Classifying and Identifying of Threats in E-mails – Using Data Mining Techniques

D.V. Chandra Shekar and S.Sagar Imambi

Abstract—**E-mail has become one of the most ubiquitous methods of communication. The large percentage of the total traffic over the internet is thee-mail. E-mail data is also growing rapidly, creating needs for automated analysis. So, to detect crime, to organize bundles of emails, a spectrum of techniques should be applied to discover and identify patterns and make predictions. Data mining has emerged to address problems of understanding ever-growing volumes of information for structured data, finding patterns within data that are used to develop useful knowledge. Earlier Statistical methods are used to characterize user behavior , classifying spam and detecting novel email viruses. However, previous techniques have not examined the contributions of these features to their classification and they need some improvements. By applying data mining classification techniques we can achieve this very efficiently. In this paper we show that Naïve Bayes classification approach is useful for predicting user's behavior and to organize the emails according to users constraints**.

Index Terms— **Data mining, email classification, Organization of emails, Naïve Bayes classification model**

I. Introduction

Electronic mail has become one of the most ubiquitous methods of communication. It is a fast, efficient and an inexpensive method of reaching out to a large number of people at the same time. The average user is often overwhelmed by the sheer amount of email sent and received.

A large part of the Internet mail traffic comprises of unsolicited bulk email or spam as it is popularly, rather notoriously known. E-mail management is a significant and growing problem for individuals and organizations. Users often find themselves expending large amounts of time and effort sifting through the mass of email messages and classifying them to their corresponding folders. Unsolicited marketing messages, or spam, account for more than half of the total daily message traffic (Message Labs, 2004). In addition, at least eight out of the ten computer worms most frequently reported during 2004 to a prominent anti-virus company spread via email (Sophos Corporation, 2004).

As a result of this, efficient automated methods for analyzing the content of e-mail messages and identifying threats from these messages are becoming imperative. The time spent in this task can be greatly reduced if traditional classification techniques can be adapted to email classification thereby automating the process with sufficient accuracy and efficiency [7]. The work done by various researchers suggests that threats are identified by reduced frequency of first-person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs [8]. The main contribution of our study is, we present a new model based on the Naïve Bayesian algorithm for classifying E-mails and identifying threats, from a company's rapidly expanding emails. The approach proposed herein is a novel one and is based on the use of mining techniques for email classification.

2. RELATED WORK

Rambow et al. apply a machine learning approach to email summarization [9]. They use RIPPER as a classifier to determine which sentences should be included in a summary. Features used for learning include linguistic features, and features describing the email and the threading structure. Such an approach requires a large number of positive examples and cannot produce summaries with varying length

Manuscript received January 13, 2008. Reviewed and accepted on January 17,2007
F. Author is Assistant Professor, Department of Computer Science , TJPS College (P.G Courses), Guntur, Andhra Pradesh–522006,INDIA,
Phone : 9440454033;e-mail : chand.info@gmail.com
S. Author is Assistant Professor, Department of Computer Science, TJPS College (P.G Courses), Guntur, AndhraPradesh-522006, INDIA
e-mail: simambi@gmail.com

based on the users request. It is also not clear how this approach can handle hidden emails. Section 6 will compare CWS with RIPPER.

Wan et al. study decision-making summarization for email conversations [2]. Email threading is used. Among the various sets of features explored, their experiments show that a centroid based method is effective.

Several systems for automatic e-mail classification have been developed. Cohen [3] used RIPPER to induce keyword-spotting rules. Bayesian approaches have been used e.g. [11] as well as Nearest-neighbor techniques [13]. Previous studies have been limited because they have failed to explore the use of feature selection and phrase representations. In this paper we show how these techniques can be used to increase classification accuracy.

Segal and Kephart [1] use the TF-IDF classifier as the means for classification in *SwiftFile*, which is implemented as an add-on to Lotus Notes. It predicts three likely destination folders for every incoming mail message. The TF-IDF classifier performs well even in the absence of large training data, and the classifier accuracy remains reasonable even as the amount of training data increases, adding to the heterogeneity of a folder. The classifier learns incrementally with every new message that is added or deleted from a folder, eliminating the need for re-training from scratch.

3. ELECTRONIC MAIL MINING

Electronic mail can be viewed as a special type of document as it is primarily text along with some identifying information unique to it (e.g., from, to, subject etc.) Email mining is one type of intelligence mining that companies are already starting to do inside their own firewall. The mining of email actually has a history of sorts, at least in forensics, surveillance, and the government intelligence community.

Fig 1 shows the need for email management

E-mail documents have several characteristics which make authorship categorization challenging compared with longer, formal text documents such as literary works or published articles. Applicability in these datasets with existing classification techniques was limited because the large number of features makes most documents undistinguishable. E-mail documents have generally few sentences /paragraphs, thus making contents profiling based on traditional text document analysis techniques, such as the \bag{of{words" representation more difficult[4].

