International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 5 Issue 3, March-2016

A Fuzzy Based Ontology Extraction For Efficient E-Mail Classification

Suma T[1], Dr. Kumara swamy Y S[2]

[1]Phd scholar, Department of CSE, JJT University, Rajasthan, India
[2]Professor & Dean (R&D), Department of CSE, Nagarjuna College of Engg. & Technology, Karnataka, India

ABSTRACT

Day to day uses of multimedia growth is very rapid in that case dependability on business message transfer, advertisement mail, promotional mail is very high in that case manage all kind of email is very necessary in busy life style. If E-mail is not categorize and managed properly in that case some important schedule also loss. If E-mails are managed properly and important information gets updated itself, it is very useful. Here author purpose is categorized the mail and update the calendar based on Natural language processing using ontology and Fuzzy logic. Extraction of concept and make cluster based on fuzzy logic and check that the each email has number of concept. The author conducted experiment evaluation to check precision and recall performance and its efficiency.

Keywords: Clustering, Fuzzy, Ontology, NLP, Semantic.

1. INTRODUCTION

NLP stands for Natural Language Processing is a technology which investigates and evaluates human language itself. It is a modern way of computation, which generalized from Artificial intelligence. NLP that includes the flexibility and behavioral competence, it also involves thinking and understanding of the cognitive and mental processes behind behavior, it is a multi-dimensional process. NLP is a powerful technique which is very useful to extract computable information from unstructured data. NLP applications as a technology-driven problem, concentrating on intrinsic factors ("precision" and "recall") as the prime factor of adoption and success. Here Electronic- mail (E-Mail) processing is our area of interest by using NLP techniques. The growth of information society has led to the rise of new communication environments and technologies. One of the most important requirements to such atmospheres is the fast access to transfer of message within second that meets end client necessities as precisely and promising as possible [1]. In recent times the major growth has been attained in creation of mail transfer schemes based on Web technologies. Web-based message or document transfer very fast, just-in-time, relevant, and at any time or from any place we can access it. As we know that as technologies is some negative things also growing simultaneously like junk –email which we don't want to receive.

Structure of e-mail system is divided into three parts

•MTA (mail transfer agent)
•MDA (mail delivery agent)
•MUA (mail user agent)

MTA work is to filter the mail and check header and body of e-mail, before sending. MDA filter the mail which it receive from MTA, lastly at client side MUA works which filtered the receive mail [2]. The volume of E-Mail transactions has seen a rapid growth. Per day Users receive no. of E-mail but unable to read or process one by one and in that case not fetch the valuable information. Therefore it is essential for NLPs to be added to the E-Mail structures based on comprehensive ontologies extracted from the E-Mails. NLPs added to the E-Mails would empower the user for real processing effective workflow management. For Example if a User from the institute receives an E-Mail regarding an exam schedule in the college normally he would have to manually process the E-Mail and update his calendar of events. Adding NLPs to the E-Mails would automatically update the user's calendar with the relevant details extracted from the ontology extraction engines. Another problem that persists is the unstructured formulation of the E-Mails which makes it difficult to extract effective ontologies

Conversion of high-dimensional data set into lower dimensional data set through type matching extraction method. Here author purpose a self-clustering type matching clustering algorithm based on fuzzy, which is an approach for E-Mail classification. Mails that are similar to each other grouped into the same cluster. Each cluster has some specific function called mean and deviation. If a mail is not similar to any existing cluster, a new cluster is created for that mail. Similarity between a mail and a cluster is predict by this mean and variation. Each cluster have an extracted type matching, this extracted type matching is weighted combination of E-mails in the cluster. Hard, soft and mixes are three ways of weighted combination of E-Mails.

More specifically, in the e-mail application, we provide content creators with a keyword extractor which allows for semi-automatic metadata annotation of the learning objects. Keyword explored in the NLP research area. In our methodology, we adopt those results to the eLearning situation by using extraction has been widely statistical measures in combination with linguistic processing to detect relevant words which are good keyword candidates. In addition, we adopt a glossary candidate detector which allows for the creation of glossaries to be linked to learning objects. The glossaries are based on the definitions of the relevant terms which are attested in the learning objects. Definition extraction is the topic of much current research and techniques have been developed to this end within the Natural Language Processing and the Information Extraction communities mainly based on grammars that detect the relevant patterns and machine learning or artificial intelligence methods. Here author proposed fuzzy ontology based E-mail extraction and their categorization, concepts and relation. Based on that author update calendar automatically and showing pending meeting date, important news etc.

