CHAPTER 2

LITERATURE SURVEY

Document Clustering

Document clustering is one of the most critical methods to systematize the documents in an unsupervised manner. In recent years, a lot of attention has received. Document clustering is one of the significant tasks in machine learning and artificial intelligence [E. Keogh et al., 1998].

Various clustering methods are suggested like K-means [Lesh N et al., 1999], naive Bayes or Gaussian mixture model [Lang K et al., 1998]. In the different perspectives, the clustering methods are categorized into agglomerative or divisive, hard or fuzzy and deterministic or stochastic. The data clustering is directly performed in the data space, but there is high dimensionality in the document space. On the other hand, the document space is always of very high dimensionality, ranging from several hundreds to thousands. This is because of the deliberation, it is essential to project the documents into a lower-dimensional subspace. Due to this, the semantic structure of the document becomes clear.

2.1 Traditional methods for Document Clustering

According to the numerous distance measures, different methods are suggested to handle the document clustering.

[L.D. Baker and McCallum A.K. 1998] suggested distributional clustering for text classification. The distributional clustering is one of the techniques for the document classification. In this method, based on the class labels associated with each word clusters words into groups. Word clustering methods generate new, reduced size, event spaces by combining similar words into groups. This can provide useful semantically related group of words. The semantic groups depend on the class labels assigned to the documents. This replicates the fact that some words that are synonymous in one context are not in another. The clusters are based on a supervised machine learning algorithm and task focused. The size of the classification model is greatly reduced because separate sets of parameters for many words are replaced with a single
set of parameters for a word cluster. There are some key benefits by using the word clustering. It achieves higher classification accuracy with smaller classification models and useful semantic word clustering. But the disadvantage of this method is not symmetric.

A typically and extensively used distance measure is the Euclidean distance. The K-means method [Lang K 1998] is one of the clustering methods which utilize the euclidean distance as a similarity measure, which diminishes the sum of the squared euclidean distance between the data points and their consequent cluster centers. K-means clustering method is a famous method for clustering the documents. In this clustering method, the clusters are characterized by centers of mass of their members. In this method, the cluster membership is assigned for each data vector to the adjacent cluster center and the center of each cluster is computed as the centroid of its member data vectors is correspondent to find a sum-of-squares cost function using coordinate descend.

[Gangavane H. N. et al., 2015] presented an approach to automatically group the retrieved documents into list of meaningful documents. There are various techniques were used for document clustering. In this method a novel approach is used for document clustering called K-means clustering, which cluster the documents based on the similarity between the terms involved in the document. Through the part of speech tagging in NLP (Natural Language Processing) extract the information in the document and by using chunking group the extracted information. This document clustering used the descriptors and descriptor extraction to enhance the performance of document clustering.

[Kadhim A.I. et al., 2014] proposed an efficient document clustering method which implements singular value decomposition and TF-IDF dimensionality reduction techniques and utilized k means algorithm for document clustering. TF-IDF represents term frequency which means the number of times a term appeared in the document and TF-IDF represents inverse document frequency which defines the number of time the documents appeared. In this document clustering is carried over by the following steps which are text pre processing, term weighting, dimensionality reduction and clustering the documents. In the text pre-processing, the features of the documents such as attributes, words, terms and tokenisation are divided from the input text to improve the quality of features and this process reduced the processing time of document clustering. In the term weighting process TD-IDF is used for binary representation of terms in the
document. Then in the dimension reduction process the redundant terms in the term weighting are reduced and cluster the documents based on k means algorithm. This improved the performance of document clustering.

