CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

This chapter reviews the various literature of document clustering. Generally, the efficient document clustering is judged by three criteria, document dimensionality, time consumption and memory. In this research, the Ontology framework is applied to document corpus to reduce dimensionality, computational time and memory utilization by different categorization methods.

The proposed document clustering techniques help to overcome the challenges available in existing document clustering. Additionally, this research gives a brief introduction to all the techniques used for document clustering. In this chapter, the document clustering is reviewed based on the following criteria, searching and querying information, optimizing storage, classification of documents and clustering of documents. In addition to that, preprocessing is done by using BOW process, StopWord elimination, and Porter Stemmer algorithm.

The feature weighting and selection are done to reduce dimensionality. Finally, the raw input corpus document is preprocessed using document frequency. After that, partitioning based rule mining clustering
algorithms such as Hybrid Scheme for Text Clustering (HSTC) model, and hybrid schemes of combining Ontology with k-means, EM, and its variant sGEM are discussed. Moreover, the frequent itemset based methods such as Apriori, FP-Growth, and FP-Bonsai Tree algorithm is used to maintain the document clustering purity with respect to the Hybrid Ontology pattern. These processes are used to improve the text document clustering efficiency of input document corpus such as Reuters-21578, 20NG, and TDT-2 datasets.

2.2 BAG OF WORDS PROCESS

The BOW is the process of optimizing the terms or information in the natural language processing, information retrieval (IR) and data corpus. In this era, several authors discuss about bag-of-words in various domains.

Martins C. A. et al. (2003) presented an efficient decomposing technique to decompose the collection of text into words using the bag-of-words. This bag-of-words approach is used to reduce dimensionality of text terms. They implemented a computational tool called PRETEXT to improve the accuracy of text representation by reducing the dimensionality of text. The PRETEXT tool is used to analyze the frequency of words and stems in a collection of text or documents. Finally, the dimensionality reduced text representation is viewed in the form of data file and attributes file. The PRETEXT files (.data and .names) are converted to attributes and data syntax, which is used by many learning systems such as See5 and CN2. At the same time, PRETEXT tool improves the accuracy of classifiers through reduction of the dimensionality of text in an efficient manner.
Lebanon G. (2007), proposed a scheme named Local Word Bag kernel (LWB kernel) model. Various bag-of-words models, namely Naive Bayes, Support Vector Machine and Maximum Entropy. Compared to the conventional methods, the LWB kernel runs to bring out more semantic information in text categorization context. These models are useful to enhance the performance of sentiment classification than the “Bag-of-Words dimensionality reduction” by the way of combining adjectives, and negations to identify the target words. The LWB kernel model produces higher sentiment classification accuracy.

Liu X. and Wan C. (2014), observed the drawbacks of traditional bag-of-words model. To overcome the drawback of traditional bag-of-words model, a novel approach keyword Query with Structure (QWS) is proposed to analyze and interpret XML keyword query. The QWS model decomposes a keyword query into several fragments (query unit). This QWS approach is used to get the semantics of a keyword query. In addition to this, a novel scoring method called semantics-oriented scoring method is proposed to find the semantic relevance of results which is based on the proposed QWS model. This semantics-aware scoring method is used to match the patterns and structural properties of results. QWS and semantics-aware scoring model shows effectiveness when compared to the other methods.

Santo M.D. and Liguori C. (2012) demonstrated the notion of mixed Graph of Terms based on a hierarchical representation, to retrieve a several number of relevant documents from the interconnected pair of words or a list of words. This technique is used to identify common patterns and make predictions from the vectors of features of weighted words. The concept of “mixed Graph of terms in a query expansion method” is very similar to that of
“Bag-of-Words”, but the proposed representations obtain more accuracy in revelation of common patterns, predictions and average precision of about 50%.

Boulis C. and Ostendorf M. (2005), proposed the filter approach for selecting the features. This approach is used to determine the redundancy of a bigram. Bigrams offer improvements over bag-of-words representation, across a variety of corpora and learning methods. This proposed system evaluates the redundancy of a bigram based only on its unigrams. From that, the entire feature combinations are examined for redundancy and make the text classification task more efficient. Therefore, the proposed filter approach to feature selection leads to get better redundancy performance and identifies non-redundant bigrams with optimal compensation. The Proposed system is applied to a different corpus such as telephone conversations collection, 20NG corpus and collection of postings to electronic forums. The proposed system gains better results in all three corpuses.

