Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market

Y.L. Hsieh[1], Don-Lin Yang[1], and Jungpin Wu[2]

[1]Department of Information Engineering and Computer Science
[2]Department of Statistics
Feng Chia University, Taichung, Taiwan 407

Abstract

To understand the causal relationship of stock market is always a top priority for investors. Most investors use some fundamental knowledge and basic analysis techniques to analyze or predict the trends. However, there are always some other factors beyond our control or unexpected events that might affect the stock market one way or the other.

After working on data mining with good results, we found inter-transaction mining can help answer the above questions in a systemic way. Our experiments show that causal relationship between upstream and downstream stocks do exist. To simplify our discussion, we focus on the electrical industrial stocks.

Keywords: **Stock, Inter-Transaction Data Mining**

1. Introduction

Every investor wants to know or predict the trends of the stock trading. Some of the investors use fundamental data analysis, such as company's annual report to predict if there is any potential profit in its future stock trading. Others use the technical approach, such as various short-term, middle-term, and long-term indicators to decide when to buy or sell stocks. However, there are other factors like economic forecast, government policy and unexpected incidents that might affect the stock market.

In the recent years, data mining techniques started to appear and derive some useful applications. One of them is to use neural network to cluster stock data in a fixed period of time and then to predict stock trends in the future [4]. Another popular application is using Web mining to search economic news for predicting the index trends the next day [3].

However, the results are not satisfactory. There are still many improvements can be made.

2. Our Approach

There are several data mining methods to find interesting rules, such as association rule and sequential pattern mining. In the association rule, one of the famous applications is MBA (Market Basket Analysis). In retailer business, association rule can be used to analyze customer's consumption behaviors and find patterns of buying habits. If some customers like to buy both A and B at the store, the owner can put A and B together to increase business volume. However, using association rules to predict stock market can only find stocks A and B might go up or down at the same day. It does not work for investors looking for future investment.

In the sequential pattern method, researchers find sequential patterns in the database which record the user's Web browsing sequence. The sequential pattern was used to help Web viewers match their needs quickly. It means past user behaviors can be used to predict future behaviors. Although the stock trading records can be taken as a group of investor's behaviors, sequential pattern does not include the time interval dimension. Using sequential pattern on the stock market, investors won't know when to buy or sell.

We choose inter-transaction mining [1] to solve above problems. The algorithm presents the rule as: if the price of IBM goes up, Microsoft's will most likely (80% of time) go up the next day [2]. In the real world, investors have a great interest in the relationship between upstream and downstream companies in the stock market.

In the past decades, one of the hot industries is technology. Some of the famous technology companies perform very well with good profits. In Taiwan's economy, technology industry plays an important role. Hence, our research goal is to find the causal relationship of stocks between upstream and downstream companies in the industrial supply chain.

2.3. Data Source and Attributes

Our source data came from the Taiwan Economic Journal (TEJ). The data attributes include opening price, highest price, lowest price and closing price for every trading day from 1971. All of stock trading data have been adjusted. The purpose was to avoid confusion. When stocks were Ex-Rights and Ex-Dividend, the prices is getting cheaper. Stock prices adjustment includes Ex-Rights, Ex-Dividend, capital reduction and stock split.

2.4. Stock Selection

The structure of the electrical industry in Taiwan's stock market can be presented in Figure 1. Table 1 shows some famous stocks we selected from the electrical industry.

If our research can find the causal relationship between Taiwan major index and electrical group index, we may help investors predict the index trend. Then, use it to call or put the index option as well as to buy or sell future contracts.

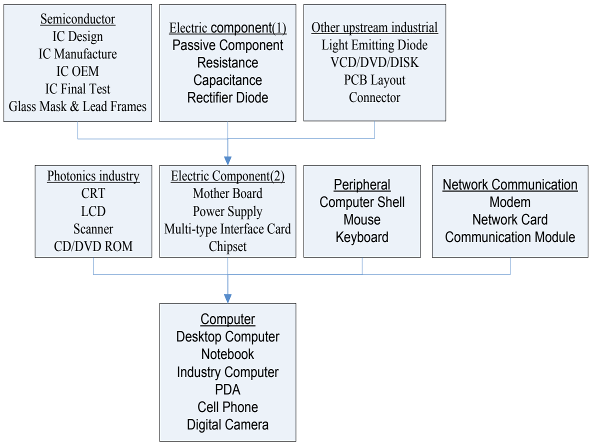


Figure 1 The structure of the electrical industry stocks

In order to verify the casual relationship between selected stocks and the other group of stocks, we add a stock from finance group and a stock from plastic and chemical group in our experiment. All of the selected stocks were shown in Table 1.

