IJCTA, 9(36), 2016, pp. 395-407

Contextual Abstraction Based Clustering Technique for Effective Text Document Mining

S.Vadivukkarasi* and Puniethaa Prabhu**

Abstract : Document clustering is considered to be the essential process in grouping the unsupervised documents for effectual applications in text mining and information retrieval. Recently, many research works has been developed for text document clustering. However, performance of clustering the text document is not effective. In order to overcome such limitation, a novel Contextual Abstraction based Document Clustering (CADC) Technique is proposed in this paper. The CADC technique is designed to improve the performance of document clustering and information gain on clustering of multidimensional data. Proposed CADC technique based on the analysis of both the sentence and document. The Proposed CADC technique comprises of two parts namely an abstraction based analysis of terms and a contextual abstraction based similarity measure. The concept which contributes to the sentence semantics is evaluated with respect to its significance at the sentence and document levels. The CADC technique can efficiently discover significant matching terms, either words or phrases, of the documents according to the semantics of the abstraction (*i.e*. concept). The similarity between documents based on contextual abstraction based similarity measure which is used for matching concepts between documents. Thus, proposed CADC technique is efficiently improves the document clustering accuracy and reduces document retrieval time in an effective manner. The performance of proposed CADC technique is analyzed with the metrics such as document abstract similarity, document clustering accuracy and information gain and document retrieval time. Experimental analysis shows that the CADC technique is able to improve the document clustering accuracy by 13% and also reduce the document retrieval time by 34% when compared to the state-of-the-art works.

Keywords : Document, clustering, Contextual, text mining, concept-based similarity measure.

1. INTRODUCTION

Document clustering is application of cluster analysis to textual documents. Many clustering algorithm is developed to overcome difficulties, when cluster hiding in subspaces. Based on an effective clustering algorithm, similarity measures on diverse document collections are developed. However by using different clustering algorithm, performance gets affected and task increases with high computational complexity. Recently, most of research works has been designed for performing document clustering. For example, semi supervised text clustering algorithm called as Seeds Affinity Propagation (SAP) was designed in [1] to reduce the computing complexity of text clustering and to improve the accuracy of text clustering. Though, SAP did not consider any structural information and all features and vectors are not computed concurrently but one at a time. MultiViewpoint-based Similarity (MVS) and two related clustering methods was presented in [2] to significantly improved clustering performance. However, Clustering algorithm performs clustering task with high computational complexity.

* Assistant Professor, Department of Computer Applications, K.S.Rangasamy College of Technology, Tiruchengode-637215.

** Professor, Department of Biotechnology, K.S.Rangasamy College of Technology, Tiruchengode – 637 215.

Semantically Document Clustering algorithm was introduced in [3] by integrating the features of Directed Ridge Regression (DRR), Fuzzy relational Hierarchical clustering (FHC) and Conceptual clustering methods. Semantically Document Clustering algorithm presents more precise document clustering with the support of fuzzy hierarchical rules. Semantic clustering and feature selection method was designed in [4] for enhancing the clustering and feature selection mechanism with semantic relations of the text documents. Concept based mining model was presented in [5] that includes of four constituents for improving the performance of text clustering quality.

A document clustering approach was developed in [6] depends on the DPM model that groups documents into an arbitrary number clusters where document words are separated according to their usefulness to discriminate the document clusters. But, the document clustering quality is poor. Projective nonnegative matrix factorization (PNMF) method called as automated graph regularized projective nonnegative matrix factorization (AGPNMF) was designed in [7] for improving the clustering performance of documents. The design of AGPNMF was used to expand the original PNMF by means of integrating the automated graph regularized constraint into the PNMF decomposition.

Semi supervised spectral clustering method called as SSNCut was introduced in [8] which can incorporate both ML and CL constraints for combining different information for document clustering. Firefly algorithm (FA) was presented in [9] that discover documents which has the maximum light intensity in a search space and characterizes it as a centroid. In FA, Documents that are similar to the centroid are located into one cluster and dissimilar in the other which results in improved precision and reduced computational complexity.

2. RELATED WORKS

A new hybrid algorithm was developed in [10] for document clustering based on cuckoo search optimization integrated with k-means method which results in improves the quality of clustering. Efficient document clustering was presented in [11] with the aid of hybridizing the traditional partitioning clustering techniques such as K-Means and Fuzzy-C Means with PSO for improving the performance of document clustering. A generalized approach was designed in [12] for clustering a set of given documents or text files or software components for reuse depends on hybrid XOR function described for the intention of discovering degree of similarity among two document sets or any two software components.

In [13], the key challenges and the significant problems in designing extraction features and clustering algorithms were described. An evolutionary approach using genetic algorithm was introduced in [14] for text document clustering. A modified WordNet-based semantic similarity measure was designed in [15] for word sense disambiguation and lexical chains are employed to mine core semantic features that express the topic of documents.

In [16], three dynamic document clustering algorithms such as TMARDC, CCMARDC and CCFICA were designed to capture the technical correlation between the documents which results in increased document clustering accuracy. Text Document Clustering Using Dimension Reduction Technique was developed in [17] to reduce the intra cluster distance between documents when maximizing the inter cluster distance by using an appropriate distance measure between documents. Idiom Semantic Based Mining Model was introduced in [18] where the documents are clustered based on their meaning using the techniques of idiom processing, semantic weights with the aid of Chameleon clustering algorithm.

Concept based indexing technique with dynamic weight was planned in [19] to effectively identify the leading concept of the document based on the back ground knowledge provided by the MeSH concept hierarchy. A Concept-based document similarity model was developed in [20] to determine the similarities of documents based on the Suffix Tree Document (STD) model.

