

# A Framework for Email Clustering and Automatic Answering Method (Formal Approach)

<sup>1</sup>Praveen Kumar<sup>2</sup>Himanshu Kumar<sup>3</sup>Remya Joseph<sup>1</sup>Mtech Software Technology, <sup>2,3</sup>Mtech IT, School of Information Technology and Engineering (SITE),

VIT University, Vellore - 632014, Tamil Nadu, India

**ABSTRACT-** The mail-box consisting of number of emails in which many attributes or files are existing in a single email and hence they can be considered as a database. It is not easy to find out the important email or message quickly or within a stipulated time. So we are using clustering techniques for grouping similar emails. The attributes of an email like sender email id, receiver email id, subject, message body, sending time, attachments, cc, bcc, etc. are used for clustering the emails. Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM. In this project we are going to add certain extra features like clustering based on the subject of the email, priority of the person who sends the message, message body contents, date and time, size of the message and also add automatic answering feature for a received mail.

**KEYWORDS-** Automatic answering, Email Attributes, Email Clustering, String Matching.

## I. INTRODUCTION

The growth of the Internet has led to an explosion in e-mail traffic and in number of the users of e-mail service. At the same time, many large-scale e-mail service providers have appeared. They have hundreds of millions of subscribers and process billions of messages: for example, in May 2001, hotmail had over 100 million users and Yahoo! Mail, in March 1999, served 45 million users with 3.6 billion mail messages [1][2]. Thus email has been an efficient and popular communication mechanism as the number of Internet users increase. Therefore, email management is an important and growing problem for individuals and organizations because it is prone to misuse. So clustering is used to group different emails so that it can be represented in a well organized manner. Clustering shows similar emails

exchanged between the users.

### A. Email as a database

Email is comprised of various types of attributes such as sender-id, receiver-id, subject, message body, cc, bcc, date and time, attachments etc. and when many emails come together in the mail-box it forms a large group of messages. So this can be considered as a huge email database consisting of a large number of email records each with a number of attributes.

### B. Email Mining

Email mining is a process of discovering useful patterns from emails. Data mining is the process of discovering new patterns from large data sets involving methods from not only statistics and artificial intelligence but also from database management. Machine learning techniques are often used to identify previously unknown patterns as well as learning known patterns. There are number of applications of email mining today. Some of the interesting email mining applications are email categorization, summarization, automatic answering and spam filtering etc. Text mining is nothing but performing various searching functions, linguistic analysis and categorizations, applying various algorithms etc. to find the similarity in texts.

### C. Clustering Email

Clustering is a technique to create groups of similar email or objects according to user requirement and put it into different folders and when it is used in email mining it is called as email clustering. In email clustering subject-based folders can be automatically

constructed from a set of messages. In this case, the main aim is to acknowledge same clusters of emails from the mail-box give an appropriate name to each cluster and put them into their respective folders.

#### D. Algorithm

Indeed different text clustering methods require several steps of preprocessing of the data. First, any non-textual information such as HTML-tags and punctuation is removed from the documents. Then, stop words such as "I", "am", "and", "ing", "are", "those", "here" etc. were also removed. Then a term is generated. It is any sequence of characters separated from other terms by some delimiter. Also a term can either be a single word or can consist of several words. The terms are reduced to their basic form applying a stemming algorithm such as brute-force algorithm. Then after that term are then compared with terms of other emails to find any similarity. This is called as string matching and certain algorithms like Jaccard algorithm, Cosine algorithm etc. are also used for this. If there is any similarity found between the strings then they are clustered using K-means clustering algorithm.

##### 1) Brute-Force Algorithm

Brute force algorithm was used for the purpose of removal of stop words. Here in this algorithm the stemmers employ a lookup table. In that lookup table there are some relations between root forms and inflected forms in other words it contains the stop words. To remove a stop word the table is explored and searched to find a matching word. If a matching word is found, the associated root form is returned. Brute force algorithm consumes a lot of storage because it has to store all the list of relations and root words in a table(relative to the work).The accuracy of the algorithm is only limited to the number of relations and root words already present in the table. Also in any given number of words in a particular language, like English, it is highly improbable to expect that all word forms can be captured and manually recorded by human action alone. So we are adding Suffix-striping algorithm and making it a Hybrid approach for stop word removal. We are keeping certain specific words like "swim", "swam" "run", "ran" etc. in lookup table and for rest we are using suffix striping for removing the suffix part of the word.

