International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 3 | ISSN : 2456-3307

Hybrid Intelligent Similarity Measure for Effective Text Document Clustering Using Neural Network Algorithm

R. Preethi[1], K.Selvi[2]
[1]P.G Student, R.M.K Engineering College, Kavaraipettai, Chennai, Tamil Nadu, India
[2]Research Scholar, Sathyabama University, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai, Tamil Nadu, India

ABSTRACT

Extensive use of World Wide Web for information search using popular search engines has turned many researchers to focus on text mining issues. Natural Language Processing required effective methods to capture the actual requirements of the user during Machine Learning. Application of genetic algorithm and similarity measure for text mining during document clustering yield significant results for WordSim353 data sets. Experiments show that application of Echo State Neural Network and Radial Basis Function to the training data set gives better clustering of text documents based on the stored weights in order to avoiding retrieval of irrelevant documents**.**

Keywords : World Wide Web, WordSim353, Clustering, Neural Network, Cyber terrorism investigation, Verb-Argument Structures, DIG, DIGBC, LSI, PCA, SVD, RFP, Textmining, Document Clustering Similarity Measure, Artificial Intelligence

I. INTRODUCTION

The concept of similarity measure in search engines is very beneficial and also being used in many applications. In order to retrieve the desired web documents based on the query, similarity measure plays an important role. The nature of the documents is different, it can be either structured or semi structured which causes the difficulty in handling. Therefore, use the concept of clustering to improve the retrieval methodology. It is a systematic way of arranging the documents that need to be provided to the users based on the query. It can be done by composing based on contents and links [12]. Cyber terrorism investigation [9], topic spotting, email routing, language guessing are some of the applications where most of the researches being done. Pre-processing and document similarity analysis are done in order to improve the accuracy and classification. In pre-processing, to find the concepts from a set of words as features is done using Verb-Argument Structures [3].In some research areas, bag-of-words [19] is found from the text documents. This huge set of words need to be reduced using feature clustering [20] [23] methods. The resultant is further examined for the document similarity [22] [23] and documents are clustered if they are similar.

Many fuzzy similarity based models and algorithms have been introduced with the very nature of its membership functions [22] [23], fuzzy association [19] [23], fuzzy c-means, production rules [17] [2]. The performance of an information system depends on the algorithms and the assessment of the performance resides in the measures applied that could be carried out effectively. The user will always like exact and perfect web pages to the query given to the system. The exact match is possible to determine whether the value of a given term matches the value specified in the query which requires a measure to satisfy the information need. Hence our key objective is to develop a good approximate match to the query.

The rapid growth of World Wide Web has imposed challenges to cluster the documents over the internet and thereby improving the efficiency. Search engines are facing difficulty in organizing the relevant documents among huge volumes of search results returned to a simple query. We require an effective method to solve the problem by clustering the documents that are similar, which helps the user in identifying the relevant data easily [6].

We surveyed on related works and mentioned in Section 2, followed by briefing the steps involved in various phases of our proposed method in Section 3. This is followed by a mathematical analysis of the results obtained during the training process for Word-Sim353 data sets, where the human ratings of the cosine similarity measures are obtained using Kardi resources [27]. Section 4 presents the results obtained from Weka tool on application of the data mining techniques for the results obtained in the training process. Finally, in Section 5 we conclude with the summary, highlighting the drawbacks and directions for future work.

Document clustering is important for quick accessing of relevant documents in the web for a given pair of words. The time taken for retrieving the documents should be in less than seconds so that the user can verify whether the documents are relevant to the query. Software and algorithm are being used by search engines which can satisfy the query from the users and return relevant documents correctly in time. Therefore, more efficient new document clustering algorithms are required than conventional clustering algorithms [17].
Ling Zhuang Honghua Dai 2004 introduced the initial points for k-means algorithm as random centers. This clustering approach is difficult for interpretation as it is unstructured and noisy [14].

Benjamin Fung, et al., 2003[3], has introduced a method for clustering the documents. A tree is constructed based on the topics and similarity is generated among clusters . Sharma et al., 2009[21] has introduced this approach for large wordsets.

