Creating a Model for Preventing Semantic Email Violations

Mr.Dimpu Sagar N
PG Scholar,Dept of CSE
Rajarajeshwari College of Engineering.
Bangalore, India.

Dr.Usha Sakthivel
HOD,Dept of CSE
Rajarajeshwari College of Engineering.
Bangalore, India.

Abstract - **Email is one of the most important Internet applications for most of the computer users. The usage of email is increasing from time to time. However, together with this growth comes a various problems, such as increase of spam and the widely spread of computer worms via emails. The current scenario poses a challenge on ways to manage email efficiently, especially in avoiding the sending and receiving of spam email. Recent development such as Semantic Web in web technology provides an infrastructure that enables web pages, databases, and services to both consume and produce data on the web. Application developer can use this information to search, filter, and prepare information in new and exciting ways to assist the web user. With these features, Semantic Email, one of Semantic Web application, is believed to be able to help control spam by using semantic filtering and filing of email. This paper reports the current practice of email violation prevention method by Internet Service Providers in Malaysia, and later proposes a measure to prevent email violation in Malaysia through Web 3.0 technology. The development of an email violation prevention model based on Web 3.0.**

I.INTRODUCTION

There has been much alarm about Internet abuse in the past decade. Claims of Internet-related crimes such as homicides, suicides, and child neglect have received widespread media attention across the globe. Almost 10 percent of adult Internet users are identified as Internet addicts and 31 percent of Facebook users admitted that they are addicted to Internet applications.

In India, according to local Internet Service Provider Jaring, reported the recent abuse of the Internet, in which 38 computer servers belonging to local educational, government and private organizations were hacked and used as the launching pads for the online abuse in 1999.

The Web technology is evolving day by day. Starting with Web 1.0, the first generation of web technology was static and read-only applications that followed strict categorization and naming of web element representations. At the transaction level, these applications are a complete backward and forward communication to the server for each of the user requests. In other words, each request had to be serviced by the server.

Web 2.0 aims to enhance creativity, information sharing and collaboration among users. User participation is exaggerated through user's involvement in the application, ability to contribute and share interest in the group that they are associated. Control over data hosted on the application provides scope for creative data management, usage and representation. This transforms application usage experience from a static to a participative operator of the application.

The current new generation of Web applications is Web 3.0. Web 3.0 extends Web 2.0 applications by completely upsetting the technology of the traditional computer application industry. Major web sites with the Web 3.0 technology will undergo transformation into web services and will effectively expose their information to the community.

This study aims to explore the current practice of Internet Service Providers (ISP) in India on handling email violation problem and their prevention method of the problem. Based on the information gathered, an email violation prevention model is developed by integrating Web 3.0 technology into the current practice framework.

II.METHODOLOGY

The paradigm of inquiry for this research is positivist and the strategy of inquiry is qualitative. The interviews have been conducted at the selected ISPs in India. The selection of the ISPs are based on their popularity among the Internet users in India, categorized by cellular broadband and digital subscriber line (DSL) as well as technology and package download speed provided.

Based on the ISPs feedback, they key point of email violation prevention is ‘filtering’. The concept of semantic email is same with the current prevention; where semantic email filters the email that is going to send out to the receiver and incoming email in the mailbox. Currently, most of the ISPs are using off the shelf filtering software and hardware to filter the incoming emails to their organization. With the input and filtering concept by ISPs, and Web 3.0 elements, an email violation prevention model based on Web 3.0 had been constructed.

III.CURRENT PRACTICE BY ISPs

Email violation definition by Internet Service Providers (ISP) is slightly different from the definition that has been found in literature review of this study. Most of the major ISP in India agreed that the email that has been categorized as email violation depends on the receiver of the email themselves. The receiver will decide whether the email that they received is an unwanted email or not and proceed to make a report to ISP or Cyber Security India. The ISP will proceed with necessary procedure to prevent the violation based on the case reported by the users.

The most common email violation cases reported to Internet Service Provider (ISP) is email spamming. The Internet user will log a report to their respective ISP if they feel that the email that they received is a spam email. For an ISP to take action on the case reported, two important information are needed, the contents of the email and the email header. The descriptions of these are elaborated in the following sub-sections.

