CHAPTER 2

LITERATURE REVIEW

This chapter deals with the literature survey about the various techniques and methodologies applied for data mining especially for clustering, specified in Chapter 1, the review has been carried out using the following aspects.

- Clustering
- Peculiar mining
- Software Agents

The present chapter gives an overview of clustering, and peculiar mining to identify the peculiarity in multi databases. It discusses the application of software agents.

2.1 CLUSTERING

Clustering analysis is an important field of research intelligence and data mining. The fundamental concept of clustering analysis is to use characteristic to calculate the degree of similar relationship among objects and to attain automatic classification. This section gives the basis of the present study on clustering, related works on clustering techniques that are related to the present research and an evaluation of clustering algorithms. Cluster analysis is defined as the study of methods for grouping, or clustering, objects depending on the measured or perceived intrinsic characteristics or similarity (Yu and Philip, 2003). Clustering analysis is an important field of research intelligence and data mining. The fundamental concept of clustering analysis is to use characteristic to calculate the degree of similar relationship among objects and to attain automatic classification.

Even though there is an increasing interest in the use of clustering methods in pattern recognition (Anderberg 1973), image processing (Jain and Flynn 1996) and information retrieval (Rasmussen 1992; Salton 1991), clustering has a rich history in other disciplines (Jain and Dubes 1988) such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing. Other terms more or less synonymous
with clustering include unsupervised learning (Jain and Dubes 1988), numerical taxonomy (Sneath and Sokal 1973), vector quantization (Oehler and Gray 1995), and learning by observation (Michalski and Stepp 1983). The field of spatial analysis of point patterns (Ripley 1988) is also related to cluster analysis.

The importance and interdisciplinary nature of clustering is evident through its vast literature. A number of books on clustering have been published (Jain and Dubes 1988; Anderberg 1973; Hartigan 1975; Spath 1980; Duran and Odell 1974; Everitt 1993; Backer 1995) in addition to some useful and influential review papers. A survey of the state of the art in clustering circa 1978 was reported in Dubes and Jain (1980). A comparison of various clustering algorithms for constructing the minimal spanning tree and the short spanning path was given in Lee (1981). Cluster analysis was also surveyed in Jain et al. (1986). A review of image segmentation by clustering was reported in Jain and Flynn (1996). Comparisons of various combinatorial optimization schemes, based on experiments, have been reported in Mishra and Raghavan (1994) and Al-Sultan and Khan (1996).

Cluster analysis is based on partitioning a collection of data points into a number of subgroups, where the objects inside a cluster (a subgroup) show a certain degree of closeness or similarity. Hard clustering assigns each data point (feature vector) to one and only one of the clusters, with a degree of membership equal to one, assuming well defined boundaries between the clusters. This model often does not reflect the description of real data, where boundaries between subgroups might be fuzzy, and where a more nuanced description of the object’s affinity to the specific cluster is required. Thus, numerous problems in the life sciences are better tackled by decision making in a fuzzy environment. Due to the advancement and development in computing and communication over wired and wireless networks, it has resulted in many pervasive distributed computing environments. The Internet, intranets, local area networks, ad hoc wireless networks etc., are some examples of the distributed computing environments. These environments often come with various distributed sources of data and computation. Mining in such situation logically calls for correct utilization of these distributed resources.
Furthermore, in most of the privacy sensitive applications, multi-party data sets gathered at different sites must be processed in a distributed manner without collecting everything to a single central site. On the other hand, the majority off-the-shelf data mining systems are planned to perform as a monolithic centralized application. They usually download the appropriate data to a centralized place and then execute data mining operations. This centralized method does not perform well in most of the emerging distributed, ever-present, probably privacy-sensitive data mining applications.

There are several clustering approaches available in the literature to cluster the document in distributed environment. But most of the existing clustering techniques suffer from a wide range of limitations. The existing clustering approaches face the issues like practical applicability, very less accuracy, scalability, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. Peculiarity rules are a new class of rules which can be discovered by searching relevance among a relatively small number of peculiar data. Peculiarity oriented mining in multiple data sources is different from, and complementary to, existing approaches for discovering new, surprising, and interesting patterns hidden in data. A theoretical framework for peculiarity oriented mining is presented. Within the proposed framework, we give a formal interpretation and comparison of three classes of rules, namely, association rules, exception rules, and peculiarity rules, as well as describe how to mine interesting peculiarity rules in multiple databases.

Clustering involves the task of dividing data points into homogeneous classes or clusters so that items in the same class are as similar as possible and items in different classes are as dissimilar as possible. Clustering can also be thought of as a form of data compression, where a large number of samples are converted into a small number of representative prototypes or clusters. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measures include distance, connectivity, and intensity.
Local clusters are generated by using the documents in the local node where as global clusters are created with documents from all peer nodes. The results of the local clusters are used for creating global clusters. Local optimization based clustering process uses initial centroid estimation in local document collection. Alternatively, global optimization technique uses centroid estimation from all peer nodes. Term weights and semantic weights are taken into consideration in this clustering process. Cluster analysis is an exploratory data analysis technique aiming at getting insight into data. Adopting an informal definition, clustering can be stated as the problem of identifying natural or interesting groups in data. It is called unsupervised learning due to the lack of any information on cluster membership: no particular assignments of data items are available and usually the number of clusters is not known in advance.

This means getting to know your data. If you can categorize, classify, and/or codify your data, you can place it into chunks that are manageable by a human. Rather than dealing with 3.5 million merchants at a credit card company, if we could classify them into 100 or 150 different classifications that were virtually dead on for each merchant, a few employees could manage the relationships rather than needing a sales and service force to deal with each customer individually. Likewise, at a university, if an alumni group treats its donors according to their classifications, part-time students might be the representatives who work with minor donors and full-time professionals might receive incoming calls from the donors whose names appear on buildings on campus.

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data
from comment or warranty cards, the retailer could develop products and promotions to
appeal to specific customer segments. While large-scale information technology has
been evolving separate transaction and analytical systems, data mining provides the link
between the two. Data mining software analyzes relationships and patterns in stored
transaction data based on open-ended user queries. Several types of analytical software
are available: statistical, machine learning, and neural networks. Generally, any of four
types of relationships are sought:

- **Classes**

  Stored data is used to locate data in predetermined groups. For example, a
  restaurant chain could mine customer purchase data to determine when customers visit
  and what they typically order. This information could be used to increase traffic by
  having daily specials.

- **Clusters**

  Data items are grouped according to logical relationships or consumer references.
  For example, data can be mined to identify market segments or consumer affinities.

- **Associations**

  Data can be mined to identify associations. The beer-diaper example is an example
  of associative mining.

- **Sequential patterns:**

  Data is mined to anticipate behavior patterns and trends.

**Different levels of analysis Exist**

- **Artificial neural networks**

  Non-linear predictive models that learn through training and resemble biological
  neural networks in structure.

- **Genetic algorithms**

  Optimization techniques that use process such as genetic combination, mutation,
  and natural selection in a design based on the concepts of natural evolution.
• **Decision trees**

Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

• **Nearest neighbor method**

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k > 1). Sometimes called the k-nearest neighbor technique.

• **Rule induction**

The extraction of useful if-then rules from data based on statistical significance.

• **Data visualization**

The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships. Organizing data into sensible grouping is an essential factor for understanding and learning. For instance, a common scheme of scientific classification puts organisms into a system of ranked taxa: domain, kingdom, phylum, class, etc. Cluster analysis is defined as the study of methods for grouping, or clustering, objects depending on the measured or perceived intrinsic characteristics or similarity (Yu and Philip, 2003). A category label is not used in cluster analysis. Clustering analysis is an important field of research intelligence and data mining. The fundamental concept of clustering analysis is to use characteristic to calculate the degree of similar relationship among objects and to attain automatic classification.
High-volume and high-dimensional data sets have been generated due to the development in storage technology and also due to remarkable growth in applications such as digital imaging, video surveillance etc. It is estimated that the digital universe consumed roughly 281 exabytes in 2007, and it is predicted to be 10 times that size by 2011. (One exabyte is $\sim 10^{18}$ bytes or 1,000,000 terabytes). Most of the information is stored digitally in electronic media, which offers huge potential for the development of automatic data analysis, classification, and retrieval techniques. In addition to the growth in the amount of data, on the other hand, the variety of available data like text, image, and video has also increased drastically. Inexpensive digital and video cameras have made available huge records of images and videos. The RFID tags or transponders have become popular because of their low cost and small size and it is very much useful for the use of millions of sensors that transmit data regularly. Several terabytes of new data have been created every day due to E-mails, blogs and billions of Web pages. These data streams are mostly unstructured which are very difficult to analyze. Thus, due to increase in both the volume and the variety of data, it is necessary that there should be advancements in methodology to automatically understand, process, and summarize the data.

Selim, S. Z et al (1984) discussed that the K-means approach is a widely used technique in cluster analysis. The author addressed about various questions about the approach. The clustering problem is primarily casted as a non convex mathematical program. Then, an exact proof of the finite convergence of the K-means-type approach is given for any metric. It is shown that under certain conditions the algorithm may fail to converge to a local minimum and that it converges under differentiability conditions to a Kuhn-Tucker point. Ultimately, a technique for obtaining a local-minimum solution is presented by the author.

Cutting et. al. (1992) describes that document clustering has not been considered as an effective information retrieval tool. This is mainly because of the two reasons:

- Clustering is too slow for large datasets
- Clustering does not significantly improve retrieval.
These problems occur only when clustering is used to improve traditional search algorithms. But, looking at clustering as information access tool, these objections have to be taken under consideration and eliminated. The authors present a document browsing technique that uses document clustering as its primary operation and also present fast (linear time) clustering techniques which support this interactive browsing paradigm.

Al-Sultana and M. Maroof Khan (1996) examine the problem of clustering m objects in c clusters. The objects are denoted by points in n-dimensional Euclidean space, and the main aim is to categorize these m points into c clusters in such a way that the distance between points within a cluster and its center is reduced. The problem is a complex optimization problem as it possesses many local minima. Various techniques have been proposed to solve this problem such as the k-means algorithm, the simulated annealing algorithm, the tab search algorithm, genetic algorithm etc. The authors examine these four techniques and compare their computational performance for the clustering problem. These techniques are tested on various clustering problems from the literature as well as several random problems.

Bradley et. al. (1998) proposed that practical statistical data clustering techniques need multiple data scans to converge. These scans become expensive for large databases. This research proposes a scalable clustering framework which requires at most one scan of the database, and is applied to the Expectation-Maximization (EM) algorithm. EM is a suitable optimization approach for building proper statistical models of the data. EM is found to be different from traditional distance-based algorithms like k-means. EM also accommodates categorical and continuous data fields. It acknowledges varying degrees of data membership in multiple clusters. The scalable approach used in this paper is depending on finding regions of the data that are compressible and regions that must be maintained in memory. The algorithm functions within the confines of a limited memory buffer. Data resolution is conserved as much as possible depending on the size of the memory buffer and the fit of the current clustering approach to the data. The framework is extended to update multiple clustering
approaches concurrently. Computational tests show that this proposed scalable system outperforms sampling-based and incremental approaches.

Jain, A.K et al (1999) proposed that clustering is the unsupervised classification of patterns into groups or clusters. Because of the wide range of utility of clustering techniques in many disciplines; the problem of clustering has been widely addressed by several researches. The author proposed an outline of pattern clustering techniques from a statistical pattern recognition viewpoint, with the aim of providing useful information and references to basic concepts accessible to most of the users of clustering techniques. The authors present taxonomy of clustering approaches, and explains recent advances and also describe some important applications of clustering algorithms in various fields like data mining, image processing etc.

Likas et al (2003) proposed the global k-means clustering technique that creates initial centers by recursively dividing data space into disjointes subspaces using a k-d tree approach. The cutting hyper plane used in this approach is the plane that is perpendicular to the maximum variance axis derived by Principal Component Analysis (PCA). Partitioning is carried out until each of the leaf nodes contains less than a predefined number of data instances (bucket size) or the predefined number of buckets has been constructed. The initial center for K-means is the centroids of data that are present in the final buckets.

Shehroz S. Khan and Amir Ahmad (2004) describes that the performance of iterative clustering techniques which converges to various local minima is based mainly on initial cluster centers. Usually initial cluster centers are chosen randomly. The author proposed an algorithm to calculate initial cluster centers for K-means clustering. This approach depends on the observations that some of the patterns are very analogous to each other and so they have similar cluster membership irrespective to the choice of initial cluster centers. Moreover, an individual attribute may give some information about initial cluster center. The initial cluster centers calculated using this approach are very close to the desired cluster centers, for iterative clustering algorithms. This process is feasible to clustering techniques for continuous data. The author showed the use of
proposed approach to K-means clustering technique. The experimental results showed significant solutions using the proposed algorithm.

Agrawal et.al. (2005) describes about data mining applications and its various requirements on clustering techniques. The main requirements considered are

- ability to identify clusters embedded in subspaces of high dimensional data,
- scalability,
- end-user comprehensibility of the results,
- non-presumption of any canonical data distribution,
- Insensitivity to the order of input records.

A clustering algorithm called CLIQUE is presented that assures the above requirements. CLIQUE algorithm finds dense clusters in subspaces of maximum dimensionality. It produces cluster descriptions in the form of DNF expressions that are minimized for ease of comprehension. The approach generates identical results irrespective of the order in which input records are presented and does not presume any particular mathematical form for data distribution. From the experimental result, it is observed that CLIQUE algorithm efficiently identifies accurate clusters in large high dimensional datasets.

K-means is the simplest and the most popular clustering technique that is widely used in various fields of science and technology. But the main limitation of the k-means approach is that it generates empty clusters based on initial center vectors. This drawback does not cause significant problem for static execution of the k-means algorithm and the problem can be overcome by executing the k-means algorithm for a number of times. But when k-means algorithm is presented as an essential part of some higher level application, this empty cluster problem causes irregular behavior of the system and may lead to overall performance degradation.

Malay K. Pakhira et al.(, 2009) proposed a modified version of the k-means algorithm that effectively eradicates this empty cluster problem. From the experimental results, it is observed that, the proposed approach shows higher performance when compared with the conventional approach. The heterogeneous data streams with
uncertainty are almost seen in most of the applications. But, the clustering quality of the existing approaches for clustering heterogeneous data streams with uncertainty is not satisfactory.

