EMAIL CLASSIFICATION with SUPERVISED LEARNING

Priyanka Megharaj, Supriya Thakur, Vaidehi Mungekar, Heena Vasani, Reena Somani
Department of Information Technology
Atharva College of Engineering, University of Mumbai, India
{megharajpriyanka, supriya.thakur.40, mungekarvaidehi, heenavasani5, reena.jagetia }@gmail.com

***Abstract* - Email classification is the technique to group and manage email messages. To gain knowledge by learning through training data is one of the rare field of email classification. Email classification system is considered as an extremely intelligent system because it can perform tasks like human brain. This paper proposes email classification system based on neural network using Back Propagation Technique (BPT).During past few years email users have increased so volume of email messages has also increased and hence many problems are getting generated like unstructured mail boxes, email overload, unprioritized email messages etc. These problems can be overcome by Email Classification.**

Keywords - ***Email classification, back propagation technique, neural network, email, automatic learning and Email management.***

I. INTRODUCTION

Email classification is defined as classifying emails based on the contents of the email including - email id, subject and email message into default and user defined categories. The most modern email classification system organizes emails into folder based on the keywords detected in it. Email classification system may face challenges because of huge and various data sets and large number of emails. The quality of training data set decides performance of email classification algorithm. An ideal data set must consist of important terms related to corresponding category. Previous research has shown that accurate result can be achieved with back propagation technique (BPT) in neural network (NN). This paper uses Back Propagation Technique in Neural Network for automatic email classification [1].

Classifying or tagging emails can be done by several methods. The process of applying these tags (i.e. classifying) may be either –

1) Manual: The message custodian, a human, applies the classification when the message is either created or received.

2) Automatic: A computer applies the classification based on rules or contextual analysis.

3) Hybrid: Combining both manual and automatic methods.

II. EMAIL CLASSIFICATION CHALLENGES

The challenges encountered in email classification are [1] -

- The number of incoming mails for different users varies and increases day by day. This results in new messages to be added and old messages to be deleted. A classification technique for varying email characteristics must be adaptable.
- Classifications of emails done manually based on personal preferences are not simple and hence this distinction has to be taken into consideration.
- The fields like subject field, Sender field, CC field, BCC field play a significant role in classification as compared to document classifications which are rich in content resulting in easy classification.
- The emails can be categorized into folders and can also be reclassified into subfolders which should be considered while classification.
  Semantics is the study of meaning and so classification can be done purely on the basis of email's body.
  Example - Design within Project folder, Coding within Project folder and many more.

III. RELATED WORK

There are many classification algorithms such as Neural Network (NN), Support Vector Machine (SVM) and Naïve Bayesian (NB) are used for effective email classification [1]. The main limitation of Back Propagation Technique is it requires more time in parameter selection and training process. Yukun et al proposed a new email classification system using a nonlinear neural network trained by Back Propagation Technique and a linear neural network trained by Perception Learning algorithm (PLA). Rule based system is semi automated system. In semi automated system users are required to specify instructions for mail application to

classify emails into different folders [7] .Limitation of Rule based system is that it is not useful for non-technical users because programming knowledge is required for writing rules. Terry proposed new system in which email is classified on the basis of its priority i.e. high or low importance. There are three approaches for email classification -

A. *Rule based Classification*

In Rule based classification emails are classified into folders. *RIPPER* learning algorithm induces 'keyword spotting rules' for email classification by William Cohen. *i-ems is* another a rule based classification system that based only on sender information and keywords.

B. *Information Retrieval Based Classification*

For every incoming mail it predicts three likely destination folders. TF-IDF calculates value for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. The TF-IDF classifier performs well not only in the absence of a large training set but also when the amount of training data increases, adding to the heterogeneity of a folder.

C. *Machine Learning Based Classification*

In this type of classification, Naive Bayes approach is used. It is used for effective training, accuracy and performing iterative learning.

IV. EMAIL CLASSIFICATION

The best known example of supervised learning is classification of text in an email. In supervised learning entire email is considered as data set and learning algorithm represented as already classified examples. This set is called as Training set. A set of emails form training set are used for testing the models performance. This set is termed as testing set.

The accuracy of our model is mainly dependent on -

- The performance of back propagation algorithm.
- The important word selection using information retrieval.
- The representativeness of the training data with respect to newly acquired email data to be classified.
- The more representative, the training data, the [11] better the performance [9].

A larger number of training examples is often better, because a larger sample is likely to be [9] more reflective of the actual distribution of the data as a whole.

Each email is considered as bag of word. Each unique word is considered as Attribute. Frequency of that attribute is considered as attribute value. Hence Email messages are represented as vector of numeric attributes. This set is termed as vector space.