Wen P. et al. (2007) proposed the Local Word Bag (LWB) model for information retrieval and text categorization, with conventional BOW model. This method is used to remove local co-occurrence pattern of words at sentence or paragraph level. In addition to this, the kernel is proposed based on the optimal partial match approximation. This kernel is useful to make hierarchical document clustering in all local bags from the text corpus. The hierarchy produced, offers better information retrieval and text categorization.
2.3 STOPWORD ELIMINATION

The StopWord elimination process in text mining is defined as the word that a search engine has been programmed to ignore, both while searching and retrieving a search query. Whenever the predefined “StopWord index” list occurs in searching documents, the search engine automatically eliminates those StopWord. It is used to reduce the irrelevant searching process, save space and search time. From that definition, several authors have discussed about the StopWord elimination process in various domains.

Fox C. (1989) discussed the removal of StopWord. StopWord should be reduced in the preprocessing step of text mining. They eliminated the rationale StopWord such as “on”, “at”, “the”, “in” and so on. These words are not essential for the mining process so they are removed from the input data.

Liu H. and Yu L. (2005) processed StopWord elimination and stemming process in document preprocessing. This preprocessing produced the initial phase for the system and eliminates the redundant word from the documents.

Williams H. E. et al. (2004) used “word-n-grams” based stoplist to remove the StopWord list. The word-n-grams are used in phrase search from the sets of words. It removes the appeared StopWord even the linguistic considerations of attentive sequence of words.
Rosell M. (2009) discussed the stop-word filtering process. StopWord filtering is used to reduce the number of non-zero entries. This system is employed to improve the cluster purity in an effective manner. The StopWord filtering helps for information retrieval in the clustering.

Lee M.D. et al. (2005) created document clustering with preprocessing of input data. The StopWord elimination is applied to the input data for eliminating the redundant list of words from the documents. In Weka, the proposed evaluation model of text document eliminates 1378 words from the input document. After preprocessing an average of 38.2 words for each document is removed with respect to the StopWord list.

Rijsbergen C. (1975) has used the StopWord to eliminate the predefined list of words from the original corpus, which is not used in the text mining tasks. This proposed StopWord elimination process is used to remove basic functional words (“on”, “at”, “the” and “in”) which occur so frequently in documents. From that, proposed system has no discriminating power and can be considered to be noise. So it has been removed from the text corpus for efficient document clustering.

2.4 PORTER STEMMER ALGORITHM

Stemming is the process of reducing the word length by eliminating the suffixes and to discover the root of a word. There are various stemming algorithms such as porter stemmer Porter M.F. (2001), Bahassine S.et al. (2014). These stemming algorithms are used to reduce the word length and to make an efficient document clustering.
Chris D.P. (1994) presented various methods for Stemming algorithms which is useful to eliminate the suffixes in an effective manner. The preprocessing stemming algorithm is used to reduce the problems in word standardization such as etymologically related words, irregularities of adding and removing suffixes. The proposed system eliminates the two stemming errors such as under-stemming errors and over-stemming errors. The under-stemming errors are eliminated by using heavy stemmer and over-stemming errors are eliminated by using light stemmer.

Ali H.M. and Long M.S. (2009) discussed the Stemming algorithms in Arabic documents classification. In the preprocessing, stemming is used to group words based on semantic similarity. The proposed affix removal algorithm is used to remove the suffixes or prefixes from words. The affix removal algorithm efficiently removes the stem words and form stemmed word called morpheme. The morpheme is efficiently used in the preprocessing of Arabic document classification.

Sitaula C. (2013) proposed hybrid algorithm for stemming which is a combination of traditional rule based system and string similarity approach. This proposed hybrid algorithm works based on the distance of stripped words and stem. The proposed algorithm produces an accuracy of 70.10%, meanwhile the total accuracy of “traditional rule based system” is 68.43%. When compared with the proposed hybrid algorithm produce higher accuracy than the traditional rule based system.
Ganesh J.A. (2011) compared various stemming techniques such as Porter Stemmer (Porter M.F. 1980) and Paice/Husk Stemmer (Chris D.P. 1994). The comparison results are produced for each stemming. These stemming techniques work is based on several statistical values. Especially N-Gram Stemmer follows the string-similarity approach. The Hidden Markov Model stemmer is used in finite-state automata. Moreover, the Corpus Based Stemmer, Context Sensitive Stemmer, Iterative Stemmer are also discussed and tabulated.