Guo-Yan Huang et. al. (2010) proposed an approach for clustering heterogeneous data streams with uncertainty called HU-Clustering. A Heterogeneous Uncertainty Clustering Feature (H-UCF) is proposed to illustrate the feature of heterogeneous data streams with uncertainty. Depending on H-UCF, a probability frequency histogram is proposed to track the statistics of categorical attributes; the approach initially creates n clusters by k-prototypes algorithm. A two phase streams clustering selection process is applied to HU-Clustering algorithm to improve the quality of clusters. Firstly, the candidate clustering is chosen through the new similarity measure; secondly, the most similar cluster for each new arriving tuple is chosen through clustering uncertainty in candidate clustering set. The experimental observations reveal that the clustering quality of HU-Clustering is significant than that of UMicro.

Clustering is a vital data mining task, which groups the data according to the similarities among the data. The partitional clustering and hierarchal clustering are combined and a novel clustering algorithm called Hierarchical Particle Swarm Optimization (HPSO) data clustering is proposed by Alam et. al. (2010). The proposed algorithm uses the swarm intelligence of cooperating agents in a decentralized environment. The performance of the HPSO is evaluated on benchmark datasets and is compared with clustering techniques like K-means, PSO clustering, Hierarchical Agglomerative clustering (HAC) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The experimental observations reveal that the effectiveness of Swarm based clustering algorithm and the capability to perform clustering in a hierarchical agglomerative manner.

Shin-Jye Lee et. al. (2010) presents a clustering-based approach to fuzzy system identification. In order to generate an effective initial fuzzy model, this paper tries to present a modular approach to identify fuzzy systems based on a hybrid clustering-based technique. Furthermore, the determination of the proper number of clusters and the suitable location of clusters are one of primary considerations on constructing an
effective initial fuzzy model. Because of the above mentioned reasons, a hybrid clustering approach concerning input, output, generalization and specialization has been introduced in this article. Moreover, the proposed clustering approach, three-part input-output clustering algorithm, incorporates a variety of clustering features simultaneously, including the advantages of input clustering, output clustering, flat clustering, and hierarchical clustering, to effectively perform the identification of clustering problem.

Most of the data has been collected in incremental mode and incremental clustering has attracted the attention of several researchers. But, only few researches have focused on partitioning categorical data in incremental mode. Designing an incremental clustering for categorical data is a vital issue. Li Taoying et al (2010) proposed an incremental clustering for categorical data using clustering ensemble. The author initially reduce redundant attributes if required, and then make use of true values of different attributes to form clustering memberships, and next use clustering ensemble to merge or divide clusters to gain optimal clustering. Ultimately, the proposed approach is applied in Yellow-Small dataset, Diagnosis dataset and Zoo dataset and results reveal the effectiveness of the proposed approach.

Clustering techniques have been widely used for pattern discovery from Web Usage Data (WUD). In e-commerce applications, clustering techniques can be used for the purpose of creating marketing policies, product offerings, personalization and Web site adaptation. A novel partitional based technique for dynamically grouping Web users based on their Web access patterns using Adaptive Resonance Theory Neural Network (ART1 NN) clustering algorithm is proposed by Raju et.al. (2011). The problem formulation and the proposed ART1 NN clustering methodology have been discussed. Experimental results reveal that the proposed ART1 NN clustering technique outperforms K-Means and SOM clustering algorithms in terms of intra-cluster and inter-cluster distances.

There are several techniques for using correlations between terms in document clustering. The method computes similarity between documents depending on the statistical correlations between their terms and then uses these similarities to group documents into clusters. So a different similarity model is examined. In VSM, terms are
independent and hence ignore any semantic relations between them. This means that the proximity between documents is not similar. Moreover, redundancy increases the dimensionality and affects the performance of algorithms.

To overcome this limitation, Wang et. al. (1985) proposed Generalized VSM (GVSM) which represents the documents as vector in a non-orthogonal basis of terms. The similarity between documents is estimated depending on the similarity between their terms. High performance is achieved by the proposed approach.

Hinneburg and Keim, (1999) stated that projection is one of the key factors for ontology based text clustering, as projection increases effectiveness and efficiency of the clustering technique. The performance of the cluster is mainly concentrated rather than the quality of the cluster. There are various projection techniques and each of the techniques is considered important as every individual technique have different output performance.

Schuetze and Silverstein, (1997) had a detailed study about the different projection schemes. The author mainly focused on local and global projection. Local projection deals with ‘how document maps with different subspace’ whereas global projection chooses the appropriate data for all documents using latent semantic indexing. This projection shows significant performance but they are not suitable for real-time application as it does not coincide with the projection most appropriate for humans to solve a particular problem.

Devaney and Ram (1997) proposed a novel feature selection for an unsupervised learning task, called conceptual clustering. The author discuss about feature selection based on an existing COBWEB conceptual clustering system. From the evaluation result, it is clearly observed that the feature selection significantly improves. The limitation of this approach is that COBWEB is not scalable like K-Means. Therefore, COSA (Concept Selection and Aggregation) and Term Selection (TES) are more suited for practical purposes of clustering in large document repositories.
Barry de Ville et al, (2002) explained the data-mining classification and predictive modeling techniques that depends on bootstrapping techniques. This describes re-use of source data frequently that can offer a holographic view of the modeled data. This holographic application is mostly used in industrial area that involves text mining warranty claims at a major international car, truck, and heavy equipment manufacturer. This paper describes the working, shows the functioning of the text mining approach.

Li haiying et. al. (2002) have discussed about real-time clustering techniques that consists of two steps. The first step involves the extraction process in which partial parser and a shallow stemmer are invoked, it can be used for both linguistic and statistical methods, in order to minimize the term variations within the returned text snippets the system introduce normalization approaches.

Crescenzi et. al. (2004) proposed an approach that automatically extracts data from large data-intensive Web sites. The “data grabber” investigates a large Web site and infers a scheme for it, describing it as a directed graph with nodes describing classes of structurally similar pages and arcs representing links between these pages. After locating the classes of interest, a library of wrappers can be created, one per class with the help of an external wrapper generator and suitable data can be extracted.

Miha Grcar et. al. (2008) proposed an approach concerning lack of software mining approaches, which means process of extracting knowledge out of source code. The author proposed a software mining task with a combination of text mining and link analysis technique. This mainly deals with interlinks between one instance to another instance. Retrieval and knowledge based approaches are the two main techniques to construct tool for software component. First approach is the natural language documentation of the software components. No interpretation of the documentation is made with this technique but information is extracted through statistical analyses of the words distribution. Manually provided information Knowledge-based systems can be significant than IR systems but they suffer from the scalability issue.
An ontology-learning framework named Link-Analysis and Text-mining toolbox (LATINO) is recently developed by Mladenic et. al. (2006). LATINO is an open source common purpose data mining platform providing text mining, link analysis, machine learning, and data visualization capabilities.

There are several techniques available in literature that can be classified into two major categories namely similarity-based approach and model-based approach. These approaches are described in Pallav Roxy and Durga Toshniwal (2009). Similarity-based technique is a pair wise similarity document clustering technique. The main goal of this technique is that the average similarities within clusters have to be maximized and the average similarities between the clusters should be minimized. Alternatively, Model-based approaches try to learn generative techniques from the documents, with each approach representing one particular document group. Several techniques are available in the literature for document clustering.

New technique such as self-organizing map Kohonen et. al. (2000), mixture of Gaussians Tantrum et. al. (2002), spherical k-mean S. Dhillon, D. S. Modha (2001), bi-secting k-means Steinbach et. al. (2000), mixture of multinomial Vaithyanathan. S and Dom.B. (2000) have being developed to improve the performance of the clustering techniques. K-means is one of the simplest unsupervised learning approaches that solve the well known clustering problem. The fundamental concept is to define k centroids, for each cluster. These centroids should be positioned far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid, then need to re-compute k new centroids. Then a new binding has to be done between the same data set points and the nearest new centroid. A loop has been created. As a result of this loop it is observed that the k centroids alter their location step by step until no more changes are made. This algorithm aims at minimizing an objective function.

The text data of text mining has turned out to be a most favorite research area. The study of the text clustering has attracted extensive consideration. An enhanced fuzzy clustering-text clustering approach is proposed by Jiabin Deng et. al. (2010) which depends on the fuzzy C-means clustering approach and the edit distance
approach. This approach makes use of the feature estimation to decrease the dimensionality of high-dimensional text vector. Since the clustering output of the conventional fuzzy C-means clustering approach requires the stability, the authors establish the high-power sample point set, the field radius and weight. Because of the boundary value attribution of the conventional fuzzy C-means clustering approach, the authors suggest the edit distance approach. The experimental outputs confirm that the application of the enhanced approach to the text clustering produces more constant and precise clustering results than the conventional FCM clustering algorithm.

An enhanced data clustering approach for mining web documents is proposed by Odukoya et al., (2010) which formulates, simulates and assess the web documents with the intention of conserving their conceptual relationships and removing the difficulty of speed while increasing accuracy. The enhanced data clustering approach was formulated with the concept of K-means algorithm. Real and fake datasets were utilized to analyze the proposed and existing approaches. The proposed approach was simulated using the fuzzy logic and statistical toolbox in Matlab 7.0. The experimental results were evaluated with the existing data clustering approaches by means of accuracy, response time, adjusted rand index and entropy as performance attributes. The output illustrates that the enhanced data clustering approach with a new initialization technique based on finding a set of medians obtained from a dimension with greatest variances. The outcomes of the simulation confirmed that the accuracy is at its maximum when the number of clusters is 3 and decreases with the increase of the number of clusters.

Semantic concepts

Majority of the available clustering algorithms in the Chinese text clustering, suffer from the drawbacks of data's scalability and the results' interpretability. Liu Jinling and Zhou Hong (2010) presents a well-organized Chinese text clustering technique depending on the semantic concepts. This technique, initiating from the text itself, by utilizing classified hierarchy Subject Word in Thesaurus of Modern Chinese, obtains the conceptional tuple from a high-dimensional text vector collection to create the high-level concept expressing clustering outcomes. Then samples are segmented according to these high-level concepts which indicate the complete text clustering
process has been done. This technique can significantly reduce the number of required data to be processed and develops the clustering approaches scalability. The experimental output shows that this clustering approach has attained an acceptable clustering outcome and better implementation efficiency.

Renchu Guan et. al. (2011) proposed a novel semisupervised text clustering approach called Seeds Affinity Propagation (SAP) based on an Affinity Propagation (AP) algorithm. There are two most important contributions in this proposed technique are:

- A novel similarity metric that captures the structural information of texts,
- A new seed construction technique to enhance the semisupervised clustering process.

Eisenhardt et. al. (2003) proposed an approach for document clustering using a distributed peer-to-peer network. The author used the $k$-means clustering approach, changed to work in a distributed P2P fashion using a probe and echo system. This is also used by Li and Morris (2002) but the problem is posed from the information retrieval perspective. Cluster signatures are generated from a subset of the document collection by centrally dividing it into clusters. Each cluster is then allotted to a node, and then documents are classified to their relevant clusters by comparing their signature with all cluster signatures.

Datta et. al. 2006 proposed an exact local approach for monitoring a $k$-means clustering, in addition to an approximate local $k$-means clustering approach for P2P networks. The P2P $k$-means approach has its fundamental in a parallel implementation of $k$-means proposed by Dhillon and Modha (2000). Though, $k$-means monitoring algorithm does not generate a distributed clustering, it assists a centralized $k$-means process and recompute the clusters by monitoring the distribution of centroids across peers, and triggering a re-clustering if the data distribution changes over time. Alternatively, the P2P $k$-means approach works by updating the centroids at each peer depending on information obtained from their immediate neighbors. The algorithm
stops when the information obtained does not result in considerable update to the centroids of all peers.

(Shih-Sian Cheng et al.,) suggested a genetic algorithm (GA) for K-means clustering. The author shares his view that in case of using string of- group-numbers encoding, the prototypes of the clusters are encoded into the chromosomes. To interchange the prototypes between two chromosomes, crossover operator is implemented. One-step K-means approach is utilized as the mutation operator. Hence the developed GA is called as the prototypes-embedded genetic K-means algorithm (PGKA). PGKA performs better than the classical K-means algorithm with the progression method of evolutionary algorithms. PGKA is more efficient and appropriate for large scale data sets when compared with other GA techniques.

Lu et. al. (2004) proposed fast genetic k-means cluster technique (FGKA). It is a quickest version of GKA and FGKA and has lot of development over GKA comprising an effective estimation of the objective value TWCV (Total Within-Cluster Variation) which reduces illegal string elimination overhead, and a generalization of the mutation operator. These enhancements resulted that FGKA runs 20 times quicker than GKA [58]. Haung et. al. (2005) Even though FGKA is outperformed by GKA considerably, it has some potential disadvantage. When the mutation possibility is low then the number of allele changes also low. The cost to compute the centroids and TWCV from score will be of more costly. To overcome the difficulties of FGKA, (Lu et al., 2004) formulated an incremental genetic k-means approach (IGKA). IGKA take over all the benefits of FGKA including the convergence to the global optimum, and when the mutation possibility is low this method outperforms FGKA. The main focus of IGKA is to incrementally compute the objective value TWCV and to cluster centroids. When the mutation possibility is lesser than the threshold then IGKA comparatively performs well than FGA. But when the mutation possibility is higher than the threshold, FGA performs better. Hence, proposed the hybrid genetic k-means algorithm (HGKA) and holds the advantages of both FGKA and IGKA and performs better in higher and lower mutation probability. Fundamental basics of these GA dependent clustering methods are k-means clustering and it can be used only for numeric data sets. Consequently, a
genetic clustering algorithm called GKMODE is proposed. GKMODE combines both the k-modes algorithm proposed by Chaturvedi et. al. (2001) and the genetic approach. GKMODE perform well in categorical data but it is very difficult to manage both the combination of numeric and categorical data.

