IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-4 E-ISSN: 2347-2693

User-Defined Classification for Email System using Back Propagation Algorithm

Arichi Arzare[1*], Suneha Chaudhari[2], Sayalee Desai[3] and Sonali Jadhav [4]

[1*,2,3,4]*Department of Computer Science, Rajiv Gandhi Institute of Technology, Mumbai, India*

www.ijcseonline.org

Received: Mar/30/2015 Revised: Apr/12//2015 Accepted: Apr/18/2015 Published: Apr/30/ 2015

***Abstract*—** These days email system is one of the major sources of communication and users' depend heavily on it. Even after the evolution of new mobile applications, social networks etc. emails are extensively used on both personal and professional platforms. Pertaining to this extensive use, inboxes these days usually become a chunk with unnecessary messages from social media, advertisements, subscriptions etc. which might not be of that much importance. Thus there's a need of classification so that the user does not have to surf through the chunk for one particularly important mail. In this paper, we propose a solution for email classification using back propagation technique which has user defined categories where word search is made on the content of the email. The output of this solution will give the user a selected number of emails according to the category he/she chooses.

***Keywords*—** *Email, Classification, User-Defined, Back Propagation, Categories*

1. INTRODUCTION

Email has been the most common and reliable means of communication. Thus managing the emails is vital as it is prone to misuse. The major problem arises in the mailbox due to disordered, congested and unstructured emails .It is very difficult to find emails with specific contents if it is not organized. Schuff et al [1] stated that "Emails are widely used to synchronize real-time communication, which is inconsistent with its primary goals". Large number of people are dependent on emails for communication due to its effectiveness, hence the probability of the inboxes getting congested is high. Messages range from small informal messages to formal notifications. Users may find it clumsy to find a previously archived message. Kushmerick [2] stated that "the ubiquity of email and its convenience as knowledge management tools make it unlikely that users' behavior will change as falling bandwidth and disk storage prices further reduce the incentive to steer away from using email as a document storage system". At this point of time it is necessary to classify emails on basis of important words, derive classes on basis of content.

2. RELATED WORK

There are a few solutions into the problems of segregating emails into folders but very little advancement in classification of emails based on the users' needs. The existing method used for email classifications is to archived

messages into folders. This is a manual classification solution, but this is not enough as folder names are not necessarily represent the actual content. Yukun et al [4] proposed a new email classification model using a linear neural network trained by Perception Learning algorithm (PLA) and a nonlinear neural network trained by Back Propagation Neural Network (BPNN) and Semantic Feature Space (SFS) method was also introduced in this classification model.

Schuff et al [1] can provides a simple way to semi automate email classification and such system require the users to define a set of instructions for the email application to segregate incoming messages into folders and order them by their priority. The disadvantages of rule-based system are that they are highly difficult for non-technical users because writing the rules require some level of programming knowledge. Bifrost [5] an email classifier avoids this difficulty by letting user define all sorting rules with a simple graphical interface. Terry et al [6] also proposed an approach by automatically checking incoming messages and making required suggestions before emails reach the user's inbox, so the priority system classifies each messages as of either high or low priority based on the users utility. Yukun et al [7] designed a system "that automatically filter spam emails by using the principal component analysis (PCA) and the Self Organized Feature Map (SOFM). In their schema, each email is represented by a series of textual and non-textual features. To reduce the number of textual features, PCA is used to select the most relevant features. Finally the output of the PCA and the non-textual features should be inputted into a well-trained SOFM to classify (spam or normal)"

3. EMAIL CLASSIFICATION

Classification of text in an email is an example of supervised learning that looks to build a model of a function that maps emails to classes. A learning algorithm is presented with a set of already classified, or labeled examples. This set is called the training set. A number of

classified emails from the training set are eradicated before the model building to be used for testing the model's performance. This set is known as the testing set. To measure the classification accuracy of our model, several models are constructed from different sections of the examples to training and testing sets. The error is then averaged over each model. This process is called n-times cross validation where "n" is the number of times the example set is sectioned. We produce 100 models for evaluation using this process and we obtained 100-times cross validation. Now our model has been built, it was used to predict the classification of future email messages. The correctness of our models are largely rely on: performance of our back propagation algorithm, vital word selection using information retrieval. The more representative, the training data, the higher the performance.

