ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 12, June 2016

A Fuzzy based Semantic Ontology Extraction for Email Classification

Suma T
PhD scholar, Department of CSE, JJT University, Rajasthan, India
Dr. Kumara swamy Y S
Professor and Dean, Department of Computer Science &Engineering
Nagarjuna College of engg. & Technology, VTU, Karnataka, India

Abstract*—Today's business world common medium of communication is email. Email is not only used in the professional way it is used in the personal interest as well. So, one of the main concern is, that user can received mails which is relevant to the user and some mail those are not all related to the user. The main thing is we need to do is to identify the important email and then make a cluster of the similar mails. Concept extraction and clustering of concept is done based on fuzzy logic, similar mail pattern is grouped in a same cluster if similarity is less than threshold value a new cluster is defined for that. From the extracted concept author establish the relationship between them and generate the result. Update the important content in the calendar which makes user tasks very easy.

***Keywords*—OWL,Concept, Fuzzy,Ontology.**

I. INTRODUCTION

Natural language processing is an interesting tool for the predication of things happen around you based on learned data. NLP is very useful in different field like text mining, data mining, here we are using concept of NLP in Email classification or enhancement of email visualization. Reading every mail is not possible for today's busy world where each day 30 to 50 mails arrived in inbox. Open each mail one by one and read it out then deletion of that mail is hurdle. Grouping of data set or text based on similarity is called making cluster. Email can also be grouped in a cluster based on their similarity. In an intra cluster similarity between mails is very high while inter-cluster mail has low similarity. Making cluster is a motive to segregate the email based on their similarity like, date, event, time etc.. Content of the email is compared based on a function called similarity function, assigning of weights to the words is done as per frequency of occurrence of words . In machine learning techniques they are used data set and then train it for the statistical model which is used for the test data set and give some result [1]. Semantic web fundamental technologies is the RDF which is called as the Resource Description Framework. Through RDF you can asserting statement for any document no matter document is web or not [2]. For semantic web language OWL is representative of Knowledge, it has predefined concept for creating ontology. In OWL it consist of the basic piece of knowledge [3]. A tool that allows the user to create a single email with a Semi-Structured content might assist the recipients with the act of filtering and prioritizing this message [4]. In the clustering of text, feature selection is an important parameter, it gives the text data properly, which shows that the effect of clustering algorithm [5]. FSM is a techniques in which text is grouped based on frequent maximal substring, it also required a very small space for the storage, but result obtained from this method is shows that it faces difficulties when the keyword are overlapped in the document and it not properly divided the document. Now a days for clustering or classification of document vector space model is used widely it represent each document as a space vector.

Fuzzy logic is very popular to solving the problem which don't have specified boundary condition, because fuzzy logic deal between 0 and 1. In this fuzzy approach automatically classifying of the email document is proposed. For fuzzy logic author used the concept f set theory for information retrieval process [6]. Relationship among the different word inside the email content is derived it means that each word has a unique value that make that word different from other words. In fuzzy set value [0,1] for any class defined that if a word has value 0 then it has no membership in that class. If a word has value 1 then it has full membership for that class. If value varies between 0 and 1 that show that it has marginal membership for that class. Here author propose work is to deal with email classification or updating the calendar based on email data. Ontology is used for the reuses of knowledge or sharing that knowledge, classification or extraction of knowledge for the email content should be done. Some concept which is usually find in the email content, that type of content can also be useful for the validation of the email. Author used concept of fuzzy logic to extract the feature and concept extraction. Word which is frequently appears considered for the features or concept extraction. Relationship establish between those extracted concepts. Distribution of word inside email data is done through the concept vector and these concept vector is processed one by one. Based on similarity index two word can assign in the same cluster, used the mean and deviation concept for defined the membership function. If a new word is arrived which is not similar to any existing cluster in that case a new cluster is defined for that.

Rest of the paper organized as follows in section two proposed model is discussed in which feature extraction, concepts and OWL concept for ontology extraction is discussed. In section 3 Result obtained from our proposed model is discussed and in the last section 4 is conclusion and future work has given.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 12, June 2016

II. PROPOSED WORK

A. Selecting Email-Based Concept

First, Concept formation approach is the extraction of basic terminology used in email which is relevant to the most of the email these basic terminology can be called as candidate. Concept vector is associated with ontology, in term of extracted entity from email corpora.

