ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 5, Issue 12, December 2016

Content based Classification of Emails for Email Filtering and Spam Detection

Prof. S.B. Madankar[1], Himani Vihare[2], Aparna Deshmukh[3], Harshal Walde[4], Ashok Gaikwad[5]

Assistant Professor, Information Technology, PVPIT, Pune, India [1]

Students, Information Technology, PVPIT, Pune, India [2,3,4,5]

Abstract: Researchers initially have addressed the problem of spam detection as a text classification or categorization problem. However, as spammers' continue to develop new techniques and the type of email content becomes more disparate, text-based anti-spam approaches alone are not sufficiently enough in preventing spam. In an attempt to defeat the anti-spam development technologies, spammers have recently adopted the image spam trick to make the scrutiny of emails' body text inefficient. The main idea behind this project is to design a spam detection system. The system will be enabled to analyse the content of emails, in particular the artificially generated image sent as attachment in an email. The system will analyse the image content and classify the embedded image as spam or legitimate hence classify the email accordingly. This experiment results show this approach can get high recognition ratio and reduce the cost of calculation.

Keywords: Spam Filtering, Content Based Filtering, Spam email detection, Machine Learning, Nearest neighbour classifier, Pattern recognition, Data mining Classification, Naïve Bayes.

I. INTRODUCTION

Electronic mails have become one of the most widely used modes of communication in the past two decades. Nowadays emails are used not only for conveying messages but also for other applications such as, official data transfer (electronic documents), highly secured data transfer (password/PIN numbers), private data transfer (personal information) and even to transfer highly sensitive information (company secrets). When a user sends an email from her mailbox, the sender guarantees the authorship of the corresponding email. The receiver can consider this email as significant as a signed letter from the sender. Even in forensic investigations or legal procedures, an email is granted a lot of significance to stand as a valid proof Existing system have poor performance.

In earlier traditional mail system, we weren't able to separate out spam mails as it only categorize mail according to its primary status. As well as existing system not able to classify mailbox more systematically.

Unsolicited commercial email, commonly known as spam, is a pressing problem on the Internet. It undermines the usability of the email system and also costs space thus delaying in system response. The goal of this paper is to design and develop a spam detection system for emails by using classifiers like NaiveBayes, NaiveBayes Multinomial, KNN, SVM.Classifying email and filtering for spam detection using content based classification algorithms for systematic arrangement of mailbox.

Mail filters can be installed by the user, either as separate programs (see links below), or as part of their email program (email client). In email programs, users can make personal, "manual" filters that then automatically filter mail according to the chosen criteria. Most email programs now also have an automatic spam filtering function. Mailbox providers can also install mail filters in their mail transfer agents as a service to all of their customers.

Besides pass, redirect, and drop actions, this kind of filters can also reject a message back to the sender, who is presumed to generate a bounce message in this case. Anti-virus, anti-spam, URL filtering, and authentication-based rejections are common filter types. Corporations often use filters to protect their employees and their information technology assets. Mail filters have varying degrees of configurability. Sometimes they make decisions based on matching a regular expression. Other times, keywords in the message body are used, or perhaps the email address of the sender of the message. More complex control flow and logic is possible with programming languages; this is typically implemented with adata-driven programming language, such as procmail, which specifies conditions to match and actions to take on matching, which may involve further matching. Some more advanced filters, particularly anti-spam filters, use statistical document classification techniques such as the naive Bayes classifier. Image filtering can also be used that use complex image analysis algorithms to detect skin-tones and specific body shapes normally associated with pornographic images.

II. WORKING

Publicly available email message data from ling spam data set is used in the paper. Downloaded data which contains a

DOI 10.17148/IJARCCE.2016.512116

series of folders with each having clean and spam message need to be restructured into clean and spam folders for training and validation purpose as a part of data preparation. Classifying email and filtering for spam detection using content based classification algorithms for systematic arrangement of mailbox.