A. Email Content

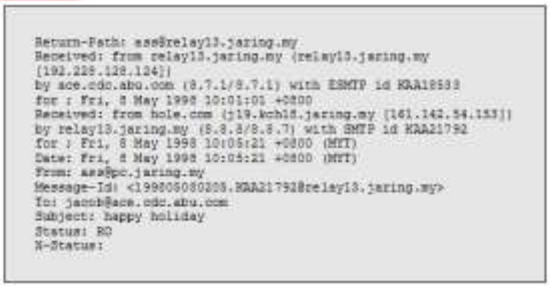
Email content is first source needed by ISP to identify whether the email received by their customer is really a spam email or not. From some of the case reported to ISP, the email is not containing any contents that harm the recipients.

The ISP will contact the individual who reported the case and get their reason why they want the ISP to take action to the mail sender. For further action, ISP will get email information from the email header.

B. Email Header

Email header (Figure 1) is the most important information to ISP to investigate the email violation case reported to them. The header contains the “name” and “address” of the sender, recipient and anyone who is being copied, the “date” and “time” the mail is sent and the “subject” of the mail. The function of the header is for the computer to route mail to the receiver. The “received:” item indicates the mailers. It shows what mailers the mail is routed through before it goes to the recipient. Usually, over the Internet, the mail will go through several mailers before it finally reaches the recipient. This information will help in tracing the source IP address of the sender.

According to the ISP, there are two types of email spammer which is real spammer and machine spammer. Real spammer means the individual that have intentioned sending email to unknown person and the receiver is not willing to receive the email. Machine spammer means the spam email sent by the spy boot or spyware and the process is hidden from the computer user. If this happened, the ISP needs to educate their customer on how to scan and clean up their computer.



```
Return-Path: ass@relay13.jaring.my
Received: from relay13.jaring.my (relay13.jaring.my
[192.228.128.124])
by ace.cdc.abu.com (8.7.1/8.7.1) with ESMTP id KAA18533
for ; Fri, 8 May 1998 10:01:01 +0800
Received: from hole.com (j19.kch18.jaring.my [161.142.54.153])
by relay13.jaring.my (8.8.8/8.8.7) with SMTP id KAA21792
for ; Fri, 8 May 1998 10:05:21 +0800 (MYT)
Date: Fri, 8 May 1998 10:05:21 +0800 (MYT)
From: asa@pc.jaring.my
Message-Id: <199805080205.KAA21792@relay13.jaring.my>
To: jacob@ace.cdc.abu.com
Subject: happy holiday
Status: RO
X-Status:
```

Figure 1. A full email header

IV.CURRENT EMAIL VIOLATION REVENTION METHOD BY ISPs

To prevent email violation in their own organization, most of the ISPs are using off the shelf software and hardware to make sure their organization is free from email violation. The way it works is almost same and it depends on the organization budget.

Basically the incoming emails will be filtered by the filtering application that has been subscribed by the organization. Most common applications are rate controls, IP analysis, sender authentication, recipient verification, virus scanning, custom policy, image analysis, and spam scoring. Figure 2 illustrates the architecture of spam firewall with the filtering application located in the defence layer.

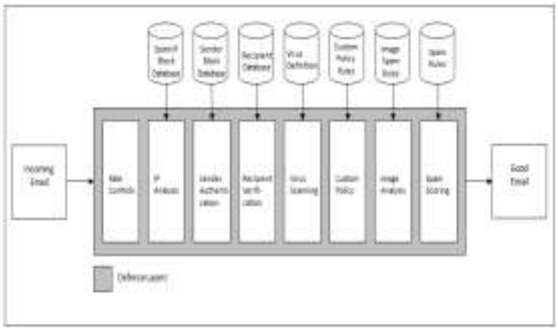


Figure 2. Spam Firewall Architecture

V.WEB 3.0 AS A MEASURE TO PREVENT EMAIL VIOLATION

The current email violation prevention by ISP is done by filtering the incoming email to their server. The concept of semantic email is also same with the current prevention, which is by filtering the email that going to send to the receiver and filtering incoming email in the mailbox. Semantic technologies promise a more flexible representation than XML-based technologies. The approach is based on Web 3.0 elements, namely RDF (Resource Description Framework), WOL (Web Ontology Language), SPARQL (SPARQL Protocol and RDF Query Language) and FOAF (Friend of a Friend). The ontology describes the main concepts of the domain and their interrelationships with a <subject-predicate-object> structure for each of the concepts. Below is a brief description of each element:

A. OWL

OWL is a family of knowledge representation languages for authoring ontologies, and is endorsed by the W3C. This family of languages is based on two semantics: OWL DL and OWL Lite semantics are based on Description Logic, which have attractive and well understood computational properties, while OWL Full uses a semantic model intended to provide compatibility with RDF Schema. OWL ontologies are most commonly serialized using RDF/XML Syntax. OWL is considered one of the fundamental technologies underpinning the Semantic Web.

