TELKOMNIKA, Vol.14, No.4, December 2016, pp. 1472~1479
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v14i4.4026

Email Classification Using Adaptive Ontologies Learning

Suma T*[1], Kumara Swamy Y. S[2]
[1]Department of CSE, JJT University, Rajasthan, India
[2]Department of Computer Science & Engineering, Nagarjuna College of Engg. & Technology, VTU, Karnataka, India
*Corresponding author, e-mail: tsumamurthy.cs@gmail.com[1], yskldswamy@yahoo.co.in[2]

Abstract

Email is a way of communication for the today's internet world, private and government sector or public sector all are used email for communication with their clients. They can freely send number of mail to their client without disturbing them. Now a day email communication is also a way of advertising, some mail is also spam, lots of social mails are there. Categorization and handling lots of email is an important task for the researches, as they all are working in this field by using the Natural language processing and ontology extraction process. User get frustrated for handling lots of mails and reading those for finding there is any important mail, sometime user delete lots of mail without reading and in that case may be some important mail which contain the important information may be about meeting, seminar etc. is also deleted. For avoiding these scenarios here auto updation of schedule calendar procedure is proposed by the author. 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. Computation overhead is also calculated for different set of mails and finds that it takes very less time in computing large email data set.

Keywords: *Concept vector, Feature SROIQ, Ontology*

1. Introduction

Uses of internet is a habit for today's generation and it's a basic needs for all organization. All organization uses E-mail for communication with their clients. Other than primary mail company send email for advertisement, for promotional and social message is also there. Number of Email receiving per day is more than 20 on an average so handling these email are now a hurdle for the user. Reading every mail is not possible and in that case some time may be important mail also user not read and left that which is not beneficial for user. As per requirement of email categorization there are various work is ongoing but its not up to the mark, they only categories the mail but after this also we are not eligible for achieving best result like number of email amount is huge. Inbox handling is the important issues. So we need a better email management system for handling huge amount of email in one day. If a people open their email id after 2 or 3 days he see lots of email is their, its human nature he can't open each email one by one in that case some important meeting detail may he miss out. Machine learning technologies used to classify based on their details. Before using it classifier also must be trained by using set of samples, preparing of training sample also needs lots of labor work [1].

Resource Description Framework (RDF) defined the semantic of the resources which is best describe on web language recommended by the World Wide Web Consortium [4]. Semantic web which is carried by the RDF data could be queried using language like SPARQL [5], there is also another technique for this semantic web learning called Web ontology Language (OWL). Ontology is frequently used in the semantic web application [6]. If two document contain different words and no word is common for that also we can establish relation between them if the context of the document is similar by the use of similar vector feature space [2]. By the uses of ontologies email validation can be done, it is very helpful in reusability of knowledge, it can share the knowledge and analyses it. It also helpful in separate the common

Received May 20, 2016; Revised October 28, 2016; Accepted November 12, 2016

functionality from different knowledge area. It extract the knowledge from email contents and defined the concept which is generally find in email documents [3]. Here author proposed work to generate the useful information from email and update that in calendar regarding any event meeting etc. let take an example user received lots of mail and if don't check it but from auto updation techniques calendar is update and user not get worry about reading each mail. The process of achieving this by the using concept of ontology base on fuzzy logic and fuzzy based feature extraction and concept finding is done. Relationship is established between obtained concept and feature, and clustering of received mail is also done by the using fuzzy logic. Based on similarity self-clustering and concept clustering is defined and reduced the effort for the email classification by the reduction in concept. Distribution of word in email set is done by using concept vector and it is processed one by one. If two word are find similar they kept in to the same cluster. By the uses mean and deviation a membership function defined for each cluster. Concept vector is defined for each cluster and if a word is not similar to any existing cluster a new cluster is defined for that word. Here number of concept fetched is not describe in advance.

Rest of the paper organization as follow: section 2 cover the various research work in the field of email classification. In section 3 all the usable concept regarding email classification is discuss, in section 4 proposed method is given and at last conclusion is given.

