Correlated Concept based Topic Updation Model for Dynamic Corpora

J. Jayabharathy
Department of Computer Science and Engineering
Pondicherry Engineering College

S. Kanmani
Department of Information Technology
Pondicherry Engineering College

N. Sivaranjani
Department of Computer Science and Engineering
Pondicherry Engineering College

ABSTRACT

A rapid growth of documents available on the Internet, digital libraries, medical documents, news wires and other scientific document corpuses has motivated the researchers to propose many text mining techniques that help users to quickly retrieve trace and summarize the information in an effective way. Topic detection is one such technique which discovers precise, meaningful and concise labels for the formulated static document clusters. This technique helps the user to navigate and retrieve the needed information quickly and efficiently. Topic updation is the process of identifying and renewing the discovered labels whenever the document clusters are updated dynamically. This paper focuses on topic updation model based on Testor theory. The proposed work is experimented using 20newsgroup and scientific literature data set. The experimental results demonstrate that the proposed algorithm exhibit better performance, compared to the existing algorithms for topic detection.

General Terms

Algorithms, Experimentation, Methods, Results.

Keywords

Document Clustering; Static Clustering; Dynamic Clustering; Topic detection; Topic updation; Testor Theory, F-Measure and Purity.

1. INTRODUCTION

Information extraction plays a vital role in today's life. Retrieving relevant documents efficiently and effectively the relevant documents from World Wide Web is a challenging issue [1]. As today's search engine does just pattern matching, retrieved documents may not be so relevant to user's query [1]. A good document clustering approach can assist computers in organizing the document corpus automatically into a meaningful cluster hierarchy for efficient browsing and navigation [1], which is very valuable for overcoming the deficiencies of traditional information retrieval methods. It is a more specific technique for unsupervised document categorization, automatic topic discovery and rapid information retrieval or filtering [2]. It breaks down huge linear results into manageable sets. It is an automatic grouping of text documents into clusters where documents of the same cluster are more similar than the documents in different clusters.

By clustering the text documents, the documents sharing the same topic are grouped together [1]. Labels were not discovered in clustering like document classification; hence clustering is known as unsupervised learning. Topic detection deals with discovering meaningful and concise labels for the clusters which are grouped using document clustering algorithm. Searching collection of documents by choosing from the set of topics or labels assigned to the clusters becomes easy and efficient. A good descriptor or the label for a cluster should not only indicate the main concept of the cluster, but should also discriminate the cluster from other clusters. Hence a proper document mining model is considered to consist of three phases Document pre-processing, Document clustering and Topic discovery [3].

As the number of documents increasing day-by-day, there arises a need to cluster the documents dynamically. Clustering documents dynamically reduces the time taken for clustering the document as it processes and assigns the newly inserted document alone into the existing clusters instead of re-clustering the entire documents in the corpora. After adding a new document in the clusters, there arises a need to update the discovered label/topic of the clusters based on the concepts of the document inserted. So, an efficient Topic Updation model helps to discover and update the label/topic for the dynamically updated clusters; which is more suitable for the dynamic environment. This work aims at proposing and implementing a Topic updation using Testor theory [4] model for discovering and updating meaningful and concise labels for the dynamically updated clusters which are grouped using Semantic Similarity based Histogram based Incremental Document Clustering (SHC) [5] and Enhanced Similarity Histogram Clustering using Intra Centroid Vector Similarity (ESHC-IntraCVS) [6] based on Semantic-based similarity for the scientific literature documents and newsgroup dataset. This proposed technique is compared with the existing Topic Discovery by clustering keywords[7] proposed by Wartena and Brussee, (2008) and TF-IDF classifier [8] by Seymore and Rosenfeld (1997) method using F-measure and Purity as evaluation metrics.

The remaining part of this paper is organized as follows. Section 2 reviews related work on Topic detection and Updation. Section 3, outlines the proposed model for dynamic Topic Updation, also, the need for considering correlated terms are briefly stated in that section. In Section 4, the experimental setup and data set descriptions have been discussed, followed by analysis of results. Finally salient conclusions are presented in section 5.

2. RELATED WORKS

Christian Wartena and Rogier Brussee proposed the discovery of Topic by Clustering Keywords [7]. This approach consists of two steps. First a list of the most informative keywords is extracted. Subsequently clusters of keywords are identified for which a center is defined, which is taken as the representation of a topic. Documents are clustered according to various similarity functions like Cosine similarity, Jensen-Shannon divergence distance. A collection of 8 Wikipedia topics are considered as data set. Considering F-measure as the performance metrics, the experiments show that discovery of topic by clustering keywords gives better results. Jenson Shannon divergence shows better performance compared to cosine similarity [1].

