

Behavior Informatics/Analytics

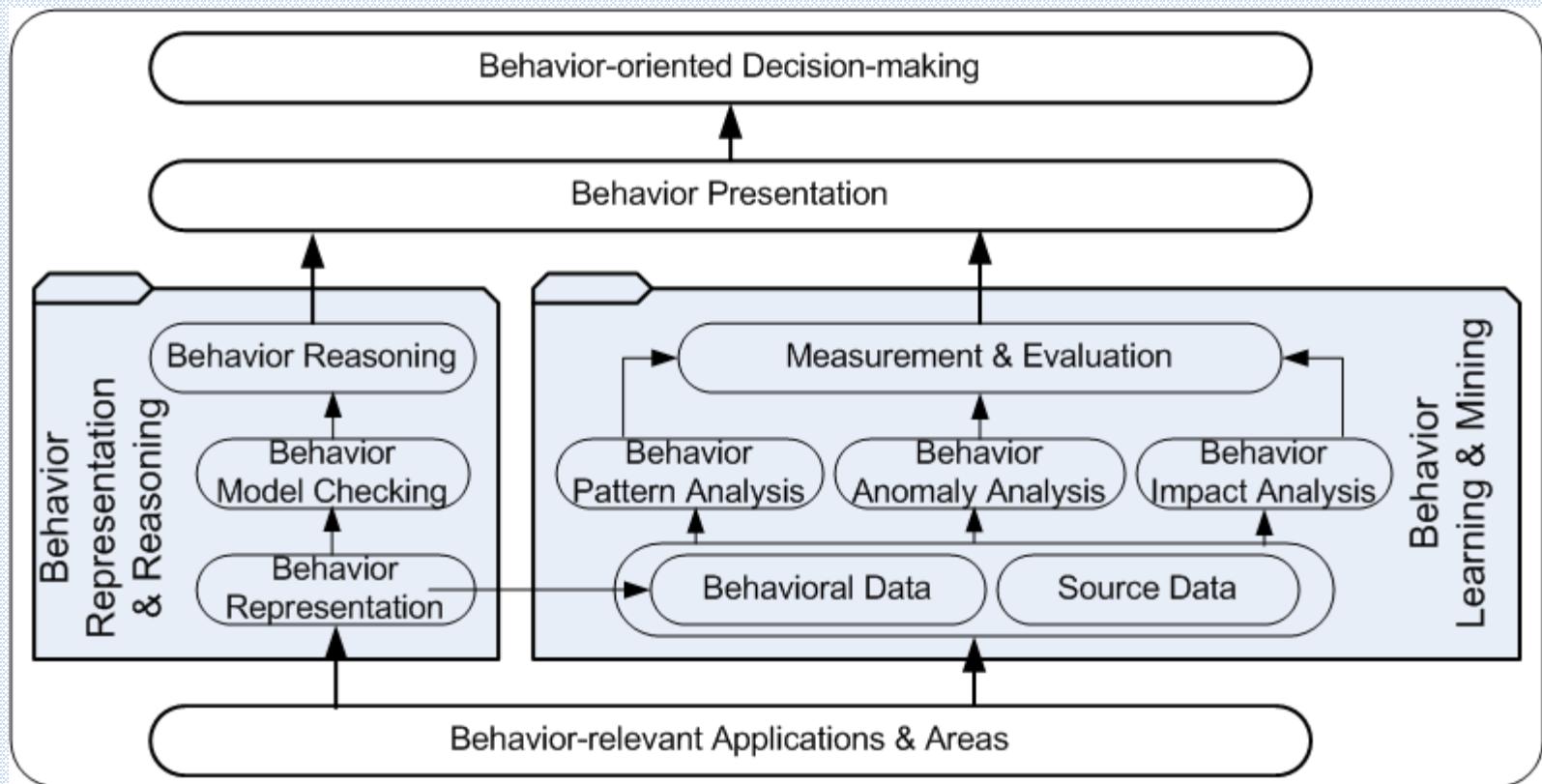
Modeling, Analysis and Mining of Complex Behaviors
Discovering Behavior Intelligence and Insights

Professor Longbing Cao

Advanced Analytics Institute
University of Technology Sydney, Australia

Behavior informatics/analytics

– Concept Map



<http://www.behaviorinformatics.org/>

BEHAVIOR INFORMATICS
...Discovering Behavior Intelligence

Outline

1

Why Behavior Informatics & Analytics?

2

What is Behavior?

3

What is Behavior Informatics & Analytics?

4

Behavior Model/Representation

5

High Impact Behavior Analysis

6

High Utility Behavior Analysis

7

Negative Behavior Analysis

8

Coupled Group Behavior Analysis

9

Enterprise Applications of Behavior Analytics

References & Slides

- <http://www-staff.it.uts.edu.au/~lbciao/publication/behaviorinformatics-tutorial-slidesx.pdf>
- <http://www-staff.it.uts.edu.au/~lbciao/publication/publications.htm>
- www.behaviorinformatics.org

Behavior Informatics Resources

Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://www.behaviorinformatics.org/

www.behaviorinformatics.org

BEHAVIOR AND SOCIAL INFORMATICS, IEEE TASK FORCE

...Discovering Behavior and Social Intelligence

MAIN MENU

- Home
- Calls for ...
- Introduction
- Research Topics
- Activities
- Projects
- Communities
- Resources
- References
- About Us
- Contact Us

LINKS

- BI2012
- BI2011
- BI2010
- AMII-SIG
- DDDM-SIG
- EDM-SIG
- MS-SIG

Welcome to IEEE Task Force on Behavior and Social Informatics



**Behavior Computing:
Modeling, Analysis, Mining and Decision**

Longbing Cao, Philip S Yu (Eds.)
Springer, 2012

First dedicated source of references for the theory and applications of behavior informatics and behavior computing..

News:

- Tutorial: [Behavior Computing: Complex Behavior Modeling, Analysis and Mining](#), VVI-IAT2012, 4 Dec 2012.
- [2012 Workshop on Behavior Informatics \(BI2012\)](#) has been accepted by VVI-IAT2012.
- The first dedicated reference to behavior informatics: [Behavior Computing](#) is available in Springer.

Opportunities:

[Call for books]: Calls for edited books, monographs and so on to the [Book Series: Advanced Studies on Behavior Informatics](#).

[Call for papers]: Call for papers to the [2012 International Workshop on Behavior Informatics \(BI2012\)](#).

<http://www.behaviorinformatics.org/>

Acknowledgement

- My ex-students, fellows and visitors: Dr Yanchang Zhao, Dr Huaifeng Zhang, Dr Zhigang Zheng, Dr Can Wang, Dr Yin Song, Dr Junfu Yin, Dr Wei Cao, Prof Xiangjun Dong, etc.

Part I.

Concepts & Representation

Learning Objectives

- Why behavior informatics?
- What is behavior?
- What are the key behavioral factors?
- What is the conceptual map of behavior informatics?
- How to represent/model behavior?
- How to check the behavior model?



In-depth behavior understanding and use: The behavior informatics approach

Longbing Cao

Faculty of Engineering and Information Technology, University of Technology, Sydney, Australia

ARTICLE INFO

Article history:

Received 10 May 2009

Received in revised form 20 March 2010

Accepted 24 March 2010

Keywords:
Informatics
Behavior analysis
Behavior informatics
Behavior computing
Decision making

ABSTRACT

The in-depth analysis of human behavior has been increasingly recognized as a crucial means for disclosing interior driving forces, causes and impact on businesses in handling many challenging issues such as behavior modeling and analysis in virtual organizations, web community analysis, counter-terrorism and stopping crime. The modeling and analysis of behaviors in virtual organizations is an open area. Traditional behavior modeling mainly relies on qualitative methods from behavioral science and social science perspectives. On the other hand, so-called behavior analysis is actually based on human demographic and business usage data, such as churn prediction in the telecommunication industry, in which behavior-oriented elements are hidden in routinely collected transactional data. As a result, it is ineffective or even impossible to deeply scrutinize native behavior intention, lifecycle and impact on complex problems and business issues. In this paper, we propose the approach of *behavior informatics* (BI), in order to support explicit and quantitative behavior involvement through a conversion from source data to behavioral data, and further conduct genuine analysis of behavior patterns and impacts. BI consists of key components including *behavior representation*, *behavioral data construction*, *behavior impact analysis*, *behavior pattern analysis*, *behavior simulation*, and *behavior presentation* and *behavior use*. We discuss the concepts of behavior and an abstract behavioral model, as well as the research tasks, process and theoretical underpinnings of BI. Two real-world case studies are demonstrated to illustrate the use of BI in dealing with complex enterprise problems, namely analyzing exceptional market microstructure behavior for market surveillance and mining for high impact behavior patterns in social security data for governmental debt prevention. Substantial experiments have shown that BI has the potential to greatly complement the existing empirical and specific means by finding deeper and more informative patterns leading to greater in-depth behavior understanding. BI creates new directions and means to enhance the quantitative, formal and systematic modeling and analysis of behaviors in both physical and virtual organizations.

© 2010 Elsevier Inc. All rights reserved.

