Domain-Driven Local Exceptional Pattern Mining for Detecting Stock Price Manipulation

Yuming Ou, Longbing Cao, Chao Luo, and Chengqi Zhang

Faculty of Information Technology, University of Technology, Sydney, Australia
{yuming,lbcao,chaoluo,chengqi}@it.uts.edu.au

Abstract. Recently, a new data mining methodology, Domain Driven Data Mining (D^3M), has been developed. On top of data-centered pattern mining, D^3M generally targets the actionable knowledge discovery under domain-specific circumstances. It strongly appreciates the involvement of domain intelligence in the whole process of data mining, and consequently leads to the deliverables that can satisfy business user needs and decision-making. Following the methodology of D^3M, this paper investigates local exceptional patterns in real-life microstructure stock data for detecting stock price manipulations. Different from existing pattern analysis mainly on interday data, we deal with tick-by-tick data. Our approach proposes new mechanisms for constructing microstructure order sequences by involving domain factors and business logics, and for measuring the interestingness of patterns from business concern perspective. Real-life data experiments on an exchange data demonstrate that the outcomes generated by following D^3M can satisfy business expectations and support business users to take actions for market surveillance.

1 Introduction

The traditional data mining is a data-driven trail-and-error process in which data is used to create and verify research innovations. Typically, it aims to build standard models to summarise training and test data well. However, the general patterns emerged from standard models is unactionable to business needs, and cannot support business users to take decision-making actions in the business world. This may be due to many reasons. One of key reasons is that the domain knowledge is underemphasised by the traditional data mining. For instance, in the area of stock market surveillance [1], the pattern that *if there is a sharp price change then an alert is generated to indicate the occurrence of price manipulation* often leads to too many false positive alerts. The reason for that is the pattern does not take the real-life factors into consideration and embed the business logics in stock markets. The simple ignorance of domain factors and pure data-centered pattern mining result in a big gap between academic deliverables and business expectations.

To deal with the above issues, Domain Driven Data Mining (D^3M) [2, 3, 4, 5] was recently proposed targeting actionable knowledge discovery for real user needs under domain-specific circumstances. It aims to narrow down the gap between academic deliverables and business expectations through catering for key issues surrounding

T.-B. Ho and Z.-H. Zhou (Eds.): PRICAI 2008, LNAI 5351, pp. 849–858, 2008.

real-world actionable knowledge discovery. (1) D^3M makes full use of both real-life data intelligence and domain intelligence, for instance, all real-life constraints [6] including domain constraints, data constraints and deliverable constraints during the whole mining process. (2) The resulting outcomes are evaluated by not only technical but also business interestingness metrics toward knowledge actionability [5]. D^3M has attracted more and more attention including workshops with KDD and ICDM.

In stock markets, the detection of intraday price manipulation is of great importance to market integrity. However, it is very challenging to effectively detect price manipulation in real markets. Based on D^3M, this paper carries out a study on mining actionable local exceptional patterns indicating price manipulation on intraday trading data for market surveillance. We fully employ domain knowledge through the mining process. We first define a five-dimension vector to represent trading orders based on domain knowledge. The order vector is designed specifically for the stock market domain and captures the characteristics of stock market properly. Furthermore, we develop two interestingness measures for pattern mining which also reflect the business concerns. We deploy our approach in mining real-life orderbook data, and evaluate the mined patterns in terms of business concerns, which shows the findings are deliverable to business users for market surveillance.

The remainder of this paper is organised as follows. Section 2 introduces the related work on the actionable knowledge discovery. In Section 3, the domain problem is carefully studied with the involvement of domain intelligence. Base on the analysis in Section 3, Section 4 proposes an approach to construct the domain-driven multidimensional sequences. To discover local exceptional patterns, two interestingness measures and an algorithm are designed for identifying such local exceptional patterns in Section 5. Experiments and performance evaluation are demonstrated in Section 6. We conclude this paper and present our future work in Section 7.

2 Related Work on Actionable Knowledge Discovery

Recently the actionable capability of discovered knowledge has been payed more and more attention [7, 8, 9], a typical work is on D^3M. D^3M aims to narrow down the gap between academia and business, and a paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge discovery to support decision making [4]. In D^3M, both data and domain knowledge make contributions to the outputs.

