A System for Email Users Clustering, Automatic Answering and Automatic Text Summarization.

Akshay Jayale[1], Sudershan Karad[2],Sudhakar Gawade[3]

[1]MIT Academy Of Engineering, Pune University
akshayuddhavraojayale@gmail.com

[2]MIT Academy Of Engineering, Pune University
karad_sudershan@yahoo.com

[3]MIT Academy Of Engineering, Pune University
sud_gawade121@gmail.com

Keywords-**Email clustering, attachments, automatic answering, text summarization, fuzzy logic, etc..**

I. Introduction

Email communication has came up as the most effectiveand 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.

The dynamic dataset is used here to find text similarities. The dynamic dataset can downloads all the attachments of email users whose email ids along with respective password are stored in the system initially.

A. Email Users Clustering

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 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 . Email users clustering can be explained as the attachments with the same concept give an appropriate name to cluster and then put all the users who are discussing the similar content.

B. Automatic answering

The automatic answering system can answer the cluster of email users who want some type of information like if some students in the college who require fee structure of current academic years, then the administrator of system can cluster the email users on word "fee structure ". The automatic system can send them link to these cluster of email users where they get fee structure of current year.

C. Automatic Text Summarization

Automatic text summarization is the summary of the source text created by machine to present the most important information in a shorter version of the original text while keeping its main content and helps the user to quickly understand large volumes of information. We propose text summarization based on fuzzy logic method to extract important sentences as a summary. Summarization by extraction is the identification of important features such as sentence length, sentence location, and term frequency, number of words occurring in title, number of proper nouns and number of numerical data.

II. Related and previous work

A. Email Users Clustering

The Enron Email Dataset Database Schema and BriefStatistical Report which show how the distribution of emails per users and showing the network how the employees are connected.

Email classification can be applied to several different applications, including filtering messages based on priority, assigning messages to user-created folders, or identifying SPAM. One major consideration in the classification is thatof how to represent the messages. Specifically, one mustdecide which features to use, and how to apply thosefeatures to the classification. Manco, et al. defined threetypes of features to consider in email: unstructured text,categorical text, and numeric data.

Unstructured text in email consists of fields like thesubject and body, which allow for natural language text ofany kind. Categorical text includes fields such as "to" and "from". These differ from unstructured text fields in thatthe type of data which can be used in them very well defined. Numeric data is data with numeric values.

B. Automatic Answering

Q&A system research received considerable attention from the research community through Text Retrieval Conference [8] Q&A track since 1999. The original aim of the track is to systematically evaluate both academic and commercial Q&A systems. Maybury has discussed the characteristics of Q&A systems and resources needed to develop and evaluate such systems. Main approaches in Q&A systems could be found in in which template-based approach discussed in detail.

Although, most Q&A systems are based on Web environments, SMS has also been used as an environment in contexts such as in learning and agriculture.

C. Automatic text summarization

Automatic text summarization dates back to the
Fifties, when Luhn created the first summarization system in 1958. Rath et al.in 1961 proposed empirical evidences for difficulties inherent in the notion of ideal summary. Both studies used thematic features such as term frequency, thus they are characterized by surface-level approaches. In the early 1960s, new approaches called entity-level approaches appeared; the first approach of this kind used syntactic analysis . The location features were used in, where key phrases are used dealt with three additional components: pragmatic words (cue words, i.e., words would have positive or negative effect on the respective sentence weight like significant, key idea, or hardly); title and heading words; and structural indicators (sentence location, where the sentences appearing in initial or final of text unit are more significant to include in the summary.

The feature extraction techniques are used to locatethe important sentences in the text. For instance, Luhn looked at the frequency of word distributions asfrequent words should indicate the most importantconcepts of the document. Some of features are usedin this research such as sentence length. Somesentences are short or some sentences are long. Whatis clear is that some of the attributes have moreimportance and some have less, so they should havebalance weight in computations and we use fuzzylogic to solve this problem by defining themembership functions for each feature.

III. PRE-PROCESSING

The pre-processing includes parsing, stemming andemail representation technique for parsed information.Parsing of email documents is required to retrieve variousattributes information separately like subject, contentsetc. Stemming is required to clean the text informationavailable in email attributes. And finally one representationtechnique is required for effective representation of emaildocuments, in which all the attributes associated with anemail can be made accessible.

For automatic text summarization there are four main activities performed in this stage: Sentence Segmentation,Tokenization, Removing Stop Word, and Word Stemming. Sentence segmentation is boundary detection and separating source text into sentence. Tokenization is separating the input document into individual words. Next, Removing Stop Words, stop words are the words which appear frequently in document but provide less meaning in identifying the important content of the document such as ‘a’, ‘an’, ‘the’, etc.. The last step for preprocessing is Word Stemming; Word stemming is the process of removing prefixes and suffixes of each word.