We have set of email N , and in that set number of e-mail is k like $e_1, e_2, \ldots . e_N$ all together with a concept vector X of s mail $x_1, x_{2,\ldots} x_k$ each email has its own specific property some email may has some property same so from set of mail let t group as $g_1, g_2, \ldots\ g_t$. We make one concepts concept for each email in X. for concepts x_m , its concepts concept Z_m is defined, by

$$Z_m =< Z_{m1}, Z_{m2}, \ldots, Z_{md} > \quad (1)$$
$$=< PL(g_1|x_m), (g_2|x_m), (g_3|x_m), \ldots, (g_d|x_m) >$$

Where,

$$PL(g_l|x_m) = \frac{\sum_{q=1}^{\kappa} e_{qm} \times \alpha_{ql}}{\sum_{q=1}^{k} e_{qm}} \quad (2)$$

For $1 \le l \le d$. Here e_{qm} indicate the number of occurrence of x_m in E-mail e_q, α_{ql} can be defined as

$\alpha_{ql} = 1$ when email e_q, belongs to group g_l, if it not belongs to any group value of $\alpha_{ql} = 0$. from the above equation 2 value obtained for email cluster which is vary between 0 and 1 this fuzzy concept similarity.

A cluster have certain number of email concept and is the product of d one-dimensional Gaussian function. Let C be a cluster containing q email concept $z_1, z_2, z_3, \ldots . z_q$. Let $z_l =< z_{l1}, z_{l2}, z_{l3}, \ldots, z_{lq}, >$, $1 \le l \le q$

B. Concept Extraction

Concept extraction can be expressed in the following form:

$$N'=NS, \quad (3)$$

Where,

$$EN = [e_1\ e_2\ \ldots .\ e_N]^S \quad (4)$$
$$N' = [e'_1\ e'_2\ e'_3\ \ldots e'_k\ ]^S \quad (5)$$
$$S = \begin{bmatrix} s_{11} & \cdots & s_{1c} \\ s_{21} & \cdots & s_{2c} \\ s_{31} & \cdots & s_{3c} \\ . & \cdots & . \\ . & \ldots & . \\ s_{n1} & \cdots & s_{nc} \end{bmatrix} \quad (6)$$
$$e_1 = [e_{m1}\ e_{m2}\ \ldots . e_{mn}] \quad (7)$$
$$e'_m = [e'_{m1}\ e'_{m2}\ e'_{m3}\ \ldots e'_{mc}] \quad (8)$$

For $1 \le m \le k$. Clearly, S is a weighting matrix. The aim of concept reduction is achieves by finding an appropriate *S* such that *c* is smaller than *n* .

- By using clustering algorithm, concepts concept have been grouped into clusters,
- Concepts in the concept vector V are also clustered according to that.
- One concept vector is assign to one cluster, so for different-different cluster we have different concept vectors.
- If we have *c* luster in that case we have *c* extracted concept vector also.
- The element of S are find based on the obtained clusters, and concept extraction will be done.
- Our proposed weighting approaches id hard, soft and mixed.
- In the hard –weighting approach, each word is only allowed to belong to a cluster, and so it only contributes to a new extracted concept. In this case element of **S** are defined as follows:

$$s_{ml} = f(x) = \begin{cases} 1, & if\ l = \arg max_1 \le \beta \le \\ 0, & otherwise. \end{cases} \quad (9)$$

If *l* is not unique in (9), one of them is chosen randomly. In the soft-weighting approach, each word is allowed to contribute to all new extracted concepts, with the degrees depending on the values of the membership functions.

The element of **S** in (8) are defined as follows

$$s_{ml} = \alpha N_l(Z_m). \quad (10)$$

Combination of hard weighting and soft weighting approach give a new approach called mixed-weighting approach. For this case, the element S in (8) are defined as follows:

$$s_{ml} = (\delta\delta) \times s^X_{ml} + (1 - \delta) \times s^R_{ml}, \quad (11)$$

Where s^X_{ml} is obtained by (10) and s^R_{ml} is obtained by (11), and ∂ is a user-defined constant lying between 1 and 0. Note ∂ is not related to the clustering but it concerns the merge of component concept in a cluster into a resulting concept. If threshold value is small, the number of cluster is small, and each cluster covers more type match. In this case, a smaller δ value favor soft-weighting and get higher accuracy. ∂ can vary between 0 and 1, as threshold increases number of cluster also increases

C. Relationship Establishment Between Concept

In Relation extraction in an email document can be done based on grammar rule or Part of speech (POS) like Noun, verb, adverb etc. Establish the relation between these tokens or word by the use of POS like which POS belongs to which Noun or verb or it belongs to adverb etc. Syntactic pattern is used to find the Part_of Relations. In RDF model data subject predicate and object is considered and it would be extracted with semantic relation between them and their domain.

