Journal of Theoretical and Applied Information Technology
30th June 2017. Vol.95. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

ENSEMBLE MULTI-LABEL TEXT CATEGORIZATION BASED ON PYRAMIDAL CLUSTER MEMBERSHIP APPROACH

[1]J. STALIN JOSE, [2]DR. P. SURESH
[1]Research Scholar, Bharathiar University, Department of Computer Science, Coimbatore, TamilNadu.
[2] Head of the Department, Department of Computer Science, Salem Sowdeswari College, Salem, TamilNadu.
E-Mail: [1] jstalinjose1505@gmail.com, [2] sur_bhoo71@gmail.com

ABSTRACT

Text Categorization is an interesting field in the study of Textual Data Mining. It has attracted an increasing popularity with its explosive growth of textual documents. The documents are connected with exclusive multitude categories i.e sports, medical, health, and Olympic Games). Text categorization paves different opportunities for creating multi-label learning approaches that specifically to textual data. Text mining defines the processes of discovering useful knowledge patterns from textual data. This is one of the factors followed in automated text categorization. It is practiced by developing novel machine learning approaches. Anyhow, the ML model generates low expressivity. The ML model established using Train-Test scenario. In case the existing model is found deficient, the Train-Test-Retrain is developed which is time consuming process. In this paper, we proposed "Pyramidal Cluster Membership Approach (PCMO)". It works in two models namely, training and testing model. The training model comprised of four phases, Pyramid-Fuzzy Transmutation, Novel k-edge classifier, Cluster to Category mapping and finding the boundaries. These estimated boundaries are applied on new textual data and the categories are assigned. Experimental results on Freebase dataset show that the proposed approach based on pyramidal membership method can achieve better classification accuracy than the traditional approaches especially that includes over-fitting document categories.

Keywords: *Textual Data Mining, Text Categorization, Membership Functions, Pyramid Structures And Machine Learning Models.*

1. INTRODUCTION

Text mining is one of the essences of data mining techniques. It belongs to the class of Information Retrieval (IR) systems [1]. In the modern world, the information should derive automatically within a stipulated time. Text mining is also known as Intelligent Text Analysis, Text Data Mining or Knowledge-Discovery in Text (KDT). An enormous amount of textual data is generated from real-life applications such as Spam Filtering, Automatic Labeling and Indexing, Document organization, Text Filtering, Word Sense Disambiguation and Hierarchical Web page categorization. According to statistics on Textual data, over 80% of the data are stored as text [2], thus, text mining is trusted to have high commercial potential value. Knowledge- Discovery in Text (KDT) is used to straighten out the trivial-pattern and semantic relationships between concepts using Natural Language Processing (NLP) [6].

Textual data is stored in three labels, namely, structured textual data, Non-Structured textual data and Quasi- structured data. In recent Research & Developments, structural data plays a vital role in Information Retrieval systems [3 - 5]. The universal issue in text mining is obvious: human percepts easily about the text and its patterns whereas computers are unable to process automatically unless it is trained [7]. The pictorial representation of text mining is given as:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

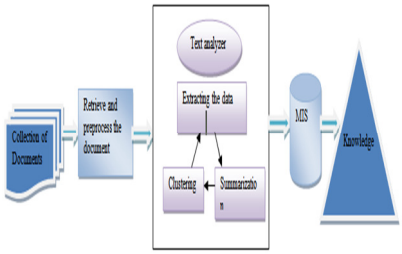

Figure 1: Workflow Of Text Mining

The proposed study is on multi-label learning process in text categorization systems. We have proposed "Pyramidal Cluster Membership Approach (PCMO)". It works in two models namely, training and testing model. The training model comprised of four phases, Pyramid-Fuzzy Transmutation, Novel k-edge classifier, Cluster to Category mapping and finding the boundaries. These estimated boundaries are applied on new textual data and the categories are assigned in testing model.

The paper is structured as follows: Section I depicts the advent of textual data and categorization in Knowledge-Discovery Text (KDT), Section II depicts the various researchers researches on efficient text categorization techniques, Section III proposes a novel algorithm to solve the issue of text categorization, Section IV justifies the proposed algorithm using various performance criteria and concluded in Section V.

2. RELATED WORK

As text mining fall under the category of data mining, the modeling approach used for categorizing the text were studied as follows:

Decision tree strategies [8] remake the manual order of the training documents in the type of a tree structure. In the tree structure, node denotes the queries and a leaf represents the particular category of dataset. The demerit in decision tree strategy is 'over-fitting'. It is simple and convenient to use. Bayesian approaches are subdivided into Naïve and Non-Naïve Bayesian approaches. The naïve approaches operate in two models namely, multivariate and multinomial models. Both the models work on terms distribution in the document [10]. An N-gram is a ceaseless [11] sequence of n characters of long portion of a content. The most successive N-grams are kept. This procedure produces an imperative number of parts contrasted with the content analysis by taking into account the punctuation and word separators, however, it is extremely tolerant to the spellings errors and it remains immaculate to all successions frames on a discrete letter (Chinese dialect, DNA arrangements…). Its achievement in dialect's recognition is headed to its execution in content categorization [12].

Basically, there are two types of vector-based methods namely, Centroid algorithm and Support Vector Machines [13]. Centroid algorithm is the simplest algorithm. In the learning model, a vector space for centroid is estimated. The centroid algorithm is efficiently used in small databases [14]. Another system, Support Vector Machines is calculated, to classify the new documents. Chaitrali S. Dangare [16] have investigated in heart disease prediction systems that incorporates a wide use of attributes to detect the similarity scores of patient having heart disease or not. D. Lavanya et al [17] suggested a hybrid approach to predict the breast cancer disease. They utilized the features of CART that includes selection and bagging techniques to measure the performance. They detected the disease at an untimely phase with accurate results. A top-down approach in the insights of decision tree algorithm was surveyed by Lior Rokach [18]. They depicted a novel approach for splitting criteria and pruning techniques.

