Clustering of Emails
Prof S Pratap Singh¹ Minal Mohite² Sangeeta Maharana³ Pornima Wadekar⁴
¹²³⁴Department of Computer Engineering
¹²³⁴SP’S IOKCOE, Pune

Abstract—The abstract Email Clustering is a technique of clustering group of emails together and to form a specific labels which will define the meaning of the mails in it. Email plays a vital role in our day to day life. Human beings can never be satisfied. The hunger never ends. Due to this, in every sector, the advancement in technology never stops. We use emails daily for our official work, personal work etc. to make email system easier, labels and clusters are generated.

Key words: Stemming, Email Clustering

I. INTRODUCTION
Email is considered to be the most common method for communication. Email is an easiest and best form of electronic messaging service. Due to this, the use of email is increasing extensively day by day. The main source for official messages nowadays is email service. As per survey, it is found that, a normal human being spends 90 min of his/her day on email. A person gets so many emails from different sites. But they are not in clustered form, the proposed system will help to cluster these data and will generate a meaningful label. Lingo algorithm is a very powerful clustering technique for generating clusters and labels.

A. Existing System:
In Existing System, the labels can be generated manually by forming folders as per the mails and place mails in the formed folder. There are many mail service providers on internet today like yahoo, MSN etc. Mails being one of the most popularly used service by all sector of life, corporate as well as personal to contact each other. And that too with no restriction on location and of course fee free of cost. Users have mail accounts on different mail servers. One cannot access email from other mail servers in existing mail accounts.

B. Disadvantages of Existing:
The disadvantages of current system are
1) Need to remember different User-Id and Passwords.
2) Waste of time creating new sessions of each service providers by logging into their respective domains.
3) More waste of Bandwidth and download capacity.
4) People cannot access mails from different mail server at the same time from a single server.
5) Labels can be generated manually only.

As we saw, that our existing system has so many drawbacks. We need to overcome these drawbacks. We can overcome these drawbacks by applying some modification.

II. RELATED WORK
As we nowadays are so much used to our emails, they have become an important part of our corporate and social life. The data are in extensive quantity and the labels or clusters which are been given by the existing system are not very useful to ease our work. To find a particular document, we have to unnecessarily sift through a random list of emails. In this paper the main focus is on Lingo clustering algorithm, which we believe is able to capture thematic threads in a search result, that is discover groups of related documents and describe the subject of these groups in a way meaningful to a human. Lingo combines several existing methods to put special emphasis on meaningful cluster descriptions, in addition to discovering similarities among documents.

III. PROPOSED SYSTEM
The proposed work of the system gives us an idea about how our system is actually going to work. The proposed system is Email Clustering System Using Lingo Algorithm which includes single Sign-In. In our proposed system third party server forms clusters according to the content of the emails that are available in Inbox. It is a desktop based application in which it will first fetch the emails from inbox and then forms cluster, to form the clusters we are using Lingo algorithm For Single Sign-On, the user will have to initially enter the username and passwords of all the email accounts he wants to access with a single login. The system will generate a unique user id and password for the user and after that the user can access all the other mail accounts with a single login, without switching from one email account to other.

A. Algorithm:
1st phase (Preprocessing)
1) Dc ← Set of input documents
2) for all d ∈ Dc do
3) perform text segmentation of d;
   {Detect word boundaries etc.}
4) if language of d recognized then
5) now apply stemming and mark stop-words in d;
   {stemming removes the ‘ing ’and maintains stems of frequent similar words.}
6) end if
7) end for

2nd phase (Frequent Phrase Extraction)
8) concatenate all documents;
9)Fc ← discover complete phrases;
10) Ff ← f : {f∈Fc ∈ frequency(f) > Term Frequency Threshold};

3rd phase (Cluster Label Induction)
11) A ← term-document matrix of terms not marked as stop-words and with frequency higher than the Term Frequency Threshold;
12) Σ, U, V ← SVD(A);
   {Product of SVD decomposition of A}
13) k ← 0;
   {Start with zero clusters}
14) n ← rank(A);
15) repeat
16) k ← k + 1;
17) \( q \leftarrow \frac{\sum_{i} F_{ki}}{\sum_{i} F_{ni}} \);  
18) until \( q < \) Candidate Label Threshold;  
19) \( F \leftarrow \) phrase matrix for \( F \);  
20) for all columns of UT \( k \) do  
21) find the largest component \( m_i \) in the column;  
22) add the corresponding phrase to the Cluster Label Candidates set;  
23) labelScore \( \leftarrow m_i \);  
24) end for  
25) calculate cosine similarities between all pairs of candidate labels;  
26) identify groups of labels that exceed the Label Similarity Threshold;  
27) for all groups of similar labels do  
28) select one label with the highest score;  
29) end for  
30) for all \( CL \in \) Cluster Label Candidates do  
31) create cluster \( C \) described with \( CL \);  
32) add to \( C \) all documents whose similarity to \( C \) exceeds the Snippet Assignment Theshold;  
33) end for  
34) for all clusters do  
35) clusterScore \( \leftarrow \) labelScore \( \times kCk \);  
36) end for  