Based on the aforementioned techniques and methods presented, in this work we propose a novel framework called Contextual Abstraction based Document Clustering (CADC) Technique is designed. The key objective of CADC Technique is to improve the document clustering accuracy and to improve

information gain of clustering of multidimensional data and also to reduce the document retrieval time. The proposed CADC technique is includes two main parts namely abstraction based analysis of terms and Contextual abstraction based similarity measure. Abstraction based analysis of terms evaluates the semantic structure of each sentence to capture the sentence concepts in given input (*i.e*. text document). Besides, contextual abstraction based similarity measure in CADC technique determines the similarity between documents which is used for clustering sets of documents with aiming at improving the document clustering accuracy.

The rest of the paper organized as follows. In Section 2, a summary of different document clustering are explained. In Section 3, the proposed CADC Technique is described with the help of neat architecture diagram. In Section 4, simulation environment is presented with detailed analysis of results explained in Section 5. In Section 6, the concluding remarks are included.

3. CONTEXTUAL ABSTRACTION BASED DOCUMENT CLUSTERING (CADC) TECHNIQUE

The Contextual Abstraction based Document Clustering (CADC) Technique is designed to improve the document clustering accuracy and information gain on clustering of multidimensional data in text document mining. Proposed CADC Technique captures the semantic structure of each term within a sentence and a document, rather than the frequency of the term within a document only. In CADC technique, each term that has a semantic role in the sentence is called as an abstraction (i.e. concept). Abstraction can be either words or phrases and are entirely dependent on the semantic structure of the sentence. When a new document is introduced to the system, the Proposed CADC Technique can detect a concept match from this document to all the previously processed documents in data set by scanning new document and extracting the matching concepts.

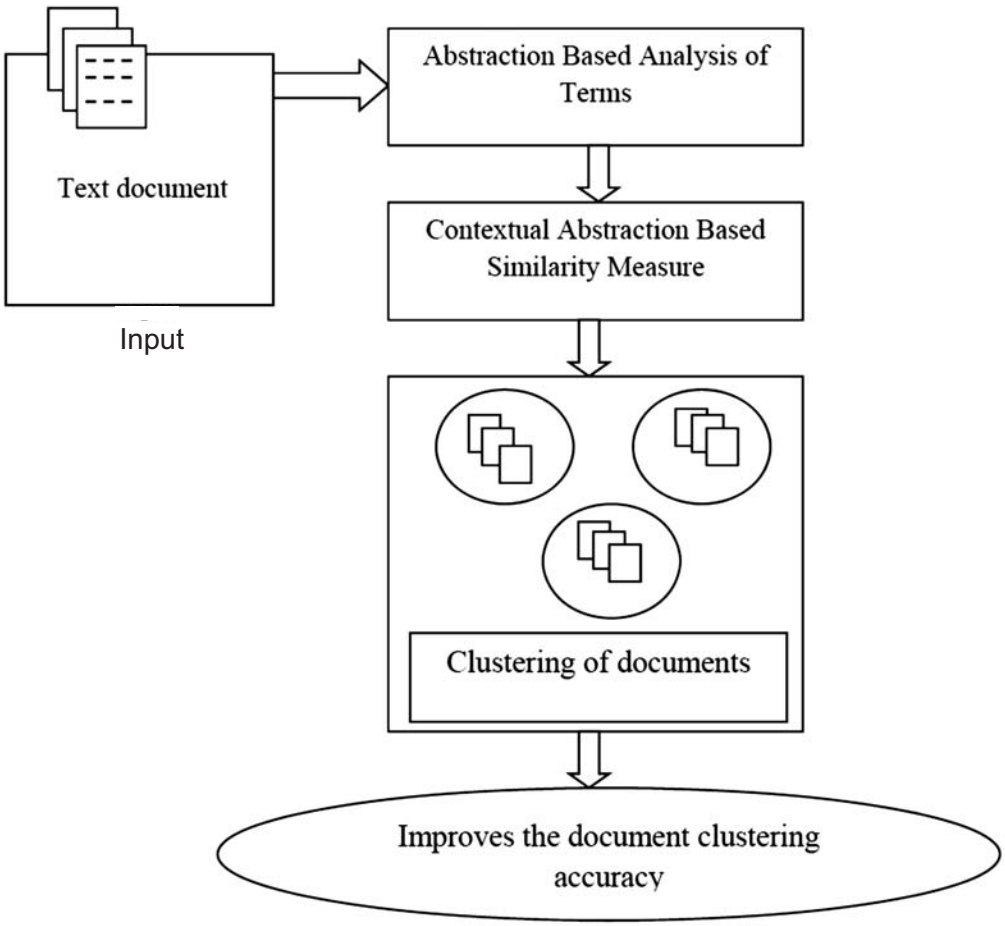

Figure 1: Contextual Abstraction based Document Clustering

In CADC Technique, the contextual abstraction based similarity measure is used for performing concept based term matching. The contextual abstraction based similarity measure outperforms other similarity measures that are based on concept analysis models of the document dataset only. The similarity between documents is based on a combination of abstraction based term analysis similarity within a sentence and contextual abstraction based term analysis similarity within a document. The block diagram of Contextual Abstraction based Document Clustering (CADC) Technique is shown in below Figure 1.

As shown in Figure 1, proposed CADC technique is initially takes the text document as input and then analyzes the semantic structure of each sentence to collect the sentence concepts in given input with the help of abstraction based analysis of terms. Then, abstraction based analysis of terms is performed to evaluate each concept at the sentence and document level. After that, contextual abstraction based similarity measures the importance of each concept with respect to the semantics of the sentence and the topic of document. In CADC Technique, contextual abstraction based similarity measure performs concept matching among documents which in turn improves the accuracy of document clustering and information gain on clustering of multidimensional data.

