www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 2 February 2015, Page No. 10475-10477

Clustering of E-mails Using Pattern Matching

Bhavana Pansare[1], Yamini Chaudhari[2], Pradnya Deshmukh[3] ,Deepti Mutalik-Desai[4], Tanuja Khedkar[5]

N.M.I.E.T Department of Computer , Pune University,
Talegaon Dabhade
yamnichaudhari93@gmail.com
N.M.I.E.T Department of Computer , Pune University,
Talegaon Dabhade
deshmukh.gauri7@rediffmail.com
N.M.I.E.T Department of Computer , Pune University,
Talegaon Dabhade
desaidipti20@gmail.com
N.M.I.E.T Department of Computer , Pune University,
Talegaon Dabhade
tanuja997@gmail.com

Abstract: *Email communication is one of the most effective and popular way of communication today. Sending and receiving of messages for exchange of information is done by people every day. E-mail communication is popularly used way of communication. E-mail data that are now becoming most important way of inter and intra organizational written communication for many companies. Clustering is defined as creating group of similar objects. The cluster shows the similar emails exchanged between the users and for finding the text similarities to cluster the users, we are using the Pattern i.e., by considering the different Threshold value the similar words exchanged between the users, Threshold value shows the frequency of the words used and we have graphically represented the cluster in the form of Bar charts.*

Keywords: *Email Clustering, Email Attributes, Text similarity, Pattern Matching*

1. Introduction

Emails are the one of the faster and economical way of communication. But the increase of email users has resulted in the dramatic increase of spam emails during the past few years. New filters need to be developed to catch spam as spammers always try to find a way to evade existing filters. Ontologies allow for machine-understandable semantics of data. For more effective spam filtering it is important to share information with each other. Many researchers have applied techniques to email for classifying emails, such as identifying spam messages. These approaches are highly effective, but many examine incoming email exclusively which are not sufficient to provide detailed information about an individual users behavior. By analyzing outgoing messages users behavior can be ascertained. of abnormal email activity; and a demonstration of the effectiveness of outgoing email analysis using an application that detects worm propagation Clustering techniques can be applied over email data to create groups of similar emails. An email is an Object consisting of several attributes like sender email-id, sending-time, receiver email-id Subject, message, and attachments etc. to discover email groups clustering is used. Generally most of the attributes in emails are text type. For measuring the similarity between pair of email objects text similarity techniques are used. Depending on the information they have exchanged and graphically representing Cluster Clustering of emails is done. The Enron email dataset is used for such research.

2. Related Work

The Enron Email Dataset Database Schema and Brief Statistical Report [1] which show how the distribution of emails per users and showing the network how the employees are connected. Use of interpersonal communication for network inference has been of interest to researchers for several decades. Early work [2] utilized email traffic to infer social networks for the purpose of discovering communities of shared interest. Email classification can be done to several different applications including assigning messages to user-created folders, filtering messages based on priority, or identifying SPAM. One major consideration in the classification is that of how to represent the messages. One must decide how to apply those features to the classification and which features to use. Manco, et al. [3] defined three types of features to consider in email: categorical text, unstructured text , and numeric data. Relationship data is an other type of information that could be useful for classification. Unstructured text in email consists of fields like the subject and body allowing for natural language text of any kind. Generally, these fields have been used in classification using a bag-of-words approach, the same as with other kinds of text classification [3, 4, and 5]. Stemming and stop word removal are often Used, as they are useful in general text classification, although their usefulness in email in particular has not yet been studied thoroughly. It has been

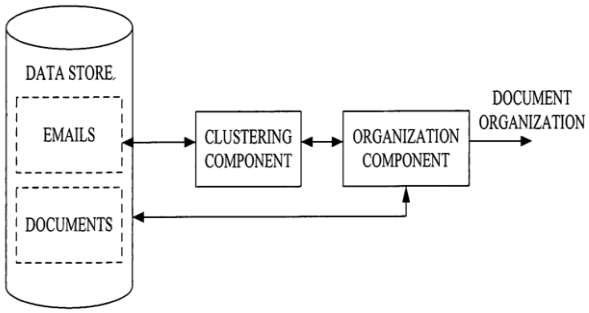
found that some of these fields are more important that others in classifying email [3]. A technique to classification the email is proposed by Martin, Sewani, Nelson, Chen, and Joseph . The proposed classification is used for identifying spam messages. Fields such as "to" and "from" [3] includes categorical text. These differ from unstructured text fields in that the type of data which can be used in them very well defined. But, these fields have typically been treated like unstructured text fields, with the components added to the bag of words [3, 4]. These fields have been found to be very useful in automatic email classification, although not as useful as the unstructured data [3].

3. Existing System

Existing system consists all the current application which we use daily for E-mail communication. It is nowadays important documents worldwide used in organization for inter and intra communication. Hence any user would like to organize his/her mailing boxes so as to have faster retrival of mail. However in existing system we hardly have such type of clustering technique used. Hence we have introduced it in our ways possible.

