An Enhancement of Clustering Technique using Support Vector Machine Classifier

Mehajabi Sayeeda
M. Tech. Scholar CSE,
TIT, Bhopal

Rachana Kamble
Asst. Prof. CSE Dept.,
TIT, Bhopal

ABSTRACT
Web surfing is very essential task of daily life for any professional person they search information regarding their field. But to get exact required information from ocean internet of data have become complex task. To manage files and information properly document clustering is a good approach. Clustering method divides text information into subgroup on basis of content based similarity. Document clustering reduces searching effort and fulfils human interest information looking for. It groups similar files together to minimize the search time and complexity. This paper gives new clustering method based on hybrid XNOR function to find degree of similarities within any two documents. Resultant similarity used for document clustering by applying SVM classifier for learning network. This paper introduces new method for document clustering by use of similarity matrix calculation and this matrix is passed for training SVM network for upcoming document classification. The results show the effectiveness of proposed work.In this paper, we describe the formatting guidelines for IJCA Journal Submission.

Keywords
Hybrid XNOR, SVM classifier, learning network, Document clustering, similarity matrix.

1. INTRODUCTION
Development of web application makes today's life very straightforward. Everybody likes to look info exploitation web. To provides user a decent navigation and browsing facilities. For this reason to arrange documents in cluster manner is nice technique to produce looking document quick access. Therefore clustering schemes are wide studied in last decade. In grouping strategies all teams of document are allotted to a special topic. The concept of grouping is such scheme that is evolved basic ideas like information representation schemes by using general idea such as data presentation methods, similarity values, cluster formation methods and algorithms used for cluster formation. Vector Space Document (VSD) method is used widely for creating documents cluster. The main concept of this method is it presentation any text file documents whatever the document is either webpage or document file. These files are stored collectively and converted as feature array collection of different keywords contain by document of given dataset. All keywords of any document generally smallest term of any document so it is used as feature of that document by VSD method. In normal methods weights of similar words are measured by term frequency (tf) and inverse document frequency (idf) of common words. The similarity membership worth inside any two documents are often derived from calculative anyone of the various similarity measures by apply the two corresponding feature values, for example cosine similarity or Jaccard relationship function live or euclidian similarity cost

The problem of finding frequent thingsets is first initiated from [1] that use frequent items to search out association rules in massive transactional databases. Cluster can be defines as [2] collection of similar document which having similar text contain documents set. In [3] the authors propose a way for locating large size of frequent word item collections. As given in [4], text files or documents can be classify by considering Gaussian membership function and creating use of it to get clusters by finding word patterns. every cluster is known by its word pattern calculated applying fuzzy primarily based Gaussian membership functions once clusters square measure formed.

In this paper the thought is to first get frequent item sets for every document by applying existing association rule finding techniques wherever according vertical or horizontal scheme. If frequent thing sets discovered in every document then we need to create such matrix contains only Boolean values in which documents are stored in row wise and frequent items are stored in column wise for every document. this is often followed by the computation of a ternary feature vector for every document set, drawn as a second array or second matrix by redefining the XNOR mechanism as hybrid XNOR logic with slight modification within the perform introducing high impedance variable as Z. the thought of most capturing is taken because the base framework for cluster [5]. The authors perform cluster by applying XOR similarity perform [6].

2. PROBLEM FORMULATION
Documents are collected as dataset in form of array {d1,d2….dN} with N documents in set, there's got to sub-group [2] the documents supported the semantic of the text contents present during a Document , assuming to need K such sub-groups, the clustering method generates C={c1,c2,…ck} clusters, with every ci being non empty.

Document clustering continues to be developing process that creates group of likely documents together. Firstly normal vector derived where document are represented as collection of words and clustering method compare common word to find similar documents Many modifications were applied on this methodology to enhance this methodology because the result set would solely offer us info on what words were present during a cluster of documents, it is not about any particular information or text of documents. Here is a requirement of additional intuitive ways that of clustering that may offer us sound information of the content present within the documents.

3. RELATED WORK
Clustering is principally established by 3 factors throughout generation of excellent quality of cluster. They're information illustration, similarity operates and agglomeration mechanism that is applied [3], [5], [8], [9]. Vector house model is standard technique for information illustration utilized in document agglomeration. During this technique all documents collected in kind assortment of vectors wherever every vector

may be document information. Within document every keyword may be define by its TF and IDF cost.

