Document Clustering in Forensic Investigation by Hybrid Approach

G. Thilagavathi,
Pg Scholar (CSE),
Sri Ramakrishna Engineering College,
Coimbatore.

J. Anitha,
Assistant Professor (IT),
Sri Ramakrishna Engineering College,
Coimbatore

ABSTRACT
Digital Forensic Investigation is the branch of scientific forensic process for investigation of material found in digital devices related to computer crimes. Digital evidence analogous to particular incident is any digital data that provides hypothesis about incident. The essential part of Digital forensic Process is to analyze the documents present on suspect's computer. Due to increasing count of documents and larger size of storage devices makes very difficult to analyze the documents on computer. To overcome these problems, a subject based semantic document clustering algorithm along with bisecting-kmeans has been proposed that allows the examiner to analyze and cluster the documents based on particular subject and also the terms that does not belong to any subject. The accuracy of clustering of documents has been improved by means of this hybrid approach.

Keywords
Digital Forensic, Stemming, Term Importance, Subject – based semantic clustering, Document-subject similarity

I. INTRODUCTION
Different people will have different needs with regard to clustering of documents. The major goal of cluster analysis is the division of group of objects into homogenous clusters. To organize large number of documents into meaningful clusters, document clustering can be used. Using clustering techniques to group the documents can improve the information retrieval system and it is an efficient way to find similar documents. Clustering is an unsupervised learning which it finds the natural grouping of instances for given unlabeled data. It can be used as stand-alone tool and considered as preprocessing step to reduce the noise. A good clustering method produces high quality clusters which it depends on similarity measure and implementation.

Document clustering is the process of grouping similar documents into cluster. The main advantage is to retrieve the information effectively, reduce the search time and space, to identify the outliers, to handle the high dimensionality of data and to provide the summary for similar documents. It provides the efficient way of representing and visualizing the documents in which it provides better navigation. The Similarity measure used to find similarity between documents, document representation, and algorithm or technique used to cluster the documents plays major role in document clustering. Document Clustering has been used in variety of application such as recommended system, search optimization, Duplicate content detection, Document Summarization and forensic Investigation.

2. RELATED WORKS
Semi-supervised document clustering [7] approaches are unable to solve the clustering problems even though it will be search based or similarity based. Semi-supervised clustering algorithms uses labeled or a pair wise constraint supports the unsupervised clustering. User supervision can be considered as alternative forms for document clustering, by means of labeling a feature by associating it with a cluster. Besides labeled documents, it also identifies the labeled features to generate cluster seeds to process the unsupervised clustering. In this process, unified framework is used which it process labeled documents and features in the form of seeding clusters and intermediate clusters can be formed from refining this information. Two methods have been used to generate cluster seeds by using labeled features.

The traditional document classification algorithms available to train the classifiers have no training data. Text Categorization along with Support Vector Machines method [10] determines the specific properties of learning based on text data and determines the importance of SVMs appropriate for task. SVMs provide better performance over the existing best performing models and perform robustly while considering the learning tasks. They are automatic and it eliminates the requirement of manually performing parameter tuning. The evidence collected based on empirical and theoretical results for text categorization, SVMs are considered as well suited method.

With respect to content, Text categorization [5] assigns texts to one or more defined categories. The most effective five different automatic learning algorithms for text categorization are examined based on training set size, and alternative document representations. Linear Support Vector Machines (SVMs) are particularly promising because they are very accurate, quick to train, and quick to evaluate.

An Event-Based Digital Forensic Investigation Framework [3] presents a framework for digital forensics that includes an investigation process model based on physical crime scene procedures. In this model, each digital device is considered a digital crime scene, which is included in the physical crime scene where it is located. The investigation includes the preservation of the system, the search for digital evidence, and the reconstruction of digital events. The focus of the investigation is on the reconstruction of events using evidence so that hypotheses can be developed and tested.

