International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Investigation of Email Scam Detection using Weka Tool

Mohammad Zuber[1], Dr. Mohammad Iliyas Khan[2], Dr. S.Veenadhari[3]
[1]*Research Scholars,* [3]*Associate Professor, AISECT University, Bhopal (M.P) India.*
[2]*Principal, Vidhyapeeth Institute, Bhopal (M.P) India.*

Abstract*: *Data mining is also being valuable to give resolutions for assault discovery and checking. Although data mining has numerous requests in defense, there are likewise thoughtful privacy fears. Because of email mining, even inexperienced users can connect data and make responsive associations. Therefore we must to implement the privacy of persons while working on practical data mining. Using K-mean clustering procedure and weka tool we executed the method of Email-mining. The WEKA tool calls the .eml file format into text converter and then processed the whole data into preprocessor output in form of .csv file format. The preprocessor output shows the graphical results of the processed email data. The goal of this implementation is to detect or filter the email addresses from which we get maximum emails.
Keys: Data mining, Email mining, Weka tool, K-mean, Clustering algorithm, Preprocessor

I. INTRODUCTION

Every day E-mail users receive hundreds of spam messages with a new content, from new addresses which are automatically generated by robot software. To filter junk with old-style approaches as black-white lists (domains, IP addresses, mailing addresses) is nearly unbearable. Request of text mining approaches to an E-mail can increase competence of a filtration of spam. Also classifying spam messages will be possible to establish thematic dependence from geographical [1]. These paper emphases on the effort complete to classify the written junk E-mails with data mining techniques. Our purpose is not only to filter messages into spam and not spam, but still to divide spam messages into thematically similar groups and to analyze them, in order to define the social networks of spammers [2]. In this study we proposed a dynamic clustering of data with simple k-means algorithm. The procedure receipts amount of clusters (K) as the contribution from the operator and the operator has to reference whether the amount of clusters is secure or not. If the quantity of clusters secure then it works similar as K-means algorithm.

II. METHODOLOGY

Initially historical method of research will be used to gather the relevant data through authentic literature, books journals etc., in this research I will study a past published research thesis related to Email Mining [3].

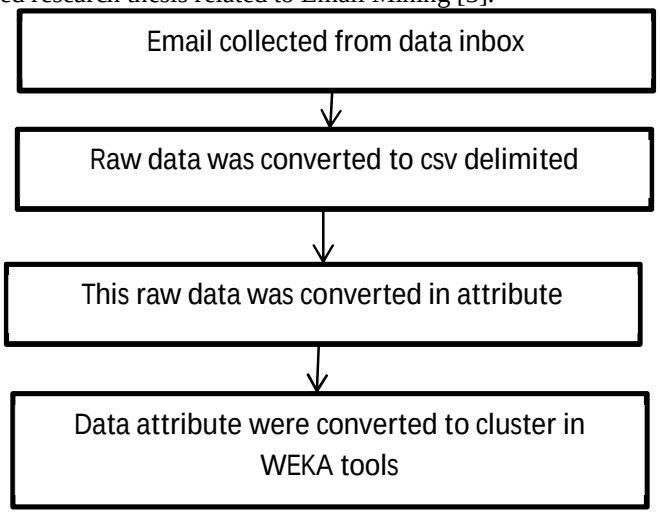

Fig.1: Data analysis process structure

After that evaluative and analytical method will be used through structured questionnaires of two patterns:

A. Primary Data

Information based on the results of analyzing data of logged emails and the research scholar will form the primary data.

B. Secondary Data

Information based on the literature, books, journals will form the secondary data.

C. Survey method will be used for the study.

D. Tool and techniques for analysis would be Reliability and Validity test.

Some of the data which was extracted as a single main from incoming mailbox different organizations or and personal mail boxes for sample I have collected data available for an society in viewpoint express data arranged this data was transformed to modest delaminated cvs format for informal data adaptation [4].

III. PROPOSED WORK

A. K-Mean Algorithm

The k-means procedure is an evolutionary procedure that improvements its name from its technique of process. The process clusters explanations into k clusters, where k is providing as an input limit [5, 11].

1) Input
2) k: the number of clusters.]
3) D: a data set covering n substances.
4) Output: A set of k clusters.
5) Method:
a) Subjectively select k objects from D as the early cluster middles.
b) Repeat.
c) Re-assign all object to the cluster to which the object is maximum alike using, created on the mean worth of the objects in the cluster.
d) Appraise the cluster means.
e) Until no change.