Equality of any two document combine is calculated by jaccard operate [4], [10] similarity live. during this technique solely word represents whole document thus not finds several valuable data from word proximity [5]. Alternative strategies supported Vector house model additionally not think about progressive process [1]. Survey of net agglomeration engines that progressive process will increase the effectiveness whereas helpful within the agglomeration schemes [11]. The foremost connected work that takes into consideration the knowledge concerning proximity of words ANd phrase primarily based analysis in an progressive means is Suffix Tree agglomeration (STC) [12].

Frequent item sets finding downside is make a case for in elaborate paper [13]. Ordinarily frequent things ar derived by association rules mining. However as technology growing this technique additionally capable to calculate frequent item kind documents and additionally for documents agglomeration and alternative mining works. Paper [14] mentions technique for agglomeration documents by scheming neighbor operate worth from given document set. Another technique for given in paper [15] for those word set which appears most of the time in document. File categorization is additionally done by Gaussian membership worth between documents. This worth is useful for agglomeration.

In [6], the categorization of documents is complete by giving Gaussian relationship and creating use of it to attain clusters by discovery term. Each cluster is outlined by its term behavior notice by Fuzzy Gaussian relationship if cluster created. a completely unique technique referred to as most Capturing is projected for computer file agglomeration in [16]. Most Capturing concerned 2 actions as call document clusters and giving cluster members. In [17], algorithmic program to look for a pattern in an exceedingly text is projected which might be wont to look for part of interest within the part repository.

Hierarchical agglomeration is preferred over non-hierarchical agglomeration as a result of in non-hierarchical agglomeration a central purpose, additionally referred to as center of mass, required to be chosen arbitrarily and also the distance from that time is calculated to cluster documents (with less distance) in one cluster[18]. Finding this central purpose poses a huge challenge. That is why non hierarchal methods are not extraordinarily in style. A comparative work on each the strategies is finished by [19. Florian Beil et. al. introduced 2 agglomeration algorithms FTC(non-hierarchical) and HFTC(hierarchical) in [19], supported the thought of frequent Term Set and analyzed their behavior.

They used the association rule mining to spot the frequent terms in documents to cluster them into clusters. Agglomertive stratified agglomeration (AHC) may be a wide used buttom up agglomeration algorithmic program. several researchers have with efficiency used this technique in an exceedingly data retrieval and information and net agglomeration [20].

Recent analysis has developed new strategies for estimating the linguistics distance between each terms and words. Rudi L. Cilibrasi et. al. introduced a replacement similarity, referred to as Normalized Google Distance(NGD) [20], to effectively capture the linguistics similarity between words and phrases supported data distance and Kolmogorove quality. Later on, Alberto J Evangelista et. al. reviewed the work of Rudi L. Cilibrasi to spice up their distance operate through elimination of random data[21]. this method to estimate the similarity between among terms clusters instead of simply 2 words. [22].

4. PROPOSED WORK

The proposed work based on similarity matrix calculation this matrix will used for final training and testing of SVM classifier. For similarity XNOR function is used to create similarity feature vector within any pair of two documents.

The textual information of any document is scan by software to classified documents. The similarity relation S can be given in Three states if document A is same as B then this relation returns output 1 if both documents having both different word than relation results 0 and if both documents doesn't has any common word then relation gives output Z. This method is given in the truth Table 1.

Table 1. Truth Table Example Of XNOR Similarity Calculation

A	B	S(A,B)
0	0	Z
0	1	0
1	0	0
1	1	1

The algorithmic program for document cluster has its given as text files with recurrent keywords and resultant as group of clusters dynamically. The approach followed could be a tabular approach. equally the algorithmic program for part cluster has its input as software package elements with properties predefined and therefore the output could be a set of extremely cohesive elements with low coupling feature.

4.1 Algorithm for Clustering

It may be used for software package part cluster, document cluster or pattern cluster normally overall block diagram of proposed algorithm is given in figure no. 1 and flowchart can be easily understand by figure no. 2.