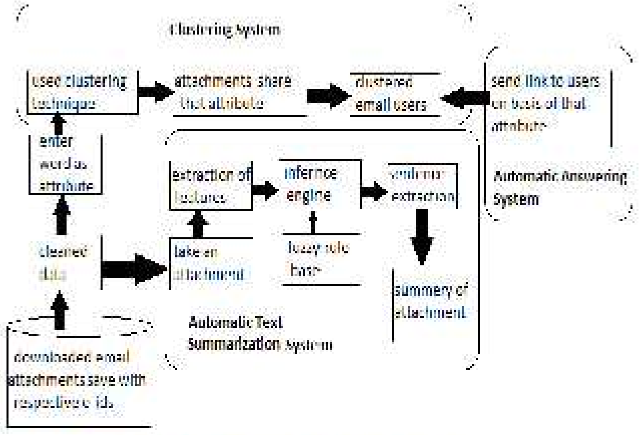
Clustering System
used clustering technique
attachments share that attribute
clustered email users
send link to users on basis of that attribute
Automatic Answering System
enter word as attribute
extraction of features
infernce engine
sentence extraction
cleaned data
take an attachment
fuzzy rule base
summery of attachment
Automatic Text Summarization System
downloaded email attachments save with respective ids

Fig.1 Clustering,Automatic Answering and Automatic Text Summarization System

A. Features In Text Summarization

It is necessary to represent the sentences as vectors of features. These features are attributes that attempt to represent the data used for their task. We focus on eight features for each sentence. Each feature is given a value between ‘0’ and ‘1’. Therefore, we can extract the appropriate number of sentences according to 20% compression rate.

There are 8 features as follows:

1) Title feature: The number of title words in sentence, the words in sentence that also occurs in title gives high score. This is determined by counting the number of matches between the content words in a sentence and the words in the title.

The score is to calculated for this feature which is the ratio of the number of words in the sentence that occur in the title over the number of words in title.

$$Score_{f1}(S_i) = \frac{No.Title\ word\ in\ S_i}{No.Word\ in\ Title} \qquad (1)$$

2) Sentence length: The number of words in sentence, this feature is useful to filter out short sentences such as datelines and author names commonly found in news articles. The short sentences are not expected to belong to the summary. Use of normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

$$Score_{f2}(S_i) = \frac{No.Word\ occurring\ in\ S_i}{No.Word\ occurring\ in\ longest\ sentence} \qquad (2)$$

3) Term weight: The average of the TF-ISF (Term frequency, Inverse sentence frequency). The frequency of term occurrences within a document has often been used for calculating the importance of sentence.

$$Score_{f3}(S_i) = \frac{Sum\ of\ TF\text{-}ISF\ in\ S_i}{Max(Sum\ of\ TF\text{-}ISF)} \qquad (3)$$

4) Sentence position: Whether it is the first 5 sentences in the paragraph, sentence position in text gives the importance of the sentences. This feature can involve several items such as the position of a sentence in the document, section, and paragraph, etc., proposed the first sentence is highest ranking. The score for this feature: we consider maximum positions of 5. For instance, the first sentence in a paragraph has a score value of 5/5, the second sentence has a score 4/5,and so on.

$$Score_{f4}(S_i) = 5/5 \text{ for } 1^{st},\ 4/5 \text{ for } 2^{nd},\ 3/5 \text{ for } 3^{rd},\ 2/5 \text{ for } 4^{th},\ 1/5 \text{ for } 5^{th},\ 0/5 \text{ for other sentences} \qquad (4)$$

5) Sentence to sentence similarity: Similarity between sentences, for each sentence s, the similarity between s and each other sentence is computed by the cosine similarity measure. The score of this feature for a sentence s is obtained by computing the ratio of the summary of sentence similarity of sentence s with each other sentence over the maximum summary.

$$Score_{f5}\ (S_i) = \frac{Sum\ of\ Sentence\ Similarity\ in\ S_i}{Max(Sum\ of\ Sentence\ Similarity)} \qquad (5)$$

6) Proper noun: The number of proper noun in sentence, sentence inclusion of name entity (proper noun). Usually the sentence that contains more proper nouns is an important one and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of proper nouns in sentence over the sentence length.

$$Score_{f6}\ (S_i) = \frac{No.\ Proper\ nouns\ in\ S_i}{Length\ (S_i)} \qquad (6)$$

7) Thematic word: The number of thematic word in sentence, this feature is important because terms that occur frequently in a document are probably related to topic. The number of thematic words indicates the words with maximum possible relativity. The top 10 most frequent content word are used for consideration as thematic. The score for this feature is calculated as the ratio of the number of thematic words in the sentence over the maximum summary of thematic words in the sentence.

$$Score_{f7}\ (S_i) = \frac{No.\ Thematic\ word\ in\ S_i}{Max(No.\ Thematic\ word)} \qquad (7)$$

8) Numerical data: The number of numerical data in sentence, sentence that contains numerical data is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data in sentence over the sentence length.

$$Score_{f8}\ (S_i) = \frac{No.\ Numerical\ data\ in\ S_i}{Length\ (S_i)} \qquad (8)$$