3.1. Abstraction Based Analysis of Terms

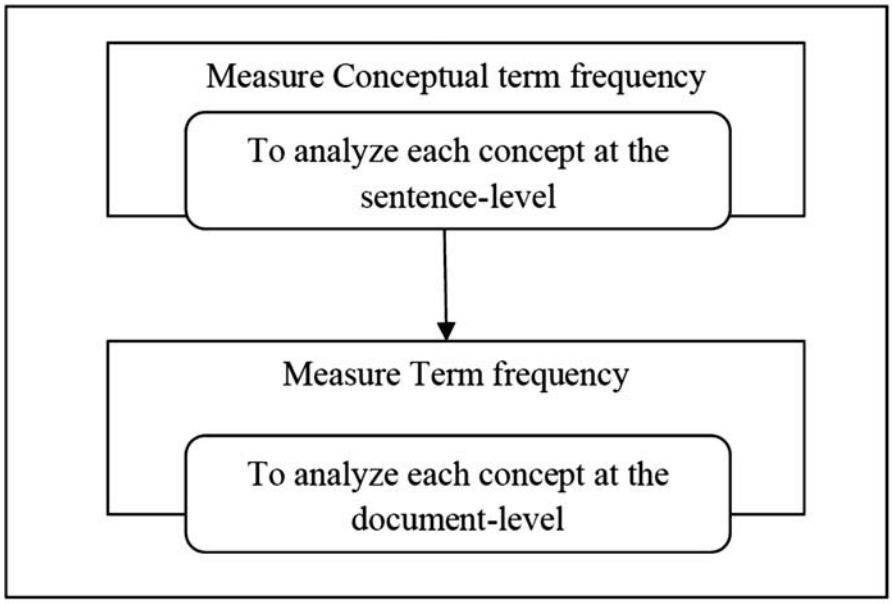

Figure 2: Abstraction Based Term Analysis in CADC Technique

In CADC Technique, The objective of this task is to achieve an abstraction based term analysis (statement or phrase) on the sentence and document levels rather than a single term analysis in the document set only. In abstraction based term analysis, conceptual term frequency '(*ctf*)' is evaluated to determine each concept at the sentence level. The '*ctf*' represent the number of occurrences of concept '*conc*' in verb argument structures of sentence sens. The concept conc which frequently appears in different verb argument structures of the same sentence sens has the primary role of contributing to the meaning of sens. The conceptual term frequency '*ctf*' value of concept '*conc*' in document '*doc*' is mathematically formulated as,

$$ctf = \frac{\sum_{n=1}^{Sens_{Sn}} ctf_n}{Sens_n} \quad (1)$$

From (1), $Sens_n$ is the total number of sentences that contain concept conc in document doc. To analyze each concept at the document level, the term frequency *tf* is measured in CADC Technique. In abstraction based term analysis, the term frequency *tf* refers number of occurrences of a concept (statement or phrase) conc in the original document. The term frequency *tf* is mathematically formulated as below,

$$tf_i = \frac{conc_{i,f}}{\sum_k conc_k} \quad (2)$$

Form (2), $conc_{(i,j)}$ denotes the number of occurrence of considered concept in document doc_j and $\sum_k conc_k$ signifies the sum of number of occurrences of all concept in document doc_j. The abstraction based term analysis using CADC Technique is shown in below Figure 2.

As shown Figure 2, CADC Technique measures the conceptual term frequency and term frequency in order to evaluate the each concept at a sentence and document level. The algorithmic process of abstraction based term analysis is shown in below Figure 3.

Input: Dataset 'DS', Documents '$doc = doc_1, doc_2,\ldots, doc_n$', Sentence '$sens = sens_1, sens_2,\ldots, sens_n$' , Concept $conc = conc_1, conc_2,..conc_n$
Output : Obtains sentence and document level concept
1. Begin
2. For each Dataset 'DS'
3. For each Documents '*doc*'
4. Measure conceptual term frequency (*ctf*) using (1)
5. Measure term frequency (*tf*) using (2)
6. End for
7. End for
8. End

Figure 3: Algorithmic Process of Abstraction Based Term Analysis

As shown in Figure 3, the algorithmic process of abstraction based term analysis contains two steps as follows. Initially, abstraction based term analysis algorithm computes conceptual term frequency with aiming at analyzing each concept at the sentence level. After that, abstraction based term analysis algorithm computes the term frequency with aiming at evaluating each concept at the document level. Thus, proposed CADC Technique is easily obtains the sentence and document level concepts in each document.

3.2. Contextual Abstraction Based Similarity Measure

In CADC Technique, abstraction (*i.e*. concepts) expresses local context information that is important in determining an accurate similarity between documents. A contextual abstraction based similarity measure based on matching concepts at the sentence and document levels. The contextual abstraction based similarity measure relies on two critical aspects. Initially, the analyzed labeled terms are the concepts that collect the semantic structure of each sentence. Secondly, the frequency of a concept is utilized to compute the contribution of the concept to the meaning of the sentence and main topics of the document. These aspects are evaluated by the proposed contextual abstraction based similarity measure which determines the importance of each concept at the document level with the support of *tf* measure and at the sentence level by *ctf* measure. The contextual abstraction based similarity measure is used information extracted from the abstraction based term analysis algorithm to evaluate the similarity between documents.

The contextual abstraction based similarity measure is a function of the subsequent factors such as (1) the number of matching concepts (match) in the verb arguments structures in each document (*doc*). (2) The total number of sentences (*sens*) in each document *doc*. (3) The total number of the labeled verb argument structures (*v*) in each sentences, (4) the term frequency tf_i of each concept $conc_i$ in each document *doc* where $i = 1, 2,..., m$. 5) The conceptual term frequency ctf_i of each concept $conc_i$ in sentence

for each document *doc* where $i = 1, 2,..., m$, 6) the length (leng) of each concept in the verb argument structure in each document *doc*. And the length (*s*) of all verb argument structure that includes a matched document concept.