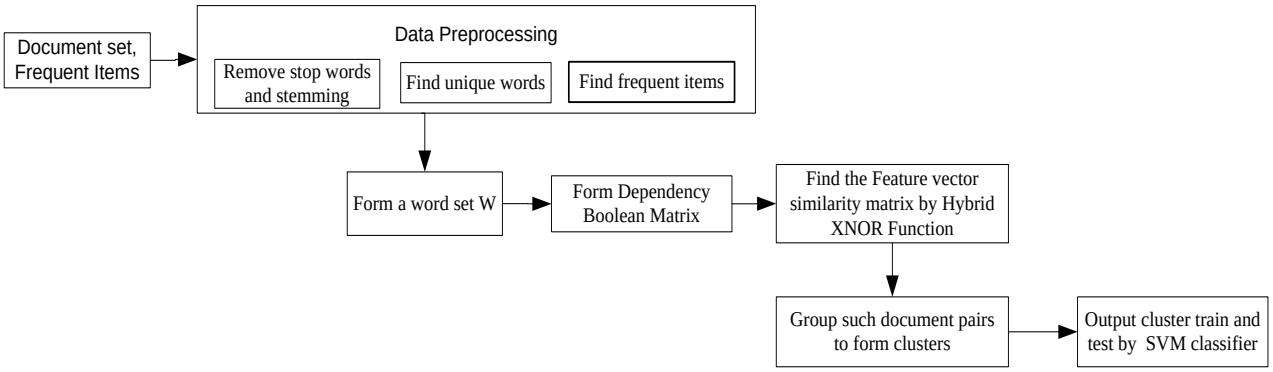

Fig 1: Proposed Work Flowchart.

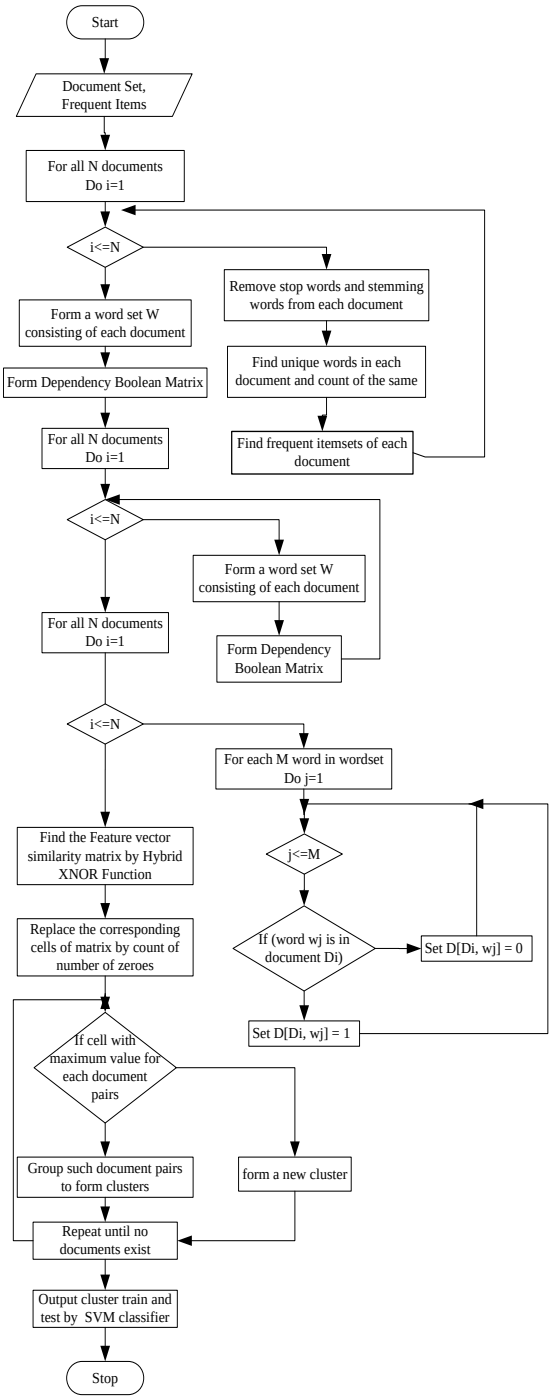

Fig 2: Proposed Work Flowchart.

Input: Document set, frequent items.

Output: set of clusters.

Begin of Algorithm

Step1: For each document D do

Begin

Step1.1 Remove stop words and stemming words from each document.

Step1.2 Find unique words in each document and count of the same.

Step1.3 Find frequent itemsets of each document

End for

Step 2: Form a word set W consisting of each word in frequent item sets of each document.

Step 3: Form Dependency Boolean Matrix with each row and column corresponding to each Document and each word respectively

For each document in document set do

Begin

For each word in word set do

Begin

If (word wk in Word set W is in document Di)

Begin

Set D[Di, wk] = 1

Else

Set D[Di, wk] = 0

End if

End for

End for

Step 4: Find the Feature vector similarity matrix by evaluating similarity value for each document pair applying

Hybrid XNOR Function defined in table 1 to obtain the matrix with feature vectors for each document pair.