Bhamidipati L.N. and Pal K.S. (2007) proposed Distribution-based Word segregation for stemming, classification and retrieval of documents. The proposed word segregation model obtains the morphological roots for the given words. Subsequently, it makes a huge reduction in dictionary size and a significant improvement in classification accuracy in 20 NG and WebKB data corpus.

2.5 FEATURE WEIGHTING AND SELECTION

Feature selection is a process of selecting important words and phrases from the dataset. In addition to this, the selection of subset of original features is called as the features extraction. The feature selections are done by using the name, entity and frequency occurrences of word or terms. In feature selection, the original features with high level of dimensionality reduction is used for preprocessing. It removes unwanted features from the document corpus. This powerful feature selection process is used for an efficient document retrieval application to simplify or speed up the computations. The efficiency and accuracy of text classification algorithms is improved by removing redundant and irrelevant terms from the corpus. The following
studies have discussed the various feature selection processes is text classification method.

Gao Z. et al. (2014) discussed the feature selection in terms of information gain. The proposed improved information gain-based feature selection method, uses the term frequency information and balance factor to improve the categorization performance in imbalanced corpus. This feature selection plays a significant role in selecting the necessary field in terms of key feature items selection.

Ghaderi M.A. et al. (2010) proposed the Ordered-Weighted Averaging (OWA). Information Gain (IG) and Chi-Square ($\chi^2$) are used in OWA approach called fusion-based feature selection to improve the classification effectiveness. This approach is applied to the dataset of Reuters-21578. Fusion-based feature selection method has improved the effectiveness of classification on Reuters-21578. The main concern is to select the best minimum subset of original dimensions by eliminating unnecessary features, to reduce the size of the model of feature selection.

Berry M.W. and Castellanos M. (2004) discussed the text mining process for selecting the information from different resources. In order to find the relevant information from the dataset corpus, the proposed text mining processes have been retrieved in an efficient manner. This unknown discovered information has 80% of relevant information retrieval based on feature selection.
Bradley P. et al. (1998) discussed the information extraction from the corpus. There is more number of Web data found in HTML pages. In HTML, “locating the minimal object-rich sub tree” and “set of objects is refined to eliminate irrelevant objects” are used to extract the information from the web pages. This information extraction is more efficient when compared to the conventional extraction techniques.

Huang S. et al. (2006) mentioned the feature selection of large dataset corpus. They proposed the novel feature co-selection for finding web document, which is called multi-type feature co-selection for clustering (MFCC). Feature co-selection is implemented iteratively and can be well integrated into an iterative clustering algorithm. The proposed system enhances the feature selection and produces better clusters in a web document.

Park S. et al. (2010) proposed the Non-Negative Matrix factorization (NMF) for document clustering. This proposed system works based on the basic of weighted semantic features and cluster similarity. NMF is used to group documents with the major topics of a document and also improves the quality of document clustering. The non-negative matrix factorization schema gave better performance than other document clustering methods.

Ribeiro N.M. et al. (2009) proposed local feature selection approach for text clustering. The proposed system compared the local approach with a global feature selection approach. The text cluster is done by using entropy based ranking higher the entropy, higher the cluster efficiency. The proposed result shows better precision in the local feature selection.
Gupta M. and Rajavat A. (2014) discussed the integrated clustering and feature selection scheme for text document clustering. The proposed clustering and feature selection is used to group text documents by using similarity. The feature selection is a method that eliminates the redundant and irrelevant items from the text document contents by statistical methods. In addition to this, the integrated feature selection method is used to improve the feature selection with semantic relation by using an Ontology representation. The integrated feature selection work is based on the terms and concept relationship. The proposed result shows that the feature selection process produced more accurate results.

Jain A.K. et al (1999) recognized new feature selection method for text clustering based on EM and cluster validity. The proposed system is used to monitor feature selection method on intermediate clustering. The Davies-Bouldin’s index is used to evaluate the intermediate feature subsets indirectly. The result generated using iterative clustering produce efficient feature selection in text clustering.