Document index graph based document clustering is put forth by Momin et al., 2006[10]. Document clustering methods are based on a single term examination of document data set. Document Index Graph (DIG) allows documents to be encoded using phrases. It focuses on improving phrase-based similarity measure. Further, a Document Index Graph based Clustering (DIGBC) algorithm is also being proposed. It incrementally form clusters based on the cluster-document similarity measure and the documents can be assigned to more than one cluster.
Muflikhah et al., 2009[11], introduced space and cosine similarity measurement. The authors used Latent Semantic Index (LSI) approach with Principle Component Analysis (PCA) or Singular Vector Decomposition (SVD). The method decreases the matrix dimension by identifying the pattern in the document collection which refers to simultaneous terms. Every technique makes use of frequency of the term as weight in Vector Space Model (VSM) with the help of fuzzy c-means technique for clustering. Web document clustering on affinity-based similarity measure is presented by Shyu et al., 2004[22]. The concept is enhanced to be used in web document clustering by establishing the approach of affinity based similarity measure, uses access patterns to find the similarities through a probabilistic model. Various experiments are analyzed based on real time dataset, and the experimental results illustrated that the presented similarity measure outperforms Euclidean distance and cosine coefficient technique under various document clustering techniques.

Eldesoky et al., 2009[5], given a novel similarity measure for document clustering based on topic phrases. The latest trend which uses the phrase to be more informative feature and taking into account the issue that led in extending the performance of document clustering. This paper used a method for analyzing the similarity measure of VSM by taking into account the topic phrases of the document and applying it to the Buckshot technique. These methods increment the values of metric in order to improve clustering effectiveness.

Cobos et al., 2010[4] used k-means, term sets, Bayesian information for web document clustering. Haojun et al., 2008[11] developed hierarchical algorithms for document clustering by using cluster overlapping rate to improve efficiency. Sub clusters (say two) are merged if they have high overlapping rate. In order to number the parameters, this paper used Gaussian mixture as a model in Expectation – Maximization techique.

Thanh Van Le et al., 2013[26] proposed a method for multi variables. This approach makes use of pre-topological way of clustering the documents without any distance measurements.
Narayanan et al., 2013[11**]** has introduced a distributed algorithm for enhancing the performance of document clustering using parameters such as quality and accuracy. This paper also explains various similarity measures such as cosine similarity, Jaccard and Pearson coefficient for comparison.

A text based clustering approach using novel similarity measure is presented by Reddy et al., 2014[16] introduced a measure to determine the commonality which is used as a similarity measure between two text files. In order to reduce the dimensionality of input (text file) this paper uses frequent item set algorithm on the initial set of files.

Karthick et al., 2014[8] examined the usage of multi-word frequency to determine the clustering performance on the hierarchical clustering technique. However, the combining effect of link and content based representation of web documents and their interrelationship are analyzed.

Peipei Li et al., 2015[14] proposed an algorithm to compute term similarity. This approach makes use of concept space by mapping the terms and then their similarity is compared in a semantic network. Modified k-medoids technique is being used. Select the center for each cluster based on the criteria of minimum of distances and then maximum of it is used from each existing centers. Assign concepts to their corresponding clusters. The iteration stops when a minimum objective function is reached. This method requires prior specification of number of clusters.

II. MATERIALS AND METHODS

Pre-processing tasks need to be performed before data mining algorithms are used on the data set. These include data cleansing, session identification, path completion, and formatting. It is done in order to improve the quality of the results. This section discusses various phases of the proposed method as shown in Figure 1.

Documents are files which contain ASCII and non ASCII characters. Sample categories are thesis, paper, journal, Invoice, quote, RFP, Proposal, Contract, Packing slip, Manifest, Report detailed & summary, Spread sheet, Waybill, Bill of Lading, Financial statement, Nondisclosure agreement, Mutual nondisclosure agreement, summons, certificate, license, gazette, white paper, application forms, user-guide, brief, mock-up, script. These documents can be created using templates. The documents can be computerized by using different types of editors starting from basic editors which use only ASCII characters to sophisticated editors which embed graphics. The graphics in a document is represented by special characters.

