A NEW APPROACH TO SPAM MAIL DETECTION

R.JENSI
Lecturer (CSE),
Dr.Sivanthi Aditanar College of Engineering,Tiruchendur.Tamilnadu,India
E_mail id: r_jensi@yahoo.co.in

Abstract-The ever increasing menace of spam is bringing down productivity. More than 70% of the email messages are spam, and it has become a challenge to separate such messages from the legitimate ones. I have developed a spam identification engine which employs naive Bayesian classifier to identify spam. A new concept-based mining model that analyzes terms on the sentence, document is introduced. . The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. The proposed mining model consists of sentence-based concept analysis, document-based concept analysis similarity measure. In this paper, a machine learning approach based on Bayesian analysis to filter spam is described. The filter learns how spam and non spam messages look like, and is capable of making a binary classification decision (spam or non-spam) whenever a new email message is presented to it. The evaluation of the filter showed its ability to make decisions with high accuracy. This cost sensitivity was incorporated into the spam engine and I have achieved high precision and recall, thereby reducing the false positive rates.

Keywords-Spam, Bayesian filter, concept-based mining model

I. INTRODUCTION

In today's highly technical world and our computer-connected society, email has become the fastest and most economical form of communication available. Spam is an unsolicited email that is sent indiscriminately to mailing lists, individuals and newsgroups. This misuse of the electronic message system is becoming rampant as spamming is economically feasible. A recent study says that more than 70% of the total messages that are sent over the internet are spam. Many anti-spam researchers believe that spam is responsible for anywhere from 35% to 65% of all email traffic on the Internet today, with a whopping annual growth rate of 15% to 20%. Many are concerned that spam could be the end of email.

Emails can be of spam type or non-spam type as shown in the Fig.1. Spam mail is also called as junk mail or unwanted mail whereas non-spam mails are genuine in nature and meant for a specific person and purpose. Information retrieval offers the tools and algorithms to handle text documents in their data vector form. The Statistics of spam are increasing in number. At the end of 2002, as much as 40 % of all email traffic consisted of spam. In 2003, the percentage was estimated to be about 50 % of all emails. In 2006, BBC news reported 96 % of all emails to be spam.

In the past few years, a number of anti-spam techniques have been proposed and used to combat spam such as whitelisting, blacklisting, and greylisting of domain names, keyword-based filtering, heuristic-based filtering, etc. All of these techniques, however, require heavy maintenance and can not achieve very high overall accuracy.

Most current spam filtering methods are based on the Vector Space Model (VSM), which is a widely used data representation for text classification and clustering. The VSM represents each document as a feature vector of the terms (words or phrases) in the document. Each feature vector contains term weights (usually term frequencies) of the terms in the document. The similarity between the documents is measured by one of several similarity measures that are based on such a feature vector. Examples include the cosine measure and the Jaccard measure.

Usually, in text mining techniques, the term frequency of a term (word or phrase) is computed to explore the importance of the term in the document. However, two terms can have the same frequency in their documents, but one term contributes more to the meaning of its sentences than the other term.

The concept-based model captures the semantic structure of each term within a sentence and document rather than the frequency of the term within a document only. Each sentence is labeled by a semantic role labeler that determines the terms which contribute to the sentence semantics associated with their semantic roles in a sentence. Each term that has a semantic role in the sentence, is called a concept. Concepts can be either words or phrases and are totally dependent on the semantic structure of the sentence. When a new email is introduced to the system, the proposed mining model can detect a concept match from this email document to all the previously processed email documents in the data set by scanning the new email document and extracting the matching concepts.

A classifier is used to classify a document to be either spam or non-spam accurately. This paper deals with the possible improvements gained from differing classifiers used for a specific task. Statistical-based Bayesian filters have become a popular and important defense against spam. Bayesian filters usually perform a dictionary lookup on each individual token and summarize the result in order to arrive at a decision.

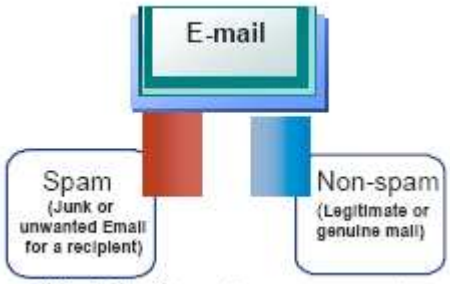

Fig. 1 : Email types (spam or non-spam)

II. Our Overall Acceleration Comes From Four Improvements:

A. Parsing:

The parsing function of the developed algorithm is as follows,

- Separate sentences
- Label terms
- Remove stop words
- Stem words

B. Concept based analysis:

In the concept0based mining model, both the verb and the argument are considered as terms. One term can be an argument to more than one verb in the same sentence. This means that this term can have more than one semantic role in the same sentence. In such cases, this term plays important semantic roles that contribute to the meaning of the sentence. In the concept-based mining model, a labeled terms either word or phrase is considered as concept.

C. Similarity measure:

A concept-based similarity measure, based on matching concepts at the sentence, document rather than on individual terms (words) only, is devised. The concept-based similarity measure relies on three critical aspects. First, the analyzed labeled terms are the concepts that capture the semantic structure of each sentence. Second, the frequency of a concept is used to measure the contribution of the concept to the meaning of the sentence, as well as to the main topics of the document. Last, the number of documents that contains the analyzed concepts is used to discriminate among documents in calculating the similarity.

D. Classification:

Bayesian filters combine all the concepts statistics of an incoming message to an overall score by a Bayesian probability calculation. Finally, filtering decision is made based on the score and a predefined threshold.

