Intra and Inter Sector Stock Price Forecasting using Association Rule Mining (IISARM)

Mr. Vishal Adheli
P. G. Scholar
N. K. Orchid College of Engineering and Technology,
Solapur, India.

Mr. Vipul Bag
Associate Professor
N. K. Orchid College of Engineering and Technology,
Solapur, India.

***Abstract*—The paper emphasizes on forecasting stock prices among the same sector known as intra sector and also among different sector known as inter sector and provides best rules using Intra and Inter Sector Association Rule Mining (IISARM) algorithm. Association rules are well suited for predicting stock market behavior. Our approach is to develop and test the algorithm that predicts the impact of company stock prices in intra and inter sector. Our approach is also to measure the impact in percentage. The object of this algorithm was to find hidden rules within sector having strong and relative association between them.**

***Keywords*— IISARM; Apriori; FP-Growth; Stock Data.**

I. INTRODUCTION

Data mining also known as knowledge discovery in large database refers to the extraction of useful implicit knowledge from large data set. It is the process of analyzing data from different source and perspective and summarizing it into useful information. It includes the process of finding patterns which can be useful for business activities.

Data mining techniques can be applied to stock market. Stock prices are very dynamic and highly fluctuating area. The investors have to take quick decision and must be very careful while investing. Sometimes his decision may take him to a heavy loss so making him to take right decision is a big challenge, so here in this research the aim would be dealing with generating promising results to ensure investor to take right decision while investing.

One of the useful techniques of data mining is Association Rule Mining which is used for predicting and analyzing behavior in the respective field. Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. It gives the relation between the two or more variables.

We have collected historical stock prices (data) of top companies in sectors like Banking, IT, Reality, Auto FMCG, Oil and Gas from Bombay Stock Exchange (BSE) and have successfully tested the outcome with intra and inter sector. We have tested on ARM Tools developed by various researchers like P. Bloomfield[4], J. Han[7],[16] and M. Hall[17] for finding frequent items and the relation between them.

II. RELATED WORK

Prediction in stock market is always a hot topic for practicing data mining technique. Many researchers have worked on predicting the closest, best and strong rule. Neural Network, Genetic Algorithm, Association, Decision Tree and Fuzzy systems are widely used to predict stock prices. Among all the above techniques, association mining is the simple and straightforward technique to find the relation between the items.

In the work carried out by P. Paranjape [14], have shown that the association rule mining with a support confidence framework can be used to build a stock market portfolio recommender system and their approach demonstrates the application of soft computing techniques like ARM and fuzzy classification in the design of an efficient recommender system.

Another related work done by You-Min Ha [13], the aim was devising an efficient and flexible approach that recommends appropriate investment types to stock investors by discovering useful rules from past changing patterns of stock prices stored in a database.

The work reported by using uncertain data [15]. They have worked on apriori algorithm and modified it so as to suite the uncertain dataset and named it U-Apriori algorithm. They identified the computational problem of U-Apriori and proposed a data trimming framework to address this issue. They proposed the LGS-Trimming technique under the framework and verified, by extensive experiments, that it achieves very high performance gain in terms of both computational cost and I/O cost. Unlike U-Apriori, LGS-Trimming works well on datasets with high percentage of low probability items. In some of the experiments, LGS-Trimming achieves over 90% CPU cost saving in the second iteration of the mining process, which is the computational bottleneck of the U-Apriori algorithm.

III. BACKGROUND

A. *Apriori Algorithm*

Developed by Agarwal and Srikant 1994 [1] is an innovative way to find association rules on large scale, allowing implication outcomes that consist of more than one item, based on minimum support threshold. Apriori is designed to operate on databases containing transactions (e.g., collections of items bought by customers). Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time a step known as candidate generation, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

B. *Association Rule Mining*

Association rule mining [1] is a technique for discovering hidden data dependencies and is one of the best known data mining techniques. The basic idea is to identify from a given database, consisting of item-sets whether the occurrence of specific items, implies also the occurrence of other items with a relatively high probability. Association rules have to satisfy constraints on measures of significance and interestingness.

```
Apriori Algorithm
procedure Apriori(T, minSupp) {
        $L_1$, {frequent items};
        for(k=0; $L_{k-1}$ != 0; k++) {
                $C_k$ = candidates generated from $L_{k-1}$
                for each transaction t in database do{
                        #increment the count of all candidates in $C_k$ that are contained in t
                        $L_k$ = candidates in $C_k$ with minSupp
                }
        }
        return $U_k L_k$;
}
```