Lior B.C. and Maimon R.O. (2009) introduced dimensionality constitutes in data mining algorithms. This dimensionality constitutes reduction of the dimensionality of data in feature selection process. The proposed system consists of three approaches for feature selection, such as wrapper, filters and embedded. The filter approach operates independently on data mining method. This method subsequently filters the undesirable features out of data before learning begins. The wrapper approach is a statistical re-sampling technique. In an embedded approach the features are specifically selected for a certain inducer; these proposed approaches efficiently selects the features in the process of learning.
2.6 DIMENSIONALITY REDUCTION

The dimensionality reduction is the process of reducing the length, cost, and size of the input data. To reduce the size of text documents, the input document must sort and reduce dimension with respect to the suffix. The dimensionality reductions are done by using Vocabulary, Term, and Entropy. If the above dimensionalities are reduced, the new feature space will have smaller dimensionalities. The new feature space typically contains the linear combination of features; it is used to eliminate noise and irrelevant terms from the input data. Generally, the deduction is based on DFT. This results in an efficient dimensionality reduction and improves the cluster efficiency. Various authors have discussed about dimensionality reduction in various domains such as Feature selection, Feature extraction, Threshold methods, Information theoretic methods, Feature clustering, Topic separation, Outlier detection, Cluster label quality and Computational efficiency. The above dimensionality reduction techniques are used in the preprocessing method to remove noisy features. The two techniques used to improve categorization, effectiveness and to reduce computational complexity, the two techniques used are redundancy analysis and conventional feature selection algorithm. These methods are applied to the well-known Reuters-21578 and 20 NG corpuses. The redundancy analysis and conventional feature selection can eliminate the redundant features dramatically.

Zhen Z. et al. (2011) discussed the dimensionality of the text documents to produce an effective dimensionality reduction. The traditional irrelevant and noise terms based classification methods are not sufficient to meet the complexity. It leads to higher cost of computation and low accuracy of text categorization. To fulfill these constraints, the proposed system used effective dimensionality reduction methods. This process achieves an enhanced
performance, complexity reduction and storage space reduction for text categorization.

Yang Y. and Pedersen O.J. (2000) proposed two dimensionality reduction fields such as feature extraction and feature weighting. In classification process, the dimensionality is reduced by feature extraction. The feature extraction consists of three approaches such as Filter approach, Maximization and weighted averaging. The Wrapper approach is used to improve the classification efficiency, and hybrid approaches were used to reduce the computational time. Meanwhile, the feature extraction consists of three approaches such as Latent Semantic Indexing for removing noise, redundancy, and word ambiguity. Independent Component Analysis transforms data into lower dimensions. Linear Discriminative Analysis extracts best features for discrimination. This approach of feature weighting is used to reduce dimensionality in the classification process. These reductions are used to reduce the time, computational resources, storage, and memory of text classification significantly.

Parimala R. and Nallaswamy R. (2012) proposed efficient dimensionality reduction by DFT. In this proposed system, the frequent and rare terms are removed from the dataset. Depending upon the document frequency reduction, the proposed system is to produce higher computation and efficient performance for classifying the text documents.

2.7 DOCUMENT FREQUENCY

Document frequency based document clustering is used to improve the feature reduction of the dataset. It is used to reduce the unnecessary features
from the document corpus. Document frequency is applied to text corpus, to create linear meaningful cluster. Many methods are available to calculate the categorical document frequency for clustering the documents. Information gain and chi-square statistic criteria are the main aspects to manage document frequency. These criteria are helpful to improve the classification performance.

Zhen Z. et al. (2011) proposed the categorical document frequency for better categorization performance, accuracy and efficiency of text categorization. The proposed filter based feature selection work is based on document frequency of the document terms. The proposed system is compared with the information gain and chi-square statistic of large-scale text data. The categorical document frequency method produced lower computational cost than the classical information gain and chi-square statistic methods.