Step 5: Replace the corresponding cells of matrix by count of number of zeroes in tri state feature vector.

Step 6: At each step, find the cell with maximum value and document pairs containing this value in the matrix. Group

such document pairs to form clusters. Also if document pair (X,Y) is in one cluster and document pair (Y, Z) is in another cluster, form a new cluster containing (X, Y, Z) as its elements.

Step 7: Repeat Step6 until no documents exist or we reach the stage of first minimum value leaving zero entry.

Step 8: Output the set of clusters obtained.

Step 9: Label the clusters by considering candidate entries.

End of algorithm

4.2 Working of proposed work

To explain proposed method well here an example with document sets having only frequent item sets. This frequent item sets are collected by applying association rule mining as mentioned.

Table 2. Documents and Corresponding Frequent Item Sets Frequent Item Sets

Documents	Frequent Itemsets
Document 1	{Viral, Blood, Infection }
Document 2	{Clinical, Viral, Blood }
Document 3	{Viral, Virus, Infection }
Document 4	{Clinical, Blood, Infection }
Document 5	{Blood, Virus}
Document 6	{Clinical, Virus, Infection }
Document 7	{Clinical, Viral, Virus }
Document 8	{Blood, Virus, Infection }
Document 9	{Clinical, Viral, Infection }
Document 10	{Viral, Virus, Infection }

This item set help to create Boolean matrix which represents each documents in rows and unique frequent items stored in column of respective rows for all documents present in document set.

Table 3. Boolean Matrix Representation of Table.2

	clinical	viral	blood	virus	infection
D1	0	1	1	0	1
D2	1	1	1	0	0
D3	0	1	0	1	1
D4	1	0	1	0	1
D5	0	0	1	1	0
D6	1	0	0	1	1
D7	1	1	0	1	0
D8	0	0	1	1	1
D9	1	1	0	0	1
D10	0	1	0	1	1

Table 4. Feature Vector Representation Of Document Set

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D1	×	{0,1,1,Z,0}=2	{Z,1,0,0,1}=2	{0,0,1,0,1}=2	{Z,0,1,0,0}=1	{0,0,0,0,1}=1	{0,1,1,0,0}=2	{0,0,1,Z,0}=1	{0,1,0,Z,1}=2	{0,1,0,0,0}=1
D2	×	×	{0,1,0,0,0}=1	{1,0,1,Z,0}=2	{0,1,0,Z,0}=2	{1,0,0,0,0}=1	{0,1,1,0,0}=2	{0,0,1,0,0}=1	{1,1,0,0,0}=2	{0,1,0,0,0}=1
D3	×	×	×	{0,0,0,0,0}=0	{Z,0,0,1,0}=1	{0,0,Z,1,1}=2	{0,1,Z,1,0}=2	{Z,0,0,1,1}=2	{0,1,Z,0,1}=2	{Z,1,Z,1,1}=3
D4	×	×	×	×	{0,Z,1,0,0}=1	{1,Z,0,0,Z}=1	{1,0,0,0,0}=1	{0,Z,1,0,1}=2	{1,0,0,Z,1}=2	{0,0,0,0,1}=1
D5	×	×	×	×	×	{0,Z,0,1,0}=1	{0,0,0,1,Z}=1	{Z,Z,1,1,0}=2	{0,0,0,0,0}=0	{Z,0,0,1,0}=1
D6	×	×	×	×	×	×	{1,0,Z,1,0}=2	{0,Z,0,1,1}=2	{1,0,Z,0,1}=2	{0,0,Z,1,1}=2
D7	×	×	×	×	×	×	×	{0,0,0,1,0}=1	{1,1,Z,0,0}=2	{0,1,Z,1,0}=2
D8	×	×	×	×	×	×	×	×	{0,0,0,0,1}=1	{Z,0,0,1,1}=2
D9	×	×	×	×	×	×	×	×	×	{0,1,Z,0,1}=2

The finally form a matrix D [n-1, n]. This matrix is collection of all n documents and stores values only in upper triangular indexes. Every index of upper triangular of matrix hold similarity values with respect to each other documents. This similarity value known as feature vector for current document.

After this feature vector table created total number of 1's is counted from three state feature vectors. Three possible values may hold any one from 0 and 1 and z.

