International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

JOINT INTELLIGENCE TEXT CLUSTERING USING DATA INFORMATION THEORETICAL MINING MODEL

[*1] Mrs. Ranjani A.M., [*2]Mrs. Shoba S.A.,

[*1]M.Phil Research Scholar, PG & Research Department of Computer Science & Information Technology Arcot Sri Mahalakshmi Women's College, Vellore, Tamil Nadu, India.

[*2]Assistant Professor, HOD of PG & Research Department of Computer Science & Information Technology Arcot Sri Mahalakshmi Women's College, Vellore, Tamil Nadu, India.

Abstract: - *Data mining from organizational data for extracting hidden knowledge is a growing field of study, which forms new study groups named knowledge discovery in database. Engineering applications of data mining which include, web mining, network mining, image mining and multimedia mining in general and text mining in particular. Predicting business flow, finding recent trends, sales and markets monitoring, forecasts, competition monitoring, expenditures control, revenues and administrative processes, finding volcanoes on Venus, Earth geophysics, earthquake photography from space, atmospheric science are some of the real world applications of data mining. Around 80% of data in the world are stored in unstructured textual format. Hence, the text mining and document clustering are major research areas in the past few years. Designing a effective and novel text mining requires high dimensionality, dynamic methodology, fast information access, specialized knowledge extraction from very huge data sets. Variety of approaches is proposed in the literature and these approaches vary with a range of traditional k-means algorithm, term based methods, pattern taxonomy model and rank based methods.*

KeyWords: *Quality of Software Product, Data Mining Measurement Feature, Testability, text mining, clustering.*

I. INTRODUCTION

The modern world uses physical, biological and social systems more effectively with the help of advanced computerized techniques. Therefore, a great amount of data being generated by such systems; it leads to a paradigm shift from classical modelling and analysis based on basic principles in developing models and the corresponding analysis directly from data.[1][4] The AJIlity to extract useful hidden knowledge in these data and to act on that knowledge is becoming increasingly important in today's competitive world[3]. The entire process of applying a computer based methodology for discovering knowledge from huge amount of data is called data mining.[2] Data mining is a very helpful tool to handle such large data set which will find the hidden, valuable and nontrivial knowledge from huge subject oriented historic data warehouse. The data mining is also known as, discovery through one or more mining methods. It is an iterative process, which is a part of knowledge discovery in database.

Data mining is an iterative process. This process is defined by discovery through either automatic or manual methods. Data mining is most useful in an exploratory analysis [3],[2], in which, there are no predetermined notions. Data mining is the technique which searches for new, valuable and nontrivial information from large volumes of data. It is a cooperative effort of humans and computers. Best results are achieved by balancing the knowledge of human experts in describing problems and goals with the search capAJIlities of computers. Brief methodologies used in data mining are provided below.[2] tested software, verifying that the system meets its specification.

A single security problem can cause severe damage to an organization by not only incurring large costs late fixes but by losing invaluable assets and credibility and leading to legal issues. Annual world-wide losses caused from cyber attacks have been reported for. The organizations must prioritize vulnerAJIlity detection[5], efforts and prevent vulnerAJIlities from being injected. One way of identifying the most vulnerable code locations is to use characteristics of the software product itself. Perhaps complex code is more likely to be vulnerable than simple code.

1.1 Data Mining to Text Mining

A data mining is the extraction of hidden knowledge (Han et al, 2011) from the analytical data. The data mining is a process that combines several

mathematical, statistical or algorithm methods, to determine a solution for a problem in a decisional universe. Data mining describes another characteristic of the process. In general, when mining data, useful and interesting information are extracted from organizational data. However, the most common is to use numeric and so, quantitative data. It requires a good knowledge of statistics. Five major functionalities of data mining are associative rules, classification hierarchies, sequential patterns, patterns of temporary series, clustering and segmentation (Berry and Kogan, 2010).

