Fragment Based Approach to Forecast Association Rules from Indian IT Stock Transaction Data

Rajesh V. Argiddi, Sulabha S. Apte

Computer Science Department,
Walchand Institute of Technology
Solapur,India

Abstract— **In this research we mainly focus on overcoming the drawbacks in FITI approach in predicting the stock market and propose a new approach called fragment based mining which gives some promising results as compared to FITI. FITI consists of all the transaction from the stock market some of which are not necessary and simply increases the overhead in processing the data, so we improve this by reducing the number of transactions using some aggregate functions, so the time needed to process the transactions will be less and generate some efficient rules from which we predict the stock market behavior. This research is purely based on the data mining technique called association mining. Association rules suites the behavior of stock market and helps in analyzing the associations among the companies. As mentioned above we propose a technique Fragment Based mining which helps in minimizing the input transaction table size which leads to reduced processing time.**

Keywords— **FITI; Fragment Based Mining, Association mining; Stock Data.**

I. 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]

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]

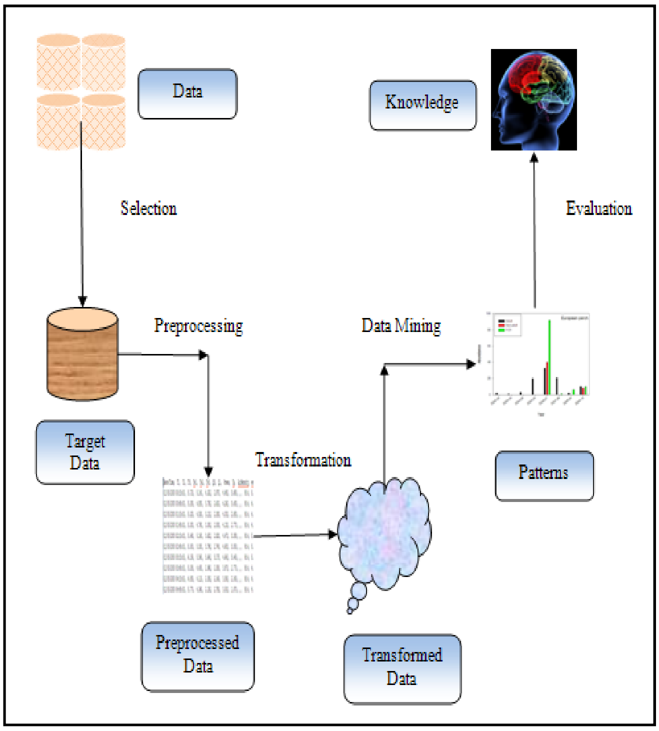

Fig. 1 KDD Process

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.

II. Related Work

In the previous research, different inter transaction techniques for multidimensional data has been proposed for data mining; Anthony J.T. Lee, Chun-Sheng Wang, Wan-Yu Weng, Yi-An Chen, Huei-Wen presented " An Efficient algorithm for mining closed inter-transaction item-sets" an ICMiner, for mining closed inter-transaction item-sets. He performed on the synthetic, real & "worst case" datasets and concluded ICMiner is more efficient than the EH-Apriori, FITI approaches.

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).

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).

III. Background

A. Association Rule Mining

Association rule mining is a technique for discovering unsuspected 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 (e.g. shopping baskets), whether the occurrence of specific items, implies also the occurrence of other items with a relatively high probability. In principle the answer to this question could be easily found by exhaustive exploration of all possible dependencies, which is however prohibitively expensive. Association rule mining thus solves the problem of how to search efficiently for those dependencies. 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. ssociation rules are implications of the form X -> Y where X and Y are two disjoint subsets of all available items. X is called the antecedent or LHS (left hand side) and Y is called the consequent or RHS (right hand side). Association rules have to satisfy constraints on measures of significance and interestingness

B. FITI(First Intra then Inter) Algorithm

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|

IV. Methodology

As FITI algorithm takes lot of time in processing the data so we focus mainly on reducing time and produce more realized association rules. Fragment Based approach works on overcoming the drawbacks of FITI approach which groups the transactions instead of considering all the transactions from the stock data.

Our goal in this research is to find association among the Small and Large Scale IT companies from Indian IT Stock data.

In this new approach we consider a single transaction as aggregation of the length of the sliding window, which helps in producing more generalized rules as compared to FITI approach. window and Small Scale SUM 2 as the second sliding window and so on.

TABLE I. INDIAN IT STOCK MARKET TRANSACTION (SMALL SCALE)

ID	Date	A1	A2	A3
1	1/1/2009	315	152	242
2	2/1/2009	320	154	240
3	3/1/2009	320	162	230
4	4/1/2009	310	157	236
5	5/1/2009	310	160	231
6	6/1/2009	315	134	223
7	7/1/2009	320	125	237
8	8/1/2009	300	135	238
………………………………………………………………………………..				
100	6/4/2009	306	140	236
101	7/4/2009	304	140	237
102	8/4/2009	300	145	239
103	9/4/2009	322	158	240

Sliding Window

Small Scale SUM 1

Small Scale SUM 2

Small Scale SUM N

This huge amount of data is minimized by performing the aggregation based on the sliding window size. Here we defined the size of the sliding window as four and move this sliding window linearly and do the same aggregation. This is expressed in the figure below i.e. Fragmented Transaction table.

TABLE II. FRAGMENTED DATA OF SMALL SCALE COMPANIES

ID	Small Scale SUM		
	A1	A2	A3
1	1265	625	948
2	1245	554	929
………………………………………………………………………………………………			
N	1232	583	952

We do the same operation as that for Small Scale companies; here too we define the Sliding window size as four, Large Scale Sum 1 as the first sliding window and Large Scale Sum 2 as the second sliding window and so on.