An enhanced prediction approach was studied in [19] educational databases to categorize the career options for end users. Based on the student's behavior, the performance chart was evaluated and suggested them an improvement way for academic with use of Rapid Miner, a data mining tool. The classification system was further tested in blood donar systems [20]. They used the advent of CART decision tree algorithm and implemented using WEKA tool. The accuracy of the system was good. K. Sudhakar et al [21] utilized the swarm based classification systems. Genetic algorithm is merged with associative classifier to categorize the best attributes for predicting the heart disease. Along with CART system, ID3 and C4.5 was studied to analyze the performance of the student's data [12]. This was further extended to the study of CHAID decision tree algorithms [13]. Smart drill [14] studied about the working of CHAID decision tree algorithm. Then, the suitability of the text categorization using classification and regression tree methods was studied in [15]. A financial based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

operational risk indicator was studied in [22] to find the similarity score of document categorization.

kNN is a lazy [9] learning algorithm. It restricts from being connected to regions where dynamic characterization is required for large databases. It is also a case-based learning method. It classifies the data points using Euclidean space distance function. Jaydeep Jalindar Patil and Nagaraju Bogiri [23] provided an automatic text categorization system using Marathi documents. The classification was done based on user's profile and their browsing history. A Label Induction Grouping (LINGO) was designed based on browsing history of users. The results shown those 200 documents were effectively categorized under 20 labelled documents. A bangla online text documents [24] was studied on four supervised learning algorithms namely, Decision Tree (C 4.5), k-NN, Naïve Bayes and Support Vector Machines. Atlast, they exhibited a better outcomes that k-NN and NB were efficient than the SVM and C4.5. The k-NN used smaller time consumption for training classifier. A knowledge based rule was generated for classifying the Hindi documents in syntactic view was presented by Neha Dixit and Narayan Choudary [25]. They created a largest lexical resource for Hindi verbs with good relevancy score. The text categorization was also performed in Tamil Documents by Aruna Devi and Saveetha R [26]. An improved C-feature extraction, a predefined category was used to classify the document in pairwise approach [27].

With an effective combination of naïve bayes and centroid techniques were used for Punjabi text classification systems [28]. It also concentrated on the ontology based systems which are used for categorizing the documents on the extracted features. Several researchers expected that the hybrid approach shows higher accurate rate, unfortunately, it depicted the lower performance. An enhanced k-NN approach was used in English and Telugu text categorization. A top-down approach [29] was utilized to classify the documents. The results exhibited a lower suitability for the testing documents. A study on neural networks for the classifying documents is conducted by [30]. A wide variety of languages was used for text classification purpose, only Tamil text documents leads to 93.33% accuracy. A statistical oriented approach was investigated by [31] using Naïve bayes and Support Vector machines to classify the Urdu documents. A reduced feature lexicon was created using various preprocessing techniques. Yet, the categorization yielded high time consumption. N-gram based classifier is used in the context of bangla text classification [32]. A new corpus was designed that does not have high compatibility rate.

3. RESEARCH METHODOLOGY PYRAMIDAL CLUSTER MEMBERSHIP APPROACH (PCMO)

3.1 Challenges

In real-world dynamic scenarios, a huge number of symbols are regularly needed to encode the data values. This makes us to confront the issue, known as "Curse of Cardinality & Curse of Dimensionality", which can engross us to design an optimal pattern of text categorization methods in terms of computational efficiency and in practical use. Definitely, given the vast data scale and the heterogeneous nature of the textual data, a new visions and strategies are required to face the challenges. Parallel Coordinates (PC) is considered as the popular visualization technique for multivariate data analysis. In this research, PC is widely used to visualize geometry that represents data under multiple domains. From the previous studies, the principle challenges existed were listed as follows:

- ✓ *Curse of dimensionality*- Adding noise to the datasets, includes a higher preprocessing techniques.
- ✓ *Curse of Cardinality*- To classify the documents or domains, a higher number of attributes are selected.
- ✓ *Data Redundancy*- This includes higher memory consumption and poor performance due to weak system designs.

3.2 Problem Formulation

In the perspective of multilabel text classification problem, we are provided with triplet {D, T, and C}, where $D=\{(d^{(1)},y^{(1)}),(d^{(2)},y^{(2)},.....(d^{(n)},y^{(n)}\}$ an order of n training patterns is, $T=\{t_1,t_2,....t_m\}$ is a set of m features, and $C=\{c_1,c_2,....c_p\}$ is an order of p categories. The attributes in T are obtained after

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

the preprocessing of D orders. The k^{th} training pattern consists of two components document d^k and the label y^k. Each document inheres to one or more than one category. Then y^k is a vector with p categories and formulated as

$$y_j^k = \{1, d^i \; inherestocategoryc_j\}$$

$$y_j^k = \{0, d^i \; notinherestocategoryc_j\}$$

In multilabel classification, several components may contain 1 in y^k. In the event of categorizing the new documents, the arrangements of the new event also inheres the triplet {D, T, C}. If the value returned for D^i is 1, then it belongs to category c_j. The pyramid oriented fuzzy proposed algorithm works in two phases namely, a) Training phase and b) Testing phase.

