METHODOLOGY

The proposed research methodology for the research is a mix of literature review, Internet surfing, and Data Mining Techniques algorithms to achieve the objectives of my research which stated earlier in objective document. Our research work basically covers following steps.

1. Data Collection

Indian stock market NSE and BSE are public entity where there are more than 6000 of companies are listed for the trading of companies share. For trading there are facilities of exchange between the buyers and sellers it may be a real or virtual. Exchanges provide the facilities of price discoveries. During the transaction the huge data are generated for the trades that take place in the stock exchange that data can be useful for my research work of relation finding between two stocks of Indian stock markets. The available data are secondary data.

2. Data Cleaning and Preprocessing

The collected data may contains the error or unwanted data which have no influence on the research work this type of data will be identified and converted into proper format on which the data mining process can be carried out with help of various tools available among which is best suited with my research problem. The preprocessing of data includes the conversion of data from one type of data format which is available from collection work into a desired format on which the analysis can perform. Mostly the stock data are available in the csv (comma separated value) file which may contains the data of the stock name, stock type, opening price, high price, low price, previous day closing price, quantity traded and trade date.

3. Data Clustering

Data clustering techniques are used to group up the relevant data sets. Hierarchical and partitioned clustering techniques are used for the clustering process. Clustering analysis segments a large set of data into subsets or clusters. Each cluster is a collection of data objects that are similar to one another within the same cluster but dissimilar to objects in other clusters. In other words, objects are clustered based on the principle of maximizing the intra-class similarity while minimizing the inter-class similarity. Sequential pattern and time-series mining example, clustering techniques can be used to identify stable dependencies for risk management and investment management. The clustering process is the complex task with high process time.

4. Applying Data Mining Algorithm

The data mining algorithm is the mechanism that creates a data mining model. To create a model, an algorithm first analyzes a set of data and looks for specific patterns and trends. The algorithm uses the results of this analysis to define the parameters of the mining model. The mining model that an algorithm creates can take various forms which include a set of rules that describe how stocks are correlated with other, a mathematical model that forecasts future trends of stock, a set of clusters that describe how the stocks in a dataset are related.
4.1. Classification of Algorithm

**Statistical Algorithms** Statistical analysis systems such as SAS and SPSS have been used by analysts to detect unusual patterns and explain patterns using statistical models such as linear models. Such systems have their place and will continue to be used.

**Neural Networks** Artificial neural networks mimic the pattern-finding capacity of the human brain and hence some researchers have suggested applying Neural Network algorithms to pattern-mapping. Neural networks have been applied successfully in a few applications that involve classification.

**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.

**Nearest neighbor method** A technique that classifies each record in a dataset based on a researcher, combination of the classes of the k record(s) most similar to it in a historical dataset. 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.

5. Analysis

Our research objective is to find the relation between stock of two different group or similar group, We have to apply the clustering analysis, regression analysis and associative rule analysis which are the techniques of data mining for huge data analysis as shown in figure below. The follow figure describes data analysis task and techniques of data mining. To perform analysis two important technical tools may be very help full to me during research work. One is Weka Data mining software and other is MATLAB.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

**MATLAB (matrix laboratory)** is a numerical computing environment and fourth-generation programming language. Developed by Math Works, MATLAB allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages, including C, C++, Java, and Fortran.
6. Testing

The analysis on the data is performed on the historical data and the outcomes of the research work must be tested on the current and future data then and then only the purpose of the research work will fulfill. The association of data will tested on the current and future transaction of the BSE and NSE listed stocks on daily transaction with the experiment in which the positive association and negative association, which are outcomes of the proposed research analysis are checked with the bear and bull environment of the stock market to make the research work genuine.

WORK PLAN

1<sup>st</sup> 6 month : Collection of research paper and review it, understanding the research subject and completion of synopsis.

2<sup>nd</sup> 6 month : Data collection and Preprocessing (cleaning, storing).

3<sup>rd</sup> 6 month : Data clustering and study analysis.

4<sup>th</sup> 6 month : Summary and Final draft of thesis.