.

4. APPLICATION OF MESSAGE CLASSIFICATION

Email classification has many direct applications such as: classification of emails according to the content, important words, phrases, to identify particular emails of interest. We have implemented supervised learning for the word extractions and also to support bigger objectives like Information extraction: procedure to extract bits of specific information from unstructured emails Classification methods: words, phrases. For example, extracting the date, time, location, important words- meeting, interview, surgery, appointment, bookings, credit card deadline etc.

.

Neural Network Approach

We use supervised learning to implement the email classification in our system. Our sample categories are: education, travel, friends, work, and gym. Our solution is based on the heuristic technique that an email can be classified as:

- Education related if there are keywords like classes, notes, lectures, teachers, project, deadline etc.
- Travel related if there are keywords like trip, flight, hotel, camping, luggage etc.
- Friends related if there are words like fun, love, picnic, movie, games group etc.
- Gym related if there are keywords like sets, abs, workout, cardio, strength, membership etc.
- Work related if there are keywords like meeting, presentation, formal, invitation etc.

We provide observations that this easy algorithm can classify a huge chunk of emails into smaller groups according to the user's convenience. With our proposed solution, we explain how word data sets can be used to retrieve emails using the content of the email.

4.1. Our Method

We use user defined "word classes" for implementing the automated classification system using neural network that can contain words and email addresses. This solution is performed on the subject, the email address and the content of the email. We searched for important words according to our word classes in email corpus [9] from Enron. The next problem is to search this body for the related words and return email data sets according to the user's query.

For example, if our system searches for the word 'illness' in the email set then the corresponding emails will be displayed under the category of doctor or hospital or whatever tag the use has defined. Also, cases of ambiguity of messages is removed since 3 different fields of the email is searched. For example, if a user makes the category 'dance class' and adds words 'class', 'routine', 'song', 'performance', 'dance' to it. The word search for 'class' might conflict with the 'education' category, and hence that email could be shown in both 'dance' and 'education' category. But this will not be the case as the other words in the word classes will put that email into the 'dance' category. So all the important words are searched and emails belonging to that class with favorable membership are displayed. Advantage of this system is that the user has to only define the word classes once, there after he can directly go into that category and see the related emails.

4.2. The learning process

We implement an associative learning technique in which matching input and output patterns train the network. These input-output pairs can be provided by the user i.e. each user can have his/her own categories according to their convenience or can be also be defined by the system.

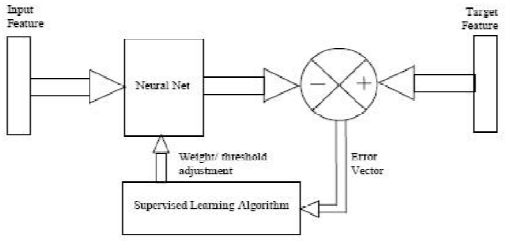

Figure 1. A sample Supervised Learning Process

To implement this search in the neural network using back prorogation technique. According to Habra [11], back propagation is a network trained to identify various patterns including images, signals and text. The data sets are used to run by a multilayer neural network once they are given by

the user. In our system, the word classes are the inputs to the NN and the classified emails are the output.

Each word in the word class acts as input node in the neural network. So these represent the neurons in the first layer. As these are checked in the email fields, each word is searched for in the subject, address, and the body of the email and all the corresponding mails with more favorable membership will go to the category containing that word class. Figure 2 shows how the inbox content is classified.

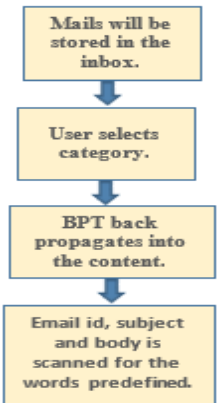

Figure 2. Work flow of the Classification.

5. OBSERVATIONS

The code was executed and debugged to obtain the outcome for the system. Various users were created and emails were transferred amongst them. We then created categories for a particular user and added words pertaining to that category.

5.1. Execution

The following screenshot shows the registration page on which the new user registers for a new account.

