Applications of Data Mining in Correlating Stock Data and Building Recommender Systems

Sharang Bhat
Student, Department of Computer Engineering
Mukesh Patel School of Technology Management and Engineering,
NMIMS (Mumbai) University
Mumbai, 400056

ABSTRACT
Since the 2008 sub-prime mortgage crisis and the global economic meltdown of 2009, investors have understood the value of historical data affecting a stock during investment. The general skeptical tendency of an investor increased in cases of mortgage bonds and general equities. Since then, the investment banks and financial institutions have increased their reliability on the available stock market data and finding correlations from the history of a stock to determine future trends and analyze current trends before making a decisions on the action to be performed on a particular stock. Using data mining techniques like clustering and association rules has helped quantitative analysts to determine correlations between the trend a stock follows and the reaction of a customer to the change in trends. This proves to be crucial in making recommendations to a customer on a particular financial product. The proposed research paper and eventual system, aims at providing a platform for data entry of large data sets of stock data, a set of input variables (which can be two or more) upon which various clustering and associative algorithmic techniques are implemented, to produce a result, similar to the result of a recommendation system, which predicts the next possible choice of a user based on historical ratings. The difference being that the ratings will be replaced by certain data points which can be system or user fed, and a correlation will be drawn between two or more variables with respect to these specific data points. The following paper lists the principles to be followed, the architecture of the system and the algorithmic techniques to be implemented.

General Terms
Data Mining Applications in chart pattern recognition and building recommender systems.

Keywords
Mortgage crisis, stock market data, correlations, clustering, recommendation system, data mining, chart pattern recognition, classification, large data sets,

1. INTRODUCTION
Data mining is the process of discovering hidden patterns, trends and knowledge from large data stores which can't generally be found by simple analysis of databases. Data mining uncovers unknown patterns, and new rules from large databases between objects that are potentially useful in making crucial business decisions.

It applies data analysis and knowledge discovery techniques under acceptable computational efficiency limitations, and produces a particular enumeration of patterns over the data. Existing systems such as Trade Station and VTT Forex trader use the same principles to provide insight into trends by using external servers to store data, and various algorithms to carry out data mining techniques.

Spatial data mining mainly aims at discovering patterns observed in images, graphs and any graphical representation of a real-world situation. Such data is known as spatial data and is obtained from satellite images, medical equipment and stock graphs. Satellites especially, produce about 2-3 Terabytes of data every hour. This data is largely composed of images. Therefore it becomes difficult to examine such data in isolation and in detail. Spatial data mining aims at extracting meaningful information from such data with an aim to automate a cumbersome process. The general issue encountered in such situations is that the user or expert must provide a hierarchy along with the data at hand. Which means, the user must provide the algorithms with spatial concept hierarchies. Thus, the quality of results generally obtained from such systems relies a lot on the hierarchy provided to the given data.

2. PRINCIPLES AND METHODS
Based on the type of knowledge that is mined, data mining can be mainly classified into the following categories [2]

1) ***Classification and prediction*** is the process of identifying a set of common features in a dataset that can be used to demarcate, describe and distinguish data from the dataset into classes or concepts. Classification generally consists of a set of target classes, attributes of data which are predictors and the data of each object which is the case. In an example to analyse credit risk in case of bank classifying loan applicants as low, medium or high risk(target classes). This model would use historical data of a customer associated with the bank along with data such as employment history, ownership or rental of property, years of residence and number and type of investments. These are used as predictors and the data associated with the customer (object) constitutes a case.

2) ***Clustering analysis*** recognizes different classes, distributes and segments data in 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 on the basis of certain common characteristics that each item possesses. These attributes are obtained from the data warehouse which is modelled from one or more databases. In other words, objects are clustered based on the principle of maximizing the intra-class similarity while minimizing the inter-class similarity. For example, clustering techniques can be used to identify dependencies of various financial products such loans and bonds to various indicators in the economy, and hence a trend can be derived from the dataset, helping the user/developer to predict future trends based on historical data of both the products. Like classification, clustering segments data into various groups. Unlike classification, clustering segments data into groups that were not previously known i.e

clustering does not use targets and the data is segmented as per inherent similarities.