- ***Associative rules:*** To seek to "find items in a transaction that can determine the presence of other items in the same transaction".
- ***Classification hierarchies:*** To create "a model based on known data" and help to explain the reason of a given classification. This model allows to "classifying new data with an existing classification". For example, to create limits for credit concession based on the transaction report of previous loans.
- ***Sequential patterns:*** To indicate behaviors or a sequence of behaviors. For example, "whenever a young woman buys leather shoes, she will also buy belts and bags in the next thirty days".
- ***Patterns in temporary series:*** To show similar occurrences in a space of time. To the above data, seasons are added—such as, in the winter, young women buy leather shoes, bags an d belts. In the summer, this pattern is inverted to sandals, bags and hats.
- ***Clustering and segmentation:*** To gather occurrences with similar characteristics. Example: based on their buying expenses, a group of consumers can be classified as "small buyers", "average buyers", or "big buyers" of a certain product.

II. Problem Statement

Given a fixed number L of classes, each document can be in multiple, exactly one or no class at all. Classes are usually semantic topic identifiers used to tag documents, newswires, web pages (Li et al, 2000).The multi-label text classification setting is modelled with an 'L' dimensional class vector y where each component can take the value of {-1 to +1}. Formally, the class vector is defined in equation (1.1) as follows:

$$y = \{-1, +1\}L \qquad (1.1)$$

Classification errors have to be counted. Indeed, the 0=1 loss function does not allow to model close misses while a reasonable distance metric is offered by the Hamming distance which counts the number of mismatches between the class vector y and the classifier output 'Ay'.

Given a set of observations (x_1, x_2, x_n), where each observation is a 'd' dimensional real vector, k-means clustering aims to partition the 'n' observations into 'k' sets $(k < n)$ $S = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares (WCSS), the computation of WCSS is given in equation (1.2).

$$\text{Arg } S \min \sum\_(i=0)^K \sum\_Xi=S) \, || X- S||^2 \quad ...(1.2)$$

2.1 The Quantitative features can be subdivided as

- Continuous values (e.g., real numbers where P_j^ R),
- Discrete values (e.g., binary numbers Pj= {0, 1}, or integers P_{jc} Z) and
- Interval values (e.g., $P_j = \{x_{ij} < 20, 20 < xi_j < 40, x_{ij} > 40\}$). The Qualitative features can be subdivided as
- Nominal or unordered values (e.g., colour are "blue" or "red") and
- Ordinal values (e.g., military rank with values "general", "colonel").

2.2 Applications of Text Mining

Few applications of text mining are given below:

- Document management which includes document organization, version control and security policy for business documents.
- Document imaging which involves the digitization of paper based document such as capturing, transforming and managing. The document imaging includes text, image, audio and video. (Chiu et al, 2012).
- Records management like long-term document archiving according to compliance policies.
- workflow management like business process management in an organization
- Web content management using the integration of contents for web publishing.
- Document-centric collaboration for document sharing for project teams.
- Natural language processing (NLP) is the attempt to extract a fuller meaning representation from free text.

2.3 Related Work

The Clustering is a well-established technique for data interpretation. It usually requires prior information,

e.g., about the statistical distribution of the data or the number of clusters to detect. "Clustering" attempts to identify natural clusters in a data set. It does this by partitioning the entities in the data such that each partition consists of entities that are close (or similar), according to some distance (similarity) function based on entity attributes (Luhr and Lazarescu, 2009).

Existing clustering algorithms such as K-means, Partioning Around Medoids (PAM), Clusterig Large Applications based RANdomized Search (CLARANS), Density Based Spatial Clustering of Applications with Noise and (DBSCAN) are designed to find clusters that fit some static models. For example, K-means, PAM and CLARANS assume that clusters are hyper-ellipsoidal or hyper-spherical and are of similar sizes. The DBSCAN assumes that all points of a cluster are density reachable and points belonging to different clusters are not

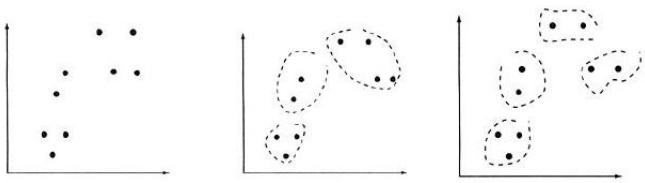

Fig : (a) Initial data (b) Output in three (c) Output in four

Many clustering algorithms that work well with traditional data deteriorate when executed on geospatial data (which often are characterized by a high number of attributes or dimensions), resulting in increased running times or poor-quality clusters. For this reason, recent research has cantered on the development of clustering methods for large, highly dimensioned data sets, particularly techniques that execute in linear time as a function of input size or that require only one or two passes through the data. Recently developed spatial clustering methods that seem particularly appropriate for geospatial data include partitioning, hierarchical, density based, grid based and cluster based analysis.