[Liu X et al., 2002] presented about the utilization of a richer feature set to symbolize each document and use the Gaussian Mixture Model (GMM) together with the Expectation-Maximization (EM) algorithm to conduct an initial document clustering. By using this initial result, a set of discriminative features is discovered for each cluster and refine the initially attained document clusters by voting on the cluster label of each document using this discriminative feature set. The process of cluster label voting and the self-refinement of discriminative feature classification are applied in an iterative manner until the convergence of document clusters. In addition to that, the model selection capability is accomplished by introducing randomness in the cluster initialization stage and then identifying a value C for the number of clusters N by which processing the document clustering process for a fixed number of times yields sufficiently similar results. By using this method, a high accuracy of document clustering is achieved and ability to estimate the number of clusters in the document corpus. But the major disadvantage is the merge decisions are based on insufficient local information but wrong decisions are not corrected later.

[Xu S et al., 2015] analysed different techniques for document clustering. In this paper three different graph partitioning algorithms, four different clustering algorithms and proposed spherical K-means algorithm by combining different techniques. The proposed spherical K-means algorithm choose n number of clusters and randomly generates n clusters from the available n number of clusters. Then find out the concept vector or from the n random points create the concept vector continued by the calculation of the cosine similarity values for document clustering. This process goes until convergence criterion is met. Thus the document clustering is done by using spherical K-means algorithm.

[Mall R and Suykens J. A. 2015] proposed kernel spectral clustering for document clustering which overcomes the problem of principal component analysis. In the primal dual optimization framework there was a major issue called principal component analysis. In this proposed method the clustering model is developed using
dual solution that determines the number of clusters with the help of evaluation metric and it defined the quality of clusters. In this paper various quality evaluation techniques are analyzed depends on the unsupervised version of recall, precision and F-measure and proposed a new cluster model called kernel spectral document clustering. Thus the documents in the same cluster have the maximum similarity and the documents in the different clusters have the maximum dissimilarity.

The document sub-space is highly dimensional. So, in order to reduce the computational complexity it is preferable to identify a low-dimensional representation of the documents so, the spectral clustering methods are analyzed.

2.2 Spectral clustering methods

By using the spectral clustering methods, the computational cost is reduced. In this method, firstly the documents are projected into low-dimensional semantic space. After that, the conventional clustering methods are used to cluster the documents.

[Andrew Y. Ng et al., 2001] analyzed the spectral clustering algorithm. In the machine learning and pattern recognition the process of finding good clusters is an important problem. A simple spectral clustering algorithm is used to find the good clusters. In this method, one uses the top Eigen vectors of a matrix derived from the distance between the points. But there is a problem to find exactly which Eigen vectors to use and how to derive clusters from them. One line of analysis builds the link to spectral graph partitioning where second Eigen vector of a graph’s Laplacian is utilized to define a semi-optimal cut. The Eigen vector is seen as a resolving a relaxation of an NP-hard discrete graph partitioning problem and it can be revealed that cuts based on the second Eigen vector give a guaranteed approximation to the optimal cut. This can be extended to clustering by building a weighted graph in which the nodes correspond to data points and edges are related to the distance between the points. The major disadvantage of this method gives poor clustering results.

[Yuqiang G & Brian K 2004] analyzed kernel K-means and spectral clustering. Clustering is one of the fundamental problems in data mining, to identify clusters kernel K-means and spectral clustering have been used that are non-linearly separable in input space, to provide an explicit theoretical connection between the kernel K-means and spectral clustering. The general function of the weighted kernel K-means clustering is to
derive the spectral clustering objective of normalized cut as a special case. To specify a
green definite similarity matrix, the results lead to a novel weighted kernel K-means
algorithm that monotonically reduces the normalized cut. It has significant implications
are i) Eigenvector-based algorithms, that are computationally prohibitive are not
necessary for minimizing normalized cuts ii) A variety of techniques such as local
search and acceleration schemes may be utilized to improve the quality as well as speed
of kernel K-means. Finally, to provide results on numerous interesting datasets
containing diametrical clustering of huge gene expression matrices and handwriting
recognition dataset, but there is a limitation that it does not suitable for large datasets.