3) ***Association rule mining*** discovers interesting correlation patterns among a large set of data items by showing frequently occurring attribute-value conditions. A typical example is market basket analysis, which is employed by retail stores such as Walmart, e-retail stores such as Amazon, Zulily and Proctor and Gamble, which analyses buying habits of customers by finding associations between different items the customers choose to purchase, to the extent of when they generally purchase the item and can help determine which promotional and discount offers will gather large volumes of sale and when.

3. SYSTEM ARCHITECTURE

This research paper deals primarily with the architecture of such a system, which integrates the use of the k means algorithm and association rules, to provide maximum support and confidence during analytics and decision making in the stock market.

Support and confidence are associated with the association rules used in the clustering and classification of data. These rules are similar to a general If-Then-Else statement in which an If statement triggers a result clause (Then).These include support and confidence.

In ARs, the threshold support and confidence are critical to verify the rule whether the rule is valid. Support indicates the percentages of records containing an item or combination of items to the total number of records. Confidence reflects how sure when the "If" part is true that the "Then" part is also true under a particular condition.

For a developer of such a system, it is important to understand the needs of an investor to use this system to devise efficient asset allocations. First, data must be collected from relevant databases and must further be qualitatively and quantitatively organised to build an efficient data warehouse for further transaction. One year of historical data is required for understanding which investment products the client might be interested in.

1. Data Selection and Pre-processing Module

This module is used to collect germane data and organize it in a suitable manner and format for mining processes and other data analytics to be carried out upon the dataset. This module aims at building a centralized data warehouse for analytics to be carried out using clustering and association rules. The data selection process aims at gathering data from sources which are heterogeneous or similar in characteristics such as client databases, customer credit files and historical data from banks containing information about customers classified as per their income and the investment products they choose as per their income. Once this data is collected, the pre-processing module is used to organize and remove redundant data from the collection and also correct any erroneous occurrences. The data transformation can enhance the capability of reading different data. The data transformation can involve different types of normalization.

2. Clustering Module

The clustering module is primarily used to reduce the processing time for the rules discovery module, as it not only increases the efficiency of performance but also makes the rules that the dataset follows easier to find. The K-means algorithm is used to partition the data into groups based on their inherent similarities.

3. Rules Discovery Module[5]

In this stage, the RDM aims at discovering the relationships in a specific cluster. The RDM can directly extract an input data set from the CM to generate useful rules. In this module, the Apriori algorithm is applied to find the frequent patterns, correlations and associations. Such rules can indicate which groups or sets of items customers are likely to purchase in a given set of clusters. After the generation of the rules, the rules will allow management to make an evaluation. Then, the sales and marketing department can use such rules for decision making in regard to a specific cluster. The following figure [1] represents the architecture of the proposed system. The system architecture is represented diagrammatically in Figure-1.

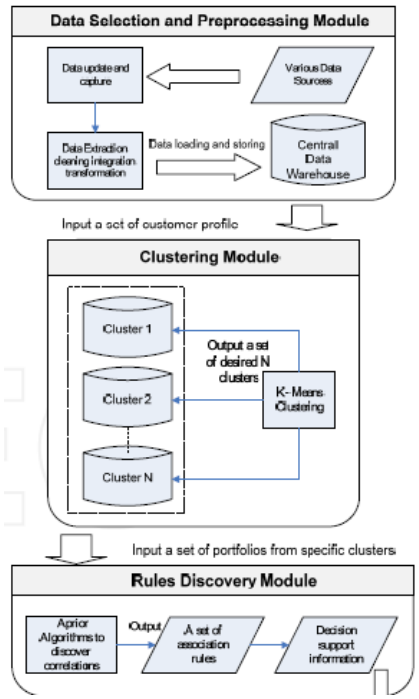

Figure 1

4. ALGORITHMS USED

K-means [3] is a simple machine learning algorithm that organizes a given dataset into meaningful clusters based on certain inherent similarities. This nonhierarchical method initially takes the number of components of the population equal to the final required number of clusters. The procedure follows a simple and easy way to categorize data into various classifications or clusters as per their similarity with respect to certain characteristics. The primary idea is to define k centroids with one for each cluster. These centroids should be placed in a carefully as placement in a different location generates a different classification. So, the better choice is to place the centroids at a distance sufficient enough so as to prevent erroneous results. The next step is to take each data point from a given data set and correlate it to the closest centroid. When no data point is remaining, an early group age is said to be done. At this point, t new centroid are generated as the centers of the newly generated subsets with each having

a centroid, k (i). After we have these t new centroids, a new correlation has to be done between the same data set points and the nearest new centroid. A nested loop is thus formed. The result is obtained until no further divisions of the subset can be made. This algorithm, in all its functioning, as a basis, aims at minimizing an *objective function*, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{t} \sum_{i=1}^{k} |X_j - C_j|^2$$