Diagram of the Proposed Work followed. Firstly paper implementing rule based filtering which will indentify the spam mails according to the cache architecture and other header based rule. Implementing content based filtering technique, i.e. function. Paper chooses this technique because it is showing the most accurate results as compared to other content based techniques Pass the unclassified mails of rule based filter through content based filtering which will finally classify the remaining mails.

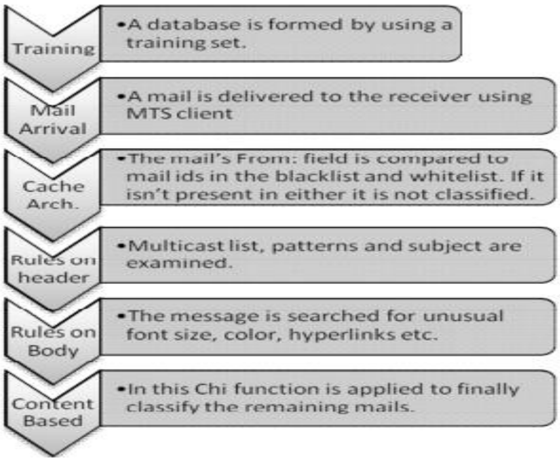
Training
•A database is formed by using a training set.
Mail Arrival
•A mail is delivered to the receiver using MTS client
Cache Arch.
•The mail's From: field is compared to mail ids in the blacklist and whitelist. If it isn't present in either it is not classified.
Rules on header
•Multicast list, patterns and subject are examined.
Rules on Body
•The message is searched for unusual font size, color, hyperlinks etc.
Content Based
•In this Chi function is applied to finally classify the remaining mails.

Fig: System Flow Diagram

III. SYSTEM ARCHITECURE

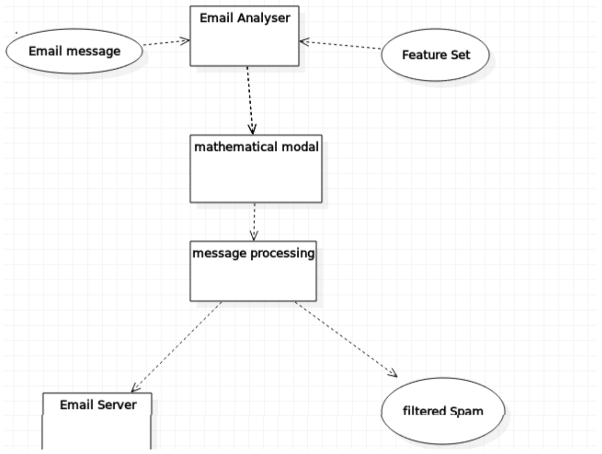

Fig: System Architecture

This diagram represents filtering of electronic mail, and detection of spam mails. Those filtered e-mails are again classified according to customized sections.
Publicly available email message data from ling spam data set is used in the paper. Downloaded data which contains a

series of folders with each having clean and spam message need to be restructured into clean and spam folders for training and validation purpose as a part of data preparation. Classifying email and filtering for spam detection using content based classification algorithms for systematic

Diagram of the Proposed Work followed. First implementing rule based filtering which will indentify the spam mails according to the cache architecture and other header based rule. Implementing content based filtering technique, i.e. function.

And then classifying emails according to contents and chooses this technique because it is showing the most accurate results as compared to other content based techniques Pass the unclassified mails of rule based filter through content based filtering which will finally classify the remaining mails Email management is a specific field of communications management for managing high volumes of inbound electronic mail received by organizations. Today, email management is an essential component of customer service management. Customer service call centers currently employ email response management agents along with telephone support agents, and typically use software solutions to manage emails.

The method of dealing with the inboxes and their organization is discussed and implemented in the thesis. It can considerably simplify handling of emails basically by their classification according to ones preferences and then dispatching to appropriate folders.

