Future Trend Prediction of Indian IT Stock Market using Association Rule Mining of Transaction data

Rajesh V. Argiddi
Assit Prof. Department Of Computer Science and Engineering,
Walchand Institute of Technology,
Solapur, India

S. S. Apte
Professor and Head of
Department Of Computer Science and Engineering,
Walchand Institute of Technology,
Solapur, India

ABSTRACT
The approach stated in this paper mainly focuses on minimizing the length of the transaction table of the stock market, based on some common features among the attributes which indirectly minimize the complexity involved in processing; we call this approach as Fragment Based Mining. This deals mainly with reducing the time and space complexity involved in processing the data. Experimentally we try to show our approach is promising one. We conclude that this approach can potentially be used for predictions and recommendations stock trading platforms.

Keywords
Apriori, FITI, Fragment Based Mining, Stock Data.

1. INTRODUCTION
Data Mining also popularly known as Knowledge Discovery in Databases (KDD) refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process. [1]

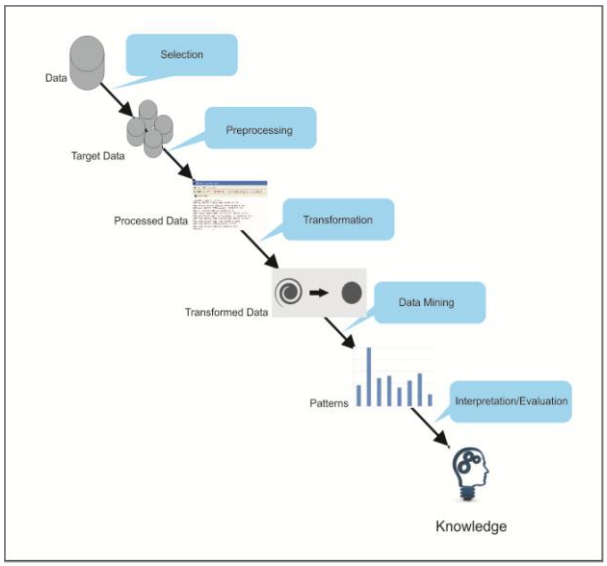

Fig 1: KDD Process

Data mining consists of useful techniques such as Clustering and Association rules, these techniques can be used to predict the future trends based on the Item-sets [6]. Clustering is used to group similar item-sets while association is used to get generalized rules of dependent variables. Useful item-sets can be obtained from huge trading data using these rules. [2]

Association mining, which is widely used for finding association rules in single and multidimensional databases, can be classified into intra and inter transaction association mining. Intra-transaction association refers to association in the same transaction; inter-transaction association indicates association among different transactions [3]. Most contributions in association mining focus on intra-transaction association also referred to traditional association mining. Inter-transaction association mining was proposed in 2000 [3] and has a broad range of applications, though its basic idea extends from intra-transaction association mining. [4]

Stock Prices are considered to be very dynamic and susceptible to quick changes because of the underlying nature of the financial domain and in part because of the mix of known parameters (Previous Day's Closing Price, P/E Ratio etc) and unknown factors (like Election Results, Rumors etc). [7]

In this research we have taken the original data sets of Bombay Stock Exchange (BSE) of different companies such as Infosys, TCS, and Oracle etc from Yahoo Finance and try to find the association among the large scale IT companies and Small scale IT companies.

As we know that there are always some dependencies between different fields in stock market. Our aim is to find whether large scale companies affect the small scale companies' shares.

Some experimental results shows that there is a strong relation between large and small scale companies, we found that major of the times when the share value of large companies go high, small scale companies shares also goes high and vice-versa.

Granule mining [4] finds interesting associations between granules in databases, where a granule is a predicate that describes common features of a set of objects (e.g., records, or transactions) for a selected set of attributes (or items). For example, a granule refers to a group of transactions that have the same attribute values. Granule mining extends the idea of decision tables in rough set theory into association mining. The attributes in an information table consist of condition attributes and decision attributes, with users' requirements.

As in granule mining, fragment based approach fragments the data sets into fragments for processing thereby reducing the input size of data sets fed to the algorithm. In contrast to granule mining, in fragment based mining the condition and decision attributes are summed for obtaining generalized association rules.