2.8 CLUSTERING ALGORITHMS-PARTITIONING ALGORITHMS

2.8.1 K-MEANS, EM, and its variants sGEM

In data mining, Cluster analysis is one of the important techniques to group the term, document, and images. Generally, clustering comprises two algorithms such as Partition algorithms and Hierarchical algorithms. Partition clustering algorithm is the process of converting huge number of dataset items into separate set of functions, meanwhile, hierarchical clustering is the process of converting the dataset into smaller hierarchical subsets. Hierarchical clustering has two processing schema such as agglomerative and divisive clustering. Both agglomerative and divisive clustering forms tree with cluster node and child cluster. Especially two or more function are combined to form the agglomerative clustering.
In distinctive character, document clustering has many processes. This algorithm is used to cluster the document with respect to Partition algorithms. Typical partition algorithms such as k-means, EM, and its variant sGEM algorithm are used to cluster the documents in an efficient manner. Several researchers have discussed about k-means, EM, and sGEM in various domains.

k-means is the most important clustering algorithm which is used to minimize the average distance of objects from their cluster centers. Xu G.X. et al. (2014) discussed text clustering with k-means and DBSCAN. In Tibetan text clustering, the proposed DBSCAN text clustering produced better results than k-means clustering.

Mahdavi M. and Hassani H.A. (2009) proposed a novel Harmony k-means algorithm for document clustering. The proposed system consists of the Harmony Search (HS) which is used to optimize document clustering by using finite Markov chain theory. The proposed Harmony k-means Algorithm produced improved quality of clusters.

Zhao Y. and Karypis G. (2001) compared the differences between partition and agglomerative clustering algorithms. The comparison results showed that the existing clustering algorithms did not generate an agglomerative tree efficiently. The comparison results suggested that the agglomerative trees offer better structure for finding the relations among documents at different levels of granularity.
Jaganathan P. and Jaiganesh S. (2013) discussed about hybrid document clustering in large datasets to improve efficiency, execution time on document clustering. The hybrid document clustering combines various processes such as Particle Swarm Optimization (PSO), k-means algorithm, PSOK (PSO and k-means), KPSO (k-means and PSO) and KPSOK (k-means and PSO k-means) algorithms. The proposed hybrid processes are applied to various text document collections. The hybrid model produced quick and better clustering in large dataset.

Sahu L. and Mohan R.B. (2014) proposed the modified Cosine Distance Measure for document clustering. The proposed modified cosine distance measure consists of k-means algorithm, Cosine distance measure, and modified Cosine distance measure. The modified cosine distance measure achieved better feature weighting, performance metric, cluster size, execution time and better cluster quality when compared to other methods.

Lv H. et al. (2013) proposed the variant of k-means, approximate k-means to cluster the sample training dataset. The approximate k-means contains a visual vocabulary of each dataset. The proposed approximate k-means from the k-decision tree based on cluster centers. In each iteration, the proposed system achieved a cluster efficiency of 31.9% compared to the AKM’s 29.8%.

Sathya M. et al. (2011) discussed efficient Link Based Search Engine for Information Retrieval by using k-means clustering. The two parts of the Information retrieval process are co-occurrence terms and k-means clustering of documents. Compared to the conventional techniques, the proposed Link Based Search Engine model clusters the documents efficiently.
Singh V.K. et al. (2011) used k-means, heuristic k-means and fuzzy C-means algorithms for clustering text documents. In the above techniques, documents are clustered with respect to TF/IDF. The proposed result of k-means, heuristic k-means, and fuzzy C-means with respect to TF/IDF representation and stemming produces the best clustering. Particularly fuzzy based clustering produces better results than both k-means and heuristic k-means in all datasets.

Dzogang C. and Fabon E. (2012) discussed the variant of k-means, spherical k-means. The spherical k-means produced similarity measure towards the most relevant dimensions. The proposed k-means achieves better clustering in large number of dimensions.

2.8.2 Expectation Maximization (EM)

Expectation Maximization is the clustering algorithm used to optimize the document cluster. The EM is defined as the generalization of k-means algorithm in document clustering. An Expectation maximization algorithm has two models such as finding model and retrieving model. These models are used to retrieve original data from the dataset corpus. In addition to this, the expectation step deals with the reassignment meanwhile maximization deals with re-computation. These steps are used in an iterative manner for successful document clustering.

Zhang K.K. et al. (2015) examined the EM algorithm to retrieve the missing data. The maximum-likelihood is used to find the categorical distribution of the data. The Expectation Maximization produced the analytical
solution computationally in a simple manner. The results show a significant improvement in clustering.

Hu C.H. et al. (2011) discussed the Recursive Expectation Maximization in recursive algorithms. The proposed Recursive Expectation Maximization is used to update the ER-based prediction model with respect to the probability value. The results are used to fine tune the ER-based prediction model and retrieve new information which is available online.