1) Training phase:

In training phase, the categorization model is developed from the group of training data. The high-dimensional and high-cardinal training documents are transmuted to low-dimensional and low-cardinal pyramid-fuzzy vectors. The training phase is splitted into four processes:

- ✓ Pyramid-Fuzzy Transmutation
- ✓ Novel k-edge classifier
- ✓ Cluster to category mapping
- ✓ Finding the boundaries.

a) Pyramid- Fuzzy Transmutation:

In the geometry view, pyramid is the 3D geometric shape formed by connecting all corners of the polygon to a central apex. We made an attempt to incorporate the fusion of Triangular Pyramid and Fuzzy oriented geometry class to find the topology of the given data. The three main components of triangular pyramid were the Triangle's formation (Highest order attributes-Subject, Object and Predicate), Lateral faces (Least priority attributes) and Apex (Combination of high and least priority order). Generally, m is large, which is the number of attributes in T. This might results in sparse apportioning of the training documents in high dimensional space. The document with m dimensions is transmutated to a pyramid-fuzzy geometric structure with p dimensions, where p is the number of categories in C. Normally, p should be smaller than m. Henceforth, cardinal-dimensional reduction is achieved. Firstly, the pyramid- fuzzy membership function is expressed as: $F_P : T * C \rightarrow [0,1]$. It is processed into two steps:

a) Finding the pyramid-based similarity relation
b) Transforming the input (pyramid-based similarity relation) to fuzzy based similarity relation.

I. Finding the pyramid based similarity relation:

Each data point in a pyramid represents the region. Each individual region will maintain a list of attributes. The attributes are the Creator, Creator_ time, Deletor, Deletor_time, Subject, Predicate, Object and Language Code. Let us assume, Subject as base of the pyramid, Object as face of the pyramid and Predicate as apex of the pyramid. Data Intensity is the study of discovering the degree of the data and its data points in the field of text categorization. Let us find the Intensity (Volume) of the data, is expressed as follows:

Data Intensity (DI) =

$$\frac{1}{6} * Sub * Obj * \Pr ed \quad (1)$$

Where Sub – Subject of the data; Obj- Object of the data and Pred- Predicate of the data.

II. Transforming Data Intensity into fuzzy based membership function:

From the step (a), we discover the topology of the data and its data points. Based on DI, the degree of membership function is originated. The fuzzy set F in a universe of discourse S is defined as the pairs:

$$F = \{\mu_F(s) / s : s \in S\} \quad (2)$$

Where $\mu_F(s) : S \rightarrow [0, 1]$ is called as the membership function of the fuzzy set F and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$\mu_F(s)$ defines the degree of membership value. Generally, the membership function build based on the system data design and decision of the preferred shape. Though, there are many shapes of membership functions, the triangular based membership function is widely adopted to meet our requirement. The triangular based membership function is given as triplet (Sub s, Obj o and Pred p).

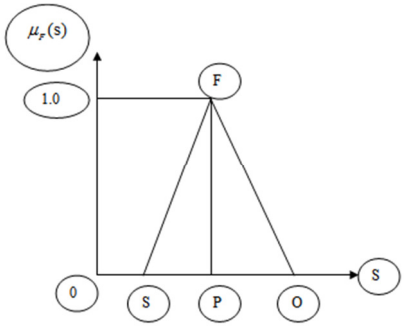

Figure 2: Triangular membership functions

From the Fig.1 the triangular function of a vector S depends on the three scalar parameters such as Sub s, Obj o and Pred p. Then the triangular-fuzzy based membership function is given as:

$$\mu_F(s) = \begin{cases} \frac{s-a}{b-a}, & a \le s \le b \\ \frac{c-s}{c-b}, & b \le s \le c \\ 0 & otherwise \end{cases} \quad (3)$$

B) Novel k-edge classifier:

Novel k-edge classifier is the second step in the training phase. The target of this proposed classifier is to select the features to build a cluster for the given data. Given n fuzzy membership vectors, the efficient edge classifier algorithm is executed to group these vectors into clusters. The features selected are the Subject, Object and Predicate. A predicate feature is the most important one among all others features because it takes the Subject features linkage to the predicate features. So, a predicate feature is used for forming the clusters. In graphical view, the relationship between subject and predicate feature is represented as follows:

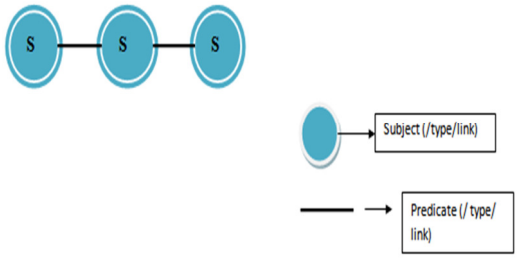

Figure 3: Nodes and Edges Relationship

An efficient classifier known as k- edge based classifier is proposed. The aim of this novel classifier is to partition the data points into a small number of clusters. The proposed classifier is depicted as follows:

a. Cluster the data into k groups.
b. Select k points at random as cluster centers.
c. Assign objects to their closest cluster center according to the Betweeness function.
d. Calculate the centrality of all objects in each cluster.
e. Repeat steps b, c and d until the same points are assigned to each cluster in consecutive rounds.

Betweenness function is given as:

$$B_i^j = \sum_{i=1}^{n} \frac{\sigma_{v_i v_j}(s)}{\sigma_{v_i v_j}} \quad (4)$$

Where S is the discourse of the n fuzzy vectors.