The conceptual term frequency (*ctf*) is a considerable factor in evaluating the contextual abstraction based similarity measure among documents. The more repeated the concept occurs in verb argument structures of a sentence in a document, the documents are higher abstractly similar. The contextual abstraction based similarity between two documents doc_1 and doc_2 is mathematically formulated as,

$$\text{similarity}_c = \sum_{i=1}^{m} \max\left(\frac{\text{leng}_i}{s_{i_1}}, \frac{\text{leng}_i}{s_{i_2}}\right) * w_{i_1} * w_{i_2} \quad (3)$$

Where, $w_{i_1} = tfw_{i_1} + ctfw_{i_1}$, $w_{i_2} = tfw_{i_2} + ctfw_{i_2}$

From (3), the concept based weight of concept i_1 in document doc_1 is represented by w_{i_1}. In determining w_{i_1}, the tfw_{i_1} value denotes the weight of concept *i* in the first document doc_1 at the document level and the $ctfw_{i_1}$ value denotes the weight of the concept *i* in the first document doc_1 at the sentence level based on the contribution of concept *i* to the semantics of the sentences in doc_1. The sum between the two values of tfw_{i_1} and $ctfw_{i_1}$ presents an accurate measure of the contribution of each concept to the meaning of the sentences and to the topics mentioned in a document. The term w_{i_2} is applied to the second document doc_2. Equation 3 gives a higher score, as the matching concept length approaches the length of its verb argument structure, since this concept tends to hold more conceptual information related to the meaning of its sentence.

The weight of concept *i* in the first document doc_1 at the document-level tfw_{i_1} is mathematically formulated as,

$$tfw_{i_1} = \frac{tf_{ij_1}}{\sqrt{\sum_{j=1}^{cn_1} (tf_{ij_2})^2}} \quad (4)$$

From (4), the tf_{ij_1} value is normalized by the length of the document vector of the term frequency tf_{ij} in the first document doc_1 where $j = 1, 2,...,cn_1$ whereas cn_1 indicate the total number of concepts which has a term frequency value in the document doc_1. Then, the weight of concept *i* in the first document doc_1 at the sentence-level $ctfw_{i_1}$ is mathematically formulated as,

$$ctfw_{i_1} = \frac{ctf_{ij_1}}{\sqrt{\sum_{j=1}^{cn_1} (ctf_{ij_1})^2}} \quad (5)$$

From (5), the ctf_{ij_1} value is normalized by the length of the document vector of the conceptual term frequency ctf_{ij} in the first document doc_1 where $j = 1, 2,...,cn_1$ whereas cn_1 represent the total number of concepts which has a conceptual term frequency value in the document doc_1. The same normalization equations are applied to the weights of the concepts in the second document doc_2. The algorithmic process of contextual abstraction based document clustering technique is shown in below Figure 4.

As shown in Figure 4, contextual abstraction based document clustering algorithm initially takes new document as input. Then, abstraction based term analysis algorithm explains the process of measuring *tf* and *ctf* of the matched concepts in the documents. Afterward, each concept in the new document is matched with the other concepts in the previously processed documents in Bag of words dataset. In proposed contextual abstraction based document clustering algorithm, each concept in the new document is matched with the other concepts by means of keeping a matching concept document list. In CADC technique, a matching concept document list contains the entry for all of the previous documents which shares a concept with the new document. After the processing of document is completed, matching concept document list outputs all the matching concepts among the new document and any previous document that shares at least one concept with the new document. Therefore, contextual abstraction based document clustering algorithm is capable of matching each

concept in a new document with all the previously processed documents with reduced document retrieval time. This in turn improves the document clustering accuracy in an effective manner.

Input : Dataset 'DS', Documents '$doc_i = doc_1, doc_2, \ldots, doc_n$', Sentence '$sens = sens_1, sens_2, \ldots, sens_n$', Concept $conc_i = conc_1, conc_2, .. conc_n$, doc_{new} is a new Document, $match_{list}$ is an empty List ($match_{list}$ is a matching concept document list), conceptual term frequency '(ctf_i)', term frequency 'tf_i', weight of concept in document at the document-level '*tfw*', weight of concept in document at the sentence-level '*ctfw*'

Output : Matched concepts document list $match_{list}$ and improved the document clustering accuracy

1. **Begin**
2. get doc_{new} is an input
3. **For** each sentence *sens* in doc_{new} **do**
2. $conc_i$ is a new concept in sentence *sens*
4. **For** each concept $conc_i \in \{conc_1, conc_2, .. conc_n\}$ in sentence *sens* **do**
5. compute ctf_i and tf_i using abstraction-based term analysis algorithm
6. End for
7. For each doc_k where $k = \{0,1,.., doc_{i-1}\}$, $conc_i$ exist **do**
8. For each concept $conc_j = \in \{conc_1, conc_2, .. conc_m\}$ in sentence *sens* **do**
9. **if** ($conc_i == conc_j$) **then**
10. measure $tfw = avg(tf_i, tf_j)$
11. measure $ctfw = avg(ctf_i, ctf_j)$
12. add concept matches document to $match_{list}$
13. **End if**
14. **End for**
15. **End for**
16. **End for**
17. **End**

Figure 4: Contextual Abstraction Based Document Clustering Algorithm

4. EXPERIMENTAL SETUP

The Contextual Abstraction based Document Clustering (CADC) Technique is implemented using Java Language. The experimental evaluation of the CADC Technique is done with standard Bag of words dataset from UCI repository. The Bag of words dataset includes five text collections in the form of bags-of-words. For each text collection, D is the number of documents, W is the number of words in the vocabulary, and N is the total number of words in the collection (below, NNZ is the number of nonzero counts in the bag-of-words). After performing tokenization and removal of stopwords, the vocabulary of unique words was short end by only keeping words that occurred more than ten times. Individual document names (*i.e*. an identifier for each docID) are not presented for copyright reasons.