2.8.3. Spherical Gaussian EM (sGEM)

The Expectation Maximization direction is not an item of information group; the corresponding hyperplane tends to produce individual clusters with wrongly partitioned contents. To resolve this type of clustering, and to improve the quality, sGEM algorithm, re-allocates the new cluster membership to make the cluster process more efficient. This method is also used in finding accuracy, complexity and time space.

Hamdan H. and Wu J. (2013) discussed the bin-EM-CEM algorithm of spherical parsimonious Gaussian mixture models. The proposed model is applied to binned data, the EM-CEM produced efficient structure to reduce the computation time, improve minimum accuracy and speed. In addition to this, the spherical parsimonious Gaussian mixture models maintain the balance between accuracy and CPU time.
2.9 FREQUENT ITEMSET BASED METHODS

The frequent itemset based mining methods were introduced by Borgelt C. (2005). In this work, the frequent itemset based mining methods are divided into three methods such as Apriori algorithm; Divide-and-Conquer strategy based FP-Growth and FP-Bonsai Algorithm. These algorithms are upgraded as hybrid approach which includes Ontology framework. In this era, various authors have discussed about frequent itemset based methods in various domains.

Apriori algorithm is an algorithm proposed by Aggrawal S. and Kaur R. (2013), it is used to mine frequent itemsets and knowledge discovery by using Apriori Algorithms.

Badal S. and Shruti T. (2010) developed the frequent data itemset mining to reduce complexity involved in data mining. The frequent data itemset in large scale database are mined based on the VS_Apriori. The proposed work demonstrates the clustering process and produce minimal time consumption and efficient clustering in intelligent data mining.

Nivet C. et al. (2010) discussed the distributed Apriori Association Rule and Classical Apriori Mining algorithms for grid based knowledge discovery. The proposed system is used to speed up data mining execution, minimize resource utilization, and to obtain efficient knowledge from the dataset corpus effectively.
Jamsheela O. and Ragu G. (2015) compared various scheme of frequent itemset mining algorithms to improve the performance of frequent pattern mining algorithms. The comparison result shows FP-Tree based approach achieves better performance than Apriori.

In association rule mining, Lin K.C. et al. (2011) proposed Improved Frequent Pattern (IFP) growth method with FP-Tree. The proposed system is used to utilize the least amount of memory, better performance, and lower searching complexity. The results show that IFP-Growth method has a better performance when compared to the FP-Tree. This method is a key tool for hybrid FP-Tree mining method, that reduces the reconstruction complexity of the frequent pattern based FP-Tree.

Yu X. et al. (2015) proposed the noise-tolerant frequent itemsets to maintain effective pruning strategies from the dataset. The proposed system is used to reduce noisy data and to mine the frequent itemsets. Proposed noise-tolerant frequent itemsets have anti-monotonicity property which creates lower efficiency, support count, and more cluster computation process. To resolve this issue, the noise-tolerant frequent itemsets has an effective pruning strategy. The results produced better efficiency, higher support count, and less cluster computation process.

Liu S.H. et al. (2014) proposed the Improved Reformative Apriori Algorithm (IOMRA) to merge maximum candidate itemsets in large dataset. To overcome the drawbacks of memory utilization, time consumption and processor speed in the existing Apriori algorithm, the IOMRA is proposed. It is used to speed up the time consumption, memory exhaustion, and execution efficiency. The final proposed result of IOMRA has reduced the amount of
memory occupied and execution time. Finally, improves the utilization ratio of the memory.

Kovacs F. and Illes J. (2013) introduced the substantial frequent itemset mining to find frequent itemsets in large itemsets. The Apriori algorithm is used to reduce the search space. Meanwhile, the count distribution algorithm is used for itemset counting synchronization. The proposed substantial frequent itemset mining framework reduces the execution time, response time when compared to the Apriori algorithm.

Vo B. et al. (2013) proposed the Hybrid Approach for Mining frequent itemsets by using an improved version of PrePost. The proposed system is used to enhance the process of creating N-lists associated with 1-itemsets. This itemsets result improves the run time execution of mining the text documents.