2. RELATED WORK

In the previous research, different data warehouse systems presented different techniques to support data mining; Ahmed et al. [9] presented the data warehouse backboned system integrated data mining and OLAP techniques. This system makes use of a router to adopt the previous mining result stored in the data warehouse, accordingly avoiding processing large amounts of the raw data. [8]

Both fundamentalists and technicians have developed certain techniques to predict prices from financial news articles. In one model that tested the trading philosophies; LeBaron et. al. posited that much can be learned from a simulated stock market with simulated traders (LeBaron, Arthur et al. 1999).

Extending the work from intra transaction to inter transaction association mining; by H. Lu, J. Han, and L. Feng (2000) for multi dimensional data set has been successfully evaluated ; this work was done for transactions of multidimensional data inter association.

M. Chen, C. Huang, proposed a technique in data mining to group the customer order in warehouse management system. This technique groups the data based on orders of customers and store it in a proper order in the warehouse.

Wanzhong Yang also proposed one innovative technique to process the stock data named Granule mining technique, which reduces the width of the transaction data and generates the association rules. [4]

Our aim is to extend the work in this field and provide some basic abstractions (Fragments).

3. BACKGROUND

3.1 Apriori Algorithm

Developed by Agarwal and Srikant 1994 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 (for example, collections of items bought by customers, or details of a website frequentation).

The algorithm attempts to find subsets which are common to at least a minimum number C (the cutoff, or confidence threshold) of the item-sets.

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. [10]

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.

3.2 FITI (First Intra then Inter)

The FITI algorithm [11] is based on the following property, a large inter-transaction item-set must be made up of large intra-transaction item-sets, which means that for an item-set to be large in inter-transaction association rule mining, it also has to be large using traditional intra-transaction rule mining methods. By using this property, the complexity of the mining process can be reduced, and mining inter-transaction association rules can be performed in a reasonable amount of time. First FITI introduces a parameter called maxspan (or sliding window size), denoted w. This parameter is used in the mining of association rules, and only rules spanning less than or equal to w transactions will be mined.

Second, every sliding window in the database forms a mega transaction. A mega transaction in a sliding window W is defined as the set of items W, appended with the sub window number of each item. The items in the mega transactions are called extended items.

Txy is the set of mega transactions that contain the set of extended items X, Y, and Tx is the set of mega transactions that contain X. The support of an inter-transaction association rule X=> Y is then defined as”

Support = |Txy| /S, Confidence = |Txy|/|Tx|

4. METHODOLOGY

There are some weaknesses in the previous FITI approaches such as time and space involved in processing the data is more. In FITI approach it is difficult to process an information table with many attributes and long intervals for inter transaction associations. This results into large amount of time and cost in processing the data.

Fragment based mining groups all the attributes once and performs the operation group wise instead of single attribute, which results into more generalized rules.

Table 1. Indian IT Stock Market Transaction Table

ID	Date	A	B	C	X	Y	Z
1	4/1/2010	128	727	100	750	2606	697
2	5/1/2010	128	742	101	756	2614	700
3	6/1/2010	133	735	107	752	2618	709
4	7/1/2010	139	724	110	733	2580	693
5	8/1/2010	130	709	108	700	2575	680

Let T= {ID1, ID2, ID3,….., IDn} be a transaction database as shown in the Table 1. In this table A, B, C, X, Y, Z is the shares from Indian IT Stock Market that represent KPIT, Mphasis, MahiStyam, TCS, Infosys, and Wipro respectively.

Here A, B, C are the Small Scale Company share and X, Y, Z represent Large Scale Company shares respectively.

Here share price refers only for the open price at the transaction data.

Here in this fragment based approach we are trying to reduce the length of the input table, so the time needed to process the table will ultimately get reduced and also be able to find more efficient rules.

Our main aim is to reduce the size of the table and increase the performance.

Table 2. SUM Function for small scale attributes

ID	Date	A	B	C	Small Scale SUM
1	4/1/2010	128	727	100	1281
2	5/1/2010	128	742	101	1309
3	6/1/2010	133	735	107	1307
4	7/1/2010	139	724	110	1314
5	8/1/2010	130	709	108	1281

In above Table 2 we add all the shares of the small scale companies and form one single SUM function, i.e. it is the aggregation of all the shares of the small scale companies.