The training model for bag of words data sets have no class labels and for copyright reasons no file names or other document-level metadata are present. Bag of words dataset has been selected because it gives a clear picture that helps in analyzing the documents on multidimensional data and is more suitable for clustering and topic modelling experiments. For each text collection, the dataset presents docword.*.txt where the bag of words file is in sparse format and vocab.*.txt in the form of the vocab file. The total number of attributes included in bag of words dataset is 1,00,000. For experimental purpose, we reviewed using 70 attributes.

The CADC Technique is conduct experimental work on metrics such as document abstract similarity, document clustering accuracy, information gain and document retrieval time. The results of proposed CADC technique are compared against with the existing methods such as Seeds Affinity Propagation (SAP) [1] and Multiviewpoint-based Similarity (MVS) [2] respectively.

5. DISCUSSION

To validate the efficiency of proposed Contextual Abstraction based Document Clustering (CADC) Technique, the comparison is made with existing two methods namely Seeds Affinity Propagation (SAP) [1] and Multiviewpoint-based Similarity (MVS) [2]. The performance of CADC technique is evaluated along with the following metrics.

5.1. Measure of document abstract similarity

In CADC technique, document abstract similarity is defined as the ratio of number of matching concepts in new document with other documents to the total number of concepts in new document. The document abstract similarity is measured in terms of percentage (%) and mathematically formulated as,

$$\text{Document abstract similarity} = \frac{\text{Number of matching concepts in new document with other documents}}{\text{Total number of concepts in new document}} * 100 \quad (6)$$

When the document abstract similarity is higher, the method is said to be more efficient.

Table 1

Tabulation for Document Abstract Similarity

No. of concepts	*Document abstract similarity (%)*		
	SAP	*MVS*	*CADC*
5	63.26	76.65	82.65
10	65.23	78.54	85.36
15	68.59	81.35	87.56
20	70.12	83.65	89.24
25	73.26	86.24	92.15
30	75.59	88.45	94.26
35	78.15	91.25	97.25

Table 1 shows the result analysis of document abstract similarity using three different methods namely, CADC technique, SAP [1], MVS [2]. From the table value, it is descriptive that the document abstract similarity using proposed CADC technique is higher as compared to the other existing methods.

Figure 5 shows the impact of document abstract similarity using three different methods versus different number of concepts taken in the range of 5-35. As shown in figure, the document abstract similarity using proposed CADC technique provides better performance as compared to other existing methods namely SAP [1], MVS [2].Besides, while increasing the number of concepts, the document abstract similarity is also increased using all three methods. But, comparatively document abstract similarity using proposed CADC technique is higher. This is due to contextual abstraction-based similarity measure in CADC

technique. In CADC technique, a contextual abstraction-based similarity measure based on matching concepts at the sentence and document levels which in turn improves the document abstract similarity. In addition, the contextual abstraction-based similarity measure is employed information extorted from abstraction-based term analysis algorithm to estimate the similarity between documents which results in improved document abstract similarity in a significant manner. As a result, proposed CADC technique is improved the document abstract similarity by 21% when compared to SAP [1] and 7% when compared to MVS [2] respectively.

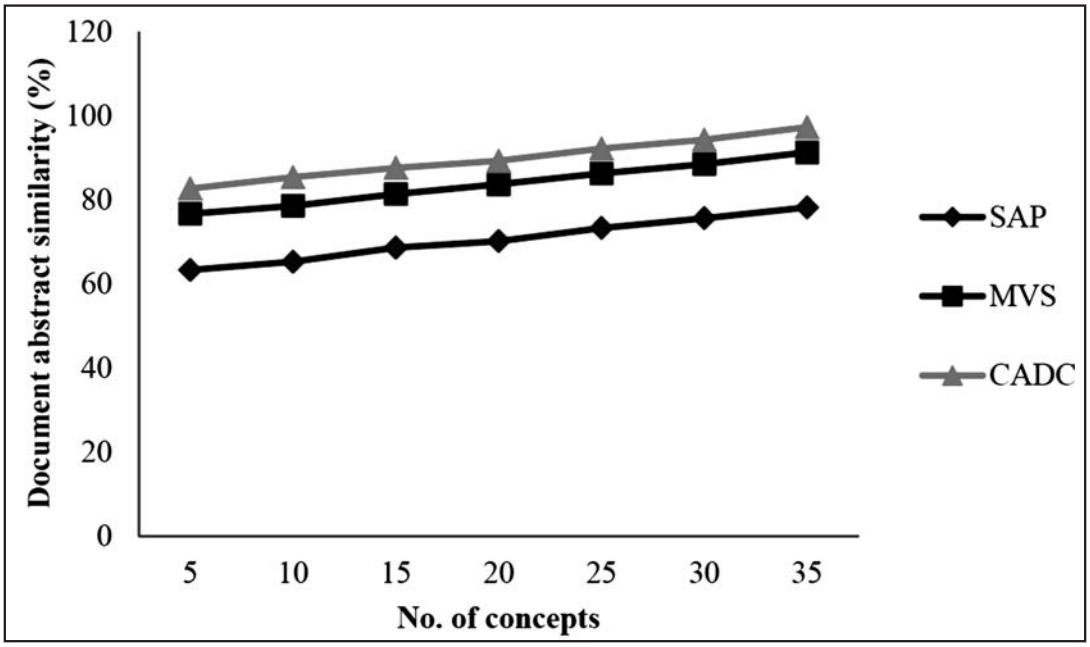

Figure 5: Measure of Document Abstract Similarity

5.2. Measure of Information Gain

In CADC technique, Information gain computes the amount of information that's gained by knowing the value of the attribute. Information gain is defined as the entropy of the distribution before the split minus the entropy of the distribution after it. Smaller the value of entropy, higher the information gain is said to be. Information gain is measured in terms of percentage (%) and formulated as,

$$IG = (\text{Entropy of whole dataset-Entropy of an attribute}) * 100 \quad (7)$$