Centrality function is given as:

$$C_j = \sum_{j=1}^{n} \frac{\sigma_{v_i v_j}(B_i^j)}{\sigma_{v_i v_j}} \quad (5)$$

Then the objective function is given as:

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$E = \sum_{j=1}^{k} \sum_{i=1}^{n} \| b_i^j - c_j \|^2 \quad (6)$$

From the proposed classifier, we can evaluate the subregions of the class.

C) Clusters to category mapping:

Classifier to category mapping is the third step in text classification process. Category is used to cover a region that consists of number of subregions. It is necessary to explore a relationship between clusters and categories. A linear mapping structure is used for assigning the categories. For a dataset d with k-edge classifier, the cluster likeliness vector is obtained from (6), which feed as an input to the mapping functions and category likeliness vector is given as $[C_1, C_2, C_3 \ldots C_p]$, output of the category mapping. The linear model of category's output is in form of

$$C_k = \gamma_{1,k} E_1(x) + \gamma_{2,k} E_2(x) + \ldots\ldots + \gamma_{K,k} E_K(x) \quad (7)$$

For every $1 \le k \le p$, the C_k contains the likelihood of x to category C_k. If C_k is near to 1, then it belongs to category C_k and if C_k is near to 0, then it belongs to category C_k. The weight γ is calculated from the training database.

D) Finding the boundaries

The category likelihood vector $[C_1, C_2, C_3 \ldots C_k]$ is generated for the database which in turn generates the output values $O_1, O_2 \ldots O_k$. The $O_{i=}$ 1 implies that it belongs to category Ck, $1 \le k \le p$. A variant number of clusters belong to 1 which means that multilabel classification is achieved.

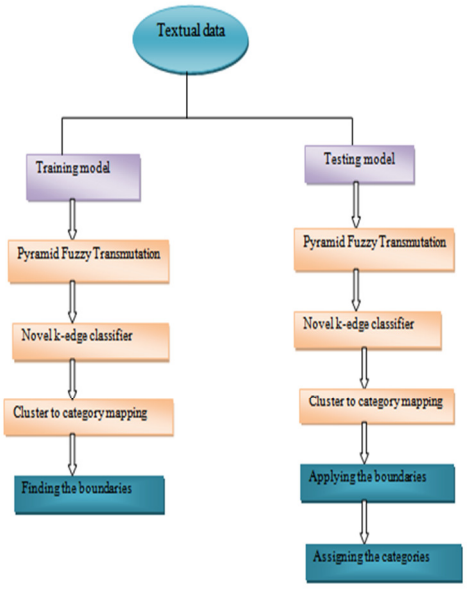

Figure 4: Proposed Architecture

4. EXPERIMENTAL RESULTS

This section depicts the performance validation of our proposed work.

4.1 Dataset Description

Freebase [33] is a public knowledge base. It compromised of semi-structured data about the real world entities and their facts. According to recent statistics on 2015, the freebase contains 48 million entities and 2.9 billion facts. It is the largest public knowledge when compared to Open Information Extraction [34], NELL [35], and DBPedia [37]. Freebase's information is organized in graph form. An object is denoted as the node with a unique machine ID. An object is connected to another object or its attribute by a link known as edges. A variant number of edges are available that denotes different facts. Several numbers of Freebase dataset are available in which we took 'Freebase Deleted Triples' dataset. It contains the information that is deleted over period of time due to its data maintenance issue. In freebase,

i) Subject: It is the ID of a freebase object.
i) Predicate: It is the human-readable ID for a freebase property.
ii) Object: It is the freebase MID for the subject schema.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1. Dataset's attribute and its value

Attribute	Value
Creation_ timestamp	1.35285E+12
Creator	/user/mwcl_wikipedia_en
Deletion_timestamp	1.35286E+12
Deletor	/user/mwcl_wikipedi
	a_en
Subject (MID)	/m/03r90
Predicate (MID)	/type/object/key
Object (MID/ Literal)	/wikipedia/en/$B816
Language_code	en

Table 2. Sample Dataset

Creation_ timestamp	Creator	Deletion_timestamp	Deletor	Subject (MID)	Predicate (MID)	Object (MID/ Literal)	Language _ Code
1.35285E+12	/user/mwcl_wikipedia_en	1.35286E+12	/user/mwcl_wikipedia_en	/m/03r90	/type/object/key	/wikipedia/en/$B816	en
1.35517E+12	/user/mwcl_musicbrainz	1.36426E+12	/user/turtlewax_bot	/m/0nncp9z	/music/recording/artist	/m/01vbfm4	en
1.17663E+12	/user/mwcl_images	1.33593E+12	/user/gardening_bot	/m/029w57m	/common/image/size	/m/0kly56	en
1.29285E+12	/user/mwcl_musicbrainz	1.36482E+12	/user/mbz_pipeline_merge_bot	/m/0fv1vl8	/type/object/type	/common/topic	en
1.3428E+12	/user/googlebot	1.34292E+12	/user/google_gardener	/m/0k7nmpn	/type/object/key	/dataworld/freeq/job_aa7b73bc-23b7-4836-925f-3235a8a42bdb_var_google_nces_university_domestic_tuition_domestic_tuition_456287_2008	en
1.20553E+12	/user/mwcl_images	1.33602E+12	/user/gardening_bot	/m/01x5scz	/common/licensed_object/license	/m/02x6b	en
1.25662E+12	/user/tvrage	1.33651E+12	/user/gardening_bot	/m/07xv13j	/type/content/uploaded_by	/m/0668v6f	en

