RESEARCH ARTICLE OPEN ACCESS

Email Clustering Using Lingo Algorithm

Miss. Sangeeta Maharana, Miss. Minal Mohite, Miss. Pornima Wadekar
Department of Computer Science and Engineering
SP's Institute of Knowledge College Of Engineering, Pimple Jagtap
Pune - India

ABSTRACT
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. Thousands of emails are present in our inbox, to access a particular email, we have to go through those random thousands of mails. The existing techniques used in generating labels and clusters are not much efficient, and therefore a new and much more efficient technique of automatic generation of labels and clusters (Lingo Algorithm)have been used in our project. The advance feature added in our project is of Single Sign-On method, which will be proved beneficial to users in many aspects.
Keywords:- Lingo Algorithm, Single Sign-On, Clustering Of Emails, SMTP, POP3.

I. INTRODUCTION

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

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 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 System

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 drawback by applying some modification to our existing system, so that the advantages of our existing system remains as it is on the other side we can also remove the demerits of our existing system.we saw mainly four major drawbacks which our existing systems are facing.

1)Need to remember different user-id and passwords

To overcome this ,we are using Single sign on method. Through this, a user does not need to remember various passwords. He/She can access all his/her account by using a single user-id and password.

2)Waste of time creating new sessions of each service providers by logging into their respective domains.

In the existing system,whenever a user logins ,mail server of that particular domain creates a new session for that specific user.Time is wasted while creating these session for different differentlogins.If there is a single login ,then there wont be different session created from different accounts.

3)More waste of bandwidth and download capacity.

The bandwidth used by single account for any transfer of messages will be more.

4) People cannot access mails from different mail server at the same time from a single server.

A person having a gmail account cant switch to yahoo without getting log out from gmail.we can achieve this by using Single Sign On technique,in which a user can simultaneously work on different mail servers without switching from one to another.

5)Labels can be generated manually .

Here we can create folders for similar content and we can place those content inside that particular folder.To do

this process automatically ,we are using lingo algorithm. Lingo algorithm generates semantic labels and clusters.
In this approach crucial is the careful selection of cluster labels – the algorithm must ensure that the labels both differ significantly from each other and at the same time cover most of documents.

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 exixting 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. 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.
7) Form the final clusters for presentation.

A. Pseudo-Code Of Lingo Algorithm[2]

Ist 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
IIndPhase(**Frequent Phrase Extraction**)
8)concatenate all documents;
9) Fc ← discover complete phrases;
10)Ff ← f : {f∈ Fc ∈ frequency(f) > Term Frequency Threshold};
IIIrdPhase**(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 ← (Fki=1 Σii)/(Fni=1 Σii);
18)until q < Candidate Label Threshold;
19) F ← phrase matrix for Ff;
20) for all columns of UT k F do
21) find the largest component mi in the column;
22) add the corresponding phrase to the Cluster Label Candidates set;
23)labelScore ← mi;
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
4th Phase (**Cluster Content Discovery**)
30: for all CL ∈ 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 ← labelScore × kCk;
37: end for

B. *PREPROCESSING*

Stemming and stop words removal 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.

Example of Stemming

Words	Stems
Manual Man Manhood Manisha	Man
Common Computer Complex	Com

Table 1. Example Of Stemming

2)*Stop Words Marking:* The other clustering algorithm usually deletes the stop words,but lingo algorithm marks the stop words inorder to generate a meaningful label.

A	An	The	And	are	As
At	As	Be	By	for	Has
He	In	Is	It	on	of

Table 2. Example Of Stop-words

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 .i.e it removes unnecessary tags,symbols and words.

B. Frequent Phrase Extraction

We define frequent phrases as recurring ordered sequences of terms appearing in the input documents. To be a candidate for a cluster label, a frequent phrase or a single term must:

- It should appear in the input documents at least certain number of times (term frequency threshold).
  -The particular word should have been occurred for frequent number of time in the whole document.
- It should not cross sentence boundaries.
  -There are sentence boundaries in a sentence ,it should not cross those sentence boundaries.
- It should be a complete phrase.
  -A complete phrase is a substring of collected text.
- It must not begin nor end with a stop word.