Where

$|X_i - C_j|^2$ is the chosen Minkowski distance between a data point X_i and the cluster center C_j

The K-means algorithm has a major disadvantage when being used to determine cluster formation in dynamic datasets and datasets containing erratic values. This is because the representative of the cluster is determined as the mean, which may or may not be representative of the cluster, as a mean tends to be affected by presence of values of greater magnitude.

The algorithm consists of the following steps

1. Place K points in the space containing the objects to be clustered. These are the first centroids obtained. Mark them as group 1.
2. Check each object and compare it with the centroids. Depending on the closeness of the object to the centroid (using a distance measure), assign that object to the group.
3. When the list of objects is exhausted, recalculate the K positions centroids.
4. Repeat steps 2 and 3 until the centroids remain constant. This results in separation of objects based on the objective function to be optimized.

Partition around medoids
PAM [2] is a k-medoid based algorithm.

The *k*-medoids [4] is a medoidshift clustering algorithm similar to the *k*-means algorithm. The *k*-medoids algorithms uses partitioning (breaking the dataset up into groups) and attempts to minimize the objective function which in this case is the distance measure between the data points in the cluster and the data points and the centroid chosen. Each cluster is known as a priori.

Each priori has a representative object, which is at first arbitrarily selected. First, the cost of each object with respect to the medoid/representative is calculated, and the smallest objects having the smallest distances are tabulated.

The clustering implementation is shown using a Dendogram (implemented in R), and shows the clustering of a random array of objects being clustered using PAM. (Figure 2 and Figure 3)

The algorithm for PAM is as follows

1. Randomly select k data points as medoids.
2. Assign each data point to the closest centroid. The closest centroid is one having minimum distance (using a distance measure) from the object or data point.
3. For each medoid *m*
    1. for each non-medoid o
        1. Swap m and o and compute total cost of the new composition
4. Select the composition with the lowest cost.
5. Repeat steps 2 to 4 until there is no movement of the medoid.

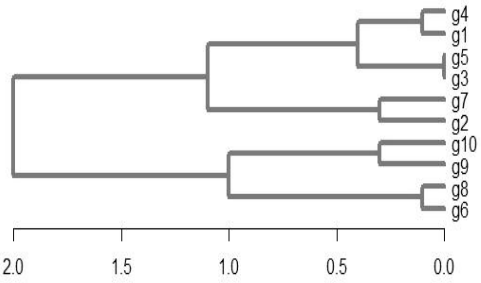

Figure 2

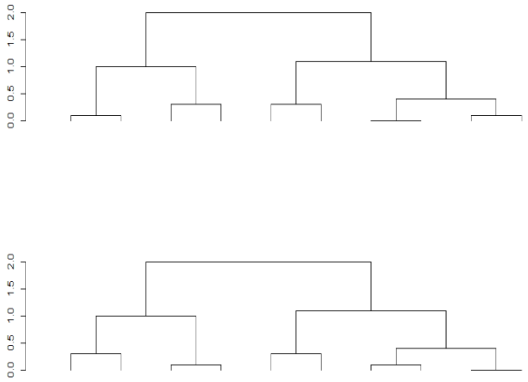

Figure 3

The quality of the cluster is obtained as a measure of the average distance between the medoid and the objects of the cluster.

The disadvantage with the PAM algorithm is that it cannot operate efficiently on large datasets, as it tends to be comprehensive with a brute force method, and takes O(n^2) time. When the dataset is linear and relatively small, PAM is an efficient and simple algorithm.