The rest of the paper is organized as follows. Section 2 highlights the previous research work carried out in the field of NLP E-Mail. The NLP E-Mail process is described in Section 3. Our proposed system, the NLP Extraction Engine is presented in the fourth section of the paper. Section 4 also discusses the preliminaries of sectioning and sub sectioning, preprocessing, fuzzy ontology extraction using the weight and association functions. Section 4 also discusses the representation of the DL obtained via SROIQ axioms. The evaluations based on the derived Ontology Extraction Engine are thereafter illustrated in the next section. Sections 6 propose the conclusion and the open issues in the areas of the NLP E-Mail.

2. LITTERAURE SURVEY

Communication through E-Mail grows very rapidly [3]. Studies show that people check their E-Mails minimum three times a day. For that reason, companies and firms send out E-Mail blasts almost every day to generate leads [4]. A user can receive, every day, hundreds of E-Mails from different sources promoting their product. Hence, there have been technologies that would help sort E-Mails to eliminate the hassle of processing each and every E-Mail. According to the various stages of email transmission and methods, this paper [2] advancing the intelligent of multi-level mail filtering scheme based on natural language processing. Its main type matchings include black and white list filter. McDowell introduced the idea of NLP E-Mail which refers to “an E-Mail message consisting of a structured query coupled with corresponding explanatory text, based on a number of NLP E-Mail Processes that represent commonly occurring workflows within E-Mail” [5]. It is the process of generation of the summary of an input document by extracting the representative sentences from it. In this paper[6], author present a technique for generating the summarization of domain specific text from a single Web document by using statistical NLP techniques on the text in a reference corpus and on the web document. Blocking spam mail is a tuff task we still receive spam mail very frequently. Here [7] an approach based on Natural Language Processing (NLP) for the penetration of spam filters is proposed. Preliminary results using Spam Assassin are provided indicating the feasibility of the proposed approach.

Statistical based method is a well-known method in spam filter plan. Bayesian founded in statistical method helps the possibility of spam keywords within E-Mail classification. Ontology is employed in one of learning instruments for E-Mail classification approach based on machine learning [8]. Here [9] Suffix Tree Clustering (STC) algorithm is presented. By the using of NLP algorithm selection of the noun, verb and entity is taken out from the given input of STC. In [11] automatic taxonomy construction is find out in which classes is formed by the grouping of nouns. Linguistic pattern also aimed at discovering taxonomic relations. [12] Here author design a frame work which predicts the read, reply, delete, or delete-without read. Author used horizontal and vertical learning approach for regularization purposes. Users always have a choice of ignoring suggestion for read, write, reply kinds of things. Author offer decent level of recall for active user but not suggest important date schedule for upcoming event. [13] in this author proposed the multi-user personalized email community detection method, for creating the group of emails based on their semantic intimacy and structure. Creating the social graph from personalized emails of multi-user is adopted here. From above survey find that in E-mail for various kinds of work is done but based on E-mail important schedule, meeting, date and time calendar is not updated itself, if calendar updated itself based on E-mail data for user it is very useful, and user not worry about reading each mail.

Type matching Clustering

An approach for reduction in type matching is called type matching clustering, here grouping of similar kind of type matching in one cluster. Clustering of E-mails can be understand as E-mail with same motive or concept can be provided an appropriate name to each cluster and then fed all message into their regarding folders. It may be clustering of the user who are chatting over similar topic [10]. In ontology Extraction input is given which is processed and formed a cluster from Enron corpus different cluster is forming based on data like ness, date , meeting etc. after clustering mechanism output like is find based on query and then after evaluation is performed.