The concept of actionability was initially studied from the interestingness perspective [10, 11, 12, 13, 14] to filter out the redundant and 'explicit' patterns through the mining process or at the stage of post-processing [15]. A pattern is *actionable* if a user can get benefits from taking actions on it (e.g., *profit* [16, 17]). Particularly, subjective measures such as *unexpectedness* [13, 14], *actionability* [2, 9, 14, 15] and *novelty* [18] were studied to evaluate the actionable capability of a pattern. However these research efforts mainly focus on the development of general business interestingness measures. Due to the absence of domain-specific knowledge, the general business interestingness measures are often inefficient when they are applied to various domains. Thus, in order to enhance the actionable capability of discovered knowledge, a reasonable assumption is that the domain-specific intelligence should be involved in the whole mining process and the business interestingness measures should also be studied in

terms of specific domain problems. In this paper, we demonstrate the development of domain-specific interestingness to measure actionability.

3 Domain Problem Definition

In the stock market, one of typical illegal trading behaviour is price manipulation. Price manipulation is defined as the trading behaviour attempting to raise or lower the price of a security for the purpose of exceptionally high profit. Price manipulation can damage the market integrity and ruin the market reputation. Consequently, any market regulators in the world are keen on developing effective detection, combating and prevention tools for price manipulation.

3.1 Current Methods

To detect price manipulation, current methods mainly focus on the sharp price changes and/or large trade volumes. When a sharp price change or a large trade volume occurs in the market, an alert is triggered and then a further investigation may be carried out. However, these methods are far from working well. As the stock market is a complex system, there are too many factors which can affect the stock price and trade volume. For example, the disclosure of a really bad news for a certain company is likely to make the stock price of that company go down dramatically. On the other hand, the experienced manipulators may manipulate the stock price without sharp price changes and large trade volumes. For example, manipulators can split a big order into several small ones to trade, or place a large-scale order on the orderbook that cannot be traded but still has great impact on the market. Obviously, the current methods are incompetent for dealing with these cases. The key reason is that current methods normally deal with price or volume only, while price manipulation is a dynamic emergence of trading behaviour that needs to be catered from microstructure perspective.

3.2 Scenario Analysis into Approach Design

Our approach is based on the following foundation: (1) analysing trading behaviour from microstructure perspective, (2) involving domain knowledge and market factors, and (3) implementing data mining process by following the principle of D^3M.

In the stock market, an order is an instruction made by a trader to purchase or sell a security under certain condition. Traders enter their orders into the market and trade with others for making money. Both the attributes of orders and the way of entering orders are consistent with traders' intention. For instance, a trader who is urgent to sell and does not care about the return much will enter a sell order with a lower price which is easy to trade. However, a trader who does care much about the return and does not want to sell in haste will enter a sell order with a higher price. In fact, as well known by domain experts, manipulators also use orders to achieve their purposes. They manipulate the market by placing a series of particular buy or sell orders into the market. These tricky orders create an artificial, false, or misleading market appearance that misguide public traders but affiliate the manipulators for opportunities to make extra profit.

The above scenario analysis shows that two items of important domain knowledge. First, there is a strong connection between a trader's intentions and his/her trading activities reflected through entering orders. Second, it is reasonable to investigate order sequences for scrutinizing price manipulations. These indicate that the trading patterns of those genuine traders are different from fraudulent manipulators'. This can be through analysing order sequences to identify the difference. Once the exceptional patterns of entering orders are detected, it is reasonable to believe that there are likely suspicious trading behaviours taking place.

Inspired by the above scenarios analysis, this paper proposes an approach to discover the local exceptional patterns of order sequences for detecting price manipulation. Our approach has the following key steps: (1) constructing order sequences based on domain knowledge and scenario analysis, (2) defining interestingness specific for mining trading patterns catering for the domain issues. The *market microstructure patterns* [19] discovered by our approach can really power market surveillance system. We interpret them in detail in the following sections.

4 Domain-Driven Multidimensional Sequence Construction

4.1 Vector-Based Order Sequence Representation

Before mining patterns, it is necessary to represent and construct order sequences in the way reflecting market mechanism and trader's intention properly. In the stock market, orders have many attributes. Some attributes have categorical values like trade direction, and other attributes have continuous values such as price and volume. Though the values of these attributes vary from order to order, they are set according to the traders' own intentions. In addition, orders have their lifecycle, and may present in certain state at a single time point:

$$\{enter,\ trade\ partly,\ trade\ entirely,\ delete,\ and\ outstanding\}.$$

Two orders with the same values of attribute when they are entered into the market, however, may pass through different stages later on under different circumstances. That means the order's lifecycle also has a connection to the trader's purpose. From the above analysis, we can learn that both the information of order's attributes and the information of orders' lifecycle are related to the trader's intention directly or indirectly. Therefore, a reliable representation of market order should meet the requirement that it covers all the information of order's attributes and order's lifecycle.

