Email Categorization using Hybrid Supervised and Unsupervised Approach

Jainesh Patel
Student, Department of Computer Engineering,
SardarVallabhbhai Patel Institute of Technology, Vasad.

Ms. Neha Ripal Soni
Asst. Prof, Department of Computer Engineering,
SardarVallabhbhai Patel Institute of Technology, Vasad.

Abstract: **As with the use of internet, use of emails increases drastically for electronic communication. This leads the mail boxes gets congested and emerged the problem of email overload, which is solved with the help of email categorization or email management. Email Categorization is multifaceted problem with many difficulties. Many schemes have been proposed for solving this problem in either supervised or unsupervised approach. With that approach once categorization model is built, it is hard to make any changes to them for handling of dynamic situations. As email replicates current information around the globe, the email content will be changed with the passage of time. Concept drift is the situation which occurs due to changes in underlying data distribution over a time period. The problem of concept drift detection and handling will occur due to dynamic nature of email. This paper proposes the dynamic hybrid scheme, combines supervised and unsupervised approach for detection and handling of concept drift. Initial classifier is built with the help of classification algorithm, and then clustering algorithm is applied in 'General' category of classifier to detect concept drift. . If it is detected then new cluster is formed for that new emerging concept and appropriate label is assigned to that cluster.**

Keywords:Email Categorization, Email Management, Email Overload, Supervised Approach, Unsupervised Approach, Hybrid Approach

I. INTRODUCTION

The immense increase in the use of internet helps globalization and rapid communication around the globe. Email or electronic mail is the mechanism to send messages or data electronically for communication provided by email service providers like Google, Yahoo, Microsoft, etc. It is simple to use and provides high reliability and high speed for data transfer typically in a few seconds. AOL research source investigated email as the most frequent used communication tool [3].

As more and more people are connected with internet, they get emails from various sources like personal messages, social activities, business activities, promotional or advertising activities, group activities, and many more. According to an estimate one may get tens to hundreds emails per day and every day and approximately billions of emails are sent across the world [1].As email services grow, increasing volumes of emails can flood users' mail boxes and leads to congestion problem. Users will not be able to view content of incoming emails and may find it difficult to get important emails in inboxes [3]. The survey also shows that almost 80% of internet users use email as a means for communication and that is why it is widely considered as most frequent communication tool in the world. Hence, a more effective and powerful mechanism for managing information in email is required.

Inboxes can be managed in different ways according to user preferences or their routine lifestyle. Generally emails are organized into different categories or classes according to some criteria or condition, this type of management of emails is called email categorization. One can do email categorization either manually by checking each and every email and then decide what to do with that email or where to put it, or by automatic system, which automatically checks emails and then puts them in appropriate place. Manual handling of this task becomes time consuming and tedious with the increase of emails in inboxes.

Nowadays emails are not only categorized as per spam and non-spam, but they are also categorized as per social, personal, educational, business, promotional, etc. For this kind of task email subject and body part is generally used, other parts are also used as per the nature of categorization required. Currently Gmail provides the facility of categorizing emails into different categories or tabs [2].

Generally machine learning algorithms are used for email categorization. Classification and Clustering algorithms are used differently for this task. If categories or classes are already defined or fixed or criteria are known in advanced then, classification algorithms are used. In order to uncover hidden similarities in email messages, which may not be described in advanced, the clustering algorithms are used.

In this paper we have proposed hybrid system,which uses techniques, classification as well as clustering for email categorization task to handle the problem of concept drift which occurs due to dynamic nature of email.

II. RELATED WORK

There are two main approaches for email categorization: Supervised and Unsupervised. Supervised approach uses different classification methods for email categorization, whileUnsupervised approach uses different clustering methods for email categorization.

Recent studies for email categorization using supervised approach are as follows: Gaunting surveyed many classification algorithms used for email categorization. They surveyed three classifiers: TF-IDF, Naïve Bayes and Support Vector Machine. TF-IDF classifiers are one of the popular classifiers used in email categorization. The major idea is to calculate the distances of incoming email with all existing categories, and then assign the email to the category with shortest distance.Naïve Bayes classifiers are simple to use and is basic algorithm for categorization. Here categories are generated from training data with the help of probability distribution. The incoming email is assigned to the category that has the highest probability to assign it. Support Vector Machine uses 'one-v/s-rest' methodology to classify emails into more than two categories. Here, the major idea is for n different categories where n>2, SVM classifier is applied n times. Each time it decides whether the incoming email belongs to certain category or not, and the probability of belonging is also calculated. Each email may be assigned to top k categories [4].

Matthew uses K-Nearest Neighbour algorithm for email categorization. Here each incoming email is compared with k classified neighbour emails and is assigned to the category according to the maximum voting of neighbour emails. They also compared the performance of k-NN with resemblance, k-NN with TF-IDF and Naïve Bayes [5].