Table 5. Feature Vector Replaced by Count of 0s From Similarity Matrix.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10
D1	x	2	2	2	1	1	2	1	2	1
D2	x	x	1	2	1	1	2	1	2	1
D3	x	x	x	0	1	2	2	2	2	3
D4	x	x	x	x	1	1	1	2	2	1

D5	x	x	x	x	x	1	1	2	0	1
D6	x	x	x	x	x	x	2	2	2	2
D7	x	x	x	x	x	x	x	1	2	2
D8	x	x	x	x	x	x	x	x	1	2
D9	x	x	x	x	x	x	x	x	x	2

After this feature vector table created total number of 1's is counted from three state feature vectors. Three states can have 0 or 1 or z as the value.

Table 6. Feature Vector Matrix Count of 0s of Dataset.

	D1	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**
D1	x	2	2	2	1	1	2	1	2	1
D2	x	x	1	2	1	1	2	1	2	1
D3	x	x	x	0	1	2	2	2	2	3
D4	x	x	x	x	1	1	1	2	2	1
D5	x	x	x	x	x	1	1	2	0	1
D6	x	x	x	x	x	x	2	2	2	2
D7	x	x	x	x	x	x	x	1	2	2
D8	x	x	x	x	x	x	x	x	1	2
D9	x	x	x	x	x	x	x	x	x	2

Now according to 1's count value any documents having highest count value become a cluster because highest count means these documents have highest words common so holding high similarity value with respect of each other. Then after removing these selected documents form feature vector matrix again having next highest count value documents becomes within group since having highest common words. This process repeated until all document not covered within groups.

The groups of clusters generated as result of proposed method are as given in Figure 2.

Cluster-1: {3, 10}

Cluster-2 :{ 1, 2, 4, 7, 8, 9}

Cluster-3 :{ 5}

This group and feature vector matrix is passed to SVM machine to create network and this network being trained by current feature vector and output group sets. After training network this network may be used to classify and other document without applying all over calculation again.

5. SIMULATION AND RESULTS

5.1 Document Representation

Throughout this paper, uses the symbols N, M, and k. where N shows the number of documents present in dataset, M represents number of key terms in dataset, and k gives maximum number of clusters may generated. Here symbol D to represent the collective document set of N documents that need to be cluster, the C1;C2;...;Ck to denote cluster name of every cluster within k clusters.

5.2 Dataset

Experiments were performed using Matlab 7.10.1 on 20 new group dataset. We compared the proposed algorithm with other competing algorithms under same experimental setting. The experimental results obtained when number of cluster are set to from 2 to 8 for 20 new group dataset data sets, the number of nearest. In all experiments, proposed algorithm performs better than or competitively with other algorithms. The details of experiments can be described as follows.

5.3 Results

During experimental analysis 10 documents are selected randomly and performed proposed work and finding clustering as XNOR similarity vector. By this vector document similarity is calculated and this similarity is used to finding clusters between selected documents. These clustered groups are classified by SVM classifier and train svm network. Performance measurement for this network testing is given as follows.

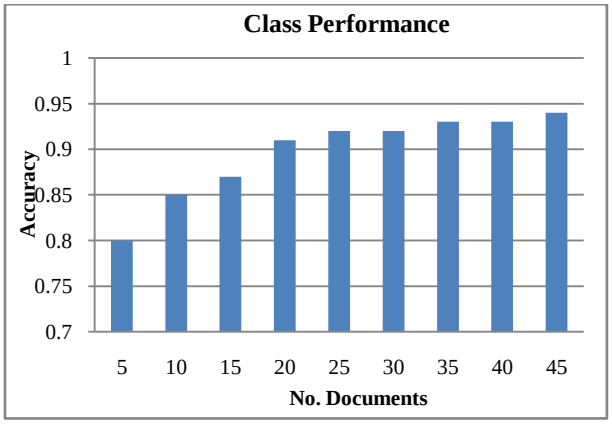

Fig 3: Accuracy comparison for SVM network during

Table 7. Performance Measurement

S. No.	**No of Documents**	**No. Of clusters**	**Class Performance**
1	5	2	0.80
2	10	3	0.85
3	15	3	0.87
4	20	3	0.91
5	25	3	0.92
6	30	3	0.92
7	35	4	0.93
8	40	4	0.93
9	45	4	0.94

6. CONCLUSION

This paper defines a brand new similarity performs to work out similarity between any two code elements or text files. AN algorithmic rule to cluster a collection of given documents or text files or code elements is meant that uses the planned similarity perform known as hybrid XNOR to search out the degree of similarity between two text documents. The given documents to algorithmic rule could be a similarity matrix and therefore the output is that the groups of clusters generated. As future aspects, the approach can be extended to classify the elements exploitation classifiers by applying symbolic logic.