Spectral clustering has been used in many applications like machine learning,
exploratory data analysis, computer vision and speech processing. Spectral clustering
specifies to a set of procedures which based on the Eigen structure of a similarity matrix
to partition points into disjoint clusters, in which the points in the same cluster have
high similarity and points in different clusters having low similarity. Spectral clustering
derive new cost functions based on measures of error between a given partition and a
solution of the spectral relaxation of a minimum normalized cut problem. With respect
to partition minimizing these cost functions leads to new spectral clustering algorithms.
With respect to the similarity matrix minimizes the cost functions that leads to
algorithms for learning the similarity matrix from fully labeled datasets.

[Xinlei C & Deng C 2011] suggested the clustering approaches the spectral
clustering which is one of the effective methods, but there is a limitation in its
applicability to large-scale problems because of its high computational complexity. For
accelerating the spectral clustering, various methods are used, but these methods
typically sacrifice quite lot information of the original data, resulting that the
degradation of performance. In order to solve the large scale clustering problems, the
landmark based spectral clustering is proposed. Particularly, chosen the representative
data points as the landmarks and represents the original data points as the linear
combinations of these landmarks. With the landmark based illustration, the spectral
embedding of the data can then be efficiently evaluated. This method scales linearly
with the size of the problem, but the main disadvantage is difficult to implement.
[Yangqiu Song et al., 2010] analyzed the spectral clustering algorithm is an effective method to find clusters than traditional algorithms. On the other hand, spectral clustering suffers from a scalability issue in both computational time and memory when a dataset size is huge in volume. To achieve clustering for the large datasets propose to parallelize both memory use and computation on distributed computers. The parallelize spectral clustering on distributed computers deals with resource bottlenecks of both memory use and computation time. Compared to the parallelizing K-means, parallelizing spectral clustering is much more challenging to perform. In the parallelization approach, distribute the n data instances onto p distributed machine nodes. In every node, the similarities between the local data and the whole set is computed in a way that uses minimal disk I/O. These two steps, together with parallel Eigen solver and distributed tuning of parameters, speed up clustering time substantially. The parallel spectral clustering outperforms k-means in finding quality clusters and that it scales well with large datasets, but the main problem is high communication cost.

Some of the spectral clustering methods are analyzed and suggested Latent Semantic Analysis is used for automatic indexing and retrieval. The particular Latent Semantic Indexing (LSI) analysis that has tried utilizes singular-value decomposition [Deerwester S.C 1990]. A large matrix of term document association data is considered and generate a semantic space where in terms and documents that are directly connected are located near one another. Singular-value decomposition permits the arrangement of the space to duplicate the main associative patterns in the data and keep away from the smaller, less important influences. Due to this, some of the terms are not appeared in the document but still end up to close the document, if that is consistent with the most important patterns of association in the data. Position in the space then provides as the new type of semantic indexing and retrieval proceeds by utilizing the terms in a query to determine a point in the documents and space in its neighborhood are returned to the user.

On the other hand, due to the high-dimensionality of the document space, a certain representation of documents typically resides on a nonlinear manifold embedded in the similarities between the data points [Andrew Y. Ng. 2001]. Unfortunately, the Euclidean distance measure is a dissimilarity measure which gives the dissimilarities between the documents. Thus, it is not able to efficiently detain the nonlinear manifold structure embedded in the similarities between them [Yuqiang G and Brian K 2004].
[Xu W et al., 2003] suggested a novel document partitioning method based on the non-negative factorization of the term-document matrix of the specified document corpus is used for the document clustering. By using the Non-negative Matrix Factorization (NMF), the latent semantic space is determined, each axis captures the base topic of a particular document cluster and each document is characterized as an additive combination of the base topics. By identifying the base topic with which the document has the largest projection value, the cluster membership of each document can be easily determined. This method differs from the latent semantic indexing method depends on the Singular Vector Decomposition (SVD) and the related spectral clustering methods in that the latent semantic space derived by NMF does not required to be orthogonal and that each document is definite to take only non-negative values in all the latent semantic directions. Compared to space derived by the SVD, an important benefit that each axis in the space derived by the NMF has much more straightforward correspondence with each document cluster. Without using additional clustering operations document clustering results can be directly derived. But the novel document partitioning method is computationally complex.