Table 1 Our selected stocks

Group	Stocks
Semiconductor industry	A01, A02
Electric component	B01, B02
Photonics industry	C01, C02
Electric Component (2)	D01, D02
Network Communication	E01, E02
Computer	F01, F02
Others	G01, G02
Finance Group	H01
Plastic and Chemical Group	I01

In Figure 1, we can see that semiconductor and electric component (2) have a direct relationship. It means semiconductor is a supplier of electric component (2). Semiconductor and other upstream industry have an indirect relationship. It means they all are members of electrical industry, but not direct supply relationship. Therefore, the relationship among the electrical industry group, finance group and plastic and chemical group is an exception relationship.

2.5. Preprocess the Selected Stock Trading Data

The source trading data include opening price and closing price. Input data must meet the requirement of our algorithm. Source data were transformed to indicate the status of going up or down of the stock prices in the respective trading days.

The transformed result is shown in Table 2. Because the algorithm can only process either up or down index, every stock must be encoded as shown in Table 3. Weekends have no trading data and are not counted. In Table 2, TID stands for Transaction ID and Itemset is a set of stocks whose prices are up or down in that day. For example, TID 1 represents the up stocks of TSMC, UMC and Index. If a trading day does not have any up or down stocks, Itemset column will remain blank. For example, TIDs 3, 4 and 8 do not have stock data of up or down.

Table 2 Transformed data

Table 3 Stock price encoding

To measure the up and down of stock prices, we define the stock UP/DOWN Percentage as:

UDP=(Closing Price – Opening Price) / Opening Price

When *UDP* > 0.005, it means the stock is up.

When *UDP* < -0.005, it means the stock is down.

For all the other cases, it means a tie. Investors are not interested in a tie. So, we remove the tie from our transformed data records as in Table 2.

After completing the data mining process, the resultant rules look as follows:

Rule1：If stock A goes up, *p*% of time stock B will go up after *d* days.

Rule2：If stock A goes up, q% of time stock B will go down after e days.

Rule3：If stock A goes down, r% of time stock B will go up after f days.

Rule4：If stock A goes down, s% of time stock B will go down after g days.

3. Mining Inter-transaction Rules

First, we define a sliding window m which is the maximum number of days the investors are interested in trading. The mining algorithm will find the rules between one and m days. The value of m must be greater than 1. Here, we use the value of five days for m which means a week.

3.1. More Definitions

Before finding interesting rules from stock trading database, there are some definitions need to be made. What rules are interesting to us? How to measure the interestingness? In a database, the frequency of a pattern is called the *support*. When the pattern's *support* value is greater than a user defined threshold, we know the user is interested in that pattern which is called a rule. The formula of *support* is $PROB(A \cup B)$.

As mentioned above, the pattern is not equal to the rule. Looking at Rules 1~4, we know that they involve the conditional probability. Users need to define a second value to represent the threshold of conditional probability as well. The threshold is called *confidence* in data mining. The formula is $PROB(B \mid A)$. It is also a user provided parameter. The inter-transaction mining will find out the rules based on the defined *confidence*.

The pattern is not meaningful for investors when stock A goes up and there is no other stock appearing after n days. Inter-transaction mining will remove this kind of patterns.

With a sliding window of m = 3, Table 4 presents an example of the inter-transaction concept by using the data from Table 2. Each transaction in Table 4 consists of MID and COL-n ($1 \le n \le m$). MID stands for the mega-transaction identifier. COL-n (n=1~3) contain the data of each mega-transaction.

Table 4 Mega-transactions

Let us define the up and down of stock trading database as D (Table 2). By observation, we find an interesting lemma from COL-1~COL-n in Table 4: $D \supseteq COL1$, $COL1 \supseteq COL2$, $COL1 \supseteq COL3$. With this interesting lemma, we are able to implement our mining algorithm more efficiently. The details are too long to describe here.