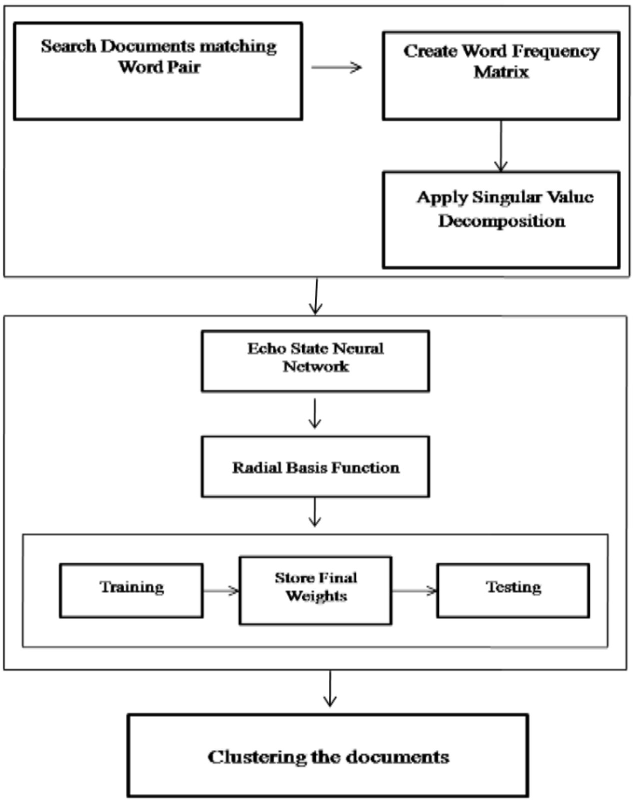

Search Documents matching Word Pair
Create Word Frequency Matrix
Apply Singular Value Decomposition
Echo State Neural Network
Radial Basis Function
Training
Store Final Weights
Testing
Clustering the documents

Figure 1. System Architecture

Documents are files which contain ASCII and non ASCII characters. Sample categories are thesis, paper, journal, Invoice, quote, RFP, Proposal, Contract, Packing slip, Manifest, Report detailed & summary, Spread sheet, Waybill, Bill of Lading, Financial statement, Nondisclosure agreement, Mutual nondisclosure agreement, summons, certificate, license, gazette, white paper, application forms, user-guide, brief, mock-up, script. These documents can be created using templates. The documents can be computerized by using different types of editors starting from basic editors which use only ASCII characters to sophisticated editors which embed graphics. The graphics in a document is represented by special characters.

Searching of similar documents involve processing of documents where elimination of information should be avoided taking care of word pair search. When the documents are properly preprocessed then the measures of similarity will be perfect, and the quality of search will be maximum. The documents should first be preprocessed. Each document can be converted into a vector D. Each vector contains more than two dimensions. Each dimension of a vector is a unique representation of a document. When the contents of two documents are almost similar, then the numerical values of the vector are also being almost same. Each feature of the vector will represent a distinct term. The term is a single word or phrase.

2.1 Preprocessing

Step 1. Pair of words is given in the search area.

Step 2. Each document is preprocessed and converted into vectors of numerical values. If the vectors of numerical values are already available in the searching folder corresponding to the available documents, then preprocessing of the documents and converting into vectors need not be carried out.

Step 3. The bags of representation of words (distinct single words) is created. During this process, irrelevant data are omitted for calculation purposes.

Step 4. Indexing the bags of words is done. In this process, tokens are created by segmenting strings by white space and punctuations called Tokenization. Each token is stemmed into its root form by converting a noun into its singular form and removing grammatical words like articles, conjunctions, pronouns called Stop Word Elimination.

Step 5. Converting vectors into a matrix is formed. In this matrix presence of absence of words corresponding to a document is indicated using 1 or 0. On the other hand, frequency of words is represented in fractions in the matrix by normalization.