When the information gain is higher, the method is said to be more efficient. The information gain rate 'IG' is obtained using the entropy of whole dataset and an attribute from bags of word dataset.

Table 2

Tabulation for Information Gain

No. of attributes	*Information gain (%)*		
	SAP	*MVS*	*CADC*
10	76	84	90
20	77	86	92
30	80	89	95
40	77	87	92
50	78	88	93
60	81	90	96
70	82	91	97

The information gain of document clustering using three different methods namely, CADC technique, SAP [1], MVS [2] is shown in Figure 6. From the table value, it is expressive that the information gain using proposed CADC technique is higher as compared to the other existing methods.

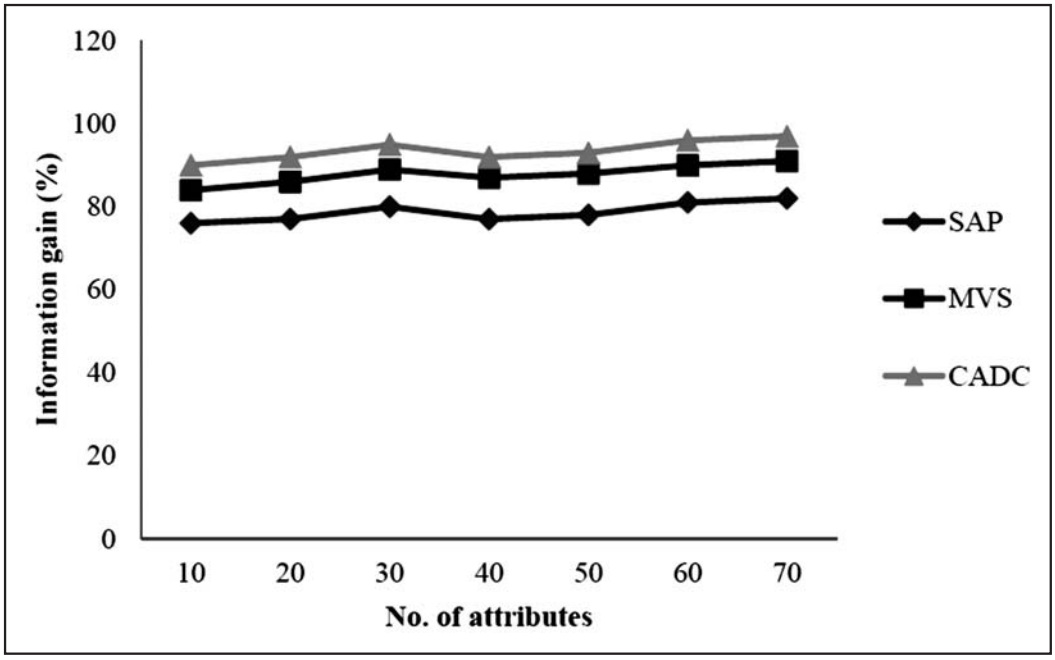

Figure 6: Measure of Information Gain

Figure 6 demonstrates the impact of information gain using three different methods versus different number of attributes in the range of 10-70. As shown in figure, the information gain using proposed CADC technique provides better performance as compared to other existing methods namely SAP [1], MVS [2]. This is because of abstraction-based term analysis algorithm and contextual abstraction-based similarity measure in CADC technique where it calculates conceptual term frequency and term frequency in order to analyze each concept at the sentence-level and document-level which in turn helps in improving the information gain. Besides, the frequency of a concept is employed to evaluate the contribution of the concept to the meaning of the sentence and the key topics of the document which in turn increase the information gain in an efficient manner. As a result, proposed CADC technique is improved the information gain by 16% when compared to SAP [1] and 6% when compared to MVS [2] respectively.

5.3. Measure of document retrieval time

In CADC technique, the document retrieval time measures the amount of time taken to cluster the document based on concepts. The document retrieval time is measured in terms of milliseconds (*ms*). Lower the document retrieval time, the method is said to be more efficient.

Table 3

Tabulation for document retrieval time

No. of concepts	*Document retrieval time (ms)*		
	SAP	*MVS*	*CADC*
5	34	26	20
10	46	38	27
15	58	50	39
20	70	62	51
25	82	74	63
30	94	86	75
35	106	98	87

Table 3 illustrates the result analysis of document retrieval time using three different methods namely, CADC technique, SAP [1], MVS [2]. We consider the framework with different number of concepts in the range 5-35 for experimental purpose using Java Language. From the table value, it is descriptive that the document retrieval time using proposed CADC technique is lower as compared to the other existing methods.

