An Association Mining Method for Time Series and Its Application in the Stock Prices of TFT-LCD Industry

Chiung-Fen Huang, Yen-Chu Chen, and An-Pin Chen

Institute of Information Management, National Chiao Tung University, Hsinchu, Taiwan
No. 1001, Dashiue Rd., Hsinchu, Taiwan 300, R.O.C.
huangcf@ms60.url.com.tw; yunchu@ms6.hinet.net

Abstract. TFT-LCD is one of industries currently promoted by the "Two Trillion and Twin Star Industries Development Plan" in Taiwan. This study endeavors to find out the stock price associations between the suppliers and manufacturers in the value chain of the TFT-LCD industry by means of data mining techniques, and meanwhile, to improve the Apriori algorithm so that it can facilitate association mining of discrete data points in a time series. An efficient data mining method which consists of two phases is proposed. In the first phase, data are classified and preprocessed using the algorithm proposed by R. Agrawal et al. (1996), then Apriori algorithm is applied to extract the strong association rules. The second phase further improves the Apriori algorithm by breaking down the traditional limitation of relying on pattern matching of continuous data for disclosing stock market behavior. By mining the association rules from the discrete data points in a time series and testing the corresponding hypotheses, statistically significant outcomes can be obtained. The proposed data mining method was applied to some real time-series of the stock prices of companies in the supply chain of TFT-LCD industry in Taiwan. It is suggested that a positive correlation does not necessarily exist between the companies' stock prices in the supply chain of TFT-LCD industry. For instance the result shows that, if the stock price of Sintek Phonrotic Corp., a company in the up stream of the value chain, soars for more than 5% in a day, the stock price of Tatung, a company in the down stream of the same value chain, may not respond positively accordingly. If an investor can short the stock of Tatung on the 7th day and long it back on the 10th day after Sintek's stock price soaring for more than 5%, the annual return of investment is 199.88% with 95% confidence interval. In conclusion, the results may reveal helpful information for the investors to make leveraged arbitrage profit investing decisions, and it might be interesting to apply this proposed data mining method to the time series in other industries or problems and investigate the results further.

Keywords: Data Mining, Association rule, Apriori Algorithm, TFT-LCD, Time series analysis.

1 Introduction

While more and more data generated in the form of time-series, there are much more needs to find frequent patterns in time-series data. Time-series data mining becomes

P. Perner (Ed.): ICDM 2004, LNAI 3275, pp. 117–126, 2004.

more and more popular in recent research areas and has broad applications like analysis of customer purchase patterns, web traversal patterns, etc.

Take the example of stock price fluctuation of TWSE. There may be some implications that when the stock price of some companies in supply chain went upwards, the stock price of the other companies in the related industry might also be affected. It is interesting to find the relationships among the stock fluctuation patterns. Learning from the association rules within these patterns might help the investors making decisions more precisely.

In this study, we try to find out the stock price association for the entire value chain in the TFT-LCD industry by means of data mining techniques, and, at the same time, to improve the Apriori algorithm so that it could be applied to the cross-day discrete discontinuous data. Two-phase experiment associated with the industry was conducted. In the first phase, data were prepared using the algorithm proposed by R. Agrawal et (1996), then Apriori algorithm is used to compute the support and confidence value. The strong association rules among the data are found accordingly. The second phase is to improve the Apriori algorithm. The idea is to break down the traditional limitation of continuous pattern when considering the stock market behavior. The results of this study reveals that a positive relationship does not necessarily exist between the companies' stock price in the supply chain of TFT-LCD industry. By using the idea, investors may make leveraged arbitrage profit investing decisions more precisely.

2 About TFT-LCD

Taiwan's flat-panel display industry developed out of the TN/STN LCD industry and expanded to the production and processing of TFT-LCD (Thin Film Transistor-Liquid Crystal Display) and upstream components, such as color filters, polarizers, backlights, and glass products. With large demand from local notebook PC manufacturers and technical advantages borrowed from the semiconductor industry, Taiwan's TFT-LCD industry has achieved its current strong position in less than five years of development. In 2001, Taiwan's TFT-LCD production reached NT$123.4 billion (US$3.65 billion), ranking third in the world and accounting for 26.1 percent of the world's total output.