2. Related Work

Based on this morphological model a system is developed which takes input i.e. an Arabic word. That means the system utilize morphological Arabic Natural Language Processing (ANLP) and translates the Arabic language into English language [21]. Word Sense Disambiguation (WSD) is crucial and its significance is prominent in every application of computational linguistics. WSD is a challenging problem of Natural Language Processing (NLP) [22]. In this work uses Hierarchical clustering algorithm with different similarity measures which are cosine. NIL is powerfull mechanishm which we can uses in various application here we are using NLP for the email data processing. Semantic email was very popular collection of semantic email and managed its task gain focused by the researches. At the time of integration different meaning is used for the notation, email is full of with meta data [7]. Construction of ontologies done for email classification, from the email data set features calculated which is used for training of data, for email classifier they used feature vector and classify email in different category [8]. Logical and decision based theoretic model is defined [9] set of update were given to the email as per certain utilities and constraint. Interference problem is describe here, logical model can be take care based on acceptable email responses. In polynomial time it is possible to generate the optimal message-handling rule for the decision theory. Support Vector Machine (SVM) used by the author for the classification and filtration of email based on text classifier, it is used for finding the unwanted email [10]. Main motive of this approached is to prevent unusual and harmful email spreading in our system and damage it. In this [11] paper author focused the for most usable task performed with email read, reply, delete, and delete without read. By used the learning concept prediction of these four task, for that horizontal and vertical learning is used. Clustering techniques is used for grouping of similar kinds of emails, email have similar kind of attributes here by using text mining similar type of email find and clustering of email is done [12]. Using speech act theory [13] categorization of email done by the used of sender intension in email system. Fuzzy based spam filtering in email which not require training of data set in advance, it used the clustering mechanism based on fuzzy logic [14]. Here [19] author gives the idea about latent semantic analysis and it is for finding the similar words in the poetry and analyzing the emotional relatedness. In [20] based on semantic language author proposed cooccurence search engine for Chinese-tibetan and monitor that bases on semantic language but they are not given detail idea about semantic annotation.

3. Proposed Work for Email Clustering

3.1. Concept Vector

Sensor Concept vector is associated with ontology, in term of extracted entity from email corpora. A concept vector is a linker between entity in an email content or document. Consideration of both semantic and lexical analysis is done in the formation of concept vector, terminologies used in email corpora is related to the concept. 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.

3.2. Concept Clustering Algorithm

In 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 pattern for each email in X. for concepts x_m , its concepts pattern Z_m is defined, by:

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

Where,

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

For $1 \leq l \leq d$. 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 $\alpha_{ql} = 0$.

Here we have concepts pattern, take an example that we have six emails $e_1, e_2, e_{3,}\ e_{4,}\ e_{5,} e_{6,}$ belonging to group $g_1, g_{1,} g_1 \ and \ g_2, g_{2,} g_2$ respectively it can be simplify as $e_1, e_2, e_{3,}$ belongs to g_1 and $e_{4,}\ e_{5,} e_{6,}$ belongs to g_2

Now occurrences of x_1 in these emails be 1, 2, 3, 4, 5 and 6 respectively. Email pattern of z_1 of x_1 can be calculated by the using equation number 2.

$$\boldsymbol{Z_1} =< 0.3,\ 0.7 > \quad (3)$$

Our motive is the making cluster, based on these email pattern. Combine the email in X into cluster based on these email patterns.

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

3.3. Adaptive-clustering

Let suppose here user don't have any idea about existing number of cluster, let suppose no cluster is exists at the beginning and cluster can be created as per needs. If a new cluster is defined a detail containing about that cluster a function is also generated. While if an email pattern is combined with existing cluster it detail containing function updated as per new entry.

Initial Condition:

1. n number of cluster
2. $C_1, C_{2,\ldots . . C_n}$ be the number of cluster
3. $mean\ value\ for\ cluster\ C_i\ is\ \alpha_i = \alpha_{i1,} \alpha_{i2,\ldots . ,} \alpha_{in}$
4. $deviation\ \beta_i = \beta_{i1}, \beta_{i2,\ldots . ,} \beta_{in}$
5. $Z_m =< Z_{m1}, Z_{m2}, \ldots . , Z_{md} >$
6. Matching or similarity between $Z_m\ to\ the\ existing\ cluster$

Similarity calculation $\delta C_i(Z_m) = \prod_{q=1}^{d} exp\left[-\left(\frac{z_{mq}-\alpha_{iq}}{\beta_{iq}}\right)^2\right]$

We have defined some threshold value ∂ between 0 and 1 if we want to create larger cluster threshold value set to small value. Z_m pass the similarity test if its value is greater than set threshold value ∂. If value of ∂ increases similarity is less and in that case number of cluster is more. If similarity test is not passes by Z_m foration of new cluster is done for that.

