Clustering of Emails

Prof S Pratap Singh[1] Minal Mohite[2] Sangeeta Maharana[3] Pornima Wadekar[4]
[1]Dean Student Welfare at IOK
[1,2,3,4]Department of Computer Engineering
[1,2,3,4]SP'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 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:

```
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;
    q ← (Fki=1 Σii)/(Fni=1 Σii);
until q < Candidate Label Threshold;
F ← phrase matrix for Ff;
for all columns of UT k F do
    find the largest component mi in the column;
    add the corresponding phrase to the Cluster Label Candidates set;
    labelScore ← mi;
end for
calculate cosine similarities between all pairs of candidate labels;
identify groups of labels that exceed the Label Similarity Threshold;
for all groups of similar labels do
    select one label with the highest score;
end for
4th Phase (Cluster Content Discovery)
for all CL ∈ Cluster Label Candidates do
    create cluster C described with CL;
    add to C all documents whose similarity to C exceeds the Snippet Assignment Theshold;
end for
put all unassigned documents in the “Others” group;
5th Phase (Final Cluster Formation)
for all clusters do
    clusterScore ← labelScore × kCk;
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

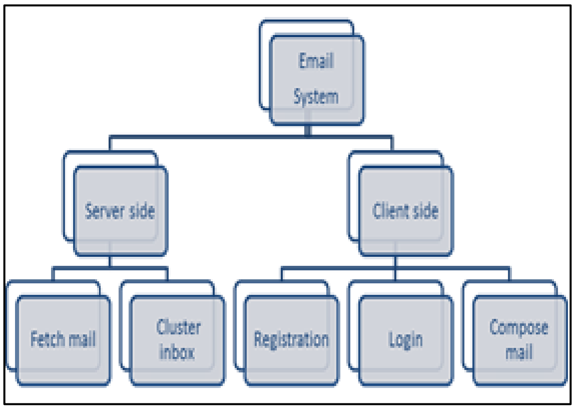

Fig. 1: Module Decomposition

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

Sr No.	Description	Observation
1	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	S identifies system set

2	S1={I,O,In,Fn, P,Sc,Fc,C}
	I={Username, Password}
	O={Successful authentication}
	In={Server connection}
	Fn={Successful
		The module that authenticates and perform single sign-on.
		Constraints
		1) User name should be greater than four letters.

	Authentication}
	P={P1,P2}
	P1=Verify whether all the fields are filled.
	P2=Verify the entered fields.
	Sc=Authentication is successful
	Fc=Authentication is unsuccessful
	I=Set of inputs to the system
	O=Set of outputs
	In=Initial Conditions
	Fn=Final Conditions
	P=Process involved in our module
	Sc=Success cases
	Fc=Failure Cases
	C=Constraints
		2) Passwords should not be less than 8 alphanumeric characters.
		3) phone number should not be equal to 10 digits.
		4) Client should be connected to the server.

3	S2={I,O,In,Fn, P,Sc,Fc,C}
	I={I1,I2,I3}
	Where let I1=mails from yahoo,
	I2=mails from gmail,
	I3=mails from Aol.
	O={L1,L2…Ln}
	Li=Generated label
	In={m1,m2,m3….mn}
	mi=unclustered set of mails
	Fn={Set of Different Labels}
		The module that Forms label.
		Constraints
		1) All mails must be fetched before label generation.

Table 1: Mathematical module

VI. Result Analysis

A. Graph:

Graph description comparing the proposed algorithm of the project and the alternative algorithms.

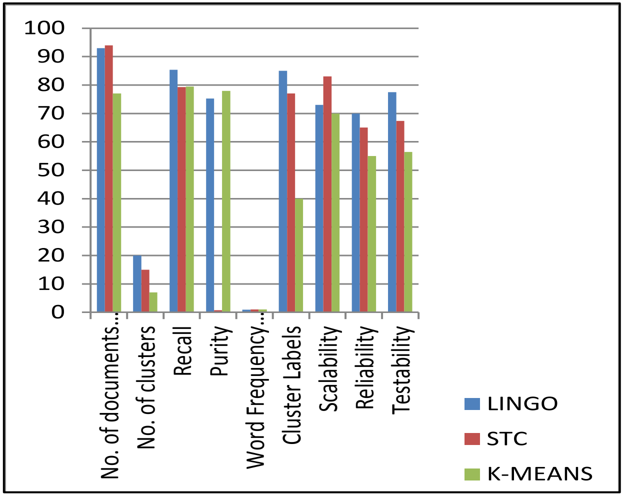

Fig. 2: Comparison of LINGO STC and KMEANS

1) LINGO-Label Induction and Grouping Algorithm
2) STC-Suffix Tree Clustering
3) K-means

B. Result:

Each algorithm has its own merits and demerits, Lingo produces high cluster diversity, the Small outliers are highlighted well, In STC and K-means algorithms the small outliers are rarely highlighted .In Lingo the number of clusters produced are more when compared to other two algorithms. With respect to the cluster labels, in LINGO they are descriptive but lengthy, not very descriptive in K-Means, but in STC cluster labels are small but very appropriate. The Scalability is high in STC compared to Lingo and K-Means. Other features of K-Means clustering are Running time: O(KN) (K = number of clusters) ,Fixed threshold ,Order dependent. Features of STC are Overlapping clusters, Non-exhaustive, Linear time, and High precision.

C. 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.

VII. Conclusion

Overlapping clusters, Non E-mail is a widely used and highly distributed application involving several actors that play different roles. These actors include hardware and software components, services and protocols which provide interoperability between its users and among the components along the path of transfer. It illustrated logical e-mail architecture and underlining various core 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 algorithm that have functionalities to automatically analyse e-mail and produce labels and clusters

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 throughtout 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

[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.