Email Mining: A Review

Mrs. Pranjal S. Bogawar[1], Dr. Kishor. K. Bhoyar[2]

[1]Information Technology, R.T.M. Nagpur University, Priyadarshini College of Engineering,
Nagpur, Maharashtra, India

[2] Information Technology, R.T.M. Nagpur University, Yashavantrao Chavhan College of Engineering,
Nagpur, Maharashtra, India

Abstract

E-mail is one of the most widely used ways of written communication over the internet, and its traffic has increased exponentially with the advent of World Wide Web. The increase in email traffic comes also with an increase in the use of emails for illegitimate purpose. Phishing, Spamming, email bombing, threatening, cyber bullying, racial vilification, terrorist activities, child pornography and sexual harassment are common examples of e-mail abuses. So, there is a need for e-mail mining. Various methods and approaches were used by the scientists for classification of email messages in above categories. In this paper we are presenting various techniques and approaches used by researchers for email mining and subsequent classification.

Keywords: *email, email mining, spam, header body, MIME, SMTP, POP*

1. Introduction

According to the survey of Radicati group from April 2010, there are about 1.9 billion users of email worldwide [1]. As the popularity of email increased, it becomes an important form of communication for many computer users, for both legitimate and illegitimate activities. Legitimate activities are like messages and document exchanges which are also misused eg., distribution of junk mails, unauthorized conveyance of sensitive information, mailing of offensive or threatening material.

Email system is inherently vulnerable to misuse for three main reasons. First, an email can be spoofed and metadata contained in its header about the sender and the path along which the message has travelled can be forged or anonymzed. An email can be routed through anonymous e-mail servers to hide the information about its origin. Second, e-mail systems are capable of transporting executables, hyperlinks, Trojan horses, and scripts. Third, the internet including email services is accessible through public places, such as net cafe and libraries. Hence there is a need of email mining. Many authors worked on emails. This paper is an overview of them.

2. E-MAIL

Electronic mail, commonly known as email or e-mail, is a method of exchanging digital messages from an author to one or more recipients. Ray Tomlinson is generally credited as having sent the first email across a network, initiating the use of the "@" sign to separate the names of the user and the user's machine in 1971, when he sent a message from one Digital Equipment Corporation DEC-10 computer to another DEC-10. The two machines were placed next to each other. Tomlinson's work was quickly adopted across the ARPANET, which significantly increased the popularity of email.

Modern email operates across the internet or other computer networks. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver and store messages [16, 17].

Email’s have specific format which is defined in RFC(Request for Comment) 5322, with multi-media content attachments being defined in RFC 2045 through RFC 2049, collectively called Multipurpose Internet Mail Extensions or MIME.

2.1. Message Format

Internet email messages consist of two major sections:

- **Header** — the message header contains control information, an originator's email address and one or more recipient addresses and descriptive information, such as a subject header field and a message submission date/time stamp which is structured into fields such as From, To, CC, Subject, Date, and other information about the email.
- **Body** — body content is unstructured text which sometimes contains a signature block at the end. This is exactly the same as the body of a regular letter.

The header is separated from the body by a blank line.

2.1.1. Message Header:

Each message has exactly one header, which is structured into fields. Each field has a name and a value. RFC 5322 specifies the precise syntax. Informally, each line of text in the header that begins with a printable character begins a separate field. The field name starts with the first character of the line and ends before the separator character ":". The separator is then followed by the field value (the "body" of the field).

Header fields: The message header must include at least the following fields:

- *From:* It is the email address, and optionally the name of the author(s).

- *Date:* The local time and date when the message was written.
- *Message-ID:* Also an automatically generated field; used to prevent multiple deliveries and for reference in In-Reply-To: (see below).
- *In-Reply-To:* Message-ID of the message which is a reply to. It is used to link related messages together. This field only applies for reply messages.