Newly formed cluster has only one member mail pattern C_v is the newly formed cluster $v = n + 1$ if Z_m passed the similarity test then new cluster is not formed and find that like its most similar to the cluster C_2. Pattern of email sorted in a decreasing order, calculation for similarity is done and compare it with all the existing cluster.

Concept Extraction

Pattern extraction can be expressed in the following form:

N'=NS (4)

Where,

$$\boldsymbol{EN} = [\boldsymbol{e_1}\, \boldsymbol{e_2} \ldots . \boldsymbol{e_N}]^{\boldsymbol{S}} \quad (5)$$

$$\boldsymbol{N}' = \left[\boldsymbol{e_1'}\, \boldsymbol{e_2'}\, \boldsymbol{e_3'}\, \ldots \boldsymbol{e_k'}\right]^{\boldsymbol{S}} \quad (6)$$

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{s_{11}} & \cdots & \boldsymbol{s_{1c}} \\ \boldsymbol{s_{21}} & \cdots & \boldsymbol{s_{2c}} \\ \boldsymbol{s_{31}} & \cdots & \boldsymbol{s_{3c}} \\ . & \cdots & . \\ . & \cdots & . \\ \boldsymbol{s_{n1}} & \cdots & \boldsymbol{s_{nc}} \end{bmatrix} \quad (7)$$

With,

$$\boldsymbol{e_1} = [\boldsymbol{e_{m1}}\, \boldsymbol{e_{m2}} \ldots . \boldsymbol{e_{mn}}] \quad (8)$$

$$\boldsymbol{e_m'} = \left[\boldsymbol{e_{m1}'}\, \boldsymbol{e_{m2}'}\, \boldsymbol{e_{m3}'}\, \ldots \boldsymbol{e_{mc}'}\right] \quad (9)$$

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

By using clustering algorithm, concepts pattern have been grouped into clusters, and concepts in the pattern vector V are also clustered according to that. One pattern vector is assign to one cluster, so for different-different cluster we have different pattern vectors. If we have *c* luster in that case we have *c* extracted pattern vector also. The element of S are find based on the obtained clusters, and pattern 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 contribute to a new extracted pattern. In this case element of **S** are defined as follows:

$$\boldsymbol{s_{ml}} = \boldsymbol{f}(\boldsymbol{x}) = \begin{cases} \mathbf{1},\ if\ l = \mathbf{arg}\,\boldsymbol{max_1} \leq \boldsymbol{\beta} \leq \boldsymbol{c}\ \left(\boldsymbol{\alpha N_\beta}(\boldsymbol{Z_m})\right), \\ \mathbf{0},\ otherwise. \end{cases} \quad (10)$$

If *l* is not unique in (10), one of them is chosen randomly. In the soft-weighting approach, each word is allowed to contribute to all new extracted patterns, with the degrees depending on the values of the membership functions.

The element of **S** in (4) are defined as follows:

$$\boldsymbol{s_{ml}} = \boldsymbol{\alpha N_l}(\boldsymbol{Z_m}) \quad (11)$$

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:

$$\boldsymbol{s_{ml}} = (\boldsymbol{\delta\delta}) \times \boldsymbol{s_{ml}^X} + (\mathbf{1} - \boldsymbol{\delta}) \times \boldsymbol{s_{ml}^R} \quad (12)$$

Where s_{ml}^X is obtained by (10) and s_{ml}^R is obtained by (11), and $\delta$is a user-defined constant lying between 1 and 0. Note δ is not related to the clustering but it concerns the merge of component

pattern in a cluster into a resulting pattern. 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. δ can vary between 0 and 1, as threshold increases number of cluster also increases.

3.4. Extraction of Relation

Let 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}@→{bird}@→{animal,animate_being}@→{organism,life_form,living_thing}@→ is a transitive , semantic relation that can be considered as IS_A of KIND OF and direction of arrow represent as upward pointing [18].