4. Proposed System

The database contains all the emails which are received and send then the next component is clustering in which the clustering algorithm is applied to emails. Next is organization in which the clustered mails are ranked and organized accordingly and the output is given to the user. The same output is again stored into database for further use.



5. Mathematical Model

1 . Let S be a system that describe Our Email system
S = ...

2. Identify input as I
S = I,....
I1:Information
I2 :Threshold value.
I3:Cluster name

3. Identify output O
S = I, O,....
O1:Clusters of E-mails.

4. Identify the processes as P
S = I, O, P,....
P0 :Seek all data of messages/E-mails
P1:Give desirable threshold value.
P2:Give the keyword upon which clustering is been done.
P3:final o/p in desired form

5. Identify failure cases as F.
S = I, O, P, F,....
F = Failure occurs if any attribute does not contain value(Threshold, keyword, priority), Validation Fails

6. Identify Success cases as s.
S = I, O, P, F, s....
s = Success occurs when Registration, validation is done, where as other attribute is filled up properly to have clustered E-mail view.

7. Identify Initial condition as Ic.
S = I, O, P, F, s, Ic
Ic = Registration is the first step

6. Algorithm

System algorithm consists of following steps:

6.1 *Email as a database*

Technically an email is constituted of number of fields or attributes like sender email-id, receiver email id ,subject, message body , attachments, cc, Bcc etc. set of emails in a mail box can be treated as number of email records with the number of attributes of an Email. In this way the emails stored in the mail box can be treated as the email database.

6.2 *Email Mining*

Data mining techniques can be applied over the email databases to discover the useful and interesting knowledge There are number of applications of email mining today. Some of the interesting email mining applications are email Clustering, email categorization, Summarization, automatic answering etc

6.3 Email Clustering

Clustering is a technique of creating group of similar objects. When clustering is used in email mining it is called as email clustering. Email Clustering can be explained as clusters of messages with the same concept give an appropriate name to each cluster and then put all messages into their corresponding folders. Clustering the users who are discussing the similar content.

6.4 Algorithms

When the mail comes into the mailbox are first step is preprocessing. Preprocessing includes Parsing, stemming and E-mail representation .For parsing and stemming we use Porter stemmer algorithm. First all the non-textual information is removed. Then stop words such as "i", "am" ,"ing" etc are removed. Now we are left with words which then undergoes for pattern matching.

IMPLEMENTATION

The proposed algorithm is implemented using open source technologies and algorithm is applied over the popular email corpus database. Java is selected as the programming languages and the other open source API's (Application Programming Interfaces) to Support the other functionalities. My Eclipse is used as a development IDE (Integrated Development Environment) for Java and library of other technologies are added as external jar (Java Archives) in the

eclipse. MyEclipse is built upon the Eclipse platform and integrates both proprietary and open source solutions into the development environment. JFree Chart is an open-source framework for the programming language Java, It is an open source library available for Java that allows users to easily generate graphs and charts. It is mainly effective for when a user needs to regenerate graphs that change on a frequent basis. JFree Chart supports time series charts, pie charts (2D and 3D), scatter plots, bar charts, line charts, and high-low-open close charts.

7. Technologies

1.1 Java

It is general purpose, object -oriented programming language developed by sun micro system of USA in 1991 which was originally called as 'Oka' by James Ghosling. The important feature of language is that it is a platform neutral language. Java is the first programming language which is not tied to any particular hardware or any OS. Programs developed in java can be executed anywhere on any system

8. Outcomes

Outcomes of the project are as follows:

1. **Apart from having a primary tab just like in Gmail which have all the mails we are also expecting to create additional customized tab.**
2. **Customer can have as many number of clusters according to need and priority.**
3. **Different databases can also be merged.**

9. Conclusion

Thus in this system instead of only sequential arrival of mails, mails are sorted and group to form clusters Clustering makes easy retrievals of any important mail from the huge/large dataset. Thus accessibility of emails become easier than the present system.

References

[1] Brief Statistical Report, /"Technical Report, Information Sciences Institute, 2004. Available at: http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf.

[2] Michael F. Schwartz and David C. M. Wood. Discovering shared interests using graph analysis. Commun. ACM, 36(8):78–89, 1993 208

[3] G. Manco, E. Masciari, M. Ru_olo, and A. Tagarelli: Towards an Adaptive Mail Classifer. AIIA 2002, Sep. 2002

[4] W. W. Cohen: Learning Rules that classify E-mail. In Proc. of the 1996 AAAI Spring Symposium in Information Access, 1996.

[5] J. Rennie: i_le: An Application of Machine Learning to E-Mail Filtering. In Proc. KDD00 Workshop on Text Mining, Boston, 2000.