

# “Critical Analysis and Study of Gmail Mail Classification Strategies”

Javed Akhtar Khan<sup>1</sup>, Dr.M.R Aloney<sup>2</sup>

<sup>1,2</sup>Dept of Computer Science &Engineering

<sup>1,2</sup>Bhagwant University ,Ajmer (INDIA)

**Abstract-** Gmail services widely used for exchanging the information. Gmail or any other mail services is able to provide the lots of services to user these are shearing the data , information, sending the file in different different format like document file format , pdf file format, Ppt, shearing the address link etc . This is the public platform for shearing the relevant information with other , promote the business idea , shearing advertisement etc , for this Gmail make a kind of changes in the Gmail Templates .In this article we are include analysis and study of Gmail mail classification polices those are already proposed by the authors.

**Keywords-** Gmail , Gmail Classification , Storage , Mail .

## I. LITERATURE SURVEY

This part of paper include the empirical study of various existing research paper those are directly related with Mail Classification . During the Literature study we are collect the some important point and suggestion for the Mail Classification based on the filtered parameters , unique polices, new frame work , novel approach etc.

Ref [1]-A Text Classification, E-mail filtering and email organization is an application rife with the potential to streamline the management of the vast amount of information that accumulates in the inbox. In this paper author categorize the prevalent popular techniques for classifying email as spam or legitimate and suggests possible techniques to fill in the lacunae in the arena of automatic management of emails. The proposed solution of this paper is suggest that context-based email organization has the most potential in improving quality by learning various contexts such as n-gram phrases, linguistic constructs or users’ profile based context to tailor his/her filtering scheme. In the introduction part of this paper include the brief information about the TC that this text classification task of automatically sorting a set of documents into categories such as topics from a predefined set. The task falls at the crossroads of information retrieval (IR) and Machine Learning (ML). during this work author are provide the result in the form of summarization , easy sorting and user- specific presentation, researcher are mainly investigate TC approaches for personalized management of all email .in this work author

also categories the inbox mail in the spam mail and non spam mail . The statistical approaches have been applied for text classification. These approaches are based on the word occurrences i.e. frequency of one or more words in a given document.

Ref[2]-The major objective of this paper is E- mail Classification. So in this paper researcher are proposed the new frame work this is used the E- Mail Classification . A Tournament-based approach is used for E- Mail Classification. In this paper the classification method passes through three phases: 1) Clustering (grouping) email folders (topics or classes) based on their token and field similarities, 2) Training binary classifiers on each class pair and 3) Applying 2-layer tournament method for the classifiers of the related classes in the resultant clusters. The first phase evolves K-mean algorithm to result in cluster sizes of 3, 4, or 5 email classes with the pair wise similarity function. The second phase uses two classifiers namely Maximum Entropy (MaxEnt) and Winnow. The third phase uses a 2-layer tournament method which applies round robin and elimination tournament methods sequentially to realize the winner class per cluster and the winner of all clusters respectively. The proposed method is tested for various K settings against tournament and N-way methods using 10-fold cross-validation evaluation method on Enron benchmark dataset. The experiments prove that the proposed method is generally more accurate than the others.

Ref[3]- Various E- Mail classification technique is introduce in this paper for classified the Spam mail messages are mainly a illegal form where thousands of users are easily reachable. This type of emails has become a serious problem. In this paper author proposed a method of classifying emails using the concept of association rules. Association rule mining is used for finding interesting hidden patterns in large transactional databases. Association rules are rules that identify associations between items in transactions. The proper exploration of Mail using the Association Rule Mining is built for emails classification system. In this paper author are make some section like Training phase in the training Phase author are include the Emails Collection, Emails Preprocessing Tokenization Process , Stop Words Elimination Process ,

Suffix Stripping Process and many more , for the implementation part of this paper author are make the frequent set and calculate the pattern and make association rule for the e-mail .

