A Novel Framework for Stock Trading Analysis Using Casual Relationship Mining

Harchana Bhoopathi
Dept. of Computer Science
Kakatiya University
Warangal, India
archana.bhupathi@gmail.com

B.Rama
Asst.Prof. Dept. of Computer Science
Kakatiya University
Warangal, India
rama.abbidi@gmail.com

Abstract—**Stock market has been around as a platform for buyers and sellers of shares. Success in stock trading, which depends on comprehensive knowhow on the trends of trading, is the dream of thousands of enthusiasts. Investors employ different means of analysis for expert decision making including periodical trade indicators, industry growth indicators, and economic forecasts. An automated approach for stock trading analysis is inevitable as there is huge amount of temporal data is involved. Several techniques came into existence as revealed in the literature. Data mining techniques are, of late, widely used for discovering trends of stock market. High-confidence rules extracted using such techniques are found instrumental to make well-informed decisions. In this paper we propose a framework with an underlying approach to have casual data mining which explores inter-transaction relationships that are latent in the stock datasets. This is used to unearth the trends in stock trading that can help investors to have strategies for profitable business. Our framework also provides placeholders for government policy and unexpected incidents to reduce error rate in the prediction of stocks. Our empirical results reveal that the proposed framework improves stock trading analysis and prediction accuracy significantly.**

Index Terms – Data mining, stock trading, casual data mining, prediction of stocks.

I. INTRODUCTION

Stock market is one of the most interested areas of research as it involves stakeholders who do business. The population that depends on stocks or shares business is increasing every year. Their lives and their wellbeing depend on the intelligent decision making. Wrong decisions are costly. Right investment at right time can give dividends [1]. However, the knowledge to make well informed decisions plays very crucial role in stock business. Therefore many researchers contributed towards making techniques or algorithms that provide needed knowhow. Data mining has been around and it is best used to process huge amount of data and obtain information that was not known earlier. Such interesting information is extracted from datasets in the form of trends or patterns that give valuable insights into the ways tickers are performing and even customer behaviour [2].

Association rule mining is part of data mining which is widely used in the real world applications. Association rule explains the relationship among item sets. Especially association rules are generated from frequent item sets. Frequent item sets are the set of item sets that appear together in space or time domain. The associations among transactions play vital role in stock market analysis [3], [4]. However, association rules are of different kinds. They are direct associations, indirect associations and exception associations. Direct association refers to the association between two or more items with respect to frequency. Even infrequent associations which are latent in the datasets can provide valuable information. Often two items are not directly associated but they are associated with some mediator. This behaviour or phenomenon is useful in stock business as a stock value can bring other two stock values to get influenced and related. This is called indirect relationship or association. Exception relationships exist between item sets containing low support and high confidence. Support and confidence are the statistical measures used in association rule mining. They are used to obtain high quality association rules.

Our contributions in this paper are described here. We proposed a framework that has provision for generating direct, indirect, and exception association rules. It also considers the events and government decisions that can influence stock market business. When all are considered to have comprehensive business intelligence, it will be useful in making well informed decisions. The remainder of the paper is structured as follows. Section II provides review of literature. Section III provides problem formulation. Section IV presents the proposed system in detail. Section V shows experimental results while section VI concludes the paper besides providing directions for future work.

II. RELATED WORKS

Tan et al. (2000) [4] studied the inter-transaction differences in stock market. They focused on in finding indirect association rules. Indirect association rule mining has significant utility in the real world. Hsieh et al. (2002) [3] explored data mining for stock market analysis in terms of downstream and upstream causal relationships. They focused on inter-transaction mining that includes time-interval dimension. The temporal relationship in stock can provide useful insights. Wan and an (2002) [5] proposed an algorithm named HI-mine for discovering indirect association rules. These association rules do not exhibit direct relationships between two items but they do it with the third item. Often this kind of relationship is valuable and provides needed business intelligence.

Vu et al. (2012) [6] focused on sentiment analysis and integrated it with stock market prediction. They used decision tree classifier. They made use of both positive and negative sentiments for prediction accuracy. Kaur and Mangat (2012) [7] discussed data mining techniques for stock price prediction, stock index prediction, portfolio management, recommender systems, and discovering patterns from stock data. Xu (2012) [8] used social media data for sentiment analysis with respect to stock volume correlation, negative and polarized detections. Korgaonkar (2012) [9] investigated the role of foreign direct investment (FDI) on financial improvement of a country and found that FDI could have positive impact. It depends on banking sector and stock market variables.