C. Singular Value Decomposition

Singular value decomposition is used to reduce the dimentionality of the matrix and thus reduce the sparsity ,SVD also has a effect of smoothing the values ,The post-SVD column dimension of the matrix are minimum of 10% of the actual column dimensions or 300.Thus if the original column dimension were more than 3000,then the matrix is reduced to 300 columns.Each row represents the content .Thus the matrix translates into context vectors at each row of the matrix which are later clustered.

D. Cluster Label Induction

Once frequent phrases (and single frequent terms) that exceed term frequency thresholds are known, they are used for cluster label induction. There are three steps to this: term-document matrix building, abstract concept discov- ery, phrase matching and label pruning.

E. Cluster Content Discovery

In the cluster content discovery phase, the classic Vector Space Model is used to assign the input documents to the cluster labels induced in the previous phase. In a way, we re-query the input document set with all induced clus- ter labels.

F. Final Cluster Formation

Finally, clusters are sorted for display based on their score, calculated using the following simple formula: Cscore = label score × kCk, where kCk is the number of documents assigned to cluster C. The scoring function, although simple, prefers well-described and relatively large groups over smaller, possi- bly noisy ones. For the time being, no cluster merging strategy or hierarchy induction is proposed for Lingo.

1. *An illustrative example*[2]

Let us assume that the following input data is given
t = 5 terms
T1: Information
T2: Singular
T3: Value
T4: Computations
T5: Retrieval

p = 2 phrases
P1: Singular Value

P2: Information Retrieval

d = 7 documents
D1: Large Scale Singular Value Computations
D2: Software for the Sparse Singular Value Decomposition
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D5: Matrix Computations
D6: Singular Value Analysis of Cryptograms
D7: Automatic Information Organization

The normalized, tfidf -weighted term-document matrix b Atfidf is shown below together with matrix U (part of the SVD decomposition):

$$Atfidf = \begin{vmatrix} 0 & 0 & 0.56 & 0.56 & 0\ 0\ 1 & 0 \\ 0.49 & 0.71 & 0 & 0 & 0\ 0.71 & 0 \\ 0.49 & 0.71 & 0 & 0 & 0\ 0.71 & 0 \\ 0.72 & 0 & 0 & 0 & 1\ 0 & 0 \\ 0 & 0 & 0.83 & 0.83 & 00 & 0 \end{vmatrix}$$

$$U = \begin{vmatrix} 0 & 0.75 & 0 & -0.66 & 0 \\ 0.65 & 0 & -0.28 & 0 & -0.71 \\ 0.65 & 0 & -0.28 & 0 & 0.71 \\ 0.39 & 0 & 0.92 & 0 & 0 \\ 0 & 0.66 & 0 & 0.75 & 0 \end{vmatrix}$$

Now, we look for the value of k – the estimated number of clusters. Let us define quality threshold q = 0.9. Then the process of estimating k is asfollows: k = 0 → q = 0.62, k = 1 → q = 0.856, k = 2 →q = 0.959 ,so the number of clusters k = 2. To find descriptions of our clusters (k = 2 columns of matrix U), we calculate $M=Uk^{T}P$,where P is a term-document–like matrix created out of our frequent phrases and terms (values in P tfidf -weighted and normalized):

$$P = \begin{vmatrix} 0 & 0.56 & 1 & 0 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0.71 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0.83 & 0 & 0 & 0 & 0 & 1 \end{vmatrix}$$

$$M = \begin{vmatrix} 0.92 & 0 & 0 & 0.65 & 0.65 & 0.39 & 0 \\ 0 & 0.97 & 0.75 & 0 & 0 & 0 & 0.66 \end{vmatrix}$$