Along with above header fields some common header fields of email which every person is using are:

- *To*: The email addresses, and optionally name(s) of the message's recipient(s). Indicates primary recipients (multiple allowed), for secondary recipients see Cc: and Bcc: below.
- *Bcc:* Blind Carbon Copy; addresses added to the SMTP delivery list but not (usually) listed in the message data, remaining invisible to other recipients.
- *Cc:* Carbon copy; many email clients will mark email in your inbox differently depending on whether you are in the To: or Cc: list.
- *Subject:* A brief summary of the topic of the message. Certain abbreviations are commonly used in the subject, including "RE:" and "FW:"
- *Content-Type:* Information about how the message is to be displayed, usually a MIME type.
- *Precedence:* commonly with values "bulk", "junk", or "list".
- *Received:* Tracking information generated by mail servers that have previously handled a message, in reverse order (last handler first).
- *References:* Message-ID of the message that this is a reply to, and the message-id of the message the previous reply were a reply to, etc.
- *Reply-To:* Address that should be used to reply to the message.
- *Sender:* Address of the actual sender acting on behalf of the author listed in the From: field.
- *Archived-At:* A direct link to the archived form of an individual email message.[18]

2.1.2 Message Body

In the SMTP (Simple mail Transfer Protocol) standard, the body is the full email message. Most modern graphic email clients allow the use of either plain text or HTML for the message body at the option of the user. HTML email messages often include an automatically generated plain text copy as well, for compatibility reasons.

In order to ensure that HTML sent in an email is rendered properly by the recipient's client software, an additional header must be specified while sending: "Content-type: text/html". Most email programs send this header automatically.

2.2. Working of E-Mail

Email relies on two basic communications protocols: SMTP (Simple Mail Transfer Protocol), which is used to send messages and POP3 (Post Office Protocol), which is used to receive messages. A simplified version of the email life cycle can be seen in Figure 1.

The most important logical elements of the Internet Mail System are:

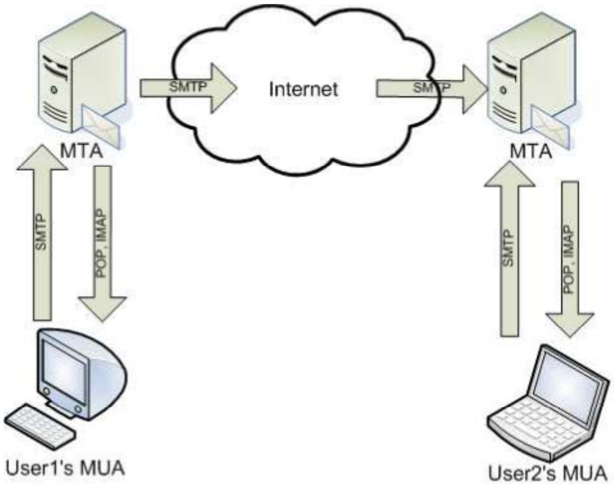

Fig 1: Life Cycle of an email [19]

1) Mail User Agent (MUA*)* – It is responsible for helping the user to read and write email messages. The MUA is usually implemented in software usually referred to as “email client”. Popular email clients are Microsoft Outlook2 and Mozilla Thunderbird3, claws mail, Zimbra Collaboration Suite etc. These programs transform a text message into the appropriate internet format in order for the message to reach its destination.

2) Mail Transfer Agent (MTA**):** It accepts a message passed to it by either an MUA or another MTA and then decides for the appropriate delivery method and the route that the mail should follow. It uses the SMTP protocol to send the message to another MTA or an MDA.

3) Mail Delivery Agent (MDA): It receives messages from MTAs and delivers them to the user’s mailbox in the user’s mail server

4) Mail Retrieval Agents (MRA): It fetches mail messages from the user’s mail server to the user’s local inbox. MRAs are often embedded in email clients [19].

2.3. Filename Extensions

Upon reception of email messages, email client applications save messages in operating system files in the file system. Some clients save individual messages as separate files, while others use various database formats, often proprietary, for collective storage. A historical standard of storage is the mbox format. The specific filename extensions are in table1.

Table 1: Specific formats used for filename extension

Email Extension	***Using Agency***
eml	Used by many email clients including Microsoft Outlook Express, Windows Mail and Mozilla Thunderbird. The files are plain text in MIME format, containing the email header as well as the message contents and attachments in one or more of several formats.
emlx	Used by Apple Mail.
msg	Used by Microsoft Office Outlook and Office Logic Groupware.
mbx	Used by Opera Mail, KMail, and Apple Mail based on the mbox format.

3. Email mining

Email mining can be considered as mining of data embedded in header and or body of the email message. Various text mining techniques which extract unknown and useful information from huge set of emails can be employed to achieve email mining. Email Mining can be considered as an application of the upcoming research area of Text Mining (TM or also known as Knowledge Discovery from Textual Data) on email data.