An input to a cluster analysis can be described as an ordered pair (X, s), or (X, d), where 'X' is a set of descriptions of samples and 's' and 'd, are measures for similarity or dissimilarity (distance) between samples, respectively in equation (2.1) and (2.2). Output from the clustering system is a partition A = {G1, G2, ..., GN} where Gk, k = 1, ..., N is a crisp subset of 'X' such that:

G1u G2u uGN = X (2.1)

G1n G2 n nGN = 0 (2.2)

The G1, G2 . Gn are the clusters. Most clustering algorithms are based on the following four popular approaches:

(1) Partitioning methods

(2) Hierarchical clustering

(3) Iterative square-error partitioned clustering

(4) Density based clustering

III. PREVIOUS IMPLEMENTATIONS

Existing text mining methods has many pitfalls like slow processing, lesser scalAJIlity, unable to solve conceptual problems like synonomy and polysemy. Therefore, this thesis focused text mining, proposed conceptual analysis and concentrated to solve conceptual problem like synonomy. Initially, this research work proposed metadata conceptual mining model for effective text mining. The proposed work executes in two phases of manipulation, which are training phase and testing phase. Initially in the pre-processing stage, the di-grams such as in, as, it; and tri-grams such as are, for, ing are removed from the documents. The proposed work created a data structure called, Significant Model

3.1 ANALYSIS OF JOINT INTELLIGENCE LEARNING METHOD

These prediction models selecting from biological neural networks are made up of real biological neurons that are physically connected or functionally-related in the human nervous system and especially in the human brain. The human brain can perform tasks much faster than the fastest existing computer, thanks to its special AJIlity in massive parallel data processing. ANN tries to mimic such a remarkable behaviour for solving narrowly defined problems, i.e. problems with an associative or cognitive tinge. To this effect, ANN have been extensively and successfully applied to the pattern (speech/image) recognition, time-series prediction and modeling, function approximation, classification, adaptive control and other areas. Neural networks are made of several processing units called neurons. Three types of neurons are distinguished: input neurons which receive data from outside the ANN and are organised in the so called input layer, output neurons which send data out of the ANN and generally comprise the output layer, and hidden neurons whose input and output signals remain within the ANN and form the so called hidden layer (or layers).

- Supervised learning or Associative learning in which the network is trained by providing it with input and matching output patterns. These input-output pairs can be provided by an external resource, or by the system which contains the network (self-supervised).
- Unsupervised learning or Self-organization in which an output unit is trained to respond to clusters of pattern within the input. In this paradigm the system is supposed to discover statistically salient features of the input population.
- Both learning paradigms discussed above result in an adjustment of the weights of the connections between units, according to some modification rule.

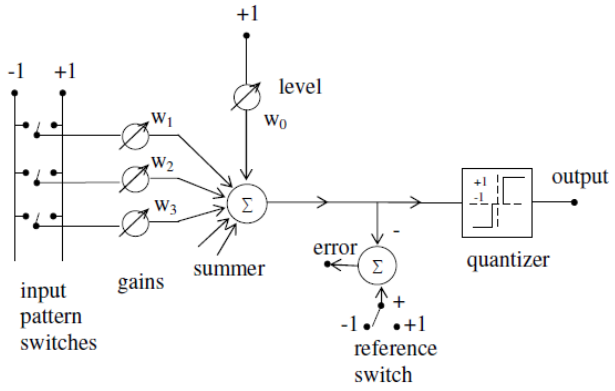

Fig : Working model of the Adaline

3.2 Levenberg-Marquardt algorithm (LM)

This is a LM algorithm with traditional forward-backward computation; for LM (and NBN) algorithm, the improved forward-only computation performs faster training than forward-backward computation for networks with multiple outputs. Now it is also only used for standard MLP networks. LM (and NBN) algorithm converges much faster than the EBP algorithm for small and media sized patterns training.

3.3 Neuron By Neuron (NBN)

Since the development of EBP-error back propagation-algorithm for training neural networks, many attempts were made to improve the learning process. There are some well-known methods like momentum or variable learning rate and there are less known methods which significantly accelerate learning rate. The recently developed NBN (neuron-by-neuron) algorithm is very efficient for neural network training. Comparing with the well known Levenberg-Marquardt algorithm.