3.5. Email Classification based on ontology

N is set of training emails, E-mail classification can be done as:

Author specify the matched threshold δ, $\alpha D_l(A_m \geq \delta)$ 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.

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∈N_c, C, D are concepts, R, S roles, a an individual, and n a positive integer.

Table 1 Terminology for ontology tool

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

4. Simultion Result And Analysis

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 mails. The BC3 Corpus [16, 17] consists of about 40 threads embodying 261 E-Mails. The BC3 corpus is a part of the W3C corpus, as shown in Figures 1-6.

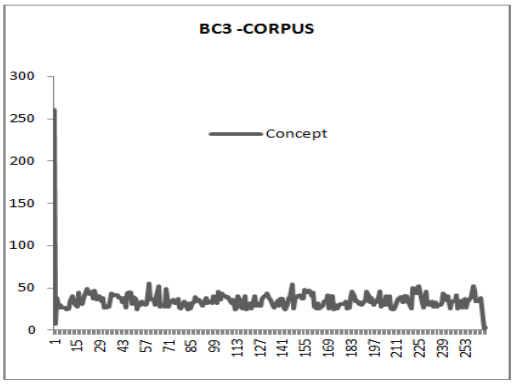

Figure 1. Concept Extracted for Each Mail in BC3 Corpus

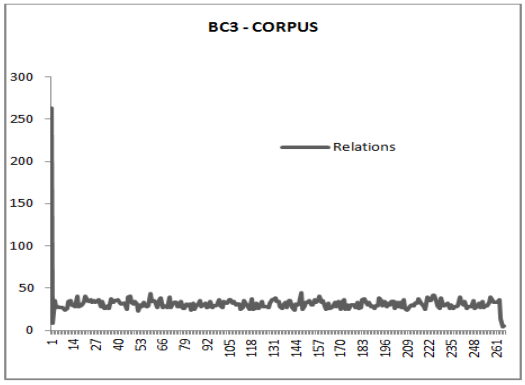

Figure 2. Relation obtained for Each Mail in BC3

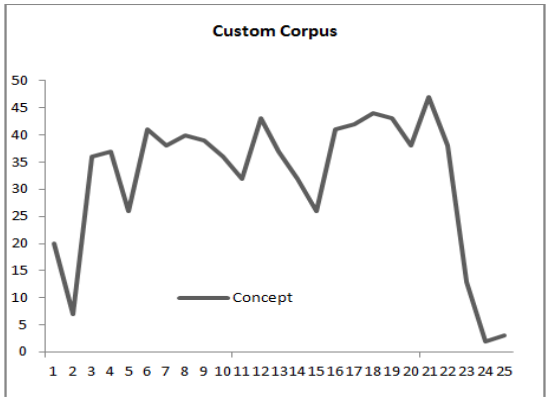

Figure 3. Concept Extracted for Each Mail in Custom Corpus

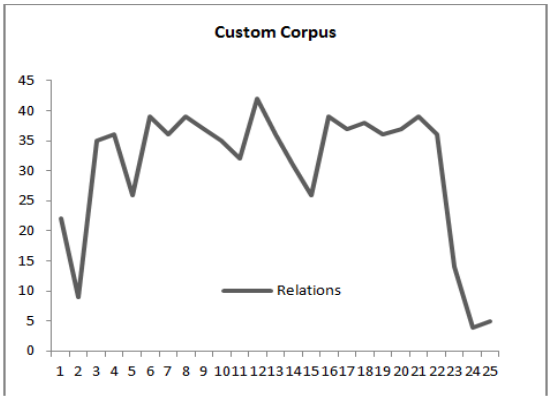

Figure 4. Relation Extracted for Mail in Custom Corpus

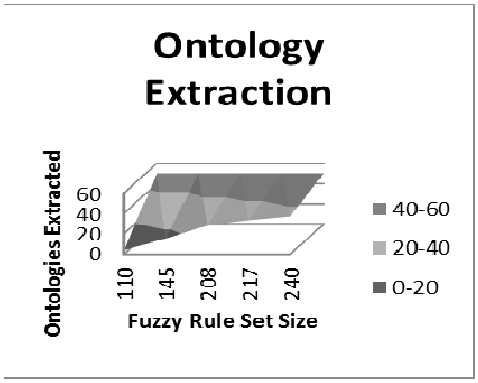

Figure 5. Ontology Extracted

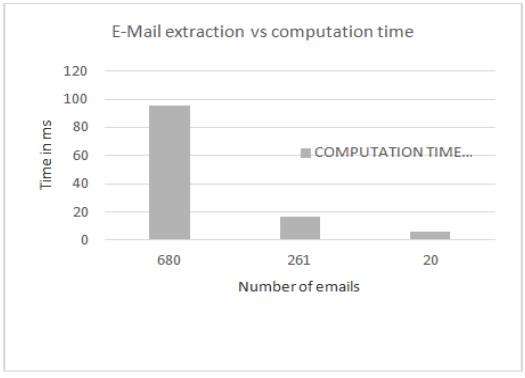

Figure 6. Computation Time For Number

Ontology representation for BC3 corpus is shown in Figure 7 where relationship establish between obtained concept.Ontology extraction for custom dataset is visualized which consist of 20 emails. The ontology extraction for email 2 is visualized in Figure 8 which consist of concept such as people, area. Relationship among concept from various emails and based on this ontology extraction the calendar is updated for every incoming and outgoing email.