However, there are some specific characteristics of email data that set a distinctive separating line between Email and Text Mining:

1. Email includes additional information in the headers of email that can be exploited for various email mining tasks.
2. Text in email is significantly shorter and, therefore, some Text Mining techniques might be inefficient in email data.
3. Email is often cursorily written and, thus, linguistic well-formedness is not guaranteed [19]. Spelling and grammar mistakes as well as nonstandard user acronyms also appear frequently.
4. Email is personal and therefore generic techniques are difficult to be effective to individuals.
5. Email is a data stream targeted to a particular user and concepts or distributions of target classes of the messages may change over time, with respect to the messages received by that user.
6. Email will probably have noise. HTML tags and attachments must be removed in order to apply a text mining technique. In some other cases, noise is intensively inserted. In spam filtering for example, noisy words and phrases are inserted, in order to mislead machine learning algorithms.
7. It is rather difficult to have public email data for experiments, due to privacy issues. This is a drawback especially for research since comparative studies cannot be conducted without public available datasets. An exception to the above statement is the Enron Corpus (Klimt & Yang, 2004), which was made public after a legal investigation concerning the Enron Corporation [19, 26].

Email mining is done by various researchers to extract different information from email. Some topics for which invention is done are discussed below.

3.1. Authorship Attribution

Email authorship attribution means identify the most plausible author of an anonymous email from a group of potential suspects. For author attribution various techniques used by various authors. The various topics on which work was done are gender, language, various writing styles.

Oliver de Vel et.al. used combination of stylometric, structural, gender preferential features, and language preferential features together with support vector machine algorithm to classify author's gender and language. The author also classified the language of the person as EFL (English as first language) and ESL (English as second language). Researchers also classified English and Arabic language [10]. Farkhund Iqbal et.al. gave a data mining technique to capture the write prints of every suspect and model it as combination of features that occur frequently in the suspect's email called frequent patterns. Every person has unique identity, features and writing styles. Writing patterns usually contain the characteristics of word usage, word sequence, compositions, layouts, common spelling and grammatical mistakes, vocabulary richness, hyphenation and punctuations.

3.2. Content Analysis

"Content analysis is a summarising, quantitative analysis of messages that relies on the scientific method (including attention to objectivity, intersubjectivity, a priori design, reliability, validity, generalisability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the context in which the messages are created or presented."

As communication on e-mail get increased, marketing, sales, customer services and dedicated call centres required to process high volumes of emails with many messages having repetitive enquiries. Indrajit Mukerjee et.al created the automatic email answering mechanism by finding the keywords of product in email's of client.

In another research of content analysis relationships of email files were found by using keyword based matching techniques. They created a tool called Email Mining Set (EMS) for analyzing email archives which includes a graphical display to explore the relationship between users and groups of email users [14].

Appavu alias Balamurugan et.al., introduced the new Ad Infinitum algorithm to classify the threatening messages. Ad infinitum would be the extension of the decision tree induction algorithm.

Vatcharaporn Esichaikul et.al. proposed a simple and intuitive model to locate significant messages and users from an analysis of email message.

3.3. Phishing

Phishing can be defined as a scam by which email users are duped into surrendering private information that will be used for identity theft. Phishing attacks use both social engineering and technical subterfuge to steal personal identity data and financial account credentials. It is one of the fastest growing scams on the Internet. The exclusive motivation of phishers is financial gain.
John Yearwood et. al. used the structural characteristics of the emails received by persons and the information derived on hyperlinks from 'Whois Database' for profiling of phishing emails. While generating profiles author used structural characteristics such as text content, vlinks, html content, script, table, images/logos, hyperlinks, form tag, fake tags etc.

3.4. Spam Filtering

Spam, also known as unsolicited bulk email (UBE), is becoming increasingly harmful for email traffics. Filtering is a simple and efficient way to combat against spam. Machine-learning-based classification algorithms are of excellent performance in filtering spam.

An email naturally has no label indicating whether it is a spam or a legitimate message. The label can only be set by humans for accuracy, which needs heavy manual labor. This challenge is especially significant for email service providers (ESPs) because the spam filtering system deployed on the email servers need to be retrained frequently to meet the rapidly changing spam. Figure 4 shows the email behavior model and its applications.