B. Weka

The workflow of Weka would be as monitors [6]:

1) Data →Pre-processing → Data Mining →Knowledge
2) The maintained data formats are ARFF, CSV, C4.5 and binary. Alternatively you could also import from URL or an SQL database.
3) Subsequently filling the data, preprocessing cleans might be used for addition/eliminating qualities, discretization, Sampling, randomizing etc.

C. Outlook Express

Microsoft Outlook and Microsoft Conversation use a exclusive email add-on format called Transport Neutral Encapsulation Format (TNEF) to grip arranging and other features specific to Outlook such as meeting requests [7]. An open-source scheme called UnDBX was also shaped, which appears to be fruitful in improving immoral databases. Microsoft has likewise free documentation which may be capable to precise some non-severe difficulties and reinstate admission to email messages, without resorting to third-party solutions. Though, with the newest informs functional, Outlook Rapid now types backup reproductions of DBX files previous to compaction. They are kept in the Recycle Bin. If a mistake happens throughout compaction and messages are misplaced, the DBX files can be derivative from the recycle bin [8, 14]. Opening or previewing the email could cause code to run without the user's knowledge or consent. Outlook Express does not properly grip MIME, and will not show the body of employed messages inline. Users become an occupied email and one add-on (one of the message text and one of the signatures) and consequently essential to open an add-on to see the e-mail [9].

IV. RESULTS AND DISCUSSION

Procedure of Email mining done weka mechanism as assumed under with results:

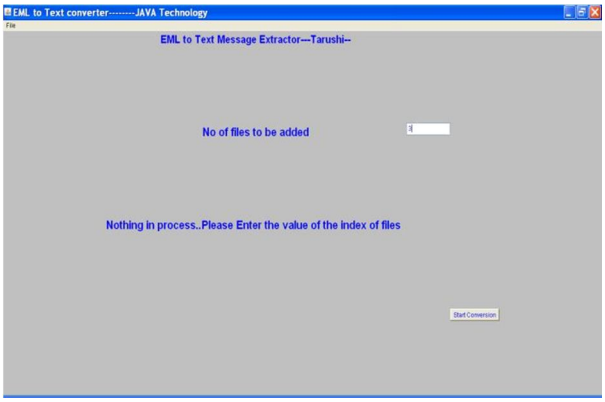

Fig.2: Weka explorer EML to text converter

The .eml to text converter is the chief dialog box of the operation work which contain that how many emails are to be improve [9].

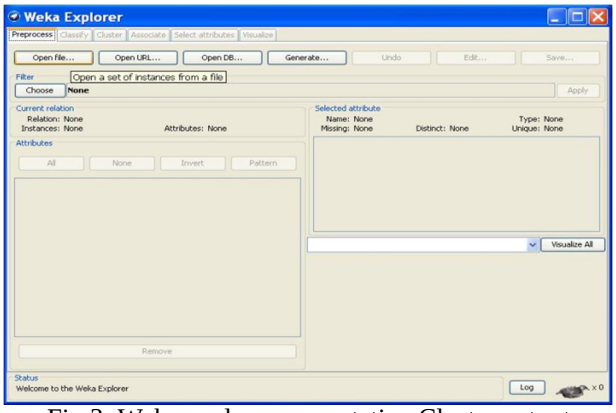

Fig.3: Weka explorer presentation Cluster output

Exposed dialog box seemed to choice the excerpt emails in form of .csv form which is situated in concluding directory. The weka tool requests the .eml file setup into text converter and then administered the complete data into preprocessor output in form of .csv file format [10] . The preprocessor output displays the graphical results of the treated email data.