B. Text Summarization based on Fuzzy Logic

In order to implement text summarization based on fuzzy logic. First, the features extracted in the previous section are used as input to the fuzzy inference system. We used Triangular membership functions. The generalized Triangular membership function depends on three parameters a, b, and c as given by

$$f(x, a, b, c) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$$

The parameters a and c set the left and right "feet," or base points, of the triangle. The parameter b sets the location of the triangle peak. For instance, membership function of number of words in sentence occurred in title is show in Figure 2

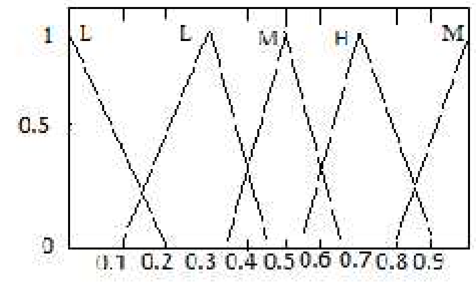

Fig. 2 Membership function of number of words in sentence occured in title

Afterword, we use fuzzy logic to summarize the document. A value from zero to one is obtained for each sentence in the output based on sentence features and the available rules in the knowledge base. The obtained value in the output determines the degree of the importance of the sentence in the final summary.

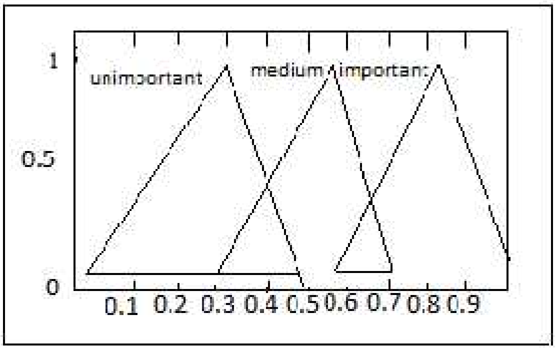

Fig 3 Membership function of number of output

The input membership function for each feature is divided into three membership functions which are composed of insignificant values (low (L) and very low (VL), Median (M) and significant values high (H) and very high (VH). For example, membership functions for title feature: No Word In Title {VL, L, M, H and VH}. Likewise, the output membership function is divided into three membership functions: Output {Unimportant, Average, and Important}. The most important part in this procedure is the definition of fuzzy IF-THEN rules. The important sentences are extracted from these rules according to our features criteria. For example our rules are showed as follow.

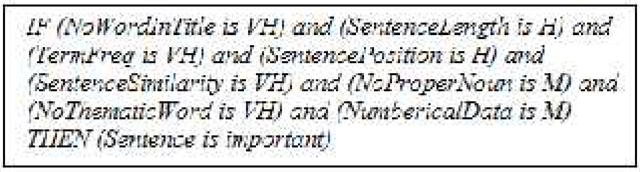
IF (NoWordinTitle is VH) and (SentenceLength is H) and (TermFreq is VH) and (SentencePosition is H) and (SentenceSimilarity is VH) and (NoProperNoun is M) and (NoThematicWord is VH) and (NumbericalData is M) THEN (Sentence is important)

Fig 4 Sample of IF-THEN rules

IV. IMPLEMENTATION

The clustering algorithm is implemented using open source technologies and algorithm is applied over the dynamic data set which we are prepared. Java is selected as the programming languages and the other open source API's (Application Programming Interfaces) to Support the other functionalities. Netbeans is used as a development IDE (Integrated Development Environment) for Java and library of other technologies are added as external jar (Java Archives) in the Netbeans.

For automatic answering we used the concept of multi threads in the java is used.

The application shall be implemented using java technology using Netbeans as IDE and JDK 1.6, JRE 6 will be the Standard kit. Programming technology is used as Java and JDBC is used as connector driver. For JDBC we shall use sql jar file.

Fuzzy logic theme is used to create parameter limits and access specifications. For data base we used My Sql 5.0.And for that we used the external jar.

V. CONCLUSION

In this paper we have studied three systems namely as clustering, automatic answering and automatic text summarization.

The clustering technique shows the email attributes and how the text similarities are used to cluster the users.

The automatic answering system shows us how to answer the group of user on the basis of attribute which is used to know purpose.

The automatic text summarization system shows automatic text summarization for important sentences extraction with the important features based on fuzzy logic; title feature, sentence length, term weight, sentence position, sentence to sentence similarity, proper noun, thematic word and numerical data.

REFERENCES

[1] LaddaSuanmali,"Feature-Based Sentence Extraction Using Fuzzy Inference rules" 978-0- 7695-3654-5/09©2009 IEEE.

[2] Jayadev Gyani, Syeda Farha Shazmeen" A Novel Approach for Clustering E-mail Users Using Pattern Matching"

[3] Inderjeet Mani and Mark T. Maybury, editors, Advances in automatic text summarization MIT Press. 1999.

[4] R. Witte and S. Bergler, "Fuzzy co reference resolution for summarization," In Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS). Venice,Italy: Università Ca' Foscari. pp.43–50. 2003.

[5]Tilani Gunawardena, Medhavi Lokuhetti, Nishara Pathirana, Roshan Ragel and Sampath Deegalla "An Automatic Answering System with Template Matching for Natural Language Questions"

[6] H. P. Luhn, "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, vol. 2, pp.159-165.1958.