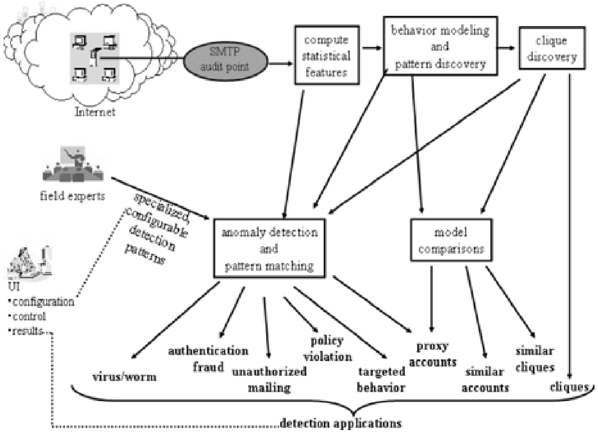

Fig. 2: Overview of email behaviour modelling, architecture and applications [20].

Yong Hu et.al suggested the fuzzy clustering method for spam and legitimate emails. They proposed spam filter consists of four components, namely, "feature extractor", "fuzzy clustering algorithm", "labeling algorithm" and "adjusting algorithm"[11].

Chun Wei et.al. is concentrating on the advanced analysis of spam emails, by considering eleven attributes from the message: message id, sender's IP address, sender email,subject, body length, word count, attachment filename, attachment_MD5, attachment size, body_URL, body_URL_domain. Some attributes were again broken. E.g. body_URL into machine name and path[13].

Salvatore J. Stolfo et.al.; gave a data mining system called EMT(Email Mining tool kit) which is used for core security applications to detect virus propagations, "spambot" activity and security policy violation. They used behaviour based analysis rather than content analysis[20, 21, 22].

4. Algorithms used in email mining

There are various algorithms used for email mining which are given below.

4.1 Support Vector Machine Algorithm:

The original SVM algorithm was invented by Vladimir Vapnik. SVM concept is based on the idea of structural risk minimisation which minimizes the generalization error. The advantage of SVM is that they do not require a reduction in number of features in order to avoid the problem of over fitting, which is useful when dealing with large dimensions as encountered in the area of text mining. It is a learning machine that classifies an input Vector X using decision function:
$f(X)= <X,W> + b$ ………(1)
SVMs are hyper plane classifiers and work by determining which side of hyper plane classifiers and work by determining which side of the hyper plane X lies. In the above formula given in eq. no. 1 the hyper plane is perpendicular to W and at a distance b/‖ W‖ from the origin.
SVM maximize the margin around the separating hyper plane. The decision function is fully specified by a subset of training samples [7, 8, 23, 24, and 25].

4.2 Naive Bayes Algorithm:

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. It is used the knowledge of prior events to predict future events. Baye's Theorem says that

$$P\left(\frac{h}{D}\right) = \frac{P(\frac{h}{D})}{P(D)} \text{.......................(2)}$$

where
P(h) : Prior probability of hypothesis h
P(D) : Prior probability of training data D
P(h/D) : Probability of h given D
P(D/h) : Probability of D given h[27]

4.3 Clustering Algorithms:

Researchers used some clustering algorithms they are

4.3.1. Expectation Maximization (EM):

The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin [31].
EM is an iterative optimization method to estimate some unknown parameters Θ, given measurement data U. However, we are not given some “hidden” nuisance variables J, which need to be integrated out. In particular, we want to maximize the posterior probability of the parameters Θ given the data U, marginalizing over J:

$$\Theta = \text{argmax}_{\Theta} \sum_{J \in \mathcal{J}^n} P(\ \Theta\ , J|U)$$

The intuition behind EM is an old one: alternate between estimating the unknowns Θ and the hidden variables J. This idea has been around for a long time. However, instead of finding the best J ∈ $\mathcal{J}$ given an estimate Θ at each iteration, EM computes a distribution over the space $\mathcal{J}$[28, 29, 30 ,31].

4.3.2. K-Means:

K-means clustering (MacQueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:
1. Each instance di is assigned to its closest cluster center.
2. Each cluster center Cj is updated to be the mean of its constituent instances [32].

4.3.3. Bisecting K-means:

It produces the clusters of the similar sizes and with smaller entropy than K-means [33].

4.3.4. Agglomerative Algorithm:

This algorithm starts with each individual item in its own cluster and iteratively merges clusters until all items belong in one cluster.