In our approach, a five-dimension vector *O(d, p, v, t, b)* is defined to represent the order. Among the five dimensions, dimension $d \in \{B, S\}$ reflects the trade direction of order, dimension $p \in \{H, M, L\}$stands for the probability that the order can be traded, dimension $v \in \{S, M, L\}$ measures the size of order, dimension $t \in \{N, O, S\}$ represents how many trades the order leads to and dimension $b \in \{C, O,\ D\}$ reflects the balance of order when the market closes.

From the above definition and formulas, it can be learned that our five-dimensional vector contains the information not only of the order's attributes but also of the lifecycle which the order has passed through. The dimension *d, p* and *v* contain the

information of order's attributions while the dimension *t* and *b* contain the information of order's lifecycle. Consequently, it satisfies the requirement of reliable representation of order.

4.2 Constructing Multidimensional Sequences

To construct order sequences, there is a need to decide the time range of sequence first. In the stock market, orders last for not more than one day. Traders can enter their orders after the market opens and the orders which have not been traded expire at the market closing time. This means that orders have only one-day impact on the market. According to this domain-specific characteristic, we construct the order sequences with time range of one day. The second issue is how to divide orders into sequences. There are so many orders placed by traders in a trading day. It is unreasonable to put all the orders into a sequence. Recall that traders use a series of orders to implement their own intentions. Consequently, it is rational to assign all the orders placed by the same trader to a sequence.

A multidimensional sequence (for short sequence) Ω is defined as the sequence of orders placed by a same trader within a trading day,

$$\Omega = \{O_1(d_1, p_1, v_1, t_1, b_1), O_2(d_2, p_2, v_2, t_2, b_2), \ldots, O_i(d_i, p_i, v_i, t_i, b_i), \ldots\} \quad (1)$$

in which O_i is the five-dimensional vector defined in Section 4.1.

5 Mining Local Exceptional Patterns

5.1 Targeted Data and Benchmark Data

In the area of market surveillance, the exceptional patterns are more interesting. However, there are two kinds of exceptional patterns: *global exceptional patterns* and *local exceptional patterns*. The *global exceptional patterns* are the patterns which are exceptional for the whole data, while the *local exceptional patterns* are the patterns which are exceptional only for the local data. As the stock market is a dynamic system changing very fast, a pattern is exceptional for this period but may be normal for another period. Besides, a pattern is exceptional for a long period but may be normal for a short period. Consequently, the *global exceptional patterns* do not mean much in our case. We are interested in identifying the *local exceptional patterns*.

The definitions of *targeted data* and *benchmark data* are based on a sliding time window. As shown in the Fig. 1, there is a sliding time window with size of $m + 1$ days. Among these $m + 1$ days falling into the sliding time window, the first m days are called *benchmark day 1, 2, …, m* respectively, and the last day is called *targeted day*. Furthermore, the data drawn from the *benchmark day i* is called *benchmark data i* while the data drawn from the *targeted day* is called *targeted data*. With the movement of the sliding time window from left-hand side to right-hand side, any day can be the *targeted day* and has its corresponding *benchmark days*. Because the *benchmark days* are the close neighbours of the *targeted day*, there are correlations between the *targeted day* and its *benchmark days*. Intuitively, the closer the *benchmark day* is, the bigger the degree of correlation is. Therefore, each *benchmark day* is assigned a weight by the following formula:

$$W_i = (1 + \gamma)^{i-1} \tag{2}$$

Where $\gamma \geq 0$ reflects the volatility of market.