Ontology Extraction

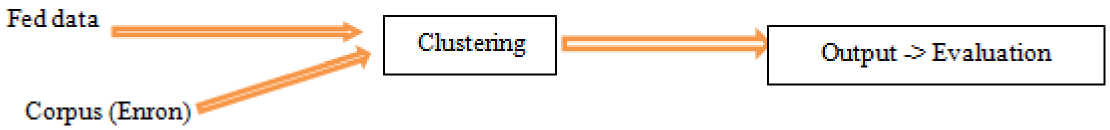

Figure. 1 Ontology extraction of E-mail

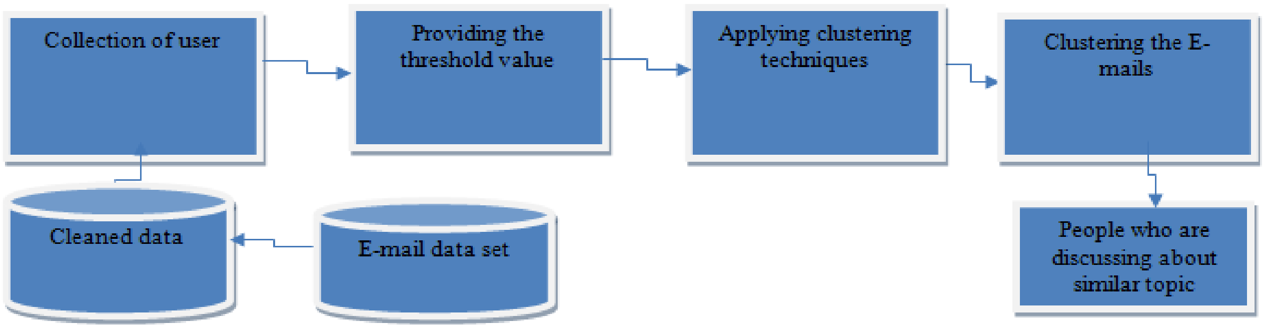
Collection of user
Providing the threshold value
Applying clustering techniques
Clustering the E-mails
Cleaned data
E-mail data set
People who are discussing about similar topic

Figure. 1 Clustering of E-mail

3. PROPOSED V2V ENVIROMENTAL MODELING IN VANETS

Frequently used words or terms are consider for NLP techniques in clustering, concept extraction classification etc. Most frequent term can be of two type Generic which is present in all most all E-mails and bespoke which depends on the sender profession.

A. Concept Extraction

In Concept Extraction can be defined in two part 1. E-mail Preprocessing 2. Handling the body part of E-mail. Concept extraction is process of fetching term from structured and semi-structured E-mail. It process the E-mail and divide header and body part of E-mail, from header part extracts terms. From the body of E-mail sentence is split into words, and phrases. Parse the sentence and find the noun and noun phrase in the sentence.

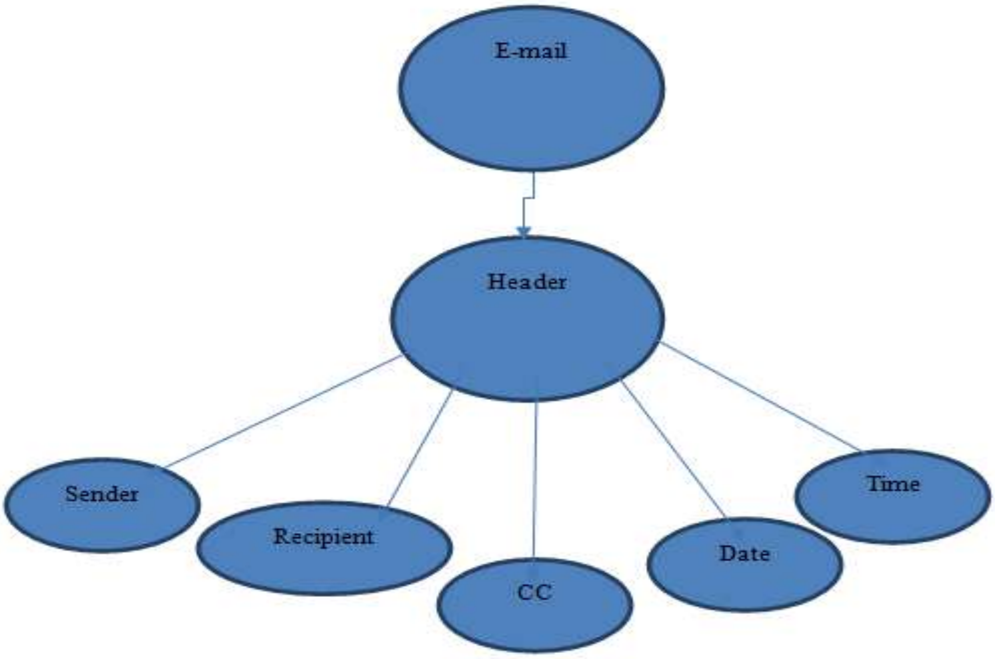

Figure. 3 Concept extraction from E-Mail

Step 1: in this step semi structured or unstructured text of emails are processed. The main aim for preprocessing of E-mail is to prepare parts of emails for ontology extraction. Sender receiver dates etc. are generated by the XML description in preprocessing step.