Genetic approach is an analytical population-dependent search technique which has three most important operators namely crossover, mutation and selection. Selection operator performs a vital role in discovering optimal result for constrained optimization complexity. Ke-Zong Tang et al., (2010) formulated an improved genetic algorithm (IGA) depending on a novel selection approach is offered to manage nonlinear programming difficulties. All the entities in selection method are characterized as a three-dimensional feature vector comprised of objective function value, the measure of constraints violations and the amount of constraints violations. Based on Pareto partial order the outstanding individuals can be differentiated through two indices. Moreover, IGA integrates a local search (LS) method into selection operation to enable us to locate possible solutions placed in adjacent areas of some infeasible solutions. Experimental observation determines that proposed IGA has enhanced robustness, efficiency and steadiness than other technique.

By knowing the value of data hidden in huge collection of databases, most of the organization has started to built data warehouses. Two characteristics are essential in the data warehousing environment namely analysis and multiple updates. These characteristics have helped to develop new approach called Data Mining. Data mining has been defined as the application of data analysis and discovery algorithms that - under acceptable computational efficiency limitations - produce a particular enumeration of patterns over the data. Numerous data mining techniques have been recognized namely clustering, classification and summarization. But the author mainly focuses on clustering technique. All the data are not updated instantly in data warehouse when insertion and deletion operations are done. Frequently these updates are gathered and changes are made to the data warehouse in a batch mode. As the databases are of very huge collection of information, it is very difficult to cluster all data for each updates. So, it is necessary to carry out these updates incrementally.
Data warehouse offers a great deal of chance for performing data mining operations like classification and clustering. Frequently these updates are gathered and applied to the data warehouse. The entire sample obtained from the warehouse by some data mining approach has to be updated as well. The new technique depending on Genetic algorithm was developed by Atul Kamble, (2010). This approach can be applied to all database includes data from a metric space. It is confirmed that the incremental approach provides better output than any other approach. Incremental Clustering using Genetic Algorithm (ICGA) is a performance estimation approach on spatial database which represents the effectiveness of the proposed approach. ICGA provides speed-up aspects compared to other clustering algorithms.

Clustering is one of the most significant data mining methods and intends to combine the data objects into a meaningful class by increasing the similarities of objects within clusters and the similarities between objects of different clusters is reduced. A clustering approach depending on Genetic k-means model that performs better for data contains both numeric and categorical attributes was proposed by Dharmendra K Roy and Lokesh K Sharma, (2010). The author formulated a customized explanation of cluster center to avoid the numeric data is the major drawback of Genetic k-mean algorithm and offers an enhanced characterization of clusters.

An n-dimension convergence approach was developed by Min-Yuan Cheng et. al. (2010) developed to follow the possible development in conventional genetic algorithm (GA) by K-means clustering method. Then the chaotic approach was exploited to avoid the new approach from premature. This approach not only maintains the fundamental search capability but also enhanced the flexibility and effectiveness of parametric method.

This approach exhibits how to enhance the GA optimizer by integrating a hybridization policy. Experimental observation shows that the hybrid Chaotic method with Genetic Algorithm (CGA) method will give good results compared to other optimization approaches. K-means clustering is combined with CGA, called KCGA, which provides greater convergence performance than the other approaches. This approach provides 84% of probability to get optimization. This approach has been tested
with many construction examples, all the results confirms that this method is very effective.

The main difficulty in most of the application is the generic drift when using GA. This implies that the search may stick in local optima without any additional development towards the optimal solution. This occurs as GA usually takes only a sample part of the individuals of a population without taking the complete search space. As a result, it is necessary to maintain population size as huge as possible to eliminate this problem. The next significant problem in this layer concerns population diversity. Many schemes have been formulated to maintain diversity which includes migrations, local selection, minimal generation gap, and local search. The utilization of GA in complex applications necessitates high computational effort to compute individuals and hence it is hard to maintain large populations. Several approaches have been implemented to estimate strength of individuals instead of directly estimating them. One probability is to believe that individuals are someway genetically associated with each other. In this method, huge population size can be managed by clustering the population into a collection of related individuals. Clustering procedures are extensively used in many real world applications like image processing, pattern recognition, classifiers, machine learning. The most significant cluster method is fuzzy c-means, which identifies the fact that clustering is a common imprecise and with different level, the object may be compatible with dissimilar clusters.

The clustering ensemble is a novel technique which is of interest in the areas of machine learning. It can integrate multiple partitions created by different clustering approaches into a single clustering solution. Genetic algorithms are the techniques with high ability to find the solution of optimization problems like the clustering ensemble problem. There are several contributions done to find consensus cluster partition by genetic algorithms; but very little approaches are available for techniques available for carrying out the initialization population and generation of initial cluster partitions in the first phase of clustering ensembles. (Ghaemi et al., 2010) proposed a new approach that used by clustering ensembles which enhance cluster partitions fitness. Moreover, the proposed approach solves the diversity clustering problem. The fitness average among
individuals generated by the proposed algorithm and other clustering algorithms which have been calculated by three different fitness functions are compared. The experimental observation on several benchmark datasets have showed that the proposed algorithm improve cluster solutions fitness.

This research proposed by Jia Zhen and Wang Yong Gui, (2010) presents a genetic clustering approach depending on dynamic granularity. From the perspective of a parallel, random search, global optimization and diversity features of genetic algorithm, it is integrated with dynamic granularity approach. In the process of granularity changing, suitable granulation can be made by refining the granularity, which can improve the efficiency and the accuracy of the clustering algorithm. From the experimental results, it is observed that the approach enhances the clustering algorithm based on genetic algorithm local search ability and convergence speed.

Chen Wei et al., (2010) proposed a technique which deals with the drawbacks of the genetic Fuzzy C Means (FCM) such as higher classification time and less clustering accuracy. An enhanced algorithm which improves the crossover, selection, and mutation parts of the GA, improves its global searching capability and eliminates the trouble in setting up genetic parameters is proposed in this approach. This enhanced approach performs FCM optimization immediately after each generation of genetic operation, which increases the converging speed. The clustering accuracy, converging speed and stability of the approach are found to be better than the conventional clustering approaches.

Chen Rui et al., (2010) proposes an enhanced genetic algorithm. This algorithm keeps the population diversity by similarity checks on the population before selection, and the approach resolves the early-maturing issue of the population evolution, and presents a formula for mutation probability connected with similarity rate and iteration times. The approach not only retains a good diversity of population, but also ensures the convergence of the approach. The experiments are conducted in UCI datasets of WINE and IRIS. From the experimental results, it is observed that, this approach shows better performance when compared to c-means clustering algorithm.
Text clustering is very important in peer-to-peer networks. The high dimensionality of documents indicates more communication could be saved if each node could get the estimated clustering result by distributed algorithm rather than transferring them into a center and perform clustering. K-means algorithm forms the basis for a lot of existing text clustering algorithms in unstructured peer-to-peer networks. The main issue in the existing approaches is that the clustering quality may decrease with the increase of the network size.

A novel text clustering approach based on frequent term sets for peer-to-peer networks is proposed by Qing He et. al., (2010). It needs relatively lower communication volume while attaining a clustering result whose quality is not affected by the size of the network. Furthermore, it provides a term set describing each cluster, which makes the user to have a clear comprehension for the clustering result, and assists the users to find resource in the network or manage the local documents in accordance with the whole network.

This paper proposed by Zhao-Kui Li and Yan Wang (2010), deals with the topology mismatch in structured P2P network, routing advanced approach using super-nodes and self-organizing clustering. This approach includes self-organizing clustering depending on the physical location data in structured P2P system and each node judges if it can be a clustering header depending on its ability. In this approach, a series of dominant nodes are chosen as super-nodes, which mostly take charge of clustering header nodes of self-organizing clustering and non-clustering nodes. From the experimental results, it is revealed that, this approach can obviously improve routing performance and solve topology mismatch.

Zhongjun Deng (2010) examined clustering algorithm in Peer-to-Peer (P2P) network. Conventional clustering approaches cannot be applied P2P systems because of its lack of central control and very large size. The author proposed a hybrid clustering approach, called P2PKMM. In each node, the K-medoids algorithm is used. Therefore, local noise can be greatly avoided. In the meantime, K-means algorithm is used between different nodes, which can be computed easily over distributed environment. The proposed approach takes an entirely decentralized approach, in which the peers
synchronize only with their immediate topological neighbors in the underlying communication network. Moreover, this approach can adapt to dynamic P2P network in which the existing nodes withdraw and new nodes are connected during the execution of the approach and the data in network gets altered. Experimental results reveal that the proposed P2PKMM approach produce accurate clustering results with high scalability.

A correlation-based clustering hierarchical P2P network approach is proposed by Yuan Li and Xia Shixiong (2010). This approach overcomes the drawbacks of poor scalability and low search efficiency in unstructured P2P networks. This approach partitions the whole unstructured P2P network into clusters formed by a number of nodes through correlation. Each cluster chooses a master node which is to be in charge of managing the whole cluster. The entire network is partitioned into 2 layers. The upper layer is a structured network which consists of the master nodes of each cluster and the lower layer is an unstructured network which contains client nodes of a cluster. The simulation results show the significant performance in the query success rate and query delay.

L-Fuzzy context was proposed by (Pollandt, 1996) as an effort to integrate fuzzy logic with FCA. The L-Fuzzy context utilizes linguistic variables, which are linguistic terms related with fuzzy sets, to characterize uncertainty in the context. But, human interpretation is necessary to define the linguistic variables. Furthermore, the fuzzy concept lattice created from the L-fuzzy context generally generates a combinatorial explosion of concepts as compared to the conventional concept lattice.

Liping Jing deals with clustering of text documents into different category groups is a vital step in indexing, retrieval, management and mining of abundant text data on the Web or in corporate information systems. Significant text clustering technique enables better information services by browsing and organizing documents into meaningful cluster hierarchies and offers a useful complement for traditional text search engines when key-word based search returns too many documents. The challenging problems in text clustering include big volume, high dimensionality and complex semantics. This paper is mainly focussed on these problems. These problems can be solved by, subspace clustering, and by ontology technique. This paper use
bottom up sub-space search approach, Iterative top-down subspace search methods, Adaptive Subspace Iteration (ASI) and Simultaneous Keyword Identification and Clustering of text documents (SKWIC).

Formal Concept Analysis (FCA) is a proper technique for data analysis and knowledge presentation. FCA is considered an efficient technique for conceptual clustering in recent years. But, as most concept lattices are fairly complex in terms of the number of concepts created, it is essential to simplify the lattice generated. In Iceberg concept lattice Stumme et. al. (2002), association rules are used to cluster concepts on the lattice, and conceptual scaling Vogt and Wille (1995) is then used to generate the concept hierarchy.

Raghu Krishnapuram et. al. (2001) presented new techniques (Fuzzy C-Medoids FCMdd and Fuzzy C Trimmed Medoids or FCTMaZ) for fuzzy clustering of relational data. The objective functions are based on choosing c representative objects (medoids) from the data set in such a way that the total dissimilarity within each cluster is minimal. The experimental observations are conducted by comparing FCMdd with the Relational Fuzzy c-Means algorithm (RFCM). The result shows that FCMdd is much faster and efficient. The worst case complexity of the algorithm is O (n) which occurs while updating the medoids in each iteration. These approaches are very well applied in Web mining applications such as categorization of Web documents, snippets, and user sessions.

A technique to identify inter-transaction fuzzy association rules that can calculate the variations of events was proposed by (Yo-Ping Huang and Li-Jen Kao, 2005). The proposed algorithm first mapped a quantitative attribute into several fuzzy attributes. A normalization process was taken to avoid the total contribution of fuzzy attributes from being larger than one. In order to mine inter-transaction fuzzy association rules, both the dimensional attribute and sliding window concepts were proposed in this technique.

A complete comparison investigation of text document clustering for a biomedical digital library MEDLINE was proposed by (Lilhoi Yoo and Xiaohua Hu,
2006). Document clustering technique can provide the advantages of better document retrieval, document browsing, and text mining in digital library. The investigation on several document clustering techniques such as three hierarchical methods such as single-link, complete-link, and complete link, Bisecting K-means, K-means, and Suffix Tree Clustering in terms of the efficiency, and the scalability etc have been done by the author. Moreover, the author uses domain ontology to document clustering technique to examine if the ontology such as MeSH, which enhances clustering, qualifies for MEDLINE articles. Ontology is a formal, explicit specification of a shared conceptualization for a domain of interest. Ontology can be used to solve conventional information retrieval problems such as synonym/hypernym/hyponym. The author conducted widespread experiments based on various evaluation metrics such as misclassification index, F-measure, cluster purity, and entropy on very large article sets from MEDLINE.

A concept about multiple-level taxonomy and multiple minimum supports to identify fuzzy association rules in a given quantitative transaction data set was proposed by Yeong-Chyi Lee et. al. (2006). The significance of different items are judged using different criteria, managing taxonomic relationships among items, and dealing quantitative data sets are three issues that generally occur in real mining applications. This fuzzy mining approach can create large itemsets level by level and then derive fuzzy association rules from quantitative transaction data.

A novel approach called cluster-based fuzzy-genetic mining technique for extracting both fuzzy association rules and membership functions from quantitative transactions was proposed by (Chun-Hao Chen et al., 2006). The technique can dynamically adjust membership functions by genetic algorithms and uses them to fuzzify quantitative transactions. It can also accelerate the evaluation procedure and keep nearly the same quality of solutions by clustering chromosomes. Each chromosome represents a set of membership functions used in fuzzy mining. This approach first divides the chromosomes in a population into k clusters by using the k-means clustering approach. All the chromosomes in a cluster then use the number of large 1-itemsets derived from the representative chromosome in the cluster and their
own suitability of membership functions to calculate the fitness values. The evaluation cost can thus be reduced due to the time-saving in finding 1-itemsets.

Fuzzy ontology techniques using intuitionistic fuzzy set for knowledge sharing on the semantic web has been proposed by (Jun Zhai et al., 2008). Ontology is implemented as a standard for knowledge representation and sharing in collaborative design on the semantic Web. Fuzzy ontology is an extended version of domain ontology for solving the uncertainty problems. The author proposes a series of fuzzy ontology techniques that consists of fuzzy domain ontology, using intuitionistic fuzzy set, and fuzzy linguistic variable ontologies, considering semantic relationships between fuzzy concepts, including set relation, order relation and equivalence relation. The fuzzy ontology establishes the fuzzy concepts as values of properties and extends the ordinary relationships to fuzzy relationships and intuitionistic fuzzy relationships.