The concept of behavior informatics

Longbing Cao:
In-depth behavior understanding and use: The behavior informatics approach. Inf. Sci. 180(17): 3067–3085 (2010)



TRENDS & CONTROVERSIES

Editors: Longbing Cao, University of Technology, Sydney, longbing.cao@uts.edu.au

Behavior Informatics: A New Perspective

Longbing Cao, University of Technology, Sydney

Behavior is a concept increasingly recognized in broad communities spreading from social to business, online, mobile, economic, and cultural domains. However, systematic and comprehensive methodologies, theories, tools, and systems aren't ready for deeply, fully, and effectively capturing, representing, quantifying, analyzing, learning, and measuring the semantics, sequencing, networking, evolution, utility and impact of individual, group, and cohort behaviors taking place in the real world. This is becoming fundamental and critical in the age of Big Data. Here, in this installment of "Trends & Controversies," we look at how *behavior informatics* targets the development of effective methodologies and techniques to tackle these issues.

social and collaborative searching activities is needed. Gabriella Pasi presents insights on engaging behaviors in information seeking, especially considering coupled behaviors within certain contexts.

Nowadays, an increasing number of users are interested in IPTV programs online, and generate massive amounts of activities. Ya Zhang and her colleagues lead a discussion about the behaviors of IPTV users that are related to system efficiency, personalization, recommendation, and targeted advertisement.

Finally, Edoardo Serra and V.S. Subrahmanian raise an interesting question: Should behavior models of terror groups be disclosed? They share their research and arguments on strategic disclosures and consequences in tackling today's terrorism.

Longbing Cao, Thorsten Joachims, Can Wang, Éric Gaussier, Jinjiu Li, Yuming Ou, Dan Luo, Reza Zafarani, Huan Liu, Guandong Xu, Zhiang Wu, Gabriella Pasi, Ya Zhang, Xiaokang Yang, Hongyuan Zha, Edoardo Serra, V. S. Subrahmanian: **Behavior Informatics: A New Perspective.** IEEE Intelligent Systems 29(4): 62-80 (2014)

Business Case

1 Develop your business case

Before you start this learning and discussion journey, please think of a business case that is familiar to you, and we will use it for discussions and exercises in the course. The best way is probably for you to choose something happening to your daily business.

2 Toy business case

I here include a toy business case for you to think or customize for the course use, if you like.

Why Behavior Informatics & Analytics?

Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, *Information Science*, 180(17); 3067-3085, 2010.

www.behaviorinformatics.org

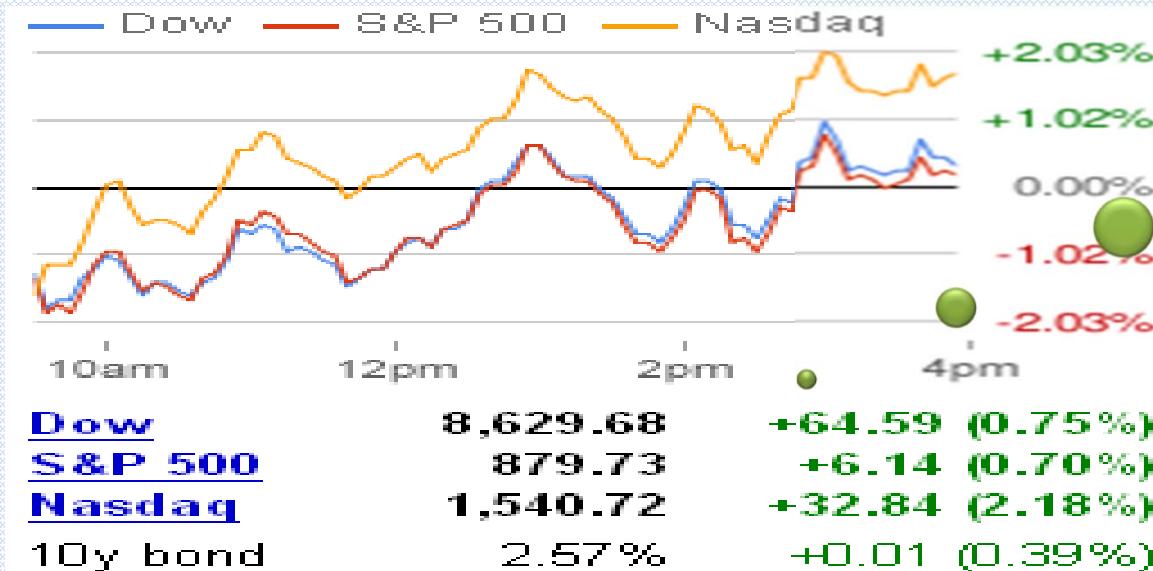
Argument 1: Behavior is ubiquitous

- Behavior is an important analysis object in
 - Consumer analysis
 - Marketing strategy design
 - Business intelligence
 - Customer relationship management
 - Social computing
 - Intrusion detection
 - Fraud detection
 - Event analysis
 - Risk analysis
 - Group decision-making, etc.

➤Customer behavior analysis
➤Consumer behavior and market strategy
➤Web usage and user preference analysis
➤Exceptional behavior analysis of terrorist and criminals
➤Trading pattern analysis of investors in capital markets

Argument 2: Major work focuses on Behavior exterior-driven analysis

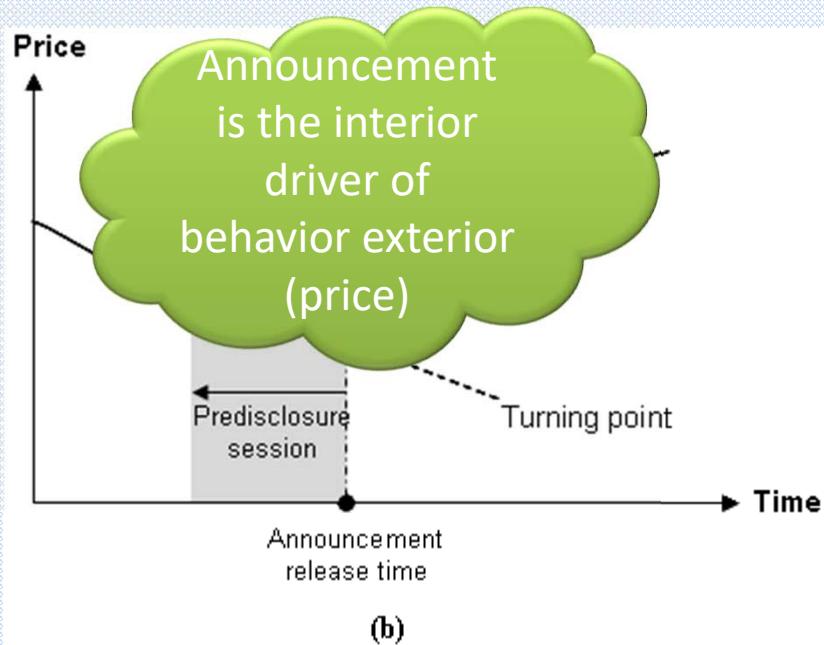
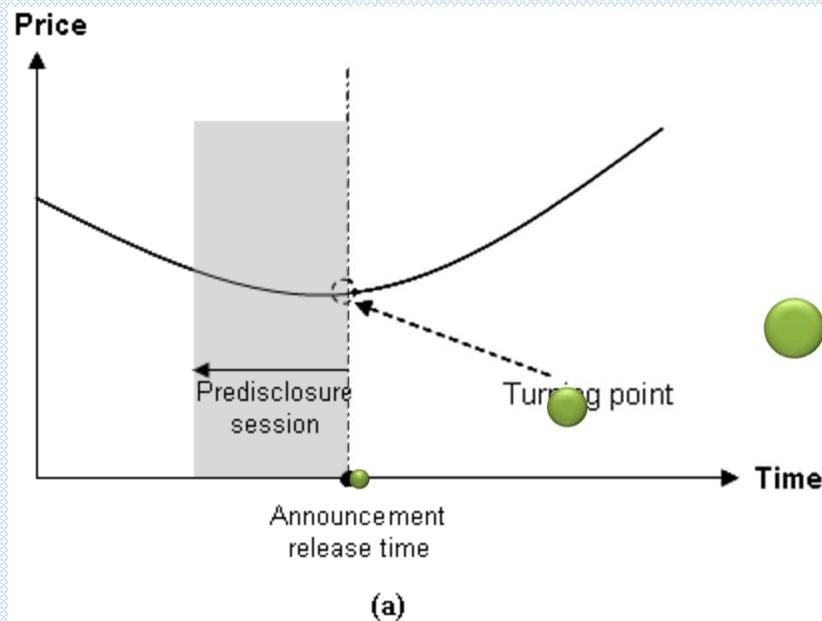
- Example 1: Price movement as market behavior