Anaya-Sánchez et al. [9] proposed a new document clustering algorithm for topic discovering and labelling which relies on both probable term pairs generated from the collection and the estimation of the topic homogeneity associated to term pair clusters. Starting from the most probable pair of terms generated from the collection of documents C, its support set (i.e. the set of documents in C that contain both terms) is built. If this set is homogeneous in content, a cluster consisting of the set of relevant documents for the content labeled by the pair is created[1]. In order to measure the homogeneity of a document collection C, entropy is usually applied over the vocabulary of the collection at hand. This approach is restricted to term-pairs in order to simplify the search of cluster labels. A document clustering algorithm for discovering and describing topics [10] is an extension of their previous work. It provides more descriptive and suitable topic labels instead of a simple term pair. Experiments carried out over TDT2 English corpus, AFP Spanish collection and Reuters-21578 show significant improvements over existing methods in terms of the standard macro- and micro-averaged F1 measures.

Hei-Chia Wang et al. [11] proposed a topic detection method based on bibliographic structures (Title, Keyword and Abstract) and semantic properties to extract important words and cluster the scholarly literature. Based on the lexical chain method combined with WordNet lexical database, the proposed method clusters documents by semantic similarity and extracts the important topics for each cluster. There are three main models in the system architecture. At first, the pre-process model collects journal papers and processes their Title, Keyword and Abstract information to prepare for the lexical chain construction. Secondly, the document representative model implements the steps to build lexical chains. WordNet is used to find the nearest hypernyms between two nouns. The tertiary model is the semantic cluster model. It calculates similarity and clusters documents. In order to take semantic features into account, in this approach a novel method to calculate semantic similarity is proposed. The similarities in the Title, Keyword and Abstract of papers will be calculated separately. These bibliographic structures are taken into consideration, with different weights given to each [1]. After the semantic similarity calculation, the HCA method is used to cluster the documents. After collecting key phrases and their respective frequencies from each of the documents in one cluster, the key phrases with the highest frequencies can be viewed as topics of the cluster. The experimental results show that the proposed method is better than the traditional TF-IDF method. Most of the researches work detecting topics for the closed set of documents in off-line mode [12-14].

AlSumait et al. proposed an online Topic model (OLDA) [15] which automatically captures the thematic patters and detects the emerging topics from the text stream during run time. This dynamic approach also provides an efficient mean to track the topics over time and detect the emerging topics in real time. This method is evaluated both qualitatively and quantitatively using Reuters and NIPS dataset. GAC-INCR [16] applies clustering algorithm (GAC) in the first phase and merges the cluster based on similarity/novelty threshold. The topics are discovered from the patterns of the formulated cluster. Ankan and Vikas [17] proposed a system to capture the themes, track the existing topics and also, maintains temporal and continuity in user views. Data set considered for their experiment analysis is TDT2 and Twitter social media. The experimental results shows better performance compared to traditional TDT. Many dynamic topic modeling system uses the variants of PLSI and LDA online topic discovery [18-22].

The survey findings of Prathima and Supreethi (2011) [23] on concept based clustering algorithms, concludes that most of the mining techniques use TF-IDF method. This method has the following issues:

- It fails to differentiate the degree of semantic importance of each term;
- It assign weights without distinguishing between semantically important and unimportant words within the document and
- It does not consider synonyms, polysemous, etc.

From the literature survey, it is understood that very few researches on topic detection are based on Testor theory; More than 60% of the dynamic topic modeling system uses the variants of PLSI and LDA online topic discovery. Many of the researches are based on Topic Detection and Tracking which is mainly based on term frequency representation. Moreover, the existing topic detection models either uses term frequency or synonyms and hypernyms representation. As scientific literature and many tracks of news documents consist of purely domain-specific technical terms, the performance of synonyms and hypernyms based clustering and topic detection may not always yield a better result. Since, Testor theory is suitable for term based and correlated concept based topic analysis the authors has focused their research towards term frequency and correlated concept based Topic Updation for monitoring the changes that occurs in the emerging information. In this regard, a domain specific dictionary has been developed by the authors to extract the related terms as correlated concepts.

3.PROPOSED TOPIC UPDATION MODEL FOR DYNAMIC CORPORA

Figure 1 shows the sequence of steps involved in Topic Updation model using correlated concepts. The document which is in unstructured format are preprocessed and converted to a structured format. The details of each module involved in the model are preprocessing, static clustering and dynamic document clustering and topic updation are discussed below.