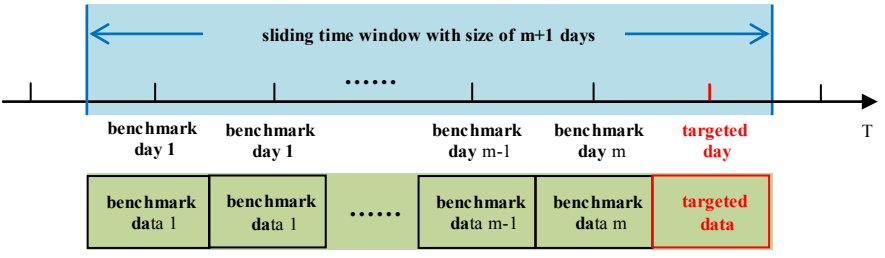

Fig. 1. Targeted data and benchmark data

5.2 Algorithms for Mining Local Exceptional Patterns

Definition 1. *Intentional Interestingness (II)*: *II* quantifies the intentional interestingness of pattern as defined in the following formula:

$$II = Sup_t \times \frac{|\Omega|}{AvgL_t} \tag{3}$$

where

- Sup_t is the support of sequence Ω in the *targeted data*,
- $|\Omega|$ is the length of sequence Ω,
- $AvgL_t$ is the weighted average length of sequences in the *targeted data*,

II is positively related to the support in the *targeted data* and the length of pattern. The idea behind this measure is very straightforward. As it is said before, traders tend to use a series of orders to implement their intentions. Therefore the order sequence with a higher support and a longer length is placed more intentionally in the *targeted day*.

Definition 2. *Exceptional Interestingness (EI)*: *EI* quantifies the exceptional interestingness of pattern as defined in the following formula:

$$EI = \frac{\frac{Sup_t}{AvgL_t} \times \sum_{i=1}^{m} W_i}{\sum_{i}^{m}(\frac{SupB_i}{AvgLB_i} \times W_i)} \tag{4}$$

- where
- $SupB_i$ is the support of sequence Ω in the *benchmark data i*,
- $AvgLB_i$ is the weighted average length of sequences in the *benchmark data i*,
- W_i is the weight for the *benchmark data i*, and
- m is the number of *benchmark days*.

EI is negatively related to the supports in the *benchmark data*. The lower support the pattern has in the *benchmark data*, the more exceptional the pattern is. It reflects how exceptional the pattern is in the *targeted day* compared with in the *benchmark days*.

Consequently, a sequence is a local exceptional pattern, if it satisfies the conditions: (1) $II \geq MinII$, and (2) $EI \geq MinEI$, where *MinII* and *MinEI* are the thresholds given by users or domain experts.

To discover the local exceptional patterns, we design an algorithm, namely *Mining Local Exceptional Patterns*, as shown in the Fig. 2.

```
ALGORITHM: Mining Local Exceptional Patterns
INPUT: trading dataset TD, order dataset OD, m, γ, MinII, MinEP
OUTPUT: local exceptional patterns LEP

LEP =∅; /*local exceptional patterns*/
BS = ∅; /*benchmark sequences*/
FOR each trading day i from trading day 1 to trading day m
        S = GenSeq(TD_i, OD_i); /*generate sequences*/
        BS = BS + S; /*add the sequences to benchmark sequences*/
ENDFOR
FOR each trading day i from trading day m + 1 to the last trading day
        S = GenSeq(TD_i, OD_i); /*generate sequences from targeted data*/
        P = MinePatterns(S) /*mine patterns from the sequences*/
        FOR each pattern P_j in P
                II_j = GetII(P_j); /* quantify the intentional interestingness*/
                EI_j = GetEI(P_j, BS, γ); /*quantify the exceptional interestingness*/
                /*add the pattern into LEP, if it meets the conditions*/
                IF II_j ≥ MinII and EI_j ≥ MinEI
                        LEP = LEP + P_j;
                ENDIF
        ENDFOR
        Replace the sequences generated from trading day i – m in BS with S;
ENDFOR
OUTPUT local exceptional patterns LEP;
```

Fig. 2. Algorithm Mining Local Exceptional Patterns

6 Experiments

Our approach has been tested on a real dataset from an Exchange. It covers 240 trading days from 2005 to 2006 for a security. There were 213,898 orders entered by traders during this period. These orders led to 228,186 trades.

Table 1 shows some samples of local exceptional patterns discovered by our approach. These patterns reflect the traders' exceptional intentions in the corresponding day. For example, in *24/05/2005*, the *intentional interestingness* and *exceptional interestingness* for pattern {*(S,M,S,O,C), (S,M,S,O,C)*} are *0.054* and *11.2* respectively, which means that this pattern indicates a strong intention and exception of trading activities for that day.