There is another interesting discovery. Given S_w that contains the mining result depending on the sliding window w (w>1), we found that the resultant rules of S_{w-1} are included in S_w. The lemma can be written as following:

$$Sn \subseteq Sm \,, (n \le m \,,\ n,m \in w).$$

According to this lemma, investors don't need to mine the rules of the sliding window n if they already have the rules from the sliding window m where n < m.

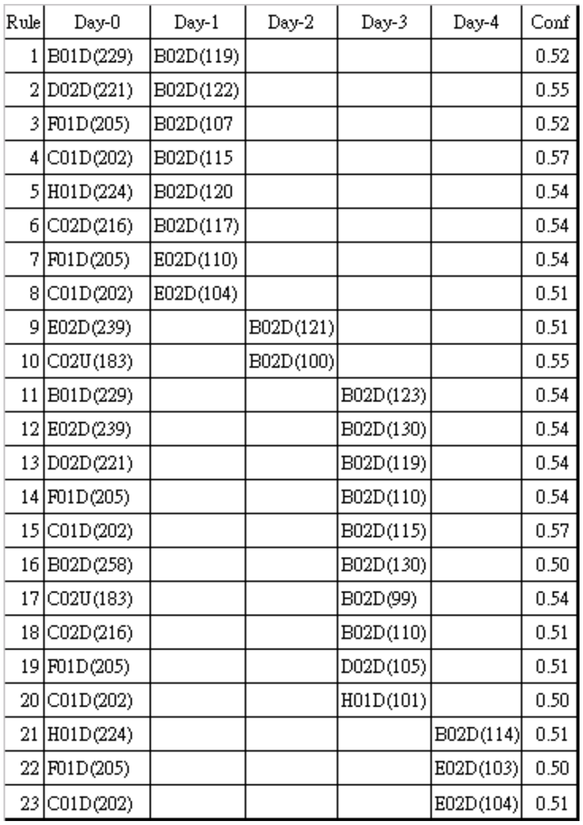
3.2. Experiments and Result Discussion

In our first experiment, we select the trading period between year 2004 and 2005. Using data preprocess to build the complete tables like Table 2 and Table 3. There are 16 stocks selected with 497 transactions between year 2004 and 2005. The following parameters are used in our experiment: sliding window = 5，support = 20%, Confidence = 50%

The first mining experiment took 14 seconds and the result was shown in Table 5. All of the resultant rules met our expectation. In the Rule 1 of Table 5, we have B01D(229) in Day-0 where B01 is the stock ID and D means that it's a downward stock with the number 229 in the parentheses indicating the frequency in the source data. The Rule1 can be explained as: when the stock B01 goes down, B02 will go down the next day and the rule's confidence is 52%. There are 17 rules in Table 5 showing all the stocks that will influence B02 to go down within a window of 5 days. We realize that the company of B02 produces passive component and is the upstream of electrical industrial. Therefore, when downstream stock prices go down, B02 stock price will go down later as a result. There are two exception relationships in the result of mining: Rule 5 and Rule 21. The other rules are both direct and indirect relationship. Since their confidence values are only between 50% and 57%, the result is not very interesting.

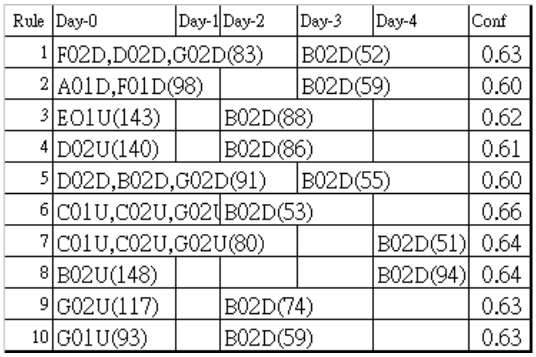
In the second experiment, we lower the support threshold value to 10% and raise the confidence threshold value to 60%. As a result, we found 10 rules as shown in Table 6. Here the confidence of mining results has been raised to 60% ~ 66%. In Day-0 of Rule 1 in Table 6, there are 3 items forming a pattern and the frequency is 83 times. The other rules are also interesting result that can influence stock B02's price to go down in this experiment.

Table 5 First mining result.