Advantages: Brute force algorithm definitely overcomes some of the challenges faced by the other approaches. Not all stop word forms in a given

language "follow the rules" appropriately. While "running" might be easy to stem to "run" in a suffix stripping approach, the alternate inflection, "ran", is not. Suffix stripping algorithms are unable to overcome this problem, and it results in increasing the number and complexity of the rules, but in brute force algorithms it requires storing a single extra relation between "run" and "ran". We improved the brute force algorithm by building it upon preliminary part-of-speech i.e. we added all possible suffixes and prefixes which will also be removed.

before stop word removal:

yes! i have the very same problem with this camera. it's been in for repairs twice. i fear my only option is to buy a new camera. i am very disappointed. i just got this camera about 6 weeks ago and have lost what would have been great moments due to the poor focusing of the camera. some are clear, some are not. not happy its just a terrible camera to deal with. disappointed.

Total no. of words: 73

After stop word removal

yes i have very same problem this camera it s been repairs twice i fear my only option is buy new camera i am very disappointed i just got this camera about weeks ago have lost what would have been great moments due poor focusing camera some clear some not not happy its just terrible camera deal disappointed

Total no. of words: 58

##### 2) Cosine Similarity Matching Algorithm

Cosine similarity is a measure of similarity between two vectors tables by measuring the cosine of the angle between them. The cosine of 0 is 1, and less than 1 for any other angle. The cosine of the angle between two vectors will determine if two vectors are pointing in the same direction or not. This is often used to compare documents in text mining. In addition, it is also used to measure cohesion within clusters in the field of Data Mining.

The cosine of two vectors is derived by using the Euclidean Dot Product formula:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta$$

Suppose the two frequency vectors are  $A$  and  $B$ , then the cosine similarity  $\theta$ , is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

If the result is  $-1$  then it means that the given string is exactly opposite to the other string, if it is 1 it means that it is exactly the same and if it is 0 this indicates that it is totally dissimilar and in-between values indicating intermediate similarity or dissimilarity. The cosine similarity can be seen as a method of

normalizing document length during comparison. The angle between the two frequency vectors cannot be more than  $90^\circ$ .

### 3) K-means Algorithm for Clustering

Suppose a set of observations are given as  $(x_1, x_2, \dots, x_n)$ , where each observation is a  $d$ -dimensional real vector,  $k$ -means clustering aims to partition the  $n$  observations into  $k$  sets  $(k \leq n) S = \{S_1, S_2, \dots, S_k\}$  so as to minimize the within-cluster sum of squares (WCSS) the formula is as follows:

$$\arg \min_S \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

Where  $\mu_i$  is the mean of points in  $S_i$ .

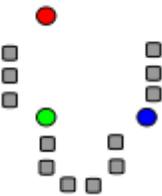
Regarding computational complexity, the  $k$ -means clustering problem is K-Means works in 4 basic steps:

1. Randomly choose  $K$  instances within the dataset and assign them as cluster centers
2. Assign the remaining instances to their closest cluster center
3. Find a new center for each cluster.
4. If the new cluster centers are identical to the previous ones, then the algorithm stops. Otherwise, repeat steps 2-4

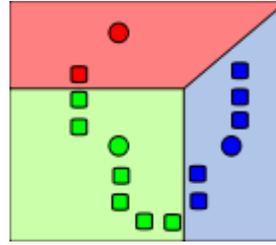
This algorithm uses an iterative refinement technique. Due to its popularity the  **$k$ -means algorithm** is also referred to as Lloyd's algorithm mainly in the computer science community.

#### Demonstration of the standard algorithm:

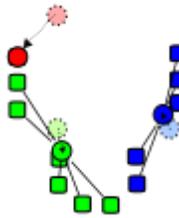
- 1) Initially  $k=3$  was randomly selected from the data set (shown in color).