(Hongwel Chen et al., 2009) proposed a fuzzy trust problem domain for P2P-based system, and compare Fuzzy Comprehensive Evaluation technique, Fuzzy Rank-ordering technique, and Fuzzy Inference approach through a concrete model. In this model, the author had applied approach to Fuzzy Comprehensive Evaluation Method for P2P-based trust system, and Blin algorithm to that of Fuzzy Rank-ordering Method, and Mamdani approach to that of Fuzzy Inference technique. Results show that different fuzzy trust technique for P2P-based system may deduce different fuzzy results.

(Matsumoto et al., 2010) proposes a prototype web search results clustering engine that improves search results by performing fuzzy clustering on web documents returned by traditional search engines, as well as ranking the results and labeling the resulting clusters. This is done using a Fuzzy Transduction-based Clustering Approach (FTCA), which uses a Transduction-based Relevance Model (TRM) to create document relevance values. These relevance values are used to cluster similar documents, rank them, and assist a term frequency based label generator. The membership degrees of documents to fuzzy clusters also help efficient detection and removal of overly similar clusters. FTCA is compared against two other web document clustering approaches namely Suffix Tree Clustering (STC) and Lingo, which are provided by the free open source Carrot2 Document Clustering Workbench. In order to compute cluster quality, an
improved version of the classic precision measurement is used to take into account relevance and fuzzy clustering, along with recall and F1 score. Experimental observations on five different datasets reveal a significant clustering quality and performance advantage over STC and Lingo approaches.

Feature clustering is a dominant technique to minimize the dimensionality of feature vectors for text classification. (Jung-Yi Jiang et al., 2011) propose a fuzzy similarity-based self-constructing approach for feature clustering. The words in the feature vector of a document set are grouped into clusters, depending on similarity test. Words that are similar to each other are grouped into the same cluster. Each cluster is featured by a membership function with statistical mean and deviation. When all the words have been put in, a desired number of clusters are formed automatically. The authors then have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this approach, the derived membership functions match closely with and illustrate appropriately the real distribution of the training data. Moreover, the user need not state the number of extracted features in advance, and trial-and-error for determining the suitable number of extracted features can then be avoided. From the experimental observation, it is clear that the proposed approach can run faster and obtain better extracted features than other techniques.

A novel hybrid genetic algorithm (GA) is proposed by (K. Krishna and Narasimha Murty, 1999). This technique detects a globally optimal separation of a given data into a described number of clusters. To create an appropriate child chromosome from the parent chromosomes, GA’s used clustering which utilizes either crossover operator or fitness function or both the methods. To avoid these costly methods, the author hybridized GA with a traditional gradient descent algorithm applied in clustering includes K-means algorithm. Hence it is named as genetic K-means algorithm (GKA). K-means operator is described as a single-step of K-means algorithm which is implemented in GKA and used as a search operator as a replacement of crossover. Distance-based-mutation is defined to be a biased mutation operator specific to clustering. It is proved that GKA converge to the global optimum based on Markov
chain theory. Experimental observation shows that GKA can converge to the finest known optimum equivalent for the given data in concurrence with the convergence result. Result also proves that GKA searches faster than other algorithms used for clustering.

Casillas et al. (2003) put forth a novel concept on document clustering using genetic algorithm. The authors present a genetic algorithm that deals with document clustering that computes an approximation of the optimum k value, and resolves the best clustering of the documents into k clusters. The author have experimented the proposed technique with sets of documents that are the output of a query in a search engine. The simulation results show that the proposed genetic algorithm attain better values of the fitness function than the well known Calinski and Harabasz, (1974) stopping rule and performs in only lesser time.

Hybrid Genetic Algorithm (HGA) proposed by Francisco Mota Filho and Fernando Gomide, (2005) uses the fuzzy c-means clustering technique during the fitness evaluation phase to decrease direct evaluations. The main goal is to enhance the processing speed of the evolutionary process, but retaining a satisfactory level of population diversity and solution quality, that is, to increase chances to acquire as good solutions as traditional GA. Previous experiments in actual practical situations, namely train scheduling and dispatch in freight railroads, have shown the effectiveness of this technique as an approach to solve complex, large scale problems. When population clustering is used, the author mainly concentrates on the quality of solutions obtained and hence investigates a set of classical optimization problems available in the literature.

(J.M Pena et al., 1999) evaluates four initialization approaches for the K-Means algorithm namely random, Forgy, MacQueen and Kaufman. Even though this approach is popular for its robustness but the performance is mainly based on two key points which are initial clustering and instance order. The author performed a many researches to illustrate the probability distribution of the square-error rates of the last clusters returned by the K-Means algorithm separately on all the initial clustering. Experimentation observations demonstrates that this method generate the K-Means in a
very efficient way and independent on any initial clustering at any order. The random and the Kaufman initialization approaches provide significant performance when compared with other approaches. Also the convergence rate of the K-Means algorithm is compared when utilizing every four initialization techniques. The result shows that the proposed approach induces the K-Means technique in a more desirable manner in accordance to the convergence rate than the random initialization method.

A semi-supervised clustering approach is developed by (Ayhan Demiriz et al., 1999) that integrates the advantages of both supervised and unsupervised learning processes to clear the classification problems. Unsupervised learning approach is utilized to cluster or segment the data and it is biased toward generating significant clusters in terms of class distribution. These clusters are utilized to judge the class of future points. For example in database advertising, this approach can be implemented to recognize and differentiate the segments of the client’s population most probably to respond to a advertising. The significant advantage of this method is that it permits unlabeled data with unknown class to be taken to increase the classification accuracy.

The main goal of an unsupervised technique is to reduce the cluster variance of the input features and a measure of cluster impurity depending on the class labels. Reducing the within cluster variance is the method of ability control to avoid over fitting. Decision tree approaches like Gini index can be utilized to measure the impurities for the output labels. A genetic algorithm utilizes the objective method to generate clusters. Non-empty clusters are tagged with the superiority class. Experimental observation shows this method increases the generalization capability by utilizing the class information when compared to unsupervised methods which depends only on the input features. The observation also determines that this approach significantly works better even with less training examples. Also, training using information from unlabeled data can increase classification accuracy on that data.

There are several approaches such as Natural Language Processing (NLP) combined with association rules Maedche and Staab, (2001), statistical model Faatz and Steinmetz, (2002), and clustering Bisson and Nedellec, (2000) that have been applied to create ontology from a concept hierarchy automatically or semi-automatically.
Clustering is one of the most efficient techniques for ontology learning. Moreover, conceptual clustering techniques such as COBWEB are powerful clustering techniques that can be used for the generation of concept representations and relationships for ontology Bisson and Nedellec (2001).

Andreas et. al. (2002) has made a discussion on the clustering technique for text data. Text clustering usually involves clustering in a high dimensional space that appears complex with considered to virtually all practical settings. Additionally, provided a scrupulous clustering outcome it is normally very tough to come up with a good clarification of why the text clusters have been created the way they are. In this paper, a novel technique is presented for applying background information during preprocessing for improving the clustering outcome and permit for selection between outcomes. The author preprocesses the input data supplied to ontology-based heuristics for feature selection and feature aggregation. Therefore, various choices for text illustrations are constructed. Based on these illustrations, the authors calculate the multiple clustering outcomes using K-Means. The achieved results by the author compare favorably with a sophisticated baseline preprocessing strategy.

Affinity-based similarity measure for Web document clustering is presented by Shyu et. al. (2004). In this paper, the concept of document clustering is extended into Web document clustering by establishing the approach of affinity based similarity measure, which makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for evaluation with the help of real data set and the experimental results illustrated that the presented similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

Web document clustering using document index graph is put forth by Momin et. al. (2006) Document clustering methods generally based on single term examination of document data set. To attain more precise document clustering, more informative feature like phrases are essential in this scenario. Therefore first part of the paper provides phrase-based model, Document Index Graph (DIG) that permits incremental phrase-based encoding of documents and capable phrase matching. It stress on
efficiency of phrase-based similarity measure over conventional single term based similarities. In the second part, a Document Index Graph based Clustering (DIGBC) algorithm is provided to improve the DIG model for incremental and soft clustering. This technique incrementally clusters documents based on presented cluster-document similarity measure. It permits assignment of a document to more than single cluster.

Cao et al. (2008) provided fuzzy named entity-based document clustering. Conventional keyword-based document clustering methods have restrictions because of simple treatment of words and rigid partition of clusters. In this paper, the author introduces named entities as objectives into fuzzy document clustering, which are the important elements defining document semantics and in many cases are of user concerns. Initially, the conventional keyword-based vector space representation is adapted with vectors defined over spaces of entity names, types, name-type pairs, and identifiers, alternative of keywords. Next, hierarchical fuzzy document clustering can be applied using a similarity measure of the vectors indicating documents.

A novel technique on ontology based classification was proposed by Lena Tenenboim et. al. (2008). The author discussed on classification of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system comprises news items from different news suppliers and distributes to each subscribed user a personalized electronic newspaper, making use of content-based and collaborative filtering techniques. The ePaper can also offer users standard version of chosen newspapers, besides the browsing abilities in the warehouse of news items. This deliberates on the automatic categorization of incoming news with the help of hierarchical news ontology. Based on this clustering technique on one hand, and on the users' profiles on the other hand, the personalization engine of the system is able to afford a personalized paper to every user onto the mobile reading device.

Zhang et. al. (2008) gives clustering aggregation based on genetic algorithm for documents clustering. In this paper, a technique based on genetic algorithm for clustering aggregation difficulty, named as GeneticCA, is provided to approximate the clustering performance of a clustering division, clustering precision is defined and features of clustering precision are considered. In the evaluation concerning clustering
performances of GeneticCA for document clustering, hamming neural network is applied to make clustering partitions with fluctuant and weak clustering performances.

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun et. al. (2008). This paper involved in analyzing and making use of cluster overlapping technique to design cluster merging criterion. In this paper, the author presented a new method to calculate the overlap rate for improving time efficiency and the veracity. The technique is uses a line to pass across the two cluster's center as an alternative of the ridge curve. Depends on the hierarchical clustering technique, the expectation-maximization (EM) method is used in the Gaussian mixture model to count the parameters and formulate the two sub-clusters combined when their overlying is the biggest.

By considering the difficulty that classical Euclidean distance metric cannot create an suitable separation for data lying in a manifold, a genetic algorithm based clustering method with the help of geodesic distance measure is proposed by Gang Li et. al. (2009). In the proposed method, a prototype-based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids. In addition, a geodesic distance based proximity measures is applied to find out the similarity between data points. Simulation results on eight standard synthetic datasets with dissimilar manifold structure illustrate the effectiveness of the algorithm as a clustering technique. Evaluating with generic k-means method for the function of clustering, the proposed technique has the potential to distinguish complicated non-convex clusters and its clustering performance is obviously better than that of the k-means method for complex manifold structures.

A Wordsets based document clustering algorithm for large datasets was proposed by Sharma et. al. (2009). Document clustering is a significant tool for applications like search engines and document browsers. It facilitates the user to comprise a better overall observation of the data contained in the documents. The available techniques of document clustering, however, do not actually consider the special difficulties of text document clustering: very high dimensionality of the document, very large size of the datasets and understandability of the cluster
explanation. Also there is a strong requirement for hierarchical document clustering where clustered documents can be browsed based on the increasing specificity of topics. Frequent Itemset Hierarchical Clustering (FIHC) is a novel data mining technique for hierarchical grouping of text documents. The technique does not provide consistent clustering results when the number of frequent sets of terms is large. In this paper the author propose WDC (Wordsets-based Clustering), an efficient clustering technique based closed words sets. WDC makes use of hierarchical technique to cluster text documents having common words.

Muflikhah et. al. (2009) proposed a document clustering technique using concept space and cosine similarity measurement. This paper aims to incorporate the information retrieval technique and document clustering technique as concept space approach. The technique is known as Latent Semantic Index (LSI) approach which used Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). The intention of this technique is to decrease the matrix dimension by identifying the pattern in document collection with refers to simultaneous of the terms. Every technique is employed to weight of term-document in vector space model (VSM) for document clustering with the help of fuzzy c-means technique. In addition to the reduction of term-document matrix, this research also utilizes the cosine similarity measurement as alternative of Euclidean distance to engage in fuzzy c-means.

ELdesoky et. al. (2009) given a novel similarity measure for document clustering based on topic phrases. In the conventional Vector Space Model (VSM) researchers have used the unique word that is contained in the document set as the candidate feature. Currently a latest trend which uses the phrase to be a more informative feature has considered; the issue that contributes in enhancing the document clustering accuracy and effectiveness. This paper presented a new technique for evaluating the similarity measure of the traditional VSM by considering the topic phrases of the document as the comprising terms for the VSM instead of the conventional term and applying the new technique to the Buckshot technique, which is a combination of the Hierarchical Agglomerative Clustering (HAC) technique and the K-
means clustering method. Such a method may increase the effectiveness of the clustering by incrementing the evaluation metrics values.

Document clustering with the help of fuzzy c-mean algorithm is proposed by Thaung et. al. (2010). Most traditional clustering technique allocate each data to exactly single cluster, therefore creating a crisp separation of the provided data, but fuzzy clustering permits for degrees of membership, to which a data fit various clusters. In this paper, documents are partitioned with the help of Fuzzy C-Means (FCM) clustering technique. Fuzzy c-means clustering is one of famous unsupervised clustering methods. But fuzzy c-means method needs the user to mention the number of clusters and different values of clusters corresponds to different fuzzy partitions earlier itself. So the validation of clustering result is required. PBM index and F-measure are helpful in validating the cluster.

With the arrival of social networks and tagging systems, Internet has recently observed a big leap in the use of Web Recommendation Systems (WRS). According to users’ interest of items and their browsing history on the World Wide Web, these systems are able to forecast and recommend items and future purchases to users. They are being used now in several domains, like news article recommendation, product recommendation, and make-friend recommendation. WRS still has several drawbacks, like sparsity and the new user problem. Moreover, WRS does not make full use and harness of the power of domain knowledge and semantic web ontologies. Mabroukeh et. al. (2011) discussed how an ontology-based WRS can use relations and concepts in ontology, along with user-provided tags, to provide top-n recommendations without the need for item clustering or user ratings. For this reason, the author also proposed a dimensionality reduction technique based on the domain ontology, to solve the sparsity problem.