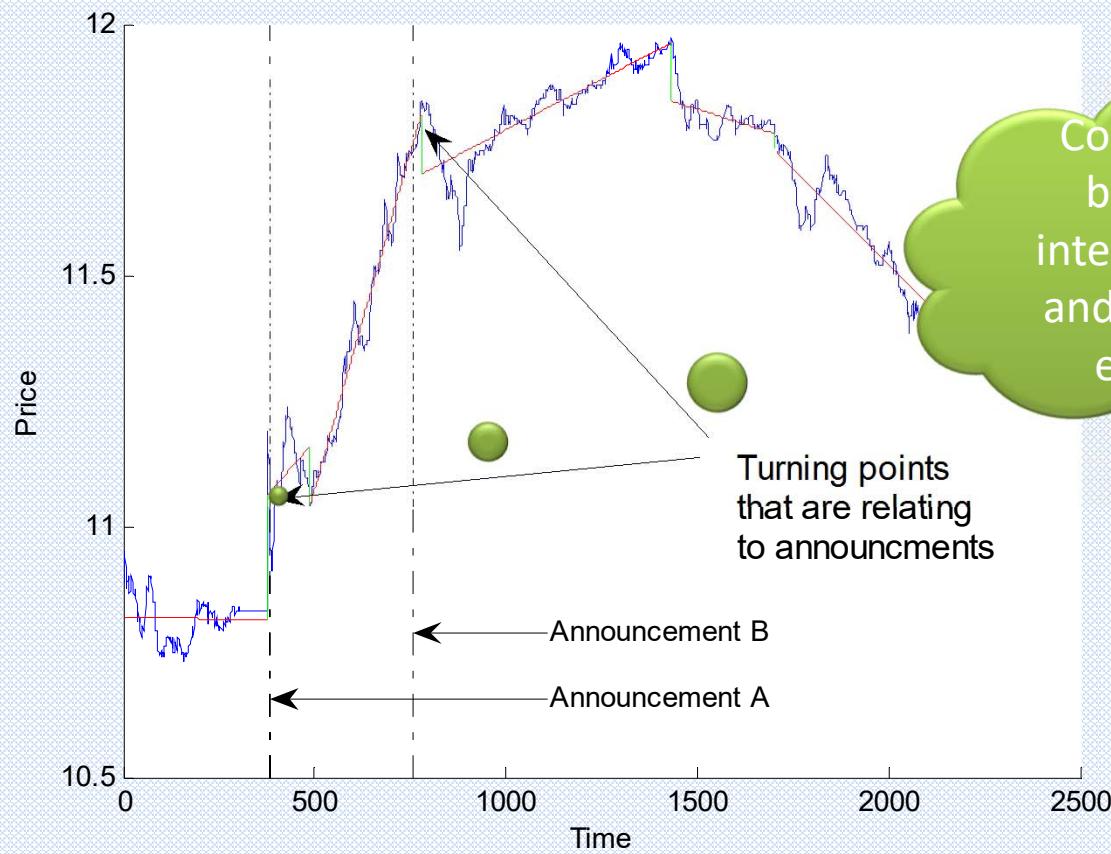


Price/index
movement
is the
behavior
exterior

Argument 3: Behavior interior-driven analysis can make difference

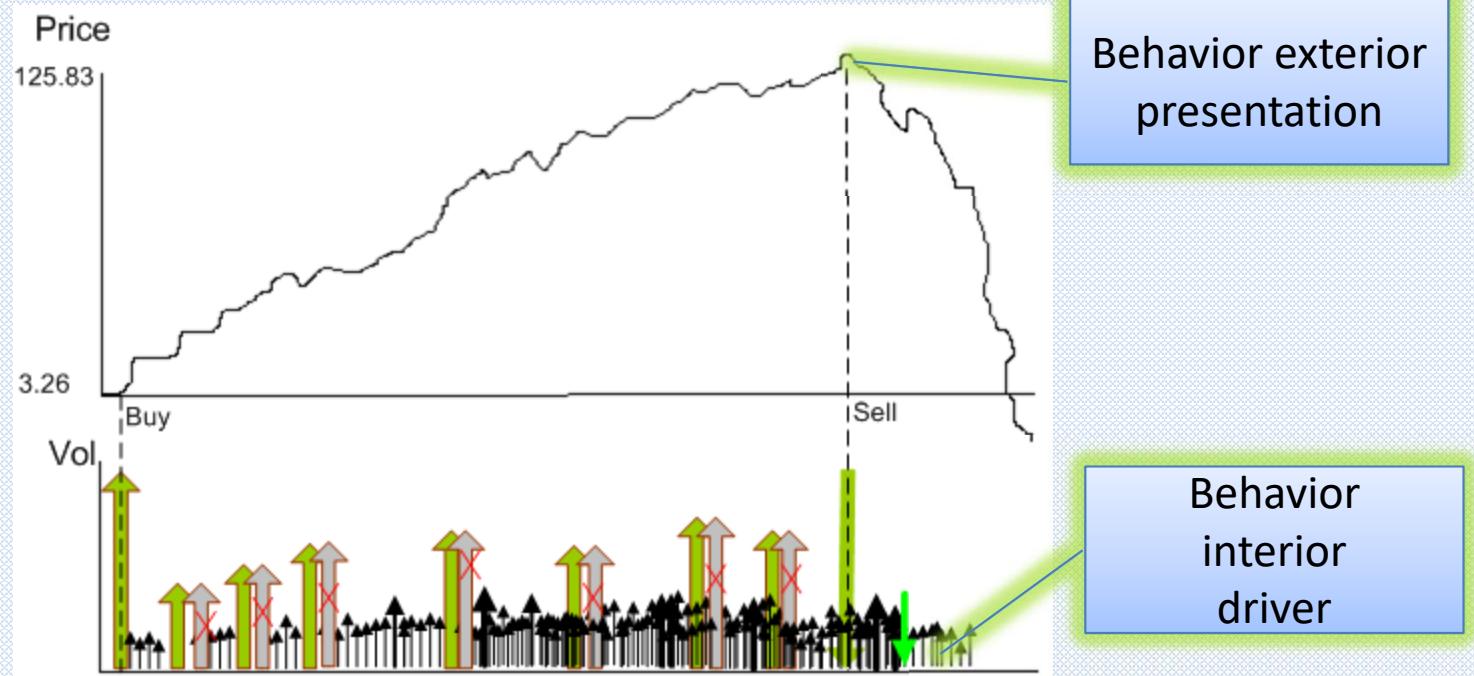
- Example 2: Announcement as market behavior driver





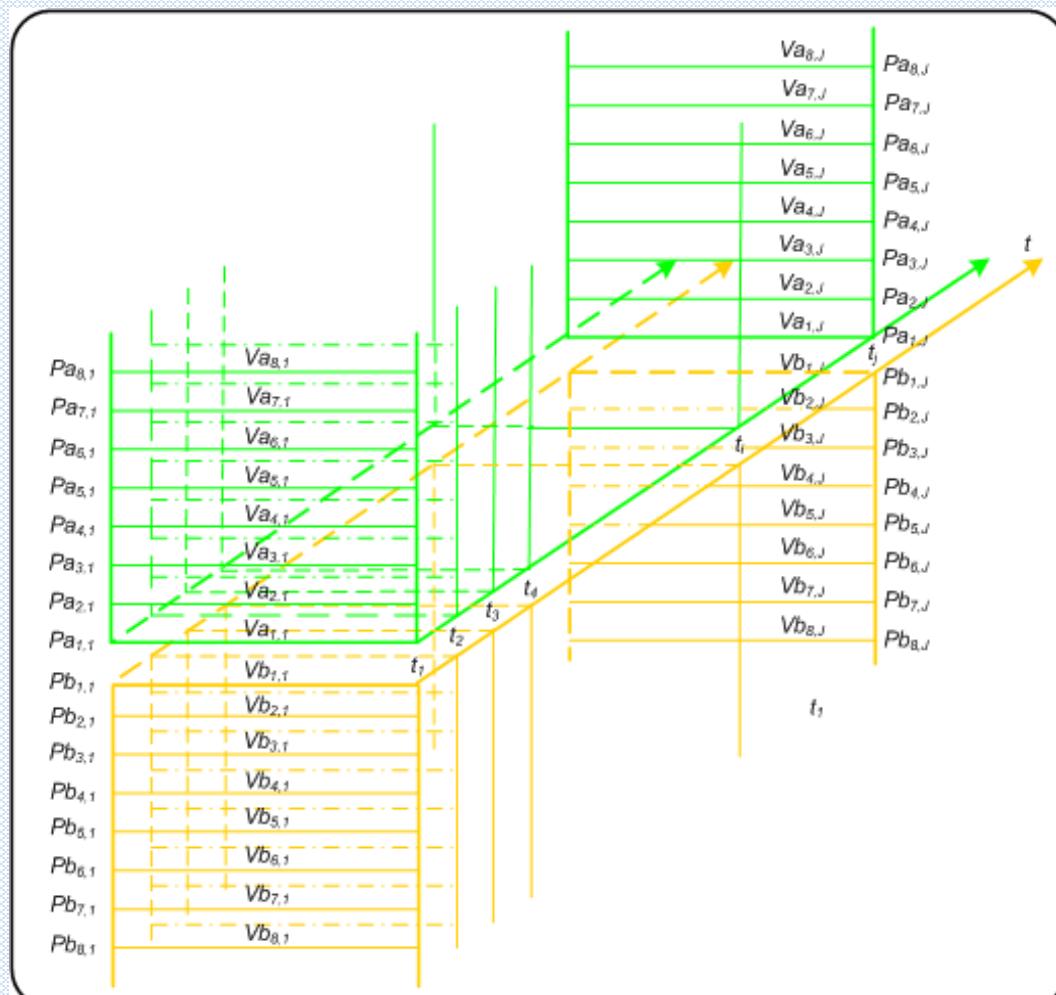
Connection
between
interior driver
and behavior
exterior