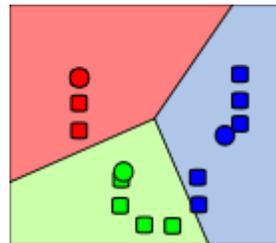
- 2) 3 clusters are created by associating every observation with the nearest mean.



- 3) The centroid of each of the  $k$  clusters becomes the new means.



- 3) Steps 2 and 3 are repeated until the clusters became more prominent.



## II. EXISTING OR RELATED WORK

Early work [2] utilized email traffic to infer social networks for the purpose of discovering communities of shared interest. Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM, or classifying according to similarity of texts etc. One major consideration in the classification is that of how to represent the messages. Also Martin, Sewani, Nelson, Chen, and Joseph [13] proposed that classification can be used for identifying Spam messages.

The message can be represented generally in three types: unstructured, categorical and numeric data as described by [3]. Unstructured text in email consists of fields like the subject, message body, etc. They are used in classification using a bag-of-words approach.

Stemming as well as stop words removal techniques are useful in general text classification and hence are often used. In previous works Porter stemming algorithm was used for removing the stop words. Categorical text includes fields such as “to” and “from”. Categorical data differ from unstructured text fields in the manner that the type of data which is used in it is very well defined. These fields have been found to be very useful in automatic email classification.

At least two research studies relating to text classification have been performed on The Enron corpus[14]. One is the automatic categorization of e-mail into folders, done by the Computer Science Department of the University of Massachusetts [7]. It has also analyzed the graph in theoretical aspect as well as spectral analysis of Enron email data. It also analyzed the Enron email dataset to discover structure within the organization. The analysis is based on constructing the graph and studying its property [8]. The other research that has been conducted is the analysis of Social Network such as [9], [10], [11] and [12] Social network analysis techniques are nothing but the methods to calculate the degree distribution of each node in networks, network density, and centrality.

### III. PROPOSED WORK

The overall approach can be explained using the diagram shown at figure-1. The approach can be divided into the four phases. The phases are pre-processing phase, Key-word matching phase, clustering determination phase and finally automatic answering phase.

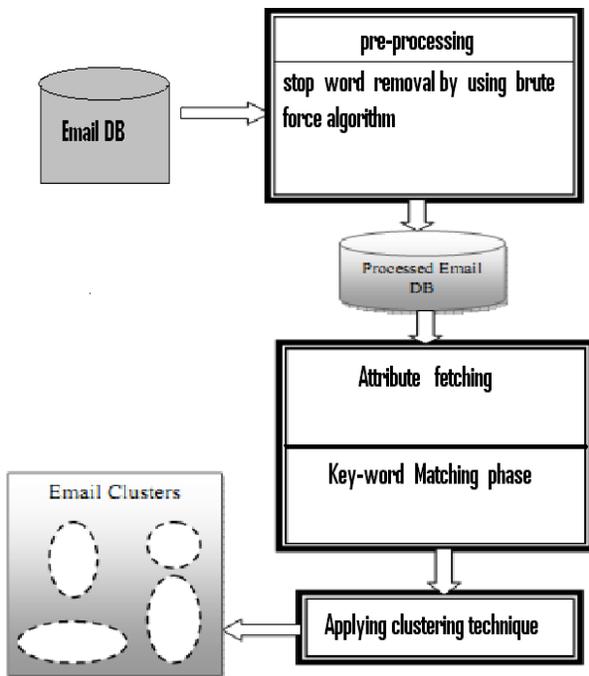
#### A. Pre-Processing

The first step is the pre-processing step of the data which in the beginning is usually available in what is called “raw format”. The raw data can be imagined as a database and it is similar to the Enron Corpus database [1]. In order to mine into “raw” data and extract knowledge from it, it is necessary to transform it into a format that is more comprehensible to the machine learning algorithms. In pre-processing phase all the email data is prepared for clustering. All the important email attributes are retrieved by parsing the email documents. Most of the attributes are of text data

type, so Brute-force stemming algorithm is used to eliminate the unwanted texts from the parsed attribute information. The pre-processing part also includes one effective representation technique of one email. Here object representation technique is used in this paper which holds all important attributes of an email.