[Xinlei C and Deng C 2011] suggested novel document clustering algorithm by using Locality Preserving Indexing (LPI) is used for the document clustering. The Locality Preserving Indexing method is different from Latent Semantic Indexing (LSI) method. This is because the Locality Preserving Indexing intends to find the local geometrical structure and it can have more discerning power, but the Latent Semantic Indexing aims to discover the global Euclidean structure. Thus, the documents associated to the same semantics are as same as to each other in the low-dimensional representation space. By identifying the optimal linear approximations to the Eigen functions of the Laplace Beltrami operator on the document manifold, the Latent Preserving Indexing is derived. The second order derivatives of the functions on the manifolds are taken by using the Laplace Beltrami operator. In the low-dimensional space, the conventional clustering methods are used like K-means for clustering the documents into semantically different classes, but the drawback in this method is that it remains unclear as how to compute the number of topics hidden in the document set and it does not overcome the necessary limitation of Euclidean distance. Additionally, the selection of the weighted functions is often a complicated task.
2.3 Correlation similarity based document clustering

Some of the methods are suggested that utilizes the correlation as a similarity measure which detains the intrinsic structure embedded in high-dimensional data, particularly when the input data is sparse. Correlation represents the strength and direction of a linear association between two random variables which discloses the nature of data represented by the classical geometric concept of an angle. This is a scale-invariant relationship measure typically used to compute the similarity between two vectors. In some cases, correlation can efficiently represent the distributional structure of the input data which conventional Euclidean distance cannot clarify.

[Ma Y et al., 2007] suggested Correlation Discriminant Analysis (CDA) as a novel discriminant learning algorithm in correlation measure space used for document clustering. The Correlation Discriminant Analysis integrates the discriminant learning with the correlation measure and searches for the optimum transformation to exploit the distinction between within-class correlation and between-class correlation. This distinction can be empirically considered as a measure for the classification performance. The CDA is a method that seeks out a global linear transformation to exploit the correlation of samples from the same class and reduce the correlation of samples from different classes in the transformed space. This method combines the similarity measure learning with the supervised discriminant learning in a very simple way to determine the nonlinear information instead of only linear information. The main difficulty of this method that it is not suitable for open-set verification applications.

[Fu Yu et al., 2008] suggested Correlation Embedding Analysis (CEA), includes both correlation mapping and discriminant analysis, boosts the discriminating power by mapping the data from a high-dimensional hyper sphere onto another low-dimensional hyper sphere and conserving the neighboring relations with local-sensitive graph model. Correlation Principal Component Analysis (CPCA) simplifies the Principal Component Analysis (PCA) algorithm to the case with data distributed on a high-dimensional hyper sphere. There are some benefits from the two facts. Directly work on normalized data, where are frequently the outputs are from data preprocessing and directly intended with the correlation metric, which is shown to be usually better than Euclidean distance for classification purpose in many real-world applications. But the major difficulty is this method is that it takes long computation time.
[Hardoon D. R. et al., 2004] suggested Canonical Correlation Analysis (CCA) method as a similarity measure for clustering the documents. The CCA method is to identify the projections for paired datasets such that the associations between their low-dimensional representatives in the projected spaces are equally exploited. Particularly, given a paired dataset consisting of matrices it would like to discover directions that exploit the correlation between the projections. CCA is a powerful statistical method, which has been applied in the field of pattern identification and machine learning methods. Rather than finding a projection of one set of data, Canonical Correlation Analysis (CCA) identifies the projections of the two sets of corresponding data X and Y into a single latent space that projects the corresponding points in the two datasets to be as nearby as possible. In the application of document clustering, whereas the document matrix X is obtainable, the cluster label (Y) is not obtainable. So, the canonical correlation analysis method cannot used for the clustering directly.