Figure-5 [2] is depicts the comparison of the running times of the PAM and CLARA algorithms on a dataset consisting of 10000 data entries. (Tested)

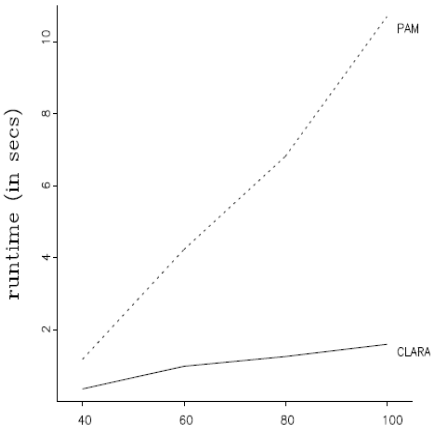

Figure 4

CLARA (Clustering Large Applications)
CLARA [2] is meant specifically for large datasets, and uses the concept of sampling to find an approximate mean or representative for the cluster. It is a continuation of the PAM algorithm. Instead of finding representatives for the entire dataset, it applies the PAM method on the sample and computes the medoids.

The point is that if the sample is drawn in a sufficiently random way, the medoids of the sample would approximate the medoids of the entire population. For better results, CLARA draws up multiple samples and PAM of those samples is found to further provide a concrete result. Experiments have reported that, 5 samples of size 40 +2k each, results were satisfactory. Figure 5 shows the formation of clusters in a dataset of the form 40+2k using the CLARA algorithm. Other methods such as QT clustering forms clusters based on a maximum clustering diameter by identifying the largest cluster below a certain specified threshold and removes its items from the data set until all items are assigned. (Figure 6)

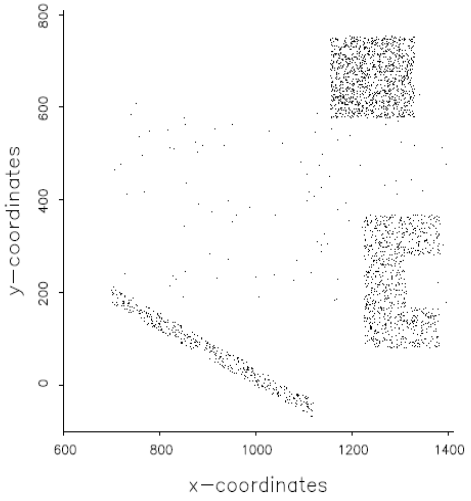

Figure 5

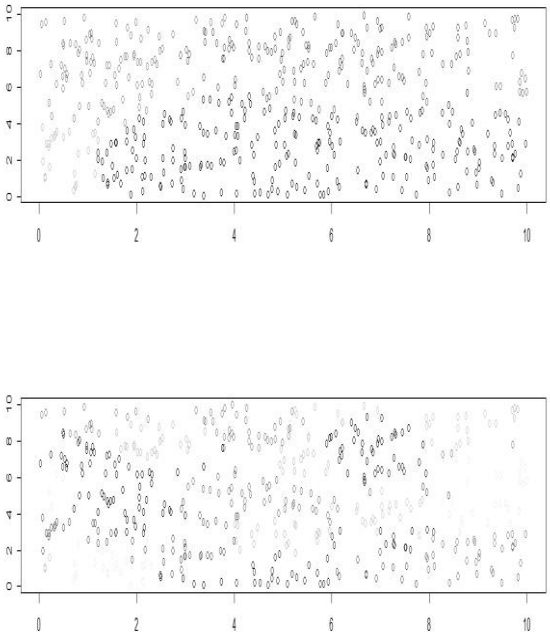

Figure 6

Choosing between algorithms [6]
The choice between using the k-means [3] and k-medoids [4] depends upon the kind of data in the dataset and the size of the dataset, and the data model [6] followed by the database from which the data warehouse is made.

If the data has inherent similarities and the clustering has already occurred in the database before using it in the data warehouse, the k-means algorithm should be used, as it only acts as a last filter before passing the modified data to the rules discovery module.

If the dataset inherently varies or is unrelated, is large and may contain redundancies, then the k-medoid algorithm proves to be more efficient as the equal distances (redundant data) are not carried further in calculation of the next medoid. This however, is an issue while using the k-means algorithm because the while calculating the mean or centre of the dataset, addition of identical values may result in an erroneous output, but this is not reflected in the deviation of the mean from any item k in the dataset. However, calculation of the mean itself is O (N^2) operation, and hence increasing the size of the dataset leads to increased computational time.