Step 2: After preprocessing NLP system analyzing of text is done based on their syntactic and semantic property. Text is preprocessed by a part-of-speech tagger and it groups the word in grammatical category. Sentence splitter, tokenizer, POS tagger etc. are the most important NLP component that is used for grouping the word and in research.

Step 3: Email is defining as a root in this ontology then concept are extracted by the help of pattern matching by XML tags at the time of preprocessing of an E-mail showing in figure 2.

B. Clustering Process

To avoid the problem of existing system Suppose we have E-mail set E, and in that set number of e-mail is k like e_1, e_2, e_3 ... . e_k , all together with a pattern vector V of n concepts $v_1, v_2, v_3 \dots . v_n$ each email belongs to one of the category so d category as $c_1, c_2, \dots c_d,$. We make one concepts pattern for each concepts in V. for concepts V_m , its concepts pattern A_m is defined, by

$A_m =< a_{m1}, a_{m2}, \dots., a_{md} >$ $=< (c_1 \vert p_m), (c_2 \vert p_m), (c_3 \vert p_m), \dots., (c_d \vert p_m) >$	(1)

Where ,

$$PL(c_l|v_m) = \frac{\sum_{t=1}^{k} e_{tm} \times \delta_{tl}}{\sum_{t=1}^{k} e_{tm}} \tag{2}$$

For $1 \leq l \leq d$. Here e_{tm} indicate the number of occurrence of v_m in E-mail e_t, δ_{tl} can be defined as
$\delta_{tl} = 1$ when email e_t, belongs to category c_l, if it not belongs to any category value of $\delta_{tl} = 0$.
Here we have n concepts pattern, take an example that we have four emails e_1, e_2, e_3 e_4 belonging to category c_1, c_1, c_2 and c_2 respectively. Now occurrences of V_1 in these emails be 1, 2, 3, and 4 respectively. Concepts pattern of A_1 of V_1can be calculated as

$$PL(c_1|v1) = \frac{1 \times 1 + 2 \times 1 + 3 \times 0 + 4 \times 0}{1 + 2 + 3 + 4} = 0.3\,,$$

$$PL(c_2|v1) = \frac{1 \times 0 + 2 \times 0 + 3 \times 1 + 4 \times 1}{1 + 2 + 3 + 4} = 0.7\,,$$

$$A_1 =< 0.3,\ 0.7 >. \tag{3}$$

Our motive is the making cluster, based on these concepts pattern. A cluster has certain number of concepts pattern and is the product of d one-dimensional Gaussian function. Let D be a cluster containing t concepts pattern $a_1, a_2, a_3, \dots . a_t$.
Let $a_l =< a_{l1}, a_{l2}, a_{l3}, \dots, a_{ld}, >,\ 1 \leq l \leq t$
Mean and deviation of D can be calculated as:
Mean $mn =< mn_1, mn_2,\ mn_3,\ \dots., mn_d, >$ and Deviation $\sigma =< \sigma_1, \sigma_2, \sigma_3, \dots., \sigma_d >$

$$mn_m = \frac{\sum_{l=1}^{t} a_{lm}}{|D|} \tag{4}$$

$$\sigma_m = \sqrt{\frac{\sum_{l=1}^{t} (a_{lm} - n_{lm})}{|D|}} \tag{5}$$

For $1 \leq m \leq d$, where $|D|$ denotes the size of D i.e. the number of concepts patterns contain in D. The fuzzy similarity of a word pattern
$A =< a_1, a_2, a_3, \dots, a_d\ >$ to cluster D is defined by

$$\alpha_D(A) = \prod_{m=1}^{d} exp\left[-\left(\frac{a_m - mn_m}{\sigma_m}\right)^2\right] \tag{6}$$