Recent studies related to email categorization using unsupervised approach are as follows: K-Means, Fuzzy C-Means, EECM are different clustering algorithms used for this task. Gunjanuses k-Means and its variant for email clustering or categorization. In this approach incoming emails are partitioned into different k clusters so that intracluster similarity is high but intercluster similarity is law. They also compared the performance of K-Means, K-Means++, Kernel-selected and their own approach and results are shown in [6].

Fuzzy C-Means and EECM techniques for email categorization.Taiwo compared the performance of K-Means and FCM approaches and concluded that FCM gives better performance though it is complex and requires more memory. EECM is developed with fuzzy inference system and separates the email input sample space based on similarity of email contents to create fuzzy rules. EECM is distance based clustering algorithm, where distance is calculated between incoming email and group centre and is assigned to group with closest distance.Taiwo performed EECM clustering for email grouping and has been proven to be better in good performance than K-Means and FCM [3].

III. PROBLEM STATEMENT

The schemes proposed so far for email categorization uses either supervised or unsupervised approach for categorization. The problem occurred with these approaches is that, once model is built for either approach it cannot modify or it is very hard to update for future requirements. As the email resemblesthe current activity by user or group of users, it provides abstract behavior about the group or user on particular concept. Hence, email contains latest information or data and model built with the help of older concepts or data may not be able to correctly categorize incoming emails, which contains sorts of latest information or new concepts.

The changes occurred in concepts or learning environment termed in machine learning is called concept drift. In machine learning concept drift refers to unforeseen changes in the distribution of underlying data that can also lead to changes in the target concept over time. Email is dynamic in nature. It arrives in stream over time.So with the passage of time, new concepts may emerge and our categorization model should correctly categorize new concepts. Older schemes may not be able to do so, there is need of dynamic scheme which can correctly categorizes new concepts as and when they emerge and can correctly categorize new incomingemails.

The issue of detecting and handling of concept drift, which occurs due to dynamic nature of email, is handled with the proposed scheme presented in this paper in next section.

IV. PROPOSED HYBRID SCHEME

The proposed Hybrid scheme, presented in this section is employed for email categorization, which is combination of classification and clustering algorithms.This scheme works in two phases. In the first phase of scheme, the initial model or classifier is built with the help of classification algorithm. K-Nearest Neighbour classification algorithm is used to construct or build initial classifier. In that model the category 'General' or 'other' contains emails not fall under any other category. All other categories have some predefined concepts associated and may classify incoming email if it contains high similarity to that concepts else it may be categorized in 'General' or 'other' category.This phase does initial classification of emails into different classes or categories.

In the second phase the clustering algorithm DBSCAN is applied to the 'General' or 'other' category of the classifier built by first phase and it checks for any concept drift scenario simulated in various parameters of clustering algorithm. If such scenario is present then algorithm forms one or more clusters of similar emails according to the concept found in it and appropriate cluster label is assigned to cluster as per concepts found in group of emails.After detection of one or more clusters the initial classifier model built in first phase is updated and newly formed clusters are assigned as new categories in the classier. With this updated model incoming emails can correctly categorizes into appropriate categories within that time frame or period. After some time period the second phase is again performed and rest process is continued. This type of arrangement keeps the model dynamic and updated and can try to correctly classify new incoming emails containing possibly different concepts.

Figure 1 shows proposed system model for email categorization. The email dataset is split into two parts: training and testing, as usually done in data mining applications. The pre-processing techniques and optional feature selection technique is applied on training email dataset. Then email is represented in vector form using Term Frequency- Inverse Document Frequency. The phase 1 classification algorithm is applied on processed email vectors and classifier is built as shown in figure 1. Then test emails are applied on classifier and that emails are classified in one of the category shown in figure 1. Phase 2 clustering algorithm is applied on emails in 'General' or 'other' category of classifier and clusters are formed as shown in figure 1. The Algorithm steps of the proposed hybrid framework are presented below.