The two significant basic measures of association rules are support(s) and confidence(c). Support(s) is defined as the proportion of records that contain X ∪ Y to the overall records in the database.

$$Support\ (XY) = \frac{\text{Support sum of XY}}{\text{Overall records in the database D}} \qquad (1)$$

Confidence(c) is defined as the proportion of the number of transactions that contain X ∪ Y to the overall records that contain X, where, if the ratio outperforms the threshold of confidence, an association rule X ➔ Y can be generated.

$$Confidence\ (X/Y) = \frac{\text{Support (XY)}}{\text{Support (X)}} \qquad (2)$$

C. *FP-Growth Algorithm*

The FP-growth algorithm is currently one of the fastest methods to frequent item set mining. In addition, projected FP-trees are (optionally) pruned by removing items that have become infrequent due to the projection (an approach that has been called FP-Bonsai). It is based on a prefix tree representation of the given database of transactions (called an FP-tree), which can save considerable amounts of memory for storing the transactions.

The basic idea of the FP-growth algorithm can be described as a recursive elimination scheme: in a preprocessing step delete all items from the transactions that are not frequent individually, i.e., do not appear in a user-specified minimum number of transactions. Then select all transactions that contain the least frequent item (least frequent among those that are frequent) and delete this item from them. This preprocessing is demonstrated in Table, which shows an example transaction database on the left. The frequencies of the items in this database, sorted descendingly, are shown in the middle of this table. If a user specifies minimal support of 3 transactions, items f and g can be discarded. After doing so and sorting the items in each transaction descendingly w.r.t. their frequencies the obtained results are the reduced database shown in Table 1 on the right.

TABLE I. TRANSACTION DATABASE (LEFT), ITEM FREQUENCIES (MIDDLE), AND REDUCED TRANSACTION DATABASE WITH ITEMS IN TRANSACTIONS SORTED DESCENDINGLY W.R.T. THEIR FREQUENCY (RIGHT).

a d f			d a
a c d e	d	8	d c a e
b d	b	7	d b
b c d	c	5	d b c
b c	a	4	b c
a b d	e	3	d b a
b d e	f	2	d b e
b c e g	g	1	b c e
c d f			d c
a b d			d b a

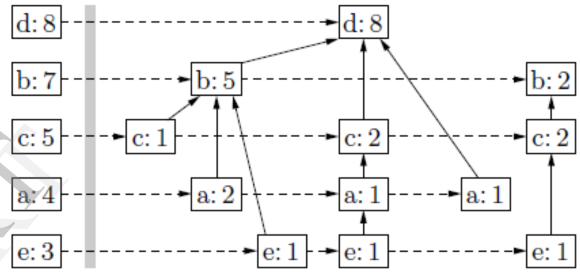

Fig. 1. FP-tree for the (reduced) transaction database shown in Table 1.

IV. METHODOLOGY

Many researchers have shown intra relations. And very few researches have worked on finding inter relations. Here the approach shows the intra and inter relation along with the measurements. The goal is to find the relation among companies in one sector i.e. Intra Sector and companies from different sector i.e. Inter Sector.

A. Preprocessing

In this approach historical prices from Bombay Stock Exchange (BSE) are collected in the tabular form. Among multiple attributes in historical data, only the date attribute and opening price attribute are considered.

Missing values make difficult to perform data analysis and in the data collected from BSE, it had some missing values. Among many solutions to the missing value problem the best solution for stock market data was to impute the nearest neighbor's value. In pre-processing the missing date stock prices values are handled by taking the previous date stock price value of that company.

B. Data Selection

First step is to fetch historical data from two different sectors. Table 1 shows the preview of historical data. Here it is shown the selection of three company stock prices. MarutiSuzuki, Tata Motors, Hero Motor Corp are from Auto sector and BOI, Axis, and BOB are from banking sector from year 2011 to 2013.

