Clumping Of Legal Documents Using Key Assumptions Segment Based Approach

M.New Begin[1] J.AnithaJosephine[2] M.Divya Bharathy[3]
Hod pg scholar pg scholar
[1, 2, 3] Vel Tech Multi Tech Dr.Rangarajan Dr.Sakunthala
[1]newbegin_m@yahoo.com [2]cinderla90@gmail.com [3]mdbharathy@gmail.com

Abstract: **In the past decades, the availability of the legal judgements that accumulate huge volumes of textual documents. Technology adoption is a widespread phenomenon in courts of law around the world. Legal documents are often multi -topical, professional, domain-specific language, contained and carefully crafted, possess a broad and unevenly distributed coverage of legal issues. Due to the complexity of the document, the examiners find it difficult to analyze. The paper proposes a segment based approach to cluster the documents. This approach extracts the relevant information from the legal documents. The approach can be proceeding with the help of key assumption that frameworks the multi-topical document and it can be segmented into smaller single topic text units. This work reduces the time consumption while the examiner, examines the documents during the analysis. The segmented documents are finally ordered into a well disciplined form using hierarchical clustering algorithm.**

Keywords: ***Clustering, segmentation, hierarchical algorithm.***

I. INTRODUCTION

For many decades, the legal documents were manually examined. Any case actually gets into court, certain legal documents must be prepared and filed with the court. Legal documents are more complex in nature as it contains a lengthy discussion about the case. In recent years, there has been a growing availability of large electronic document collections that have increased the need for the development of mountable computational methods for their manually examined documents. By being able to organize large document collections into thematically consistent groups, has emerged as a key enabling technology and is used to cluster the documents and shows the main content of the case. One of the most significant applications of topic segmentation is the Topic Detection and Tracking (TDT) task, as described in [4]. In this paper we cintend to develop a document clustering approach for collections in which each document can potentially belong to multiple topics and it is segmented using segment based approach [8]. By using this approach, the automatic filtering of the data was involved in drawing out relevant information from the document. After retrieving the information from the document it automatically clustered based on its lawsuits. For clustering the documents the hierarchical algorithm is implemented[Fig:1]. In data mining, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering is a method for decision structures in a network. The method organise the network into a hierarchical of groups according to a individual weight function. The organised group of data is represented in a tree form called Dendrogram.