Here we can see that $0 \leq \alpha_D(A) \leq 1$. A concepts pattern near to the mean of a cluster is regarded to be very similar to this cluster i.e., $\alpha_D(A) \approx 1$. A concepts pattern far distant from a cluster is not similar to this cluster i.e., $\alpha_D(A) \approx 0$ for example we can consider that M1 is existing cluster which has mean vector $mn_1 =< 0.4, 0.6 >$ and a deviation vector $\sigma_1 =< 0.2, 0.3 >$ now based on this data fuzzy similarity of the concepts type marching A_1 shown in (3) to cluster D_1 becomes

$$\alpha_{D_1}(A_1) = exp\left[-\left(\frac{0.3-0.4}{0.2}\right)^2\right] \times exp\left[-\left(\frac{0.7-0.6}{0.3}\right)^2\right] \tag{7}$$
$$= 0.7788 \times 0.8948 = 0.6969$$

C. Relation Extraction

Extraction of relation is based on part of speech (noun, pronoun, verb, adverb etc.) to find the relationship between these tokens. To extraction the Part of relations lexicon syntactic pattern is used and for finding the IS_A pattern WORDNET is used. Output obtained from this process is used to establish semantic relation between them and their domain, these would be formed subject predicate and object or Resource Description Framework (RDF).

D. Self Adaptive Clustering

In this clustering approach user has not need to any idea about how much cluster is there in advance. If in the beginning no cluster is exist it then it will be created as per requirement. For each concepts pattern is performed and check the similarity of concepts through the pattern whether it is match with existing cluster or a new one is created. If a new cluster is created, all the necessary function regarding with new cluster is initialized. Otherwise if a concepts pattern matched with existing cluster in that case member function of that cluster updated accordingly. Let C be the number of cluster existing currently, the clusters are $D_1, D_2, D_3, \ldots. D_c$ respectively. Each cluster D_l, has mean $mn_l =< mn_{l1}, mn_{l2}, mn_{l3}, \ldots., mn_{ld}, >$ and deviation $\sigma_l =< \sigma_{l1}, \sigma_{l2}, \sigma_{l3}, \ldots., \sigma_{ld} >$.

Here step wise methodology calculation for self-Adaptive Clustering

Step 1: No of concepts pattern matching initially is n
Step 2: No of category d
Step 3: threshold : ρ
Step 4: Initial deviation: σ_0
Step 5: initial no. of cluster k=0

Input:

$V_m =< v_{m1}, v_{m2}, \ldots, v_{md} >, 1 \leq m \leq n$

Output:
Clusters :

$D_1, D_2, D_3, \ldots. D_c$

Procedure self-Adaptive Clustering-Algorithm

```
For each word pattern$V_m$, 1$\leq m \leq n$
$temp\_V = \{D_l | \alpha D_l(A_m \geq \rho, 1 \leq l \leq c)\}$;
If $(temp_V == \emptyset)$

A new cluster$D_q, q = c + 1$, is created by
Else let $D_s \in temp\_V$ be the cluster to which $A_m$ is closed by
Incorporate $A_m$ into $D_s$ by
End if;
```

End for;

Return with the created c clusters;
End procedure

High degree of similarity is found in concepts pattern in a cluster with each other. When a new concept is found, the existing clusters can be adjusted or new cluster can be created. We apply heuristic to find out the order of concepts; we arranged all the patterns, in decreasing order. In this more significant pattern will be fed in first and likely becomes the core of the underlying cluster. , let $A_1 =< 0.2,0.4,0.8 >, A_2 =< 0.2,0.2,0.4 >$, and $A_3 =< 0.7,0.2,0.2 >$ be three concepts patterns. The largest components in these concepts pattern are 0.8, 0.4, and 0.7, respectively. The concepts list is 0.8, 0.7, 0.4 so the order of feeding is A_1, A_3, A_2. This heuristic seems to work very fine. We have to find pattern for every input concept with existing concepts in cluster.

4. RESULT AND ANALYSIS:

The Enron Corpus considered was obtained from the UC Berkeley Enron Email Analysis Project [20] consists of about 680 email mainly focusing on Business Communications eliminating personal messages. The ontology extraction engine and visualization tool is realized using C#.Net built on the Visual Studio 2010 platform. The Semantic E-Mails were pre-processed for Spelling and Grammar Check using Microsoft Office libraries integrated with the application. The English Dictionary available within Microsoft Office package was utilized as a benchmark. The ontology engine used for evaluation is also interfaced to the Outlook Mail Client and the activity semantic details were successfully updated to the Calendar. Notification remainders could also be enabled for user. E-Mails have been clustered based on the pattern extracted and the NLP visualization clearly demonstrates the relation between the concepts extracted. Here Author uses Enron E-mail data set, and try to find out the exact data set based on these detail like Date, time, venue, meeting etc. and find the number of concept in each mail and the performance of proposed methodology is evaluated with existing methodology [15].