In the last decade, people with the help of new and advanced email clients, have started to enter HTML code in order to enrich their plain text messages with different styles of text, different fonts, images, links etc. In fact, this is achieved by sending an HTML page as an attachment which every contemporary email client with a build-in web browser is able to present. In an email analysis procedure, HTML is most of the times not exploited to obtain knowledge and is removed using HTML parsers in order to keep the text contained in the HTML document. Sometimes HTML tags are used as characteristics (attributes) of the email.

The pre-processing includes parsing, stemming and email representation technique for parsed information. Parsing of email documents is required to retrieve various attributes information separately like subject, contents etc. Stemming is required to clean the text information available in email attributes. And finally one representation technique is required for effective representation of email documents, in which all the attributes associated with an email can be made accessible. After pre-processing the processed email dataset can be ready for the data mining operation.[19]



**Figure-1 the overall clustering approach**

Another decision that has to be made is whether to treat words with common stem as one. In that case a vocabulary of stems is built and words like “program”, “programming”, “programmer” are treated and counted as one. A common pre-processing step is the removal of commonly used words (like articles, prepositions etc) which we call “stop-words”. These can be removed by using Brute-force algorithm. That could be effective in most applications because these are words that appear in natural language independently of topic. Hence, they lack discriminative power.

After removing the stop words by using the Brute-force stemming algorithm the refined data is stored in a proposed database which can be further used for the next steps i.e. keyword matching.

### B. Keyword Matching

The second phase is keyword matching in which the attributes are selected from email objects and their pair wise similarity is measured using suitable text similarity algorithm such as Cosine similarity matching algorithm. Here we first consider the email according to their subject and message body. After that the pre-defined keywords can be matched in both the attributes i.e. subject and message body. If the keywords are matched in both the attributes then we proceed to the next step that is clustering.

### C. Email Clustering

The third stage for the proposed technique is applying the clustering over the similarity information measured in second stage. Based upon the similarity information clustering algorithms can be applied to generate the email clusters. The clusters are created using K-means algorithm Euclidean method the similar emails are stored into one folder and give the appropriate name to that folder.

The growth of email usage has forced users to find ways to organize archive and manage their emails more efficiently. Many of them are organizing incoming messages into separate folders. For example: folders can be topic-oriented like “work”, “personal” and “funny”, people-specific like “John” and “Mary” or group-of-people-specific like “colleagues”, “family” and “friends”. Some users are archiving their messages according to importance and thus maintain folders like “urgent”, “for future reference”, “spam” etc. To achieve this, many users create some so-called rules to classify their mail. Those are heuristic rules searching for keywords in the message and then taking an action like moving or copying to a folder, deleting or forwarding the message etc.

```

if(sender="John Smith" OR sender="Mary Smith")
then (moveInto FAMILY)
  
```

```

if(body contains "call for papers")
then{(moveInto CFP)
  (forwardTo "COLLEAGUES")}[19]
  
```

where FAMILY, and CFP are folders and COLLEAGUES is a group of people (practically a list of addresses the user has created for mass emailing).

### D. Automatic Answering

This is the last phase and here the automated replies are formulated and decided for which group what message should be sent. Large companies usually maintain email centers (in conjunction with “call centers”) with employees committed to answer incoming messages. Those messages usually come from company clients and partners and many times address the same problems and queries. Automatic email answering is an effort to build mail centers or personalized software that will be able to Email Mining analyze an incoming message and then propose or even send an applicable answer.