Kumari et al. (2013) [10] focused on retail forecasting by using data mining technique and neural network. Their focus was to build a model for prediction that can help improve customer satisfaction. Pham et al. (2013) [11] proposed interestingness measures and defined an algorithm for efficient sequential rule mining for prediction of stock market. Interestingness measure is used to obtain quality rules. Karabulut (2013) [12] studied Gross National Happiness (GNH) through social medium like Facebook with respect to stock market analysis. Prasanna and Ezhilmaran (2013) [13] discussed data mining techniques for predicting stock market. Al-Radaideh et al. (2013) [2] investigated stock prediction using decision rule mining.

Das and Uddin (2013) [14] proposed a methodology for stock market analysis using data mining techniques. They opined that neural network technologies were used more for prediction of stock markets. Cao (2013) [15] studied complex social and behaviour problems using data mining techniques and related them with stock markets. Evangelopoulos et al. (2013) [16] studied micro and macro messages in social media to predict stock prices. They developed a framework that contains micro and macro information dissemination and processing. Sherdiwala (2014) [17] studied different algorithms in data mining domain for stock market analysis. The techniques are explored include decision trees, neural network, clustering, association rules, and factor analysis. Kumar and Choudhary (2014) [18] focused on data mining in mobile devices and presented different frameworks for mining with mobile networks.

Kuisyte (2014) [19] studied the meaning of efficient stock market with Baltic economies and the historical statistics. They used dummy variable approach for effects of day of the week with respect to stock prices. Desai and Gandhi (2014) [20] performed sentiment analysis on stock market by exploiting SentiWordNet. Ganguly and Busch (2014) [21] built a project in Python for stock market analysis. They also focused on customer behaviour in stock market business. Thakkar et al. (2014) [22] focused on intra-day transactions to predict stock markets.

Preethi and Santhi (2015) [1] studied different mechanisms in data mining for stock market forecasting. The techniques include neuro-fuzzy system, Hidden Markov Model (HMM), time series analysis using random walk, moving average, regression method, and ARIMA model. Navale et al. (2016) [23] combined data mining and artificial intelligence for prediction of stock market and felt for the need for further research in the area. Borde et al. (2016) [24] presented various techniques for predicting future closing price of stocks to analyse significant increases or decrease in prices. In this paper we proposed a framework for finding upstream and downstream causal relationships in a comprehensive fashion.

III. PROBLEM FORMULATION

Data mining techniques such as association rule mining is widely used for discovering knowledge from databases. Associations in transactions can provide hidden information in the form of trends or patterns. A database D can have set of associations A that can be discovered by generating rules R. The problem of mining association rules can be divided into two parts. In the first part, all item sets with given support are obtained from data source. Then the frequent item sets are used to generate association rules. The problem with this kind of mining association rules is that any item set which does not have the support is considered an uninteresting item set. However, we believe that associations can be of different kinds. They are known as direct, indirect and exception associations. With respect to stock market data, it is essential to have comprehensive business intelligence before making decisions. Therefore it is inevitable to have a framework that can cater to the needs of such expert decision making. The rationale is described here. In stock market dataset item sets that are infrequent also can provide information needed towards converging decisions. Two items a and b have no direct relationship but they may have strong relationship through another item set Y. In this case the a, b pair is said to have indirection relationship or association. Mining such rules is the focus of this paper. We defer the details of direct and exception relationships besides effect of events and government decisions are deferred to our next research paper.

IV. PROPOSED FRAMEWOK

In this paper, we proposed a framework that has provision for finding causal relationships in terms of direct, indirect and exception relationships. When all these relationships are investigated, it can lead to more intelligent decision making in stock market business. In addition to this certain events and unexpected decisions made by government can have impact on stock market. When all these are considered, the error rate in the prediction can be minimized. Though the framework has provision for discovering different relationships, in this paper we present finding indirect relationships of stock tickers.