In the proposed system, the k-medoids clustering technique is used in the form of the PAM algorithm [2]

A tabular comparison of K-means and K-medoid algorithms is given in Table 1.

Table 1

Parameter	**K-Means Clustering**	**K-Medoids Clustering**
Running Time	*O*(i k n)	O(ik(n-k) ^{2})
Predefined Constants	Requires number of clusters to be defined beforehand	K is first randomly chosen, increases as number of medoids grow
Similarity Measure	Variables to be correlated have to be defined previously	Can be used with any similarity measure
Dependence on Hierarchy	May fail to converge - it must only be used with distances that are consistent with the mean	Distances do not vary with respect to the centroid as the Medoid is first randomly chosen ,and is changed until average distance of items from medoid is smallest possible
Response to Outliers	Unable to handle noisy data and outliers as one large value may shift the centroid	A medoid is more robust to outliers than a mean as magnitude of values do not affect selection of the cluster representative

5. ASSOCIATION RULE MINING

Association rule mining [5] is a classic data mining techniques which is used to highlight patterns in a given dataset.. Association rules are formed by analyzing a given dataset for frequent if/then patterns and using the criteria *support* and *confidence* to identify the most important relationships. *Support* is an indication of how frequently the items appear in the database. *Confidence* indicates the number of times the if/then statements have been found to be true.

Apriori [7] is an algorithm for mining frequently occurring items in a dataset. It follows a simple divide and conquer principle by first dividing the dataset into a number of smaller subsets and finding association rules amongst them. Then, these subsets are combined with other subsets and the rules are applied on the resultant subsets. This process is repeated until the association rules are formed for the entire data set. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis. The Apriori first finds all combinations of patterns or items that occur with certain minimum frequency (support) and then calculates rules that express the co-occurrence of the same rules on a larger part of the dataset. The algorithm is not used to find outliers in a given dataset as it finds the most relevant and possible occurrences. The rules of technical analysis in stock trading state that price always moves in trends, and trends observed in the past are bound to occur at some point depending on how the next trading cycle is determined. Given the high occurrence of similar trends in a daily stock graph, the Apriori algorithm can find frequently occurring prices at a given point of time, which can further help an investor in determining when to buy, sell or hold a stock. The following flowchart depicts the flow of the algorithm.

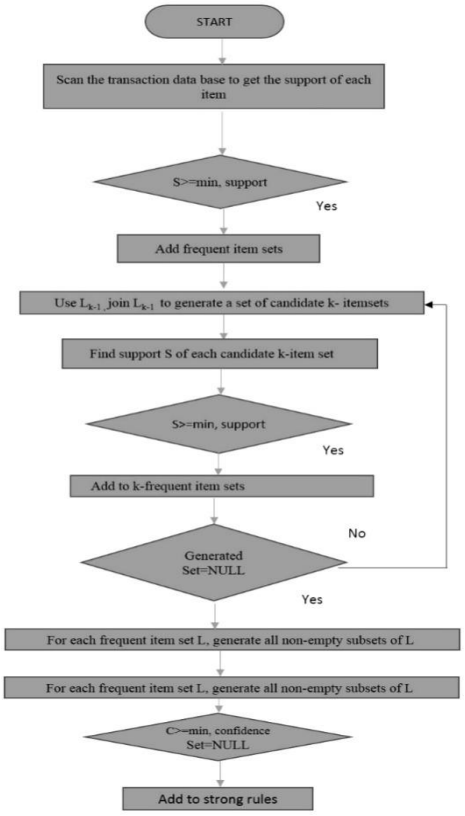

Figure 7

Figure 8 depicts a set of 10000 random transactions on a database of customers in a bank. (Implemented using R, AdultUCI dataset). The x axis denotes the items i.e customers and the y axis depicts the frequency of transactions on each object.

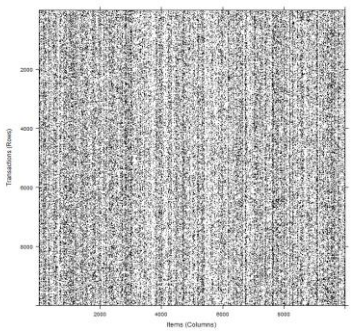

Figure 8

Figure 9 depicts the scatter graph of the association rules found in the dataset along with the support and confidence levels. Figure 10 shows the two-key plot for the same.