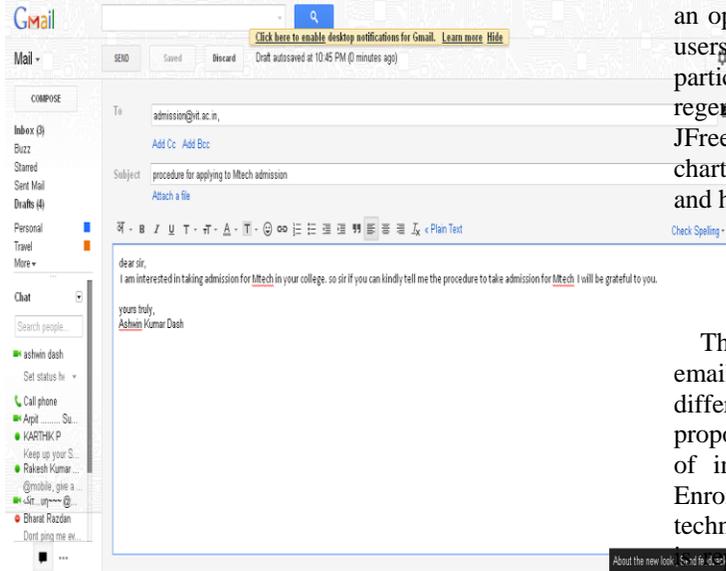


Fig.2



Fig.3

IV. IMPLEMENTATION

The proposed algorithm is implemented using open source technologies and algorithm is applied over the popular email corpus database. Java [15] is selected as the programming languages and the other open source API's (Application Programming Interfaces) to Support the other functionalities. My Eclipse [16] 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 [18] and integrates both proprietary and open source solutions into the development environment. JFreeChart [18] 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 particularly effective for when a user needs to regenerate graphs that change on a frequent basis. JFreeChart supports pie charts (2D and 3D), bar charts, line charts, scatter plots, time series charts, and high-low-open-close charts.

V. EXPERIMENTAL ANALYSIS

The inbox folder is selected for experiment. All the emails from the inbox are taken and are grouped into different clusters. To evaluate the accuracy of the proposed email clustering and finding the similarity of interest in discussion after pre-processing the Enron Email dataset[14], applying the clustering techniques to cluster the Emails and Users, the cluster represented in the graphical form. Here we have assigned numbers to clusters and no. of emails in a particular cluster is shown in the table.