Ref[4]- A general concept of Mail mining and Text Mining is include in this survey paper .In this paper researcher are work in five major area these are email mining tasks, namely, spam detection, email categorization, contact analysis, email network property analysis and email visualization. This work done In order to facilitate better usage of emails and explore business potentials in emailing, various data mining techniques have been applied on email data. The brief explanation of all five area Spam detection is to detect unsolicited bulk emails. It is the most important task in email mining. Email categorization (or email filling) is to organize emails into different categories. Categorizing emails manually becomes a heavy burden for email users when the amount of emails grows fast. Contact analysis is to identify special email contacts or groups of email contacts by analyzing the contacts' characteristics from the email contents or email social relations, Email network property analysis is to analyze the critical properties of an email network, such as general network structures, relation strength and organizational structures. Email visualization is to use visualization techniques to help users identify, retrieve and summarize useful information hidden in numerous emails. Email visualization can help to tackle the challenges like how to optimize email interface to provide users more convenience and how to display useful hidden information in one's email account. A good user interface design can provide users understandable access to their email accounts and benefit users in synthesizing information and deriving insights from emails. This survey is organized mainly according to email mining tasks. In this paper author proposed the new framework is shown figure 1.1 .

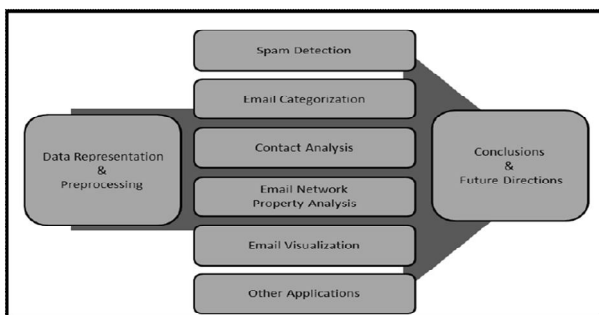


Figure 1.1 Framework For Mail Classification

This paper also include review of some widely used techniques, and include introduction of software tools available for mail classification .

Table 1.1 List of All Technique

S.No	Mail Domain	Approach Used	List of Methods
1	Spam detection	1-Content Based Detection 1.1-Classification Methods 1.1.1 Naive Bayes Classifiers 1.1.2-Support Vector Machines Classifiers 1.1.3-Rule Based Classifiers 1.1.4-Semi-supervised Clustering Methods 2-Semi-supervised Clustering Methods 2.1 K-means algorithm 3-Sender Based Detection	3.1 - Classification Methods 3.1.1- K-Nearest Neighbor (K-NN) Classifiers 4-Sender Based Detection 4.1- K-Nearest Neighbor (K-NN) Classifiers 4.2-Semi-supervised Clustering Methods 4.3- Email Sender Reputation Analysis Methods
2	Email categorization	1-Classification Methods 1.2-Naive Bayes Classifiers	1.3-Support Vector Machine Classifiers
3	Contact analysis	1-Contact Identification Classification Methods Support Vector Machines Classifiers (SVM)	2- Social Network Analysis (SNA) Methods 3- Contact Categorization 3.1- Clustering Methods 3.1(a)- Girvan-Newman Algorithm
4	Email network property analysis	1-Social Network Analysis (SNA) Methods	2-Classification Methods 2.1-SVM classifier
5	Email visualization	1-Survey and Quantitative Analysis Methods	--

Ref[5]- A Mail classification for the Spam Data Domain is proposed In this paper, spam dataset is analyzed using TANAGRA data mining tool to explore the efficient classifier for email spam classification. Researcher are applied the various classification algorithms over the Spam Data Domain and cross validate the Spam mail Finally, best classifier for email spam is identified based on the error rate, precision and Recall .In this paper author are used the 15 different classification algorithm for the evaluation. during this evaluation process researcher are make a static of all algorithm working process and make conclusion , this paper is mainly work for the spam detection in this paper author are used the spam data set from UCI machine learning repository is taken as input data for analyzing the various classification technique using TANAGRA data mining TOOL. In this work, feature construction and feature selection is done first to select the relevant features for classification. After feature extraction, 15 different classification algorithms are taken for evaluation. In this evaluation process, different features are considered for choosing best spam filtering algorithm. Finally performance evaluation is done to analyze the various classification algorithms to select the best classifier for spam emails.