Rows of matrix M represent clusters, columns – their descriptions. For each row we select the column with maximum value – the two clusters are: Singular Value (score: 0.92) and Information Retrieval (score: 0.97). We skip label pruning as it is not needed in this example. Finally, documents are assigned to clusters by applying matrix Q, created out of cluster labels, back to theoriginal matrix b Atfidf.

$$\begin{vmatrix} 0 & 0.56 \\ 0.71 & 0 \end{vmatrix}$$

$$Q = \begin{matrix} 0.71 & 0 \\ 0 & 0 \\ 0 & 0.83 \end{matrix}$$

$$C = \begin{vmatrix} 0.69 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0.56 \end{vmatrix}$$

Information Retrieval [score: 1.0]
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D7: Automatic Information Organization
Singular Value [score: 0.95]
D2: Software for the Sparse Singular Value Decomposition
D6: Singular Value Analysis of Cryptograms
D1: Large Scale Singular Value Computations
Other: [unassigned]
D5: Matrix Computations

IV. 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. Application

The application of lingo algorithm is in each and every field of networking where the job of the server is to form clusters of the mails received in the inbox of an account and finally reduces space and time complexity.Large number of mails can be handled simultaneously by the server depending on the network load.Also the Account holder doesn't get distributed while the clustering is in progress.So it saves a large amount of time of administrator and also the space on the server.

B. Architeture

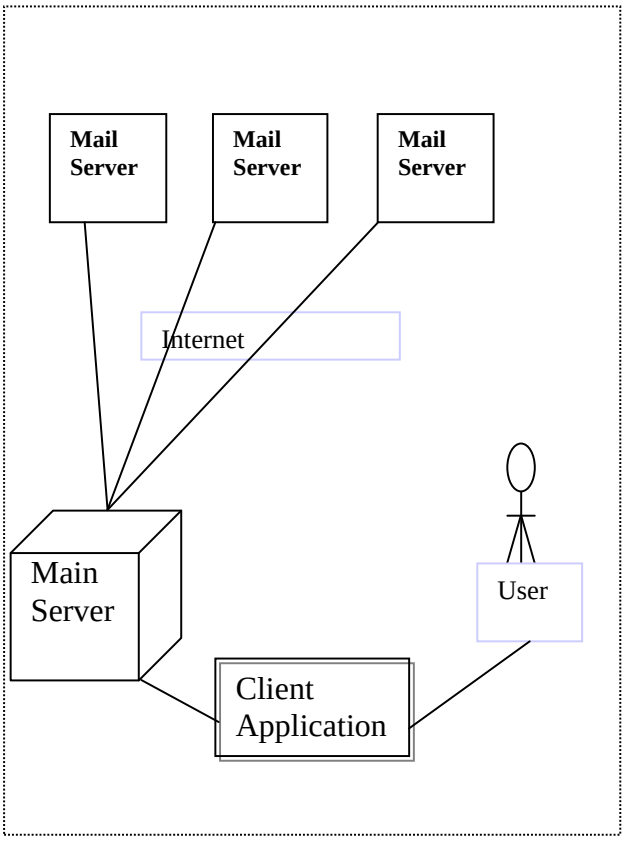

Fig 1. Architecture Of Proposed Email System

All account holders store their information by registering themselves at server side. It request for the validation of the user's authentication. Administrator maintains the database which contains user information and information about account holders. Admin view the request and gives response to it accordingly. The architecture above shows that the overall external structure of our system .In this we can see a main server which is our proposed system. This main server is connected to three different mail servers. They are connected through internet. Client application is an intermediate between user and main server. User sends the inbox to the main server .The main server then clusters the inbox and send the clustered inbox to the user.