With the commencement of the "Two Trillion and Twin Star" program, measures are being taken to offer more incentives to businesses, resolve patent-related problems, and strengthen R&D and training of new talents. These actions will not only help free Taiwan's TFT-LCD industry from technological dependence and its current shortage of skilled staff, but also give it a more completely integrated structure. Through the cooperation of government and industry, the TFT-LCD industry is expected to achieve a value of NT$1.37 trillion (US$40.53 billion) by 2006, making Taiwan the largest TFT-LCD supplier in the world.

3 Data Mining

3.1 Data Preprocessing

The task of data preparation is to transform the transactions in the database into the format we need. For example, in order to find the association rules of the fluctuations

for daily stock prices, we need to transform the daily data of stocks into daily fluctuation rates.

There are three steps in the phase of data preparation. First we compute the fluctuation for each attribute values, to get rid of noisy data. Second, we transform the attribute values in the transaction, in the form of fluctuation. Then, we classify the fluctuation rates into several classes and label them as defined before.

3.2 Modified Apriori Algorithm

The first algorithm of mining association rules, called Apriori algorithm, was proposed by Agrawal and Srikant in 1994[15]. The methodology is so important in frequent pattern mining that there are many approaches adapting from it.

The Apriori algorithm employs iterative, level-wise approach, where frequent k-itemsets are deduced from (k-1)-itemsets. Let Lk be the frequent k-items, and Ck bethe candidate k-itemsets. In the first pass, it counts the support of data items to determine the frequent 1-itemsets. Then, it uses frequent itemsets Lk-1 found in the (k-1)th pass to generate the candidate items Ck and then scan the database, D, to counts the supports of candidates in Ck, Let Lk be the itemsets in Ck with their supports no less then the minimum support.[12]. In order to fit our experimental model, we modify Apriori algorithm and it is described in Figure3.1.

```
Input: Database, D, of transactions; minimum support
threshold, min_sup.
Output: L, frequent itemsets in D.
Method:
  Dim array [x, y]
  For (day = 1; EOF; day ++) {
     Data_preprocess;
     C_k = apriori_gen(L_k-1, min_sup);
     Gen_report;
  }
  L_1= find_frequent_1-itemsets(D);
  For (k=2; L_k-1 ≠∅ ;k++) {
     C_k = apriori_gen(L_k-1, min_sup);
     For each transaction t ∈ D{//scan D for counts
        C_t = subset(C_k,t);
             //get the subsets of t that are candidates
        For each candidate c ∈ C_t
            c.count++;
      }
      L_k = {c ∈ C_k|c.count ≥  min_sup}
  }
return L= ∪ _kL
Procedure apriori_gen (L_k-1:frequent(k-1)-items; min_sup:
minimum support threshold)
  For each itemset l_1 ∈ L_k-1
     For each itemset l_2 ∈ L_k-1
        If (l_1[1] = l_2[1] ) ∧ (l_1[2] = l_2[2] ) ∧…∧ (l_1[k-
2] = l_2[k-2] ) ∧ (l_1[k-1] < l_2[k-1] ) then {
```

```
            C=  l1 ∪ l2 ;//join step:generate candidates
            If has_infrequent_subset(c,Lk-1) then
              Delete c; // prune step:remove unfruitful
candidate
            Else add c to Ck;
          }
  return Ck;
Procedure has_infrequent_subset (d:candidate k-
itemset;Lk-1:frequent (k-1)-itemsets); //use prior
knowledge
  For each (k-1)-subset s of c
     If s ∉Lk-1 then
        Return TRUE;
  Return FALSE;
```

Fig. 3.1. Modified Apriori algorithm

3.3 Interested Measures

Although a data mining process may generate a large number of patterns, typically only a small fraction of these patterns will actually be of interest to the given user. Thus, users need to further confine the number of uninteresting patterns returned by the process. This can be achieved by specifying the interested measures that estimate the simplicity, certainty, utility, and novelty of patterns. In this paper, we use novelty to test the measures of pattern interest.

Novel patterns are those that contribute new information or increased performance to the given pattern set. For example, a data exception may be considered novel in that it differs from that expected based on a statistical model to user beliefs. Another strategy for detecting novelty is to remove redundant patterns. If a discovered rule can be implied by another rule that is already in the knowledge base or in the derived rule set, then either rule should be reexamined in order to remove the potential redundancy.