4.4 Behaviour Based Models:

1) The *user cliques* model profiles a user’s communication groups that naturally occur in her or his email communication history. These cliques models provide important information to rank order the relative importance of individuals in an organization.
2) The *Hellinger distance* model profiles the distribution of the frequency of communication of users, and the variability of frequency, between a user and his/her correspondents. The recipient frequency analysis also identifies the relative importance of various email users. By extending the analysis to compute the response rates to a user’s typical recipients, one can learn relative rank ordering of various people.
3) The *cumulative distribution* model profiles the rate at which a user sends emails to distinct parties in sequential order. A virus would generally not know this statistics and so would violate the user’s typical behaviour while propagating itself to new victim

5. Conclusions

Email is very popular and necessity of many users. The paper gives an overview of email, its message format and working. Now a day’s criminal are also using emails. So, email is considered as powerful evidence. In this paper, have presented the research carried out in the field of email mining. Researchers worked on the email body and found gender, language, writing styles of author. They found the relationships between users, threatening messages. Researchers also found the profiles of phishers. In this paper we have also presented the overview of spam messages and spam filters. Some spam filters used contents of emails; some used link information, while others used behavioural information.

6. References

[1] http://www.radicati.com
[2] Farkhund Iqbal, Rachid Hadjidj, Benjamin C.M. Fung, Mourad Debbabi, “A novel Approach of mining write prints for authorship attribution in email forensic”, Digital Investigation(Elsevier Journal)2008,pp42-51
[3] Farkhund Iqbal, Hamad binsalleeh, Benjamin C.M. Fung, Mourad Debbabi, “Mining writeprints from anonymous emails for forensic investigation”, Digital Investigation(Elsvier Journal)2010, pp1-9
[4] Robert Layton, Paul Watters, Richard Dazeley, “Authorship Attribution for Twitter in 140 characters or less”, second Cybercrime and Trustworthy Computing workshop, 2010 IEEE, pp-1-8
[5] Nouf Al Fe’ar, Einas Al Turki, Asma Al Zaid, Mashael Ai Duwais, “E-Classifier: A Bi-Lingual Email Classification System”, Information technology, ITSim2008, IEEE,Vol-2, pp-1-4
[6] Indrajit Mukerjee, Mohammad Al-Fayoumi, P.K. Mahanti, “Content Analysis based on text Mining using genetic algorithm”, Second internation conference on Computer technology and Development(ICCTD 2010),IEEE, pp -432-436
[7] Oliver de Vel, Malcolm Corney, Alison Anderson, and George Mohaly, “ Language and Gender Author Cohort Analysis of Email for Computer Forensics”, In Proc. Of digital forensic workshop(2002),pp-1-16
[8] O.De Vel, A. Anderson, M. Corney, G. Mohaly, “Muti-Topic E-mail Authorship Attribution Forensic”, ACM Conference on Computer Security- Workshop on data Mining for Security Applications, November 8,2001,pp: 1-7
[9] John Yearwood, Musa Mammadov, Arunava Banerjee, “ Profiling Phishing Emails Based on Hyperlink Information”, International conference on Advances in Social Network Analysis and Mining, 2010 IEEE,pp-120-126
[10] Nouf Al Fe’ar, Einas Al Turki, Asma Al Zaid, Mashael Ai Duwais, “E-Classifier: A Bi-Lingual Email Classification System”, Information technology, ITSim2008, IEEE,Vol-2, pp-1-4
[11] Yong Hu, Ce Guo, Xiangzhou Zhang Zhihui Guo, Jing Zhang, Xiaojuan He, “An Intelligent Spam Filtering System Based on Fuzzy Clustering”, 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, pp515-519
[12] Appavu alias Balamurugan, Rajaram,Muthupandian and Athiappan, “ Automatic mining of threatening e-mail using

Ad Infinitum algorithm, In International Journal of Information Technology,Vol14. No.2,2008

[13] Chun Wei, Alan Spragur, Gary Warner, Anthony Skjellum, “Mining spam Email to identify Common Origins for Forensic Applications”, ACM 2008, pp: 1433-1437

[14] Hongjun Li, Jiangang Zhang, Haibo Wang, Shaoming Huang,” A Mining Algorithm For Email’s Relationships Based on Neural Networks”,International Conference on computer science and Software Engineering ,2008 IEEE, pp- 1122-1125