Topic driven clustering Algorithm [16] organizes the document collection according to the given set of topics. The resultant cluster corresponds to given topics and documents in same cluster are similar to cluster topic. It takes the advantages of both supervised and unsupervised documents and performs well with topic prototypes of different levels of specificity. The traditional Partitional [8, 9, and 12] and

Hierarchical Clustering algorithms [6, 15, and 4] cannot accept any feedback from user and it will not take the advantage of user provided initial subject definition. Enhanced Topic-based Vector Space Model (eTVSM) [13] defines relation of term-topic by using algorithm with WordNet [11].

Topic based vector space model (TVSM) [2] proposed a method which compares the document based on content. From theoretical point of view, TVSM enables the integration of several natural language processing algorithms into one model. It calculates the document similarity with in relational database by using plain SQL.

3. PROPOSED SYSTEM

In the proposed system, investigator initially defines the subject vectors i.e. the terms belong to particular subject such as hacking. Then suspect document set is preprocessed in which parsing, indexing and analyzing to find the terms. Preprocessing involves the removal of stop and stem words. Indexing provides how many times the particular term appear with in document. Weight for each term in each document is provided by term frequency and Inverse document frequency.

In Subject vector space model (SVSM), generation of vectors and document similarity function will be done in which it finally forms the cluster based on similarity between the documents. Vector generation process is done by comparing the terms with Extended Synonym List, WordNet and by Top Frequent terms. Similarity measure between the terms is provided by Jacquard Coefficient in Top Frequent terms. Document-subject similarity function determines the similarity of term with in subject and within document. By means of Document-subject clustering, similar documents are clustered based on subjects. There are some terms which do not belong to any subject. To improve the accuracy Bisecting k-means algorithm is used to form the clusters from Generic cluster (ref fig1).

3.1. Initial Subject Definition

The forensic investigator is involved in clustering the collection of documents based on their related subject(S), initially the investigator defines the set of terms that belongs to particular subject. For each input term, part of speech (PoSt.) is provided to improve the accuracy of clustering algorithm. Clustering algorithm utilizes Extended Synonym list tool to generate the terms related to subject.

3.2. Pre-Processing

Pre-Processing has to be done to reduce the noise, dimensionality, computational complexity and loss of information.

3.2.1 Tokenization

The process of breaking stream of text into words or phrases in to tokens is called Tokenization. In a document, tokenization separates the sequence of characters in to tokens by using punctuation marks and white space as separators. For instance, Consider the string “Ram, Sunny and Hill” produces the tokens such as: “Ram”, “Sunny”, and “Hill”.

3.2.2 Stop words Removal

To save space and to speed up searching process, the words which are considered as less important should be removed. Any group of words can be chosen as stop words such as ‘the’,’at’,’on’,’which’, etc.

3.2.3 Stemming

Stemming algorithm is used to reduce the word to its root or stem. The key terms used in document are expressed by stem rather than original words. Porter Stemming Algorithm [18] is used here to remove the stem words. For example, consider the words “playing”, “played”, “play”, and “player” can be reduced to the root word, “play”. Once after preprocessing, the unique terms in a document set are represented as T. Then it involves the indexing and weight assignment of the terms in each document.

3.2.4 Indexing

Indexing defines how many times the particular term will appear with in document.