In figure 4 processing time of Enron Corpus is shown. The processing time taken for 84, 110, 120 and 133 E-mails are 20739, 27.084, 28.515 and 31.512 seconds respectively. In figure for each mail number of extracted Concept is showing. In figure 5 Enron corpus data set is presented and here it showing that per email number of concept is extracted. For particular mail maximum number of extracted concept is approximately 380. In figure 6 the ontology extraction process is shown by varying fuzzy rule set size.

.

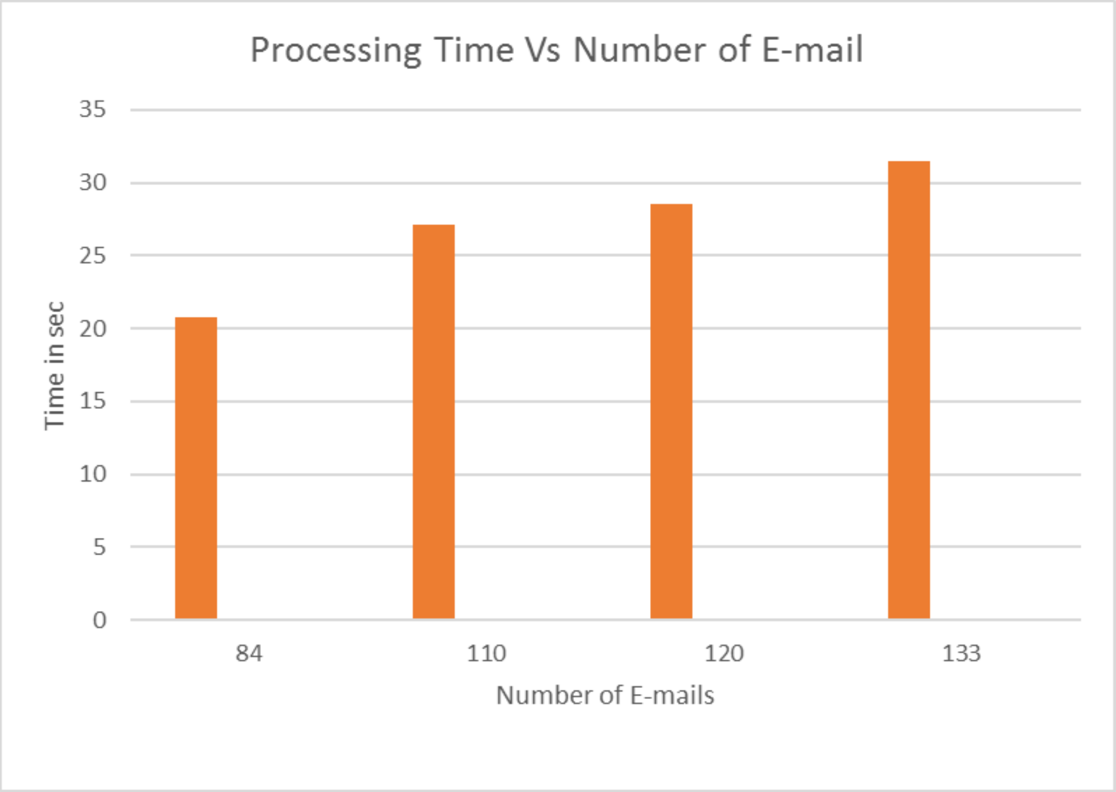

Figure. 2. Processing time of Enron corpus

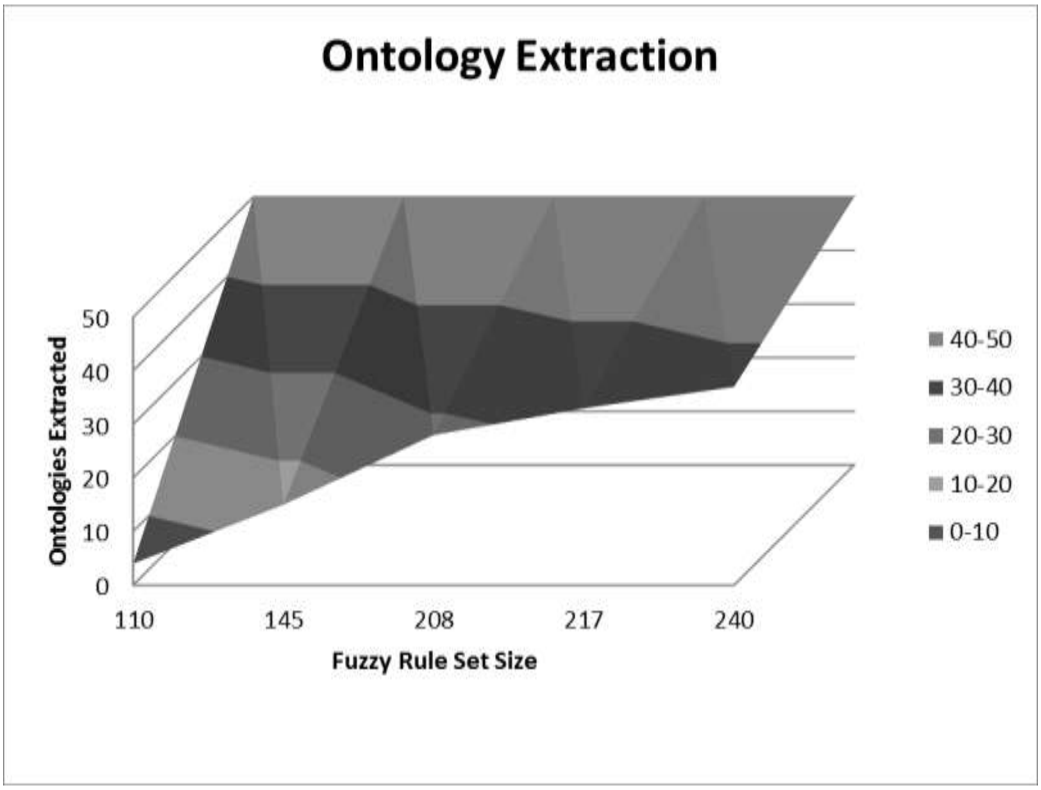

Figure. 5 Fuzzy Rule Set VS Ontologies Extracted

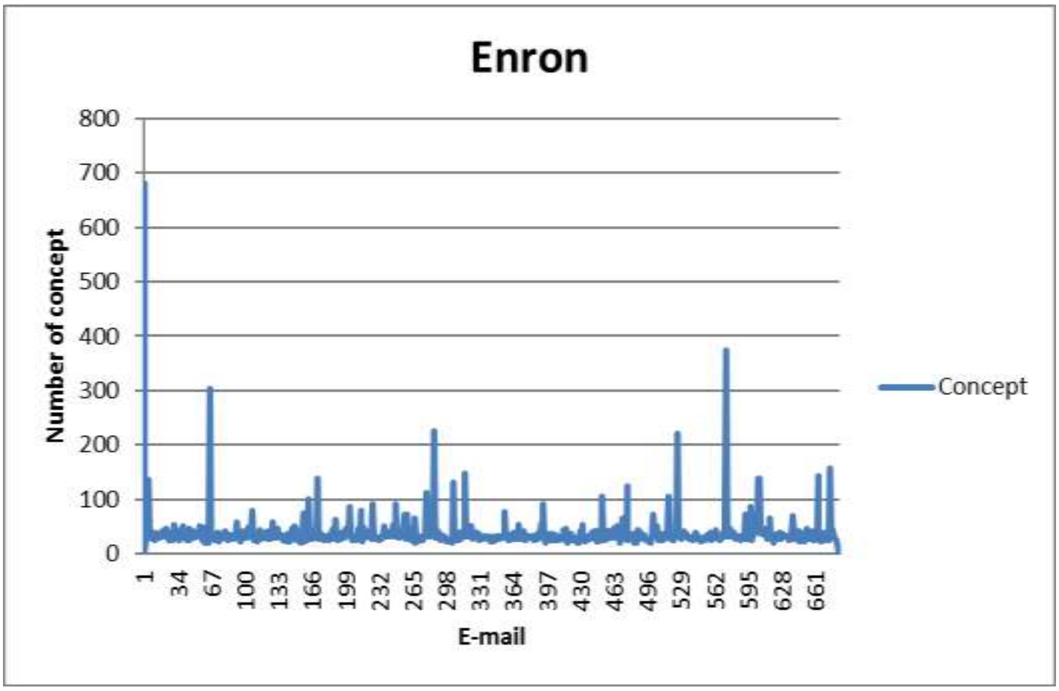

Figure. 6 Number of E-mail and Concept Extracted

CONCLUSION

Handling of E-mail is tuff task for today's scenario, E-mail is a way of communication for personal and business related data. One organization can mail schedule of meeting to every employee at a same time. But due huge amount of mail in the mailbox per day employee may not aware this schedule and not prepare for that. To overcome these problem here author give a new techniques for E-mail classification. E-mail classification is done based on fuzzy self-constructing concept clustering (FFC) algorithm, Concepts which are matched or similar to each other are grouped together and put it into the same cluster. Cluster membership can be defined by the value of mean and deviation. If a concept is not similar to any existing cluster a new cluster is