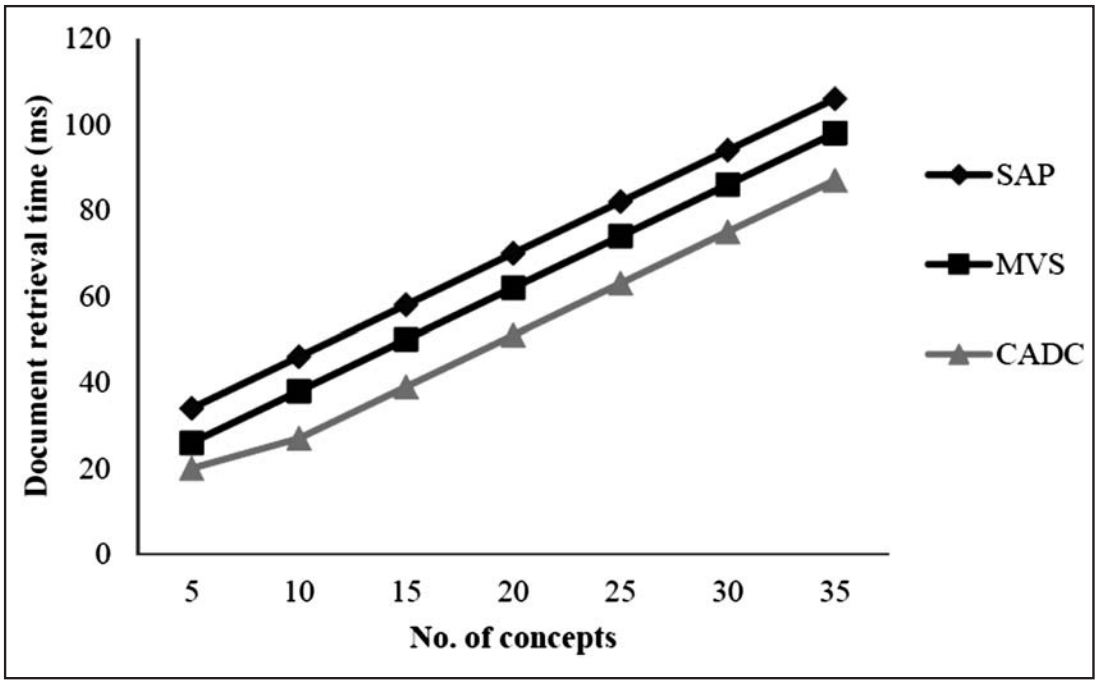

Figure 7: Measure of document retrieval time

Figure 7 reveals the impact of document retrieval time using three different methods versus different number of concepts in the range of 5-35. As shown in figure, the document retrieval time using proposed CADC technique provides better performance as compared to other existing methods namely SAP [1], MVS [2]. Besides, while increasing the number of concepts, the document retrieval time is also gets increased using all three methods. But, comparatively document retrieval time using proposed CADC technique is reduced. This is because of contextual abstraction based document clustering algorithm in CADC technique where it efficiently match each concept in a new document with all the previously processed documents with reduced document retrieval time. As a result, proposed CADC technique is reduced the document retrieval time by 43% when compared to SAP [1] and 24% when compared to MVS [2] respectively.

5.4. Measure of document clustering accuracy

In CADC technique, document clustering accuracy is defined as the ratio of number of correctly clustered documents based on abstraction to the total number of concepts in new document. The document clustering accuracy is measured in terms of percentage (%) and mathematically formulated as,

$$\text{Document clustering accuracy} = \frac{\text{Number of correctly clustered documents based on concepts}}{\text{Total number of concepts in new document}} * 100 \quad (7)$$

When the document clustering accuracy is higher, the method is said to be more efficient.

Table 4 and Figure 10 shows the impact of document clustering accuracy using three different methods namely, CADC technique, SAP [1], MVS [2]. As shown in figure, document clustering accuracy using CADC technique is provides better performance as compared to two other existing methods namely SAP [1], MVS [2]. This because of the contextual abstraction-based similarity measure is employed in CADC technique. With the support of contextual abstraction-based similarity measure, CADC technique

is significantly performs concept matching and concept-based similarity calculations between documents. This in turn improves the document clustering accuracy in an effective manner. As a result, proposed CADC technique is improved the document clustering accuracy by 18% when compared to SAP [1] and 7% when compared to MVS [2] respectively.