Most of Web Services exist without explicit associated semantic descriptions. Hence many services that are related to a specific user service request may not be regarded during service discovery. (Paliwal et al., 2011) address the problem of Web Service discovery given non-explicit service description semantics that match a specific service request. This proposed approach to semantic based Web Service discovery
involves semantic based service categorization and semantic enhancement of the service request. The author proposes a solution for achieving functional level service categorization based on an ontology framework. Moreover, the author uses clustering for accurately classifying the Web Services depending on service functionality. The semantic based categorization is performed offline at the Universal Description Discovery and Integration (UDDI). The semantic enhancement of the service request attains a better matching with relevant services. The service request enhancement involves expansion of additional terms (retrieved from ontology) that are deemed relevant for the requested functionality. A proficient matching of the enhanced service request with the retrieved service descriptions is obtained using Latent Semantic Indexing (LSI). The experimental results validate the effectiveness and feasibility of the proposed approach.

Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form LHS→RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. the diagnosis of a tumor sample from which a profile was obtained). In this paper, association rule mining techniques that have been recently developed and used for genomic data analysis have been reviewed and discussed. There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as gene expression microarrays. These large genomic data sets are information-rich and often contain much more information than the researchers who generated the data might have anticipated. Such an enormous data volume enables new types of analyses, but also makes it difficult to answer research questions using traditional methods. Global Gene expression data can be a valuable tool in understanding of genes, biological networks, and cellular states. Analysis of these massive genomic data has two important goals: First goal is try to determine how the expression of any particular gene might affect the expression of other genes; the genes
involved in this case could belong to the same gene network. By a gene network, we mean a set of genes being expressed together in a non-random pattern. Second goal of expression data analysis is to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells.

Most of the available gene-expression data-analysis methods are based on clustering algorithms that try to establish synexpression groups [4], that is, groups of genes whose expression is correlated in different biological situations. The basis for all clustering algorithms is their ability to generate groups of genes that fulfill two related constraints: maximum intragroup similarities and minimum intergroup similarities. Although such algorithms have been quite successful, most notably in the molecular profiling of human cancers [5], their biological validity can be questioned when the identification of molecular networks is the goal. In this context, they have three main drawbacks. First, a gene which functions in numerous physiological pathways will have to be clustered in one and only one group. Second, no relationship can be inferred between the different members of a group. That is, a gene and its target genes will be co-clustered, but the type of relationship cannot be rendered explicit by the algorithm. Third, most clustering algorithms will make comparisons between the gene-expression patterns in all the conditions examined. They will therefore miss a gene grouping that only arises in a subset of cells or conditions. To overcome these problems, the potential impact of the association-rule discovery (ARD) technique is investigated. This is an unsupervised data-mining technique that seeks descriptive rules in potentially very large datasets [4]. This method should resolve the above drawbacks of existing clustering approaches for the following reasons. First, any gene can be assigned to any number of rules as long as its expression fulfills the assignment criteria. This means that a gene involved in many synexpression groups will appear in each and every one of those groups, without limitation. Second, rules are orientated (If ... then ...) and thus to a certain extent describe the direction of a relationship. For example, in the overall dataset, a specific subset of cells exhibit highly characteristic patterns of gene expression, the algorithm should be able to detect it. Last but not least, by focusing on
strong rules, the biologist does not have to browse and study a huge number of redundant rules.

Shih-Sian Cheng et al. suggested a genetic algorithm (GA) for K-means clustering. The author shares his view that in case of using string of group-numbers encoding, the prototypes of the clusters are encoded into the chromosomes. To interchange the prototypes between two chromosomes, crossover operator is implemented. One-step K-means approach is utilized as the mutation operator. Hence the developed GA is called as the prototypes-embedded genetic K-means algorithm (PGKA). PGKA performs better than the classical K-means algorithm with the progression method of evolutionary algorithms. PGKA is more efficient and appropriate for large scale data sets when compared with other GA techniques.

Lu et. al. (2004) proposed fast genetic k-means cluster technique (FGKA). It is a quickest version of GKA and FGKA and has lot of development over GKA comprising an effective estimation of the objective value TWCV (Total Within-Cluster Variation) which reduces illegal string elimination overhead, and a generalization of the mutation operator. These enhancements resulted that FGKA runs 20 times quicker than GKA [58] Haung et. al. (2005) Even though FGKA is outperformed by GKA considerably, it has some potential disadvantage. When the mutation possibility is low then the number of allele changes also low. The cost to compute the centroids and TWCV from score will be of more costly. To overcome the difficulties of FGKA, Lu et al., 2004) formulated an incremental genetic k-means approach (IGKA). IGKA take over all the benefits of FGKA including the convergence to the global optimum, and when the mutation possibility is low this method outperforms FGKA. The main focus of IGKA is to incrementally compute the objective value TWCV and to cluster centroids. When the mutation possibility is lesser than the threshold then IGKA comparatively performs well than FGA. But when the mutation possibility is higher than the threshold, FGA performs better. Hence, proposed the hybrid genetic k-means algorithm (HGKA) and holds the advantages of both FGKA and IGKA and performs better in higher and lower mutation probability. Fundamental basics of these GA
dependent clustering methods are k-means clustering and it can be used only for numeric data sets.

Consequently, a genetic clustering algorithm called GKMODE is proposed. GKMODE combines both the k-modes algorithm proposed by Chaturvedi et. al. (2001) and the genetic approach. GKMODE perform well in categorical data but it is very difficult to manage both the combination of n numeric and categorical data. Genetic approach is an analytical population-dependent search technique which has three most important operators namely crossover, mutation and selection. Selection operator performs a vital role in discovering optimal result for constrained optimization complexity.

Ke-Zong Tang et al., (2010) formulated an improved genetic algorithm (IGA) depending on a novel selection approach is offered to manage nonlinear programming difficulties. All the entities in selection method are characterized as a three-dimensional feature vector comprised of objective function value, the measure of constraints violations and the amount of constraints violations. Based on Pareto partial order the outstanding individuals can be differentiated through two indices. Moreover, IGA integrates a local search (LS) method into selection operation to enable us to locate possible solutions placed in adjacent areas of some infeasible solutions. Experimental observation determines that proposed IGA has enhanced robustness, efficiency and steadiness than other technique.

By knowing the value of data hidden in huge collection of databases, most of the organization has started to built data warehouses. Two characteristics are essential in the data warehousing environment namely analysis and multiple updates. These characteristics have helped to develop new approach called Data Mining. Data mining has been defined as the application of data analysis and discovery algorithms that - under acceptable computational efficiency limitations - produce a particular enumeration of patterns over the data. Numerous data mining techniques have been recognized namely clustering, classification and summarization. But the author mainly focuses on clustering technique. All the data are not updated instantly in data warehouse when insertion and deletion operations are done. Frequently these updates are gathered
and changes are made to the data warehouse in a batch mode. As the databases are of very huge collection of information, it is very difficult to cluster all data for each updates. So, it is necessary to carry out these updates incrementally.

Data warehouse offers a great deal of chance for performing data mining operations like classification and clustering. Frequently these updates are gathered and applied to the data warehouse. The entire sample obtained from the warehouse by some data mining approach has to be updated as well. The new technique depending on Genetic algorithm was developed by Atul Kamble, (2010). This approach can be applied to all database includes data from a metric space. It is confirmed that the incremental approach provides better output than any other approach. Incremental Clustering using Genetic Algorithm (ICGA) is a performance estimation approach on spatial database which represents the effectiveness of the proposed approach. ICGA provides speed-up aspects compared to other clustering algorithms.

Clustering is one of the most significant data mining methods and intends to combine the data objects into a meaningful class by increasing the similarities of objects within clusters and the similarities between objects of different clusters is reduced. A clustering approach depending on Genetic k-means model that performs better for data contains both numeric and categorical attributes was proposed by Dharmendra K Roy and Lokesh K Sharma, (2010). The author formulated a customized explanation of cluster center to avoid the numeric data is the major drawback of Genetic k-mean algorithm and offers an enhanced characterization of clusters.

An n-dimension convergence approach was developed by Min-Yuan Cheng et. al. (2010) developed to follow the possible development in conventional genetic algorithm (GA) by K-means clustering method. Then the chaotic approach was exploited to avoid the new approach from premature. This approach not only maintains the fundamental search capability but also enhanced the flexibility and effectiveness of parametric method.

This approach exhibits how to enhance the GA optimizer by integrating a hybridization policy. Experimental observation shows that the hybrid Chaotic method
with Genetic Algorithm (CGA) method will give good results compared to other optimization approaches. K-means clustering is combined with CGA, called KCGA, which provides greater convergence performance than the other approaches. This approach provides 84% of probability to get optimization. This approach has been tested with many construction examples, all the results confirms that this method is very effective.

The main difficulty in most of the application is the generic drift when using GA. This implies that the search may stick in local optima without any additional development towards the optimal solution. This occurs as GA usually takes only a sample part of the individuals of a population without taking the complete search space. As a result, it is necessary to maintain population size as huge as possible to eliminate this problem. The next significant problem in this layer concerns population diversity. Many schemes have been formulated to maintain diversity which includes migrations, local selection, minimal generation gap, and local search. The utilization of GA in complex applications necessitates high computational effort to compute individuals and hence it is hard to maintain large populations. Several approaches have been implemented to estimate strength of individuals instead of directly estimating them. One probability is to believe that individuals are someway genetically associated with each other. In this method, huge population size can be managed by clustering the population into a collection of related individuals. Clustering procedures are extensively used in many real world applications like image processing, pattern recognition, classifiers, machine learning. The most significant cluster method is fuzzy c-means, which identifies the fact that clustering is a common imprecise and with different level, the object may be compatible with dissimilar clusters.

The clustering ensemble is a novel technique which is of interest in the areas of machine learning. It can integrate multiple partitions created by different clustering approaches into a single clustering solution. Genetic algorithms are the techniques with high ability to find the solution of optimization problems like the clustering ensemble problem. There are several contributions done to find consensus cluster partition by genetic algorithms; but very little approaches are available for techniques available for
carrying out the initialization population and generation of initial cluster partitions in the first phase of clustering ensembles.

Ghaemi et al. (2010) proposed a new approach that used by clustering ensembles which enhance cluster partitions fitness. Moreover, the proposed approach solves the diversity clustering problem. The fitness average among individuals generated by the proposed algorithm and other clustering algorithms which have been calculated by three different fitness functions are compared. The experimental observation on several benchmark datasets have showed that the proposed algorithm improve cluster solutions fitness.

This research proposed by Jia Zhen and Wang Yong Gui, (2010) presents a genetic clustering approach depending on dynamic granularity. From the perspective of a parallel, random search, global optimization and diversity features of genetic algorithm, it is integrated with dynamic granularity approach. In the process of granularity changing, suitable granulation can be made by refining the granularity, which can improve the efficiency and the accuracy of the clustering algorithm. From the experimental results, it is observed that the approach enhances the clustering algorithm based on genetic algorithm local search ability and convergence speed.

Chen Wei et al., (2010) proposed a technique which deals with the drawbacks of the genetic Fuzzy C Means (FCM) such as higher classification time and less clustering accuracy. An enhanced algorithm which improves the crossover, selection, and mutation parts of the GA, improves its global searching capability and eliminates the trouble in setting up genetic parameters is proposed in this approach. This enhanced approach performs FCM optimization immediately after each generation of genetic operation, which increases the converging speed. The clustering accuracy, converging speed and stability of the approach are found to be better than the conventional clustering approaches.

Chen Rui et al., (2010) proposes an enhanced genetic algorithm. This algorithm keeps the population diversity by similarity checks on the population before selection, and the approach resolves the early-maturing issue of the population evolution, and
presents a formula for mutation probability connected with similarity rate and iteration times. The approach not only retains a good diversity of population, but also ensures the convergence of the approach. The experiments are conducted in UCI datasets of WINE and IRIS. From the experimental results, it is observed that, this approach shows better performance when compared to c-means clustering algorithm.

There are several approaches such as Natural Language Processing (NLP) combined with association rules Maedche and Staab, (2001), statistical model Faatz and Steinmetz, (2002), and clustering Bisson and Nedellec, (2000) that have been applied to create ontology from a concept hierarchy automatically or semi-automatically. Clustering is one of the most efficient techniques for ontology learning. Moreover, conceptual clustering techniques such as COBWEB are powerful clustering techniques that can be used for the generation of concept representations and relationships for ontology Bisson and Nedellec (2001).

Andreas et. al. (2002) has made a discussion on the clustering technique for text data. Text clustering usually involves clustering in a high dimensional space that appears complex with considered to virtually all practical settings. Additionally, provided a scrupulous clustering outcome it is normally very tough to come up with a good clarification of why the text clusters have been created the way they are. In this paper, a novel technique is presented for applying background information during preprocessing for improving the clustering outcome and permit for selection between outcomes. The author preprocesses the input data supplied to ontology-based heuristics for feature selection and feature aggregation. Therefore, various choices for text illustrations are constructed. Based on these illustrations, the authors calculate the multiple clustering outcomes using K-Means. The achieved results by the author compare favorably with a sophisticated baseline preprocessing strategy.

Affinity-based similarity measure for Web document clustering is presented by Shyu et. al. (2004). In this paper, the concept of document clustering is extended into Web document clustering by establishing the approach of affinity based similarity measure, which makes use of the user access patterns in finding the similarities among Web documents through a probabilistic model. Various experiments are conducted for
evaluation with the help of real data set and the experimental results illustrated that the presented similarity measure outperforms the cosine coefficient and the Euclidean distance technique under various document clustering techniques.

Web document clustering using document index graph is put forth by Momin et. al. (2006) Document clustering methods generally based on single term examination of document data set. To attain more precise document clustering, more informative feature like phrases are essential in this scenario. Therefore first part of the paper provides phrase-based model, Document Index Graph (DIG) that permits incremental phrase-based encoding of documents and capable phrase matching. It stress on efficiency of phrase-based similarity measure over conventional single term based similarities. In the second part, a Document Index Graph based Clustering (DIGBC) algorithm is provided to improve the DIG model for incremental and soft clustering. This technique incrementally clusters documents based on presented cluster-document similarity measure. It permits assignment of a document to more than single cluster.