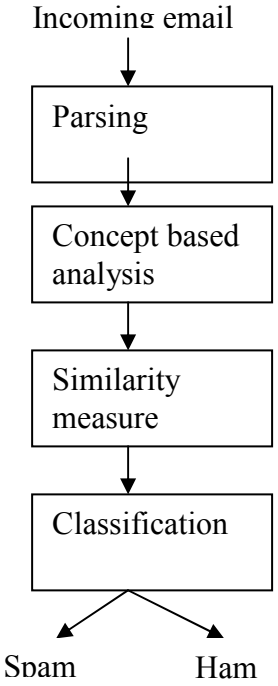

Fig 2. Stages of filtering process

III. Related Work

A. CONCEPT-BASED MINING MODEL

This paper uses the concept-based model [1] for labeling terms based on semantic structure of sentences and documents.

The concept-based mining model consists of sentence-based concept analysis, document-based concept analysis similarity measure, as depicted in Fig. 2. An email text document is the input to the proposed model. Each document has well-defined sentence boundaries. Each sentence in the document is labeled automatically based on the PropBank notations. After running the semantic role labeler, each sentence in the document might have one or more labeled verb argument structures. The number of generated labeled verb argument structures is entirely dependent on the amount of information in the sentence. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based mining model on sentence, document levels.

The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document levels rather than a single-term analysis on the document only.

B. Sentence-Based Concept Analysis

To analyze each concept at the sentence level, a new concept-based frequency measure, called the conceptual term frequency ctf is proposed. The ctf calculations of concept c in sentence s and document d are as follows:

C. Calculating ctf of Concept c in Sentence s

The ctf is the number of occurrences of concept c in verb argument structures of sentence s. The concept c,

which frequently appears in different verb argument structures of the same sentence s, has the principal role of contributing to the meaning of s. In this case, the ctf is a local measure on the sentence level.

D. Calculating ctf of Concept c in Document d

A concept c can have many ctf values in different sentences in the same document d. Thus, the ctf value of concept c in document d is calculated by:

$$ctf = \frac{\sum_{n=1}^{sn} tf_n}{sn}, \qquad (1)$$

where sn is the total number of sentences that contain concept c in document d. Taking the average of the ctf values of concept c in its sentences of document d measures the overall importance of concept c to the meaning of its sentences in document d. A concept, which has ctf values in most of the sentences in a document, has a major contribution to the meaning of its sentences that leads to discover the topic of the document. Thus, calculating the average of the ctf values measures the overall importance of each concept to the semantics of a document through the sentences.

E. Document-Based Concept Analysis

To analyze each concept at the document level, the concept based term frequency tf, the number of occurrences of a concept (word or phrase) c in the original document, is calculated. The tf is a local measure on the document level.

The concept-based analysis algorithm describes the process of calculating the ctf, tf , and df of the matched concepts in the documents. The procedure begins with processing a new document which has well-defined sentence boundaries. Each sentence is semantically labeled according to [23]. The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations. Each concept (in the for loop, at line 5) in the verb argument structures, which represents the semantic structures of the sentence, is processed sequentially. Each concept in the current document is matched with the other concepts in the previously processed documents. To match the concepts in previous documents is accomplished by keeping a concept list L, which holds the entry for each of the previous documents that shares a concept with the current document.

After the document is processed, L contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, L is output as the list of documents with the matching concepts and the necessary information about them. The concept-based analysis algorithm is capable of matching each concept in a new document d with all the previously processed documents in O(m) time, where m is the number of concepts in d.

F. A Concept-Based Similarity Measure

A concept-based similarity measure, based on matching concepts at the sentence, document rather than on individual terms (words) only, is devised.

The concept-based measure exploits the information extracted from the concept-based analysis algorithm to better judge the similarity between the documents

$$sim_c(d_1, d_2) = \sum_{i=1}^{\cdots} max\left(\frac{t_{i_1}}{Lv_{i_1}}, \frac{t_{i_2}}{Lv_{i_2}}\right) \times weight_{i_1} \times weight_{i_2}, \qquad (2)$$

$$weight_i = (tf\ weight_i + ctf\ weight_i) * log\left(\frac{N}{df_i}\right). \qquad (3)$$

5. Statistical Bayesian Filtering Algorithms :

1. 2.1. Naive Bayes (NB) Algorithm :

Naive Bayes algorithm is the simplified version of Bayes theorem with the assumption of feature independence. It computes the probability of a Class {Spam, Legitimate} .The statistic in which we are mostly interested for a token T is its spamminess, calculated as follows:

$$S[T] = \frac{C_{spam}(T)}{C_{spam}(T) + C_{ham}(T)} \qquad (4)$$

where $C_{spam}(T)$ and $C_{ham}(T)$ are the number of spam or ham messages containing concepts T, respectively.

CONCLUSION:

This paper has been developed keeping in mind today's fast changing world. In this paper, an email clustering method is proposed to efficient detect the spam mails. The proposed technique includes the concept-based mining for filtering spam email. Bayesian classifier is used to efficient filtering.

REFERENCES:

1. Shady Shehata, Fakhri Karray and Mohamed S. Kamel, Fellow, "An Efficient Concept-Based Mining Model for Enhancing Text Clustering ", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010.

2. Paul Graham. A plan for spam. In Reprinted in Paul Graham, Hackers and Painters, Big Ideas from the Computer Age, O Really, 2004,2002.

2. David D. Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In Proceedings of ECML-98, 10th European Conference on Machine Learning, number 1398, pages 4{15. Springe Verlag, Heidelberg, DE, 1998.

3.Gary Robinson. A statistical approach to the spam problem. Linux J.,2003(107), March 2003. Gary Robinson. Spam detection, January 2006.

4. Hall, R.J. How to Avoid Unwanted Email.41(3):88–95, 1998.

5. Ahmed Obied. Bayesian Spam Filtering.