- ***Forward Computation:*** In the forward computation, the neurons connected to the network inputs are first processed so that their outputs can be used as inputs to the subsequent neurons. The neurons are then processed as their input values become available.
- ***Backward Computation:*** The sequence of the backward computation is opposite to the forward computation sequence. The process starts with the last neuron and continues toward the input. The vector 8 represents signal propagation from a network output to the inputs of all other neurons. The size of this vector is equal to the number of neurons.
- ***Jacobian Element Computation:*** After the forward and backward computation, all the neurons outputs y and vector 8 are calculated. By applying all training patterns, the whole Jacobian matrix can be calculated and stored.

IV. PROPOSED ANALYSIS JOINT INTELLIGENCE LEARNING METHOD

Textual pattern mining is one of the major research areas in the field of data mining. The data mining is an emerging technique which applies many approaches and methods from another field of study and the data mining is implemented in another area to learn hidden knowledge. In this proposed work, ANN is used for learning textual pattern in the Metadata conceptual mining model. The proposed learning algorithm is called as, analysis of bilateral intelligence, is used to identify and classify the synonymy of the sentences. The proposed method provides efficient learning which identifies patterns which have synonymy and the convergent of the training algorithm is very fast than existing methodology. From the results, it is concluded that the performance of proposed AJI is optimized. Hence, the proposed Metadata conceptual mining model with AJI learning will provide optimality than existing clustering algorithm.

4.1 Analysis of Joint Intelligence (AJI)

The proposed ANN based unsupervised learning, is termed as, Analysis of Joint Intelligence (AJI). The AJI applies the learning process to identify two equivalent terms which have the same meaning. AJI contains text

documents as datasets, improving accuracy of text clustering which is the required output and achieving error free clustering in a shorter time is the goal.The working model of the proposed AJI Learning method is explained in the following sections: The sigmoid function which shown in equation (4.1) is applied in the proposed AJI,

$$X_a = \frac{1}{1+e-x} \quad ...(4.1)$$

Where the inputs are 'x' which is connected to the hidden layer from input layer. The connection has weights 'rai', between inputs to hidden layer. And the output of the neurons referred as 'sba' is computational values between output and hidden layer. Where, 'b' neurons in the output layer, 'a' neurons in the hidden layer and 'i' neurons in the input layer.

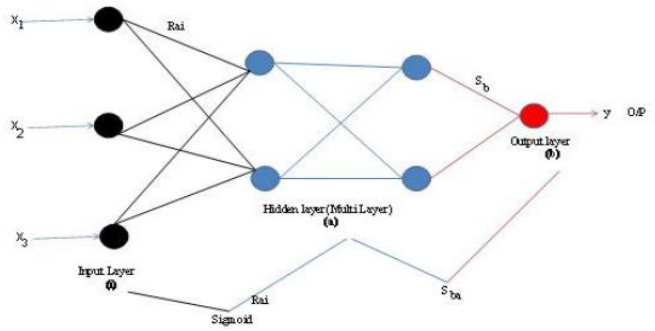

Fig : Design of Neuron Model

Step 1: Initial Phase

The proposed AJI has implemented from well known initial phase. In the initial phase, the values of the weights are assigned. Let the values are 'R' and 'S'. 'R' is a value of the hidden layer and input layer. 'S' is a value of output layer - hidden layer respectively. The other constants are penalty constant, which is defined as u; and the number of iterations, which is called an epoch, is initialized in the system. The weight vectors 'R' and 'S' are to be optimized in order to minimize the error function. The second stage involves a backward transmission, which passed through the network after the error was computed. The error signal is passed to each unit in the network and the appropriate weight changes are calculated.

Step 2: Weight adjustments Phase

This weight adjustment step is processed based on sigmoid activation function, shown in the first phase. The weight of a connection is adjusted by an amount proportional to the product of an error signal calculated in the second stage of the first phase. On the neuron, the unit 'k' receiving the input and the output of the unit 'j' is sending this signal along the connection.

Step 3: Optimization of Output Layer Weights

$$S_{Optium} = A^{-1} x\ B$$

The concept of state is fundamental to this description. The state vector or simply state, denoted by 'xb', is defined as the minimal set of data that is sufficient to uniquely describe the unforced dynamical behaviour of the system; the subscript 'b' denotes discrete time. In other words, the state is the least amount of data on the past behaviour of the system that is needed to predict its future behaviour. Typically, the state 'xb' is unknown. To estimate it, use a set of observed data, denoted by the vector 'yb'.