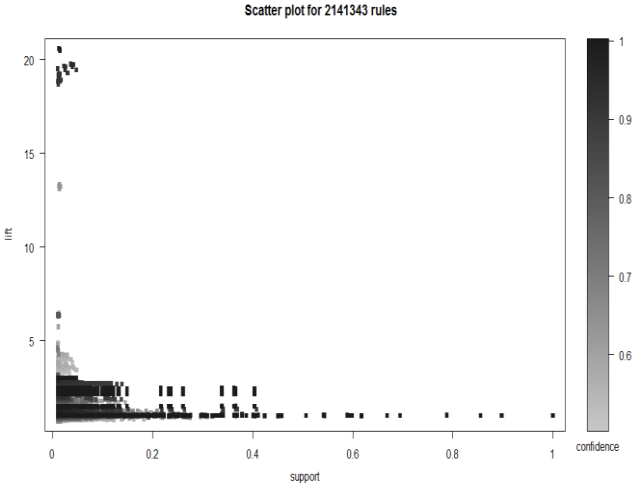

Figure 9

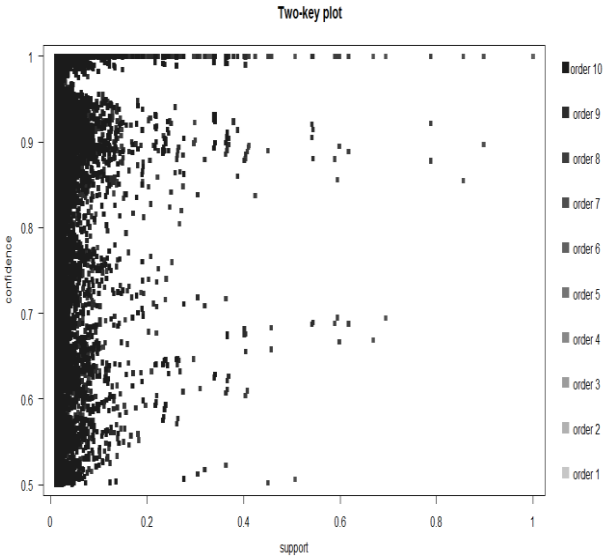

Figure 10

6. CONCLUSION AND FUTURE SCOPE

Advantages of the proposed system

1. The system can be used to find a correlation between the change in trends of certain financial products and the behaviour of consumers towards these changes.
2. The system can be implemented over any web server and can be run as a distributed system over a number of clients and hence can run on commodity hardware.
3. The algorithms used, namely, K-means/Medoid and Apriori follow linear time and logarithmic time respectively.
4. The aforementioned algorithms are able to determine a cluster of similar objects from a large dataset.
5. Using the k-medoid method, the two algorithms do not depend on the order or hierarchy of the data.
6. As the methods for calculating distances is generally using a Minkowski distance such as Euclidian of Manhattan distance, the method is simple and yields efficient results.

Inferences

- A meaningful correlation can be found between two distinct datasets
- The correct algorithm for the given problem depends upon the kind and the size of data
- Abstraction is key to correct trend prediction, especially in large data sets.
- The system can be used to provide a correlation between two variables which can provide deeper insights into a large dataset.
- A trend can be detected by forming correlations between various item sets in the data warehouse.
- The warehouse can be created by selecting relevant data, and this process of selection can be carried out using the pre-processing module of the financial data mining system.

With the increase of economic globalization and the need to move away from speculation of trends and tendency to 'gamble' with the causality of change in pattern and trends in the stock market, the need to analyze stock market data to provide accurate and dependable trend prediction and chart pattern recognition has emerged. With two thirds of the world's stock markets and economies running purely on algorithms to preclude market corrections and change in trends, data mining applications in the stock market have started to grow. With the decrease in cost of storage and the increase in processing power, coupled with the development of efficient algorithms to analyze large data sets, data mining can now successfully 'crunch' or collaborate variables from the dataset to provide trends. This has led to the development of a concept called big data which does not consider the causality of an occurrence, but only detects the occurrence by attempting to find correlations between the provided variables in the data set and then use this history of the dataset to predict future occurrences of the trend. Using the aforementioned methods, clustering can be obtained from a dataset having two or more variables. Using two variables is a straightforward method of feeding in the values of x and y and applying the algorithms. However, when more than two variables are used, they are paired with one another, and the PAM and CLARA algorithms are used on them. The result of these algorithms is the medoid of the datasets. This results in the formation of another dataset which is consists of medoids of the primary dataset, on which the two algorithms can be used, with lesser number of variables. Thus, the algorithm can be used recursively on data objects with more than two variables. Such a system is based on systems such as the Google Flu Trends [8] and Fare cast which crunch historical data to predict future trends.