B. RDF

RDF is a family of W3C specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modelling of information that is implemented in web resources; using a variety of syntax formats. RDF data model is not different from classic conceptual modelling approaches such as Entity - Relationship or Class diagrams, as it is based upon the idea of making statements about Web resources, in the form of subject-predicate-object expressions. These expressions are known as triples in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object [10].

For example, one way to represent the notion “The lecturer with qualification PhD” in RDF is as the triple: a subject denoting “the lecturer”, a predicate denoting “with qualification”, and an object denoting “PhD”. RDF is an abstract model with several serialization formats (i.e. file formats), and so the particular way in which a resource or triple is encoded varies from format to format.

C. SPARQL

SPARQL is a computer language that able to retrieve and manipulate data stored in RDF format. SPARQL can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions. SPARQL also supports extensible value testing and constraining queries by source RDF graph. The results of SPARQL queries can be results sets or RDF graphs.

D. FOAF

FOAF is a machine-readable ontology describing persons, their activities and their relations to other people and objects. Anyone can use FOAF to describe him or herself. FOAF allows groups of people to describe social networks without the need for a centralized database.

FOAF is a descriptive vocabulary expressed using the RDF and the WOL. Computers may use these FOAF profiles to find, for example, all lecturers in India, or to list all people both you and a friend of yours know. This is accomplished by defining relationships between people. Each profile has a unique identifier (such as the person's e-mail addresses, a staff ID, or a URI of the homepage or weblog of the person), which is used when defining these relationships. The FOAF dataset follows the Linked Data paradigm and participates in the Linking Open Data project by linking to other datasets.

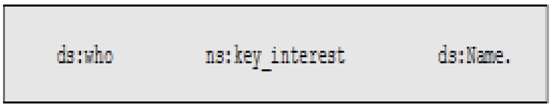
We can still maintain the current prevention because it is applied on the network area while semantic filtering applied on the email application itself. The overall concept of prevention model based on Web 3.0 and the prevention model is discussed in chapter 5.

VI.EMAIL VIOLATION PREVENTION MODEL BASED ON WEB 3.0

The email violation prevention model based on Web 3.0 will be applied in the email application. Mail sender will filter the receiver by criteria (e.g. name, faculty, academic interest). No email address will be displayed to prevent it spread out and misuse by anonymous. Once the mail receiver list generated, the mail sender will compose the email and send it. Before the sent email reach the receiver inbox, a personal filtering that has been preset by the receiver will filter the email. If the sent mail match with the preset criteria, it will store in the inbox else it will store in transit box. Transit box is an inbox for the email that has been sent by the recognized sender but not meet the filtering criteria. A notification email also will be sent to sender to notify that the email that has been sent earlier is not match with the receiver's filtering criteria.

Assume that the entire name list together with their profiles, email address and interests (objects in the domain) are stored in the database. We call the data in the database the static facts. The ontology is used in conjunction with domain rules to generate derived facts based on the data of the domain. We use CWM (Closed World Machine), a description logic based tool of W3C, to do the rule evaluation and application, and thus the Domain Ontology is created. The static facts and the derived facts are then converted into RDF. This Domain Ontology is used by the email system to filter the receiver list.

The schema of the ontology helps define domain specific relations that are relevant and semantically meaningful for the domain. For example, we have defined schemas called "ds" and "ns" with relations that can be used to state triples such as in Figure 3.



```
ds:who          ns:key_interest          ds:Name.
```

Figure 3. Domain specific relations

The semantics of the relation "key_interest" is specific to a domain and helps the natural language engine to filter based on the semantics provided by the relation.

We use SPARQL to query the RDF data in memory and to fetch relevant data. SPARQL provides the ability and the flexibility to perform generic queries. The general structure of the query is (subject, predicate, object). We have identified seven types of queries for the subject-predicate-object (hence forth referred to as <s-p-o>) structure of our ontology; these are: s (only subject); p (only predicate); o (only object); s-p (subject and predicate); s-o (subject and object); p-o (predicate and object); s-p-o (subject, predicate and object specified). After the concepts are identified from the input sentence, the concepts are classified as subject, predicate or object. The actual query is formulated by binding the value of the concept raised (and classified as s-p-o) in the input sentence to the generic SPARQL query that is one of the above seven types, in order to formulate the precise query and filter the recipients.