Adrian F. et al. (2014) proposed secure frequent itemset mining to improve the privacy in preserving frequent itemset mining. Secure frequent itemset mining creates the Secure Multiparty Computation model. This model consists of Apriori based distributed, privacy preserving frequent itemset mining algorithms. The proposed model results have a better security scheme and mines in an efficient manner.

2.10 FP-GROWTH ALGORITHM

Cobos C. et al. (2010) proposed the new description-centric algorithm which consists of k-means algorithm, Frequent Term Sets, and
Bayesian Information Criterion. All three models are used to form automatic clusters. The results produced by “description-centric algorithm” have high dimensionality for each vocabulary.

### 2.11 FP-BONSAI TREE

FP-Bonsai is the process of identifying frequent itemsets from large amount of dataset. FP-Bonsai is used to reduce unnecessary itemset from dataset corpus and links corresponding item from the dataset frequent itemset tree.

Pears R. and Kutty S. (2007) presented a novel approach that combines the preprocessing and user-defined constraints to prune the itemset. This proposed system is used to build an optimized FP-Tree. FP-Bonsai is applied to the dataset corpus to reduce the number of itemset. This system forms the efficient FP-Bonsai tree which contains the reduced constraints and unnecessary itemset.

Bonchi F. et al. (2003) presented the ExAnte and FP-Bonsai. Both systems are combined with monotone and anti-monotone constraints into base association rule mining algorithms such as Apriori and FP-Growth. FP-Bonsai approach prunes the FP-Growth algorithm. However, it performs less efficiently in terms of pruning at higher level selectivity. The proposed result of Ex Ante and FP-Bonsai algorithm without affecting preprocessing achieves anti-monotone constraints with the help of association rule mining.
Borgelt C. (2005) projected FP-Trees; it is used to optimize dataset corpus, removes unnecessary items and infrequent past observation items from the dataset by using FP-Bonsai. The proposed result shows that the FP-trees produce efficient result in any dataset corpus.

Korczak J. and Skrzypczak P. (2011) discussed the FP-Bonsai algorithm and the pruning methodology for data reduction. The pruned FP-tree called FP-Bonsai improves the efficiency of the algorithm. The proposed results reduce the time and memory complexity of FP-Growth algorithm.

Fatima P. (2012) proposed IMine index with FP-Bonsai Tree. The I-Tree works with the huge amount of information for storing, modifying, and obtaining information from a database. The proposed system is used to improve the search operation, fast mining, and efficient indexing. From the results, it is clearly understandable that, IMine index with the FP-Bonsai tree reduces long delays, memory and reduce empty linked lists.

2.12 HYBRID ONTOLOGY

Vadasa T.S. and Jhab J. (2014) discussed the various text document clustering and Ontology based techniques for text clustering. The comparison mainly focuses on semantics terms based constraints with Ontology and produced clusters in an effective manner. In text document clustering, Ontology plays an important role to reduce the issues related to the selection of features, Dimensionality of feature space process, Clustering process, and Clustering algorithm.
Wei T. et al. (2015) discussed about the semantic relationships among words. A semantic approach for text clustering is to improve the quality of text clustering. Ontology hierarchical structure is used to extract disambiguated core features and improves clustering performance significantly.

Jing L. et al. (2006) used the basis of Ontologies-based distance measure method to improve the performance of text clustering. The proposed Ontology-based Distance Measure depends upon mutual information matrix of textual data. This textual data is formed by Ontologies using the vector space model. The proposed results shows Ontologies-based distance measure makes text clustering approach perform better.

Ragunath R. and Sivaranjani N. (2015) proposed the hierarchical representation of Ontology model for concept extraction. Ontology-based summarizations compute a set of features for each sentence. The proposed ‘Ontology Based Text Document Summarization System’ use the concept terms in developing the text clustering with better performance.


Paralic J. and Kositial I. (2003) discussed about Webocrat-like approach based on Ontology. The proposed system achieves better retrieval
efficiencies in terms of recall and precision. Additionally, TF-IDF and LSI approaches are used to improve Ontology based clustering.

Zhang L. and Wang Z. (2010) introduced OFW-Clustering for Feature weight calculation. The proposed OFW-Clustering using Ontology semantic node to calculate feature weight to produce accurate clusters. The Euclidean Distance is used in the proposed ‘Ontology-based Clustering Algorithm with feature weights’ to measure the similarity measure. The result produced by OFW-Clustering is more efficient than the other conventional methods.