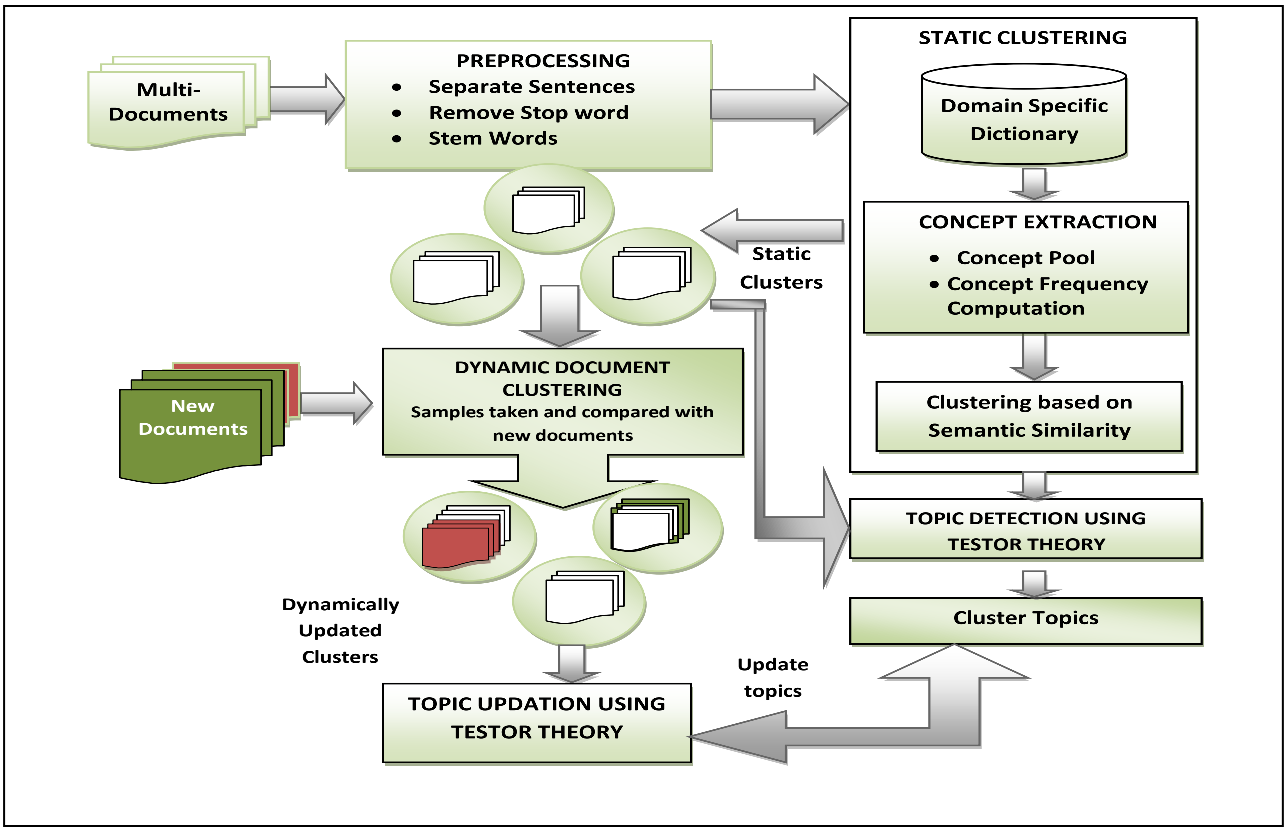
Multi-Documents
PREPROCESSING
Separate Sentences
Remove Stop word
Stem Words
STATIC CLUSTERING
Domain Specific Dictionary
CONCEPT EXTRACTION
Concept Pool
Concept Frequency Computation
Clustering based on Semantic Similarity
Static Clusters
New Documents
DYNAMIC DOCUMENT CLUSTERING
Samples taken and compared with new documents
TOPIC DETECTION USING TESTOR THEORY
Cluster Topics
Dynamically Updated Clusters
Update topics
TOPIC UPDATION USING TESTOR THEORY

Figure 1. Proposed Model for Topic Updation

3.1Preprocessing

Preprocessing involves: tokenization, removing stopwords and stemming.

Tokenization, is the process of splitting the sentences into separate tokens (http://nlp.stanford.edu/software/tokenizer). For example, “this is a paper about topic updation” is split as: this\is\paper\about\topic\updation. Stop words are frequently occurring words that have little or no discriminating power, such as: \a", \about", \all", etc., or other domain-dependent words. Stop words are often removed[1]. Stemming is the process of removing the affixes in the words and producing the root word known as the stem. [24]. Typically; the stemming process is performed to transform the words into their root form. For example: connected, connecting [1].