Step 4: Test for Completion

RMS error (ERMS) was then calculated comparing the 'Rtest matrix with 'Soptimum' matrices calculated in Step 3.

a. *ERMS< E*

The hidden layer weight matrix 'R' is updated 'R'= 'Rtest. Decrease the influence of the penalty term by decreasing 'n', Proceed to Step 5.

b. ERMS > E

Step 5: Process Termination

If the RMS error is not within the desired range, repeat Step 3, else the training process is ceased. After the successful completion of the training phase, the sample real time data are given as input of the system.

Type of ANN Model	% RMS Error in Estimation	% RMS Error in Elimination
NBN Model	7.83	5.15
2HLANN Model	7.23	8.65
Proposed AJI Learning Model	4.60	4.75

Table : Summary of Errors (in %)

NO.OF EPOCH	NBN	2HLANN	PROPOSED AJI
50	0.175	0.15	0.13
100	0.14	0.11	0.08
150	0.10	0.08	0.05
200	0.075	0.06	0.025
250	0.04	0.025	0.010

Table 1.2: Comparison of error growth on Proposed model Vs Existing Models

System will choose the comparatively best path. This thesis used 60% dataset for training and 40% dataset for testing.

IMPLEMENTATION ALGORITHM

ETM-ILM ALGORITHM

Training Algorithm

Step 1 : Apply preprocessing

Step 2 : Prepare STL for each field of study

Step 3 : Check the metadata stored in each STL is unique and primary data

Step 4 : Verify that all training documents are read then go to step 9, otherwise continue step 5.

Step 5 : Calculate the number of matching terms in the given document which matching the STL is 'm' and calculate the total number of terms in the given document is 'n'.

Step 6 : Compute 'mda', where 'mda' =m/n

Step 7: Sort the 'mda' in decreasing order and check the terms which has higher 'mda' terms in the STL, if available, go to step 8 otherwise go to step 9.

Step 8 : Update these new terms to concern STL and go to step 3

Step 9 : Apply AJI learning algorithm

Step 10: Compare output with a minimum threshold. If outputs above threshold go to step 11, otherwise go to step 12.

Step 11: Update STL

Step 12: Go to the testing process

Testing Algorithm

Step 1: Apply pre-processing

Step 2: Collect STL for each field of study from training algorithm

Step 3: Check the metadata stored in each STL is unique and primary data

Step 4: Apply AJI and verify interestingness of each keyword presented in the STL. If the confidence of interestingness is not in acceptable level, which may be removed from concern STL.

Step 5: Apply the input test document

Step 6: Read each term in every STL and calculate the number of matching terms in the given test document. In which, the terms are matching with the STL is 'm' and calculate the total number of terms in the given test document is 'n'.

Step 7: Calculate 'mda', 'mda' = -n

Step 8: Sort the 'mda' in decreasing order

Step 9: Check the terms which have higher 'mda'

Step 10: Check this highest 'mda' term is available in the given STL, if available, go to step 10 otherwise gotostep11.

Step 11: Classify the given test document as the field of matching STL

Step 12: Identify next higher 'mda' term until end of the test document and go to Step 9

V. EVALUATION RESULT

The proposed ETM-ILM is implemented in Mat Lab (MATLAB). Performance analysis and comparison of proposed work with existing TBM, CBMM and PTM are computed..

Field of Study	Data Set	TBM	CBMM	PTM	Proposed ETM-ILM
Electrical	IEEE	0.697	0.741	0.780	0.833
	ACM	0.767	0.812	0.834	0.890
	Scopus	0.724	0.807	0.828	0.887
Electronics	IEEE	0.688	0.731	0.753	0.825
	ACM	0.757	0.801	0.836	0.876
	Scopus	0.715	0.797	0.821	0.861
Civil	IEEE	0.756	0.804	0.847	0.903
	ACM	0.832	0.881	0.906	0.962
	Scopus	0.785	0.875	0.899	0.944
Computer	IEEE	0.746	0.793	0.835	0.892
	ACM	0.821	0.869	0.898	0.950
	Scopus	0.775	0.864	0.886	0.930
Mechanical	IEEE	0.736	0.783	0.807	0.880
	ACM	0.810	0.858	0.892	0.936
	Scopus	0.765	0.853	0.871	0.918