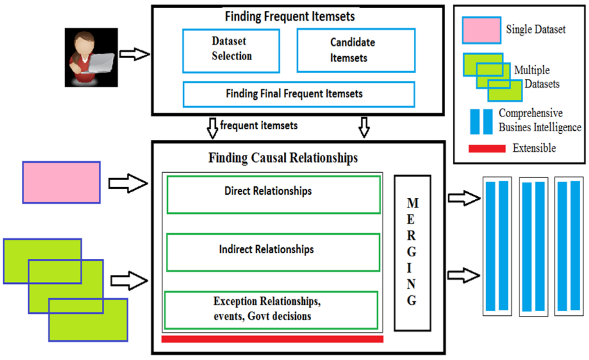
Finding Frequent Itemsets
Dataset Selection
Candidate Itemsets
Finding Final Frequent Itemsets
frequent itemsets
Finding Causal Relationships
Direct Relationships
Indirect Relationships
Exception Relationships, events, Govt decisions
MERGING
Single Dataset
Multiple Datasets
Comprehensive Busines Intelligence
Extensible

Figure 1: Proposed Framework for Comprehensive Stock Business Intelligence

As shown in Figure 1, it is evident that the basis for mining causal relationships is to have frequent item sets from given stock dataset. The ensuing section provides more details on stock dataset used for empirical study. The framework has provision to analyse one or more stock datasets. After extracting frequent item sets using any of the existing data mining algorithms, it focuses on mining direct, indirect and exception relationships from the frequent item sets. The resultant rules obtained are integrated to form comprehensive business intelligence. As far as implementation is concerned, in this paper, we present only the study on indirect relationships. In other words our focus is on inter-transactions that have no direct relationship but through a mediator. This can show potential tickers that have influence on others.

A. *Indirect Association Rule Mining Algorithm*

Mining indirect association rules is done using the following algorithm. It has two phases. In the first phase frequent item sets are mined and then they are further used to generate indirect associations and rules are generated.

```
Input: Stock Dataset SDB, support sup, confidence con
Output: Indirect Association Rules R
Initialize support vector S
For each item pair in SDB
    Find support value
    Add support to S
End for
For each item pair in SDB
    IF support<sup THEN
        Prune it from S
    END IF
End For
For each item pair in SDB
    Find mediator
    Extract indirect relationship
    Add relationship rule to R
End For
Return R
```

As shown in the above algorithm there are two important phases. The first step focuses on candidate generation and the second step is for pruning. It makes use of an algorithm like Apriori to extract frequent item sets. This knowhow is used in the algorithm for finding indirect associations. Item sets that do not have mediator dependence are pruned and final associations that have relationships through mediator are extracted and rules are generated based on the support and confidence. The support vector is constructed by scanning entire dataset once. For each infrequent item pair (a, b) finding mediator is an iterative process that is expensive. The complexity of this process depends on the number of frequent item sets that contain the items such as a or b.

V. EXPERIMENTS AND RESULTS

We made experiments with a prototype application which demonstrates the proof of concept. To evaluate the utility of the proposed algorithm, we made empirical study with S&P 500 stock market dataset which is of 2009 and 2010. The dataset excerpt is shown in Table 1. The experiments are made on a PC with Inter Core i54210U CPU at 1.70 GHz speed and 4.0 GB RAM running Windows 10 64 bit operating system. The aim of the experiments is to find out infrequent item sets (a, b) and find mediator through which they are related.

We made two sets of experiments considering five tickers in each experiment.

A. S&P 500 Stock Market Data Set

We collected stock market dataset from [25]. It contains 122574 instances. Each instance has values for attributes such as Date, Ticker (Stock Symbol of Company), Open, High, Low, Close, and Volume for the day. The data is a text file and the fields are delimited by comma while the records are delimited by carriage return. The data collected is from August 21 2009 to August 20 2010.

Date	Ticker	Open	High	Low	Close	Volume
20090821	A	25.6	25.61	25.22	25.55	34758
20090824	A	25.64	25.74	25.33	25.5	22247
20090825	A	25.5	25.7	25.225	25.34	30891
20090826	A	25.32	25.6425	25.145	25.48	33334
20090827	A	25.5	25.57	25.23	25.54	70176
20090828	A	25.67	26.05	25.63	25.83	39694
20090831	A	25.45	25.74	25.31	25.68	51064
20090901	A	25.51	26.33	25.48	25.85	66422
20090902	A	25.97	25.97	24.96	25.22	64614
20090903	A	25.47	25.54	25	25.29	46369
20090904	A	25.37	25.92	25.1475	25.86	32556
20090909	A	26.31	27.19	26.16	27.15	36764
20090910	A	27.08	27.88	26.94	27.86	42987
...	...	...	...	...	...	...
20100806	ZMH	54.08	54.3	53.26	53.98	15890
20100809	ZMH	54.39	54.49	53.72	53.99	12170
20100810	ZMH	53.61	54.4	53.29	53.9	21266
20100811	ZMH	53.21	53.21	51.89	52.01	33017
20100812	ZMH	51.45	52.32	51.31	52	28473
20100813	ZMH	51.72	51.9	51.38	51.44	14561
20100816	ZMH	51.13	51.47	50.6	51	13489
20100817	ZMH	51.14	51.6	50.89	51.21	20498
20100819	ZMH	51.63	51.63	50.17	50.22	18259
20100820	ZMH	50.03	50.55	49.48	49.82	17792