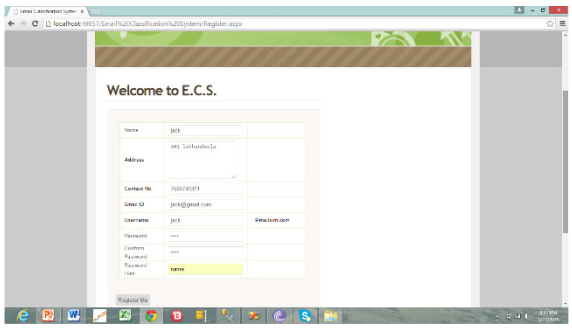

Figure 3. Registration page

User fills in details such as name, address, contact number, alternate email id, username, password and password hint. In case the user enters a username which has already been taken, a username invalid indicator appears. Similar warnings appear in case of invalid phone number. A new account is created once the user clicks on Register.

The user has to login by entering his email address in xxxx@mailsim.com format followed by password.

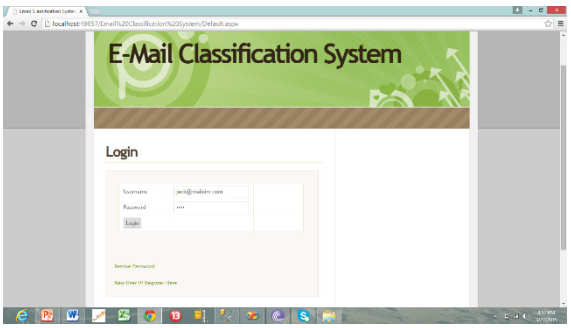

Figure 4. Login page

Once the user logs in he is all set to access and utilize all the functionalities of the website.

Figure 5. Welcome page

The homepage includes the various tabs such as compose, inbox, draft, sent mail, trash, address book, change password and logout.
User can compose email through the compose option.

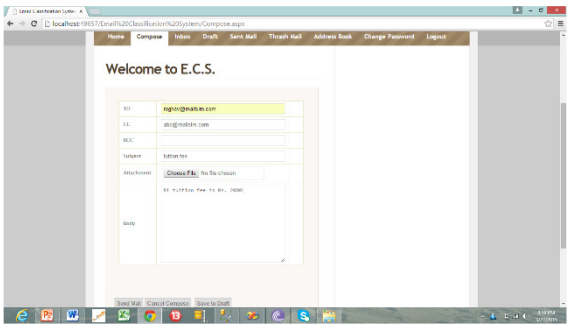

Figure 6. Compose mail

User has to enter the email address of the person he has to send the mail to. She/he can enter CC/BCC as well. Followed by subject. The main content of the email is written in the body. User can send the mail once done.
If the user wishes to save the message as a draft he can do so. The emails can be viewed later through the draft tab and edited or sent.

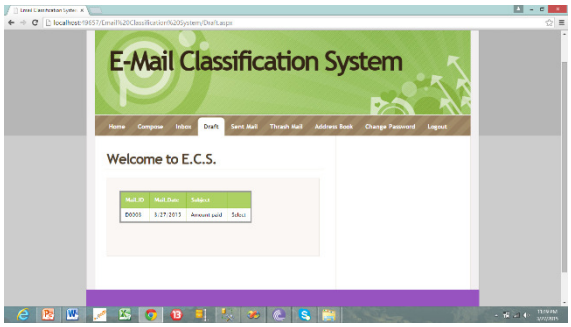

Figure 7. Draft mail

Outgoing mail can be tracked through the sent mail option.

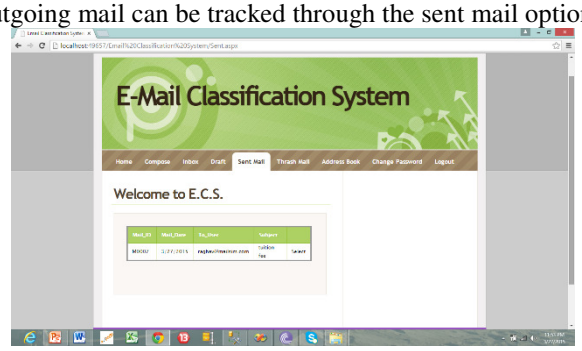

Figure 8. Sent mail

All the incoming emails are received in the inbox. The user can check them by clicking on the inbox tab. All the mails are arranged according to their dates. The recent ones being on the top. If the user wishes to view emails only from a specific group or type there needs to be a provision for the same. That's where our classification system comes into the picture. Next to the messages in the inbox there is an option which says "My Classification". The user has to click on that link to view all his tabs.