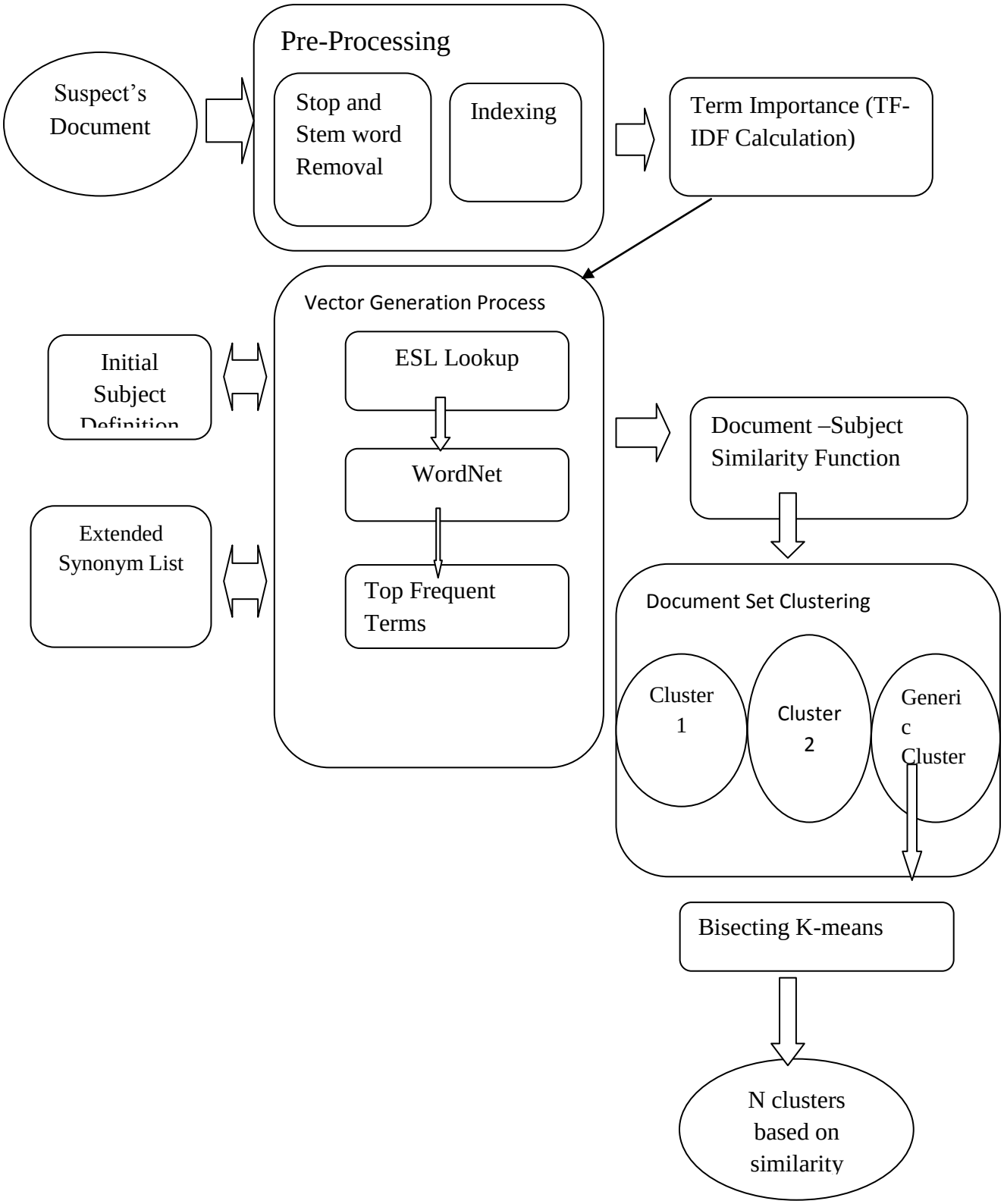

Fig 1: Architecture of SVSM

3.2.5 Term Importance

Term importance defines the influence of term with in a document. Weight and significance of term can be computed by TF-TDF Calculation.TF-IDF is used to determine the weight of each term in information retrieval and text mining. It evaluates the importance of word is to a document with in a collection. Term frequency determines how frequently particular term appears in document. Inverse document frequency evaluates how important the term is.