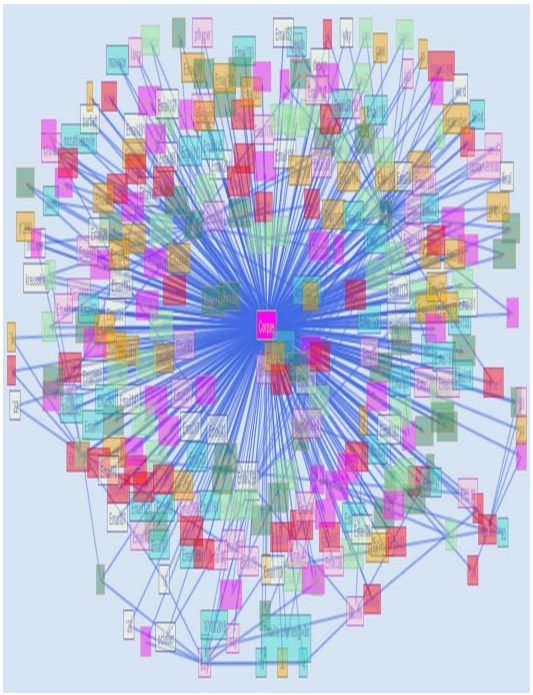

Figure 7. Schematic representation of ontology extraction for BC3 corpus dataset

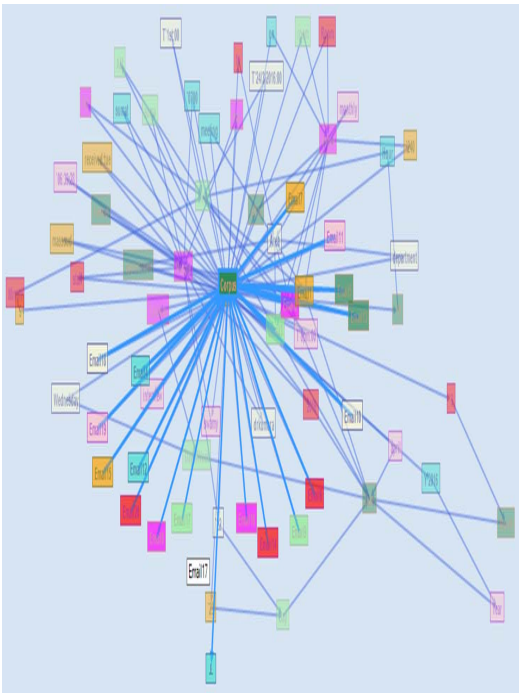

Figure 8. Schematic representation of ontology extraction for Custum Corpus corpus dataset

5. Conclusion

Email is primary source of communication for the business organization, we are dependent on it for our communication to the higher authority. Organization can send various types of email for transferring some information calling for meeting, some organization send promotional email, some are for advertising their products. So handle of all kinds of email and read each email is a hurdle for user and may escape an important mail in that case. To avoid these case and better management of email, here based on concept extraction clustering is done for that if a new mail is arrives is similarity for that mail pattern is matched with existing cluster, if similarity value is more than or equal to the threshold value it accepted for that cluster if similarity value is not matched with any cluster value a new cluster is defined and initialize some member function for that. Extracted concepts relationship is establish and generate the auto updation is done with calendar. We are also calculating computation over head for the different number of mails.