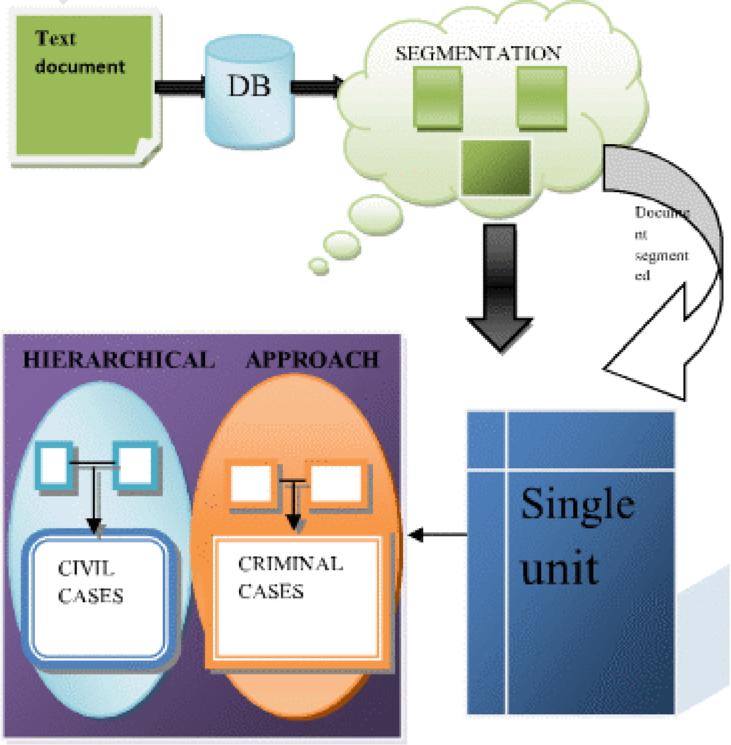

Figure 1: Architecture Diagram

II. RELATED WORK

The way to classify and grouping a document into segments is an essential task, including information retrieval, summarization, and text consideration. The main applications of topic segmentation is the Topic Detection and Tracking (TDT)[1].The fact that related or similar words and phrases tend to be repeated in topically coherent segments and segment boundaries often correspond to a change in the vocabulary[5]. Other approaches rely on complementary semantic knowledge extracted from dictionaries and thesauruses [1,3], or from collocations collected in large form, which use further field knowledge such as the use of the meaning. Clustering is an active area of research and a variety of algorithms have been developed in recent years [4]. Clustering algorithms can be categorized into agglomerative schemes or partitioning schemes and hard or soft clustering. The document clustering that process in a way that it clusters the text document in such a way it as a first step it break the multi paragraph into small segment and then it groups into a single unit. Their framework, an induction process is introduced to map the segment sets clustering solution to document level clusters in order to provide the user with a more useful organization of the input texts. In the case addressed by the authors, the text/topic segmentation algorithm assumes that documents are multi topical. It further assumes that document paragraphs represent coherent topics and topics shift on or around paragraph boundaries. Issues of scale are the prime differences between this work and our own.

III. PROBLEM STATEMENT

A. Document Organization and Browsing

The hierarchical group of documents into consistent categories can be very useful for systematic browsing of the document collection. A classical example of this is the Scatter/Gather method, which provides a systematic browsing technique with the use of clustered group of the document collection.

B. Corpus Summarization

Clustering techniques provide a logical summary of the collection in the form of cluster-digests or word-cluster, which can be used in order to provide summary insights into the overall content of the underlying quantity. Variants of such methods, especially sentence clustering, can also be used for document summarization, a topic, discussed in detail. The problem of clustering is that it arises in decline in the dimensions and in the topic model. Reduction in the dimensionality can summarize in several of corpus documents.

C. Document Classification

While clustering is inherently an unproven learning method, it has to improve the quality of the results in its supervised variant. In particular, word-clusters and co-training methods can be used in order to improve the categorization precision of supervised application with the use of clustering techniques.

IV. PROPOSED METHOD

In this section, we propose the segment based approach to cluster the document and then describe the proposed algorithm, hierarchical clustering algorithm which divides the lawsuits according to the phases and clusters the deed. The framework of the proposed work consists of

i) Segment based approach
ii) Hierarchy approach

A. Segment Based Approach

The key assumption underlying this segment-based document clustering outline is that multi-topic documents can be decomposed into smaller single-topic text units that the clustering of these segment-sets can lead to an overlapping clustering solution of the original documents that exactly reflects the collection of the topics that they contain [2, 3, 4, 5, and 6]. The various steps implicated in our segment-based document clustering agenda are illustrated. In the first step, each document is decomposed into a set of segments that correspond to the blocks of the initial document. These segments are measured to be topically consistent and to cover the entire document. In the second step, segments that are related to the same topic are grouped commonly by clustering the various segments of each document. The segment-sets are designed to merge the various topics that are situated at different parts of the original document. The clustering of each document's segments can be performed using an overlapping clustering method, with the later allowing for the assignment of a segment to multiple topically related segment sets. The segment-set identification is performed separately for each document, which is both computationally resourceful and also allows the use of various sophisticated clustering methods. In the third step, each of the set is kept as a single document by using the document clustering algorithm. The segment-set results the designs and identifies the multi-document which breaks in a single topic. Finally, the fourth step is considered to use the various topics identified by clustering the segmented document by obtaining the way out for overlapping the clustering documents. This can be done by suggesting the cluster for every record based on the set of segment. The following Diagram shows the segment based approach [Fig:2].