TABLE II. DATA PREVIEW FROM 2 DIFFERENT SECTORS

Id	Date	Auto Sector			Banking Sector		
		Maruti Suzuki	Tata Motors	Hero Motor Corp	BOI	Axis	BOB
1	01/01/2011	376	455	451.25	66.5	138.4	234.95
2	02/01/2011	374	456.25	464	69.95	138.75	231
3	05/01/2011	382.1	456	469.4	72.5	153.5	240
4	06/01/2011	382.5	459	468.15	72.9	144	226
5	07/01/2011	365	455.5	449.9	71.5	142.9	227
6	08/01/2011	386	451.7	470	71	141	225.4
7	09/01/2011	404.95	474	479	74.7	144	235.8
:							
531	20/12/2013	1732	371.25	2119	219	1295.1	655
532	23/12/2013	1781.1	368	2084	210	1255	640
532	24/12/2013	1812	372.5	2129.9	217.05	1286	650
534	25/12/2013	1804	374.5	2168	221.3	1291	659
535	26/12/2013	1790	371.5	2139	228	1289.7	648.5
536	27/12/2013	1781.25	368.1	2098	229.25	1311.3	653.3
537	30/12/2013	1776	371.45	2093	238.65	1298.1	655
538	31/12/2013	1775	374.9	2067	235.4	1298	643

C. Data Conversion

The next step is to find the percentage change (Δ). The formula,

$$\%\ \text{change}(\Delta) = \frac{(\text{t2} - \text{t1})}{\text{t1}} * 100$$

Where,

t1 is item from row1 and t2 is item from row 2,

is used to get the stock price change within a day in percentage. The values after converting are show in the table 2.

TABLE III. % CHANGE DATASET CONVERTED FROM ORIGINAL DATASET

Id	Date	Auto Sector			Banking Sector		
		Maruti Suzuki	Tata Motors	Hero Motor Corp	BOI	Axis	BOB
1	02/01/2004	-0.53	0.27	2.82	5.18	0.25	-1.68
2	05/01/2004	2.16	-0.05	1.16	3.64	10.63	3.89
3	06/01/2004	0.10	0.65	-0.26	0.55	-6.18	-5.83
4	07/01/2004	-4.58	-0.76	-3.89	-1.92	-0.76	0.44
5	08/01/2004	5.75	-0.83	4.46	-0.69	-1.32	-0.70
6	09/01/2004	4.90	4.93	1.91	5.21	2.12	4.61
:							
530	20/12/2006	1.76	1.68	2.86	6.56	3.60	1.55
531	23/12/2006	2.83	-0.87	-1.65	-4.10	-3.09	-2.29
532	24/12/2006	1.73	1.22	2.20	3.35	2.47	1.56
533	25/12/2006	-0.44	0.53	1.78	1.95	0.38	1.38
534	26/12/2006	-0.77	-0.80	-1.33	3.02	-0.10	-1.59
535	27/12/2006	-0.48	-0.91	-1.91	0.54	1.67	0.74
536	30/12/2006	-0.29	0.91	-0.23	4.10	-1.00	0.26
537	31/12/2006	-0.05	0.92	-1.24	-1.36	-0.00	-1.83

As there is no previous stock price value for first row in the dataset, it is not possible to find the percentage change value. So in this process, first row data ID1 is ignored and not considered in further processing.

D. Creating set of transaction

The association rule mining algorithm requires transaction data. Further to create a transaction, the process is to take the serial number of the company from the percentage change data and include that serial number in the itemset, if it qualifies the threshold(∂) value. Like,

1,2,3,5
2,3,5,
1,3
...etc

So to the solution was assigning number to the selected company. So 1 is assigned to MaruthiSuzuki, 2 to TataMotors, 3 to HeroMotorCorp, 4 to BOI, 5 to Axis and 6 to BOB. If company's percentage change qualifies the ∂ then that particular company's id is considered in the transaction or null otherwise. Likewise based on this ∂ value filtering is done and a text file is generated. The text file contains only the itemsets that qualify the ∂ value. This generated file is then provided to the Association Rule Mining algorithm for finding frequent items and to find the relation between them.

E. Association Rules generation using FP-Growth

Rules are generated by applying FP-Growth algorithm [7][16]. The actual results provided are like

1 ➔ 2 support: 30.016% confidence: 61.35%

4 ➔ 1 support: 30.016% confidence: 58.43%

3 ➔ 1 support: 30.55% confidence: 58.96%

1 ➔ 3 support: 30.55% confidence: 62.45%

Here the rules indicate the relation between the two companies with support and confidence. In the above rules 1 ➔ 2, 1 is interpreted as and name of company with ID1 and 2 indicates the name of company with ID2. So here 1 is A1 and 2 is A2. Similarly 4 is B1 and 1 is A1. The result of association rule mining is not pleasant. So interpretation of rules to its meaning has to be done so to understand the rules.

F. Interpretation of generated rules:

Interpreting the rules gives the clear knowledge of the algorithm. After interpreting, the rules it looks like,

Rules Showing Positive Trend

- If MarutiSuzuki goes up TataMotors goes up by ∂ with support sup% and confidence conf%.