References

[1] Y Hu, C Guo, X Zhang, Z Guo, J Zhang, X He. *An Intelligent Spam Filtering System Based on Fuzzy Clustering*. Fuzzy Systems and Knowledge Discovery, 2009, FSKD '09. Sixth International Conference on. Tianjin. 2009: 515-519.

[2] Kolcz A, Chowdhury A, Alspector J. *The impact of feature selection on signatu detection*. CEAS. 2004.

[3] Kazem Taghva, Julie Borsack, Jeffrey S. Coombs, Allen Condit, Steven Lumos, Thomas A.Nartker. *Ontology-based Classification of Email*. ITCC, IEEE Computer 2003 Society. 2003: 194-198.

[4] Lassila O, Swick R. *Resource Description Framework (RDF) Model and Syntax Specification.* W3C Recommendation. 2004.
[5] Prud'hommeaux E, Seaborne A. *SPARQL Query Language for RDF*. W3C Candidate Recommendation. 2008.
[6] Gruber R. Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*. 1995; 43: 907-928.
[7] L Mcdowell, O Etzioni, A Halevy, H Levy. *Semantic email.* In WWW '04: Proceedings of the 13th international conference on World Wide Web. New York, NY, USA. 2004: 244-254.
[8] K Taghva, J Borsack, J Coombs, A Condit, S Lumos, T Nartker, *Ontology-based classification of email*. Information Technology: Coding and Computing [Computers and Communications], 2003. Proceedings. ITCC 2003. International Conference on. 2003: 194-198.
[9] Luke McDowell, Oren Etzioni, Alon Halevy. *Semantic Email: Theory and Applications*. Department of Computer Science & Engineering, University of Washington. Seattle, WA. 2004.
[10] Qing Yang, Fang-Min Li. *Support vector machine for customized email filtering based on improving latent semantic indexing*. Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on. Guangzhou, China. 2005; 6: 3787-3791.
[11] Dotan Di Castro, Zohar Karnin. *You've got Mail, and Here is What you Could do With It!nAnalyzing and Predicting Actions on Email Messages*. WSDM'16. San Francisco, CA, USA. 2016.
[12] SF Shazmeen, J Gyani. *A novel approach for clustering e-mail users using pattern matchin*g. Electronics Computer Technology (ICECT), 2011 3rd International Conference on. Kanyakumari. 2011: 205-209.
[13] VR Carvalho, WW Cohen. *On the collective classification of email "speech acts".* Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. Salvador, Brazil, 2005.
[14] Y Hu, C Guo, X Zhang, Z Guo, J Zhang, X He. *An Intelligent Spam Filtering System Based on Fuzzy Clustering*. Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference on. Tianjin. 2009: 515-519.
[15] Horrocks I, Kutz O, Sattler U. *The even more irresistible SROIQ*. 2006.
[16] Ulrich J, Murray G, Carenini G. *A Publicly Available Annotated Corpus for Supervised Email Summarization.* AAAI08 EMAIL Workshop. Chicago, USA. 2008.
[17] http://bailando.sims.berkeley.edu/enron/enron_with_categories.tar.gz
[18] Fellbaum C. WordNet: An electronic lexical database. 1998
[19] Wujian Yang, Lianyue Lin. Process Improvement of LSA for Semantic Relatedness Computing. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2014; 12(4): 1045-1052.
[20] Lirong Qiu. Website Resource Monitoring Platform Supporting Tibetan and Uyghur Language Based on Semantics. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2013; 11(8): 4766-4773.
[21] Abdulrahman Ahmed Alzand, R Ibrahim. *Diacritics of Arabic Natural Language Processing (ANLP) and its quality assessment*. Industrial Engineering and Operations Management (IEOM), 2015 International Conference on. Dubai. 2015: 1-5.
[22] N Patel, B Patel, R Parikh, B Bhatt. *Hierarchical clustering technique for word sense disambiguation using Hindi WordNet*. 2015 5th Nirma University International Conference on Engineering (NUiCONE). Ahmedabad. 2015: 1-5.