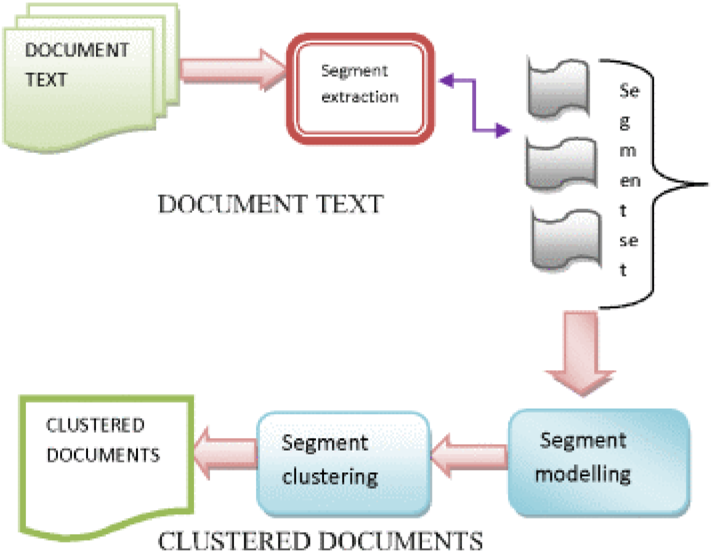

Figure 2: Document Segment Using Segment Based Approach

a) Segment – Set

This approach is based on breakdown of each document and to generate a set of views over the document according to the topics that it contains, each segment set should contains some parts of the original document, also it should contain the non-trivial quantity of the original document. As it contains different documents at different places, also it includes the text from different parts of the document. To achieve this, we developed a segment -set identification scheme that works into the following process (figure: 2). It breaks each document into paragraph based segments, and then clusters these segments into related groups according to their content. Each of these segment-clusters becomes a segment-set for the document. The paragraph based segment definition assumes that paragraphs can be easily identified in a document and that each paragraph is small enough to contain matter relevant to a single topic, since a paragraph is generally a self-contained unit of a discourse. Segment clustering algorithms that produce an overlapping clustering solution, which in turn leads to overlapping segment sets. By allowing the segment set, it increases the robustness of the approach as we rely less on the ability of the clustering algorithms to (i) correctly identify the number of topics present in a document and (ii) group together all the relevant segments. Specifically, we can cluster the segments in a relatively large number of overlapping clusters (i.e., more clusters than the expected topics). Due to this “over-clustering”, each cluster will tend to contain segments from a similar topic; and due to Overlapping , each cluster will still be sufficiently large to contain enough information about the topic.

B.HIERARCHICAL CLUSTERING

Document clustering is a process of organizing the documents into different clusters, such that the documents with in the cluster are more similar compare to the documents in the other cluster [7]. A hierarchical clustering is a sequence of partitions in which each partition is nested into the next partition in the sequence. It produces a set of nested clusters organized as a hierarchical tree. It can be visualized as a Dendrogram. It is a tree like structure which keeps the records in a combine and splitting of the document. A Dendrogram consists of layers of nodes, each representing a cluster. Lines connect nodes representing clusters which are nested into line another. Cutting a Dendrogram horizontally creates a clustering.Clustering obtained by cutting the dendrogram at a desired level each connected component forms a cluster[8].Hierarchical clustering can either be agglomerative or divisive depending on whether one proceeds through the algorithm by adding links to or removing links from the network[8].Fast and high-quality document clustering algorithms play an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. Then, starting with all the nodes if the group of connections are disconnected then we can pair the nodes according to the weight of the node. As associates are added, connected subsets start to group. This shows the structure of the network.