V. RESULTS

There are two parameters based on which our systems performance is based Precision and Recall. Let us consider D be the set of documents, set A of documents was retrieved for users query is the set of documents that are relevant to documents present in D. RA, be the intersection of R and A. Definition — Precision is the fraction of the retrieved documents which is relevant: Precision = |RA|/|A| Definition — Recall is the fraction of the relevant documents which has been retrieved: Recall = | RA |/|R| Consider a dataset for email system which contains 100 emails regarding different topics. Data is fetched from the email server to the database and the lingo algorithm works on these emails content. Hence we present result of our system. We calculated the Precision and Recall measures for each and every cluster shown by our system.

VI. CONCLUSION

An email clustering approach is proposed through this project.The main aim is to perform text similarities on emails.It shows how text similarities can be used to form clusters in emails. The proposed technique is implemented using open source technology java.The mails will be finally in a clustered format so as to minimize the job of searching for the users. Hence this will also reduce the consumption and make browsing user friendly. The future scope of the work could be incorporating the similarity of the email attachments etc. for the more accurate clustering of the emails. The other direction of the proposed work could be applying the proposed email similarity function for the more email mining operations like thread summarization, automatic answering, and applying the same Techniques for other Email datasets for participating all the attributes of the emails and achieving more accurate results.

ACKNOWLEGEMENT

During the course of writing our paper I had the genuine pleasure to work with a number of people without whose support completion of this project would be barely possible. First of all, I wish to express my sincere appreciation to Mr.Pratap Singh for enlivening our interest in clustering and for the guidelines and advice he gave us throughout the development of the project. I would also like to thank Ms.SabaSiraj, project in charge , for her valuable insights and observations. We wish to pay special gratitude to our principal Dr.R.SJahagirdar 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

[1] LUIA FILIPE DA CRUZ NASSIF AND EDUARDO RAUL HRUSCHKA **"Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection."**ieee transactions on information forensics and security,1, january 2013**,** vol. 8, no **.**

[2] Decomposition Stanis law Osin´ski, Jerzy Stefanowski, and Dawid Weiss **"Lingo: Search Results Clustering Algorithm Based on Singular Value"**Institute of Computing Science, Poznan´ University of Technology, ul. Piotrowo 3A, 60–965 Poznan´,Poland,Email:stanislaw.osinski@man.poznan.pl,{jerzy.stefanowski,dawid.weiss}@cs.put.poznan.pl

[3] Peter Hannappel, Reinhold Klapsing, and GustafNeumann,"**MSEEC—a multi search engine with multiple clustering‖ ."** Proceedings of the 99 Information Resources Management Association Conference , May 1999.

[4] Zhang Dong **"Towards Web Information Clustering‖** ", PhD thesis, Southeast University, Nanjing, China, 2002

[5] Irmina Mas lowska,"**Phrase-Based Hierarchical Clustering of Web Search Results‖** ", In Proceedings of the 25th European Conference on IR Research, ECIR2003, volume 2633 of Lecture Notes in Computer Science, pages 555–562, Pisa, Italy, 2003. Springer.

[6] Stanislaw Osinski, Jerzy Stefanowski and DawidWeiss**,"Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition‖ "**, Institute of Computing Science,Pozna´n University ofTechnology, ul.Piotrowo 3A, 60–965 Pozna´n, Poland

[7] Stanislaw Osinski and DawidWeiss,PoznanUniversity**,"‖ A concept Driven Algorithm For Clustering Search Result‖ ",**2005 IEEE.

[8] Claudio Carpinato, Stanislaw Osinski and DawidWeiss,"**A Survey Of Web Clustering Engines‖ "**, 2009 ACM. .

[9] Dawid Weiss and Jerzy Stefanowski.**"Web search results clustering in Polish:Experimental evaluation of Carrot ".** In Proceedings of the New Trends in IntelligentInformation Processing and Web Mining Conference, Zakopane, Poland,2003.

[10]. Oren Zamir and Oren Etzioni. Grouper: "**a dynamic clustering interface to Websearch results".** Computer Networks (Amsterdam, Netherlands: 1999), 31(11–16):1361–1374, 1999.