TABLE 3. SUM Function for large scale attributes

ID	Date	X	Y	Z	Large Scale SUM
1	4/1/2010	750	2606	697	6379
2	5/1/2010	756	2614	700	6442
3	6/1/2010	752	2618	709	6444
4	7/1/2010	733	2580	693	6361
5	8/1/2010	700	2575	680	6281

In above Table 3 we add all the shares of the large scale companies and form one single SUM function, i.e. it is the aggregation of all the shares of the large scale companies.

TABLE 4. Small scale and large scale sum

ID	Small Scale SUM	Large Scale SUM
1	1281	6379
2	1309	6442
3	1307	6444
4	1314	6361
5	1281	6281

The fragment based approach divides the attributes into two tiers: Small Scale and Large Scale SUM attributes. This innovation can largely reduce the number of extended item sets, therefore we can largely reduce the number of extended item sets, therefore we can use large intervals for inter transaction association mining in real application.

In above Table 4, ID1 represents transaction one and ID 2 represent the transaction two.

Let Δ be the differences for the attribute values among inter transactions. Assume 1, 0 illustrates the increase and decrease respectively.

Let ΔID1 be the difference between ID2 and ID1, where ΔID1=ID2-ID1. For Small Scale, Δ Small Scale1=Small Scale2-Small Scale1=1309-1281=28, because ΔSmall Scale>=0, therefore ΔSmall Scale1=1, similarly ΔLarge Scale3=Large Scale4-Large Scale3=6361-6444=-83, as ΔLarge Scale3<0, therefore ΔLarge Scale3=0. In this fashion we converted the above table 4 to Table 5.

TABLE 5. Converted Transaction Table

ID	Small Scale SUM	Large Scale SUM
1	1	1
2	0	1
3	1	0
4	0	0
5	--	--

Now according to our approach we will consider only those transactions whose both small scale and large scale SUM is same i.e. both are 1, 1 or 0, 0 respectively.

This we do because we are only interested in finding the association if both small and large scale companies increase or decrease at the same time.

TABLE 6. Transaction Accepting Rule

Input1	Input2	Transaction
1	1	Accept
1	0	Reject
0	1	Reject
0	0	Accept

So the original transaction Table I will get minimized as shown in the Table 6.

Table 6. Fragmented Transaction Table

ID	Date	A	B	C	X	Y	Z
1	4/1/2010	128	727	100	750	2606	697
4	7/1/2010	139	724	110	733	2580	693

5. EXPERIMENTS AND RESULTS

5.1 FITI Algorithm

In this method we have collected last 3 years data of Indian IT Stock Market from Yahoo Finance and converted that into a tabular format and applied FITI algorithm.

Input Data:

ID	KPIT	Mphasis	Mahi Stym	TCS	Infosys	Wipro
1	0	0	1	1	0	0
2	0	1	0	0	0	1
3	0	1	0	0	1	0
4	0	0	1	1	0	1
5	1	1	0	0	0	0
. . . .						
729	0	0	1	1	1	1
730	0	0	0	1	0	1
731	1	1	0	1	1	0

Output Association Rules before applying Fragment Based Mining:

1. TCS=1 (↑) ==> Infosys=1 (↑) conf: (0.74)
2. Infosys=1 (↑) ==> TCS=1 (↑) conf: (0.73)
3. Infosys=0 (↓) ==> TCS=0 (↓) conf: (0.72)
4. TCS=0 (↓) ==> Infosys=0 (↓) conf: (0.7)
5. KPIT=1 (↑) ==> Mphasis=1(↑) conf: (0.63)
6. Mphasis=0 (↓) ==> KPIT=0 (↓) conf: (0.63)
7. Mphasis=1 (↑) ==> KPIT=1(↑) conf: (0.6)
8. KPIT=0 (↓) ==> Mphasis=0 (↓) conf: (0.59)
9. Wipro=1 (↑) ==> Infosys=1 (↑) conf :(0.57)
10. Infosys=1 (↑) ==> Wipro=1 (↑) conf: (0.56)

The first association rule shows that TCS and Infosys have .74 confidences, that if TCS goes high (↑) then Infosys will also go high (↑).

And the 6th association rule shows that Mphasis and KPIT has .60 confidence, that if Mphasis goes low (↓) then KPIT will also goes low (↓).

5.2 Fragment Based Approach

After applying the fragmentation rule we get the following minimized table. Now we apply the Apriori on this processed data and find the association rules among the attributes.