The components at each iterative step are always a subset of other structures [9]. Hierarchical clustering is an agglomerative (top down) clustering method. As its name suggests, the idea of this method is to build a hierarchical way of clusters, representing it between the individual and combining clusters of data based on similarity. In the initial stage of grouping the data, the algorithm will come across the two similar data points and merge it to form a new false-data point. Each of these steps takes the next two nearby data points and merges them. This process is generally continued until there is one large cluster containing all the original data points. Hierarchical clustering results in a "tree"[10], showing the relationship of all of the original points. The agglomerative approach is used as it the bottom-up method and it uses the nearest cluster algorithm to group it, where we start on with *n* singleton clusters and successively merge clusters to produce the other ones. The divisive approach can also be used which is a top-down method that splits the documents into separate group. In general, all agglomerative algorithms usually yield the same results if the clusters are compact and well separated. The space and time where c is the number of clusters and d is the distance between them.

Algorithm:1 Merging Document Algorithm

```
STEP 1: Start
STEP 2: initialize the cluster;
STEP 3:Where c' = n keep the distance as   d
 STEP 4: do
 STEP 5: c' = c' - 1
 STEP 6:  Find the next immediate  document with the distance
 STEP 7: combine $D_i$ and $D_j$
 STEP 8:  until c = c'
 STEP 9:  return c clusters
 STEP 10: End
```

Sometimes the clusters may found as flat cluster as in that situation we can cut the clusters buying this way,
D= arg d' min[RSS [d'] + d']
Using this formula we can cluster the document

V. Result discussion

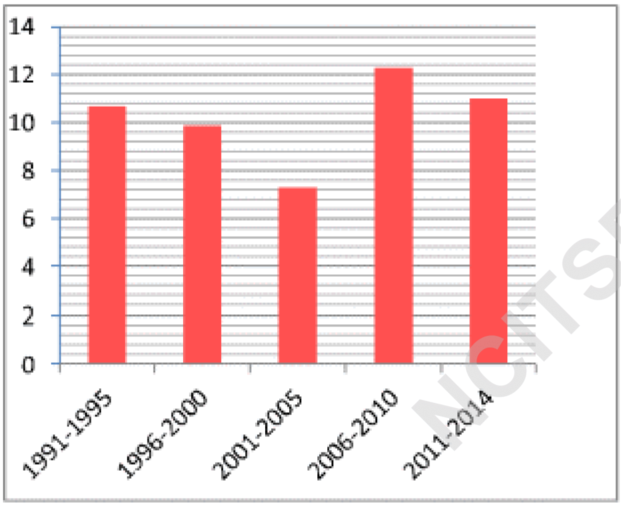

Figure 3: Before Clustering the Documents

In figure 3, all the legal documents are represented using the bar diagram. This is the illustration made before the clustering of the documents. The representation is made for the years 1991 to 2014.

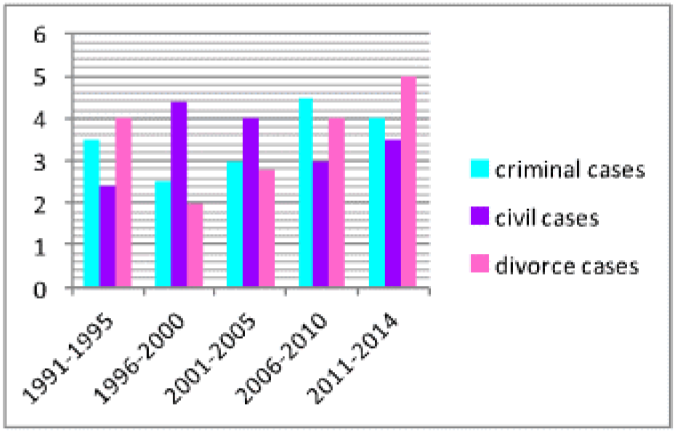

Figure 4: After Clustering the Documents

In figure 4, all the legal documents are represented using the bar diagram. This is the illustration made after the clustering of the documents. The representation is made for the years 1991 to 2014.