4th Phase (Cluster Content Discovery)  
30) for all \( CL \in \) Cluster Label Candidates do  
31) create cluster \( C \) described with \( CL \);  
32) add to \( C \) all documents whose similarity to \( C \) exceeds the Snippet Assignment Theshold;  
33) end for  
34) put all unassigned documents in the “Others” group;  

5th Phase (Final Cluster Formation)  
35) for all clusters do  
36) clusterScore \( \leftarrow \) labelScore \( \times kCk \);  
37) end for  

B. Lingo Algorithm:  
A brief algorithm of the current Lingo is given below:  
1) Preprocess documents  
   - Extract frequent phrases and single words as cluster label candidates.  
   - Determine the assigned documents for each label candidate.  
   - Filter out the label candidates that contain less number of documents than the minimum cluster size threshold.  
2) Build the term-document matrix using the stems of the label candidates (except the stop words in the label candidates)  
3) Reduce the term-document matrix to the term-abstract concept matrix according to the desired cluster count base threshold.  
4) Match the abstract concepts with the cluster label candidates.  
5) Select the cluster label candidates that matched with an abstract concept as the labels of the determined clusters.  
6) Merge clusters that share higher percentage of documents than the cluster merging threshold.  

C. Preprocessing:  
Stemming and stop words removals are very common operations in Information Retrieval. Interestingly, their influence on results is not always positive in certain applications stemming yielded no improvement to overall quality.  
1) Stemming:  
The main aim of stemming is to reduce derivationally relate forms of words to a common base form by finding the roots i.e stem of a word. Stemming, is a technique for finding a semantic representation of an inflected word (usually a lemma) to decrease the impact of a language’s syntax.  
2) Stop Words Marking:  
The other clustering algorithm usually deletes the stop words, but lingo algorithm marks the stop words in order to generate a meaningful label.  
3) Text-Segmentation:  
Text-segmentation is a technique for dividing text into words and sentences that has many implementations.  
4) Text – Filtering:  
It filters the documents  

IV. MODULE DECOMPOSITION  

System is divided into five main modules  

1) Registration module  
2) Login  
3) Fetch mail  
4) Compose mail  
5) Cluster inbox  

A. Server Side Module:  
1) Fetch Mail Module:  
   - Fetch mails from third party server  
   - Synchronisation  
2) Cluster Inbox:  
   - Label generation.  
   - Clustering of mails.  

V. MATHEMATICAL MODULE  

<table>
<thead>
<tr>
<th>Sr No.</th>
<th>Description</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Let ( S ) be the System ( S={S1, S2, S3 } ) Where ( S1 )- module that authenticates ( S2 )- the module that forms label ( S3 )- the module that forms cluster</td>
<td>( S ) identifies system set</td>
</tr>
<tr>
<td>2</td>
<td>( S1={I,O,In,Fn,P,Sc,Fc,C} ) ( I={\text{Username, Password}} ) ( O={\text{Successful authentication}} ) ( In={\text{Server connection}} ) ( Fn={\text{Successful} } ) The module that authenticates and perform single sign-on.</td>
<td>The module that authenticates and perform single sign-on.</td>
</tr>
</tbody>
</table>

Constraints  
1) User name should be greater than four letters.
Primary actors in the system are hardware and software components, modules and protocols used in the system. It presents the meta-data contained in e-mail message and various techniques used for e-mail forensics. The paper also introduces several clustering algorithms that have functionalities to automatically analyse e-mail and produce labels and clusters

We hereby are thankful to all the members involved in the completion of paper. The one who guided us at each and every step. First of all, I wish to express my sincere appreciation to Prof. S Pratap Singh (Internal Guide of our project) for enlivening our interest in clustering and for the guidelines and advice he gave us throughout the development of the paper. I would also like to thank Ms. Saba Siraj, project incharge, for her valuable insights and observations. We wish to pay special gratitude to our principal Dr. Damodar Garkal and HOD Prof. Ritesh Thakur.

VIII. ACKNOWLEDGEMENT

We hereby are thankful to all the members involved in the completion of paper. The one who guided us at each and every step. First of all, I wish to express my sincere appreciation to Prof. S Pratap Singh (Internal Guide of our project) for enlivening our interest in clustering and for the guidelines and advice he gave us throughout the development of the paper. I would also like to thank Ms. Saba Siraj, project incharge, for her valuable insights and observations. We wish to pay special gratitude to our principal Dr. Damodar Garkal and HOD Prof. Ritesh Thakur. Thanks for their time and effort devoted to evaluation of Lingo. Finally, it is difficult to explain how grateful I am to my nearest and dearest whose endless patience and personal involvement helped me to complete this challenging venture.
REFERENCES