- Short-term manipulation behavior as cause



Argument 4: Need to consider behavior context

- Microstructure data



Big Data-based Behaviour Analysis

1. Advertising

- 1. Advertising Agencies
- 2. Advertising Systems
- 3. Architecture for the Future
- 4. Multi-Media Publications
- 5. Publish and Subscribe (Basic)
- 6. Publish and Subscribe (Financial Portal)
- 7. Publish and Subscribe (using Tags)
- 8. Web Site Analytics

2. Air Transport

- 1. Aircraft Parts and Orders
- 2. Airline Bookings in Africa
- 3. Airline Operations
- 4. Airline Reservations
- 5. Airline Travel Routes
- 6. Airport in a Box NEW
- 7. Airport Management
- 8. BAA (British Airports Authority)
- 9. Enterprise Data Model for Air Travel
- 10. Heathrow Airport

3. Assets

- 1. Asset Management
- 2. Assets
- 3. Assets and Locations
- 4. Assets Maintenance
- 5. Assets Schedules

4. Banks

- 1. Banking Acquisitions
 - 2. Bank and Branches
 - 3. Banking Data Warehouses
 - 4. Investment Banks
 - 5. Online Banking
 - 6. Retail Banks
 - 7. Retail Bank Accounts for Husbands and Wives
- NEW

1. Customers

- 1. Clients and Fees
- 2. Clients and Locations
- 3. Clients, Services and Fees
- 4. CRM
 - Call Centers
 - Customer Metrics
 - Marketing Data
 - Marketing Data Architecture
 - Microsoft Dynamics CRM
 - Personalization
 - Top-Level Data Model
 - Template for a Source Systems Data Dictionary
- 5. Customer Experience Management
- 6. Customer Management Systems
- 7. Customers with multi-lingual B2C
- 8. Customers at a Bank
- 9. Customers at a Bank (retail)
- 10. Customers at a Bookstore
- 11. Customers at a Call Center
- 12. Customers and Addresses
- 13. Customers and Campaigns
- 14. Customers and Car Hire
- 15. Customers and Car Parts
- 16. Customers and Car Sales
- 17. Customers and Car Servicing
- 18. Customers and Credit Cards
- 19. Customers and Dept Stores
- 20. Customers and Deals (UML)
- 21. Customers and Deliveries
- 22. Customers and e-Commerce
- 23. **Customers and Financial Services**
- 24. Customers and Frozen Yoghurt Shops
- 25. Customers and Games Shops
- 26. Customers and Hairdressers

1. Parties

- 1. Parties, Roles and Customers
- 2. Party - IIA Insurance
- 3. Party Master Index

2. Pharmaceuticals

- 1. Pharma and Biotechnology Information
- 2. Pharmacies and Generics
- 3. Pharmacies and Medical Clinics
- 4. Pharmacies and Prescriptions
- 5. Pharmaceuticals Data Warehouse
- 6. Pharmaceutical Companies
- 7. Pharmaceutical Vendors Visits
- 8. Pharmaceutical Supplies

3. Payments

- 1. Customers and Payments (e-Gov't)
- 2. Invoices and Payments
- 3. Future of Payments
- 4. Medical Laboratories with Payments
- 5. Payments Subject Area
- 6. Tracking Multiple Job Payments

4. People

5. Police

- 1. Baltimore Police Department
- 2. Police Canonical Data Model
- 3. Police Departments
- 4. Police Generalized Data Model
- 5. Police Information Reports
- 6. Police MDM Data Model
- 7. Police Mobile Application
- 8. Tracking Evidence
- 9. Traffic Cops and Tickets

6. Products

- 1. Bill of Materials

Observation: Traditional analysis on behavior

- Empirical, qualitative, psychological, social etc
- Behavior-oriented analysis was usually conducted on **customer demographic and transactional data** directly
 - Telecom churn analysis, **customer demographic data and service usage data** are analyzed to classify customers into loyal and non-loyal groups based on the dynamics of usage change
 - Outlier mining of trading behavior, **price movement** is usually focused to detect abnormal behavior

so-called behavior-oriented analysis is actually not on customer behavior-oriented elements, rather on straightforward customer demographic data and business usage related appearance data (transactions)

Problems with traditional behavior analysis

- Customer demographic and transactional data is not organized in terms of behavior but **entity relationships**
- Human behavior is *implicit* in normal transactional data: **behavior implication**
 - cannot support in-depth analysis on **behavior interior**: focus on **behavior exterior**
 - Cannot scrutinize **behavioral actor's belief, desire, intention and impact** on business appearance and problems

Such behavior implication indicates the limitation or even ineffectiveness of supporting behavior-oriented analysis on transactional data directly.

Genuine behavior analysis does matter

- Behavior plays the role as **internal driving forces or causes** for business appearance and problems
- Complement traditional pattern analysis solely relying on demographic and transactional data
- Disclose **extra information** and **relationship** between behavior and target business problem-solving

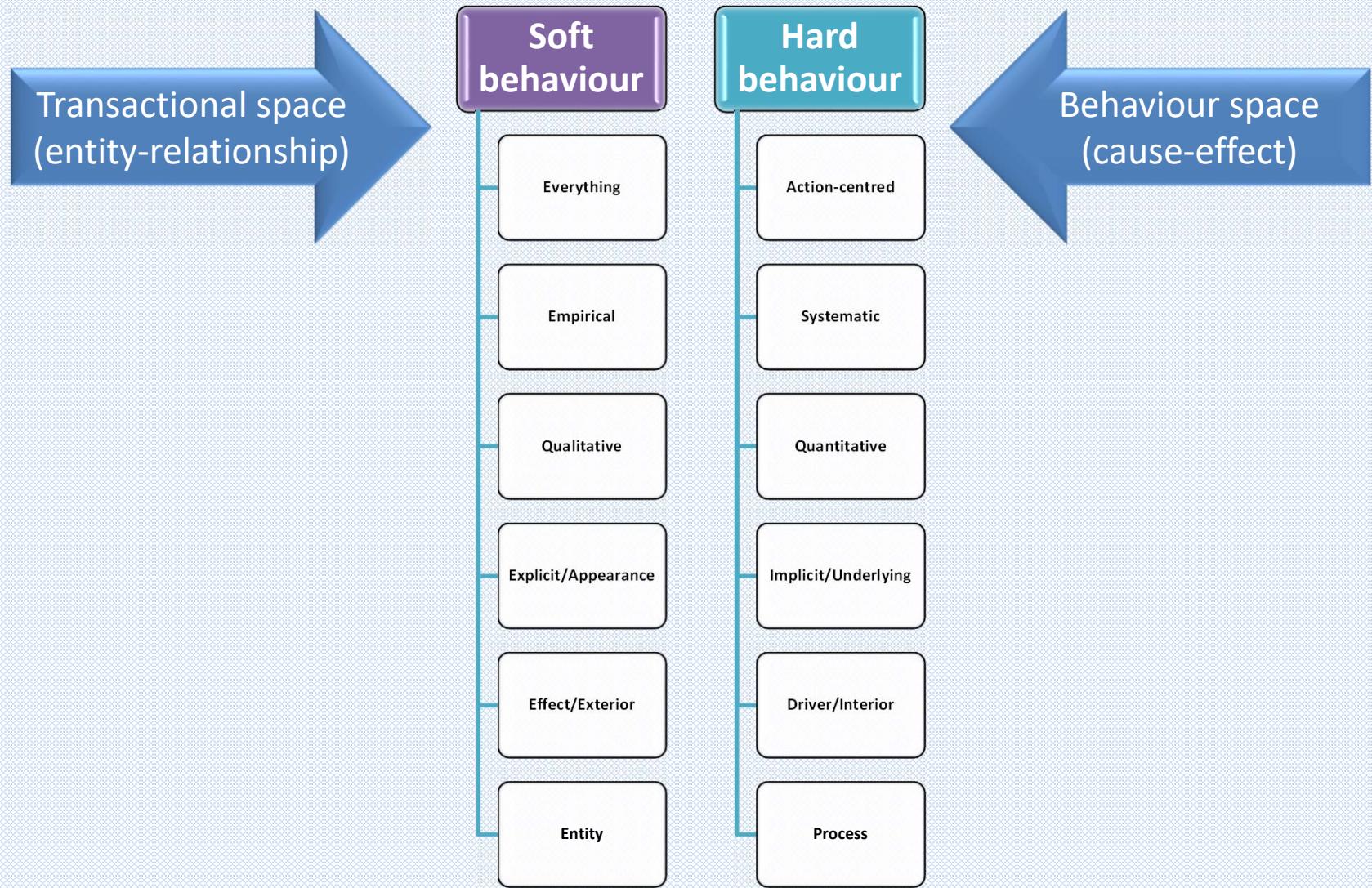
A multiple-dimensional viewpoint and solution may exist that can uncover problem-solving evidence from not only demographic and transactional but behavioral (including intentional, social, interactive and impact aspects) perspectives

Support genuine behavior analysis

- Make behavior ‘**explicit**’ by squeezing out behavior elements hidden in transactional data
- *A conversion from transactional space to behavior feature space is necessary*
- Behavioral data:
 - *behavior modeling and mapping*
 - organized in terms of behavior, behavior relationship and impact

Explicitly and more effectively analyze behavior patterns and behavior impacts than on transactional data

Behavior: soft vs. hard



Discussion 1: Behaviour in your organisation

- 1 List the business lines (drill down to specific business areas) in your organization where behaviour could be an important aspect/asset

- 2 Use a few keywords in a dot point format to describe behaviour analytics tasks conducted at your organization

What is Behavior?