4.2 Statistical Tool Description

SPSS is a window based program which is widely used in Predictive Data Mining systems. The unique property of SPSS is that, it can handle a large volume of data. Mostly, it is useful for the study of social sciences. In the field of Predictive Data mining, SPSS tool is widely adopted in today's environment. As this research concentrated on the Text mining, we make a novel attempt to use SPSS in this study. Since, statistics are a good complement to data mining. First of all, the string variables are recoded into numeric value.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

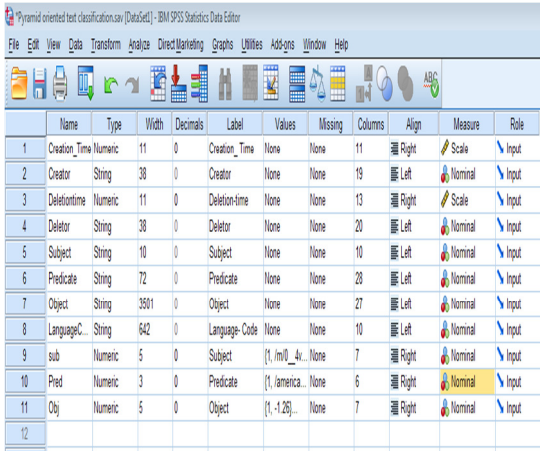

Figure 5: Recoding The String Variable Into Numeric Value In SPSS.

a) Finding the Data Intensity

Data Intensity is the study about the structure of the data points in the datasets. The importance of the Subject, Predicate and Object is studied to categorize the text. Subject is the important factor for text categorization.

Table 3: Frequencies Estimation For The Subject, Predicate And Object

Statistics

		Subject	Predicate	Object
N	Valid	943431010	943431010	943431010
	Missing	0	0	0
Mean		41948.29	579.17	18875.83
Std. Deviation		15011.263	208.749	11146.908
Percentiles	10	19375.00	197.00	4498.00
	20	27824.00	451.00	7031.00
	30	34332.00	537.00	11789.00
	40	39621.00	584.00	15012.00
	50	44387.00	597.00	18049.00
	60	48778.00	751.00	22145.00
	70	52826.00	751.00	23854.00
	80	56518.00	752.00	30379.00
	90	60084.00	754.00	34791.00

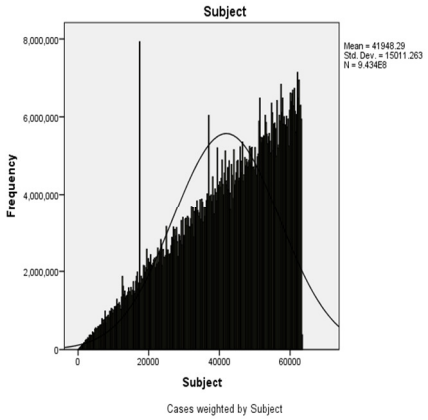

Figure 6: Histogram chart for subject

Figure 6 depicts the importance of the Subject. Based on the frequency of the cases, the histogram of the subject is studied.

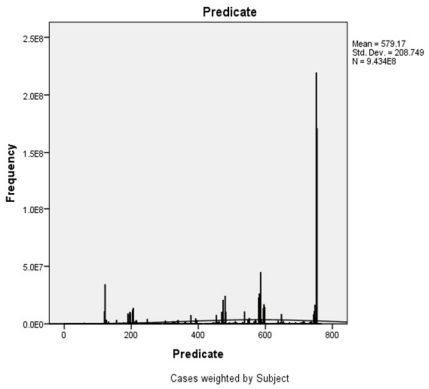

Figure 7: Histogram Chart For Predicate

Figure 7 depicts the importance of predicate. Based on the frequency of the cases, the histrogram of the predicate is studied.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

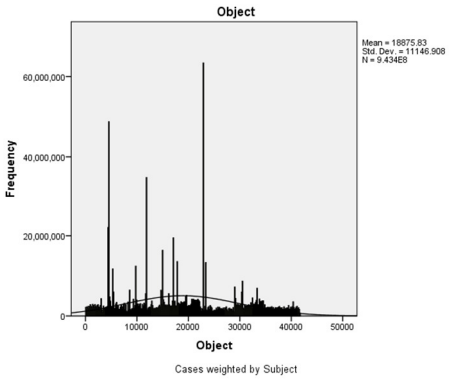

Figure 8: Histogram Chart For Object

Figure 8 depicts the importance of the object. Based on the frequency of the cases, the histogram is studied.

From the histograms study of the features Subject, Object and Predicate, we can predict the pyramid oriented memberships between them.

b) Forming the clusters:

The parameter setting for the k-edge classifier is given as No. of clusters N_c= 10, Maximum no of iterations= 25. The final center clusters is presented as:

Table 4: Mean Value Presented In The Final Cluster Centers

Cluster. No	Weight (γ) (expressed in scientific notation)
1	$1.2 * 10^8$
2	$1.4 * 10^8$
3	$1.1 *10^8$
4	$7.0* 10^8$
5	$1.1 *10^8$
6	$5.6 *10^8$
7	$1.3*10^8$
8	$5.5* 10^8$
9	$4.2 *10^7$
10	$8.2*10^8$

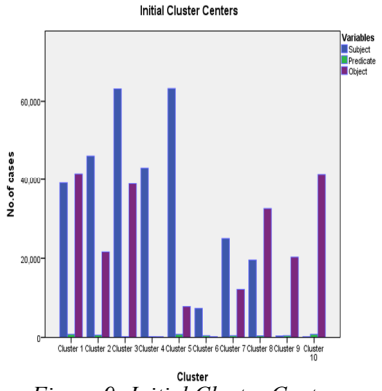

Figure 9: Initial Cluster Centers

Figure 9 depicts the study of the attributes Subject, Predicate and Object before the formation of the clusters. It is inferred that there is an improper order of the features. This type of disorder arrangement guides the issue of data redundancy. Data redundancy refers as the user can search the same subject with different objects.