3.2 Need for correlated terms [25]

There are many existing clustering algorithms that take synonyms and hypernyms for vector representation. In this study, the authors have considered *crtv* as concepts for clustering to improve the efficiency of clustering the documents both statically and dynamically. The idea of considering terms and related terms as concepts based on semantic similarity has been carried out for extracting topic from the clustered documents [3]. The proposed topic updation takes this idea of considering *crtv* as concepts for static clustering and topic detection and applies the same concept for clustering and updating topic to the document clusters dynamically. Considering terms or synonyms and hypernyms for information extraction leads the following issues [25]:

Case 1: Words have multiple meanings, hence diversifies the information extraction.

E.g. Bat: represents the cricket bat or a kind of a bird.

Whereas using correlated concept extraction algorithm the term “bat” is related to the domain what it refers to.

Case 2: Considering terms or synonyms of the terms limits the search space of the domain.

E.g. wireless: first sense medium of communication.

Multiple terms equivalent to the term “wireless” is extracted as correlated concepts instead of restricting to one term.

Similarly, synonyms of the term “*wireless*” is extracted from WordNet as: “*first sense medium of communication*”, whereas, taking related terms like “*wireless*”, “*communication*”, “*protocol” “mobile communication*” etc. will be extracted as concepts, which gives better accuracy and improves the efficiency of information extraction[3]. For example, *sports article* contains terms like: a *ball, bat, wicket, run, batsman, over* etc. Taking synonyms/hypernyms as concept, will not give better performance since the meaning of these terms are not literally same. If we consider the technically related terms i.e. *crtv*, all the above mentioned terms will be grouped together as a single concept which refers sports related to the concept – *cricket*. Similarly the synonym for the term “*farmer”* from WordNet is extracted as: *“a person Title who operates a farm*”. But using the proposed model the concept will be extracted as *“farmer”, “crops”, “fertilizer”, “land” and “farm”.* Clustering the document using this extraction procedure would improve the performance of the resulting cluster, than that of the cluster generated by existing works [25].

Concept extraction algorithm: description [25]

Considering the extraction of synonyms or hypernyms as concepts degrades the efficiency of the results in the case of scientific literature and news group dataset because of the fact that the documents speak more about scientific or technical terms. Concept extraction is based on our previous work [3] where Correlated concepts are nothing but the terms and their related terms. For Concept extraction, domain specific dictionary is used where terms related to each domain is kept along with the definition of the term. For e.g. the terms A and B are taken as a concept; if term A is in the definition of term B or vice versa combines A and B as a single concept else add the definition of A and B as separate concept to the concept list. E.g. Considering share market as the term in the news documents, the related terms are share, shareholder, money, market. The documents containing these words are grouped together as share market which forms the cluster.

3.3.Static document clustering

The processed documents are clustered using a Bisecting K-means clustering algorithm in order to group similar documents. Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations of the same cluster are similar in some sense. The Bisecting K-means method splits a large cluster into two sub-clusters and this step is repeated for several times, until the K numbers of clusters are formed with high similarity [26].

3.4 Topic Detection

The correlated concept based method and Testor theory [4] are used to cluster documents and discover topics respectively.

Testor Theory

Testor theory [4] is used to discover each cluster a tag after clusters are formed. Topic detection by Testor theory involves construction of a learning matrix and a comparison matrix.

***Learning Matrix*:** For each cluster C, a learning matrix LM(C) is constructed whose columns are the most frequent concepts in the representative C, and its rows are the representatives of all clusters, described in terms of these columns. In order to calculate the typical Testors, two classes are considered in the matrix LM(C). The first class is only formed by C and the second one is formed by the other cluster representatives. The goal is to distinguish cluster C from other clusters.

***Comparison Matrix*:** Comparison matrix could be a matrix of similarity or a matrix of dissimilarity depending on the type of comparison criteria that are applied for each feature. In this case, the features that describe the documents are the concepts and its values are the frequency of concepts. The comparison criterion applied to all the features is:

$$d(V_{ik}, V_{jk}) = \begin{cases} 1 \; if \; V_{ik} - V_{jk} \geq \delta \\ 0 \quad\quad otherwise \end{cases} \quad (1)$$

where v_{ik}, v_{jk} are the frequencies in the cluster representative i and j in the column corresponding to the concept c respectively, and δ is a user-defined parameter. As it can be noticed, this criterion considers the two values (frequencies of the concept c_k) different if the concept c_k is frequent in cluster i and not frequent in cluster j. From this comparison matrix, the most representative concept c in cluster C is obtained, which is used to tag the cluster.

