets are the set of emails that gives us a classification result. The test data is essentially the email that will run through our system which we test for classified as spam or not. This will be an continuing test process and so, the test data is not finite for the reason that of the learning procedure, and may be merge with the training data. To do that, the training data sets should be customized as a suitable input format. To query the test email in Jena, ontology should be formed based on the categorization result. To generate ontology, an ontology language is required. Resource Description Framework (RDF) will be use to create an ontology. The categorization result of RDF format will be inputted to Jena, and inputted RDF will deploy throughout Jena, finally, an ontology will be created. An ontology generated in the form of RDF data model is the base on which the incoming mail is verified for its legitimacy. According to the assertions that we can conclude from the outputs of Jena, the email can be defined as spam or legal.

The email is essentially 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 legitimate. The system updates at times the datasets with the emails surreptitious as spam when user spam report is requested. Then, increased training datasets are inputted to Weka to get a new categorization result. Through this cycle, the number of ontology will be increased. Finally, this spam filtering ontology will be modified for each user.

Modified ontology filter would be different with each other depending on each user's background, preference, hobby, needs etc. That means one email might be legal for person A and as spam for person B. The ontology evolves at times and adaptively. The training datasets and then the test email is mainly input to the system. The test email is the first set of emails that the system will categorize and learn and after a assured time, the system will take a diversity of emails as input to be filtered as legitimate or spam. The ontology technique provide us facility to decide the way diverse 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 (legal) and non-legitimacy (illegal) of the emails. Another main part of using this ontology to decide whether a new email is legitimate or spam. 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 legitimate [18].

4.1 Implementation and Architecture for system

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 essentially the email will run through our system which we test for whether categorized correctly as spam or not. This will be an continuing 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 J48

ISSN 2348 - 7968

classification. To do that, the training dataset should be modified as a compatible to query the test

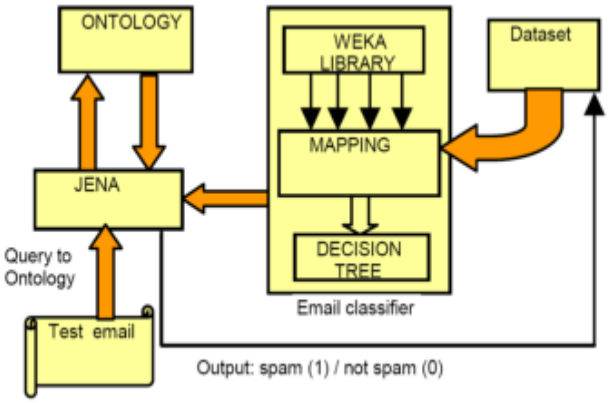

Figure 1. Filtering Architecture

email in Jena, an ontology should be formed based on the categorization result. To generate ontology, an ontology language was required. RDF was used to create an ontology. The classification result in the form of RDF file format was inputted to Jena, and inputted RDF was deployed through Jena, finally, an ontology was created. Ontology generated in the form of RDF 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 frequently is the training dataset and then the test email. The test email is the first set of emails that the system will organize 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 classification values for features on the basis of which the decision tree will classify, will first be given to get the same. The classification results need to be converted to an ontology. The decision result which we obtained J48 categorization was mapped into RDF file. This was given as an input to Jena which then mapped the ontology for us. 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 nonlegitimacy
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 choice that the new email is a spam or not. The main way where user can let the system know would be through a GUI or a command line input with a simple 'yes' or 'no'. This would all be a part of a full fledged working system as opposed to our prototype which is a basic research model.

5. OBJECTIVES

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 fixed because of the learning process, 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 in Jena, an ontology should be created based on the classification result. To create ontology, an ontology language is required. Resource Description Framework (RDF) will be use to create an ontology. The classification result of RDF format will be inputted to Jena, and inputted RDF will deploy through Jena, finally, an ontology will be created. An ontology generated in the form of RDF 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 legitimate.

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 legitimate. The system updates periodically the datasets with the emails classified as spam when user spam report is requested. Then, increased training datasets are inputted to Weka to get a new classification result. Through this procedure, the number of ontology will be increased. Finally, this spam filtering ontology will be customized for each user.

Customized ontology filter would be different with each other depending on each user's background, preference, hobby, 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 legitimate. 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 legitimate. 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 legitimate [18].

For this proposed work it is an effort to address some of the following issues related to:

1. In this project we will develop a customized ontology filter based on explicit user's profile. Hence, enhanced spam filtering rate can be achieve.
2. Classification correctness can be improved initially by pruning the tree and using better categorization algorithms, more number and feature elements, etc.
3.Use an ontology to help classifying emails on basis of text.

6. REFERENCES

[1] Aery, M., and Chakravarthy, S. eMailSift: Email Classification Based on Structure and Content. Proceedings of The 5th IEEE International Conference on Data Mining (ICDM05), Clearwater Beach, FL, 18-25. 2005.

[2] K.R.Ananthapadmanaban, Dr.S.K.Srivatsa. Personalization of user Profile: Creating user Profile Ontology. Proceeding of International Journal of Computer Applications (0975–8887) Volume 23– No.8, June 2011

[3] Giorgio Fumera, Ignazio Pillai, Fabio Roli, Spam Filtering Based On The Analysis Of Text Information Embedded Into Images. Proceedind of

ISSN 2348 - 7968

Journal of Machine Learning Research 7 (2006) 2699-2720.
[4] Gruber, T. What is an Ontology? http://wwwksl. stanford.edu/kst/what-is-an-ontology. html.
[5] Gunther, O. Environment information systems. Springer, 1998.
[6] Ferris Research. Spam Control: Problems & Opportunities, 2003.
[7] Jangbokim, Kihyun Chung, And Kyunghee Choi, Spam Filtering With Dynamically Updated URL Statistics. Published by the ieee computer society 1540-7993/07/$25.00 © 2007 IEEE.
[8] International Data Group. Worldwide email usage 2002-2006: Know what's coming your way, 2002.
[9] Lourdes Araujo and Juan Martinez-Romo, Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models. Proceeding of IEEE Transactions on information forensics and security, vol.5, no. 3, September.
[10] Pundt, H., and Bishr, Y. Domain ontologies for data sharing: An example from environmental monitoring using field GIS. Computer & Geosciences, 28, 98-102, 1999.
[11] Seongwook Youn, Dennis McLeod, Spam Email Classification using an Adaptive. Proceedind of Journal of Software, vol.2, no.3, September.
[12] 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.
[13] Steve Webb, Calton Pu, Predicting Web Spam with HTTP Session Information. CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
[14] 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.
[15] Taiwo Ayodele, Shikun Zhou, Rinat Khusainov,Email Classification Using Back Propagation TechniqueInternational Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 1/2, March/June 2010.
[16] Susan Gauch, Jason Chaffee, Alexander Pretschner, Ontology-Based User Profiles for Search and Browsing.
[17] Shankar, S., and Karypis, G. Weight adjustment schemes for a centroid based classifier. Computer Science Technical Report TR00-035, 2000
[18] Youn, S., and McLeod, D. Efficient Spam Email Filtering using an Adaptive Ontology. Proceedings of 4th International Conference on Information Technology: New Generations (ITNG07), Las Vegas, NV, April, 2007.