Ref[6]-This paper include the Email Classification using the concept of Data Reduction Method, the effective and efficient email classification technique those support the filtering method is introduce in the paper .This paper include the innovative filtering technique using instance selection method (ISM) to reduce the pointless data instances from training model and then classify the test data. The objective of ISM is to identify which instances (examples, patterns) in email corpora should be selected as representatives of the entire dataset, without significant loss of information. This paper proposes effective and efficient email classification techniques based on data filtering method into the training model. The main focus of this paper is to reduce the instance of the email

corpora from training model using ISM, which is less significant in relation to the classification.

Ref [7]- This paper surveyed Naïve Bayes, Support Vector Machine, K-Nearest Neighbor from supervised and K-Means, Fuzzy C-Means, Evolving Email Clustering Method from unsupervised algorithms for email management. These algorithms are discussed in brief along with their working strategy and their performance for email domain. The email management is the task of managing emails into different categories or classes or folders according to some criteria or condition. The email management task is talked with the help of machine learning algorithms. This paper include the survey of many classification and clustering algorithms applied for email management.

**Analysis Table – This is Analysis of above study for mail Classification in the tabular form, this table is include the**

RefNo	Major Objective Cover	A Novel Approach	Algorithm used	Model
Ref[1]	Text classification of e-mail, email text filtering. The main objective of this paper is finding suggest that context based email organization.	Some of popular technique are mention, with automatic management of e-mail	---	Improving quality by learning various contexts like n-Gram Phrases Linguistic Constructs and user profile and some of the filtering scheme are used
Ref[2]	E-mail Classification Approach are used that is based on the rule of any tournament. In other word we can say that this is clustering base approach.	A Tournament Based Method are used, or we can say the elimination round are used for filtering the data or e-mail	A different Algorithm are used for the different phase like first make a cluster, apply the K mean algorithm and then round robin method for the elimination	NO
Ref[3]	This paper is also based on the Classifying E-mail	Use the Data mining Technique Data Mining Association Rule	Classification Algorithm Used	NO make only the Association Rule using the Frequent itemset
Ref[4]	This is a survey paper in this paper author are work for the five major area these area are 1-Spam Detection 2-Email categorization 3-contact analysis 4-Email network property 5-Email- Visualization	Email Mining with five parameter. Mention above	This is a survey paper in this paper researcher include the so many algorithm that is based on the various different approach.	NO
Ref[5]	Spam Data set are analyzed	ANAGRA Data mining Tool used	Classification Algorithm used	No
Ref[6]	Email Classification	Used the concept of Data Reduction Method. The main focus of this paper is to reduce the instance of the email corpora from training model using ISM, which is less significant in relation to the classification.	A data filtering method are used in this paper. Classification Algorithm are used	NO
Ref[7]	This is survey paper based on the Supervised and unsupervised Algorithm in this paper author are make a Email management Task.	In this paper author are make a Email management and make a survey of some classification algorithm.	NO supervised and unsupervised algorithm are used	NO

International Journal of Engineering and Technology, Vol.3, No.2, April 2011

[2] Ref[2]- Sabah Sayed et al “THREE-PHASE TOURNAMENT-BASED METHOD FOR BETTER EMAIL CLASSIFICATION” International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.6, November 2012 DOI : 10.5121/ijaia.2012.3606

[3] Ref[3]- Mohammed A et al “Emails classification by data mining techniques “ Journal of Babylon University/Pure and Applied Sciences/ No.(2)/ Vol.(22): 2014 .

[4] Ref[4]-Guanting et al “ Email Mining: Tasks, Common Techniques, and Tools” School of Computing Science, Simon Fraser University, Burnaby BC, CANADAReceived Jan 17, 2012 Revised Mar 04, 2013 Accepted Apr 06, 2013

[5] Ref[5]- R. Kishore et al “Comparative Study on Email Spam Classifier using Data Mining Techniques” IAENG IMECS2012 ISBN-9789851925114

[6] Ref[6]- Rafiqul Islam et al “Email Classification Using Data Reduction Method “School of Information Technology Deakin University, Burwood 3125, Victoria, Australia

[7] Ref[7]- Jainesh Patel et al “Survey of Supervised and Unsupervised Algorithms in Email Management” International Journal of Scientific & Engineering Research, Volume 5, Issue 3, March-2014 ISSN 2229-5518

**REFERENCE**

[1] Ref[1]- Upasana et. al. “A Review of Text Classification Approaches for E-mail Management” IACSIT