The idea of Support vector machines is also used for classification once clusters square measure shaped. The search

quality is reduced by exploitation the algorithmic rule [12] wherever ever necessary as a part of part retrieval.

Using support vector with this classifier could smart technique for classification whenever cluster creates. SVM network is trained through given document files feature and cluster name can be used for any newly document matrix is result by utilized trained network for upcoming document. Therefore time of feature matching can be minimize..

7. REFERENCES

[1] Janruang, J., Guha, S.: Semantic suffix tree clustering. In: First IRAST International Conference on Data Engineering and Internet Technology, DEIT (2011)

[2] Muhammad Rafi, Mehdi Maujood ,Murtaza Munawar Fazal, Syed Muhammad Ali, "A comparison of two suffix tree-based document clustering algorithms", IEEE, 2010.

[3] Hammouda, K., Kamel, M.: "Efficient document indexing for web document clustering". IEEE Transactions on Knowledge and Data Engineering 16(10), 1279–1296 (2004)

[4] Huang, A.: "Similarity measures for text document clustering", pp. 49–56 (2008)

[5] Hammouda, K., Kamel, M.: "Phrase-based document similarity based on an index graph model". In: Proceedings of 2002 IEEE International Conference on Data Mining ICDM, pp. 203–210 (2002)

[6] Jung-Yi Jiang et.al A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 23, NO. 3, MARCH 2011

[7] Crabtree, D., Gao, X., Andreae, P.: "Improving web clustering by cluster selection". In: Proceedings of 2005 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 172–178 (2005)

[8] Chim, H., Deng, X.: "A new suffix tree similarity measure for document clustering". In: Proceedings of the 16th International Conference on World Wide Web, WWW 2007, pp. 121–130. ACM, New York (2007)

[9] Chim, H., Deng, X.: "Efficient phrase-based document similarity for clustering". IEEE Transactions on Knowledge and Data Engineering 20(9), 1217–1229 (2008)

[10] Joydeep, A.S., Strehl, E., Ghosh, J., Mooney, R.: "Impact of similarity measures on web-page clustering". In: Workshop on Artificial Intelligence for Web Search, AAAI, pp. 58–64 (2000)

[11] Carpineto, C., Osinski, S., Romano, G., Weiss, D.: A survey of web clustering engines. ACM Computing Surveys 41, 1–38 (2009)

[12] Zamir, O., Etzioni, O.: Grouper: A dynamic clustering interface to web search results. In: Proceedings of the Eighth International World Wide Web Conference, pp. 283–296. Elsevier, Toronto (1999)

[13] R. Agrawal, T. Imielinski, A. Swami. Mining association rules between sets of items in very large databases, Proceedingsof the ACM SIGMOD Conference on Management of data, 1993, pp. 207–216

[14] Congnan Luo, , Yanjun Li, Soon M. Chung. Text document clustering based on neighbors , Data & Knowledge Engineering (68), 2009,1271–1288

[15] Tianming Hu,Sam Yuan Sung, Hui Xiong, Qian Fu. Discovery of maximum length frequent itemsets, Information Sciences (178), 2008,69–87

[16] Wen Zhanga,, Taketoshi Yoshida, Xijin Tang, Qing Wang. Text clustering using frequent itemsets, Knowledge-Based Systems 23 (2010) 379–388

[17] Radhakrishna.V,C.Srinivas, C.V.Guru rao. High Performance Pattern Search algorithm using three sliding windows, International Journal of Computer Engineering and Technology , Volume 3,issue 2, 2012 , pages 543-552. Impact factor 3.85.

[18] Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, vol. 14, pp. 849–856. MIT Press (2001)

[19] Beil, F., Ester, M., Xu, X.: Frequent term-based text clustering. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2002, pp. 436–442. ACM, New York (2002)

[20] Cilibrasi, R., Vitanyi, P.: The google similarity distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)

[21] Kjos-hanssen, B., Evangelista, A.J.: Google distance between words. Computing Research Repository abs/0901.4 (2009).

[22] Rachana Kamble, Mehajabi Sayeeda, Clustering Software Methods and Comparison, Volume 5 Issue 6, Pages 1878-1885 IJCTA November-December 2014.