4 Experiment

4.1 Data Preparation

This research is base on the closing stock price of the TFT-LCD industry supply chain related corporations. Considering the date of Optimax Co.(3051)'s stock going public, this study use the stock closing price from Oct. 28th,2002 to Jan. 15th 2004 as the basis of the experimentation data.

Eight companies were selected for the experiment. These corporations are all possess a supply chain relationship with CPT(Chunghwa Picture Tubes, LTD) TFT manufacturing. There are three companies in the upstream and four in the downstream of the supply chain.

In order to have a macroscopic concept for the study, all the data are divided into two categories. One is the whole data of 8 corporations. The other is to select one corporation from each of the three stream levels in the TFT-LCD industry supply

chain: the upstream - supplier level, midstream - manufacturing level, and downstream - client level, accordingly. There will be twelve different combinations. These combinations of data are classified further.

Data need to be partitioned in order to quantify the quantitative attributes [19]. Here, the stock price movement of an attribute value is equally parted into several classes, as shown in Table 4.1. Each item of the data is represented by 5 digits, the first four digits stand for the stock number, and the fifth digit, A to G, represents the class that the item's stock price movement falls to.

Table 4.1

	Class 3	Class 4	Class 5	Class 6	Class 7
A	7 ~ 2.33	7 ~ 3.5	7 ~ 4.2	7~ 4.67	7 ~ 5
B	2.3 ~ -2.3	3.5 ~ 0	4.2 ~ 1.4	4.6 ~ 2.3	5 ~ 3
C	-2.33 ~ -7	0~ -3.5	1.4 ~ -1.4	2.34 ~ 0	3~ 1
D		-3.5 ~ -7	-1.4 ~ -4.2	0 ~ -2.3	1~ -1
E			-4.2 ~ -7	-2.3 ~ -4.6	-1 ~ -3
F				-4.7 ~ -7	-3 ~ -5
G					> -7

(%)

4.2 Experiment Result

At the first phase of the experiment, seven strong association rules are found in the group consisting of all the eight companies (Table 4.2), six of the rules are left after applying Novelty test. The result reveals that there exists a stronger association, more than 66%, between the midstream and downstream in the TFT-LCD industry supply chain. By contrast, in the group consisting of data from each supply chain layer, two association rules of each in class 6 and class 7 are found. The findings are quite interesting and worth a further study.

Table 4.2

strong association rules		
2301B <- 2442B 2475B	(42.7%, 85.2%)	3
2301C <- 2442C 2475C	(22.0%, 71.2%)	5
2301C <- 3034C 2475C	(20.3%, 70.5%)	5
2301D <- 2442D 2352D	(13.7%, 63.4%)	7
2301D <- 3049D	(33.3%, 56.0%)	6
2442B <- 2475B 2301B	(20.3%, 68.9%)	4
2442B <- 2475B 2371B	(21.7%, 66.2%)	4

In order to have a further study about these four strong association rules, and to find out a discontinuous relationship, we modify the Apriori algorithm so that it can be applied to time series data, and also to discover the strong association rules among the cross-day transactions.

Table 4.3

Class 7 strong Association Rules	
2352D <- 2475C	(14.3%, 53.5%)
2301D <- 3051E	(25.3%, 52.6%)
2301D <- 2475D	(30.0%, 50.0%)
2301D <- 3049A	(10.0%, 56.7%)
2301D <- 2475D 3034D	(11.7%, 54.3%)
2352D <- 3051A	(10.0%, 53.3%)

In the second phase experiment of this study, the modified Apriori algorithm is used to mine the transaction data from the next day to the eleventh day. The result reveals that there is less than 20% confidence level between itemsets 2442 and 2475, which are the midstream and downstream companies accordingly in the TFT-LCD industry supply chain, in the following ten transaction days. Yet, there are two interesting rules finding in class 7 after the cross-day mining. Table 4.4 an illustrate the outcome.

