Email Clustering System

Prof. S. S. Shinde, Professor, Department of Information Technology, Marathwada Mitra Mandal’s Institute of technology, Pune
H. Sardhara, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune
V. Suryawanshi, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune
A. Waje, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune
B. Raut, Student, Department of Computer Engineering, Marathwada Mitra Mandal’s Institute of technology, Pune

ABSTRACT

Emails have been considered as a useful resource for research in fields like link analysis, social network analysis and textual analysis. Email mining is a process of discovering useful patterns from emails. Clustering techniques can be applied over email data to create groups of similar emails. An email can be represented as an Object consisting of several attributes like sender email-id, receiver email-id Subject, message, sending-time, and attachments etc. clustering is used to discover email groups. Generally the most of the attributes in emails are text type, so text similarity techniques are used for measuring the similarity between pair of email objects. Clustering the emails and the user depend on the information they have exchanged and graphically representing Cluster. The Enron email dataset is a touchstone for such research.

Keywords
POP3, IMAP, SMTP, OTP, Cluster.

1. INTRODUCTION

Email communication has come up as the most effective and popular way of communication today. People are sending and receiving many messages per day, exchanging files and information. E-mail data that are now becoming the dominant form of inter and intra organizational written communication for many companies. Clustering is a technique of creating group of similar objects. The cluster shows the similar emails exchanged between the users and finding the text similarities to cluster the users, we are using the Pattern i.e., the similar words exchanged between the users by considering the different Threshold values, here Threshold value shows the frequency of the words used and we have Graphically represented the cluster in the form of Bar charts.

2. PRESENT THEORY AND PRACTICES

Clustering: Search results clustering problem is defined as an automatic online grouping of similar documents in a search result list returned from a search engine. In this paper we present Lingo a novel algorithm for clustering search results, which emphasizes cluster description quality[1][3]. We feel this problem has not been sufficiently solved in the previous research resulting in too long, ambiguous and very often even meaningless group labels. Lingo combines several existing methods to put special emphasis on meaningful cluster descriptions, [1] in addition to discovering similarities among documents.

This similarity is in form of phrase which will be counted in terms of frequency or we can call it threshold value of that phrase which can be use further in formation of different clusters [3]. The documents that contain this phrase are combined together to form the cluster.

3. PROBLEM STATEMENT

Most of the People use Email Services freely made available by various companies like Google, Yahoo, Rediff, Hotmail etc. They have more than one account and a heavy load of mails on each account which is tedious to sort through and have to browse through less important mails to get to the urgent ones resulting in waste of time. We aim to develop a software tool, which will run on a mail server, which synchronizes the mail from all other mail server accounts to get the urgent ones resulting in waste of time. We aim to develop a software tool, which will run on a mail server, which synchronizes the mail from all other mail server accounts to get to the urgent ones resulting in waste of time. We aim to develop a software tool, which will run on a mail server, which synchronizes the mail from all other mail server accounts to get to the urgent ones resulting in waste of time.

4. PROJECT SCOPE

All the Third-party-mail servers are connected to the main server. The main server synchronizes with the third-party servers and clusters the retrieved mails periodically. The user accesses the main server to view the clustered email and to carry out any advanced mail operations. User Classes and Characteristics.

- SERVER :- Maintains the database which contains user mails, user information, third-party server information and information about clustering. Fetch the mails from third-party servers.
- USER :- All users store their information by registering themselves at server side. Access their mail-accounts of other mail servers.
5. SYSTEM FEATURES

This include Functional requirement of system and major services provided by proposed system. Services includes reduce manual work, time saving, easy interface and cost efficiency. Customer’s database security is our main priority. For this we have used encryption and decryption for mail and also we can use OTP service for login security. Main priority is customer’s interest, we provide business logic for implementing different aspects for customer. The user will register with his access credentials i.e. username and password. He will be provided the userid which he has registered previously in order to login in our application. When the login process will done, all the mails from third party server will be fetched to database and by applying clustering, all mails will be categorized in different labels which will shown to the user in application.

6. PROPOSED SYSTEM

Fig. Creation of Cluster

To find the required mail from bunch of mails is very difficult job for user. To overcome this problem and reduce the overhead of user, we introduce the system that will categories all the mails according to their content and make easy for the user to go through the important mails. In our project, we are going to first fetch all the mails from user’s mail id by registering him to our application. The working of fetching mails will be done by POP3 or IMAP protocol [2][6]. Then all the mails will be encrypted and stored in database [4]. When the request for mails will be send by the application, all the mails will be applied to clustering algorithm and each mail will assign a label. According to the content of the mails, the labels will be assigned and the particular mail will be stored to its assigned label. All the mails will be shown to our application inside its assigned label.

7. SYSTEM ARCHITECTURE

Fig. Architecture of System

8. ADVANTAGES

1. Easy access to mails from all different mail accounts of the user.
2. Display the inbox as clustered messages.
3. One can see the graphical representation of all mails like a Graph or a Pai-chart.
4. One can send mails through this application.

9. LIMITATIONS

1. There is negotiable chance that a mail will be categorized in a different label.

10. EXPECTED RESULT

The expected result of this project is that the client will get an application that will run on client side and cluster all the mails of user and will put them in appropriate label. This will scale down the overhead for searching mails of user.

8. ACKNOWLEDGMENT

We would like to express our gratitude to all those who helped us to complete this work. We want to thank our guide Prof. S.S.Shinde and all others.

REFERENCES

[1] Stanislaw Osinski, Jerzy Stefanowski and Dawid Weiss, "Lingo: Search Results Clustering Algorithm Based on Singular Value
Decomposition”, Institute of Computing Science, Pozna University of Technology, Poland 2011.