A. *Algorithm*

1. For every message M
2. Let **FW** = N most frequent words in the message
3. Iterate over all activities and for each
4. activity **AC**
5. Let **AFW** = common words in
6. activity AC
7. If (**FW** = **AFW** then
8. mark the activity AC
9. update the message activity as **AC**
10. create a rule that states // machine learning
11. for each message received that has some
12. words like **FW**
13. **AC** is the activity for this message
14. Else create a new activity
15. End

Figure 1 - Classification Algorithm

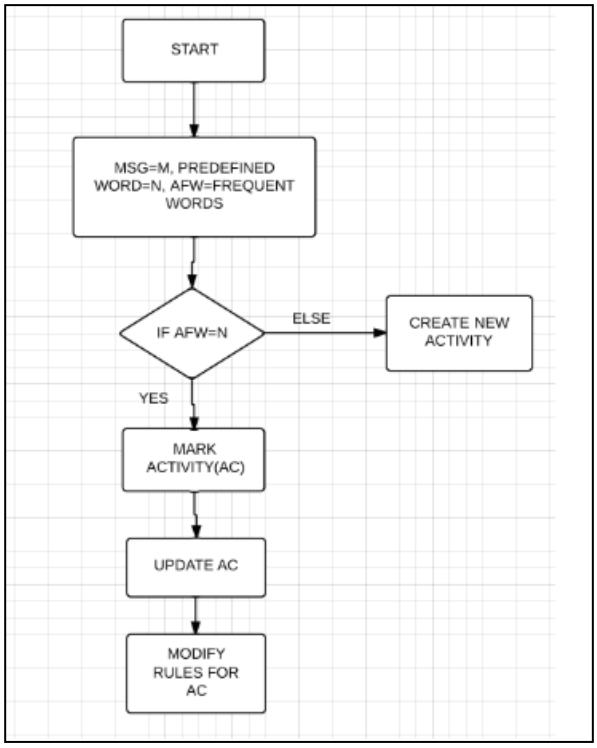

Figure 2- Flow chart of Classification Algorithm

B. Methodology

The system architecture of Email classification system consists of five stages i.e. Pre-processing, Extraction Filtering, Clustering and Summarization.

1) Pre-processing - The first basic step is to preprocess and extraction of the relevant contents from different mail recipients. This step includes applying Stemming algorithm which does the following tasks:-

- Elimination of special characters.
- Converting all uppercase letters to lowercase [5].
- Eliminating all non-letter characters [10].
- Removal of stop words.

2) Extraction - *E*-mail messages have subject, from/to mail identifiers and content irrespective of mail servers, by default. Once the mail messages are classified into different categories, extraction identifies the components of the mail messages. The extraction task is done based on three different levels namely based on email's subject (subject level), mail recipient information (mail recipient level) and based on the contents (content level). Each approach has its own significance.

3) Filtering - Filtering is done based on the list of filter words analyzed or mined by manual examination from the training corpus. These filter words are stored in filter databases for providing assistance in the process of filtering the terms or contents or mails [3]. We categorize some filter words as context sensitive words. For example if subject or content contains the word 'urgent', 'immediate', users tend to open such mails at first.

4) Clustering - The content of each email was analyzed, with each word is represented as a vector model. These words are represented with a vector whose each element corresponds to a particular word and indicates whether that word occurs or not in the text or the number of times it occurs.

5) Summarization - After the contents have been clustered, they are summarized. Summarization by extraction involves scoring the sentences and picking up the most important sentences. The following steps are adopted to summarize the contents -

- Calculation of frequency weights of each token.
- Scoring the terms in the document.
- Ranking of documents based on weights.
- Reproducing the results based on the user requirements.

C. Applications

The Email classification system has many applications such as -

- Emails are classified on the basis of contents such as - Critical, Urgent, very important, Important and Not important categories [1].
- Classification of streams of emails, with various important words, phrases to identify particular items of interest.
- Classifications of emails can also help the user to know about important messages &/or meetings. For example. If the user has a mail with information saying "FLIGHT AT 11PM" then the system can remind the user about the flight every 2 hours without the user actually setting a reminder.

We support higher level objectives and word extractions like [7]:

- Classification methods: These methods are used to identify those parts of the messages which qualify for extraction.
  For Example: extracting date, time, location or other important words like meeting, flight times, debit-credit card deadline, job interviews, surgeries etc.
- Information extraction: this process of extraction is done by extracting bits of information from email messages.

V. ARTIFICIAL NEURAL NETWORK

Artificial Neural Network (ANN) is information possessing system inspired from biological nervous system such as brain. Neural networks have been successfully applied to a variety of real world classification tasks in industry, business and science. Classification is one of the most active research and application areas of neural networks.

In Classification and clustering Neural Network is used as important data mining tool. NN usually learns by examples. NN is supplied with enough examples, it should be able to perform classification and even discover patterns in data.

Basic NN is composed of three layers, input, output and hidden layer. Each layer can have number of nodes and nodes from input layer are connected to the nodes from hidden layer. Nodes from hidden layer are connected to the nodes from output layer. Those connections represent weights between nodes.

