Semantics-Based Spam Detection by Observance of Outgoing Message

C.Selva Karthika
PET Engineering College, India.

Abstract—The existing spam detection system are mostly keyword-based and find the spam message present in the outgoing message by matching the keyword. The quality of result provided by traditional keyword-based spam detection is not optimal for finding the spam information present in the message. The semantic based spam detection can provide efficient solution for finding spam information present in the outgoing message. This paper explains how semantic approach can be used for detecting spam information present in the outgoing message. This paper also defines a general framework by which spam detection can be made by monitoring outgoing message. This framework is very much useful for various approach such as tokenization, stopword removal, semantic checking and information retrieval. Thus it supports gradual transition from keyword based spam detection to semantic based ones.

Keywords—Information retrieval, Semantic web, Tokenization, Stop word.

I. INTRODUCTION

Spam has become an enormous problem affecting Internet users and broadband service providers. Well-known viruses, worms, and Trojan horses get the headlines, but spam is arguably a more pervasive and insidious threat because it affects every Internet user-directly or indirectly-and it lacks a comprehensive solution analogous to antivirus software programs. Spam frustrates users by overloading their e-mail boxes with volumes of useless and unwanted messages.

Thus we focus on the detection of the Spam message generating computers in a network that are involved in the spamming activities through Semantic Web Technology.

The existing spam detection systems are mostly keyword-based and identify relevant emails. Keyword-based search, in spite of its merits of expedient query for information and ease-of-use, has failed to represent the complete semantics contained in the content and has led to the following problems (1) keywords could represent only fragmented meanings of the content, and the content identified through keywords did not always meet the exact spam mails. (2) Due to synonym and polysemy in human language, spam detection through keywords can only cover information containing the same keyword, while other information with similar semantics but different keywords has been completely left out.

Semantic search uses semantics, the science of meaning in language, to detect highly accurate compromised machines in the network. Here, WordNet is used to get the semantics of the spam word.

II. RELATED WORKS

Ming-Yen Chen et al. (2009) introduce a semantic enabled information retrieval in which a web corpus is taken and the related information is retrieved. The limitation of this work is that it didn't deal with the Synonyms or Synsets.

Zongli Jiang et al. (2009) introduce the concept of category attribute of a word. Here, the useless results can be removed from the search results thereby improving the retrieval efficiency. Latent Semantic Analysis is a method that can discover the underlying semantic relation between words and documents. Singular Value Decomposition is used to analyze the words and documents and get the semantic relation finally.

Gang et al. (2009) proposed a method to enhance the information retrieval recall and precision. To filter out the document which has smaller related degree with original query, the scores of search results document is re-calculated by use of ontology semantic similarity.

Hongwei Yang et al. (2010) enable the users to find the relevant documents more easily and also help users to form an understanding of the different facets of the query that have been provided for web search engine. A popular technique for clustering is based on K-means such that the data is partitioned into K clusters. In this method, the groups are identified by a set of points that are called the cluster centers. The data points belong to the cluster whose center is closest.

III. PROPOSED SYSTEM

The proposed system uses semantic based technique to find the spam information present in the outgoing message.

The user query is first processed by message preprocessing module in which various techniques such as stop word list removal and stemming is done and the resulting output is sent to semantic extraction and spam identification module. In this module, the semantic element and their semantic relations are analyzed by using wordnet. The wordnet is used to get synsets related to the spam identified in the outgoing message.
IV. IMPLEMENTATION

The system is implemented by using a corpus of 100 spam related word. The email is given as input and the processing steps are explained below. The final output obtained is the identification of spam information present in the message and the source which send the spam message. The implementation steps are

- Message Preprocessing
- Semantic Extraction
- Spam Identification

1) Message Preprocessing

In this module, the given email message is analyzed with the words present in corpus directors. It involves 2 steps

- Tokenization
- Stopword removal

Tokenization is the process of converting the user input into basic terms format. This involves splitting the E-mail content into words by removing symbols, punctuations, etc.

Stopword list removal is done by removing irrelevant terms like pronouns, articles and symbols and to convert variants of verbal nouns and participles into their original form for the purpose of reducing the volume of terms being processed. Since topics in a specialized field are usually expressed in nouns while their associations are expressed in verbs, it is therefore necessary to retrieve nouns and verbs from these processed terms with part of speech analysis before proceeding with the subsequent content summarization phase.

1) Semantic Extraction

- E-mail refinement

The e-mail is passed through stopword list to remove stop words. Then, stemming is done to retrieve only subject. This is passed to the WordNet to get more senses. For example, the word ‘Software’ has 3 senses such as software, package and program. In the keyword based search, only the
word ‘software’ will be taken but not its senses. Hence, different words expressing the same meaning will not be taken and so all the spam mail won’t be detected. Hence, the tokens of the e-mail are passed to the WordNet and more senses are considered by the search engine in semantic e-mail refinement.

![Fig 5: Semantic Extraction and Determination](image)

2) Synsets Generation

The goal of WordNet is the creation of dictionary and thesaurus which could be used intuitively. The next purpose of WordNet is the support for automatic text analysis and artificial intelligence. WordNet is a lexical database for English language. It groups English words into sets of synonyms called Synsets, provided short, general definitions and records the various semantic relations between these synonym sets.

The purpose is two-fold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet distinguishes between nouns, verbs, adjectives and adverbs because they follow different grammatical rules. Every Synset contains a group of synonymous words or collocations (a collocation is a sequence of words that go together to form a specific meaning such as ‘car pool’); different senses of a word are in different Synsets.

3) Spam Identification

The output of the semantic pattern generated should be given as input for the spam identification module. Here, the machine in the network is assumed to be either having spam or no spam. The machine is compromised if it involved in spamming. Let \( X_i \) for \( i = 1, 2, \ldots \) denote the successive observations of a random variable \( X \) corresponding to the sequence of messages originated from semantic pattern. We let \( X_i = 1 \) if message \( i \) from the machine is a spam, and \( X_i = 0 \) otherwise. The identification module assumes that the behavior of a compromised machine is different from that of a normal machine in terms of the messages they send. Specifically, a compromised machine will with a higher probability generate a spam message than a normal machine. Formally,

\[
P_r(X_i = 1|H_1) > P_r(X_i = 1|H_0),
\]

where \( H_1 \) denotes that machine \( m \) is compromised and \( H_0 \) that the machine is normal. The spam identification problem can be formally stated as follows: as \( X_i \) arrives sequentially at the detection system, the system determines with a high probability if machine \( m \) has been compromised. Once a decision is reached, the identification module reports the result, and further actions can be taken, e.g., to clean the machine.

V. CONCLUSION

In this study, a semantic based spam detection system is developed. In addition to semantic-based spam zombie detection system, the proposed system has significant novelties: a semantic extraction and determination model which employs WordNet to generate more semantics, thereby solving the problem of inaccurate spam detection.

REFERENCES


[10] Wei-Dong Fang, Ling Zhang, Yan-Xuan Wang, Shou-Bin Bong, “Toward a semantic search engine based on ontologies” Network Engineering and Research center, South China University of Technology, Guangzhou 510640, China.