[Joel W. R. et al., 2006] developed a new term weighting method which is called as Term Frequency – Inverse Corpus Frequency (TF-ICF). This method is used to resolve the difficulty of identifying and organizing information from dynamic document streams. This new term weighting method does not need the term frequency information from other documents into the set and it process the document streams in linear time. The issue in the dynamic document clustering is rectified by utilizing a term weighting method. In this method, the inverse document frequency of TF-IDF is restored by Inverse Corpus Frequency (ICF) values, so that the computational complexity is reduced.

[Rahman, N. A 2015] presented Complete Linkage Clustering algorithm with Cosine Coefficient for document clustering of Malay translated Hadith documents. In this method, documents are clustered based on complete linkage clustering technique. The documents in the same cluster have the maximum similarity and the documents in the other clusters have the maximum dissimilarity. From the analysis of the proposed method it was known that for the small size of cluster, precision score is decreased and recall score is increased.

2.4 Ridge Regression based methods

[Al-Hassan and Yazid M 2010] suggested Ridge regression estimator is an alternative measure to the Ordinary Least Squares estimator (OLS) in the presence of multi-collinearity. In the presence of multi-collinearity, the ordinary least squares
estimator of regression coefficients tends to become unstable. The variance of the estimates of some of the regression coefficients becomes large. To enhance the ordinary least squares estimator many attempts have to be made. There are two approaches. One is centers on discovering estimators which have smaller mean squared error than the ordinary least squares estimators. Even though multi-collinearity is often the situation where the aforementioned estimators are used, this method does not directly address itself to the issue of multi-collinearity. Among these estimators, the ridge estimator points indirectly to the issue of multi-collinearity by constraining the length of the coefficient estimator. In contrast, the second approach handles in a direct manner with the dependency nature of the descriptive variables.

[Pasha G. R. and Muhammad A. A. 2004] presented Multi-collinearity is a one of the main problems in evaluation and calculation it increasing the variance of least squares estimator of the regression coefficients and tending to make least squares estimates that are too large in absolute value. If the two explanatory variables are concerned, there is no declaration that any of the pair wise correlation coefficients will be large. To distinguish the multi-collinearity Variance Inflation Factor (VIF) is set. The Ridge regression is the most popular method in the class of biased estimators. The problem of multi-collinearity is solved by the Ridge Regression by adding a small quantity to the diagonal. In the presence of multi-collinearity the ridge estimator is much more stable than the ordinary least squares estimator.

[Peter E 2013] suggested the kernel ridge regression method is a regularized least square method for classification and regression. Kernel ridge regression is a data-rich nonlinear forecasting tool, which is appropriate for many dissimilar contexts. In this work, the influence of the choice of kernel and the setting of tuning parameters are considered on forecast accuracy. The linear version is equivalent to Fisher’s discriminant for categorization. The non-linear version is related to a Support Vector Machine (SVM), apart from that a dissimilar objective is being optimize, which does not put emphasis on points close to the decision boundary. The main idea of kernel ridge regression is to utilize a flexible set of non-linear functions and to over fitting by penalization, in a way that decreases the computational complexity. This is accomplished
by mapping the set of predictors into a high-dimensional space of non-linear functions of the predictors. By selecting a mapping in a convenient way, Computational traceability is achieved so, the computations in high dimensional space are actually prohibited.

2.5 Sentence based clustering methods

Sentence clustering plays a momentous role in text processing activities. Due to the including of the sentence clustering into extractive multi-document summarization seeks to reduce the issues like content overlap, leading to better coverage [Hatzivassiloglou V et al., 2001], [Zha H 2002], [Radev D. R. 2004] and [Aliguyev R. M 2009]. Most of the general text mining tasks utilize the sentence clustering methods. For example, consider web mining [Kosala R and Blockeel H 2000] the particular objective is to find some new information from a set of documents initially recovered in response to some query. By clustering the sentences of the documents, it would instinctively anticipate at least one of the clusters to be intimately associated to the concepts described by the query terms. However, other clusters may contain information pertaining to the query in some way hitherto unknown to us and in such a case it would have successfully mined new information.