What is behavior?

- An abstract behavior model
 - Demographics and circumstances of belief
 - Associates of a behavior may form into network;
 - Social behavioral network consists of sets organized in terms of certain social relations
 - Impact, costs, risk and trust of behavior

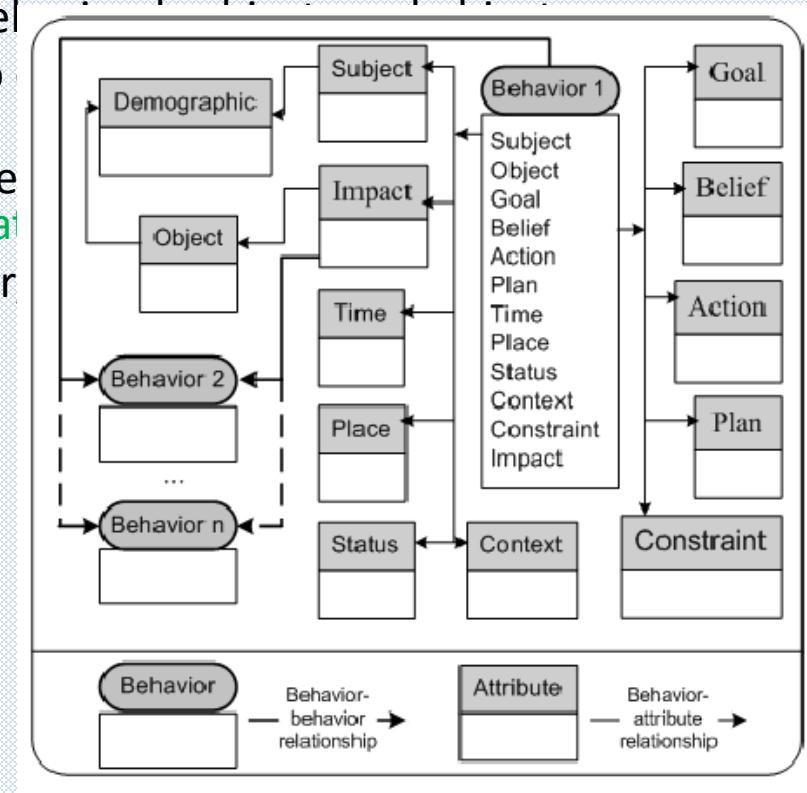


Figure 1. An Abstract Behavioral Model

- Behavior instance: **behavior vector**

$$\vec{\gamma} = \{s, o, e, g, b, a, l, f, c, t, w, u, m\}$$

- basic properties
- social and organizational factors

- Vector-based behavior sequences

$$\vec{\Gamma} = \{\vec{\gamma}_1, \vec{\gamma}_2, \dots, \vec{\gamma}_n\}$$

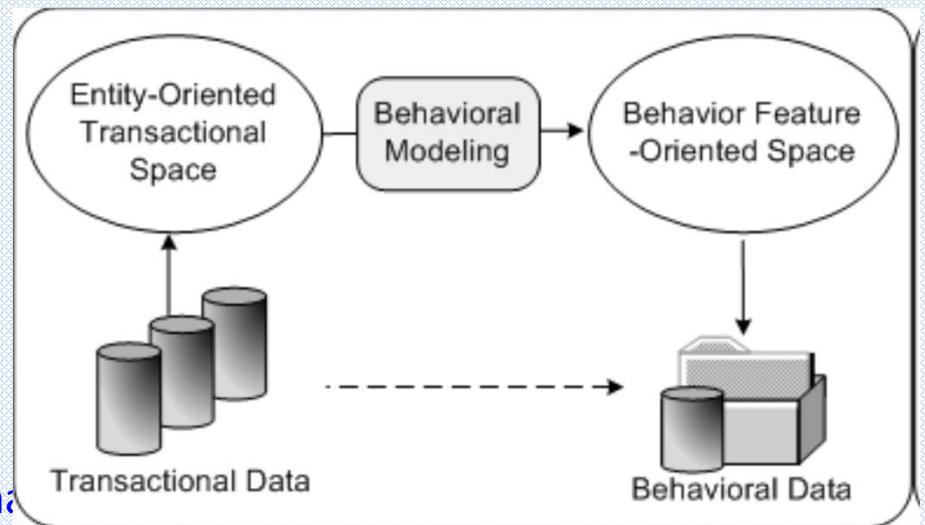
- Vector-oriented patterns

- Vector-oriented behavior pattern analysis
 - Behavior performer:
 - Subject (s), action (a), time (t), place (w)
 - Social information:
 - Object (o), context (e), constraints (c), associations (m)
 - Intentional information:
 - Subject's: goal (g), belief (b), plan (l)
 - Behavior performance:
 - Impact (f), status (u)
- *New methods for vector-based behavior pattern analysis*

Behavioral data

- Behavioral elements hidden or dispersed in transactional data
- *behavioral feature space*

- Behavioral data modeling
- Behavioral feature space
- Mapping from transactional to behavioral data
- Behavioral data processing
- Behavioral data transformation

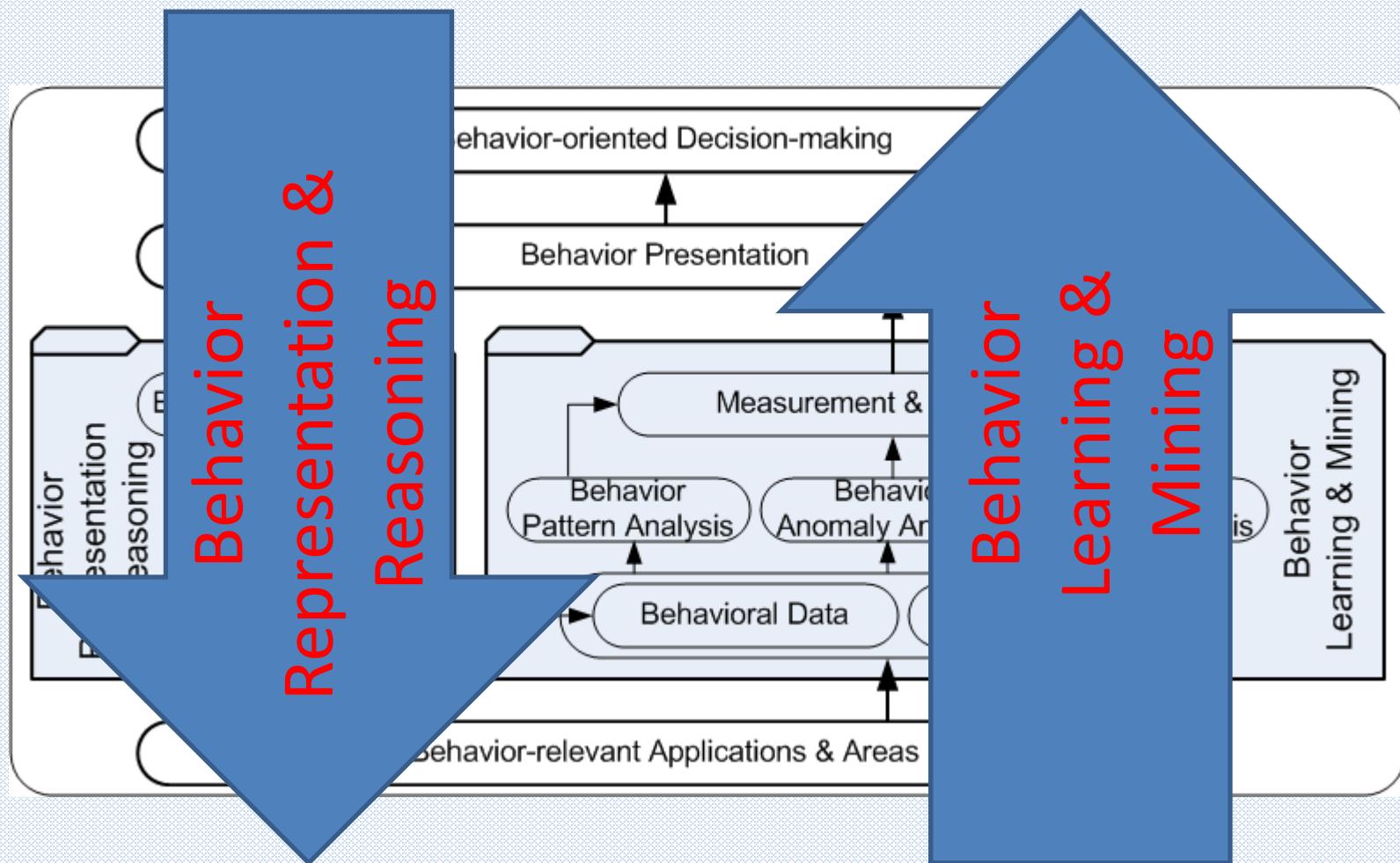


What is Behavior Informatics and Analytics?