Field of Study	Data Set	TBM	CBMM	PTM	Proposed ETM-ILM
Electrical	IEEE	0.329	0.214	0.191	0.125
	ACM	0.317	0.178	0.160	0.116
	Scopus	0.412	0.380	0.358	0.260
Electronics	IEEE	0.325	0.211	0.202	0.123
	ACM	0.313	0.176	0.159	0.114
	Scopus	0.407	0.375	0.355	0.256
Civil	IEEE	0.357	0.232	0.213	0.136
	ACM	0.344	0.193	0.174	0.125
	Scopus	0.447	0.412	0.393	0.282
Computer	IEEE	0.352	0.229	0.213	0.134
	ACM	0.339	0.191	0.165	0.123
	Scopus	0.441	0.407	0.370	0.278
Mechanical	IEEE	0.348	0.226	0.185	0.132
	ACM	0.335	0.188	0.167	0.122
	Scopus	0.435	0.401	0.366	0.275

Table : Comparison of Entropy of existing methods Vs proposed Methods

This ANN based learning model is implemented using Neural Network Tool Box in MatLab (MatLab). In the training algorithm, the goal is assigned as "0.01" and the epoch is assigned as 250. Table '4.1' shows the %RMS error in the Estimation and Elimination of NBN, 2HLANN and the proposed learning model. The estimation error identifies a number of documents and our terms identified in the clustering model. The elimination error defines the mismatch ratio for document clustering. Comparisons of RMS error in estimation and elimination for proposed AJI learning model Vs existing models is shown in and also %error in estimation and elimination for proposed AJI learning model Vs existing models is shown The results are shown in Table and performance is shown in Figure, it is concluded that the performance of proposed AJI learning model always performs better than the existing methodology. the proposed AJI learns the synonymy better than the existing systems. From this, it is concluded that the proposed AJI performs better than existing systems. The AJI shows around 30% improvement in the estimation and around 23% improvement in the elimination.

The percentage RMS error in estimation is reached at 7.83% in NBN, 7.23% in 2HLANN whereas; it is only 4.60% in the proposed learning model. The percentage RMS error in elimination is reached at 5.15% in NBN, 8.65% in 2HLANNwhereas; it is only 4.75% in the proposed learning model.

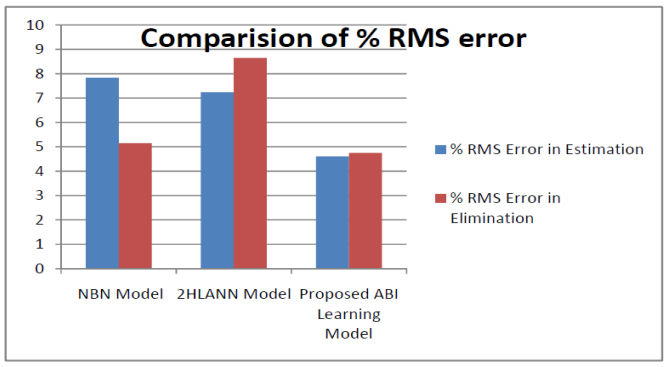

Fig : Comparison of % RMS error in Estimation and Elimination for proposed model vs. existing models

The proposed ABI learning method improved estimation, elimination and accuracy of the system. The estimation is improved around 25% than NBN and 33% than 2HLANN. Similarly the elimination is improved around 30% than NBN and 33% than 2HLANN. The accuracy of the proposed system also improved which is shown in the error rate and learning rate based on the approach.