Table 4

Tabulation for document clustering accuracy

Methods	*Document clustering accuracy (%)*
SAP	76.23
MVS	85.69
CADC	92.45

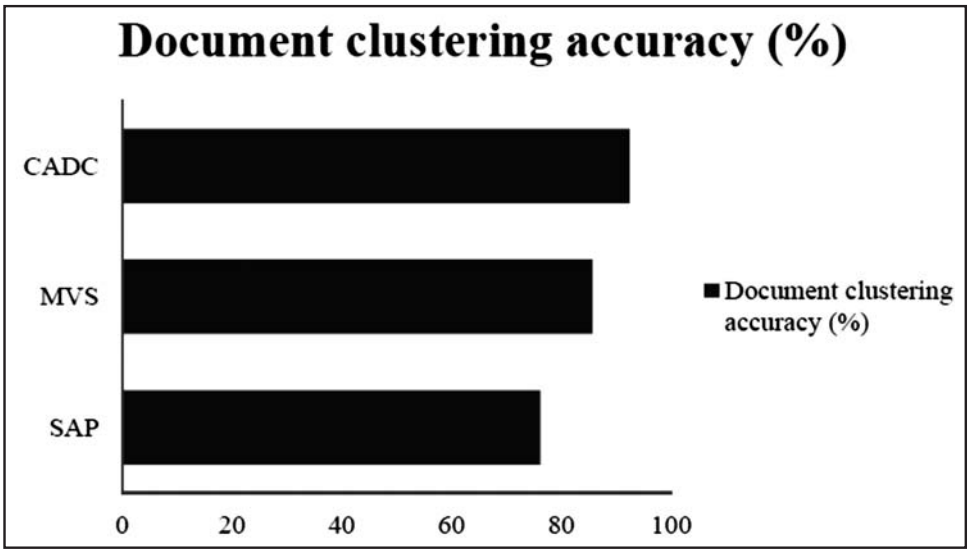

Figure 8: Measure of document clustering accuracy

6. CONCLUSION

In this paper, an effective novel framework is designed called as Contextual Abstraction based Document Clustering (CADC) to improve the document clustering accuracy and information gain on clustering of multidimensional data and to reduce the document retrieval time. The CADC technique is initially takes the text document as input. Then, CADC technique calculated the conceptual term frequency and term frequency to evaluate each concept at the sentence level and document level with aid of abstraction based analysis of terms. Finally, contextual abstraction based similarity measure accomplishes concept matching and concept based similarity calculations between documents in an effective manner which results in improved document clustering accuracy. The proposed CADC technique is implemented by using Java Language. The performance of CADC technique is tested with the metrics such as document abstract similarity, document clustering accuracy, information gain and document retrieval time. With the experiments conducted for CADC technique, it is observed that the document clustering accuracy is provided more accurate results as compared to existing methods. The experimental results show that CADC technique is provides better performance with an improvement of document clustering accuracy by 13% and also reduced document retrieval time by 34% when compared to state-of-the-art works.

7. REFERENCES

1. Renchu Guan, Xiaohu Shi, Maurizio Marchese, Chen Yang, and Yanchun Liang, "Text Clustering with Seeds Affinity Propagation", IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 4, April 2011
2. Duc Thang Nguyen, Lihui Chen and Chee Keong Chan, "Clustering with Multiviewpoint-Based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, June 2012.

3. R.Nagaraj, X.Agnise Kalarani, "Semantically Document Clustering Using Contextual Similarities", International Journal of Applied Engineering Research, Volume 11, Number 1 (2016) pp 71-76
4. J.Sathya Priya, S.Priyadharshini, "Clustering Technique in Data Mining for Text Documents", International Journal of Computer Science and Information Technologies, Vol. 3(1), 2012, 2943-2947
5. Pradnya Randive, Nitin Pise, "Improving Text Clustering Quality by Concept Mining", Journal of Engineering Research and Applications, Vol. 3, Issue 5, Sep-Oct 2013, pp.1701-1704
6. Ruizhang Huang, Guan Yu, Zhaojun Wang, Jun Zhang, and Liangxing Shi, "Dirichlet Process Mixture Model for Document Clustering with Feature Partition", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 8, August 2013
7. Xiaobing Pei, Tao Wu, Chuanbo Chen, "Automated Graph Regularized Projective Nonnegative Matrix Factorization for Document Clustering" IEEE Transactions on Cybernetics Vol.44, No. 10, January 2014
8. Jun Gu, Wei Feng, Jia Zeng, Hiroshi Mamitsuka, and Shanfeng Zhu, "Efficient Semisupervised MEDLINE Document Clustering With MeSH-Semantic and Global-Content Constraints", IEEE Transactions On Cybernetics, Vol. 43, No. 4, August 2013
9. Athraa Jasim Mohammed , Yuhanis Yusof, Husniza Husni, "Weight-Based Firefly Algorithm for Document Clustering", Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013),Volume 285 of the series Lecture Notes in Electrical Engineering, Springer, pp 259-266, 2013
10. Ishak Boushaki Saida1, Nadjet Kamel, and Bendjeghaba Omar, "New Hybrid Algorithm for Document Clustering Based on Cuckoo Search and K-means", Recent Advances on Soft Computing and Data Mining, Springer, Volume 287 of the series Advances in Intelligent Systems and Computing, pp 59-68, 2014
11. Stuti Karol, Veenu Mangat, "Evaluation of text document clustering approach based on particle swarm optimization", Central European Journal of Computer Science, Springer, Volume 3, Issue 2, pp 69-90, June 2013
12. Vangipuram Radhakrishna, C. Srinivas, C.V.Guru Rao, "Document Clustering Using Hybrid XOR Similarity Function for Efficient Software Component Reuse", Procedia Computer Science, Elsevier, Volume 17, Pages 121-128, 2013
13. Qusay Bsoul, Juhana Salim, Lailatul Qadri Zakaria, "An Intelligent Document Clustering Approach to Detect Crime Patterns", Procedia Technology, Elsevier, Volume 11, 2013, Pages 1181-1187
14. Ruksana Akter, Yoojin Chung, "An Evolutionary Approach for Document Clustering", IERI Procedia, 2013 International Conference on Electronic Engineering and Computer Science (EECS 2013), Elsevier, Volume 4, 2013, Pages 370-375
15. Security in Wireless Sensor Networks: Key Management Module in EECBKM"Presented in International Conference on World Congress on Computing and Communication Technologies on Feb 27- & 28 and 1st march 2014, on St.Joseph college,Trichy.
16. Jayaraj Jayabharathy, Selvadurai Kanmani, "Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature", Decision Analytics, Springer, 2014
17. A. Sudha Ramkumar, B. Poorna, Text Document Clustering Using Dimension Reduction Technique", International Journal of Applied Engineering Research, Volume 11, Number 7 (2016), pp 4770-4774
18. B. Drakshayani, E. V. Prasad, "Semantic Based Model for Text Document Clustering with Idioms", International Journal of Data Engineering (IJDE), Vol. 4, Issue 1, 2013, pp 1-13
19. Sapna Gupta, Vikrant Chole, "Document Clustering Using Concept Weight", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, 2014, pg. 1207-1210
20. P. Perumal, R. Nedunchezhian, M. Indra Priya, "Concept-Based Document Similarity Based on Suffix Tree Document", International Journal of Computer Science & Engineering Technology (IJCSET), Vol. 3 No. 10, Oct 2012.