Cao et al.(2008) provided fuzzy named entity-based document clustering. Conventional keyword-based document clustering methods have restrictions because of simple treatment of words and rigid partition of clusters. In this paper, the author introduces named entities as objectives into fuzzy document clustering, which are the important elements defining document semantics and in many cases are of user concerns. Initially, the conventional keyword-based vector space representation is adapted with vectors defined over spaces of entity names, types, name-type pairs, and identifiers, alternative of keywords. Next, hierarchical fuzzy document clustering can be applied using a similarity measure of the vectors indicating documents.

A novel technique on ontology based classification was proposed by Lena Tenenboim et. al. (2008). The author discussed on classification of news items in ePaper, a prototype system of a future personalized newspaper service on a mobile reading device. The ePaper system comprises news items from different news suppliers and distributes to each subscribed user a personalized electronic newspaper, making use of content-based and collaborative filtering techniques. The ePaper can also offer users standard version of chosen newspapers, besides the browsing abilities in the warehouse.
of news items. This deliberates on the automatic categorization of incoming news with the help of hierarchical news ontology. Based on this clustering technique on one hand, and on the users' profiles on the other hand, the personalization engine of the system is able to afford a personalized paper to every user onto the mobile reading device.

Zhang et. al. (2008) gives clustering aggregation based on genetic algorithm for documents clustering. In this paper, a technique based on genetic algorithm for clustering aggregation difficulty, named as Genetic CA, is provided to approximate the clustering performance of a clustering division, clustering precision is defined and features of clustering precision are considered. In the evaluation concerning clustering performances of Genetic CA for document clustering, hamming neural network is applied to make clustering partitions with fluctuant and weak clustering performances.

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun et. al. (2008). This paper involved in analyzing and making use of cluster overlapping technique to design cluster merging criterion. In this paper, the author presented a new method to calculate the overlap rate for improving time efficiency and the veracity. The technique is uses a line to pass across the two cluster's center as an alternative of the ridge curve. Depends on the hierarchical clustering technique, the expectation-maximization (EM) method is used in the Gaussian mixture model to count the parameters and formulate the two sub-clusters combined when their overlying is the biggest.

By considering the difficulty that classical Euclidean distance metric cannot create an suitable separation for data lying in a manifold, a genetic algorithm based clustering method with the help of geodesic distance measure is proposed by Gang Li et. al. (2009).

In the proposed method, a prototype-based genetic illustration is used, where every chromosome is a sequence of positive integer numbers that indicate the k-medoids. In addition, a geodesic distance based proximity measures is applied to find out the similarity between data points. Simulation results on eight standard synthetic datasets with dissimilar manifold structure illustrate the effectiveness of the algorithm as
a clustering technique. Evaluating with generic k-means method for the function of clustering, the proposed technique has the potential to distinguish complicated non-convex clusters and its clustering performance is obviously better than that of the k-means method for complex manifold structures.

A Wordsets based document clustering algorithm for large datasets was proposed by Sharma et. al. (2009). Document clustering is a significant tool for applications like search engines and document browsers. It facilitates the user to comprise a better overall observation of the data contained in the documents. The available techniques of document clustering, however, do not actually consider the special difficulties of text document clustering: very high dimensionality of the document, very large size of the datasets and understandability of the cluster explanation. Also there is a strong requirement for hierarchical document clustering where clustered documents can be browsed based on the increasing specificity of topics. Frequent Itemset Hierarchical Clustering (FIHC) is a novel data mining technique for hierarchical grouping of text documents. The technique does not provide consistent clustering results when the number of frequent sets of terms is large. In this paper the author propose WDC (Wordsets-based Clustering), an efficient clustering technique based closed words sets. WDC makes use of hierarchical technique to cluster text documents having common words.

Muflikhah et. al. (2009) proposed a document clustering technique using concept space and cosine similarity measurement. This paper aims to incorporate the information retrieval technique and document clustering technique as concept space approach. The technique is known as Latent Semantic Index (LSI) approach which used Singular Vector Decomposition (SVD) or Principle Component Analysis (PCA). The intention of this technique is to decrease the matrix dimension by identifying the pattern in document collection with refers to simultaneous of the terms. Every technique is employed to weight of term-document in vector space model (VSM) for document clustering with the help of fuzzy c-means technique. In addition to the reduction of term-document matrix, this research also utilizes the cosine similarity measurement as alternative of Euclidean distance to engage in fuzzy c-means.
ELdesoky et. al. (2009) given a novel similarity measure for document clustering based on topic phrases. In the conventional Vector Space Model (VSM) researchers have used the unique word that is contained in the document set as the candidate feature. Currently a latest trend which uses the phrase to be a more informative feature has considered; the issue that contributes in enhancing the document clustering accuracy and effectiveness. This paper presented a new technique for evaluating the similarity measure of the traditional VSM by considering the topic phrases of the document as the comprising terms for the VSM instead of the conventional term and applying the new technique to the Buckshot technique, which is a combination of the Hierarchical Agglomerative Clustering (HAC) technique and the K-means clustering method. Such a method may increase the effectiveness of the clustering by incrementing the evaluation metrics values.

Document clustering with the help of fuzzy c-mean algorithm is proposed by Thaung et. al. (2010). Most traditional clustering technique allocate each data to exactly single cluster, therefore creating a crisp separation of the provided data, but fuzzy clustering permits for degrees of membership, to which a data fit various clusters. In this paper, documents are partitioned with the help of Fuzzy C-Means (FCM) clustering technique. Fuzzy c-means clustering is one of famous unsupervised clustering methods. But fuzzy c-means method needs the user to mention the number of clusters and different values of clusters corresponds to different fuzzy partitions earlier itself. So the validation of clustering result is required. PBM index and F-measure are helpful in validating the cluster.

With the arrival of social networks and tagging systems, Internet has recently observed a big leap in the use of Web Recommendation Systems (WRS). According to users’ interest of items and their browsing history on the World Wide Web, these systems are able to forecast and recommend items and future purchases to users. They are being used now in several domains, like news article recommendation, product recommendation, and make-friend recommendation. WRS still has several drawbacks, like sparsity and the new user problem. Moreover, WRS does not make full use and harness of the power of domain knowledge and semantic web ontologies.
Mabroukeh et. al. (2011) discussed how an ontology-based WRS can use relations and concepts in ontology, along with user-provided tags, to provide top-n recommendations without the need for item clustering or user ratings. For this reason, the author also proposed a dimensionality reduction technique based on the domain ontology, to solve the sparsity problem. Most of Web Services exist without explicit associated semantic descriptions. Hence many services that are related to a specific user service request may not be regarded during service discovery.

Paliwal et al. (2011) proposed the problem of Web Service discovery given non-explicit service description semantics that match a specific service request. This proposed approach to semantic based Web Service discovery involves semantic based service categorization and semantic enhancement of the service request. The author proposes a solution for achieving functional level service categorization based on an ontology framework. Moreover, the author uses clustering for accurately classifying the Web Services depending on service functionality.

Semantic based categorization is performed offline at the Universal Description Discovery and Integration (UDDI). The semantic enhancement of the service request attains a better matching with relevant services. The service request enhancement involves expansion of additional terms (retrieved from ontology) that are deemed relevant for the requested functionality. A proficient matching of the enhanced service request with the retrieved service descriptions is obtained using Latent Semantic Indexing (LSI). The experimental results validate the effectiveness and feasibility of the proposed approach.

Association rules, used widely in the area of market basket analysis, can be applied to the analysis of expression data as well. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. An association rule has the form LHS→RHS, where LHS and RHS are disjoint sets of items, the RHS set being likely to occur whenever the LHS set occurs. Items in gene expression data can include genes that are highly expressed or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. the diagnosis of a tumor sample from which a profile was obtained). In this paper,
association rule mining techniques that have been recently developed and used for genomic data analysis have been reviewed and discussed. There has been a great explosion of genomic data in recent years. This is due to the advances in various high-throughput biotechnologies such as gene expression microarrays.

Exception rules have been previously defined as rules with low interest and high confidence. In this paper a new approach to mine exception rules will be proposed and evaluated. Interconnection between exception and negative association rules will be considered. Based on the knowledge about negative association rules in the database, the candidate exception rules will be generated. A novel exceptionality measure will be proposed to evaluate the candidate exception rules. The candidate exceptions with high exceptionality will form the final set of exception rules. Algorithms for mining exception rules will be developed and evaluated.

Data Mining is a process of discovering new, unexpected, valuable patterns from existing databases [2, 5]. Though data mining is the evolution of a field with a long history, the term itself was only introduced relatively recently, in the 1990s. Data mining is best described as the union of historical and recent developments in statistics, artificial intelligence, and machine learning. These techniques are then used together to study data and find previously hidden trends or patterns within. Data mining is finding increasing acceptance in science and business areas that need to analyze large amounts of data to discover trends, in which they could not otherwise find. Different applications may require different data mining techniques. The kinds of knowledge that could be discovered from a database are categorized into association rules mining, sequential patterns mining, classification and clustering [2].

Association rule is an implication of the form X=>Y, where X and Y are database itemsets. The example could be supermarket items purchased together frequently. Two measures have been developed to evaluate association rules, which are support and confidence. Association rules with high support and confidence are referred to as strong rules [1, 2, 3, 4]. Negative association rule is an implication of the form X=>~Y, ~X=>Y, ~X=>~Y, where X and Y are database itemsets, ~X, ~Y are negations
of database items. Negative association rules consider both presence and absence of items in the database record and mine for negative implications between database items.

In association rules mining only the rules with high support and high confidence are considered as interesting rules. The generated patterns represent the common trends in the databases and are valuable for the marketing campaigns. The rules with low support are just as valuable as they may contain unusual, unexpected knowledge about databases. Exception rules have been defined as rules with low support and high confidence [6]. A traditional example of exception rules is the rule Champagne => Caviar. The rule may not have high support but it has high confidence. The items are expensive so they are not frequent in the database, but they are always brought together so the rule has high confidence. Exception rules provide valuable knowledge about database patterns. In this paper a new approach to mine exception rules will be proposed and evaluated. An interconnection between exception and association rules will be considered.

Based on the knowledge about negative association rules in the database, the candidate exception rules will be generated. A novel exceptionality measure will be proposed to evaluate the candidate exception rules. The candidate exceptions with high exceptionality will be listed in the search algorithm output as exception rules. The proposed method for mining exceptions is highly valuable in fraud detection systems, security surveillance systems, fire prevention systems etc.

Sultana and M. Maroof Khan (1996) examine the problem of clustering m objects in c clusters. The objects are denoted by points in n-dimensional Euclidean space, and the main aim is to categorize these m points into c clusters in such a way that the distance between points within a cluster and its center is reduced. The problem is a complex optimization problem as it possesses many local minima. Various techniques have been proposed to solve this problem such as the k-means algorithm, the simulated annealing algorithm, the tab search algorithm, genetic algorithm etc. The authors examine these four techniques and compare their computational performance for the clustering problem. These techniques are tested on various clustering problems from the literature as well as several random problems.
Jain, A.K et al (1999) proposed that clustering is the unsupervised classification of patterns into groups or clusters. Because of the wide range of utility of clustering techniques in many disciplines; the problem of clustering has been widely addressed by several researches. The author proposed an outline of pattern clustering techniques from a statistical pattern recognition viewpoint, with the aim of providing useful information and references to basic concepts accessible to most of the users of clustering techniques. The authors present taxonomy of clustering approaches, and explains recent advances and also describe some important applications of clustering algorithms in various fields like data mining, image processing etc.

Likas et al (2003) proposed the global k-means clustering technique that creates initial centers by recursively dividing data space into disjoints subspaces using a k-d tree approach. The cutting hyper plane used in this approach is the plane that is perpendicular to the maximum variance axis derived by Principal Component Analysis (PCA). Shehroz S. Khan and Amir Ahmad (2004) describes that the performance of iterative clustering techniques which converges to various local minima is based mainly on initial cluster centers. Usually initial cluster centers are chosen randomly. The author proposed an algorithm to calculate initial cluster centers for K-means clustering. This approach depends on the observations that some of the patterns are very analogous to each other and so they have similar cluster membership irrespective to the choice of initial cluster centers.

Moreover, an individual attribute may give some information about initial cluster center. The initial cluster centers calculated using this approach are very close to the desired cluster centers, for iterative clustering algorithms. This process is feasible to clustering techniques for continuous data. The author showed the use of proposed approach to K-means clustering technique. The experimental results showed significant solutions using the proposed algorithm.
Agrawal et.al. (2005) describes about data mining applications and its various requirements on clustering techniques. The main requirements considered are,

- ability to identify clusters embedded in subspaces of high dimensional data,
- scalability,
- end-user comprehensibility of the results,
- non-presumption of any canonical data distribution,
- Insensitivity to the order of input records.

K-means is the simplest and the most popular clustering technique that is widely used in various fields of science and technology. But the main limitation of the k-means approach is that it generates empty clusters based on initial center vectors. This drawback does not cause significant problem for static execution of the k-means algorithm and the problem can be overcome by executing the k-means algorithm for a number of times. But when k-means algorithm is presented as an essential part of some higher level application, this empty cluster problem causes irregular behavior of the system and may lead to overall performance degradation.

Malay K. Pakhira et al.,( 2009) proposed a modified version of the k-means algorithm that effectively eradicates this empty cluster problem. From the experimental results, it is observed that, the proposed approach shows higher performance when compared with the conventional approach.

Guo-Yan Huang et. al. (2010) proposed an approach for clustering heterogeneous data streams with uncertainty called HU-Clustering. A Heterogeneous Uncertainty Clustering Feature (H-UCF) is proposed to illustrate the feature of heterogeneous data streams with uncertainty. Depending on H-UCF, a probability frequency histogram is proposed to track the statistics of categorical attributes; the approach initially creates n clusters by k-prototypes algorithm. A two phase streams clustering selection process is applied to HU-Clustering algorithm to improve the quality of clusters. Firstly, the candidate clustering is chosen through the new similarity measure; secondly, the most similar cluster for each new arriving tuple is chosen through clustering uncertainty in candidate clustering set.
Optimization (HPSO) data clustering is proposed by Alam et. al. (2010). The proposed algorithm uses the swarm intelligence of cooperating agents in a decentralized environment. The performance of the HPSO is evaluated on benchmark datasets and is compared with clustering techniques like K-means, PSO clustering.