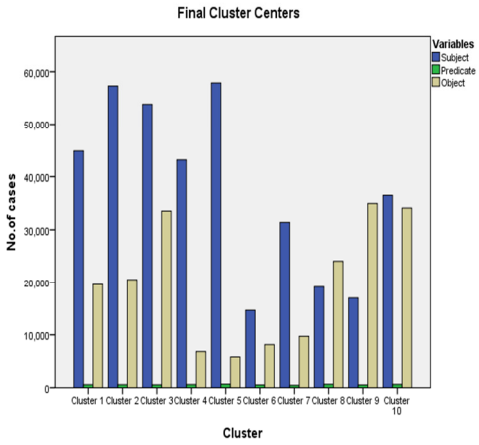

Figure 10: Final Cluster centers

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 5: Objective function E from its clusters

Cluster No.	Subject	Predicate	Object
1	44834	577	19565
2	57302	593	20580
3	53839	568	33525
4	43130	601	6827
5	57894	668	5803
6	14652	548	8142
7	31416	457	9719
8	19126	645	24080
9	16987	549	34955
10	36495	633	34101

5. *Cluster to category formation:*

In accord to eqn. (6), the categories are assigned for 10 clusters. The weight for each cluster is tabulated:

Table 6: Weight (γ) is estimated.

Case No.	Cluster Number of Case	Distance of Case from its Classification Cluster Center
1	8	1930.00476
2	5	1664.42269
3	8	9699.80961
4	4	2938.25256
5	5	5799.54444
6	6	8445.10716
7	7	4099.89202
8	1	3362.77668
9	4	6890.79935
10	4	6066.60116

Case No.	Cluster Number of Case	Distance of Case from its Classification Cluster Center
1	8	1930.00476
2	5	1664.42269
3	8	9699.80961
4	4	2938.25256
5	5	5799.54444
6	6	8445.10716
7	7	4099.89202
8	1	3362.77668
9	4	6890.79935
10	4	6066.60116

Figure 11: Weights of each cluster

Then the category likelihood vector is $E^{(x)}$, $1 \le x \le 10$ are computed.

Journal of Theoretical and Applied Information Technology
30th June 2017. Vol.95. No 12