Table-1. Number of Emails Cluster wise

Cluster No.	1	2	3	4	5
No. of Emails	38	98	69	37	30

Figure 4 Tabular representation of no. of emails

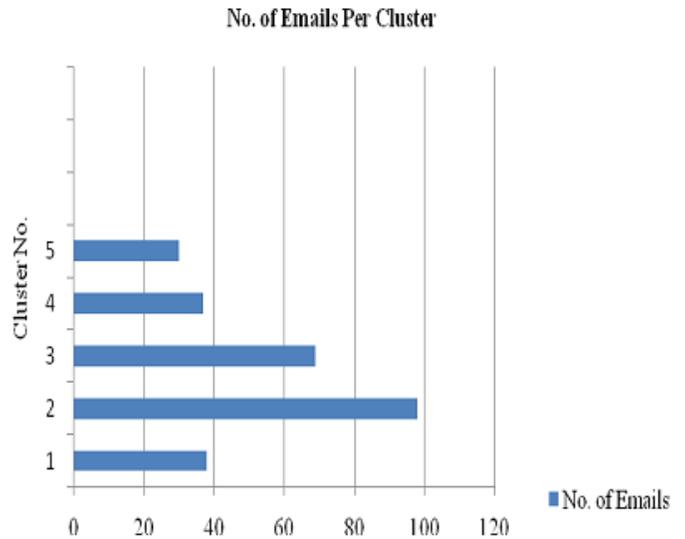


Figure 5 graphical representation

sender_id	reciever_id	subject	message_body	size of message	date & time	auto_answer
rites@gmail.com	admission@vit.ac.in	admission procedure for mtech	mtech.so sir please kindly tell me the procedure to take admission for mtech.		260 12/10/2010 & 9.30 a.m	ok ,you will check our college website www.vit.ac.in.
rajesh@gmail.com	admission@vit.ac.in	respected, sir i so much interested for mtech.sir please suggest me for to take admission in mtech	respected, sir i so much interested for mtech.sir please suggest me for mtech admission .		250 14/7/2010 & 10.20 a.m	
chikun@yahoo.com	admission@vit.ac.in	please suggest me for mtech to apply for admission into mtech.	dear,sir i chikun from bhubaneswar want to take admission into mtech.sir can u help me how take dear,sir i interested for higher study.and i want to join your college for mtech.sir can u guide me what to do to take admission		260 2/10/2010 & 12.30 a.m	
praveen@yahoo.com	admission@vit.ac.in	what is the admission process for mtech.	respected,sir i karthik from orissa ,bhubaneswar interested to know about placement status for btech.sir please tell me about placement percentage.		220 28/10/2010 & 2.30 p.m	ok ,you will check our college website www.vit.ac.in
karthik@rediffmail.com	placement@vit.ac.in	placement status for btech.	placement percentage.		260 1/10/2010 & 12.00 a.m	www.vit.ac.in
aswini@gmail.com	placement@vit.ac.in	what is the placement record for btech	respected,sir i want to know about placement statistics for btech.sir can u please guide me about placement statistics.		250 /10/2010 & 2.00 a.m	
rashmirekha@gmail.com	placement@vit.ac.in	dear sir, i rashmirekha want to know about placement statistics of previous years for btech.sir can u want to know about placement news for btech.	please tell me whats the placement percentage.		275 7/10/2010 & 12.30 p.m	
ramesh_bhatt@gmail.com	m_sinha@vit.ac.in	Dear sir, i ramesh bhatt want to apply for the post of assistant professor.i send my resume through this attachment.sir please applying for the post of assistant check ur mail.	hello,sir i smrutirekha sahu so much interested for the post of assistant professor.i send my resume to your mail id .sir please check it.		260 24/10/2010 & 12.30 p.m	thanks for applying .i will contact u as soon as possible
smruti@gmail.com	m_sinha@vit.ac.in	assistant professor in your college	hello,sir i himanshu pradhan want to apply for the post of assistant professor.i send my resume.sir please reply me		250 5/10/2010 & 7.30 a.m	
himanshu_pradhan@gmail.com	m_sinha@vit.ac.in	apply for assistant professor in software technology			250 21/10/2010 & 8.30 a.m	

**Figure 6 Sample data after clustering**

## VI. CONCLUSION AND FUTURE SCOPE

We have tested our Approach on popular Email dataset and very interesting results have been found. To evaluate the accuracy of the proposed email clustering and finding the similarity of interest in discussion after pre-processing the Email dataset, applying the clustering techniques to cluster the Emails and Users, the cluster is represented in the graphical form for different threshold values, threshold value specifies the no. of times the Email user has used a particular word. Also an automatic answering concept was implemented.

An email clustering Approach is proposed and implemented to show text similarities. The proposed technique shows the email attributes and how the text similarities are used to cluster the users. The proposed technique is implemented using open source technology such as java and an email dataset is selected for the experiment. The future scope of the work could be incorporating the similarity of the email attachments applying the proposed email similarity function for the more email mining operations like thread summarization and finding similarity between images in attachments.