3.5 Dynamic Document Clustering

Dynamic Document Clustering is the process of inserting the newly arrived documents to the appropriate existing cluster such that the formulated cluster will have a high intra- cluster similarity, and less inter-cluster similarity. At first the new documents are preprocessed and then it is clustered based on the dynamic technique. The issues that are to be addressed are [25]:

- Effectiveness: How accurately the newly arrived documents are inserted to the existing clusters.
- Insertion Order Issue: Pattern of arrival of new documents should not affect the correctness of the clusters.

The new documents are assigned to the existing cluster, one by one in recursive steps. The new documents are assigned to a cluster dynamically at run time without the need for re-clustering. As a result the existing clusters are updated and the final clusters are obtained.

Semantic similarity histogram based incremental document clustering (SHC) algorithm

Gad and Kamel [5] proposed an incremental clustering algorithm based on Phrase-Semantic Similarity Histogram (PSSM). This algorithm integrates the text semantic to the incremental clustering process. The clusters are represented using semantic histogram which measures the distribution of semantic similarities within each cluster. The PSSM which is based on single word analysis and phrase analysis, assigns and adjusts the term weight (word/phrase) based on its relationships with semantically similar terms that occur together in the document. As soon as the new document is incrementally added to the cluster, the semantic histogram ratio is calculated and the insertion order problem is addressed by making bad documents that reduce the cluster cohesiveness to leave, and reassign them to a more appropriate cluster [25].

Enhanced similarity histogram clustering using intra centroid vector similarity (ESHC-intra CVS) algorithm

Gavin and Yue [6] proposed an enhanced incremental clustering approach to develop a better clustering algorithm that helps to organize the information available on the internet in an incremental fashion in a better way. This enhanced algorithm works with the idea that the cluster that contains a large number of similar documents to the current document being clustered will have a centroid vector that has a high similarity to the current document. Therefore, the cluster whose centroid vector is most similar to the document's vector representation is the one that most likely to contain the maximum number of documents that are more similar to the current document. Adding the new document to this cluster (when possible) will probably give the greatest benefit to that cluster and the entire dataset. This approach uses the same pair-wise document similarity representation and distribution approach and also uses additional information about the cluster to determine the best cluster to place the new document [25].

3.6 Topic Updation

This section describes about the proposed Topic Updation model using Testor theory [4] for discovering and updating meaningful and concise labels for the dynamically updated clusters using correlated concepts. The new incoming documents are grouped using SHC and ESHC based on Semantic-based similarity for the scientific literature documents and newsgroup dataset. The following steps are implemented to update the existing cluster topic.

The problem of topic updation could be modeled as: The existing cluster needs to update the topic set of the cluster. Let C= {C_1, C_2 ….C_k} be the non overlapping clusters. The feature based unique topics are identified using Testor theory T= {T_1, T_2…T_k} for the formulated k clusters. T_i = {t_1, t2...t_j} let T_i represents topic set of i^{th} cluster, which consists of j number of labels. . Let d_{new} = {d_{n1}, d_{n2}… d_{ns}} be the new document(s) that are to be inserted the existing cluster set C= {C_1, C_2 ….C_k}. Ci= {d_i, d_j ...d_l, d_{ns}} where d_{ns} is the new document which is inserted to the existing cluster C_i iff sim(C_i,d_{ns}) > sim(C_j,d_{ns}) where j=1 to k. Update the topic set of i^{th} cluster as Ti = { t_1, t2...t_j,t_{j+1}…t_l, where { t_j,t_{j+1}…t_l} are the topics identified for the new document d_{ns}.

The topic updation could be addressed as:

i) The top correlated concepts are extracted from the new documents

ii) Search each entry in the learning matrix, (matrix constructed during static topic detection algorithm)

iii) Insert the new concepts to the learning matrix if the concepts are not represented

iv) The comparison matrix was recomputed using the comparison criterion

v) Choose the most prominent topics and label the clusters

4. EXPERIMENTAL RESULTS

4.1 Data set

The data set used for the experimental analysis contains 500 abstract articles collected from the Science Direct digital library. The articles are classified according to the Science Direct classification system into four major categories: computer networks and communications, nuclear and high energy physics, economics and econometrics, and civil and structural engineering. In addition, to that 20Newgroups is considered as another data, set for the result analysis which consists of more than 1000 news articles related to Sports, Political and Share market tracks.