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

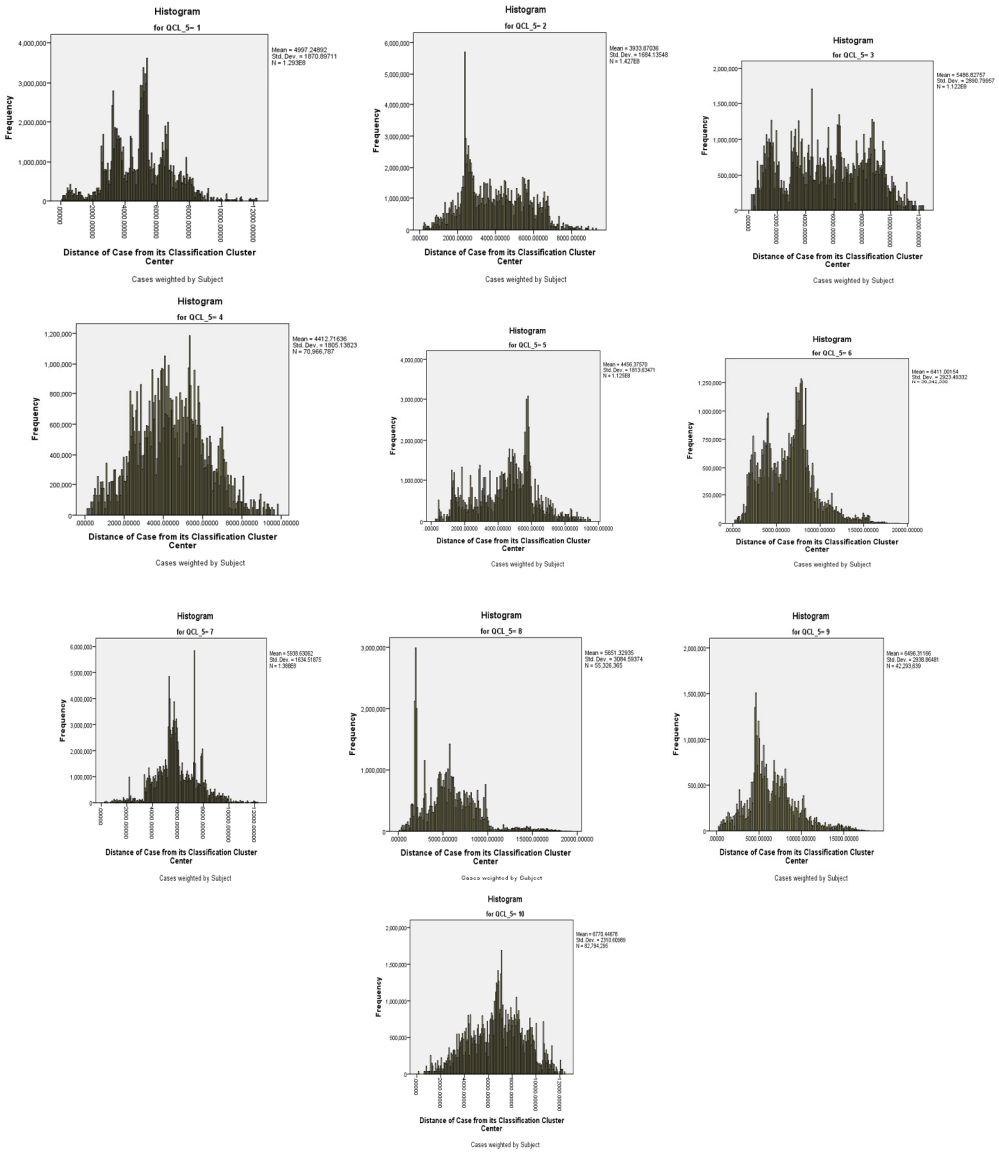

Figure 12: Distances of the cases in each cluster

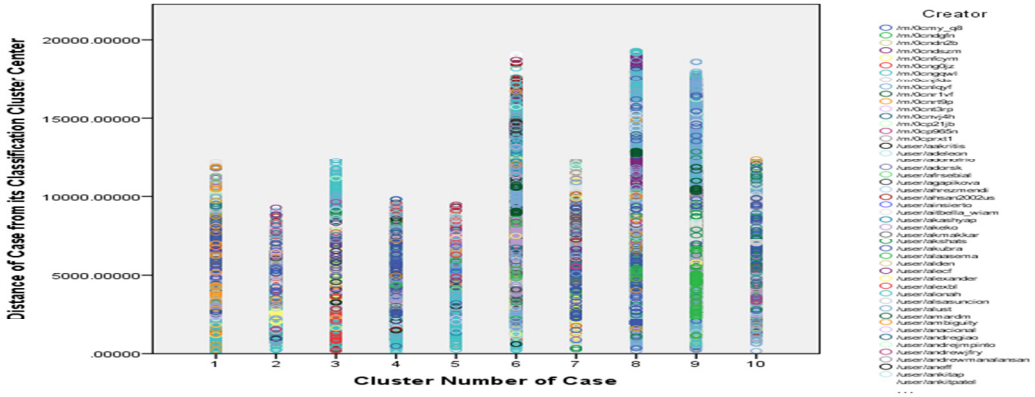

Figure 13: Cases Presented In Each Cluster Based On The Distance Value

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. *Finding the boundaries*

Table 7: Finding The Boundaries For The Categories

Categories	Mean	Min. Boundaries	Max. Boundaries
1	4997.2	66.70	12170.47
2	3933.8	220.20	9303.82
3	5486.8	179.65	12667.96
4	4412.7	101.32	9841.40
5	4456.3	238.14	9494.77
6	6411.0	214.86	19096.56
7	5938.63	262.23	12216.63
8	5651.32	178.24	19335.85
9	6496.31	232.52	18649.21
10	6770.44	169.42	12407.39

In the testing phase, the new document is again processed as follows:

i) Studying the data topology
ii) Estimating pyramid membership functions
iii) Estimation k-edge classifier
iv) Cluster to category mapping
v) Applying the boundaries from the training phases
vi) Assigning the categories

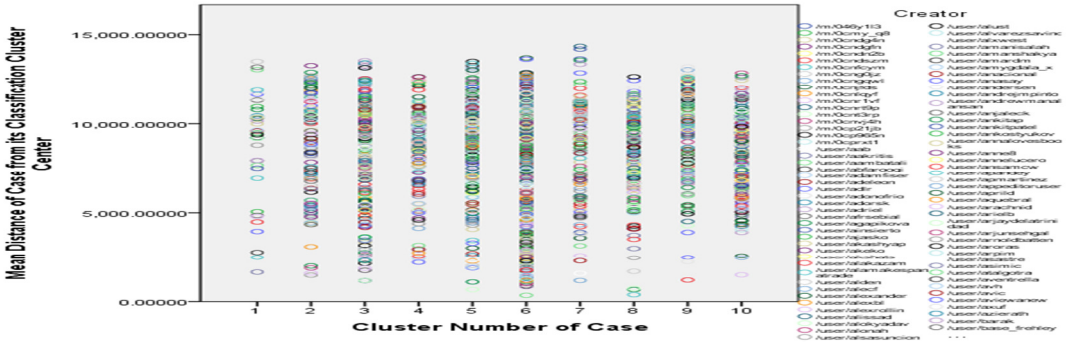

Figure 14: New Data Is Categorized Based On The Boundaries Values

5. CONCLUSION

In this paper, we present a novel framework for multi-label text classification scheme. The main motivation for the research was to develop the frameworks using pyramidal clustering techniques. Firstly, the topology of the data points is studied using triangular pyramidal approaches. The pyramidal view is combined with fuzzy membership functions to select the required features i.e. Converting high dimensional to low dimensional space. These low dimensional vectors are grouped and allocated to the clusters. Based on the distance from its cluster centers, the boundaries are estimated for assigning the categories. These boundary values are applied to the new document after the cluster formation and category is assigned. Experimental results were carried out in Freebase dataset. The attributes selected are Subject, Predicate, Object and Creation-time. The novel framework was very stable and reliable. The proposed approach based on pyramidal membership method can achieve better classification accuracy than the traditional approaches especially that includes over-fitting document categories.

REFERENCES:

[1] Bruno Trstenjaka, Sasa Mikacb, Dzenana Donkocm, "KNN with TF-IDF Based Framework for Text Categorization", 24th *DAAAM International Symposium on Intelligent Manufacturing and Automation*, 69: 1356 – 1364, 2014.

[2] Michal Hrala and Pavel Kral, "Evaluation of the Document Classification Approaches", doi: 10.1007/978-3-319-00969-8_86, 2013.