Define Is_A relationship between Noun and it can be found by checking for hypernyms in WordNet For example:

The obtained output hypernymically related synset can be reconstructed by the trail of hypernymically related synsets let take an example:

$\{robin, redbreast\}@ \rightarrow \{bird\}@ \rightarrow$
$\{animal, animate\_being\}@ \rightarrow$
$\{organism, life_{form}, living_{thing}\}@ \rightarrow$

is a transitive, semantic relation that can be considered as IS_A of KIND OF and direction of arrow represent as upward pointing [7]

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 12, June 2016

D. Ontology Analysis

By the help of ontology creation engine. Relationship establish between obtained concept Ontology may include concept and relationship among those concept, concept and obtained relationship merged and make domain ontology. Once ontology is identified it written in OWL format. Ontology must be domain specific for the alignment of extracted concept from the emails.

Author specify the matched threshold ∂ , $\alpha D_i(A_m \geq \partial)$. Assume that n clusters are obtained for the words in the concept vector X. then we find weighting matrix **S** and convert N to N' by the use of SROIQ ontology tool extraction.

Table 1 Ontology Table

Name	Syntax	Semantics
Concepts		
atomic concept	A	$A^I(given)$
nominal	$\{a\}$	$\{a^I\}$
top concept	T	Δ^I
negation	$\neg C$	Δ^I / C^I
conjunction	$C \sqcap D$	$C^I \cap D^I$
existential restriction	$\exists R.C$	$\{x \mid R^I(x, C^I)$
min cardinality	$\geq nS.C$	$\{x \mid \| S^I(x, C$
exists self	$\exists S.Self$	$\{x \mid < x, x >$
Axioms		
Complex role inclusion	$\rho \sqsubseteq R$	$\rho^I \subseteq R^I$
disjoint roles	$Disj(S_1 S_2)$	$S_1^I \cap S_2^I = \emptyset$
concept inclusion	$C \sqsubseteq D$	$C^I \subseteq D^I$
concept assertion	$C(a)$	$a^I \in C^I$
role assertion	$R(a, b)$	$(a, b) \in R^I$

The syntax and semantics of SROIQ is summarized in Table 1. The set of SROIQ concepts is recursively defined using the constructors in the upper part of the table, where $A \in N_c$, C, D are concepts, R, S roles, a an individual, and n a positive integer.

III. RESULT AND ANALYSIS

Here author used ontology extraction engine and visualization tool which is run on visual studio 2010 platform. 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 concept extracted and the NLP visualization clearly demonstrates the relation between the concepts extracted. The BC3 Corpus [8] [9] consists of about 40 threads embodying 261 E-Mails. The BC3 corpus is a part of the W3C corpus.

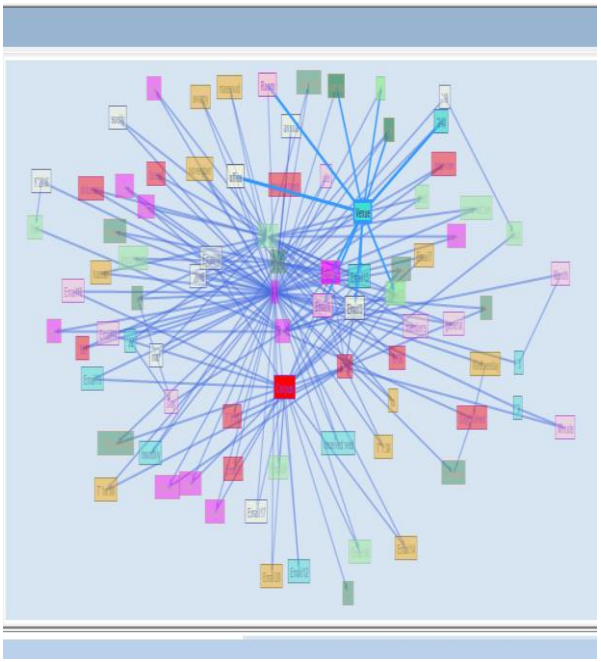

Fig 1 obtained Ontology

In above figure ontology creation is obtained for set of email and relationship is establish.

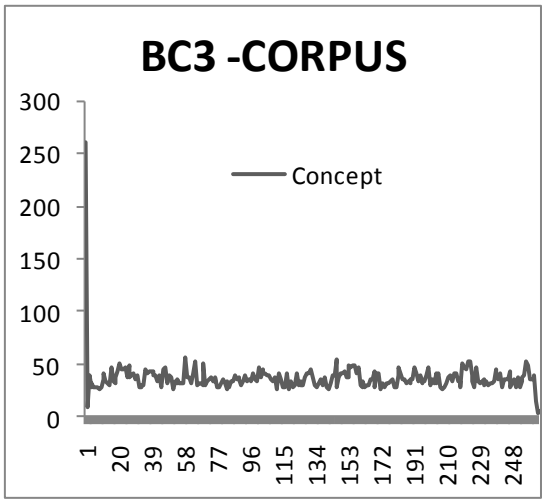

Fig 2 Concept Extracted For Each Mail in BC3 Corpus

Extracted concept in fig 2 shows that for BC3 corpus for 261 emails. For each email number of concept is extracted which is showing in the above fig. Based on this extracted concept Relation is extracted which is given in below fig 3. In figure 4 ontology is extracted based on fuzzy rule set obtained ontology will apply for the clustering mechanism.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 5, Issue 12, June 2016