TABLE III. INDIAN IT STOCK MARKET TRANSACTION TABLE (LARGE SCALE)

ID	Date	B1	B2	B3
1	1/1/2009	2745	1701	835
2	2/1/2009	2755	1675	815
3	3/1/2009	2760	1590	825
4	4/1/2009	2767	1650	817
5	5/1/2009	2689	1725	835
6	6/1/2009	2735	1698	820
7	7/1/2009	2679	1699	814
8	8/1/2009	2714	1690	810
………………………………………………………………………………………..				
100	6/4/2009	2686	1655	825
101	7/4/2009	2695	1724	840
102	8/4/2009	2699	1710	865
103	9/4/2009	2729	1709	849

Sliding Window

Large Scale SUM 1

Large Scale SUM 2

Large Scale SUM N

The Fragmented data of this huge table is expressed in the table below.

TABLE IV. FRAGMENTED DATA OF LARGE SCALE COMPANIES

ID	Large Scale SUM		
	B1	B2	B3
1	11027	6616	3292
2	10817	6812	3279
…………………………………………………………………………………………………			
N	10809	6798	3379

Now to generate rules among small and large scale companies data we perform inter transactions among the both i.e. transaction 1 from small scale companies is related with transaction 4 from large scale companies and so on.

TABLE V. INTER-TRANSACTION AMONG SMALL AND LARGE SCALE COMPANIES

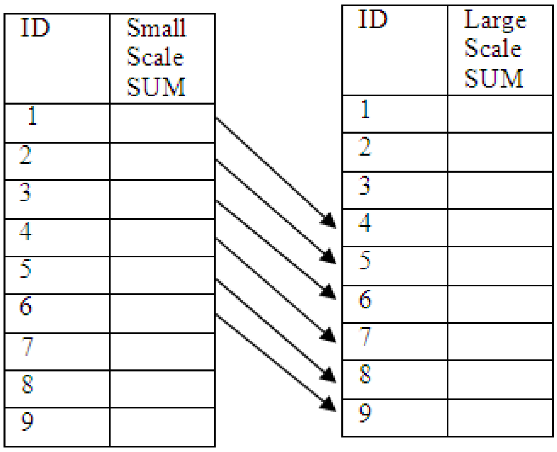

TABLE VI. INTER-TRANSACTION DATA

ID	Small Scale SUM			Large Scale SUM		
	A1	A2	A3	B1	B2	B3
1	1265	625	948	11030	6712	3239
2	1245	554	929	11133	6706	3260
3	1230	596	990	10990	6731	3271
4	1312	610	901	10925	6690	3195
5	1269	612	893	11002	6721	3189
6	1280	615	940	11104	6740	3245

This inter-transaction data need to be converted into 1's and 0's i.e. with gain or loss of transaction. This we do by performing ID1= ID2-ID1, if ID1>0 then 1 otherwise 0, in this manner we convert the above table into table below.

TABLE VII. CONVERTED TABLE

ID	Small Scale SUM			Large Scale SUM		
	A1	A2	A3	B1	B2	B3
1	0	0	0	1	0	1
2	0	1	1	0	1	1
3	1	1	0	0	0	0
4	0	1	0	1	1	0
5	1	1	1	1	1	1
6	--	--	--	--	--	--

V. EXPERIMENTS AND RESULTS

A. *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.

Here KPIT,Mphasis and MahiStym belong to Small Scale Companies where as TCS, Infosys and Wipro belong to Large Scale Companies respectively.

Input Data:

ID	KPIT	Mphasis	MahiStym	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 has .74 confidence, 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 (↓).

B. *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.

Here KPIT,Mphasis and MahiStym belong to Small Scale Companies where as TCS, Infosys and Wipro belong to Large Scale Companies respectively.

Fragmented Input Data:

ID	**KPIT**	**Mphasis**	**MahiStym**	**TCS**	**Infosys**	**Wipro**
1	0	1	0	0	0	1
2	0	0	0	0	0	0
3	0	0	1	1	0	0
4	1	0	1	0	0	1
5	0	1	1	0	1	0
		.			.	
			.			
		.			.	
			.			
		.			.	
			.			
181	1	1	1	0	0	1
182	0	1	1	0	0	1
183	1	0	0	0	0	0

Output Association Rules after applying Fragment Based Mining:

1. Infosys=0 97 ==> TCS=0 75 conf:(0.77)
2. TCS=0 100 ==> Infosys=0 75 conf:(0.75)
3. TCS=1 82 ==> Infosys=1 60 conf:(0.73)
4. Mphasis=0 84 ==> KPIT=0 61 conf:(0.73)
5. KPIT=1 80 ==> Mphasis=1 57 conf:(0.71)
6. Infosys=1 85 ==> TCS=1 60 conf:(0.71)
7. Mphasis=0 84 ==> MahiStym=0 58 conf:(0.69)
8. MahiStym=1 81 ==> Mphasis=1 55 conf:(0.68)
9. MahiStym=0 101 ==> KPIT=0 67 conf:(0.66)
10. KPIT=0 102 ==> MahiStym=0 67 conf:(0.66)

The first association rule shows that Infosys and TCS has .77 confidence, that if Infosys goes high (↑) then TCS will also go high (↑).

And the 8th association rule shows that MahiStym and Mphasis has .68 confidence, that if MahiStym goes low (↓) then Mphasis will also goes low (↓).

VI. CONCLUSIONS

By some experimental analysis we find Fragment Based approach generated more generalized rules as compared to FITI approach. Also time needed to process the data is less as we reduced the size of the input table. The rules generated from Fragment Based approach can be recommended to the customers who invest their money in the stock market.

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.ed Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.