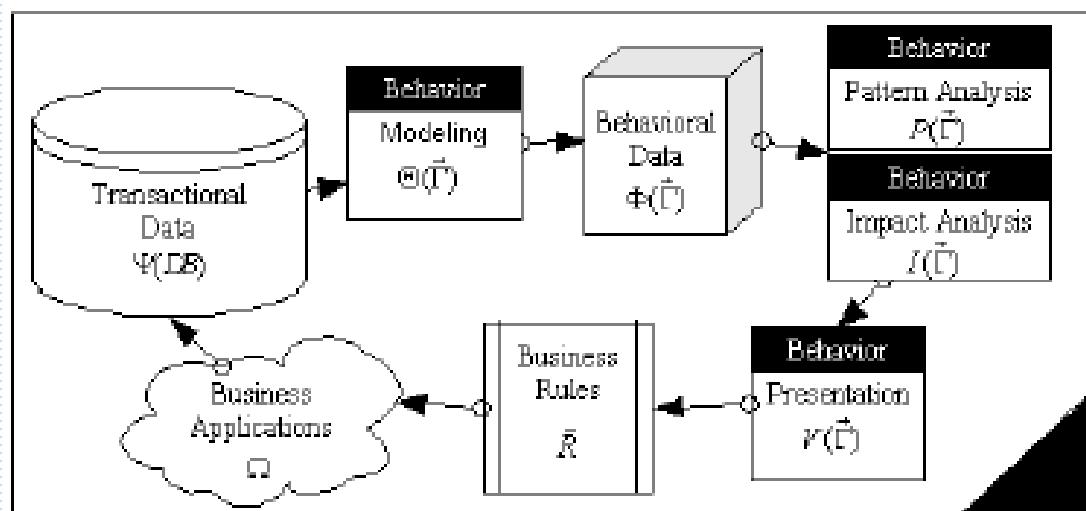
Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, *Information Science*, 180(17); 3067-3085, 2010.

www.behaviorinformatics.org

Behavior informatics – Concept Map



Behavior analysis process



$$BIA : \Psi(DB) \xrightarrow{\Theta(\vec{\Gamma})} \vec{\Gamma} \xrightarrow{\Omega, e, c, t_i()} \tilde{P} \xrightarrow{\Lambda, e, c, b_i()} \tilde{R}$$

BIA PROCESS: The Process of Behavior Informatics and Analytics

INPUT: original dataset Ψ ;

OUTPUT: behavior patterns \tilde{P} and operationalizable business rules \tilde{R} ;

Step 1: Behavior modeling $\Theta(\vec{\Gamma})$:

Given dataset Ψ ;

Develop behavior modeling method θ ($\theta \in \Theta$) with technical interestingness $t_i()$;

Employ method θ on the dataset Ψ ;

Construct behavior vector set $\vec{\Gamma}$;

Step 2: Converting to behavioral data $\Phi(\vec{\Gamma})$:

Given behavior modeling method θ ;

FOR $j = 1$ to $(count(\Psi))$

 Deploy behavior modeling method θ on dataset Ψ ;

 Construct behavior vector $\vec{\gamma}_j$;

ENDFOR

Construct behavior dataset $\Phi(\vec{\Gamma})$;

Step 3: Analyzing behavioral patterns $P(\vec{\Gamma})$:

Given behavior data $(\Phi(\vec{\Gamma}))$;

Design pattern mining method $\omega \in \Omega$;

Employ the method ω on dataset $\Phi(\vec{\Gamma})$;

Extract behavior pattern set \tilde{P} ;

Step 4: Converting behavior patterns \tilde{P} to operationalizable business rules \tilde{R} :

Given behavior pattern set \tilde{P} ;

Develop behavior modeling method Λ ;

Involve business interestingness $b_i()$ and constraints c in the environment e ;

Generate business rules \tilde{R} ;

Behavioral representation

- (Behavior modeling)
 - describing behavioral elements
 - extracting **syntactic and semantic relationships** amongst the elements
 - presentation and construction of behavioral sequences and **properties**
 - unified mechanism for describing and presenting behavioral elements, properties, behavioral impact and patterns

Behavioral impact analysis

- Behavioral instances that are associated with high impact on business processes and/or outcomes
- Modeling of behavioral impact
 - Behavior impact analysis
 - Behavioral measurement
 - Organizational/social impact analysis
 - Risk, cost and trust analysis
 - Scenario analysis
 - Cause-effect analysis
 - Exception/outlier analysis and use
 - Impact transfer patterns
 - Opportunity analysis and use
 - Detection, prediction, intervention and prevention

Behavioral pattern analysis

- Behavioral patterns without the consideration of behavioral impact
- Analyze the relationships between behavior sequences and particular types of impact

- Emergent behavioral structures
- Behavior semantic relationship
- Dynamic behavior pattern analysis
- Detection, prediction and prevention
- Demographic-behavioral combined pattern analysis
- Cross-source behavior analysis
- Correlation analysis
- Social networking behavior
- Linkage analysis
- Behavior clustering
- Behavior network analysis
- Behavior self-organization
- Exceptions and outlier mining

Behavioral Anomaly Analysis

- Abnormal behavior
- Abnormal + normal behaviors
- Abnormal group behavior

Behavioral intelligence emergence

- Behavioral occurrences, evolution and life cycles
- Impact of particular behavioral rules and patterns on behavioral evolution and intelligence emergence
- Define and model behavioral rules, protocols and relationships, and
- Their impact on behavioral evolution and intelligence emergence

Behavior networking

- **Intrinsic mechanisms** inside a network
 - behavioral rules, interaction protocols, convergence and divergence of associated behavioral itemsets
 - effects such as network topological structures, linkage relationships, and impact dynamics
- **Community** formation, pattern, dynamics and evolution

- Intrinsic mechanisms inside a network
- Behavior network topological structures
- Convergence and divergence of associated behavior
- Hidden group and community formation and identification
- Linkage formation and identification
- Community behavior analysis

Behavioral simulation

- Observe the dynamics,
- The impact of rules/protocols/patterns, behavioral intelligence emergence, and
- The formation and dynamics of social behavioral network
 - Large-scale behavior network
 - Behavior convergence and divergence
 - Behavior learning and adaptation
 - Group behavior formation and evolution
 - Behavior interaction and linkage
 - Artificial behavior system
 - Computational behavior system
 - Multi-agent simulation

Behavioral presentation

- presentation means and tools
 - describe the motivation and the interest of stakeholders on the particular behavioral data
 - traditional behavior pattern presentation
 - visual behavioral presentation
 - Rule-based behavior presentation
 - Flow visualization
 - Sequence visualization
 - Dynamic group formation
 - Visual behavior network
 - Behavior lifecycle visualization
 - Temporal-spatial relationship
 - Dynamic factor tuning, configuration and effect analysis
 - Behavior pattern emergence visualization
 - Distributed, linkage and collaborative visualization

Discussion 2: What is the behavior in your organization

- 1 Write a few keywords (dimensions and aspects), or a diagram, to explain what is behaviour on your mind

- 2 List three aspects that you believe are the most important in discussing behaviour

Behavior Modeling and Representation

Formalization and Verification of Group Behavior Interactions

Can Wang, Longbing Cao, *Senior Member, IEEE*, and Chi-Hung Chi

Abstract—Group behavior interactions, such as multirobot teamwork and group communications in social networks, are widely seen in both natural, social, and artificial behavior-related applications. Behavior interactions in a group are often associated with varying coupling relationships, for instance, conjunction or disjunction. Such coupling relationships challenge existing behavior representation methods, because they involve multiple behaviors from different actors, constraints on the interactions, and behavior evolution. In addition, the quality of behavior interactions are not checked through verification techniques. In this paper, we propose an ontology-based behavior modeling and checking system (OntoB) to explicitly represent and verify complex behavior relationships, aggregations, and constraints. The OntoB system provides both a visual behavior model and an abstract behavior tuple to capture behavioral elements, as well as building blocks. It formalizes various intra-coupled interactions (behaviors conducted by the same actor) via transition systems (TSs), and inter-coupled behavior aggregations (behaviors conducted by different actors) from temporal, inferential, and party-based perspectives. OntoB converts a behavior-oriented application into a TS and temporal logic formulas for further verification and refinement. We demonstrate and evaluate the effectiveness of the OntoB in modeling multirobot behaviors and their interactions in the Robocup soccer competition game. We show, that the OntoB system can effectively model complex behavior interactions, verify and refine the modeling of complex group behavior interactions in a sound manner.

Index Terms—Behavior interaction, coupling relationship, group behavior, model checking.

I. INTRODUCTION

BEHAVIOR refers to the action or reaction of any material under given circumstances and environment. It is intrinsic in many areas, and behavior analysis has become a fundamental topic which has been increasingly investigated as an essential activity in many fields, from social and behavioral sciences to computer science [1], [2]. In Google, the keyword “behavior” attracts 379 000 000 hits while “behavior interaction” achieves 202 000 000 results, searched on 4th Dec. 2014.

Manuscript received February 14, 2014; revised June 10, 2014; accepted November 8, 2014. Date of publication February 24, 2015; date of current version July 15, 2015. This paper was recommended by Associate Editor Y. Wang.

C. Wang and C.-H. Chi are with the Digital Productivity Flagship, Commonwealth Scientific and Industrial Research Organisation, Sandy Bay, TAS 7005, Australia (e-mail: caswang613@gmail.com).