2.2 Computing ANN algorithm

The matrices generated above are used as input for computing Echo State Neural Network algorithm (ESNN). During the training process, various parameters of the network are optimized and as a result of which the network passes through a learning phase. The types of algorithms that are used for training the ANN topology are Radial Basis Function (RBF) and Echo Sate Neural Network (ESNN).

Echo State Neural Network can best fit in multi-dimensional space after ensuring the best match to the training data. This unique feature enables it to handle problems related to classification and Clustering. The spline function applied to the training data set is:

$$f(x) = x^2 \log(x) \tag{1}$$

where x represents nodes of the network which is given as input to the neural network algorithm. Echo State Network provides better clustering abilities for a newly created data set with a complete cluster specification. Gradually, a weighed matrix is generated using echo state property after the training process and the whole training set is passed through the network.

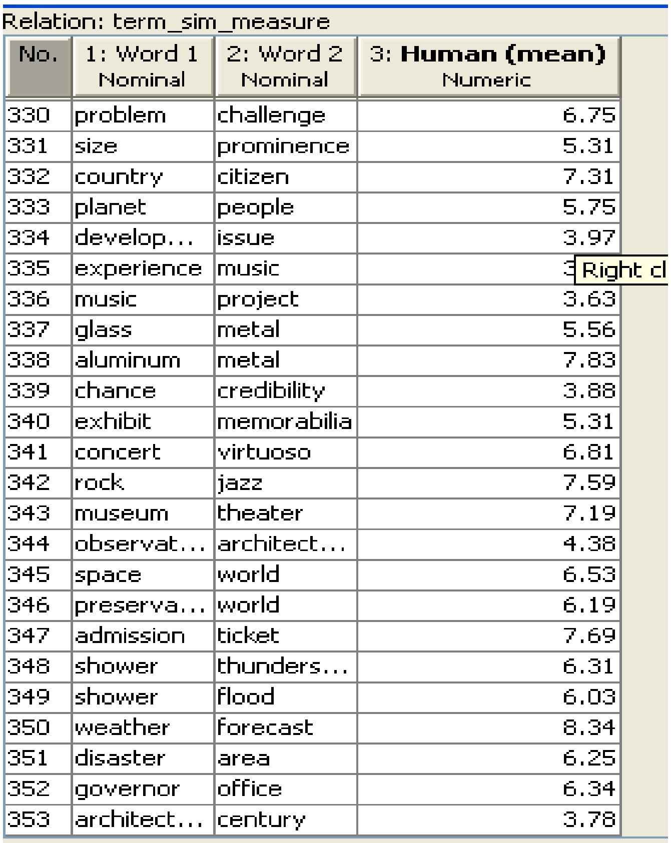
In this algorithm, training patterns with input features representing a document is used. Radial basis function uses exponential function as its activation values in the hidden layer of the ANN topology. Centre patterns are created from the training patterns. The summation of the distance between each training pattern and all the centre patterns are found. The summed values are passed over an activation function to obtain RBF output in the hidden layer. A bias value of '1' is used in the hidden layer output. Similarly, for all the remaining training patterns, the RBF outputs are obtained as the outputs in the hidden layer. All these RBF outputs from all the training patterns are further processed with the assigned target values to obtain a set of final weights.
The algorithms are developed based on different weight updating rules. In each weight updating rules, errors are calculated in the forward process of the ANN and weight updation is done during the reverse or recurrent process. In the training process of the ANN algorithms, the connections among the nodes between layers are represented by matrices. In most of the algorithms, the matrices are initialized with random numbers. At the end of the training process, the matrices contain final weight. During the testing of the ANN or testing the retrieving of the documents corresponding to word pair, final weights are used for processing with the vector corresponding to word pair. In another method, final weights are obtained from the pattern itself without any initialization of the weight matrices.

2.3 Clustering The documents

Based on the stored final weights, documents are clustered using Expectation-Maximization. This is done using weka tool. It produces a maximum number of fifteen clusters in order to perform effective retrieval of documents.