Table 4.4. 3049 A → 2301 D

Day	X < 1%	X < -1%
1	73.30%	23.30%
2	53%	30.00%
3	50.00%	23%
4	50.00%	10.00%
5	66.70%	40%
6	76.70%	20.00%
7	76.70%	37%
8	63.30%	23.30%
9	63%	23%
10	63.30%	50%

4.3 Investitive Application

The results from the above experiment could help the investors to make proper financial decisions and make leveraged arbitrage profit. By applying the strong association rules revealed in section 4.3, the investor could see the relationship of the stock prices between the upstream and downstream companies in the supply chain of TFT-LCD industry, and make the proper investment decisions. For example, if the investor uses the rule 3049A-> 2301D to invest on Taiwan stock market, he might watch the trend of the stock market, and take some actions as follows:

1. Wait for the stock price of Sintek Co.(3049), the upstream company in the TFT-LCD industry, raises for over 5%.
2. Short sell the stock of Tatung, the downstream company in the TFT-LCD industry, on the 6th or 7th day after Sintek's stock price soaring for more than 5%.

3. Long the stock of Tatung back on the 10th day. If the investors do take such action by applying the association rule, as described above, the return on investment will be significant.

Table 4.5 shows the average transaction return and annual return on investment of the investment example.

Table 4.5

invest launch day	Day 1st	Day 6th	Day 7th
Trans. return rate	3.07%	1.86%	1.64%
year return rate	112.13%	170.02%	199.88%

4.4 Statistical Test

Again, H_0 is rejected, meaning rule Day 7^{th} is better than Day 6^{th} .

By the result of statistical hypothesis test, rule Day 7^{th} is the best association rule in the three rules found in the second phase of the experiment. Therefore, in TFT-LCD industry, if one short sells the stock of the target company on the 7^{th} day and long it back on the 10^{th} day after the stock price of the upstream of the company soaring for more than 5% in one day, he would gain the greatest return on investment.

Table 4.6

	Day 1st	Day 6th	Day 7th
Sample num.	29	29	29
average	3.07	1.86	1.64
Stdev	10.44	6.67	6.08

(%)

$$H_0 : \mu_{6th} \le \mu_{1st}, \alpha = 0.05$$

$$Z^\circ = \frac{\overline{3.09 - 1.86}}{\sqrt{\frac{(10.44)^2}{29} + \frac{(6.67)^2}{29}}} = 5.25$$

$$H_0 : \mu_{6th} \le \mu_{1st}$$

So H0 is rejected, which implies that rule Day 6th is better than Day 1st.

Next, the same statistical test is performed to compare rule Day 6th and Day 7th.

$$H_0 : \mu_{7st} \le \mu_{6th}, \alpha = 0.05$$

$$Z^{\circ} = \frac{\overline{1.86 - 1.64}}{\sqrt{\frac{(6.08)^2}{29} + \frac{(6.67)^2}{29}}} = 14.2229$$

Again, H0 is rejected, meaning rule Day 7th is better than Day 6th.

By the result of statistical hypothesis test, rule Day 7th is the best association rule in the three rules found in the second phase of the experiment. Therefore, in TFT-LCD industry, if one short sells the stock of the target company on the 7th day and long it back on the 10th day after the stock price of the upstream of the company soaring for more than 5% in one day, he would gain the greatest return on investment.

5 Conclusion and Future Work

Data mining is a process of uncovering relationships, patterns and trends among huge volumes of data and then transforming them to valuable information that can leverage business intelligence and improve the process of making decisions [24]. A fundamental problem in data mining is finding frequent patterns in large datasets. The Apriori algorithm is one of several different algorithms that have been proposed to find all frequent patterns in a dataset, yet it is usually used on datasets containing continuous data. In this study, a modified Apriori algorithm is proposed to mine the association rule in time series data, and furthermore, to apply it to the cross-day discrete stock market data to find out some valuable information. It presents a data mining methodology to analyze discontinuous cross-day time-series data. The findings of this study might be helpful when making financial investment decisions:

1. In the same industry, there exists a stronger association for stock prices in vertical companies (companies in different streams in the supply chain) than in horizontal ones (companies in the same stream level in the supply chain).
2. A positive correlation does not necessarily exist between the companies' stock prices in upstream and downstream in the supply chain.
3. The study breaks down the traditional limitation of relying on pattern matching of continuous data in the same period of time. For example, when upstream company Sintek Co. (3049) has a stock price raising in one day, the stock price of downstream company Tatung Co. (2301) would decline for more than 1% with 50% confidence interval, in the 10^{th} day after.
4. Applying the association rules found in the experiment, if an investor can short sell the stock of Tatung on the 7^{th} day and long it back on the 10^{th} day after Sintek's stock price soaring for more than 5%, the annual return on investment is 199.88% with 95% confidence interval.