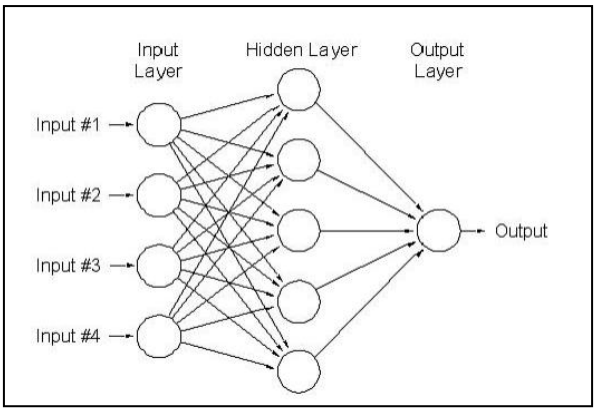

Fig 3. Neural Network

We embedded technique Term Frequency Inverse Document Frequency (TF-IDF) which determines what word in corpus is more suitable in query. This technique is invented by Ramos. TF-IDF calculates values for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in the relationship with the document given by high value of TF-IDF of that corresponding word. With our proposed solution, we define informally, that query of word retrieval can be described as the task of searching a collection of email data set for specific instances of the content.

The methodology that we proposed for our email classification is a supervised learning [1]. To categorize our mails into different categories, we use heuristic techniques and the facts that if -

1) Urgent: If the mail has contents like meeting deadlines, vital appointments, doctor's appointment then the mail is categorized in URGENT category.

2) Critical: If the mail has contents like accidents or loss of life's then the mail is categorized into CRITICAL category.

3) Very Important : If the mail has contents like meeting with CEO, important conference meetings, decisions to be taken quickly; then the mail is categorized into VERY IMPORTANT category.

4) Others: If the mail content does not have any important message or any deadline message or any loss of life; then the mail is categorized in OTHERS category.

Neural networks using back propagation technique can be successfully used 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. The effects of various features selection, hidden weight calculations techniques are observed.

The rate of success is the average of correct classification over all data. All emails from a given user are separated in two equal sets training and test. The training data is used to train the neural network (NN) and our NN knows the patterns that make emails to be categorized into default categories.

A multilayer neural network runs by the data provided by user for classifying emails to particular category. The inputs of the NN are the word importance in email messages and the output is the importance.

Each word in an email represents an input node in the neural network. Therefore the number of neurons in the first layer equals to the number of words in the input vector.
It is required to have input output data and the data is found from the email messages (words in the email content).For the output layer, nodes are filled with binary. Based on the category of email, output node gets the value.

VII. FUTURE SCOPE

- The email classification scheme can be implemented for all those incoming emails which enter in a particular pattern. These patterns can be in terms of timing or completion of incomplete data after particular time period.
- More priority based classification must be there in terms of time period namely critical, urgent, very important and others.
- Time required to classify the incoming emails, to train the classifier can be reduced.

VIII. CONCLUSION

Email Classification System is the design and implementation of a system to group and summarize email messages. The system uses the subject and content of email messages to classify emails based on user's activities and generate summaries of each incoming message with unsupervised learning approach.

Email Classification System provided a useful to generate accurate email categories. The characters of emails are analyzed and the email conversation structure is studied. Email classification is based on heuristic technique with the used of Term Frequency Inverse Document Frequency (TF-IDF) to determine what words in a corpus of email messages might be more favorable to use in a query, a neural network based system is implemented for automated email classification into user defined 'word classes' and Back Propagation Technique is able to learn technique in an associative learning approach, in which the network is trained by providing it with input and matching output patterns.

ACKNOWLEDGMENT

This paper describes the research done at Atharva College of Engineering in department of Information Technology. We would thank our project guide Prof. Reena Somani for guiding us. We are also eager & glad to express our gratitude towards the Head of our Dept. Prof. Jyoti Chinchole & all the project coordinators. We would also like to deeply acknowledge our respected Principal Prof. Dr. Shrikant Kallurkar & the Management of Atharva College of Engineering.

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] 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.

[4] Yun, F.Y., Cheng, H.L., Wei, S. (2008). Email Classification Using Semantic Feature Space, Proceedings of the 2008 International Conference on Advanced Language Processing and Web Information Technology, IEEE Computer Society Washington, DC, USA, pp.32-37.

[5] United States Patent Application Publication (10) Pub. No.: US 2012/0271626 A1 Baba et al. (43) Pub. Date: Oct. 25, 2012.

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

[7] Youn, S. a. (2006). A Comparative Study for Email Classification. JOURNAL OF SOFTWARE , 2 (3), 1-13.

[8] Aery, M. a. (2005). eMailSift: Email Classification Based on Structure and Content. In Proceedings of the Fifth IEEE international Conference on Data Mining (pp. 18-25). Washington, DC: IEEE Computer Society.

[9] 'Repeated Labeling Using Multiple Noisy Labelers' Panagiotis G. Ipeirotis Foster Provost Victor S. Sheng Jing Wang September 9, 2010.

[10] Authorship Attribution Patrick Juola Department of Mathematics and Computer Science, Duquesne University, 600 Forbes Avenue, Pittsburgh, PA 15282, USA, juola@mathcs.duq.edu

[11] Email Classification Using Back Propagation Technique Taiwo Ayodele, Shikun Zhou, Rinat Khusainov Department of Electronics and Computer Engineering University of Portsmouth, United Kingdom.