## REFERENCES

- [1] Brief Statistical Report, "Technical Report, Information Sciences Institute, 2004. Available at:  
[http://www.isi.edu/~adibi/Enron/Enron\\_Dataset\\_Report.pdf](http://www.isi.edu/~adibi/Enron/Enron_Dataset_Report.pdf).
- [2] Syeda Farha Shazmeen and Jayadev Gyani "A Novel Approach for Clustering E-mail Users Using Pattern Matching" 978-1-4244-8679-3/11/\$26.00 ©2011 IEEE
- [3] G. Manco, E. Masciari, M. Ruolo, and A. Tagarelli: Towards an Adaptive Mail Classifier. AIAA 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.
- [6] Y. Diao, H. Lu, and D. Wu: A comparative study of classification-based personal email Filtering. In Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD'00), pages 408-419, Kyoto, JP, 2000.
- [7] Ron Bekkerman et al, "Automatic Categorization of Email into Folders: Benchmark Experiments on Enron and SRI Corpora, Technical Report, University of Massachusetts, available at:  
<http://www.cs.umass.edu/~ronb/papers/email.pdf>.
- [8] Naresh Kumar Nagwani and Ashok Bhansali, "An Object Oriented email Clustering Model Using Weighted Similarities between Emails Attributes", International Journal of Research and Reviews in Computer Science (IJRRCS) Vol. 1, No. 2, June 2010.
- [9] L.Adamic, O.Buyukkokten, and E.Adar, "A Social Network Caught in the Web," First Monday, Vol.8, No.6,2003.
- [10] S.A.Golder, D.Wilkinson, and B.A.Huberman, "Rhythms Of social interaction: Messaging within a massive online Network," Proceedings of Third International Conference on Communities and Technologies, pp.41-66, Springer, June, 2007.
- [11] C.Lampe, N.Ellison, and C.Steinfield, "A familiar Face (book): Profile elements as signals in an online social Network," Proceedings of conference on Human Factors in Computing Systems, pp.435-444. ACM Press, 2007.
- [12] K.Yuta,K. N.Ono, and Y.Fujiwara, "A gap in the Community-size distribution of a large scale social Networking site," arXiv.org: physics/0701168, 2007.
- [13] Steve Martin, Anil Sewani, Blaine Nelson, Karl Chen, Anthony D. Joseph, "Analyzing Behavioural features for Email Classification", Second Conference on Email and Anti-Spam CEAS 2005, The International Association for Cryptologic Research and The IEEE Technical Committee on Security and Privacy 2005.
- [14] Bryan Klimt and Yiming Yang, "The Enron Corpus: A New Dataset for Email Classification research", European Conference on Machine Learning, Pisa, Italy, 2004.
- [15] Java, An open source programming language – <http://java.sun.com>
- [16] MyEclipse, MyEclipse is a commercially available Java IDE and AJAX IDE- <http://www.myeclipseide.net/>
- [17] Eclipse, Open language editor for programming <http://www.eclipse.org/>
- [18] JFreeChart, JFreeChart is an open source library available for Java that allows users to easily generate graphs and charts <http://www.jfree.org/index.html/>
- [19] Email Mining: Emerging Techniques for Email Management Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas Aristotle University of Thessaloniki, Department of Informatics, Greece
- [20] Microsoft, "MSN Hotmail Tops 100 Million User Milestone", REDMOND, Wash., 2001.
- [21] J.Robert von Behren, Steven Czerwinski, Anthony D. Joseph, Eric A. Brewer, and John Kubiatiowicz, "NinjaMail: the Design of a High-Performance Clustered, Distributed E-mail System", In Proceeding of International Workshops on Parallel Processing 2000, Toronto, Canada, August 21-24, 2000, pp. 151-158

**First Author: Mr. Praveen Kumar:** I was Born In Sultanpur(U.P.), India. I Received My B.TECH Electronic and Communication Engineering Degree From Integral University,Lucknow,U.P. , India in 2010. Currently I Am Pursuing My M.Tech In Software Technology At Vit University, Vellore, Tamil Nadu, India. I am a good player in Badminton and Cricket. My Major Research Interests are in Information Security,Software Development and Testing,Cloud Computing,Operating System,Data Structure and Algorithm, RDBMS, AGILE Methodology . I have Completed Successfully Research Project Entitled SQL-Injection Vulnerability Scanning Tool for Automatic Creation of SQL-Injection Attack on JSP.Currently I am doing my final project on Design and Development of Cloud-Computing Based District Resource Planning System.



**Second Author : Himanshu Kumar :** I have completed my B.E in computer science and Engineering Degree From Manipal University,Manipal,Karnataka,India in 2009. Currently I am pursuing M.Tech In information Technology(Networking) at VIT University,Vellore,Tamil Nadu, India.My Major Research area are Information

Security,Software Development and Testing, Cloud computing ,wireless adhoc network,Operating System,Data Structure and Algorithm, RDBMS. I have Completed Successfully Research Project Entitled rainbow table to crack password using md5 hashing algorithm..Currently I am doing my final project on web usage mining trends and navigational pattern.



**Third Author: Remya Joseph:** I have completed my B.E in computer science and Engineering Degree From Mahatma Gandhi (MACE) University,kothamangalam,Kerala,India in 2010. Currently I Am Pursuing My M.Tech In information Technology(Networking) at Vit University,Vellore,Tamil Nadu, India.My Major Research area are in Information Security,Software development and Testing,Cloud computing,wireless adhoc network,Operating System,Data Structure and Algorithm, RDBMS. I have Completed Successfully Research Project Entitled Efficient web usage mining in formal concept analysis.Currently I am doing my final project on automating IOT Cases in Robot Framework.