The classification of emails is a hierarchical system of categories used to organize messages according to their content, so that any email can be easy found. The following example illustrates a typical hierarchical structure of a mailbox

IV. METHODS

The In this paper different combinations of algorithms are used namely. The spam problem is treated as a classification problem, i.e., a pattern recognition problem. The user needs only to decide whether an email is spam or not. An intelligent agent will learn from his decisions to sort out whether a future email is spam or not. , a simple metric model was proposed to learn to distinguish between triangles and rectangles from their pixel representation.

In the testing stage, the test objects are simply assigned to the class of the nearest mean vector. Other methods of assignment can be considered, particularly in cases where there are many clusters in the same class of objects. For efficiency reasons, in this paper, we shall use the simple 1-nearest-neighbor rule to the mean vectors. The proposed learner here is extremely quick to recognize and at the same time naive, partly because it does not consider multiple clusters within the same class distribution.

V. CONCLUSION

The Naive Euclidean training procedure runs extremely fast for the spam detection problem and yet the correct classification rate is reasonable. It could serve as a baseline in terms of CPU time and accuracy against which other learning methods can be compared. With the increasing importance of email and the incursions of Internet marketers, unsolicited commercial email (also known as spam) has become a major problem on the Internet. To detect image spam, computer vision and pattern recognition techniques are also required, and indeed several techniques have been recently proposed. The proposed framework exploits both.

ACKNOWLEDGMENT

We take this golden opportunity to owe our deep sense of gratitude to our project guide **Prof. S.B. Madankar,** for his instinct help and valuable guidance with a lot of encouragement throughout this paper work, right from selection of topic work up to its completion. Our sincere thanks to Head of the Department of Computer Engineering and Information Technology **Prof. B, K. Sarkar** who continuously motivated and guided us for completion of this paper. We are also thankful to all teaching and non-teaching staff members, for their valuable suggestions and valuable co-operation for partially completion of this work. We specially thank to those who helped us directly-indirectly in completion of this work successfully.

REFERENCES

[1] Massaging anti-abuse working group, “Email metrics report”, http://www.maawg.org/, 2006.

[2] Surmacz, R. Tomasz, “Reliability of e-mail delivery in the era of spam”, International Conference on Dependability of Computer Systems, DepCoSRELCOMEX’07, pp. 198 – 204, Jun 2007.

[3] Calton Pu and Steve Webb, “Observed trends in spam construction techniques: A case study of spam evolution”, proceedings of 3rd conference on e-mail and anti-spam, CEAS’2006, 2006.

[4] Joshua Goodman, Gordon V. Cormack, and David Heckerman, “Spam and the ongoing battle for inbox”, communications of ACM, 50(2): 25-33, 2007.

[5] Mikko Siponen and Carl Stucke, “Effective anti-spam in companies: An international study”, proceedings of HICSS’06, vol 6, 2006.

[6] J. Smith and I. Fujinaga, “A review of authorship attribution,” Technical Report, 2008.

[7] M. Can, “Authorship attribution using principal component analysis and competitive neural networks,” Mathematical and Computational Applications, vol. 19, no. 1, pp. 21–36, 2014.

[8] F.Sebastiani,“Machine learning in automated text categorization, ”ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.

[9] C. H. Ramyaa and K. Rasheed, “Using machine learning techniques for stylometry,” in Proceedings of International Conference on Machine Learning, 2004.

[10] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, p. 27, 2011.

[11] R. Goodman, M. Hahn, M. Marella, C. Ojar, and S. Westcott, “The use of stylometry for email author identification: a feasibility study,” Proc. Student/Faculty Research Day, CSIS, Pace University, pp. 1–7, 2007.

[12] K. Calix, M. Connors, D. Levy, H. Manzar, G. MCabe, and S. Westcott, “Stylometry for e-mail author identification and authentication,” Proceedings of CSIS Research Day, Pace University, 2008.

[13] J. Webber, “A programmatic introduction to neo4j,” in Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity. ACM, 2012, pp. 217–218.

DOI 10.17148/IJARCCE.2016.512116