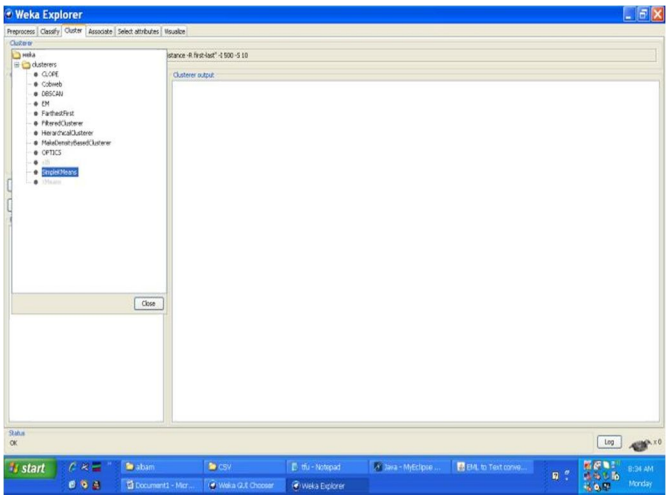

Fig.4 Weka explorer to choice K-mean algorithm

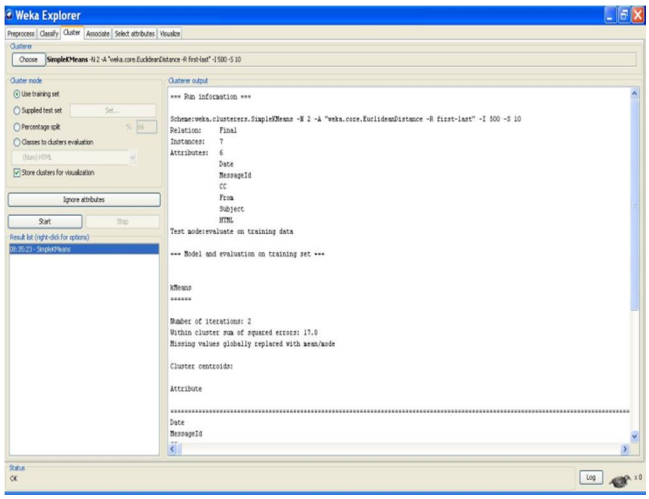

Fig.5: Weka explorer presentation Cluster output

Cluster output shows

Parameters	Output
Instances	7-number of live matches
Attribute	6-selected 6 attribute are Date, Message id, CC, From, Subject, HTML
No. of Iterations	2

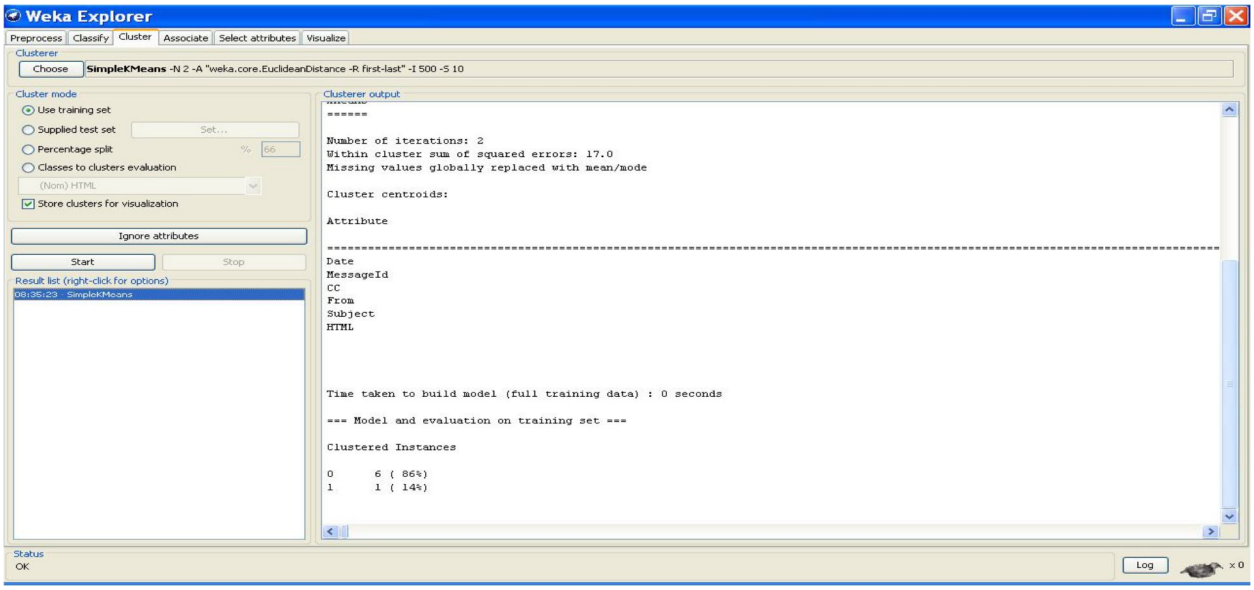

Fig.6: Weka explorer viewing Cluster output

Cluster output shows

Parameters	Output
Instances	6(86%), means 6 out of 7 instances are identical
Instances	1(14%), means 1 out of 7 instances
No. of Iteration	2

The graphical results shows the count and percentage values of instances on the basis of attributes selected. Attributes are the objects of email we select to compare [13]. The percentage values show the percentage amount of instances out of total. These examples shows the incidence of email data which benefits in noticing the email positions that are sending extreme number of spam mails or unwelcome mails. These are used to do email mining and filtering of emails [12].