[Li Y et al., 2006] suggested the sentence similarity measure to improve the clustering efficiency. Sentence similarity measures play an important role in text-oriented applications in areas such as text mining, Web page recovery. The Sentence similarity measures are significant techniques in webpage interval for improving the efficiency of the information retrieval, where titles are used to signify documents in the named page discovering task. The main concept in the sentence similarity measure is to evaluate the similarity between very short texts, principally, the sentence length. The algorithm is presented which considers word order information and the semantic information disguised in the sentences. By using information from a structured lexical database and from corpus statistics, the two sentences semantic similarity is to be computed. The utilization of lexical database facilitates this method to model human common sense knowledge and the integration of corpus statistics allows this method to be flexible to various domains. This method can be used in a various types of applications that include text knowledge representation and discovery. But the disadvantage of this method is the similarity method is only suitable for application domains not for other domains.
[Kotlerman L et al., 2012] presented a new sentence clustering method according to the projecting sentences over term clusters. They main aim of sentence clustering is to group sentences with identical meanings into clusters. The general sentence similarity measures are cosine. This measure is used to define the level of similarity over bag-of-words encoding of the sentences. This method includes the external knowledge to solve the lexical unpredictability and small corpus size and utilize the general sentence clustering methods in two real-life industrial datasets.

[Mehwish A et al., 2010] presented a new method for measuring semantic similarity of two blog-posts is proposed in this work. Sentence similarity measure is used in various areas such as language modeling, word sense disambiguation, document clustering and search filtering. This method is the sentence-based algorithm which extracts the noun-phrases, verb-phrases and common bag-of-words to calculate the similarity. This method is a combination of corpus based and dictionary based semantic mining. This method is further used to find influential bloggers from blogging community. This includes sentence based similarity measure as one of the effectual influence measuring factors utilized in an effective algorithm for deducing influential bloggers’ list.

Text summarization is the method of automatic creation of compressed version of a given document conserving its information content. There are two categories of summarization. One is extractive summarization and another one is abstractive summarization. Extractive summarization methods make simpler the difficulty of summarization into the problem of choosing a representative subset of the sentences in the original documents. Abstractive summarization may create new sentences, unseen in the original sources. [Rasim A et al., 2009] developed the sentence based extractive document summarization. The extractive summarization systems are usually based on techniques for sentence removal and the main intent is to cover the set of sentences that are most significant for an overall understanding of a given document. In this work, an unsupervised document summarization method is used which creates the summary by clustering and mining sentences from the original document. For this, a new criterion functions are used to cluster the sentences. Similarity measures play gradually an important role in document clustering. Here, a discrete differential evolution algorithm is developed to optimize the criterion functions.
[Richard K 2012] suggested a novel clustering method to robotically group the similar sentences according to the sentences’ part-of-speech syntax. This algorithm creates and combines together the clusters by utilizing a syntactic similarity metric according to the hierarchical organization of the parts-of-speech.

2.6 Semantic based clustering methods

Semantic similarity or semantic relatedness is a one of the significant measures in the clustering process. The idea of the semantic similarity is to compute the distance between them is according to the likeness of their meaning or semantic content as opposed to similarity which can be computed concerning their syntactical representation (e.g. their string format). These are mathematical tools which are utilized to compute the strength of the semantic association between units of language, concepts or instances, through a numerical explanation attained based on the comparison of information supporting their meaning or describing their nature.

Concretely, semantic similarity can be computed by describing a topological similarity, by utilizing ontology to describe the distance between terms/concepts. For example, a naïve metric is used to compare the concepts ordered in a partially ordered set and describes as nodes of a directed acyclic graph (e.g., a taxonomy), would be the shortest-path linking the two concept nodes. According to the text analyses, semantic association between units of language (e.g., words, sentences) can also be computed by using statistical means like vector space model to relate words and textual contexts from a suitable text corpus.