Shin-Jye Lee et. al. (2010) presents a clustering-based approach to fuzzy system identification. In order to generate an effective initial fuzzy model, this paper tries to present a modular approach to identify fuzzy systems based on a hybrid clustering-based technique. Furthermore, the determination of the proper number of clusters and the suitable location of clusters are one of primary considerations on constructing an effective initial fuzzy model.

Because of the above mentioned reasons, a hybrid clustering approach concerning input, output, generalization and specialization has been introduced in this article. Moreover, the proposed clustering approach, three-part input-output clustering algorithm, incorporates a variety of clustering features simultaneously, including the advantages of input clustering, output clustering, flat clustering, and hierarchical clustering, to effectively perform the identification of clustering problem.

Li Taoying et al (2010) proposed an incremental clustering for categorical data using clustering ensemble. The author initially reduce redundant attributes if required, and then make use of true values of different attributes to form clustering memberships, and next use clustering ensemble to merge or divide clusters to gain optimal clustering. A novel partitioned based technique for dynamically grouping Web users based on their Web access patterns using Adaptive Resonance Theory Neural Network (ART1 NN) clustering algorithm is proposed by Raju et.al. (2011). The problem formulation and the proposed ART1 NN clustering methodology have been discussed.

There are several techniques available in literature that can be classified into two major categories namely similarity-based approach and model-based approach. These approaches are described in Pallav Roxy and Durga Toshniwal (2009). Similarity-based technique is a pair wise similarity document clustering technique. The main goal of this
technique is that the average similarities within clusters have to be maximized and the average similarities between the clusters should be minimized.

Dom.B, (2000) have being developed to improve the performance of the clustering techniques. K-means is one of the simplest unsupervised learning approaches that solve the well known clustering problem. The fundamental concept is to define k centroids, for each cluster.

These centroids should be positioned far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid, then need to re-compute k new centroids. Then a new binding has to be done between the same data set points and the nearest new centroid. A loop has been created. As a result of this loop it is observed that the k centroids alter their location step by step until no more changes are made. This algorithm aims at minimizing an objective function.

The study of the text clustering has attracted extensive consideration. An enhanced fuzzy clustering-text clustering approach is proposed by Jiabin Deng et. al. (2010) which depends on the fuzzy C-means clustering approach and the edit distance approach. This approach makes use of the feature estimation to decrease the dimensionality of high-dimensional text vector. Since the clustering output of the conventional fuzzy C-means clustering approach requires the stability, the authors establish the high-power sample point set, the field radius and weight. Because of the boundary value attribution of the conventional fuzzy C-means clustering approach, the authors suggest the edit distance approach. The experimental outputs confirm that the application of the enhanced approach to the text clustering produces more constant and precise.

An enhanced data clustering approach for mining web documents is proposed by (Odukoya et al., 2010) which formulates, simulates and assess the web documents with the intention of conserving their conceptual relationships and removing the difficulty of speed while increasing accuracy.
Barry de Ville et al. (2002) explained the data-mining classification and predictive modeling techniques that depends on bootstrapping techniques. This describes re-use of source data frequently that can offer a holographic view of the modeled data. Li haiying et. al. (2002) have discussed about real-time clustering techniques that consists of two steps.

The first step involves the extraction process in which partial parser and a shallow stemmer are invoked, it can be used for both linguistic and statistical methods, in order to minimize the term variations within the returned text snippets the system introduce normalization approaches. This avoids redundancy.

A novel approach called cluster-based fuzzy-genetic mining technique for extracting both fuzzy association rules and membership functions from quantitative transactions was proposed by (Chun-Hao Chen et al., 2006). The technique can dynamically adjust membership functions by genetic algorithms and uses them to fuzzify quantitative transactions. It can also accelerate the evaluation procedure and keep nearly the same quality of solutions by clustering chromosomes. Each chromosome represents a set of membership functions used in fuzzy mining. This approach first divides the chromosomes in a population into k clusters by using the k-means clustering approach. All the chromosomes in a cluster then use the number of large 1-itemsets derived from the representative chromosome in the cluster and their own suitability of membership functions to calculate the fitness values. The evaluation cost can thus be reduced due to the time-saving in finding 1-itemsets.

Crescenzi et. al. (2004) proposed an approach that automatically extracts data from large data-intensive Web sites. The “data grabber” investigates a large Web site and infers a scheme for it, describing it as a directed graph with nodes describing classes of structurally similar pages and arcs representing links between these pages. Miha Grcar et. al. (2008) proposed an approach concerning lack of software mining approaches, which means process of extracting knowledge out of source code. The author proposed a software mining task with a combination of text mining and link analysis technique.
Feature clustering is a dominant technique to minimize the dimensionality of feature vectors for text classification. (Jung-Yi Jiang et al., 2011) propose a fuzzy similarity-based self-constructing approach for feature clustering. The words in the feature vector of a document set are grouped into clusters, depending on similarity test. Words that are similar to each other are grouped into the same cluster. Each cluster is featured by a membership function with statistical mean and deviation. When all the words have been put in, a desired number of clusters are formed automatically. The authors then have one extracted feature for each cluster. The extracted feature, corresponding to a cluster, is a weighted combination of the words contained in the cluster. By this approach, the derived membership functions match closely with and illustrate appropriately the real distribution of the training data.

An ontology-learning framework named Link-Analysis and Text-mining toolbox (LATINO) is recently developed by Mladenic et. al. (2006). LATINO is an open source common purpose data mining platform providing text mining, link analysis, machine learning, and data visualization capabilities.

There are several techniques available in literature that can be classified into two major categories namely similarity-based approach and model-based approach. These approaches are described in Pallav Roxy and Durga Toshniwal (2009). Similarity-based technique is a pair wise similarity document clustering technique. The main goal of this technique is that the average similarities within clusters have to be maximized and the average similarities between the clusters should be minimized. Alternatively, Model-based approaches try to learn generative techniques from the documents, with each approach representing one particular document group. Several techniques are available in the literature for document clustering.

An enhanced fuzzy clustering-text clustering approach is proposed by Jiabin Deng et. al. (2010) which depends on the fuzzy C-means clustering approach and the edit distance approach. This approach makes use of the feature estimation to decrease the dimensionality of high-dimensional text vector. An enhanced data clustering approach for mining web documents is proposed by (Odukoya et al., 2010) which formulates, simulates and assess the web documents with the intention of conserving their conceptual relationships and removing the difficulty of speed while increasing accuracy.

Liu Jinling and Zhou Hong (2010) presents a well-organized Chinese text clustering technique depending on the semantic concepts. This technique, initiating from the text itself, by utilizing classified hierarchy Subject Word in Thesaurus of Modern Renchu Guan et. al. (2011) proposed a novel semi supervised text clustering approach called Seeds Affinity Propagation (SAP) based on an Affinity Propagation (AP) algorithm. Eisenhardt et. al. (2003) proposed an approach for document clustering using a distributed peer-to-peer network. The author used the $k$-means clustering approach, changed to work in a distributed P2P fashion using a probe and echo system. This is also used by Li and Morris (2002) but the problem is posed from the information retrieval perspective. Cluster signatures are generated from a subset of the document collection by centrally dividing it into clusters.

Datta et. al. 2006 proposed an exact local approach for monitoring a $k$-means clustering, in addition to an approximate local $k$-means clustering approach for P2P networks. The P2P $k$-means approach has its fundamental in a parallel implementation of $k$-means proposed by Dhillon and Modha (2000).

A novel text clustering approach based on frequent term sets for peer-to-peer networks is proposed by Qing He et. al., (2010). It needs relatively lower communication volume while attaining a clustering result whose quality is not affected by the size of the network. This paper proposed by Zhao-Kui Li and Yan Wang (2010), deals with the topology mismatch in structured P2P network, routing advanced approach using super-nodes and self-organizing clustering. This approach includes self-
organizing clustering depending on the physical location data in structured P2P system and each node judges if it can be a clustering header depending on its ability.

Zhongjun Deng (2010) examined clustering algorithm in Peer-to-Peer (P2P) network. Conventional clustering approaches cannot be applied P2P systems because of its lack of central control and very large size. A correlation-based clustering hierarchical P2P network approach is proposed by Yuan Li and Xia Shixiong (2010). This approach overcomes the drawbacks of poor scalability and low search efficiency in unstructured P2P networks. L-Fuzzy context was proposed by (Pollandt, 1996) as an effort to integrate fuzzy logic with FCA. The L-Fuzzy context utilizes linguistic variables, which are linguistic terms related with fuzzy sets, to characterize uncertainty in the context.

In Iceberg concept lattice Stumme et. al. (2002), association rules are used to cluster concepts on the lattice, and conceptual scaling Vogt and Wille (1995) is then used to generate the concept hierarchy.

Raghu Krishnapuram et. al. (2001) presented new techniques (Fuzzy C-Medoids FCMdd and Fuzzy C Trimmed Medoids or FCTMaZ) for fuzzy clustering of relational data. A technique to identify inter-transaction fuzzy association rules that can calculate the variations of events was proposed by (Yo-Ping Huang and Li-Jen Kao, 2005).

A complete comparison investigation of text document clustering for a biomedical digital library MEDLINE was proposed by (Llhoi Yoo and Xiaohua Hu, 2006). A concept about multiple-level taxonomy and multiple minimum supports to identify fuzzy association rules in a given quantitative transaction data set was proposed by Yeong-Chyi Lee et. al. (2006). A novel approach called cluster-based fuzzy-genetic mining technique for extracting both fuzzy association rules and membership functions from quantitative transactions was proposed by (Chun-Hao Chen et al., 2006).

Fuzzy ontology techniques using intuitionistic fuzzy set for knowledge sharing on the semantic web has been proposed by (Jun Zhai et al., 2008). Ontology is implemented as a standard for knowledge representation and sharing in collaborative design on the semantic Web.
Hongwel Chen et al., (2009) proposed a fuzzy trust problem domain for P2P-based system, and compare Fuzzy Comprehensive Matsumoto et al., (2010) proposes a prototype web search results clustering engine that improves search results by performing fuzzy clustering on web documents returned by traditional search engines, as well as ranking the results and labeling the resulting clusters.

Feature clustering is a dominant technique to minimize the dimensionality of feature vectors for text classification. (Jung-Yi Jiang et al., 2011) propose a fuzzy similarity-based self-constructing approach for feature clustering.

A novel hybrid genetic algorithm (GA) is proposed by (K. Krishna and Narasimha Murty, 1999). This technique detects a globally optimal separation of a given data into a described number of clusters.

Casillas et al. (2003) put forth a novel concept on document clustering using genetic algorithm. The authors present a genetic algorithm that deals with document clustering that computes an approximation of the optimum k value, and resolves the best clustering of the documents into k clusters. The author have experimented the proposed technique with sets of documents that are the output of a query in a search engine. The simulation results show that the proposed genetic algorithm attain better values of the fitness function than the well known Calinski and Harabasz, (1974) stopping rule and performs in only lesser time.

Hybrid Genetic Algorithm (HGA) proposed by Francisco Mota Filho and Fernando Gomide, (2005) uses the fuzzy c-means clustering technique during the fitness evaluation phase to decrease direct evaluations. J.M Pena et al., (1999) evaluates four initialization approaches for the K-Means algorithm namely random, Forgy, MacQueen and Kaufman. Even though this approach is popular for its robustness but the performance is mainly based on two key points which are initial clustering and instance order.

A semi-supervised clustering approach is developed by (Ayhan Demiriz et al., 1999) that integrates the advantages of both supervised and unsupervised learning.
processes to clear the classification problems. Shih-Sian Cheng et al. suggested a genetic algorithm (GA) for K-means clustering. The author shares his view that in case of using string of- group-numbers encoding, the prototypes of the clusters are encoded into the chromosomes.

Lu et. al. (2004) proposed fast genetic k-means cluster technique (FGKA). It is a quickest version of GKA and FGKA and has lot of development over GKA comprising an effective estimation of the objective value TWCV (Total Within-Cluster Variation) which reduces illegal string elimination overhead, and a generalization of the mutation operator. These enhancements resulted that FGKA runs 20 times quicker than GKA [58]

Haung et. al. (2005)

To overcome the difficulties of FGKA, (Lu et al., 2004) formulated an incremental genetic k-means approach (IGKA). GKMODE combines both the k-modes algorithm proposed by Chaturvedi et. al. (2001) and the genetic approach.

Ke-Zong Tang et al., (2010) formulated an improved genetic algorithm (IGA) depending on a novel selection approach is offered to manage nonlinear programming difficulties.

All the entities in selection method are characterized as a three-dimensional feature vector comprised of objective function value, the measure of constraints violations and the amount of constraints violations.

The new technique depending on Genetic algorithm was developed by Atul Kamble, (2010). This approach can be applied to all database includes data from a metric space. It is confirmed that the incremental approach provides better output than any other approach. A clustering approach depending on Genetic k-means model that performs better for data contains both numeric and categorical attributes was proposed by Dharmendra K Roy and Lokesh K Sharma, (2010). The author formulated a customized explanation of cluster center to avoid the numeric data is the major drawback of Genetic k-mean algorithm and offers an enhanced characterization of clusters.
An n-dimension convergence approach was developed by Min-Yuan Cheng et al. (2010) developed to follow the possible development in conventional genetic algorithm (GA) by K-means clustering method. Ghaemi et al., (2010) proposed a new approach that used by clustering ensembles which enhance cluster partitions fitness.

Jia Zhen and Wang Yong Gui, (2010) presents a genetic clustering approach depending on dynamic granularity. Chen Wei et al., (2010) proposed a technique which deals with the drawbacks of the genetic Fuzzy C Means (FCM) such as higher classification time and less clustering accuracy.

Chen Rui et al., (2010) proposes an enhanced genetic algorithm. This algorithm keeps the population diversity by similarity checks on the population before selection, and the approach resolves the early-maturing issue of the population evolution, and presents a formula for mutation probability connected with similarity rate and iteration times.