III. RESULTS AND DISCUSSION

Figure 2 shows the word pairs with human ratings. These words are found to be semantically similar in most of the documents. This similarity is used for constructing the similarity matrix. Generally, it contains 1.0s along the diagonal .Even though two words are found to be identical; they may differ in their meaning but their value remains the same (i.e. 1.0) which is contradictory that can be considered for further research in the area of semantic similarity.



Relation: term_sim_measure

No.	1: Word 1 Nominal	2: Word 2 Nominal	3: Human (mean) Numeric
330	problem	challenge	6.75
331	size	prominence	5.31
332	country	citizen	7.31
333	planet	people	5.75
334	develop...	issue	3.97
335	experience	music	3 Right cl
336	music	project	3.63
337	glass	metal	5.56
338	aluminum	metal	7.83
339	chance	credibility	3.88
340	exhibit	memorabilia	5.31
341	concert	virtuoso	6.81
342	rock	jazz	7.59
343	museum	theater	7.19
344	observat...	architect...	4.38
345	space	world	6.53
346	preserva...	world	6.19
347	admission	ticket	7.69
348	shower	thunders...	6.31
349	shower	flood	6.03
350	weather	forecast	8.34
351	disaster	area	6.25
352	governor	office	6.34
353	architect...	century	3.78

Figure 2. Word pairs and their similarity scores

Table 1 is given below which contains sample word pairs experimented by us from WorSim353 data set. Correlation coefficient and various error rate is calculated using weka tool. Various investigations found that the human scoring has consistently high correlations. In both RG and WordSim353, the confidence levels are large and significant in their difference.

Table 1. Error Rates with Correlation Coefficient

Average Target Value : 0.5758305038549534	
Inverted Covariance Matrix: Lowest Value = -0.34971644612476366 Highest Value = 0.6502835538752364	
Inverted Covariance Matrix * Target-value Vector: Lowest Value = -0.2588713208324032 Highest Value = 0.27371229405845615	
Time taken to build model: 0.66 seconds Scheme: weka.classifiers.functions.LinearRegression	
Correlation coefficient	0.9632
Mean absolute error	0.6996
Root mean squared error	0.87
Relative absolute error	39.0574 %
Root relative squared error	40.0484 %
Coverage of cases (0.95 level)	100 %
Total Number of Instances	353

This proposed approach is suitable for any context, as query processing does not require scheme of classification. Table 2 summarizes the error rates for 353 classified instances for the given word-sim353 data. Linear regression classification algorithm yields an average value of 0.58 approximately with the absolute error rate of 40 %.

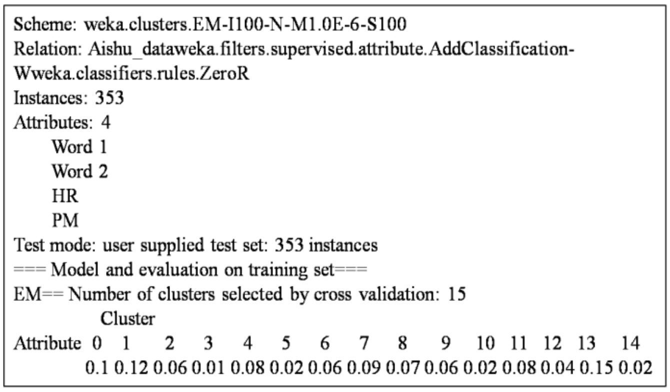
Table 2 exhibits the training results acquired on the application of RBF using Weka tool for 5 iterations. Further indices shows that the calculated scores would allow us a comparison for the taxonomy based glosses.

Table 2. Training the Dataset Using RBF

Scheme: weka.classifiers.functions.RBFNetwork -B 2 -S 1 -R 1.0E-8 -M -1 -W 0.1
Relation: Data-weka.filters.supervised.attribute.AddClassification-Wweka.classifiers.rules.ZeroR
Instances: 353 Attributes: 4 Test mode :10-fold cross-validation
Radial basis function network
(Linear regression applied to K-means clusters as basic functions):
Linear Regression Model SM = -2.1219 * pCluster_0_0 + 2.1212 * **pCluster_0_1 + 5.5318**
Time taken to build model: 0.08 seconds