Fragmented Input Data:

ID	KPIT	Mphasis	Mahi Stym	TCS	Infosys	Wipro
1	0	1	0	0	0	1
2	0	1	0	1	1	1
3	0	0	1	0	0	0
4	1	1	0	1	0	0
5	1	1	1	0	0	0
. . . .						
492	0	1	1	0	1	0
493	1	1	1	1	1	1
494	1	1	1	1	1	1

In Fragment Based Approach we can observe the input size of the processed data is reduced from 731 rows to 494 rows, i.e. near about 33.33% data redundancy has been achieved. This technique produces most efficient rules as shown below compared to FITI approach.

Output Association Rules after applying Fragment Based Mining:

1. TCS=1 (↑) ==> Infosys=1 (↑) conf: (0.83)
2. Infosys=0 (↓) ==> TCS=0 (↓) conf: (0.81)
3. Infosys=1 (↑) ==> TCS=1 (↑) conf: (0.79)
4. TCS=0 (↓) ==> Infosys=0 (↓) conf: (0.76)
5. KPIT=1 (↑) ==> Mphasis=1 (↑) conf: (0.69)
6. Mphasis=0 (↓) ==> KPIT=0 (↓) conf: (0.69)
7. Mphasis=1(↑) ==> KPIT=1 (↑) conf: (0.62)

8. KPIT=0(↓) ==> Mphasis=0 (↓) conf: (0.62)

9. KPIT=0(↓) ==> MahiStym=0 (↓) conf: (0.61)

10. MahiStym=0(↓) ==> KPIT=0 (↓) conf: (0.6)

The first association rule shows that TCS and Infosys have .83 confidences, that if TCS goes high (↑) then Infosys will also go high (↑).

And the 6th association rule shows that Mphasis and KPIT has .69 confidence, that if Mphasis goes low (↓) then KPIT will also goes low (↓).

6. CONCLUSION

On subsequent evaluation we find that fragment based approach as a promising one for extracting some association rules of predictive nature from Indian IT Stock Market which could be used for prediction or recommendations in Stock trading platforms and packages. We presented the result implementation of fragment mining algorithm. This experiment showed that fragment based mining algorithm gets accurate results with less time and space complexity as compared to FITI algorithm. In future this technique can be applied on other Stock Market fields such as Banking, Textiles, and Marketing sectors efficiently.

7. REFERENCES

[1] Osmar R.Zaiane, "Principles of Knowledge Discovery in Databases", 1999.

[2] Dattatray P.Gandhmal, Ranjeetsingh Parihar,and Rajesh Argiddi "An Optimized approach to analyze stock market using data mining technique", IJCA, ICETT 2011.

[3] H. Lu, J. Han, and L. Feng (2000). "Beyond intratransaction association analysis: mining multidimensional intertransaction association rules." ACM Transactions on Information Systems 18(4): 423-454.

[4] Wanzhong Yang, "Granule Based Knowledge Representation for Intra and Inter Transaction Association Mining", Queensland University of Technology, July 2009.

[5] J. Dong and M. Han (2007). IFCIA: An Efficient Algorithm for Mining Intertransaction Frequent Closed Item sets. The fourth international conference on fuzzy systems and knowledge discovery, China.

[6] Gebouw D, B-3590 Diepenbeek, Belgium "Building an Association Rules Framework to Improve Product Assortment Decisions" 2004.

[7] Eugene F. Fama "The Behavior of Stock Market Prices", The Journal of Business, Jan 1965.

[8] R. S. Monteiro, G. Zimbrão, H. Schwarz, B. Mitschang, and J. M. Souza (2005). "Building the Data Warehouse of Frequent Itemsets in the DWFIST Approach." Foundations of Intelligent Systems 3488: 294-303.

[9] K. M. Ahmed, N. M. El-Makky, and Y. Taha (1998). Effective data mining: a data warehouse-backboned architecture. The 1998 conference of the Centre for Advanced Studies on Collaborative research, Toronto.

[10] Professor Lee "Apriori Algorithm Review for Finals" Spring 2007.

[11] Ole Kristian Fivelstad "Temporal Text Mining" Norwegian University of Science and Technology, June 2007.

[12] M. Chen, C. Huang, H. Wu, M. Hsu, F. Hsu (2005). A Data Mining Technique to Grouping Customer Orders in Warehouse Management System. The Fourth IEEE International Workshop on Soft Computing as Tran disciplinary Science and Technology.