L. Cao is with the Advanced Analytics Institute, University of Technology, Sydney, NSW 2007, Australia.

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSMC.2015.2399862

In both natural and social sciences and applications, multiple behaviors from one or multiple actors often interact with one another, which are called coupled behaviors or group behavior interactions. They play important roles in group-based activities such as social networking and multirobot teamwork. These coupled behaviors and behavior interactions may form interior driving forces that shape underlying businesses, such as in online community and social networks [3], or may even cause challenging problems like group-based manipulation by a group of traders [4] or serious traffic jams resulting from haphazard interactions between vehicles traveling in different directions toward an intersection. With the deepening and widening of complex networking, coupled behaviors, or group behavior interactions are increasingly seen in both mainstream and emerging situations, in particular, in enterprise applications, organizations, complex systems, online, and social communities.

We illustrate coupled behaviors and behavior interactions using the example of multirobot soccer game in Fig. 1. As shown in Fig. 1, two teams participate in a Robocup soccer competition (<http://www.robocup.org/>) with four Sony AIBO robots in each group. The robot players operate on their own without any external control, either by humans or by computers. They communicate with each other by wireless or by using the speakers and microphones. Their interactions include the collaborations among different actions of the same robot, e.g., one of the robots kicks the ball after it gets a message; and distinct operations conducted by different robots, such as sending messages between different players. As shown in the scenario described by Ros and Veloso [5], a team of robots intelligently cooperate with one another and self-adjust their own activities; the successful task execution and problem resolution rely on the proper implementation of an individual robot’s activities as well as collaborative interactions between robots. If a robot undertakes tasks without appropriate arrangement and coordination with the other robots, the Robocup is likely to be unsuccessful, even though every robot performs perfectly. This example shows that group actors and behaviors by the same or different actors within the group are often coupled in different forms of interactions [6], and it is essential to identify, represent, and verify how the robots interact to ensure the performance of a multirobot system.

To enable the above behavior interaction-oriented systems to work properly, a fundamental task is to develop effective behavior representation tools to capture, formalize, and verify behavioral elements, coupling relationships, and interactions between behaviors, in both qualitative and quantitative

Behavior representation

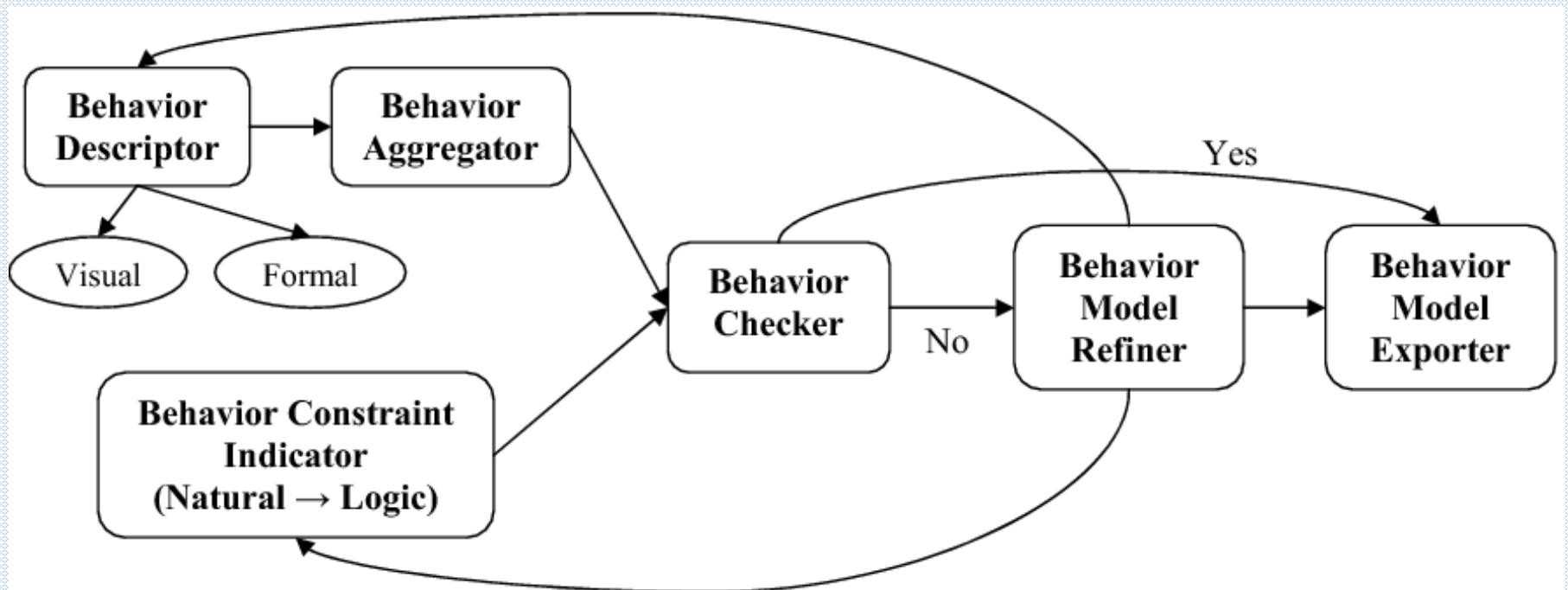
Can Wang, Longbing Cao, Chi-Hung Chi: **Formalization and Verification of Group Behavior Interactions**. *IEEE T. Systems, Man, and Cybernetics: Systems* 45(8): 1109–1124 (2015)

Can Wang, and Longbing Cao.
Modeling and Analysis of Social Activity Process, in Longbing Cao and Philip S Yu (eds) *Behavior Computing*, 21–35, Springer, 2012

Issues Addressed

- How to represent behaviors?
- Behavior ontology
- Behavior process and interaction
- Behavior interaction relationship
- How to model check behavior models built?
- Case studies of behavior modeling

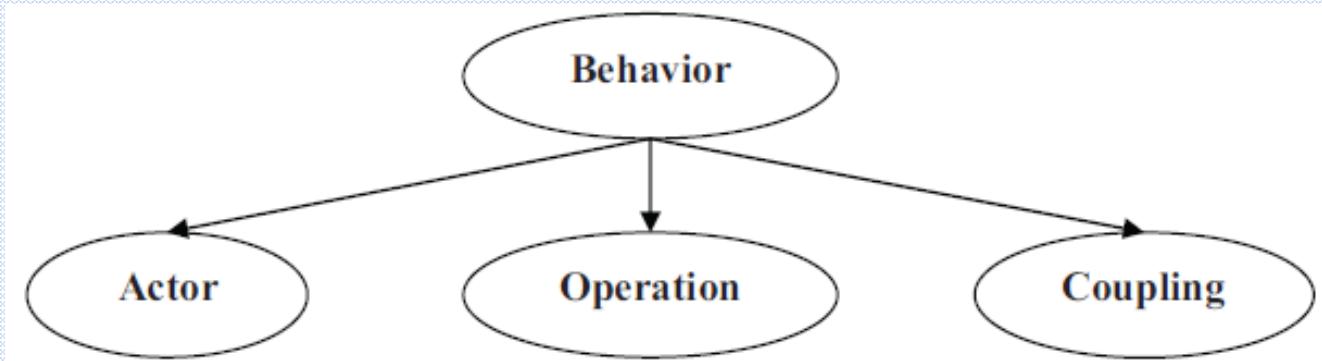
Behavior Modeling and Checking Framework



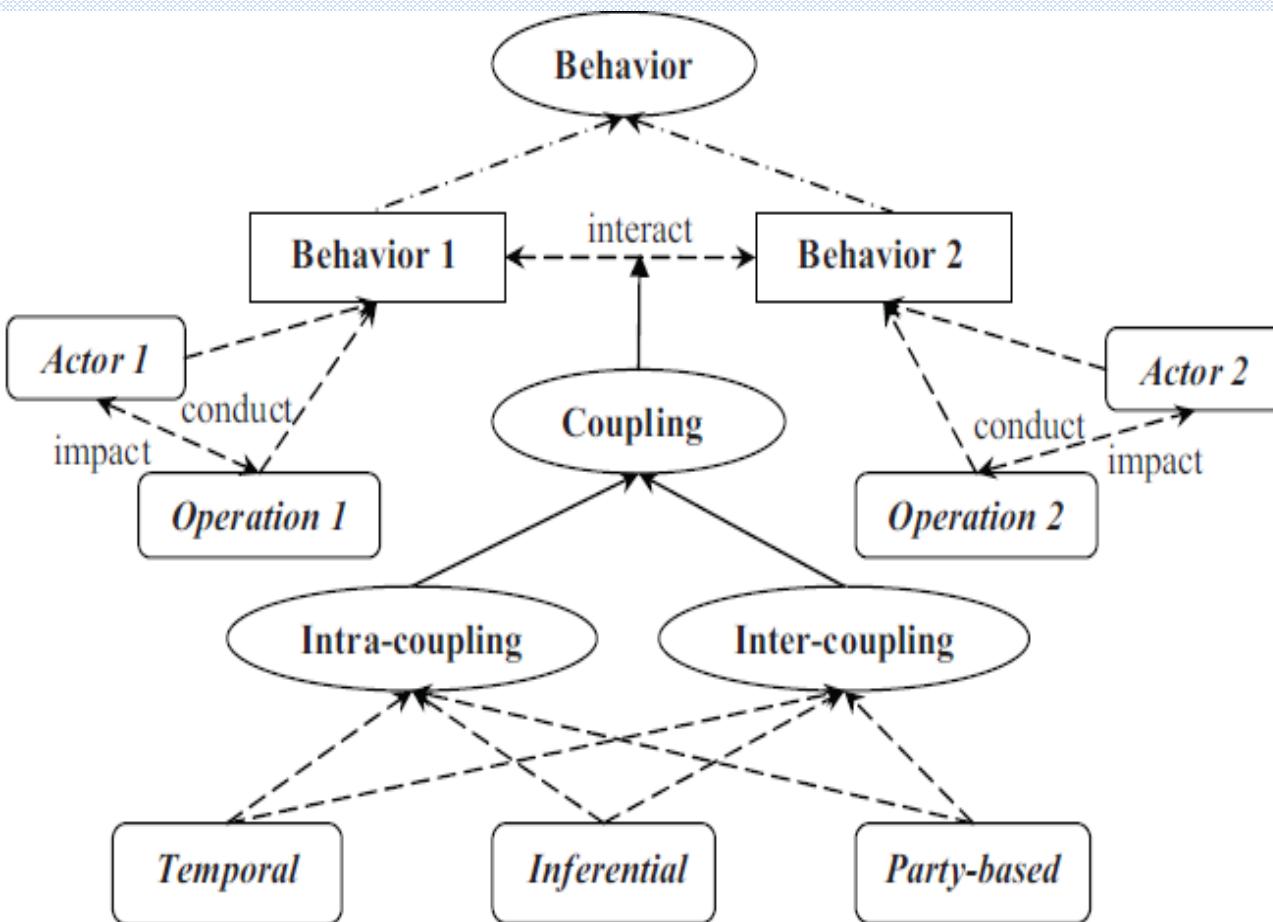
Ontology-based Behavior Modeling and Checking