Specifically, each weighted term frequency must be determines as

$$\mho(t,d)= tf(t,d)\times idf(t,D) \quad (1)$$

Where tf (t, d) represents frequency of term t in document d and idf(t,D) denotes the inverse document frequency of term t in the document set D

$$idf(t,D) = 1+\log(|D|/1 + Freq(t,D)) \quad (2)$$

where |D| denotes the count of documents in the document set D, and Freq (t,D) is the count of documents in D that denotes the term t. There will be more number of terms generated for each subject even though it is limited by threshold value. To reduce the noise, dimensionality reduction technique such as term variance is used to limit the terms that improves the efficiency and effectiveness of clustering algorithm

3.3. Vector Generation Process

Subject vector space model (SVSM) is an algebraic model based on combination of vector space model [14] and Topic vector space model [2].Vector generation Process takes place in SVSM. In this process terms are compared with ESL, WordNet and Top Frequent terms to check whether the term present in these or having similarity. If it does not exist the term must be added to initial subject definition by means of extending the list. Threshold value should be specified for limiting the terms in subject vector. Weight of term in subject vector is equal or less than 1.While constructing the subject by generating expansion vectors, the weight of each term can be evaluated from weight of term in previous expansion vector.

Vector generation process can be done by comparing the term with following three steps.

Step1: ESL Lookup

For each input term, expansion vector is generated by looking related words for each term in Extended Synonym List (ESL).Extended Synonym List is an incident specific list which contains forensic-related synonyms and acronyms terms which can be used by criminals and it will not present in dictionaries.

Step2: WordNet Synonym

By using WordNet, it determines the synset for each input term and generates the expansion vector using related terms taken from Lesk's word sense disambiguation technique [1]. Word sense disambiguation technique determines part-of-speech (PoS) tagging and identifies the sense of context .WSD provides multiple meanings for words in order to find the most accurate or relevant meaning of a word in a sentence. PoS tagging is also used to label a word in a sentence with its corresponding lexical category such as Noun, Verb, etc. To perform the PoS tagging, Brill's tagger is used.

Step 3: Top Frequent Terms

The most frequently used terms are determined from top ranked documents in document set. Jaccard coefficient is used to measure the similarity between the terms. Jaccard coefficient measures the similarity between terms as the intersection divided by the union of objects .For text document, the Jaccard coefficient compares the total weight of shared terms to the total weight of terms that are found in either of the two documents but should not be shared terms.

Jaccard-coefficient measure is provided by:

$$SIM_j\ (\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a}.\overrightarrow{t_b}}{|\vec{t}_a|^2+|\vec{t}_b|^2-\vec{t}_a.\vec{t}_b} \quad (3)$$

The similarity measure Jaccard-coefficient should have range between 0 and 1.

3.4. Subject-Semantic Clustering Algorithm

Document –Subject Similarity Function [17] measures the similarity of term with in document and also with in subject. It returns the value that lies between 0 and 1.Similarity function is provided by:

$$\mathrm{Sim}(d, s_i) = \frac{1}{\Psi_d}\sum_{t \in T} \mho_{t,s_i} \times \mho_{t,d} \quad (4)$$

Where $\mho_{t,s_i}$,$\mho_{t,\ d}$are the weight of term t in subject s_i and document d, respectively.

After the generation of expansion vector, algorithm will perform the clustering and generates the cluster c.

Document Clustering Algorithm [17]:

Require $|s_i| < \delta$

For each subject $s_i \in S$ do

For each document $d \in D$ do

If $\mathrm{Sim}(d,s_i) > \tau$ then

$c_i \leftarrow d$

End if

End for

End for

Output= $\{c_1,c_2,c_3 \ldots.\}$

3.5. Bisecting-k means

There are some terms which do not belong to any subject i.e. Generic Cluster. Bisecting-k means algorithm [19] is used to improve the accuracy. Bisecting k-Means is the combination of k-Means and hierarchical clustering. It begins with single cluster which it has all objects in it.

Steps of Bisecting-k means algorithm:

Initially cluster should be chosen to be split.

By using basic k-means algorithm to find 2 sub clusters.

Repeat Bisecting process for iterative times.

Then take the split that produces the highest overall clustering similarity.

Repeat steps 1, 2 and 3 till desired count of clusters limit is reached.

By means of using Bisecting k-means to determine the term belongs to clusters in addition to subject based clustering.