4.2 Performance metrics

F-measure and Purity are the performance measures used to evaluate the quality of document clustering. F-measure combines the Precision and Recall from information retrieval process [24]. Each cluster is treated as if it were the result of a user query, and each class as if it were the desired set of documents, for a query [26]. The recall and precision of that cluster for each given class are calculated. More specifically, F-measure for cluster *j* and class *i*is calculated as follows:

$$Recall(i,j) = \frac{n_{ij}}{n_i} \qquad (2)$$

$$Precision(i,j) = \frac{n_{ij}}{n_j} \qquad (3)$$

$$F(i,j) = \frac{\left(2 * Recall(i,j) * Precision(i,j)\right)}{Presicion(i,j) + Recall(i,j)} \qquad (4)$$

Where n_{ij} is the number of members of the class *i* in cluster *j*, n_j is the number of members of cluster *j* and n_i is the number of members of class *i*[26]. For each class, only the cluster with highest F-measure is selected. Finally, the overall F-measure of a clustering solution is weighted by the size of each cluster:

$$F(S) = \frac{1}{n}\sum_{j=1}^{n} \frac{n_j}{max\left(F(i,j)\right)} \qquad (5)$$

The purity measure evaluates the coherence of a cluster, that is, the degree to which a cluster contains documents from a single class [27]. Given a particular cluster C_i of size n_i the purity of C_i is formally defined as:

$$P(C_i) = \frac{1}{n} max\left(n_i^h\right) \qquad (6)$$

Where $max(n_i^h)$ is the number of documents that are from the dominant class in cluster C_i and n_i^h represents the number of documents from cluster C_i assigned to class *h*. The overall purity of a clustering solution is:

$$Purity(S) = \frac{1}{n}\sum_{i-1}^{n} max\left(n_i^h\right) \qquad (7)$$

4.3 Implementation Procedure

Initially, text documents which have been collected from various sources were accumulated in a database. Then, pre-processing was carried out by considering the various stages like: tagging by means of Stanford POS tagger tool, stop word removal and stemming, based on Porter Stemmer algorithm and morphological capabilities of WordNet. Then the documents are represented as correlated concept vector (crtv). These documents are clustered using Bisecting K-means algorithm which generates K number of clusters. For the formulated cluster, using Testor theory, topic are identified and assigned. The experiments are conducted by varying the number of new documents from 50 to 500 that are to be inserted in the existing clusters. SHC and ESHC are implemented using correlated concept based representation (crtv). Using SHC and ESHC the new documents are clustered and inserted to the existing clusters. For updating the topics, Testor theory is applied. The existing Topic Discovery by clustering keywords [7], and TF-IDF classifier [8] algorithms as originally proposed by the various authors were implemented in the above environment.

For implementing the existing algorithms the preprocessing as outlined in this work along with dataset chosen for the study were used. By varying the number of clusters the results of the proposed and existing algorithms are measured. These algorithms are implemented in JDK 1.7 environment using Net Beans IDE.

4.4 Results and Discussion

The performance analysis of Topic Discovery by clustering keywords [7] and TF-IDF classifier [8] and the topic updation using Testor Theory for SHC and ESHC clustering algorithms are categorized into two classes:

i) Based on F-measure and Purity analysis for Scientific Literature

ii) Based on F-measure and Purity analysis for Newsgroup

F-measure and Purity analysis for Scientific Literature Dataset

The quality of the formulated cluster and topic has been assessed based on F-measure and purity as performance metrics. Figures 2 and 3 shows the results of the proposed correlated term based algorithms and term based topic detection algorithms. Both SHC and ESHC algorithms give better results compared to TF-IDF classifier and Topic detection by clustering keywords. This is because the data set chosen for these experiments are domain-specific documents which consist of more scientific and technical terms compared to English literary terms, contained in the other dataset.