Similarly the other rule states

- If BOI goes up MarutiSuzuki goes up by ∂ with support sup% and confidence conf%.
- If MarutiSuzuki goes UP by ∂ or more then BOI also goes UP by ∂ or more with sup% Support and conf% Confidence.
- If MarutiSuzuki and B1 goes UP by ∂ or more then Axis also goes UP by ∂ or more with sup% Support and conf% Confidence.

Rules Showing Negative Trend

- If BOI goes DOWN by ∂ or less then HeroMotorCorp also goes DOWN by ∂ or less with sup% Support and conf% Confidence.
- If TataMotors and HeroMotorCorp goes DOWN by ∂ or less then Axis and BOI also goes DOWN by ∂ or less with sup% Support and conf% Confidence.

These above rules are difficult to understand to a person who has no relation with data mining. So interpreting the above output is required so as to be well-informed to the user.

V. Experiments and Results

By the methodology of our approach, various rigorous experiments are carried out. For our first case in detail, refer the methodology.

Test Case 1:
Inter Sector: Auto and Bank.
Company: MarutiSuzuki, TataMotors, HeroMotorCorp, BOI, Axis, BOI.
Historical Dates: From 20011 – 2013
∂: 0.05%
Rules Found:

Rules Showing Positive Trend

- If BoI goes UP by 0.05% or more then BoB also goes UP by 0.05% or more with 39.38% support and 70.91% confidence.
- If BoB goes UP by 0.05% or more then BoI also goes UP by 0.05% or more with 39.38% support and 70.50% confidence.
- If MarutiSuzuki goes UP by 0.05% or more then TataMotors also goes UP by 0.05% or more with 38.32% support and 68.44% confidence.
- If BoI goes UP by 0.05% or more then TataMotors also goes UP by 0.05% or more with 37.54% support and 67.60% confidence.
- If BoI goes UP by 0.05% or more then AxisBank also goes UP by 0.05% or more with 37.22% support and 67.02% confidence.
- If TataMotors goes UP by 0.05% or more then MarutiSuzuki also goes UP by 0.05% or more with 38.32% support and 66.48% confidence.

Rules Showing Negative Trend

- If BoB goes DOWN by 0.05% or less then BoI also goes DOWN by 0.05% or less with 38.56% support and 70.13% confidence.
- If BoI goes DOWN by 0.05% or less then BoB also goes DOWN by 0.05% or less with 38.56% support and 69.84% confidence.
- If TataMotors goes DOWN by 0.05% or less then MarutiSuzuki also goes DOWN by 0.05% or less with 36.25% support and 67.01% confidence.
- If BoB goes DOWN by 0.05% or less then AxisBank also goes DOWN by 0.05% or less with 36.21% support and 65.86% confidence.
- If BoI goes DOWN by 0.05% or less then AxisBank also goes DOWN by 0.05% or less with 36.30% support and 65.75% confidence
- If AxisBank goes DOWN by 0.05% or less then BoI also goes DOWN by 0.05% or less with 36.30% support and 65.47% confidence.

Test Case 2:
Inter Sector: IT and Oil-Gas
Company: TCS, Infosis, ONGC, Petronet
Historical Dates: From 20011 – 2013
∂: 0.1%
Rules Found:

Rules showing positive trend.

1. If Infy goes UP by 0.1% or more then TCS also goes UP by same percentage or more with 38.14% support and 69.49% confidence.
2. If TCS goes UP by 0.1% or more then Infy also goes UP by same percentage or more with 38.14% support and 67.03% confidence.

Rules showing negative trend.

1. If TCS goes DOWN by 0.1% or less then Infy also goes DOWN by same percentage or less with 35.05% support and 66.27% confidence.
2. If Infy goes DOWN by 0.1% or less then TCS also goes DOWN by same percentage or less with 35.05% support and 65.88% confidence.
3. If Infy goes DOWN by 0.1% or less then ONGC also goes DOWN by same percentage or less with 30.99% support and 58.24% confidence.
4. If Petronet goes DOWN by 0.1% or less then ONGC also goes DOWN by same percentage or less with 31.46% support and 55.99% confidence.
5. If ONGC goes DOWN by 0.1% or less then Petronet also goes DOWN by same percentage or less with 31.46% support and 54.32% confidence.
6. If ONGC goes DOWN by 0.1% or less then Infy also goes DOWN by same percentage or less with 30.99% support and 53.51% confidence

Test case 3:
Inter Sector: Bank and Oil-Gas
Company: Bank of Maharashtra, Central Bank, SBI, ONGC, Petronet.
Historical Dates: From 20011 – 2013
∂: 0.1%

Rules Found:

Rules showing positive trend.