The notion of semantic similarity is more precise than semantic relatedness, as the latter include concepts as anonymity and metonymy, while similarity does not. On the other hand, much of the literature uses these terms interchangeably, along with terms like semantic distance. Specifically, semantic similarity, semantic distance and semantic relatedness all mean, "How much does term A have to do with term B?" The answer to this question is typically a number between -1 and 1, or between 0 and 1, where 1 indicates tremendously high similarity.

[Ming-Wei C et al., 2008] presented the significance of semantic representation. Text classification is typically seen as the difficulty of assigning a label to data. The semantic information obtainable in the label is disregarded whilst making this decision and the
labels are treated as atomic identifiers. This requires the use of annotated data to train classifiers that map documents to these identifiers. Additionally, humans can carry out text categorization without seeing even one training example. The notion of Data less categorization is a model of classification that does not require annotated training data. This method is based on the use of a source of world knowledge to analyze both labels and documents from a semantic point of view, permitting us to learn classifiers. Such analysis by a semantic interpreter facilitates us to evaluate the concepts discussed by the document to perform classification. The data less classification can classify text without any annotated data. The main restriction is less accuracy in classifiers.

[Anna H et al., 2008] presented document clustering with active learning by using Wikipedia. Text document clustering approach groups documents with same themes together, whereas keeps documents with dissimilar topics in separate. Wikipedia has been applied as a background information base to a variety of text mining problems, but a few of the attempts have been made to exploit it for document clustering. For document clustering, the semantic knowledge is developed in Wikipedia and facilitates the automatic grouping of documents with similar themes. Firstly, the Wikipedia is used to make a concept-based representation of a text document, with each concept associated to a Wikipedia article. After that, the semantic relatedness between Wikipedia concepts is exploited to find the pair-wise instance-level restraints for supervised clustering towards the direction specified by the constraints.

[Jayaraj J and Selvadurai K 2014] presented the semantic features for clustering the documents. Due to including of the semantic features it will enhance the accuracy of the documents retrieval by using the method of clustering and which will also pave the way to systematize and recover the documents more proficiently, from the large obtainable corpuses. However, the clustering based on the semantic features improves the quality, but still scalability is a main problem. In this work, three dynamic document clustering methods are suggested namely Term frequency based Maximum Resemblance Document Clustering (TMARDC), Correlated Concept based fast Incremental Clustering Algorithm (CCFICA) and Correlated Concept based Maximum Resemblance Document Clustering (CCMARDC). In the above three methods, the TMARDC algorithm is used based on the term frequency, in which the CCMARDC and CCFICA are based on Correlated terms concept extraction algorithm.
[Krishnan R et al., 2007] proposed a method of improve the accuracy of the clustering method by utilizing Wikipedia to enhance the illustration of the short texts to be clustered. In this method, the conventional bag of words representation of text items is enlarged with the titles of select Wikipedia articles. For text Clustering, Wikipedia can be used as the additional feature source particularly when very few lines of text are obtainable for each item. To estimate this method, a relative study is used by denoting a gathering of short news articles by utilizing the bag of words method. By using each representation, different clustering algorithms are executed. For most of the clustering algorithms the higher accuracy is attained by using this method, but the limitation in this method is that the time taken for computation is more.

Due to the high-dimensionality of the text data and the complicated semantics of the natural language the document classification presents problematical challenges. [Pu W and Carlotta D 2008] suggested the conservative document demonstration a word-based vector, where each and every dimension is related with a term of the dictionary including all the words that emerge in the corpus. In order to enhance the prediction abilities of classification algorithms, it is essential to implant semantic information and intangible patterns. In the bag of Words method the background knowledge calculated from Wikipedia is entrenched into a semantic kernel, which is then used to develop the representation of documents. This methodology is able to continue multi-word concepts unbroken and it confines the semantic nearness to synonyms and accomplishes word sense disambiguation for polysemous terms. Thus, the same enrichment modus operand could be extended to progress the clustering of documents, when certainly class labels are not obtainable and also too expensive to acquire.