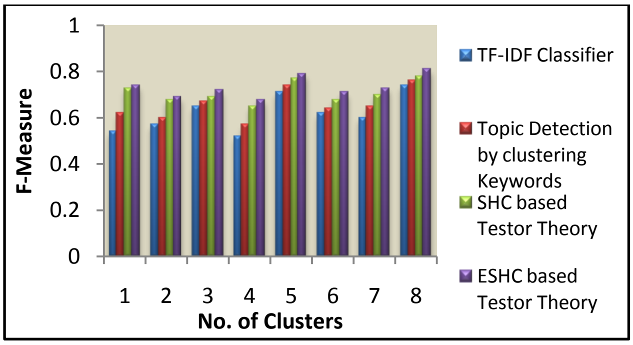

Figure 2. F-measure comparison for scientific dataset

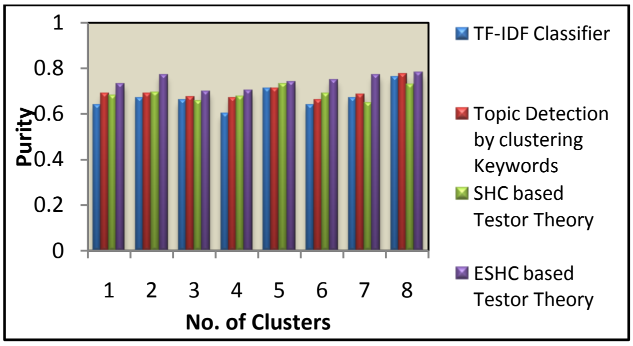

Figure 3. Purity comparison for scientific dataset

F-measure and Purity analysis for Newsgroups Dataset

The Figures 4 and 5 show the F-measure and Purity comparison for the existing and proposed algorithms. From these, charts it is inferred that the quality of SHC and ESHC based Testor theory algorithms are better than the existing Term based algorithm Topic detection algorithms. The performance of the Topic detection is evaluated between two categories of algorithms as: Concept based and Term frequency based algorithms. From the above figures it is also inferred that SHC and ESHC based Testor theory compared with TF-IDF and Topic by clustering keyword algorithm gives less improvement for newsgroup dataset. The drop in the performance of the 20newsgroup dataset is due to the dominance of English literary terms in the documents, rather than technical terms. Since the above dataset consists of more literary terms, existing algorithm works better compared to the proposed algorithms considered in this study.

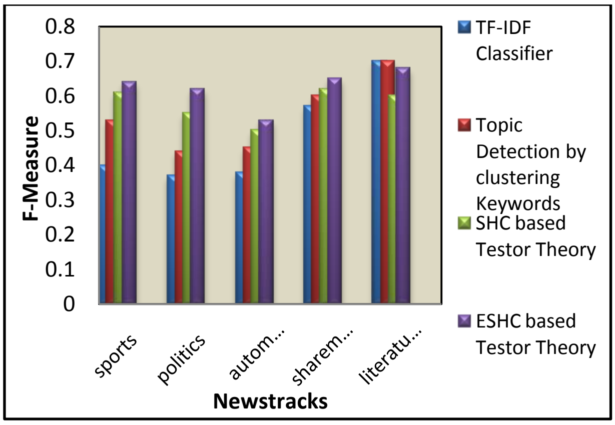

Figure 4. F-measure comparison for Newsgroup Dataset

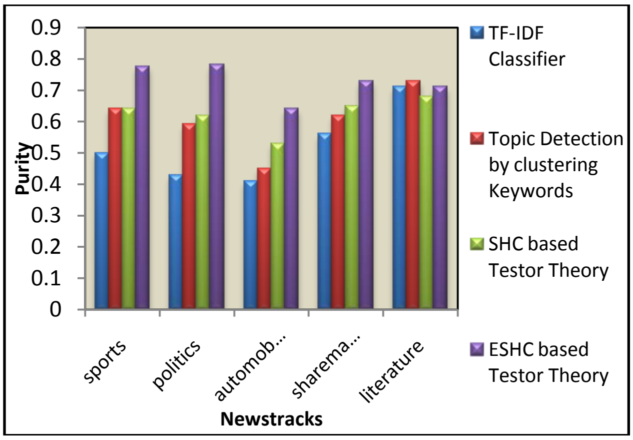

Figure 5. Purity comparison for Newsgroups dataset

5. CONCLUSION

The emphasis of the present work is Topic updation Correlated based Concept algorithm, using Testor Theory. In general the documents are represented as TF-IDF, whereas, in this study the documents are represented by means of correlated term vector (crtv). This representation helps the user to capture the technical correlation between the documents. From the comparative analysis it can concluded that considering crtv representation for topic updation leads to promising results especially for scientific literature. Sometimes the results from the Newsgroup dataset are not promising, due to the need for relatively more English literary terms, rather technical terms. In future, it is proposed to extend concept extraction based on significant phrases in documents, and also by incorporating semantic relations like hyponymy, holonymy, and meronymy.

6. REFERENCES

[1] Jayabharathy. J, Kanmani. S and Ayeesha Parveen. A. 2011, "*A Survey of Document Clustering Algorithms with Topic Discovery*", *Journal of Computing*, Vol. 3, No. 2, pp. 21-27.

[2] Kim. N, Tam. N and Van. N. 2013, " *Document Clustering Using Dirichlet Process Mixture Models of Von Mises –Fisher Distributions*", Proceedings of the Fourth Symposium on Information and Communication Technology, pp 131-138 .

[3] Jayabharathy. J, Kanmani. S, and Ayeshaa Parveen. A. 2011. “Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature”, IEEE 3rd International Communication Software and Networks (ICCSN), pp 425 – 429.