The stock market has been earlier claimed to be as an example of the random walk hypotheses[10] , which states , on the basis of an experiment, that the stock market is as unpredictable as flipping a coin. However, as more comprehensive test results conducted by Professors Andrew W. Lo and Archie Craig MacKinlay in their publication, a Non Random Walk Down Wall Street[11], it was proved through a null hypothesis test, that the stock market was in fact, predictable by considering the historical data associated with an index or a security.

Automated computer programs are developed using data mining and predictive technologies for the trade markets. Here the historic data is the key and is used to develop prediction models. Using data mining techniques correlate stock data helps investors to understand hidden patterns from the historic data and in turn enables them to do a prediction or forecast of their future investments/trade. Data analysis and record of various global events are usually used to develop the prediction model which supports numerically and graphically.

The proposed system is designed to run on commodity hardware and requires a platform which supports object oriented principles and data structures like hash tables and indices. The data must be fed to the system via a warehouse which can be built using the same platform, but must gather data from a database, which generally runs a DBMS (MS-Access).

Hadoop and Mongo can be further used to connect the system across a network which runs similar processes to gather accurate real time results.

Furthermore, the proposed algorithms allow similar stocks to be classified as per their price movement, clustered as per similarity and trends followed. The Rules Discovery Module forms associations between the change in prices and the time period. The highlight of the system is that the price can be correlated with any independent variable which may or may not be directly related or affecting the stock price. Future work will also involve presenting the analysis as a derivative or insight derived from the values generated by the modules of the system.

For over half a century, financial experts have considered the movements of markets as a random walk. With Big Data and number crunching now removing causality from a system and ending the dependence on theories that attempt to understand change in the market, such systems aim at increasing the data points to answer a different question i.e. What rather than Why.

7. REFERENCES

[1] Mak, George T.S. Ho and S.L. Ting, 'A Financial Data Mining Model for Extracting Customer Behaviour' ,Mark K.Y, Convoy Financial Services Holdings Limited, Hong Kong, Department of Industrial and

Systems Engineering, The Hong Kong Polytechnic University, China, 23rd October 2011.

[2] Raymond T.Ng, Department of Computer Science 'Effective and Efficient Clustering methods for spatial databases',5th October 2012, IEEEXplore publication, University of British Columbia

[3] K-MeansAlgorithm, http://www.math.le.ac.uk/people/ag153/homepage/KmeansKmedoids/Kmeans_Kmedoids.html

[4] K-Medoid Algorithm, http://swg-matlab.blogspot.com/2010/01/k-medoid-algorithm.html

[5] Association Rules and Mapping,http://www.computerscijournal.org/dnload/Mehzabin-Shaikh-and-Gyankamal-J-Chhajed/OJCSV05I02P263-267.pdf

[6] Data Model Concepts http://shodh.inflibnet.ac.in/jspui/bitstream/123456789/1617/5/05_methodology.pdf

[7] Apriori Algorithm, http://www.codeproject.com/Articles/70371/Apriori-Algorithm

[8] Detecting influenza epidemics using search engine query data-Google Research, publication in *Nature*, 19th February 2009',Jeremy Ginsberg, Matthew H. Mohebbi,

[9] Dongsong Zhang and Lina Zhou ,IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, VOL. 34, NO. 4, NOVEMBER 2004

[10] 'Discovering Golden Nuggets: Data Mining in Financial Application'

[11] 'An Analysis of the Random Walk Hypothesis based on Stock Prices, Dividends, and Earnings', Risa Kavalerchik, Peter Rousseau

[12] 'A Non-Random Walk Down Wall Street'-Andrew W. Lo & A. Craig MacKinlay