formed for that word. Based on this concept extraction and cluster classification E-mail data fetched and make the schedule from extracted concept. The experiment result shows the efficiency improvement of proposed methodoly considering clssification and fuzzy rule set against ontology creation ovehead is reduced considering the existing methodologies. In future work author used the concept of feature extraction and classification of E-mails for multi lingual ontology extraction which is an area that need to addressed.

.

REFERENCES

[1]. B. Ruttenbur, G. Spickler, and S. Lurie, "eLearning – The Engine of the Knowledge Economy", Morgan Keegan &Co. Inc. eLearning Industry Report, 109 p, 2001.

[2]. Haiyan Kang and Xiaojiao Yuan, "Natural Language Processing Technologies for Multi-Level Intelligent Spam Mail-Filter", International Journal of Machine Learning and Computing, Vol. 4, No. 3, June 2014.

[3]. Taghva, K., Borsack, J., Coombs, J.S., Condit, A., Lumos, S.E., and Nartker, T.A. "Ontology-based Classification of E-Mail", In Proceedings of ITCC, (pp. 194-198), 2003.

[4]. Scerri, S., Giurgiu, I., Davis, B., and Handschuh, S. Semanta "NLP E-Mail in Action". In Proceedings of ESWC, (pp. 883-887), 2009.

[5]. McDowell, L., Etzioni, O. & Halevy A., "NLP E-Mail: Theory and Applications. Journal of Web NLPs" pp. 153-183, 2004.

[6]. Anusha Bagalkotkar, "A Novel Technique for Efficient Text Document Summarization as a Service", Third International Conference on Advances in Computing and Communications, 2013.