[Anna H et.al 2009] presented clustering which is necessary in data mining, primarily for handling large scale data. When it is applied to the documents, it automatically groups ones with similar themes together whereas separating those with different topics. The creation of brief representation of a document is a main problem for clustering and also in other fields like text documents, such as information retrieval, text categorization and information extraction. This work shows how the Wikipedia and the semantic knowledge should be used for document clustering. For this, firstly, create a concept-based document demonstration by mapping the terms and phrases within the documents to their consequent articles in Wikipedia. The document representation with
these clusters also facilitates consequent clustering to relate documents that do not overlap in the original word space. A similarity measure is also used to compute the semantic relatedness between concept sets for two documents. The drawback in this method is difficult to implement this concept.

[Eibe Frank et al., 2009] suggested a genetic algorithm based clustering method which is called as GA clustering and also with ontology is used in this method. Generally, the ontology measures can be partitioned into two types. Thesaurus-based methods and corpus-based methods. The advantage of hierarchical structure and the broad coverage taxonomy of WordNet is taken as advantage of the thesaurus-based ontology. On the other hand, the corpus-based method is very difficult to handle the practical application. A transformed Latent Semantic Analysis (LSA) method is used as the corpus-based method. Furthermore, considering the influence between the diversity of the population and the discerning pressure, a self-adaptive evolution method is put forward in this article.

[Rekha B and Renu D 2010] presented Frequent Concepts based Document Clustering (FCDC) method for clustering the documents. In this method, a clustering algorithm is used with frequent concepts. But in the conventional text mining methods, frequent items are used. Various clustering methods are used for handling with the documents as bag of words and remove the significant associations between words like synonyms. This FCDC algorithm uses the semantic relationship between words to generate concepts. It develops the WordNet ontology in turn to generate low dimensional feature vector which permits us to develop a proficient clustering algorithm. This method utilizes a hierarchical approach for clustering the text documents having common concepts. FCDC is found more precise, scalable and effectual.

[Kumar M et al., 2015] analysed various document clustering algorithm to organize the documents into a specified type of group. Vector space model is a method where the documents are described as a vector. The similarity values between every two documents are defined by this proposed method. One of the issues involved in document clustering is the huge dimension of data. The dimension is defined as the term present in the document. The high dimension of data represents the high document vector space. By pre-processing the vector space the high dimensionality will be
reduced. This can be done by the various techniques like stemming and filtering. Then the term frequency vectors subsequent to each document are created. But in the proposed method the similarity values are determined by comparing the term frequencies vectors of each pairs in the document. It utilized three similarity matrices to define the similarity values are minimum-match, average-match and maximum-match.

[Stanchev L 2016] proposed an algorithm that finds the meaning of terms in the document based on similarity graph for document clustering without the intervention of human. The degree of semantic similarity between the terms in the documents is determined by using graph that used WordNet details or information. The edges in the graph is defined as the probability value which defines the destination node interest in the concept illustrated by the source node. In this method, the graph is build based on the term in the documents and the documents are represented as nodes in the graph. Thus the semantic distance between the two documents are evaluated by using similarity degree values of the terms in the document.

[Rahmawati D 2015] proposed feature based clustering methods for document clustering. Document clustering is a method used to group the documents based on term frequencies or sentence frequencies or any other consideration of document characteristics. In this proposed method, feature based clustering used K means clustering to split sequential data of features. In this, features are represented as sequence of words. In the collection of unstructured documents the optimal features are extracted by pre-processing the documents. The extracted features are represented by two different techniques are maximal word sequence and frequent word sequence. In this proposed method, maximal frequent sequence is used for feature representation. Additionally a framework was proposed to carry out feature based clustering using maximal frequent sequence as features.

**Chapter Summary**

Chapter two describes the previous work in document clustering. In the previous methods, k-means method, document partitioning method, Distributional Clustering techniques are analyzed. In the spectral clustering methods, Latent Semantic Analysis, Locality preserving indexing, Correlation Discriminant Analysis is suggested. The works related sentence based clustering and semantic relatedness is also described.