Figure 9. Inbox

Once the user clicks on the My Classification icon, a list of all the predefined as well as user defined tabs will be displayed.

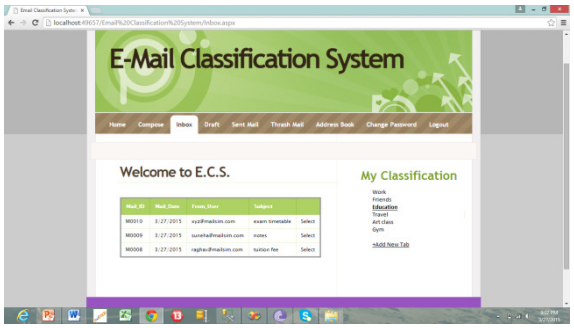

Figure 10. Selection of Category

Once user selects the tab, the back propagation theorem sets into action and classifies the email. The required mail is scanned and picked from the inbox and displayed under the selected tab. All the emails belonging to the selected category will be displayed. In the above picture, you can see that all the mail belonging to the Education tab has been filtered from the inbox and displayed. The scanning and classification takes place on basis of the predefined classes. Whichever tab the user selects, the back propagation theorem sets in and classifies the email for the user.
There are certain basic predefined tabs while the rest can be created by the user as per his convenience. The Add new tab option can be used to create new tabs.
Efforts are made to make sure the classification is accurate and all the concerned email are displayed.

6. Conclusion

In this paper, we have explained and demonstrated how to generate user defined email categories. We analyze the characters and words of emails and classify them successfully. We build a disciplined structure in which our classification is based on heuristic technique with the use of Back propagation theorem to determine what words in a corpus of email messages might be more favorable to use in a query. We also implement a neural network based system for automated email classification into user defined "word classes" and our BPT implemented was able to learn technique in an associative learning approach, in which the network is trained by providing it with input and matching output patterns.

We have shown that neural networks using back propagation technique can be successfully implemented for semi-automated email classification into meaningful words. The back propagation is based on learning by example and outperforms several other algorithms in terms of classification performance.

This classification system eases a user's usage of an email system. Though email categories have existed in previous email systems, those were predefined tabs. In our system the user can easily create his own tab and by providing the key words, the tab is generated permanently. It is a onetime process and once tab is generated the user doesn't have to look back. The email will be classified every time the tab is selected by backtracking and comparing input and output patterns. The system has proved to be an excellent advancement in the field of email classification and has received a positive feedback.

7. REFERENCES

[1] Schuff, D., O. Turetke, D. Croson, F 2007, 'Managing Email Overload: Solutions and Future Challenges', *IEEE Computer Society, vol. 40, No. 2, pp*. 31-36.

[2] Kushmerick, N., Lau, T. 2005, '*Automated Email Activity Management: An Unsupervised learning Approach', Proceedings of 10th International Conference on Intelligent User Interfaces*, ACM Press, pp. 67-74.

[3] Helfman, J., Isbell, C. 1995, 'Ishmail: Immediate Identification of Important Information', AT&T Labs.

[4] Boone, G. 1998, 'Concept Features in Re: Agent, An Intelligent Email Agent', *Proceedings of 2nd International Conference on autonomous agents,* ACM Press, pp.141-148.

[5] Balter, O., Sidner, C. *Bifrost Inbox Organizer: Giving Users Control over the Inbox*. In *In Proceedings of the Second Nordic Conference on Human-Computer interaction*. 2002. Aarhus, Denmark: ACM Press.

[6] Ramos, J. (2002). *Using TF-IDF to Determine Word Relevance in Document Queries,* Department of Computer Science, Rutgers University, Piscataway, NJ, 08855.

[7] Yukun, C., Xiaofeng, L., Yunfeng, L. (2007). *An Email Filtering Approach Using Neural Network*, Springer Berlin, pp. 688-694.

[8] United States. The Board of Trustees of the University of Illinois. (2003). *D2K™ Data to Knowledge™ Text Mining: Email Classification.*