By means of data mining technology, the study is focused on finding the companies' stock price association in the supply chain of TFT-LCD industry. The proposed data mining method could be applied to the time series data in other industries or problems and investigate the results further. Besides data mining methodology, some artificial intelligence technology, such as fuzzy, genetic algorithm, or neural network, may facilitate revealing more helpful information for

the investors to make leveraged arbitrage profit investing decisions. It is hoped that this study will call people's attention to this opportunity.

References

1. Dan Braha and Armin Shmilovici, "Data mining for improving a cleaning process in the semiconductor industry", IEEE transactions on semiconductor manufacturing, vol. 15, No. 1, Feb 2002, 91-102
2. Michael Goebel and Le Gruenwald, "A survey of data mining and knowledge discovery software tools", SIGKDD Explorations. ACM SIGKDD. June 1999, Vol. 1. Issue 1, 20-33.
3. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery: An overview" in Advances in Knowledge Discovey and Data Mining, MIT Press, 1996, 1-36.
4. M. J. Berry and G. Linoff, Data Mining Techniques. New York:Wiley, 1997.
5. T. M. Mitchell, Machine Learning. New York: McGraw-Hill,1997
6. D. Braha, Ed., Data Mining for design and Manufacturing: Methods and Applications. Boston, MA: Kluwer Academic, 2001.
7. Goebel, M., and Gruenwals, L. A Survey of Knowledge Discovery and Data Mining Tools. Technical Report, University of Oklahoma, School of Computer Science, Norman, OK, Feb 1998.
8. Pawlak, Z. "Rough Sets: Theoretical Aspects of Reasoning About Data." Kluwer, Boston, 1991
9. Ron Kohavi and Foster Provost, "Applications of Data Mining to Electronic Commerce", Data mining and Knowledge discovery, Vol. 5,2001, 5-10.
10. Kamber, M., Han, J., and Chiaung, J. Y. "Using Data Cubes for Metarule-Guided Mining of Multi-Dimensional Association Rules." Technical Report U-Sfraser-CMPT-TR, 1997-10, Simon Fraser University, Burnaby, May 1997.
11. Lippmann, R.P. "An Introduction to Computing with Neural Nets." IEEE ASSP Magazine, Apr 1987, 4-22.
12. Hui Ching Han, "Mining Association Rules among Time-series Database", University of Taiwan, 2002.
13. J. Han, H. Lu, L. Feng, "Beyond Intra-transactional Association Analysis:Mining Multi-dimensiional Inter- transaction Association Rule," ACM Transactions on Information Systems, Vol. 18, No. 4, Oct 2000, 423-454
14. R.Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases, " Proc. ACM Special Interest Group on Management of Data, May 1993, 207-216
15. R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Database," Proc. 20th Int'l Conf. On very Large Databases, Santiage, Chile, Sep 1994, 478-499
16. J. Han and M. Kamber, "Data Mining: Concepts and Techniques ", Morgan Kaufmann, San Francisco, 2000.
17. A. Savesere, E. Omiecinski, and S. Navathe, "An Efficient Algorithm for Mining Association in Large Database," Proc. 21st Int'L Conf. On Very Large Databases, Zurich, Swizerland, 1995.
18. J. Han, J.Pei, and Y.Yin, "Mining Frequent Patterns without Candidate Generation," Proc. ACM Special Interest Group on Management Of Data, May 2000, 1-12

19. R. Srikant and R. Agrawal, “Mining Quantitative Association Rules in Large Relational Tables,” Proc. ACM Special Interest Group on Management Of Data, Vol. 25, No. 2, June 1996, 1-12
20. Sector review, “Global TFT-LCD Industry 2004-05: From tightness to excess, and back.”, Semiconductor Devices, Asia Pacitic / Taiwan, 3 March 2004.
21. Taiwan Stock Exchange Corporation, http://www.tse.com.tw, Jan 2004.
22. Industrial development bureau ministry of economic affairs, http://www.moeaidb.gov.tw/idy/index1.jsp
23. Kenneth Janda, “Univariate Statistics”, Northwestern University, http://www.janda.org
24. T.M. Mitchell, Machine learning and data mining, Communications of the ACM 33 (1990), 296-310