Table 1: Shows an Excerpt of S&P 500 Stock Dataset

As shown in Table 1, the instances with 2009 stock data through 2010 stock data are presented. However, it is an excerpt from original dataset collected from [25]. It has more than 100 tickers for which transactions are available across the 2 years period. This dataset is considered suitable for finding inter-transaction relationships for finding infrequent items and obtaining relationship between them through a mediator.

B. Empirical Results

This section provides results of the two experiments made for finding indirect relationships among tickers in the given dataset. The part of data used for first experiment is shown in Table 2.

Day	INTU	IFF	IGT
20090922	29	27	20
20090923	30	30	20
20090924	27	28	24
20090925	25	24	22
20090926	28	25	25

Table 2: Shows Input Transactions for Experiment 1

As shown in Table 2, three tickers are used for empirical study. They are INTU, IFF and IGT. In this we consider only open price of each Ticker. These are the stock symbols for which indirect associations are extracted and the results are presented in Table 3.

A	B	Mediator	Support		Confidence	
			(a, mediator)	(b,mediator)	(a, mediator)	(b,mediator)
25	24	20	1	1	0.5	1.0
27	25	20	2	1	1.0	1.0
28	27	22	1	1	0.5	0.5
29	28	24	1	1	1.0	0.5
30	29	25	1	2	0.5	1.0

Table 3: Results of Indirect Relationships for Experiment 1

The item pair a and b values along with mediator are presented. The support for a and mediator and b and mediator is also shown. In the same fashion, there is confidence for a and mediator and then b and mediator as presented in the table.

Day	M	LUK	LTD
20090922	15	25	15
20090923	14	26	16
20090924	16	24	14
20090925	19	27	17
20090926	18	28	20

Table 4: Input Transactions for Experiment 2

The second experiment is based on the three ticker values extracted from S&P 500 stock market dataset. The tickers considered for this experiment are M, LUK and LTD. In this we consider only open price of each Ticker. The results of indirect associations between item pairs are presented in Table 5.

A	**B**	**Mediator**	**Support**		**Confidence**	
			(a, mediator)	**(b, mediator)**	**(a, mediator)**	**(b, mediator)**
14	24	14	2	1	1.0	1.0
15	25	15	1	1	1.0	0.5
16	26	16	2	1	1.0	1.0
18	27	17	1	1	1.0	1.0
19	28	20	1	1	1.0	1.0

Table 5: Results of Experiment 2 with Indirect Associations through Mediator

As shown in Table 5, is evident that for every item pair a and b, the mediator is identified and the support and confidence for a and mediator, b and mediator are presented. The results reveal the relationship among stock symbols considered for the empirical study.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a framework for finding causal relationships in stock transactions. Inter-transactions relationships in stock dataset provide latent trends that can help in making strategic decisions. The causal relationships are in the form of direct, indirect and exception relationships among stock performances across different tickers. Success in stock trading depends on comprehensive knowhow on the trends of trading. Many techniques came into existence to discover hidden trends in stock market data. Data mining techniques have been around to obtain actionable knowledge from real world datasets. In this paper we implemented a framework with an algorithm to find out indirect relationships among stock transactions. Especially the related businesses can have indirect relationships. Two stock tickers with no direct relationship are related through third ticker. This kind of relationship can provide valuable business intelligence. We proposed an algorithm that finds causal relationships in the form of indirect relationships. The empirical results revealed that the proposed algorithm is useful in analysing causal relationships. The research in this paper is limited to finding indirect relationships. We intend to extend it further to study direct and exception relationships besides taking unexpected incidents into consideration for finding causal relationships.

REFERENCES

[1] G. Preethi, B. Santhi, “Stock Market Forecasting Techniques: A Survey,” Journal of Theoretical and Applied Information Technology. 46, p1-7, 2012.

[2] Qasem A. Al-Radaideh, Adel Abu Assaf And Eman Alnagi, “Predicting Stock Prices Using Data Mining Techniques,” The International Arab Conference on Information Technology. 2, p1-4, 2013.

[3] Qian Wan and Aijun an, “An Efficient Approach to Mining Indirect Associations,” computer science p1-26, 2002.