Graphical representation of cluster comparison

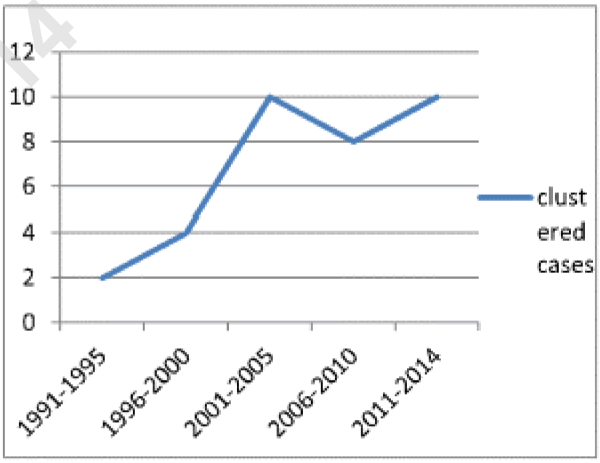

Figure 5: Graphical Comparison of Documents before Clustering

In figure 5, all the legal documents are compared using the graph. This is the illustration made before the clustering of the documents. The representation is made for the years 1991 to 2014.

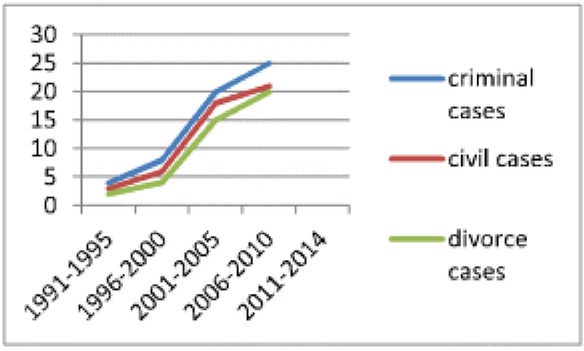

Figure 6: Graphical Comparison Of Documents after Clustering

In figure 6, all the legal documents are compared using the graph. This is the illustration made after the clustering of the documents. The representation is made for the years 1991 to 2014.

VI. CONCLUSION

We presented an approach that applies document clustering methods through segmentation approach of computerized legal documents. Normally the text documents are clustered to divide it into groups where it can characterize the topic that are differs from the topics characterized by the other groups. The legal documents are very large in amount which is a challenging position to cluster it. Here we try to demonstrate the hierarchical algorithm as it divides the law suits accordingly. In future we can scope for the high authentication for the legal documents which makes the judgements easy.

REFERENCES

[1] D. Beeferman, A. Berger, and J. Lafferty."A model of lexical attraction and repulsion".In Proceedings of the ACL, vol 1,pp: 373–380, 1997.

[2] J. Allen, et al. "Topic detection and tracking pilot study final report". In Proc. of the DARPA Broadcast News Transcription and understanding Workshop, 1998.

[3] F. Choi, P. Wiemer-Hastings, and J. Moore."Latent semantic analysis for text segmentation", In Proceedings ofEMNLP, vol 1,pp: 109–117, 2001.

[4] Brants T, Chen F, Tsochantaridis I , "Topic-based document segmentation with probabilistic latentsemantic analysis". In: Proceedings of the 11th ACM international conference on information and knowledgemanagement (CIKM),vol 1, pp:211–218, 2002

[5] Ueda N, Saito K (2002) Single-shot detection of multiple categories of text using parametric mixtureModels. In: Proceedings of the 8th ACM international conference on knowledge discovery and dataMining (KDD), pp:626–631

[6] Blei DM, Ng AY, Jordan MI, Latent Dirichlet allocation. J Mach Learn Res 3:993–1022, 2003.

[7] http://en.wikipedia.org/wiki/Cluster_analysis.

[8] http://en.wikipedia.org/wiki/Hierarchical_clustering

[9] https://sites.google.com/site/dataclusteringalgorithms /hierarchical-clustering-algorithm

[10] http://nlp.stanford.edu/IRbook/html/htmledition/ hierarchical-agglomerative-clustering-1.html.