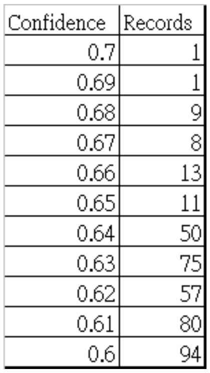
Rule	Day-0	Day-1	Day-2	Day-3	Day-4	Conf
1	B01D(229)	B02D(119)				0.52
2	D02D(221)	B02D(122)				0.55
3	F01D(205)	B02D(107				0.52
4	C01D(202)	B02D(115				0.57
5	H01D(224)	B02D(120				0.54
6	C02D(216)	B02D(117)				0.54
7	F01D(205)	E02D(110)				0.54
8	C01D(202)	E02D(104)				0.51
9	E02D(239)		B02D(121)			0.51
10	C02U(183)		B02D(100)			0.55
11	B01D(229)			B02D(123)		0.54
12	E02D(239)			B02D(130)		0.54
13	D02D(221)			B02D(119)		0.54
14	F01D(205)			B02D(110)		0.54
15	C01D(202)			B02D(115)		0.57
16	B02D(258)			B02D(130)		0.50
17	C02U(183)			B02D(99)		0.54
18	C02D(216)			B02D(110)		0.51
19	F01D(205)			D02D(105)		0.51
20	C01D(202)			H01D(101)		0.50
21	H01D(224)				B02D(114)	0.51
22	F01D(205)				E02D(103)	0.50
23	C01D(202)				E02D(104)	0.51

Table 6 Second mining result with higher confidences.



Rule	Day-0	Day-1	Day-2	Day-3	Day-4	Conf
1	F02D,D02D,G02D(83)			B02D(52)		0.63
2	A01D,F01D(98)			B02D(59)		0.60
3	EO1U(143)		B02D(88)			0.62
4	D02U(140)		B02D(86)			0.61
5	D02D,B02D,G02D(91)			B02D(55)		0.60
6	C01U,C02U,G02U		B02D(53)			0.66
7	C01U,C02U,G02U(80)				B02D(51)	0.64
8	B02U(148)				B02D(94)	0.64
9	G02U(117)		B02D(74)			0.63
10	G01U(93)		B02D(59)			0.63

Table 7 Third mining result.



Confidence	Records
0.7	1
0.69	1
0.68	9
0.67	8
0.66	13
0.65	11
0.64	50
0.63	75
0.62	57
0.61	80
0.6	94

In the final experiment we simply lower support threshold value further to 5% hoping to find more high-confidence rules. The result has 399 rules. They are displayed in Table 7 with a simpler form showing the number of records in terms of the confidence levels. The highest confidence was up to about 70%. As expected, the time is longer than the last two.

4. Conclusion and Future Work

Based on our experiment results, there are causal relationships in the selected stocks. They are mixed with direct, indirect and exception relationships. Obviously, the relationships exist in some specific stocks, but not all of stocks have the relationship which investors are interested in. Even with the use of synthetic data, we are sure to find the casual relationships and their time sequence. Especially, there is no need to assume any specific stock's price going up or down in order to find the stocks whose prices go up or down several days later.

In [3] they used Web data to perform stock market forecast with an accuracy rate of 46.7%. Using our method the best prediction can reach 70% while a random approach is about 33%.

Although useful rules can be found by using inter-transaction mining, they are not real return of investment (ROI) for investors. In addition to help in making investment strategies, further study is required to improve the usage of our method and extend the working of causal relationship chains.

Acknowledgement: This research was partially supported by the National Science Council, Taiwan, under grant number NSC94-2213-E-035-002.

5. Reference

[1] Anthony K.H. Tung, Hongjun Lu, Jiawei Han, and Ling Feng, "Efficient Mining of Inter-transaction Association Rules," IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No.1, Jan/Feb 2003.

[2] H. Lu, J. Han, and L. Feng, "Stock Movement and n-Dimensional Intertransaction Association Rules," Proc. 1998 SIGMOD Workshop Research Issues on Data Mining and Knowledge Discovery, Vol. 12, pp. 1-7, June 1998.

[3] B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, W. Lam, "Daily Stock Market Forecast from Textual Web Data," IEEE International Conf. on Systems, Man, and Cybernetics, V.3, pp. 2720-2725, 1998.

[4] S.Y. Zeng, C.H. Hsueh, S.D. Lee, "Using SOM to Study the Decision Making on Stock Investment," (in Chinese) Proceedings of the Seventh Conference on Artificial Intelligence and Applications (TAAI2002).