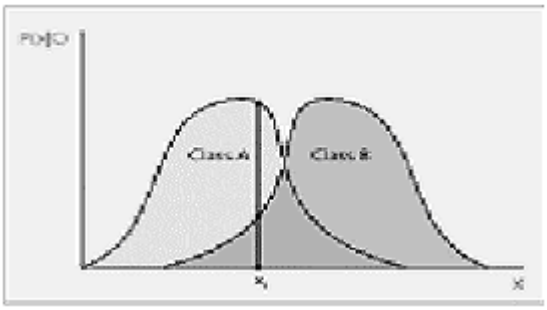
4 THE NAÏVE BAYESIAN CLASSIFICATION

Classification is an important data mining problem. The input is a dataset of training records (also called training database), wherein each record has several attributes. Attribute with numerical domains are numerical attributes and attributes whose domains are non non-numerical are categorical attributes. There is also a distinguished attribute called the class label. This classification aims at building a concise model that can be used to predict the class label of future, unlabeled records. Many classification models including Naive Bayes, Decision tree, support vector machine, and Neural networks have been proposed.[13,14]

Naïve Bayesian algorithms build a model to classify new examples based on observed probabilities and supporting evidence from the training data. The data set for this type of problem is one with input attributes and a known outcome, or class. Figure 4.1 shows the distribution of an input associated with each class, for example, given the variable X with a value at Xi the probability of it being in Class A is greater than the probability of it being in Class B. In

mathematical terms: If one knows how **P(X | C)** and the densities **P(xi)** and **P(cj)** (prior probabilities) are known, then the classifier is one which assigns class **cj** to datum **xi** if **cj** has the highest posterior probability given the data.

Fig 2 Distribution of input



4.1 SELECTED ATTRIBUTES

Number of emails sent: Worms and spam-bots tend to send emails faster than the average user.

Number of unique email recipients: This feature counts addresses in the To:, CC:, and BCC: headers. The frequency with which one sends mail to distinct users captures an important aspect of email behavior.

Number of unique sender addresses: Many users have multiple active accounts on the same machine. However, a single machine sending from a large number of addresses at a high rate could indicate that the machine is compromised.

Average number of words/characters per subject, body; average word length: These features capture trends in email wording that could separate normal email from malicious activity, and among users.

Variance in number of words/characters per subject, body; variance in word length:

These types of features have been used in previous work with some success to detect the behavior of email viruses.

Ratio of emails with attachments: Most users do not send large amounts of consecutive emails with attachments, whereas most worms do.

Topic of email: Subject of the email or more repeated words.

4.2 CLASSIFICATION ARCHITECTURE

In Figure 4.4 we present a simple architecture of text classification systems. There is a pool of emails which represents the content at hand that can either be stored on disk, or could come from data streams or the web. There are standard preprocessing steps applied to this document corpus, followed by an appropriate choice of token models, representation methods, and labeling systems. Classification models are chosen to operate on train-validation-test splits, and classifiers are learned and stored.

E-mail preprocessing: E-mail preprocessing involves the process of transforming the training dataset into a representation suitable for Bayesian algorithm. This stage extracts the informational words from the data set.

Figure 4.2 shows an example e-mail, which includes many typical noises (or errors) for mining.

1. On Wednesday February 2007 13:39:42-0500, "X"
2. YYY@Domain.Com
3. Today there will be bomb blast in parliament house in India at 11.46 am. Stop it if you could. Cut relations with the U.S.A
4.—
5.
---
6. Best regards
7. X
8

Figure 3. Example of e-mail message.

Figure 4.3 shows an ideal output of cleaning of the e-mail in Figure 4.1 within it; the non-text parts like header, signature and quotation have been removed.

1. bomb
2. blast

Figure 4. Cleaned e-mail message.

Text transformations: The conversion of text records to vectors of numeric attributes is a multi-staged process of feature construction which employs several key transformations. There are many approaches to this process – one of the most common is to treat each text record as a "bag of words." Each unique word constitutes an attribute The number of occurrences of a word in a document is the attribute's value for that document. Emails are therefore represented as vectors of numeric attributes where each attribute value is the frequency of occurrence of a distinct term.

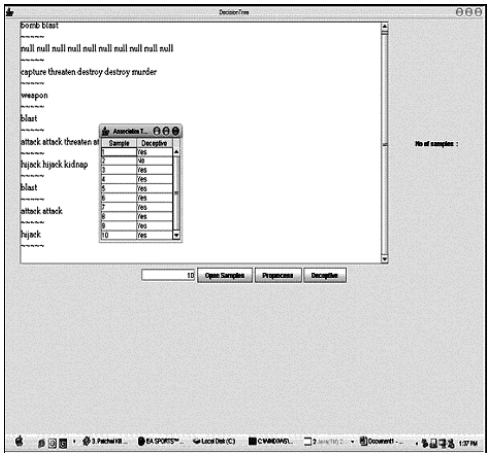


Fig 5 Email Classification process

5 EXPERIMENTAL RESULTS

The application of data mining to the task of e-mail classification is done; experiments were carried out on a small e-mail corpus. In order to conduct an experiment setting, different sets of 2000 e-mails are used: a mixture containing 1000 suspicious e-mails and 1000 normal e-mails. The system was trained with the training dataset and accuracy was calculated. When the training process was finished, the best quality rules were taken as the final classification rules.