Table 1. Local exceptional pattern samples (m = 10, γ = 0.01, *MinII* = 0.025, and *MinEI* = 5); AR stands for the security's *abnormal return* compared with *market return*

Date	Local Exceptional Patterns	*II*	*EI*	|Return| %	|AR| %
05/01/2005	{(B,H,S,N,D),(B,H,S,N,D),(B,H,S,N,D),(B,H,S,N,O)}	0.026	+∞	1.68	0.77
14/01/2005	{(B,H,S,N,D),(B,H,S,N,D),(B,H,S,O,C)}	0.025	7.6	2.68	1.38
18/01/2005	{(S,M,S,N,D),(S,M,S,N,D),(S,M,S,N,D),(S,M,S,O,C)}	0.026	10.8	2.25	1.85
20/01/2005	{(S,M,S,N,D),(S,H,S,O,C)}	0.038	5.4	2.98	1.56
28/01/2005	{(B,H,S,O,C),(B,M,S,O,C)}	0.030	8.2	2.63	1.97
01/02/2005	{(B,H,S,N,D),(B,H,S,N,D),(B,H,S,N,D)}	0.030	5.2	0.79	0.93
13/05/2005	{(S,M,S,N,D),(S,M,S,N,D),(S,M,S,N,O)}	0.026	5.1	0.95	2.24
24/05/2005	{(S,M,S,O,C),(S,M,S,O,C)}	0.054	11.2	6.82	6.38
16/06/2005	{(B,M,S,O,C),(S,M,S,N,O)}	0.025	6.1	2.64	2.47
17/06/2005	{(S,H,S,N,O)}	0.033	19.0	1.88	1.00
01/07/2005	{(B,H,S,N,D),(B,H,S,N,D),(B,H,S,O,C)}	0.028	54.0	2.21	1.28
13/07/2005	{(B,M,S,N,O)}	0.028	5.6	9.55	9.12
15/09/2005	{(B,H,S,N,O),(B,H,S,N,O)}	0.035	8.8	1.43	3.49
05/12/2005	{(S,M,S,O,C),(S,M,S,O,C)}	0.035	8.6	1.85	3.41

Figs 3 and 4 illustrate the performance of our approach under different thresholds of *MinII* and *MinEI*.

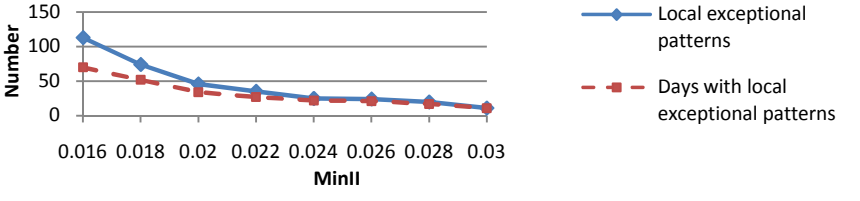

Fig. 3. Number of local exceptional patterns and number of days with local exceptional patterns for different *MinII* when m = 10, γ = 0.01 and *MinEI* = 5

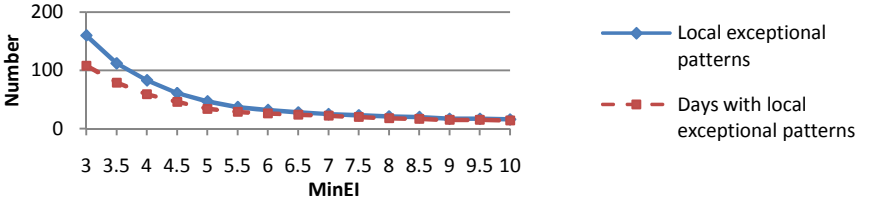

Fig. 4. Number of local exceptional patterns and number of days with local exceptional patterns for different *MinEI* when m = 10, γ = 0.01 and *MinII* = 0.02

To evaluate the actionability of our findings in the business world, we calculate the absolute *return* and *abnormal return* of the security. In the stock market, *return* refers to the gain or loss for a single security over a specific period while *abnormal return*

indicates the difference between the *return* of a security and the *market return*. As shown in Table 1, the absolute *return* and *abnormal return* on *24/05/2005* are as high as *6.82%*, *6.38%* respectively, which are aligned with the values of *II* and *EI* for pattern {*(S,M,S,O,C), (S,M,S,O,C)*}. These results from both technical and business sides present business people strong indicators showing that there likely is price manipulation on that day.

7 Conclusion and Future Work

Often the outputs of traditional data mining cannot support people to make decision or take actions in the business world. There is a big gap between academia and business. However, this gap can be filled in by domain-driven data mining (D^3M) which generally targets the actionable knowledge discovery under domain-specific circumstances. In D^3M, the domain-specific characteristics are considered and domain knowledge is also encouraged to involve itself in the whole data mining process. Consequently the D^3M-based mining results have potential to satisfy the business expectations.