4. PERFORMANCE EVALUATION

The dataset used here is Classic3 dataset where a large number of documents are available for usage and they are analyzed offline. The three parameters which are used to evaluate the performance for Subject-Semantic Clustering Algorithm along with Bisecting k-means are Precision, Recall and F-measure. Let L= $\{k_1, k_2, \ldots\}$ is the set of classes, C= $\{c_1, c_2, \ldots\}$ is set of clusters and N is number of clusters.

Accuracy can be determined from these measures.

Precision:

Precision is the fraction of the documents retrieved that are relevant to the user's information need.

$$\text{Precision}(k_i, c_j) = \frac{true\_positive}{true\_positive + false\_positive} \quad (5)$$

Recall:

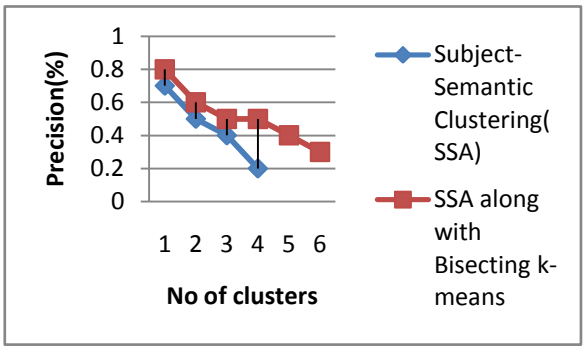

Fig.2. Performance Comparison based on Precision

Recall is the fraction of the documents that are relevant to the query that are successfully retrieved.

$$\text{Recall}(k_i, c_j) = \frac{true\_positive}{true\_positive + false\_negative} \quad (6)$$

F-Measure:

F-Measure is a measure of testing the accuracy. To compute the weighted mean, it takes both precision and recall values.

$$\text{F-Measure}(k_i, c_j) = \frac{2 \times \text{Precision}\ (k_i,c_j) \times \text{Recall}\ (k_i,c_j)}{\text{Precision}\ (k_i,c_j) + \text{Recall}\ (k_i,c_j)} \quad (7)$$

From Fig 4.1, Clusters formed are based on subjects. While considering intra-similarity, SSA with Bisecting K-means achieves high precision value of 10% more than SSA

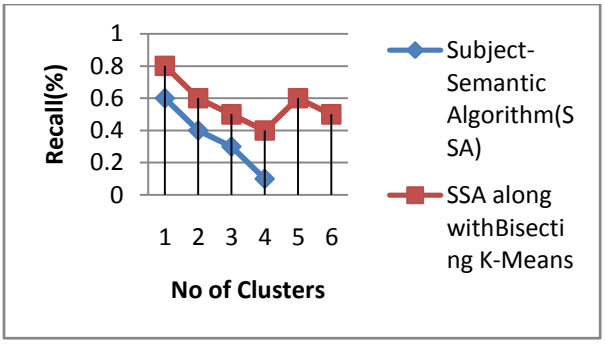

Fig.3. Performance Comparison based on Recall

Fig 4.2 shows SSA along with Bisecting K-means achieves high recall value of 20% more than SSA.

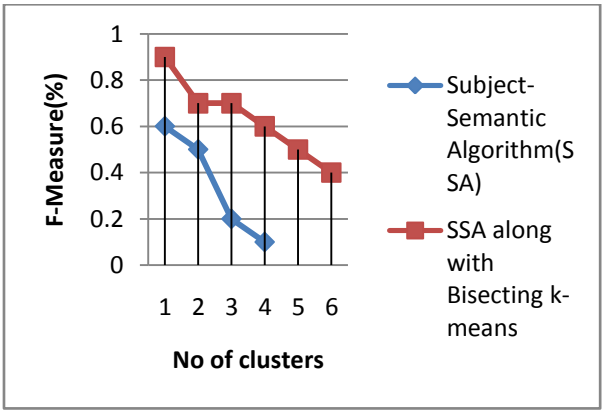

Fig.4. Performance Comparison based on F-Measure

Based on intra-similarity, SSA with Bisecting K-means achieves high F-measure value of 40% more than SSA has been shown on fig 4.3.Thus accuracy got improved based on precision, recall and f-measure values.