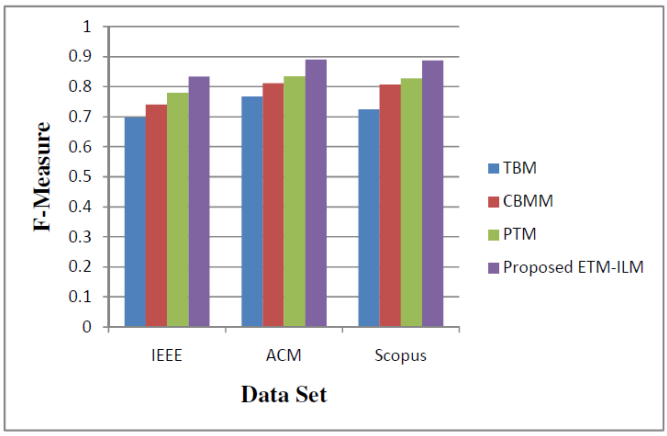

Fig : Comparison of F-Measure on Electrical data

F-Measure improvement in the proposed ETM-ILM method is seen from Table 5.4 in all the fields of study over other existing methods. The F-Measure of the proposed ETM-ILM is improved than TBM as a minimum of 16% than existing system and it leads to maximum of 23%. The F-Measure of the proposed ETM-ILM is improved than CBMM as a minimum of 8% than existing system and it leads to maximum of 13%. The F-Measure of the proposed ETM-ILM is improved than PTM as a minimum of 5% than existing system and it leads to maximum of 10%.

CONCLUSION

the research work focused data theoretical mining model which is proposed to solve synonymy problems in addition to basic requirements such as fastest processing, concept based knowledge extraction and interesting pattern. This research work is effective clustering using data theoretical mining model. The performance of DTMM is improved using ABI learning algorithm. The detailed design and methodologies are explained in the previous chapters. The convergence of the proposed ABI and existing learning models are compared and concluded that the proposed AJI provides optimal result within few iteration of training. The proposed AJI learns the synonymy better than the existing systems. From this, it is concluded that the proposed ABI performs better than existing systems. From the result and performance analysis, it is concluded that the proposed DTMM with AJI learning algorithm (ETM-ILM) is proved better result than existing and notable recent works in document clustering field of domain.

Future Work:

In our research all the technique has been implemented in MatLab and tested with small number of lesser file size documents. This algorithm can also be implemented in the middleware (embedded text mining) with large file size. This technique may be extended for text mining in very large text files considering block by block and combining their results. The theoretical problem, polysemy is yet to be solved effectively. This work may lead an interesting research in solving that problem also in near future.

REFERENCES:

1.Agrafiotis, D. K and H. Xu, "A Self-Organizing Principle for Learning Nonlinear Manifolds," Proc. Nat'l Academy of Sciences USA, Vol. 99, no. 25, pp. 15869-15872, 2002.

2. Ahmad, A., Dey, L., "A k-mean clustering algorithm for mixed numeric and categorical data", Data & Knowledge Engineering (Elsevier), Vol. 63, pp. 503–527, 2007.

3. Ahmed, M.A., Haidar, Azah Mohamed, Aini Hussain, and Norazila Jaalam "Artificial Intelligence application to Malaysian electrical power system", Journal of Expert Systems with Applications (Elsevier), Vol. 37, pp. 5023-5031, 2010.

4. Ali M. and Babak, A. "A new clustering algorithm based on hybrid global optimization based on a dynamical systems approach algorithm", Expert Systems with Applications (Elsevier), Vol. 37, pp. 5645-5652, 2010.

5. Andrew, K. and Wenyan, L. "Short-term prediction of wind power with a clustering approach", Renewable Energy (Elsevier), Vol. 35, pp. 2362-2369, 2010.

6. Andrew,S and Khaled,A., "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering, Vol. 25 , No. 1 , pp.62 – 75, 2013.

7. Association for Computing Machinery (ACM), available on web, dl.acm.org. Barros, R.C. , Basgalupp, M.P. ; de Carvalho, A.C.P.L.F. ; Freitas, A.A, "A Survey of Evolutionary Algorithms for Decision-Tree Induction" IEEE

8. Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 42 , No.3 , pp. 291 – 312, 2012.

9. Berry. M.W., Kogan,J., "Text Mining: Applications and Theory",wiley, 1st Edition, 2010.

10. Birant, D., Kut, A.,"ST-DBSCAN: An algorithm for clustering spatial–temporal data", Data & Knowledge Engineering (Elsevier), Vol.60 pp. 208–221, 2007.

11. Brown, S and Forouraghi, B., "Concept Classification using a hybrid data mining model", 21st International Conference on Tools with Artificial Intelligence, pp. 375-378, 2009.

12. Cai, D., He. X and Han. J, "Document Clustering Using Locality Preserving Indexing", IEEE Transaction on Knowledge and Data Engineering., Vol.17, No.12, pp.1624-1637, 2005.