[4] Li. F, Zhu. Q and Lin.X. 2009. “Topic Discovery in Research literature Based on Non-negative Matrix Factorization and Testor theory”, IEEE Asia-Pacific Conference on Information Processing.

[5] Gad, W. K., & Kamel, M. S. 2010. “Incremental clustering algorithm based on phrase- semantic similarity histogram”, Proceedings of the Ninth International Conference on Machine Learning and Cybernetics, Vol. 11 No.14, pp. 2088–2093.

[6] Gavin, S., & Yue, X. 2009. “Enhancing an incremental clustering algorithm for Web page collections”, IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies, pp. 81–84.

[7] Wartena. C and Brussee. R (2008), “Topic Detection by Clustering Keywords”, IEEE 19th International Conference and Expert System Application.

[8] Seymore. K and Rosenfeld. R 1997, “Largescale Topic Detection And Language Model Adaptation”, Technical Report CMU-CS-97-152,

[9] Anaya-Sánchez. H, Pons-Porrata. A, and Berlanga-Llavori. R (2008). “ A New Document Clustering Algorithm for Topic Discovering and Labeling”, CIARP 2008, Progress in Pattern Recognition, Image Analysis and Applications Lecture Notes in Computer Science Vol 5197, pp 161-168.

[10] Anaya-Sánchez. H¨ , Pons-Porrata. A and Berlanga-Llavori. R 2010. “^{A} document clustering algorithm for discovering and describing topics”, Pattern Recognition Letters, Elsevier Volume 31, No. 6, 15, pp 502–510.

[11] Wang. H, Huang. T, Guo. J and Li. S 2009, “Journal Article Topic Detection Based on Semantic Features”, 22nd IEA/AIE, Proceeding In Springer, Next-Generation Applied Intelligence, Lecture Notes in Computer Science, Volume 5579, 2009, pp 644-652.

[12] Song. X, Lin. C, Tseng. B, and. Sun. M. 2005, “Modeling and predicting personal information dissemination behavior,” Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data mining.

[13] Wang. X, Mohanty. N, and McCallum. A. 2005, “Group and topic discovery from relations and text,” The 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Workshop on Link Discovery: Issues, Approaches and Applications, pp. 28-35.

[14] Wang. X and McCallum. A. 2006, “Topics over Time: A Non- Markov Continuous-Time Model of Topical Trends,” ACM SIGKDD international conference on Knowledge discovery in data mining.

[15] AlSumait. L, Barbar´a. D, Domeniconi. C , On-Line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking”, IEEE International Conference on Data Mining, pp. 3-12.

[16] Yang. Y, Pierce. T, and Carbonell. J 1998. A Study on Retrospective and Online Event Detection. In SIGIR, 1998.

[17] Saha. A and Sindhwani. V. 2012, “Learning Evolving and Emerging Topics in Social Media: A Dynamic NMF approach with Temporal Regularization”, ACM Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, USA, February pp. 8-12.

[18] Chou. T and Chang. M. 2008. “Using Incremental PLSI for Treshhold-Resilient Online Event Analysis”. IEEE transactions on Knowledge and Data Engineering.

[19] Gohr. A, Hinneburg. A, Schult. R, and Spiliopoulou. M 2009. “Topic evolution in a stream of documents”. In SDM.

[20] Matthew D. Homan, David M. Blei, and Frances Bach 2010. “Online learning for latent dirichlet allocation”. In NIPS, 2010.

[21] Blei. D and Laferty. J 2006. “Dynamic topic models”. In ICML, 2006.

[22] AlSumait. L, Barbara. D, and Domeniconi.C 2008. “On-line Emerging Topics in IBM Tweets lda: Adaptive topic models for mining text streams”. In ICDM.

[23] Prathima, Y., & Supreethi, K. P. 2011. “A survey paper on concept based text clustering”, International Journal of Research in IT & Management, vol. 1 No.3,pp. 45–60.

[24] Frakes, W. B., & Fox, C. J. 2003. Strength and Similarity of Affix Removal Stemming Algorithms ACMSIGIR Forum, pp. 26–30.

[25] Jayabharathy. J and Kanmani. S, “Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature”, Journal on Decision Analytics, Springeropen, Vol.1, Issue.3. doi:10.1186/2193-8636-1-3

[26] Steinbach, M., Karypis, G., & Kumar, V. 2000. *A Comparison of Document Clustering Techniques* (pp. 1–2). International Conference on Data Mining: Knowledge Discovery and Data Mining (KDD) Workshop on Text Mining.

[27] Huang. A .2008, “Similarity Measures for Text Document Clustering”, NZCRSC’08, April 2008.