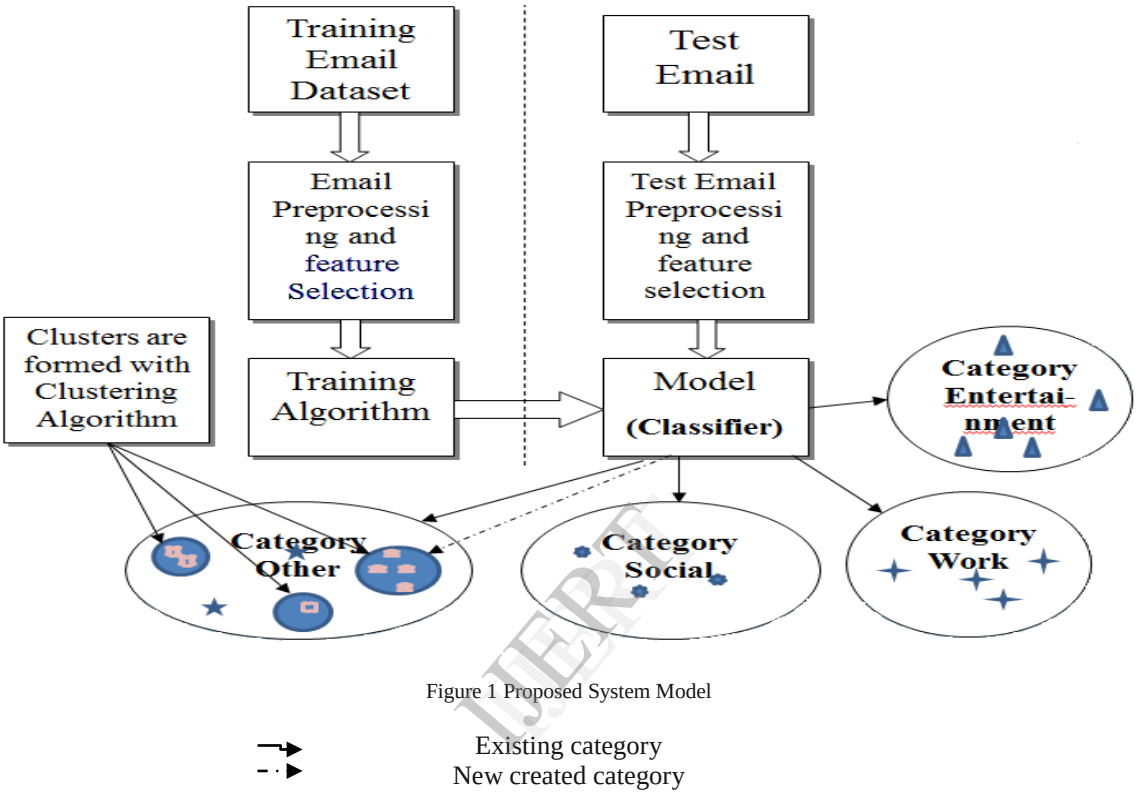

Figure 1 Proposed System Model

A. PROPOSED ALGORITHM

Input: Email Dataset

Output: Emails organized into different categories.

1) Apply pre-processing steps on dataset.
   a. Tokenization
   b. Stop Words Removal
   c. Stemming
2) Represent the features using TF-IDF.
3) Use K-NN classification algorithm to build the classifier model.
4) Classify incoming email as per model build in step 3.
5) Apply DBSCAN clustering algorithm to form cluster of emails in 'General' category.
6) If number of emails within the cluster exceeds the threshold value T, then concept drift is detected and new category for that cluster is created and classifier model is updated.

Assumptions

a. Feature selection technique can also be applied after pre-processing to improve system performance.
b. In general emails are categorized into different categories and emails that are not categorized fall into 'General' category.
c. Threshold value T is calculated by minimum number of emails that can form new cluster.

CONCLUSION

Email categorization is a rich problem with many difficulties. Many schemes have been proposed to solve this multifaceted problem. Many supervised and unsupervised approaches are used for this task. This paper combines both approaches and builds a hybrid approach which detects and handles concept drift situation which is so common in email domain due to dynamic nature of email. This paper proposed a scheme which detects concept drift from 'General' or 'Other' category and forms a one or

more clusters according to the concepts found, then appropriate label is assigned to that clusters.Thus this paper solves the problem of static model of email categorization and builds a dynamic model for email categorization as well as detects and handles concept drift situation dynamically.

REFERENCES

[1] Radicati, S., Hoang, Q.: "Email statistics report", 2012-2016. The Radicati Group, Inc., London , 2012.
[2] A new inbox that puts you back in control, http://www.gmailblog.blogspot.in/2013/05/a-new-inbox-that-puts-you-back-in.htm
[3] TaiwoAyodele, Shikun Zhou, RinatKhusainov, "Evolving Email Clustering Method for Email Grouping: A Machine Learning Approach", IEEE, 2009.
[4] Guanting Tang, Jian Pei, Wo-Shun Luk, "Email mining : tasks, common techniques, and tools", Springer , 2013.
[5] Matthew Chang,Chung Keung Poon, "Using phrases as features in email classification", The Journal of Systems and Software, Elsevier , 2009.
[6] GunjanSoni, C.I. Ezeife., "An Automatic Email Management Approach Using Data Mining Techniques", Springer-Verlag Berlin Heidelberg, 2013. pp. 260-267.
[7] Jiawei Han, MichelineKamber and Jian Pei, Data Mining Concepts and Techniques 2nd Edition, Morgon Kaufmann, 2006.
[8] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/
[9] Enron Email Dataset,http://www.cs.cmu.edu/~./enron/