There are several approaches such as Natural Language Processing (NLP) combined with association rules Maedche and Staab, (2001), statistical model Faatz and Steinmetz, (2002), and clustering Bisson and Nedellec, (2000) that have been applied to create ontology from a concept hierarchy automatically or semi-automatically. Moreover, conceptual clustering techniques such as COBWEB are powerful clustering techniques that can be used for the generation of concept representations and relationships for ontology Bisson and Nedellec (2001). Andreas et. al. (2002) has made a discussion on the clustering technique for text data. Text clustering usually involves clustering in a high dimensional space that appears complex with considered to virtually all practical settings.

A document clustering method based on hierarchical algorithm with model clustering is presented by Haojun et. al. (2008). This paper involved in analyzing and making use of cluster overlapping technique to design cluster merging criterion.

By considering the difficulty that classical Euclidean distance metric cannot create an suitable separation for data lying in a manifold, a genetic algorithm based clustering method with the help of geodesic distance measure is proposed by Gang Li et. al. (2009). A Word sets based document clustering algorithm for large datasets was proposed by Sharma et. al. (2009).

Muflikhah et. al. (2009) proposed a document clustering technique using concept space and cosine similarity measurement. This paper aims to incorporate the information retrieval technique and document clustering technique as concept space approach. ELdesoky et. al. (2009) given a novel similarity measure for document clustering based on topic phrases.

2.2 PECULIARITY MINING WITH CLUSTERING

Peculiarity mining is one of the most important issues in data mining. The growing interest in the study of peculiarity mining has been credited to the availability of a large amount of individual behavioral data. Some objects containing common behavioral patterns in the dataset are dramatically different from other individual objects and show their peculiarities. It is very important for behavior analysis to mine these peculiar objects' groups as this has great potential in practice. However, to the best of our knowledge, it has not been explored before. In this thesis, we identify this interesting and practical problem of peculiarity mining: mining peculiarity groups and defining a measurement of the degree of peculiarity. As the first attempt to tackle the problem, we present a set-value-oriented day-by-day behavioral data expression mode considering that daily behaviors with respect to an object should be recorded as a set of behaviors, and devise a peculiarity group mining algorithm in view of the set-value-oriented data expression which cannot be very well handled by existing methods.
Furthermore, we show that our method is practical and efficient using real datasets. It is necessary to define and analyse the notion of peculiarity at both the attribute-value level and the record level. Attribute-value level peculiarity analysis deals with local peculiarity, while the record level peculiarity analysis concerns the global peculiarity which can be studied based on local peculiarity. The research is implemented and checked in single system, P2P and in a distributed system. According to the number of data records the percentage of the peculiarity increases. The peculiarity in this thesis is a DNA pattern which says the level of AIDS formation in a human body.

Ribeiro, Kaufman and Kerschberg,[1995] have described a way of multi-database mining by incorporating primary and foreign keys, as well as developing and processing knowledge segments[1]. Wrobel[1997], has extended the concept of foreign keys to include foreign links, since multi-database mining also involves accessing non-key attributes. Aronis et al. introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases spread across the network.

Liu, Lu and Yao [1998],have proposed an alternative multi-database mining technique that selects relevant databases and searches only the set of all relevant databases. Their work has focused on the first step in multi-database mining, which is the identification of databases that are most relevant to an application. A relevance measure was thus proposed to identify relevant databases for mining with an objective to find patterns or regularity within certain attributes. This can overcome the drawbacks that are the result of joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application.

Zhong et al.[1999] have proposed a way of mining peculiarity patterns from multiple statistical and transaction databases based on previous work. A peculiarity pattern is discovered from the peculiar data by searching the relevance among the peculiar data. A data item is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like an exception pattern from the viewpoint of describing a
relatively small number of objects, the peculiarity pattern represents a well-known fact with common sense, which is a feature of the general pattern. Wu and Zhang[2001] have advocated an approach for identifying patterns in multi-database by weighting.

Kargupta [2001], had built a collective mining technique for distributed data. Grossman have built a system, known as Papyrus, for distributed data mining. Existing parallel mining techniques can also be used to deal with multi-databases.

Patterns (rules) with low support have been considered by many researchers. Two examples are the studies of exception rules and emerging patterns. They share a common feature with peculiarity rules in the sense that all describe a relatively small number of objects. They differ in the way in which peculiar data are collected, interpreted, and used, as well as the interpretation of corresponding rules. An exception rule is an amendment to, or a clarification of, a more general rule. The peculiar data covered by an exception rule is obtained from the subset of data covered by the general rule. On the other hand, a peculiarity rule stands on its own and has a common-sense interpretation, like ordinary association rules.

Dong and Li proposed a framework for discovering emerging patterns. Their method is essentially a study of the change of supports in different data sets. A large change suggests an interesting emerging pattern. Since emerging patterns with large supports are perhaps well-known facts, they concentrated on emerging patterns with small supports. In particular, they suggested that patterns with low support, such as 1 to 20 percent, can give useful new insights about data. Unfortunately, such patterns are difficult to discover by traditional association rule mining methods. Their approach provides another use of peculiar data, which is different from our consideration of peculiarity rules. In general, it may be desirable to have a unified framework within which many different uses of peculiar data can be studied.

A notion related to peculiarity is noise, which is an unavoidable problem in real-world databases. Although noise may appear as peculiar data, one may identify noise based on domain knowledge or Meta knowledge of the database. In this paper, we concentrate on defined peculiar data, which are characterized by attribute values and the
distribution of values in a database. The task of differentiating actual peculiar data and noise is left to domain experts.

K-means is the simplest and the most popular clustering technique that is widely used in various fields of science and technology. The medical industry is also increasing with the data for AIDS patients. It is difficult for classifying and finding the DNA pattern of the AIDS documents. We use pattern matching and/or document clustering analysis in the research area of artificial intelligence and data mining. Its fundamental task is to utilize the alphabets to compute the percentage of related relationship between the records or the documents and to accomplish automatic classification without earlier knowledge. Document clustering is to utilize clustering technique to gather the documents of high resemblance collectively by computing the documents resemblance.

There are several pattern matching and clustering approaches available in the literature to fetch the relevant data, record or the document in distributed environment. But most of the existing mining techniques suffer from a wide range of limitations. The existing mining approaches face the issues like practical applicability, very less accuracy, scalability, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. Also, mining need to mine the peculiar data from the dataset. Whenever we use data mining techniques it gives the 80% relevant and 20% irrelevant data from the dataset, but there is no peculiarity. Here the speciality and the main objective of the thesis is bring the peculiar data from the dataset.

Peculiarity represents a new interpretation of interestingness, an important notion identified in data mining. Peculiarity may be hidden in a relatively small number of records. Peculiarity oriented mining focuses on some interesting data in order to find novel and interesting rules. Peculiarity rules are new surprising, interesting patterns which can be discovered from peculiar data. This can overcome the drawbacks that are the result of joining all databases into a single huge database upon which existing data mining techniques or tools are applied. The approach is effective in reducing search costs for a given application. Although it looks like an exception pattern from the viewpoint of describing a relatively small number of objects, the peculiarity pattern
represents a well-known fact with common sense, which is a feature of the general pattern.

Wu and Zhang (2001) have advocated an approach for identifying patterns in multi-database by weighting. Kargupta (2001), have built a collective mining technique for distributed data. Grossman have built a system, known as Papyrus, for distributed data mining.

Dong and Li proposed a framework for discovering emerging patterns. Their method is essentially a study of the change of supports in different data sets. Ribeiro, Kaufman and Kerschberg, (1995) have described a way of multi-database mining by incorporating primary and foreign keys, as well as developing and processing knowledge segments[1]. Wrobel (1997) , has extended the concept of foreign keys to include foreign links, since multi-database mining also involves accessing non-key attributes.

Aronis et al. introduced a system called WoRLD that uses spreading activation to enable inductive learning from multiple tables in multiple databases spread across the network. Liu et al proposed a relevance measure for identifying relevant databases as the first step for multi base mining. Ribeiro described a way of extending the INLEN system for multi database mining by the incorporation of primary and foreign keys as well as the development and processing of knowledge segment. Zhong proposed a way for peculiarity oriented multi database mining[3] .

The success of any algorithm for identifying peculiar data depends on the quality of the database. While multi database technology can support many multi database applications, it would be useful and necessary to mine these multi-databases to enable efficient utilization of the data. Thus the development of multi-database mining is a challenging task. A peculiarity rule is useful and meaningful and which gives some explanation or justification. In many cases, the explanation or justification of a peculiarity rule cannot be obtained from a single database, it can be obtained from other databases.
The aim of this research work is to find peculiar data in multiple relations for a given discovery task when foreign link relationships exist. Zhong et al. [1999] have proposed a way of mining peculiarity patterns from multiple statistical and transaction databases based on previous work. A peculiarity pattern is discovered from the peculiar data by searching the relevance among the peculiar data. A data item is peculiar if it represents a peculiar case described by a relatively small number of objects and is very different from other objects in a data set. Although it looks like an exception pattern from the viewpoint of describing a relatively small number of objects, the peculiarity pattern represents a well-known fact with common sense, which is a feature of the general pattern.

Wu and Zhang (2001) have advocated an approach for identifying patterns in multi-database by weighting. Kargupta (2001), have built a collective mining technique for distributed data. Grossman have built a system, known as Papyrus, for distributed data mining. Existing parallel mining techniques can also be used to deal with multi-databases.

They share a common feature with peculiarity rules in the sense that all describe a relatively small number of objects. They differ in the way in which peculiar data are collected, interpreted, and used, as well as the interpretation of corresponding rules. An exception rule is an amendment to, or a clarification of, a more general rule. The peculiar data covered by an exception rule is obtained from the subset of data covered by the general rule. On the other hand, a peculiarity rule stands on its own and has a common-sense interpretation, like ordinary association rules.

A notion related to peculiarity is noise, which is an unavoidable problem in real-world databases. Although noise may appear as peculiar data, one may identify noise based on domain knowledge or Meta knowledge of the database. In this paper, we concentrate on defined peculiar data, which are characterized by attribute values and the distribution of values in a database. The task of differentiating actual peculiar data and noise is left to domain experts.
There are several pattern matching and clustering approaches available in the literature to fetch the relevant data, record or the document in distributed environment. But most of the existing mining techniques suffer from a wide range of limitations. The existing mining approaches face the issues like practical applicability, very less accuracy, scalability, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. Also, mining need to mine the peculiar data from the dataset.

There are several pattern matching and clustering approaches available in the literature to fetch the relevant data, record or the document in distributed environment. But most of the existing mining techniques suffer from a wide range of limitations. The existing mining approaches face the issues like practical applicability, very less accuracy, scalability, more classification time etc. Thus a novel approach is needed for providing significant accuracy with less classification time. Also, mining need to mine the peculiar data from the dataset.

2.3 SOFTWARE AGENTS

Software agent is a software entity that functions continuously and autonomously in a particular environment, which may contain another agents and processes.

Shoham (1990) proposed a new programming paradigm called Agent-Oriented Programming (AOP) AOP is a specialization of Object-Oriented Programming paradigm and it allows programming of agents in terms of their mental states. This includes components like beliefs, decisions, capabilities and obligations. In this paradigm agent programs control agents and communications primitives such as informing, requesting and offering can be used in interaction Shoham also developed a prototype of agents may have numerous objectives and goals, and although they are individual and autonomous they still often have to collaborate and communicate with other agents in order to achieve their objectives.
Dogac and Singil state that in order to collaborate with others, agents are required to discover the existence, network addresses, capabilities and/or roles of other agents. Also communicate with other agents through an agent-independent, that is a standard agent communication language.

Maea, (1995), proposed Autonomous agents are computational systems, that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed.

Brenner et al, (1998), proposed intelligent software agents are defined as being a software program that can perform specific tasks for a user and possesses a degree of intelligence that permits it to perform parts of its task autonomously and to interact with its environment in a useful manner.

Jennings and Wooldridge (2002), proposed an agent-based system is one in which the key abstraction used is that of an agent. It contains any non-zero number of agents. These agents work together to solve a problem. This helps to visualize and develop complex systems by dividing the complex problem into smaller ones each of which can be solved using specialized agents.

Durfee et al, (1989), proposed a multi-agent system can be defined as a loosely coupled network of problem solvers, that work together to solve problems that are beyond the individual capabilities or knowledge of each problem solver. These problem solvers agents are autonomous and may be heterogeneous in nature. This is because the system may be constituted by agents developed by different people. The system is also open with respect to the agents as new agents can become part of the system and leave the system dynamically.

This has the advantage that multi-agent systems can solve problems that have not been envisage during system development unlike the distributed problem solving systems. The disadvantage is that due to the autonomous and heterogeneous nature of
the agents, it cannot be assumed that all agents are interested in achieving an overall goal in the most efficient manner.

Conflicting goals occur as every agent of a multi-agent system is primarily interested in achieving its own particular goal. Hence, coordination is an imminent aspect of MAS in order to resolve conflicts and work together in a coordinated manner.

Singh (2004) proposed an agent-based system development is about creating new techniques and methodologies that exploits the essential abstractions of agency in the software. This is because; the software to be designed has to deal with new concepts such as agent, goal, task, services, organization, interactions, environment etc. Also, the agent-based software is usually used to solve complex and distributed problems.

Donald A. Norman (1997) proposed his work on 'How might people Interact with agents'. Also specifies that agents are capable of taking over human tasks and interact with people in human-like ways perhaps with a form of natural language.

Laurel (1997) discussed about the characteristics of agents. The following four key characteristics of interface agents. When considered for an individual agent, these properties apply to its interface.

- Agency
- Responsiveness
- Competence
- Accessibility.

When an agent has to interact with users, it requires language ability to support for natural form of interaction. An agent has to interact with the user to obtain specification of the task it has to perform, to convey the result of the task execution to the user and to provide or obtain any suggestions to or from the user. In this research data mining agents, Interface agent, data agent and query agents are used to improve the classification accuracy of peculiar mining.
2.4 SUMMARY AND DISCUSSION

From the above reviewed literature, it is clear that no researchers carried out the research to improve the classification accuracy in peculiarity mining. The software agents are also not used in any of the existing approaches. This research work addresses the issues by improving the classification accuracy with the help of various types of software agents. This chapter has discussed about the literature survey on the clustering techniques, peculiar mining techniques and software agent’s. The next chapter proposes the methodology of the proposed work and discusses about mining peculiar data using agents.