Some of the rules generated for threat identification are

Rule 1: If Bomb=Yes then mail=threat

Rule 2: If Bomb=No and Attack=Yes then mail=threat else if Attack=No and Blast=yes the mail=threat

Rule 3: If Blast=No and Hijack=Yes then mail= threat else if Hijack=No and Murder=Yes then mail=threat

Rule 4: If Murder=No and Death=Yes then mail=threat else if Death=No and Terrorist=Yes then mail=threat else if Terrorist=No and Destroy=yes then mail=threat

This is the output that is obtained during the execution:

After classification of emails(trained data set) according to the specified attributes we got the results as

Table 1 classification result based on selected attributes

Topic Category	Author Category ACi (i = 1; 2; 3)			Topic Total
	Author AC1	Author AC2	Author AC3	
Movie	15	21	21	59
Food	12	21	25	58
Travel	3	21	15	39
Author Total	30	63	63	156

To evaluate the classifier on testing data, we defined an accuracy measure as follows.

Accuracy(%)=correctly classified emails /Total emails*100

Table 2 Classification result based on data size

Data size	%
500	93.4
1000	96.2
1500	95.17
2000	95.6

6. CONCLUSION

Certainly, the presence of an automated system can reduce the time to classify mails, identify threats among them. An email classifier is one of the critical tools needed for the effective management of information in the Internet age. In this paper, we have employed Bayesian classification approach to detect threats in e-mails and classify them to predefined classes All the e-mails were classified as threat mails (1) or not (0).We found that 58% of emails are send on topic food, and movies and 63% are for same author. Further we may adopt more classification techniques with additional parameters.

7 REFERENCES:

[1]R. B. Segal and J. O. Kephart. Swift: An intelligent assistant for organizing e-mail. *Proceedings of AAAI 2000 Spring Symposium on Adaptive User Interfaces*, pages 107{112, 2000.

[2] Stephen Wan and Kathleen McKeown. Generating overview summaries of ongoing email thread discussions. In Proceedings of COLING'04, the 20th International Conference on Computational Linguistics, August 23–27 2004.

[3] Owen Rambow, Lokesh Shrestha, John Chen, and Chirsty Lauridsen. Summarizing email threads. In HLT/NAACL, May 2–7 2004.

[4] Salvatore J. Stolfo, Shlomo Hershkop, Chia-Wei Hu, Wei-Jen Li, Olivier Nimeskern, and Ke Wang. Behavior-based modeling and its application to email analysis. ACM Trans. Inter. Tech., 6(2):187–221, 2006.

[5] Fayyad, W.; Piatetsky-Shapiro, G.; Smyth, P., 1996, "From data mining to knowledge discovery: An overview.", In: "Advances in Knowledge Discovery and Data Mining", W. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.), AAAI/MIT Press, Cambridge/USA, pp. 1 – 34

[6]The Journal of Business Law: "E-mail: Speed, Convenience,and Liability," March 2005. http://iblsjournal.typepad.com/illinois_business_law_soc/2005/03/ email_speed_con.html

[7] InternetNews.com (JupiterMedia): "E-Mail Archiving Stealing ECM Spotlight?" January 2005. http://www.internetnews.com/storage/article.php/3465781

[8] E. Crawford, J. Kay and E. McCreath. Iems – the intelligent email sorter. In *19th Int. Conf. on Machine Learning*, 2002.

[9] N. Ducheneaut and V. Bellotti. E-mail as habitat: an exploration of embedded personal information management. *Interactions v.8, n.5, pp.30-38*, 2001.

[10] S. Scott and S. Matwin. Feature engineering for text classification. In *Proc. 16th ICML*, 1999.

[11] D. J. Cook and L. B. Holder. Graph based data mining. *IEEE Intelligent Systems*,15(2):32{41, 2000.

[12] J. HAN, M. KAMBER, Data Mining Concepts and Techniques. *Morgan Kaufmann Publishers*, 2004.

[13] Kamber M, Winstone L and Hanj – Generation and decision Tree Induction – Efficient classification in Data mining

[14] Meyer, T. A., & Whateley, B. (2004). SpamBayes: Effective open-source, Bayesian based, email classification system. CEAS.

[15] Steve Whittaker and Candace Sidner. Email overload:exploring personal information management of email.In CHI '96: Proceedings of the SIGCHI conference onHuman factors in computing systems,
pages 276–283,1996.