1. If MahaBank goes UP by 0.1% or more then CentralBank also goes UP by same percentage or more with 36.89% support and 67.26% confidence.
2. If ONGC goes UP by 0.1% or more then SBI also goes UP by same percentage or more with 33.33% support and 60.77% confidence.
3. If SBI goes UP by 0.1% or more then ONGC also goes UP by same percentage or more with 33.33% support and 59.03% confidence.
4. If Petronet goes UP by 0.1% or more then SBI also goes UP by same percentage or more with 32.52% support and 58.77% confidence.
5. If CentralBank goes UP by 0.1% or more then Petronet also goes UP by same percentage or more with 31.55% support and 58.04% confidence.
6. If SBI goes UP by 0.1% or more then Petronet also goes UP by same percentage or more with 32.52% support and 57.59% confidence.

Rules showing negative trend.

1. If MahaBank and CentralBank goes DOWN by 0.1% or less then SBI also goes DOWN by same percentage or less with 32.66% support and 82.75% confidence.
2. If MahaBank and SBI goes DOWN by 0.1% or less then CentralBank also goes DOWN by same percentage or less with 32.66% support and 80.23% confidence.
3. If ONGC goes DOWN by 0.1% or less then SBI also goes DOWN by same percentage or less with 37.31% support and 65.14% confidence.
4. If ONGC goes DOWN by 0.1% or less then CentralBank also goes DOWN by same percentage or less with 36.22% support and 63.24% confidence

VI. EVALUATION OF ACTUAL RESULTS

We have tested the correctness of the approach by taking the last two year data and applying the algorithm, and checking the result with the recent 2 months stock prices. We found that our approach provides good and satisfying results, enough for the investor to get the idea.

The predicted rules are shown in the graph. By the approach, it has been devised that in the next 30 days, our approach found to be true with confidence 70% and more.

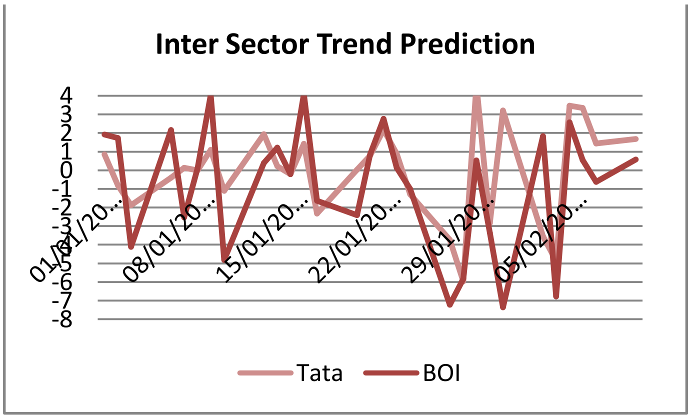

Figure 2: Chart showing the relation in future for TataMotors and BOI.

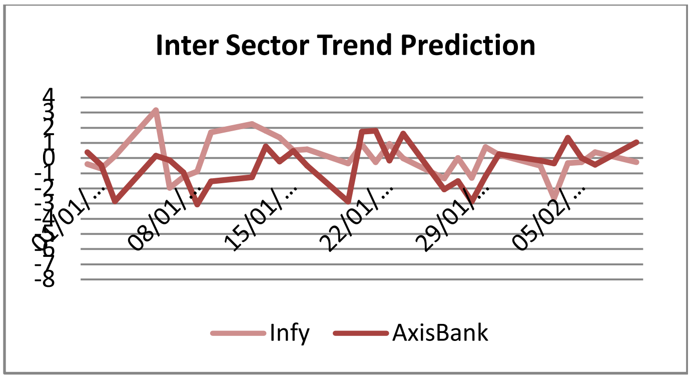

Figure 3: Chart showing the relation in future for Infosys and Axis Bank.