5. CONCLUSION AND FUTURE WORK

In this paper subject-based semantic document clustering along with bisecting k-means algorithm has been proposed for digital forensic investigation by using data mining to support forensic investigation. Term synonyms are extracted by using WordNet and appropriate sense for each term can be determined by word sense disambiguation technique .Document-subject similarity function provides the similarity of terms with in document and subject. Finally Subject – Semantic Document clustering Algorithm produces the clusters. Along with that Bisecting k-means algorithm is additionally added to improve the accuracy in addition to subject-based clustering. Precision, Recall and F-measure are the measures used to improve performance of clustering.

The main advantage of document clustering here is an unsupervised i.e. number of clusters will be formed based on subject and it cannot be predetermined or given by users. Time taken to perform the clustering takes less time based on subjects declared. The future work is to extend these ideas in to Ontology-based document clustering with limited terms to avoid noise in Digital Forensics Investigation.

6. REFERENCES

[1] S. Banerjee, Adapting the lesk algorithm for word sense disambiguation to wordnet, Ph.D. thesis, University of Minnesota, 2002.

[2] J. Becker, D. Kuropka, Topic-based Vector Space Model, Proceedings of the 6th International Conference on Business Information Systems, Colorado Springs,2003.

[3] B.D. Carrier, E.H. Spafford, An event-based digital forensic investigation framework, Proceedings of the 4th Digital Forensic Research Workshop, 2004.

[4] G. Costa, G. Manco, R. Ortale, E. Ritacco, Hierarchical clustering of xml documents focused on structural components, Data & Knowledge Engineering 84(2013) 26–46.

[5] S. Dumais, J. Platt, D. Heckerman, M. Sahami, Inductive learning algorithms and representations for text categorization, Proceedings of the 7th International Conference on Information and Knowledge Management, ACM, New York, NY, USA, 1998.

[6] B.C.M. Fung, K. Wang, M. Ester, Hierarchical document clustering using frequent item sets, Proceedings of the 3rd SIAM International Conference on Data Mining (SDM), SIAM, San Francisco, CA, 2003.

[7] Y. Hu, E.E. Milios, J. Blustein, Semi-supervised document clustering with dual supervision through seeding, Proceedings of the 27th Annual ACM Symposium on Applied Computing, SAC '12, ACM, New York, NY, USA, 2012.

[8] A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31 (8) (2010) 651–666.

[9] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[10] T. Joachims, Text categorization with support vector machines: learning with many relevant features, Proceedings of the 10th European Conference on Machine Learning, Springer-Verlag, London, UK, 1998.

[11] G.A. Miller, WordNet: a lexical database for English, Communications of the ACM 38 (1995) 39–41.

[12] R.T. Ng, J. Han, Efficient and effective clustering methods for spatial data mining, Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc.,San Francisco, CA, USA, 1994.

[13] A. Polyvyanyy, D. Kuropka, A Quantitative Evaluation of the Enhanced Topic-based Vector Space Model, Universitätsverlag Potsdam, 2007.

[14] G. Salton, Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley Longman Publishing Co., Inc.,Boston, MA, USA, 1989.

[15] M. Steinbach, G. Karypis, V. Kumar, A Comparison of Document Clustering Techniques, 2000.

[16] Y. Zhao, G. Karypis, Topic-driven clustering for document datasets, Proceedings of the SIAM Data Mining Conference (SDM), 2005.

[17] Gaby G. Dagher, Benjamin C.M. Fung ,"Subject-based semantic document clustering for digital forensic investigations ",Data & Knowledge Engineering 86 (2013) 224–241

[18] M.F.Porter. The Porter Stemming Algorithm. www.tartus.org/martin/PorterStemmer

[19] B.S.Vamsi Krishna, P.Satheesh, Suneel Kumar R," Comparative Study of K-means and Bisecting k-means Techniques in Wordnet Based Document Clustering", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-6, August 2012