[4] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava, “Indirect Association: Mining Higher Order Dependencies in Data,” springer. P632-637, 2000.

[5] Y.L. Hsieh, Don-Lin Yang and Jung pin Wu, “Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market,” computer science p1-4, 2002.

[6] Tien Thanh Vu,Shu Chang,Quang Thuy Ha and Ni gel Collier, “An Experiment in Integrating Sentiment Features for Tech Stock Prediction in Twitter,” Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, p23-38, 2012.

[7] Savinderjit Kaur And Veenu Mangat, “Applications Of Data Mining In Stock Market,” Journal of Information and Operations Management. 3, p1-3, 2012.

[8] Feifei Xu, “Data Mining in Social Media for Stock Market Prediction,” Dalhousie University Halifax, Nova Scotia, p1-84, 2012.

[9] Chaitanya Korgaonkar, “Analysis of the impact of financial development on Foreign Direct Investment: A Data Mining Approach,” Journal of Economics and Sustainable Development. 3 (6), p1-10, 2012.

[10] Archana Kumari, Umesh Prasad and Pradip Kumar Bala, “Retail Forecasting using Neural Network and Data Mining Technique: A Review and Reflection,” International Journal of Emerging Trends & Technology in Computer Science. 2 (6), p1-4, 2013.

[11] Thi-Thiet Pham, Jiawei Luo, Tzung-Pei Hong and Bay Vo, “An Efficient Algorithm For Mining Sequential Rules With Interestingness Measures,” International Journal of Innovative Computing, Information and Control. 9 (12), p1-14, 2013.

[12] Yigitcan Karabulut, “Can Facebook Predict Stock Market Activity,” Goethe University Frankfurt, p1-58, 2013.

[13] Dr.D.Ezhilmaran and S.Prasanna, “An analysis on Stock Market Prediction using Data Mining Techniques,” International Journal of Computer Science & Engineering Technology. 4, p1-3, 2013.

[14] Debashish Das and Mohammad Shorif Uddin, “Data Mining and Neural Network T Techniques in Stock Market Prediction: A Methodological Review,” International Journal of Artificial Intelligence & Applications. 4, p1-11, 2013.

[15]Longbing Cao, “Non-IIDness Learning in Behavioral and Social Data,” The Computer Journal Advance Access published August, p1-13, 2013.

[16] Nicholas Evangelopoulos, Michael J Magro and Anna Sidorova, “The Dual Micro/Macro Informing Role of Social Network Sites: Can Twitter Macro Messages Help Predict Stock Prices,” Informing Science: the International Journal of an Emerging Transdiscipline. 15, p1-22, 2012.

[17] Kainaz Bomi Sherdiwala, “Data Mining Techniques in Stock Market” Indian Journal of Applied Research. 4 (8), p1-3, 2014.

[18] Ashish Kumar and Dr. Kavita Choudhary, “A survey: Data Mining System for Mobile Devices,” International Journal of Enhanced Research in Science Technology & Engineering. 3 (2), p293-300, 2014.

[19] Viktorija Kuisyte, “Is the Baltic Stock Market Efficient,” Norwegian University of Life Sciences NMBU School of Economics and Business, p1-78, 2014.

[20] Ruchi Desai, Prof.Snehal Gandhi, “Stock Market Prediction Using Data Mining,” International Journal of Engineering Development and Research. 2, p2-5, 2014.

[21] Ishani Gangly and Michael Busch, “STOMpy: Stock Market data mining using Python,” The International School Bangalore. 3, p1-4, 2014.

[22] Dr. Rahul G. Thakkar, Mr. Vimal Patel, Dr. Manish Kayasth, “Model to Predict Stock Price with Respect to Day of the Week,” International Journal of Advanced Research in Computer Science and Software Engineering. 4 (11), p1-5, 2014.

[23] G. S. Navale Nishant Dudhwala Kunal Jadhav, “Prediction of Stock Market using Data Mining and Artificial Intelligence,” International Journal of Computer Applications. 134, p1-3, 2016.

[24] Dr. Swapna Borde, Austrin F. Dabre, Harsh M. Kamdar and Rahul Y. Purohit, “Financial Stock Price Forecast Using Classification,” International Journal of Advanced Research in Computer and Communication Engineering. 5 (4), p1-4, 2016.

[25] Historical Data, “Historical Data for S&P 500 Stocks,” Retrieved from http://pages.swcp.com/stocks/ , 2016.