[3] Ashis Kumar Mandal1and Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", *International Journal of Artificial Intelligence & Applications (IJAIA)*, 5(5), 2014.

[4] Erlin, Unang Rio, "Text Message Categorization of Collaborative Learning Skills in Online Discussion Using Support Vector Machine", *2013 International*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Conference on Computer, Control, Informatics and Its Applications , 2013.

[5] Joachims, T, "Transductive inference for text classification using support vector machines", *Proceedings of ICML-99, 16th International Conference on Machine Learning, eds. Morgan Kaufmann Publishers*, San Francisco, US: Bled, SL, 1999, pp. 200–209

[6] Addis, A., "Study and Development of Novel Techniques for Hierarchical Text Categorization", *PhD Thesis Electrical and Electronic Engineering Dept.*, University of Cagliari, Italy, 2010.

[7] Feldman, R & Sanger, J, "The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data", *Cambridge University Press New York*, 2006.

[8] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz, "A decision-tree-based symbolic rule induction system for text categorization", IBM Systems Journal, 2002.

[9] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer, "KNN Model-Based Approach in Classification doi: 10.1007/978-3-540-39964-3_62: 986-996, 2003.

[10] C. C. Aggarwal, and C. Zhai, "Mining text data", doi: 10.1007/978-1-4614-3223-4, 2012

[11] Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, Mohammed Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", *4th International Conference on Electrical Engineering and Informatics* , pp. 734-739, 2013.

[12] Z. Wei, D. Miao, J.-H. Chauchat, "N-grams based feature selection and text representation for Chinese Text Classification", International *Journal of Computational Intelligence Systems*, 2 (4), 2009, pp. 365-374.

[13] T. Joachims, "A statistical learning model of text classification for support vector machines", *in: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, 2001, pp. 128-136.

[14] Forman, G. "An Experimental Study of Feature Selection Metrics for Text Categorization", *Journal of Machine Learning Research,* 2003, pp. 1289-1305.

[15] Chaitrali S. Dangare and Sulabha S. Apte, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques", *International Journal of Computer Applications*, 47(10), 2012.

[16] D.Lavanya and Dr.K.Usha Rani, "Ensemble Decision Tree Classifier for Breast Cancer Data", *International Journal of Information Technology Convergence and Services (IJITCS)*, 2(1), 2012.

[17] Elakia, Gayathri, Aarthi, Naren J, "Application of Data Mining in Educational Database for Predicting Behavioral Patterns of the Students", *International Journal of Computer Science and Information Technologies,* 5(3), 2014, pp. 4649-4652.

[18] T. Santhanam and Shyam Sundaram, "Application of CART Algorithm in Blood Donors Classification ", *Journal of Computer Science,* 6 (5), 2010, pp. 548-552.

[19] Jaydeep Jalindar Patil, Nagaraju Bogiri, "Automatic Text Categorization Marathi Documents ", *International Journal of Advance Research in Computer Science and Management Studies*, 3(3), 2015.

[20] Ashis Kumar Mandal, Rikta Sen, "Supervised Learning Methods for Bangla Web Document Categorization", *International Journal of Artificial Intelligence & Application* (IJAIA), DOI:10.5121/ijaia.2014.5508, 2014.

[21] Neha Dixit, Narayan Choudhary, "Automatic Classification of Hindi Verbs in Syntactic Perspective", International Journal of Emerging Technology and Advanced Engineering, 2014, pp: 2250- 2459

[22] Aruna Devi, K., Saveetha, R, "A Novel Approach on Tamil Text Classification Using C-Feature", International *Journal of Scientific Research & Development*, ISSN: 2321-0613, 2014.

[23] Nidhi, Vishal Gupta, "Punjabi Text Classification using Naïve Bayes, Centroid and Hybrid Approach", *International journal on Computer science and Information technology*, DOI: 10.5121/csit.2012.2421, 2012.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[24] Nidhi, Vishal Gupta, "Algorithm for Punjabi Text Classification", *International Journal of Computer Applications*, ISSN: 0975-8887, 2012.

[25] Nadimapalli V Ganapathi Raju, "Automatic Information Collection & Text Classification for Telugu Corpus using K-NN ", *International Journal of Research in Computer Application & Management*, ISSN: 2231-1009, 2011.

[26] K. Rajan , "Automatic classification of Tamil documents using vector space model and artificial neural networks", *Expert Systems with Applications*, ELSEVIER, 2009.

[27] Abbas Raza Ali, Maliha Ijaz, "Urdu Text Classification", FIT'09, December 16-18, 2009, CIIT, Abbottabad, Pakistan.

[28] Meera Patil, Pravin Game, "Comparison of Marathi Text Classifiers", ACEEE *International Journal on Information Technology*, DOI: 01.IJIT.4.1.4, 2014.

[29] Bijal Dalwadi, Vishal Polara, Chintan Mahant, "Review: Text Categorization for Indian Language ", *International Journal of Engineering Technology, Management and Applied Sciences*, ISSN: 23349-4476, 2015.

[30] Bhumika, Prof. Sukhjit Singh Sehra, Prof. Anand Nayyar, "A Review Paper on Algorithms Used for Text Categorization", *International Journal of Application or Innovation in Engineering Technology & Management,* ISSN: 2319-4847, 2013.

[31] https://www.freebase.com/

[32] http://ai.cs.washington.edu/projects/open-information-extraction

[33] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an Architecture for Never-Ending Language Learning", *In AAAI*, 5(3) , 2010.

[34] Wiki.dbpedia.org