Correlation coefficient Mean absolute error Root mean squared error Relative absolute error Root relative squared error Total Number of Instances	0.9146 0.7058 0.9078 39.198% 41.633% 353	Number of iterations=5 Within cluster sum of squared errors = 682.37397

Figure 3 shows the documents containing the required word pairs are clustered using Expectation-Maximization method. It makes use of 353 instances for testing attributes. It produces a maximum of fourteen numbers of clusters in order to perform effective retrieval of documents.



```
Scheme: weka.clusters.EM-I100-N-M1.0E-6-S100
Relation: Aishu_dataweka.filters.supervised.attribute.AddClassification-
Wweka.classifiers.rules.ZeroR
Instances: 353
Attributes: 4
      Word 1
      Word 2
      HR
      PM
Test mode: user supplied test set: 353 instances
=== Model and evaluation on training set===
EM== Number of clusters selected by cross validation: 15
           Cluster
Attribute 0   1    2    3    4    5    6    7    8    9    10   11   12   13   14
         0.1 0.12 0.06 0.01 0.08 0.02 0.06 0.09 0.07 0.06 0.02 0.08 0.04 0.15 0.02
```

Figure 3. Cluster documents usingExpectation-Maximization

IV. CONCLUSION

This paper emphasizes the applicability of artificial neural networks in clustering the documents.The major three phases specified in this paper have to be considered for clustering the document using neural network algorithms. Experiments conducted on 353 word pairs show that the proposed method outperform. The outputs of ANN can be as well interpreted whether the documents retrieved or clustered are relevant to the words. Future work includes the application of other ANN algorithms that can be implemented for clustering the similar documents. The limitation is that it cannot make distinction between primary and secondary categories.

Ouyang, D., J. Bartholic and J. Selegean, 2005. Assessing Sediment Loading from Agricultural Croplands in the Great Lakes Basin. Journal of American Science, 1(2): 14-21.

V. REFERENCES

[1] F. Beil, M. Ester, and X. Xu, "Frequent term-based text clustering", In Proceedings of 8th International Conference on Knowledge Discovery and Data Mining, 2002.

[2] A. Budanitsky, and G. Hirst, "Evaluating wordnet-based measures of semantic distance", Comput. Linguistics, vol. 32, no. 1, pp. 13–47, 2006.

[3] C.M. Benjamin Fung, Wang Ke, and Ester Martin, "Hierarchical document clustering using frequent item sets", In Proceedings SIAM International Conference on Data Mining, pp. 59-70, 2003.

[4] C. Cobos, J. Andrade,W. Constain., M. Mendoza., and E. Leon, "Web document clustering based on global-best harmony search, k-means, frequent term sets and bayesian information criterion", IEEE Congress on Evolutionary Computation, pp. 1-8, 2010.

[5] A.E. Eldesoky, M. Saleh, and N.A. Sakr, "Novel similarity measure for document clustering based on topic phrases", International Conference on Networking and Media Convergence, pp. 92-96, 2009.

[6] Haojun Sun, Zhihui Liu, and Lingjun Kong, "A document clustering method based on hierarchical algorithm with model clustering", 22nd International Conference on Advanced Information Networking and Applications, pp. 1229-1233, 2008.

[7] Han-Saem Park, Si-Ho Yoo, and Sung-Bae Cho, "Evolutionary Fuzzy Clustering Algorithm with Knowledge-Based Evaluation and Applications for Gene Expression Profiling", Journal of Computational and Theoretical Nanoscience, vol. 11, no. 4, pp. 524-53, 2005.

[8] S. Karthick, S.M. Shalinie, A. Eswarimeena, P.Madhumitha, T.N. Abhinaya, "Effect of multi-word features on the hierarchical clustering of web documents", Recent Trends in Information Technology(ICRTIT) Internaltional Conference, pp. 1 – 6, 2014.

[9] Ling Zhuang, and Honghua Dai, "A maximal frequent item set approach for web document clustering", In Proceedings of the IEEE Fourth

International Conference on Computer and Information Technology, 2004.