In this paper, we have investigated local exceptional trading patterns for detecting stock price manipulations in real-life orderbook data by utilizing D^3M. The main contributions of this paper are as follows: (1) studying exceptional trading patterns on real-life intraday stock data that is rarely studied in current literature, (2) proposing an effective approach to represent and construct trading orders, (3) developing an effective interestingness and algorithms for mining and evaluating local exceptional patterns, and (4) testing the proposed approach in real-life data by considering market scenarios and business expectations.

In the future, we plan to improve the representation of orders by including more domain-specific characteristics. Besides, we also expect to develop more interestingness measures reflecting the business concern.

Acknowledgement

The project is partially supported by Australian Research Council Discovery grants DP0773412, LP0775041 and DP0667060.

References

1. http://www.marketsurveillance.org/
2. Cao, L., Zhang, C.: Domain-driven Data Mining: A Practical Methodology. Int'l J. Data WareHousing and Mining 2(4), 191–196 (2006)
3. Cao, L., Yu, P.S., Zhang, C., Zhang, Y., Williams, G.: DDDM 2007: Domain Driven Data Mining. ACM SIGKDD Explorations 9(2), 84–86 (2007)
4. Cao, L.: Domain-Driven actionable knowledge discovery. IEEE Intelligent Systems 22(4), 78–89 (2007)
5. Cao, L., Luo, D., Zhang, C.: Knowledge actionability: satisfying technical and business interestingness. Int. J. Business Intelligence and Data Mining 2(4), 496–514 (2007)

6. Cao, L., Luo, C., Zhang, C.: Developing Actionable Trading Strategies for Trading Agents. In: IAT 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, pp. 72–75 (2007)
7. Ankerst, M.: Human Involvement and Interactivity of the Next Generation's Data Mining Tools. In: Workshop on Research Issues in Data Mining and Knowledge Discovery joint with DMKD 2001, Santa Barbara, CA (2001)
8. Aggarwal, C.: Towards Effective and Interpretable Data Mining by Visual Interaction. ACM SIGKDD Exploration Newsletter 3(2), 11–22 (2002)
9. Cao, L., Zhang, C.: The Evolution of KDD: Towards Domain-driven Data Mining. Int. J. Pattern Recognition and Artificial Intelligence 21(4), 667–692 (2007)
10. Freitas, A.A.: On Objective Measures of Rule Surperisingness. In: Zytkow, J., Quafafou, M. (eds.) PKDD 1998, vol. 1510, pp. 1–9. Springer, Heidelberg (1998)
11. Hilderman, R.J., Hamilton, H.J.: Applying Objective Interestingness Measures in Data Mining Systems. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 432–439. Springer, Heidelberg (2000)
12. Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing Subjective Interestingness of Association Rules. IEEE Intelligent Systems 15(5), 47–55 (2000)
13. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as A Measure of Interestingness in Knowledge Discovery. Decision and Support Systems 27, 303–318 (1999)
14. Silberschatz, A., Tuzhilin, A.: On Subjective Measures of Interestingness in Knowledge Discovery. Knowledge Discovery and Data mining, 275–281 (1995)
15. Yang, Q., Yin, J., Lin, C., Chen, T.: Postprocessing Decision Trees to Extract Actionable Knowledge. In: Proc. ICDM 2003, pp. 685–688. IEEE Computer Science Press, Los Alamitos (2003)
16. Ling, C., Sheng, W., Bruckhaus, T., Madavji, N.: Maximum Profit Mining and Its Application in Software Development. In: Proc. SIGKDD 2006, pp. 929–934. ACM Press, New York (2006)
17. Wang, K., Jiang, Y., Tuzhilin, A.: Mining Actionable Patterns by Role Models. In: ICDE 2006, p. 16. IEEE Computer Science Press, Los Alamitos (2006)
18. Tuzhilin, A.: Knowledge Evaluation: Other Evaluations: Usefulness, Novelty, and Integration of Interesting News Measures. In: Handbook of Data Mining and Knowledge Discovery, pp. 496–508 (2002)
19. Cao, L., Ou, Y.: Market Microstructure Patterns Powering Trading and Surveillance Agents. Journal of Universal Computer Sciences (to appear, 2008)