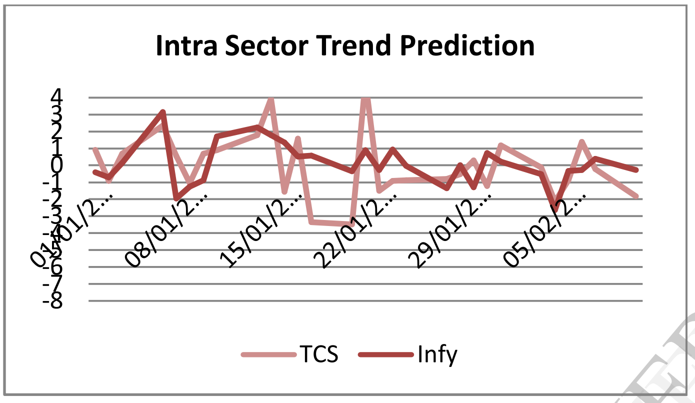

Figure 4: Chart showing the relation in future for TCS and Infosys.

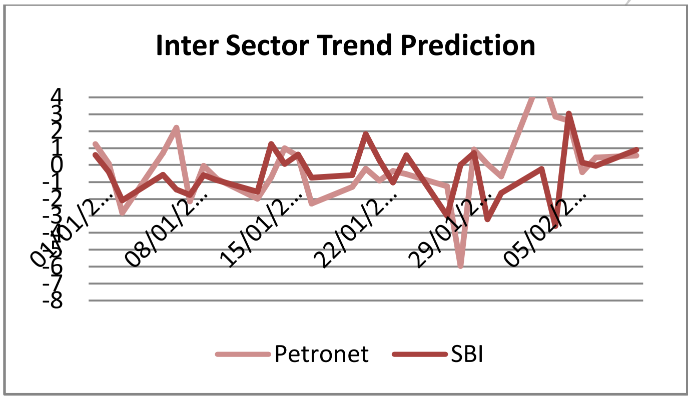

Figure 5: Chart showing the relation in future for Petronet and SBI.

VII. Conclusion

Our aim is to predict close and strong rules and help the investor or individual to have the basic idea on taking decision whether to invest or not. After testing our approach on BSE stock data, we found satisfying results. Our approach shows, how one company stock price can affect other company stock price both in intra-inter sector. Future work will be to improve the performance and design a stock recommendation system that can be useful for every investor.

References

[1] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the International Conference on Very Large Data Bases, VLDB, 1994, pp. 487–499.

[2] R. Agrawal, K. Lin, H.S. Sawhney, K. Shim, Fast similarity search in the presence of noise, scaling, and translation in time-series databases, in: Proceedings of the International Conference on Very Large Data Bases, VLDB, September 1995, pp. 490–501.

[3] R. Agrawal, R. Srikant, Mining sequential patterns, in: Proceedings of the International Conference on Data Engineering, 1995, pp. 3–14.

[4] P. Bloomfield, "Fourier Analysis of Time Series", Wiley, 2000.

[5] T. Fu, T. Cheung, F. Chung, C. Ng, An innovative use of historical data for neural network based stock prediction, in: Proceedings of the International Conference on Information Sciences, JCIS, October 2006.

[6] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufman,2001.

[7] J. Han, J. Pei, and Y. Yin, "Mining Frequent Patterns without Candidate Generation", Proc. 2000 ACMSIGMOD Int. Conf. on Management of Data (SIGMOD'00), Dallas, TX, May 2000.

[8] J. Han, Hong Cheng, Dong Xin, Xifeng Yan, "Frequent pattern mining: current status and future directions".

[9] Y. Kwon, S. Choi, B. Moon, "Stock prediction based on financial correlation", in: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO, June 2005, pp. 2061–2066.

[10] E. Saad, D. Prokhorov, D. Wunsch II, "Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks", IEEE Transactions on Neural Networks (1998) 1456–1470.

[11] V.Umarani, Dr.M.Punithavalli, "A study on effective mining of association rules from huge databases".

[12] W. Yang, Yuefeng Li and Yue Xu, "Granule Based Inter-transaction Association Rule Mining".

[13] Y. Min Ha, Sanghyun Park, Sang-Wook Kim, Jung-Im Wonb, Jee-Hee Yoo,A stock recommendation system exploiting rule discovery in stock databases.

[14] P. Paranjape-Voditel, U. Deshpande, "A stock market portfolio recommender system based on association rule mining".

[15] Chun-Kit Chui, Ben Kao, Edward Hung "Mining Frequent Itemsets from Uncertain Data"

[16] C. Borgelt "An implementation of the FP-growth algorithm" ACM New York, NY, USA ©2005 ISBN: 1-59593-210-0.

[17] Mark Hall Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer Peter Reutemann, Ian H. "The WEKA Data Mining Software: An Update".