[10] B.F. Momin, P.J. Kulkarni, and A. Chaudhari, "Web document clustering using document index graph", In Proceedings IEEE International Conference on Advanced Computing and Communications, 2006.

[11] L. Muflikhah, and B. Baharudin, "Document clustering using concept space and cosine similarity measurement", International Conference on Computer Technology and Development, vol. 1, pp. 58-62, 2009.

[12] N. Narayanan, J. E. Judith and J. JayaKumari, "Enhanced distributed document clustering algorithm using different similarity measures", Information & Communication Technologies (ICT), IEEE Conference, pp. 545-550, 2013.

[13] H.A. Nguyen, and H. Al-Mubaid, "New ontology-based semantic similarity measure for the biomedical domain", In Proc. IEEE GrC, pp. 623–628, 2006.

[14] Peipei Li, Haixun Wang, K.Q. Zhu Zhongyuan Wang, Xuegang Hu and Xindong Wu, "A large probabilistic semantic network based approach to compute term similarity", IEEE Transaction on Knowledge and Data Engineering, vol. 27, pp. 2604-2617, 2015.

[15] J. Prasannakumar, and P. Govindarajulu, "Duplicate and near duplicate documents detection", A Review European Journal of Scientific Research ISSN 1450-216X vol. 32, no. 4, pp. 514-527, 2009.

[16] G.S. Reddy, T.V. Rajinikanth, A.A. Rao, "A frequent term based text clustering approach using novel similarity measure", Advanced Computer Conference (IACC), IEEE International, pp. 495-499, 2014.

[17] Ruxixu and Donald Wunsch, "A survey of clustering algorithms", IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645-678, 2005.

[18] S. Satwardhan, Incorporating dictionary and corpus information into a context vector measure of semantic relatedness, Master's thesis, Univ. Minnesota, Minneapolis, 2003.

[19] K. Selvi, and R.M. Suresh, "Context similarity measure using fuzzy formal concept analysis", In Proc. of The Second Int'l conference On Computer Science and Engineering and Information Technology CCSEIT, pp. 416-423, 2012.

[20] K. Selvi, and R.M. Suresh, "An efficient technique to implement similarity measures in text document clustering using artificial neural network algorithm", Research Journal of Applied Sciences Engineering and Technology, vol. 8(23), pp. 2320-2328, 2014.

[21] A. Sharma, and R. Dhir, "A wordsets based document clustering algorithm for large datasets", In Proceeding of International Conference on Methods and Models in Computer Science, 2009.

[22] M.L. Shyu, S.C. Chen, M. Chen, and S.H. Rubin, "Affinity-based similarity measure for web document clustering", IEEE International Conference on Information Reuse and Integration, pp. 247-252, 2004.

[23] K.M. Sim, and P.T. Wong, "Toward agency and ontology for web-based information retrieval", IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 34, no. 3, pp. 257–269,2004.

[24] Y. Syed Mudhasir, and J. Deepika, "Near duplicate detection and elimination based on web provenance for efficient web search", In the Proceedings of International Journal on Internet and Distributed Computing Systems, vol. 1, no. 1, pp. 22-32, 2011.

[25] Ted Pedersen, V.S. Serguei. Pakhomov, Siddharth Patwardhan, and Christopher G. Chute, "Measures of semantic similarity and relatedness in the biomedical domain", Journal of Biomedical Informatics, vol. 40, pp. 288-299, 2007.

[26] Thanh Van Le, Trong Nghia, Hong Nam Nguyen, Tran Vu Pham, "An efficient pretopological approach for document clustering", Intelligent Neyworking and Collabrative Systems (INCoS), 5th International Conference , pp. 114 – 120, 2013.

[27] http://people.revoledu.com/kardi/tutorial/Similarity/Stringinstance.html#TextSimilarityCalculator.

[28] Xinjuan Peng, Lijun Cai, Bo Liao, Haowen Chen, and Wen Zhu, "Detecting the Maximum Similarity Bi-Clusters of Gene Expression Data with Evolutionary Computation", Journal of Computational and Theoretical Nanoscience, vol. 11, no. 7, pp. 1585-1591, 2014.