Bloehdorn A. et al. (2004) discussed the text document knowledge; this frequently appears based on the unsupervised text categorization. Ontology-based text mining works to learn the target Ontology from text documents with improving the effectiveness of text categorization.

Tar H. and Nyunt T. (2011) discussed Ontology-based concept weighting with k-means. The large amount of information is clustered with respect to text documents with the help of Ontology-based concept weighting. The proposed system achieves accurate document clustering by k-means with Ontology. The semantic problem resolves Google search engine in an effective manner.

Luo L.F. et al. (2002) resolved the concept-based model problem using domain-dependent ontologies by self-organizing tree algorithm. The concept-based model improved the effectiveness of dataset such as Reuters-21578 by hierarchical agglomerative clustering.
2.13 PRUNING

Pruning is the process to filter out necessary terms, and remove irrelevant terms from the documents. It is used to identify the document set and select global best from the document terms. This selection classifies the words with low frequency and high frequency. Words with low frequency are unnecessary and are removed by pruning.

Julian S. and Kazakov D. (2004) proposed pre-specified threshold pruning to remove unnecessary notices in the document clustering, meaningless data, which helps in reducing the number of dimensions of documents.

Alexander S. et al. (2000) proposed a technique of web page clustering by pruning all edges except the nearest-neighbors of each vertex. The insignificant or too generic word is also removed based on the pruning conditions (occurs less than 0.01 and more than 0.10 times respectively). The result of the proposed pruning method produced higher pruning in Web-page clustering.

Malik H.H. and Kender J.R. (2006) proposed the Hierarchical Document Clustering and Itemset Pruning. This proposed system deals without inspecting clusters, the Hierarchical Document Clustering selects nodes such as multiple parent, and child nodes automatically. It also works with respect to the Inner term set removal, document duplication and pruning of itemsets. The pruning works with the quality of clustering by decreasing all itemset and increasing the minimum support threshold.
Salton G. and McGill M.J. (1983) discussed the information retrieval from the word dictionary. In this system, StopWord elimination and Pruning are used to remove excess appearing words in the documents. The modern retrieval of information is used to remove unnecessary words from the word dictionary efficiently.

Huang A. et al. (2009) presented three processes to explore and identify the document. Depending upon the central thread, the documents are pruned based on the list of concepts to filter out input document that does not relate to the original document. The phrases are matched against a huge vocabulary of Wikipedia snapshot and contain five million distinct documents. The proposed model produces the best exploration and identification of documents.

Baghel G. and Dhir F.R. (2010) discussed the accuracy of clustering by Pruning the tree. The proposed system works with the help of creating topic hierarchy by joining similar clusters. The proposed “Concepts Based Document Clustering Algorithm”, investigates the influence of document frequency of a term for pruning, this helps to produce an effective clustering.

2.14 SUMMARY

The Document Cluster process is significant with two domains based techniques such as rule mining based text clustering and frequent pattern based text mining. Ontology framework has been applied to each technique to reduce dimensionality, memory utilization, and time consumption. This literature review focuses on various clustering techniques with different author perspectives. Each author represents different methodology for document
clustering. This chapter focuses on Bag of Words process, StopWord Elimination, Porter Stemmer, Feature Weighting and Selection, Dimensionality Reduction, Document Frequency, Clustering Algorithms, Partitioning Algorithms, Hybrid Schemes Combining Ontology with K-Means, EM and sGEM, frequent itemset-based methods Apriori Algorithm, FP-Growth Algorithm, FP-Bonsai Tree, and Datasets used, Hybrid ontology, Pruning, Purity, Similarity Measures, Similarity Measures for Text Document Clustering and Term Weighting. Various research works and real time corpus are reviewed with implementation results. The Hierarchical Document Clustering selects nodes such as multiple parent, and child nodes automatically. It also works with respect to the Inner term set removal, document duplication and pruning of itemsets. The result of this review provides a thorough understanding of the existing technique and the methods to enhance it. This result summary gave way to the new techniques (proposed from the combinational process of existing works). To the best of the knowledge, a comparison between traditional clustering, ontology-based clustering, and hybrid systems have not yet been proved (Ramasubramanian C. and Ramya R. 2013). The present research is designed to do such a performance comparison between two areas such as partition based clustering and frequent based itemset clustering with ontology framework.