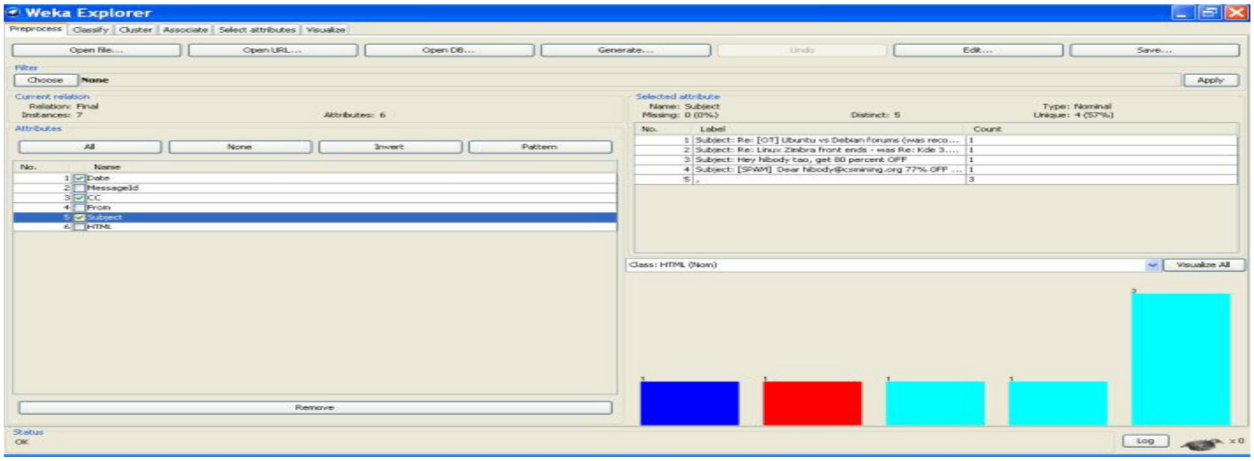

Fig.7: Weka explorer display preprocessor output

Cluster output shows

Parameters	Output
Mails	3
Attribute	3 out 7
Instances	4 different and 3 identical