FOAF predicates a person's express properties, such as name, email address, gender, birthday, interests, projects, and associates. By spidering the Semantic Web and collecting the information contained in FOAF files, we can build a collection of data about people and their interests. This information can be used to email people with a given interest, who know people who know a particular person, and so on. Figure 4 illustrates the email sending process and relates it with Semantic Web elements.

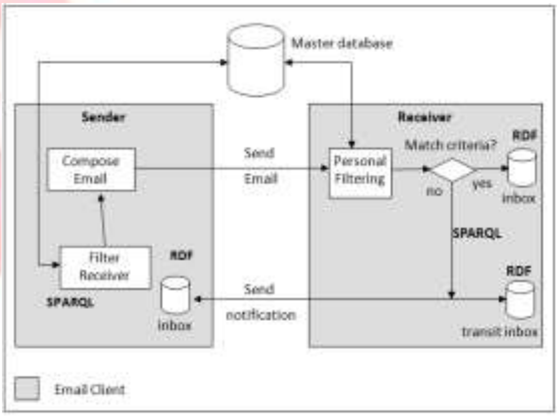


Figure 4. Email violation prevention model by filtering the receiver of the e-mail based on preset criteria.

The overall view of the proposed email violation prevention model together with the current prevention model is illustrated in Figure 5.

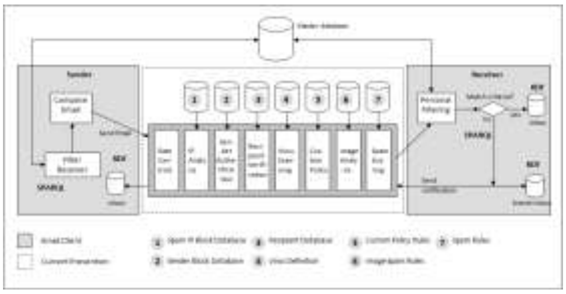


Figure 5. Overall view of the proposed email violation prevention model together with the current prevention model.

VII.CONCLUSION

This seminar discusses the email violation definition in the perspective of ISP. The definition is slightly different from what we got in the literature review since it is depends on the customer of the ISP. The current practice by the ISP in handling the email violation cases by their customer is by studying the email contents. If the email is categorized as spam email, the ISP will check the full email header to get the more information such as IP address and routing information. For email violation prevention in their own organization, the ISPs are using off the shelf software and hardware. The number of software used is depends on the organization budget. An email violation prevention model based on Web 3.0 has been constructed and explains in this chapter. The model is adapting the Web 3.0 elements to prevent email sender sends email to the wrong recipients.

REFERENCES

[1] Cooper, A., Morahan-Martin, J., Mathy, R. M., & Maheu, M. (2002) “Toward an increased understanding of user demographics in online sexual activities.” Journal of Sex & Marital Therapy, 28, 105-129

[2] Vanden Boogart MR (2006) “Uncovering the social impacts of Facebook on a college campus”, Retrieved March 12, 2009 from http://hdl.handle.net/2097/181.

[3] Chia, A. “Number of Internet abuse cases quadruples to 2”, 106 New Straits Times, 10th October 2003, Retrieved April 1, 2009 from http://www.cybersecurity.org.my/en/knowledge_bank/news/2003/main/detail/919/index.html.

[4] Shannon, V. (2006) “A 'more revolutionary' Web”, Retrieved March 15, 2009 from http://www.nytimes.com/2006/05/23/technology/23iht-web.html

[5] Kiefer, C. (2007) “Imprecise SPARQL: Towards a Unified Framework for Similarity Based Semantic Web Tasks”, Department of Informatics, University of Zurich.

[6] Hendler, J. (2009). “Web 3.0 Emerging. Digital Object Identifier”, 111-113

[7] Fensel, D. (2002). “Web Intelligence”, IEEE/WIC/ACM International Conference on Dital Object Identifier.

[8] Dumbill, E. (2002). “XML Watch: Finding friends with XML and RDF” Retrieved April 28, 2009 from http://www.ibm.com/developerworks/xml/library/x-foaf.html.

[9] Smith, Michael K.; Chris Welty, Deborah L. McGuinness (2004) “OWL Web Ontology Language Guide”. Retrieved July 27, 2009 from http://www.w3.org/TR/owl-guide/

[10] Harth, A., Decker, S. (2005). “Optimized Index Structures for Querying RDF from the Web”, 71-80.