[15] Appavu alias Balamurugan, Rajaram,Muthupandian and Athiappan, “ Automatic mining of threatening e-mail using Ad Infinitum algorithm, In International Journal of Information Technology,Vol14. No.2,2008

[16] "The Technical Development of Internet Email" Craig Partridge, April–June 2008, pp.5

[17] “The First network Email”, http://openmap.bbn.com/~tomlinso/ray/firstemailframe.html

[18] P. Resnick, “Internet Message Format”, Qualcomm Incorporated, October 2008, http://tools.ietf.org/html/rfc5322

[19] Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas, “Email Mining: Emerging Techniques for Email Management”, 2006, pp- 1-32

[20] Salvatore J. Stolfo, Sholomo Hershkop, Chia-Wei Hu, Wei-Jen Li, Olivier Nimeskern, and Ke Wang, “ Behavior Based Modeling and its Application to Email Analysis ”,ACM transactions on Internet technology, Vol6, No.2 May 2006, pp187-221

[21] Salvatore J. Stolfo, Sholomo Hershkop,Ke Wang, Oliver NImeskern, Chia Wei Hu, “ Behavior Profiling of Email”, ISI 2003,LNCS2665, pp74-90

[22] Salvatore J. Stolfo, Shlomo Hershkop, “Email Mining Toolkit Supporting Law enforcement Forensic Analyses”, ACM international conference Proceeding series(2005) , pp -221-222

[23] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, “A Practical Guide to Support Vector Classification”, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, April 2010

[24] Steve Gunn, “Support Vector Machines for Classification and Regression”, ISIS Technical Report http://www.svms.org/tutorials/Gunn1998.pdf, 14 May 1998, pp:1-52

[25] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, New York. 1995.

[26] Enron Email Dataset, http://www.cs.cmu.edu/~enron/

[27] Naive Baye’s Classification Algorithm, http:// software.ucv.ro /~cmihaescu/ro/teaching/AIR/docs/Lab4-NaiveBayes.pdf ,pp1-17

[28] Frank Dellaert, “The Expectation Maximization Algorithm”, http://www.cc.gatech.edu/~dellaert/em-paper.pdf,pp1-7

[29] Chuong B Do & Serafim Batzoglou, “What is the expectation maximization algorithm?”, Nature Publishing Group http://www.nature.com/naturebiotechnology , 2008,pp1-3

[30] Sean Borman, “The Expectation Maximization Algorithm A short tutorial”, http://www.seanborman.com/ publications/ EM_algorithm.pdf, January 09, 2009 ,pp1-9

[31] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm”, Journal of the Royal Statistical Society. Series B (Methodological)
Vol. 39, No. 1 (1977), pp. 1-38

[32] Kiri Wagstaff,Claire Cardie, Seth Rogers Stefan Schroedl, “Constrained K-means Clustering with Background Knowledge”, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584.

[33] Sergio M. Salveresi and Daniel Boley, “A comparative analysis on the bisecting K-means and PDPP clustering algorithms”, http://www-users.cs.umn.edu/~boley/ publications / papers / savaresi04.pdf, pp1-18

[34] Devi Prasad Bhukya and S. Ramachandram, “Decision Tree Induction: An Approach for Data Classification Using AVL-Tree”, International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010, pp 660-665

Mrs. Pranjal S. Bogawar has completed her B.E. (Computer technology) in 2000 from Rashtrasant Tukadoji Maharaj Nagpur University and M.E.(Computer science and Engineering) in 2009 from Sant Gadge Baba Amravati University of India. She is assistant professor in the department of Information Technology of Priyadarshini College of Engineering, R.T. M. Nagpur University at Maharashtra, India. She published 3 national, 3 international and one journal paper. Her research interests are Database Management System, Data mining and Email mining. She is Life member of Indian Society for Technical Education.

Prof. Kishor Bhoyar has completed his B.E. (Computer Science & Engineering) and M.E.(Computer Tech.) Degrees from Dr. Babasaheb Ambedkar Marathwada University and Swami Ramanad Teertha Marathwada University of India respectively in the years 1990 and 2001 respectively. He has achieved his Ph.D. degree in Computer Science and Engineering from Vishweswarayya National Institute of Technology, Nagpur, India in the year 2010. He is a professional member of ACM, Associate member of Computer Society of India and Life member of Indian Society for Technical Education. His areas of interest include Image Processing, Soft Computing and Data Mining.