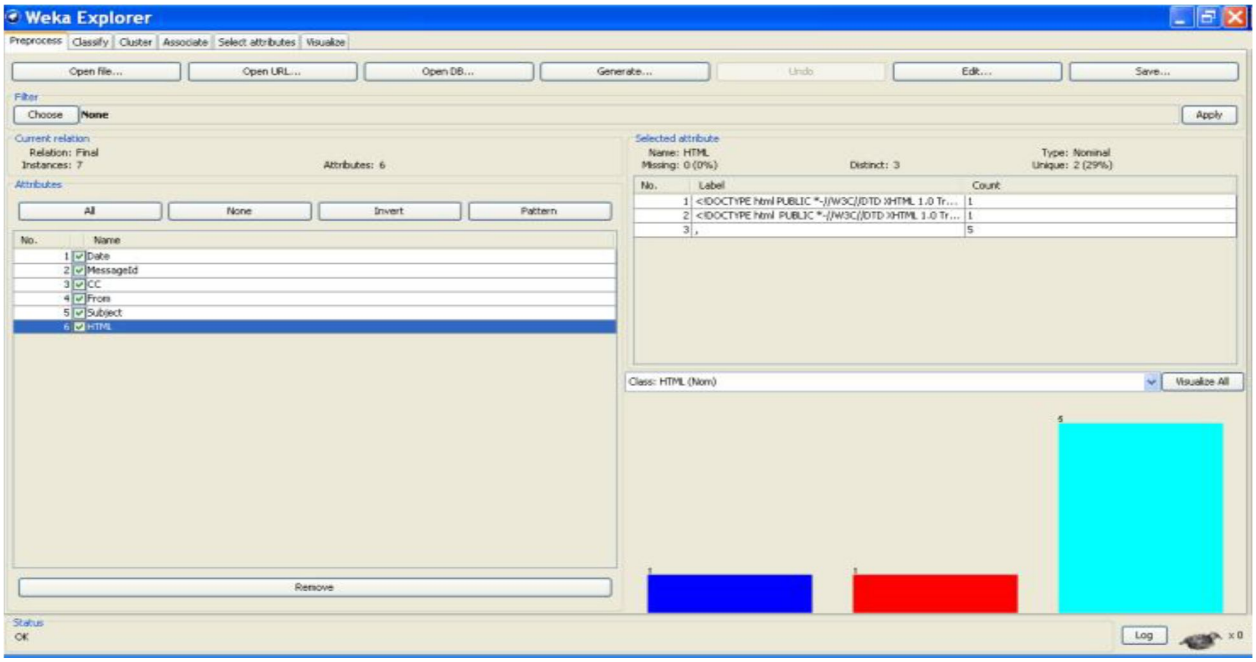

Figure 8: Weka explorer display Preprocessor output

Cluster output shows

Parameters	Output
Mails	3
Attributes	6 out of 7
Instances	2 different and 5 identical

In second case we selected all 6 attributes and then processed the data of three mails. It resulted into 2 different instances and 5 Similar instances of ",". And all those are shown in graphs of different colors.

V. CONCLUSION AND FUTURE WORK

Approaching to conclusion the chief theme of this effort is to perceive or filter the emails to examine the specific email address from which the quantity of receiving emails is maximized. A novel impression that appears promising is Semantic Email has been proposed. The technique used is resultant to more secured feature to detect spam mails and their source address. The software implemented could be used to detect those methods and integrate them into useful and accurate email-mining which will let people take back control of their mailboxes. Forthcoming references have numerous choices similarly the current work is done for solitary email-id and the data treated is done for solitary email-id. But the effort could be complete by taking multiple email-ids. Moreover this filter evaluation technique is processed by taking just six attributes. But we can take much more number of attributes to improve the filtered results. The efficiency of text converter which converts the outlook express file format into WEKA tool file formats. As in the current effort the conceal code created perform the text transformation from .eml file format to .csv file format. But in additional references the file transformation could be execute for numerous file formats of WEKA tool like .arss. In this work we have selected K-mean algorithm to procedure the clusters of data in WEKA tool. But in future work we can use more efficient existing or proposed algorithm to process the email data.

REFERENCES

[1] A. Anderson, M. Corney, O. de Vel, and G. Mohay."Identifying the Authors of Suspect E-mail". Communications of the ACM, 2001.
[2] Shlomo Hershkop, Ke Wang, Weijen Lee, Olivier Nimeskern, German Creamer, and Ryan Rowe, "Email Mining Toolkit Technical Manual". June 2006) Department of Computer Science Columbia University.
[3] Bron, C. and J. Kerbosch. "Algorithm 457: Finding all cliques of an undirected graph." (1973).
[4] Ding Zhou et al and Ya Zhang, "Towards Discovering Organizational Structure from Email Corpus". (2005) Fourth International Conference on Machine Learning and Application.
[5] Giuseppe Carenini, Raymond T. Ng and Xiaodong Zhou , "Scalable Discovery of Hidden Emails from Large Folders". Department of Computer Science, University of British Columbia, Canada.
[6] Hung-Ching Chen el al, "Discover The Power of Social and Hidden Curriculum to Decision Making: Experiments with Enron Email and Movie Newsgroups". Sixth International Conference on Machine Learning and Applications.
[7] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz and Anand Swaminathan, "Mining Email Social Networks". (May 22-23, 2006). Dept. of Computer Science, University of California, Davis.
[8] Rong Qian, Wei Zhang, Bingru Yang, "Detect community structure from the Enron Email Corpus Based on Link Mining".(2006) Sixth International Conference on Intelligent Systems Design and Applications.
[9] Deepak P, Dinesh Garg and Virendra K Varshney, "Analysis of Enron Email Threads and Quantification of Employee Responsiveness". IBM India Research Lab, Bangalore - 560 071, India.
[10] Ziv Bar-Yossef et al, "Cluster Ranking with an Application to Mining Mailbox Networks". (2006) Sixth International Conference on Data Mining.
[11] Hua Li et al, "Adding Semantics to Email Clustering". (2006) Sixth International Conference on Data Mining.
[12] http://libguides.sjsu.edu/a-z - The SJPL library database
[13] Shlomo Hershkop et al, "Email Mining Toolkit Technical Manual".Version 3.6.8 -June 2006
[14] http://www.apachefriends.org/en/xampp-windows.html, XAMPP source

I am Mohammad Zuber, Phd Scholors in Computer science & engineering from Aisect University, Bhopal (M.P) india.