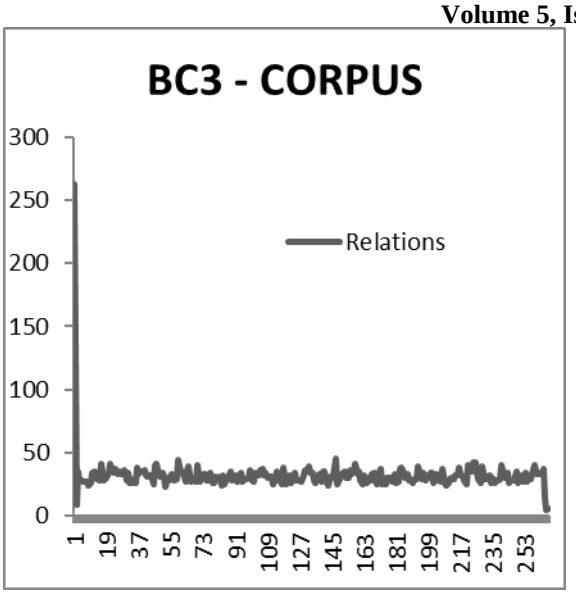

Fig 3 Relation obtained for Each Mail in BC3

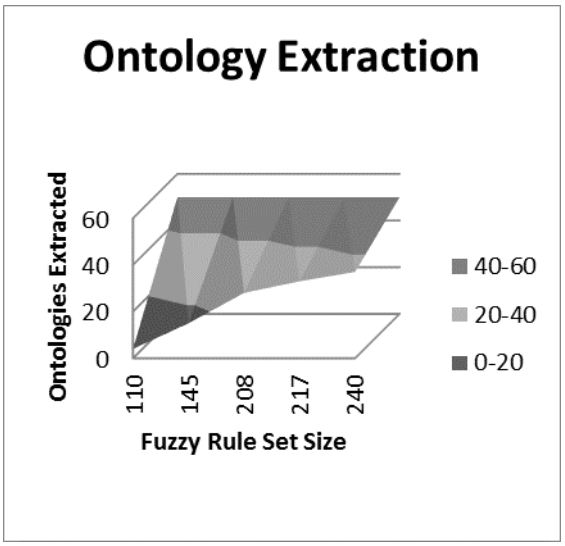

Fig 4 Ontology Extracted

IV. CONCLUSION

Classification and clustering of email based on fuzzy logic which is very useful for the document which not has a sharp edge like yes or no in logic like 0 and 1. Instead of that fuzzy logic worked on variable edge like between 0 and 1. Boundary condition is not sharp it's vary and for that case fuzzy is used for similarity purposes. Fuzzy is used for concept extraction and clustering that email data set. Find similarity and grouped inside a cluster for similar thing if value is not matched for any predefined cluster than make a new cluster and initialized function for that cluster. relationship is establish between that concept and auto update the mail content in calendar for important data

REFERENCES

[1] K. W. Wong, T. Chumwatana and D. Tikk, "Exploring the use of fuzzy signature for text mining," Fuzzy Systems (FUZZ), 2010 IEEE International Conference on, Barcelona, 2010, pp. 1-5.

[2] Klyne, G., Carol, J.: Resource Description Framework (RDF): Concepts and Abstract Syntax. In: McBride, B. (ed.) W3C Recommendation (2004).

[3] P. Mitra, C. Murthy, and S.K. Pal, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. Concept Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, Mar. 2002.

[4] Kalyanpur, A., et al.: SMORE -Semantic Markup, Ontology, and RDF Editor (1998).

[5] Smith, M.K., Welty, C., McGuiness, D.L.: OWL. Web Ontology Language Primer. W3C Recommendation (2004).

[6] L.A. Zadeh, "Fuzzy Sets," in D. Dubois, H. Prade, and R.R. Yager, editors, Readings in Fuzzy Sets for Intelligent Systems, Morgan Kaufmann Publishers, 1993.

[7] Fellbaum, C.: WordNet: An electronic lexical database (1998) WordNet is available from, http://www.cogsci.princeton.edu/wn.

[8] Ulrich J., Murray G., Carenini G., A Publicly Available Annotated Corpus for Supervised Email Summarization AAAI08 EMAIL Workshop, Chicago, USA, 2008.

[9] http://bailando.sims.berkeley.edu/enron/enron with categories.tar.gz.