[7]. Yugesh Madhavan, "Penetration Testing for Spam Filters", 33rd Annual IEEE International Computer Software and Applications Conference, 2009.

[8]. Laclavik, M. and Maynard, D. "Motivating Intelligent E-Mail in Business: An Investigation into Current Trends for E-Mail Processing and Communication Research" In Proceedings of CEC. (pp. 476-482), 2009.

[9]. Yaohong JIN, "A Topic Detection and Tracking method combining NLP with Suffix Tree Clustering", International Conference on Computer Science and Electronics Engineering, 2012.

[10]. Shazmeen, S.F.; Gyani, J., "A novel approach for clustering e-mail users using pattern matching," in Electronics Computer Technology (ICECT), 2011 3rd International Conference on , vol.6, no., pp.205-209, 8-10 April 2011.

[11]. HEARST, M.A. (1992): Automatic Acquisition of Hyponyms from Large Text Cor- pora. In: Proceedings of the 14th International Conference on Computational Linguistics (COLING), 1992.

[12]. Dotan Di Castro Yahoo Labs, "You've got Mail, and Here is What you Could do With It! Analyzing and Predicting Actions on Email Messages" WSDM'16, San Francisco, 2016.

[13]. Waqas Nawaz, Kifayat-Ullah Khan, Young-Koo Lee, "A Multi-User Perspective for Personalized Email Communities" JournalofExpertSystemswithApplications, arXiv160200479N, 2016.

[14]. www http://bailando.sims.berkeley.edu/enron/enron_with_categories.tar.gz.

[15]. Majdi Beseiso1, Abdul Rahim Ahmad2, Roslan Ismail, "A New Architecture for Email Knowledge Extraction" International Journal of Web & Semantic Technology (IJWesT) Vol.3, No.3, July 2012.