Behavior Visual Descriptor

- **Actor**: refers to the subject(s) or object(s) of a behavior, for example, organizations, departments, systems, agents and people involved in an activity or activity sequence.
- **Operation**: represents activities, actions or events in a behavior or behavior sequence.
- **Coupling**: refers to the interaction between behaviors, including connections between actors and/or operations of either one or multiple actors.

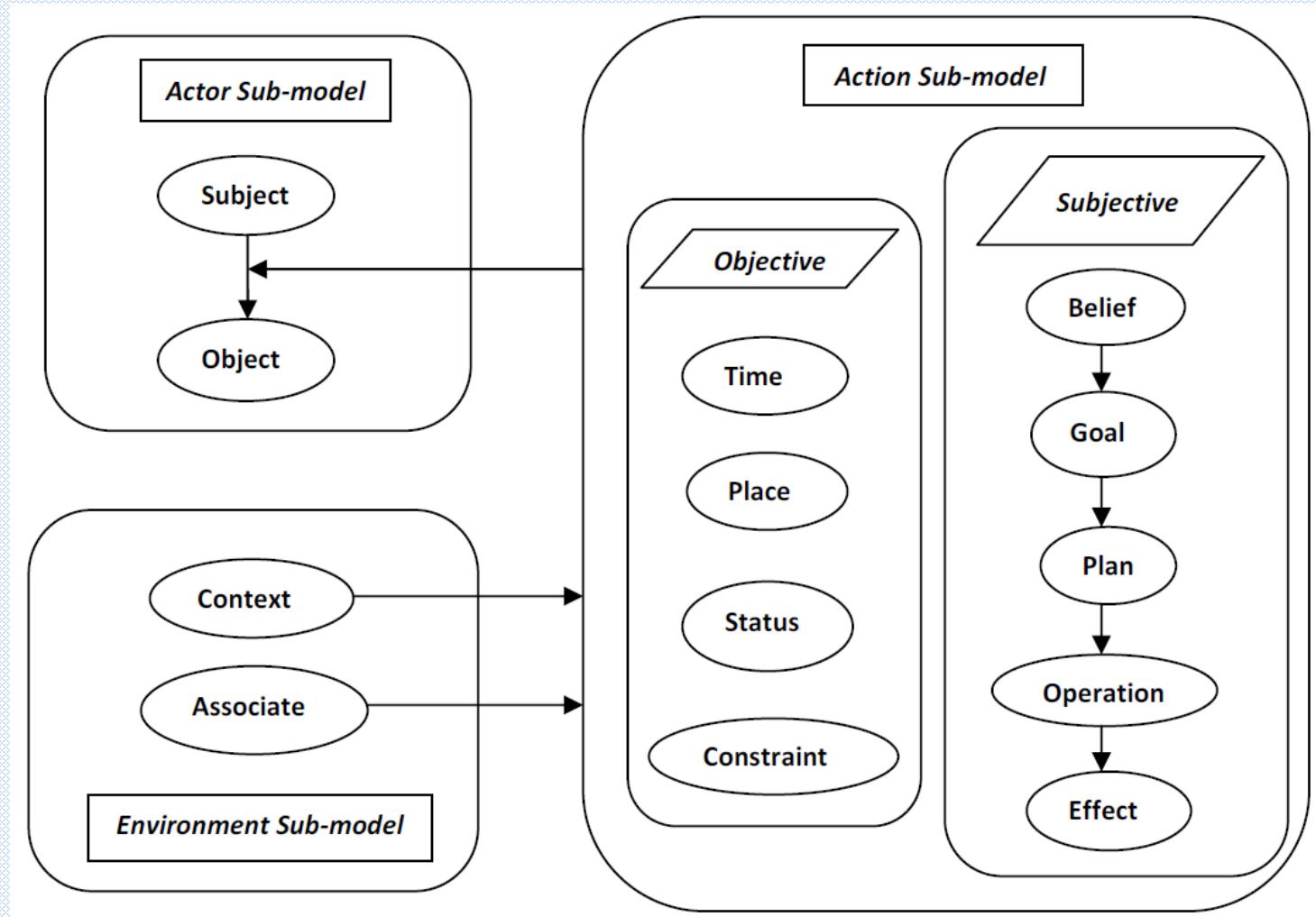


Behavior Visual Descriptor

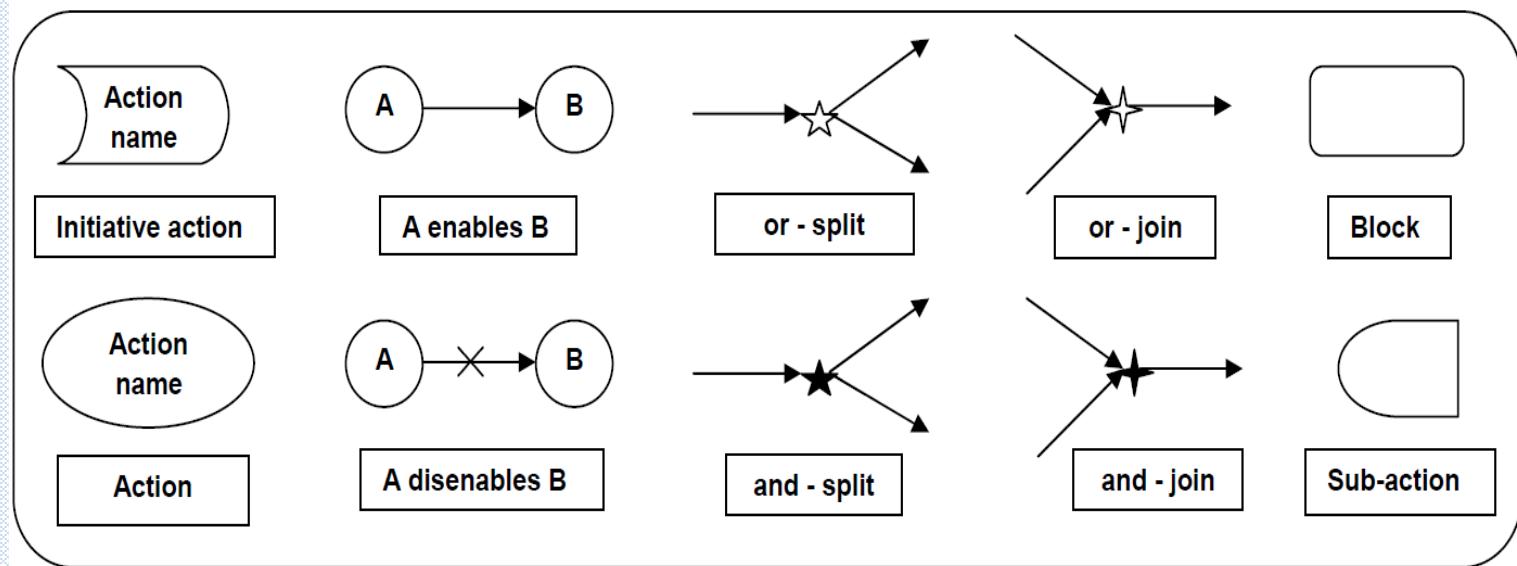


- **Instance Of** Connecting instances (in Rectangle) to their corresponding classes
- **Subclass Of** Linking a subclass (in Oval) to its parent class
- **Object Property** Denoting the relationships between instances, between an object and its properties (in Rounded Rectangle), or between properties.

Overall Single Behavior Model

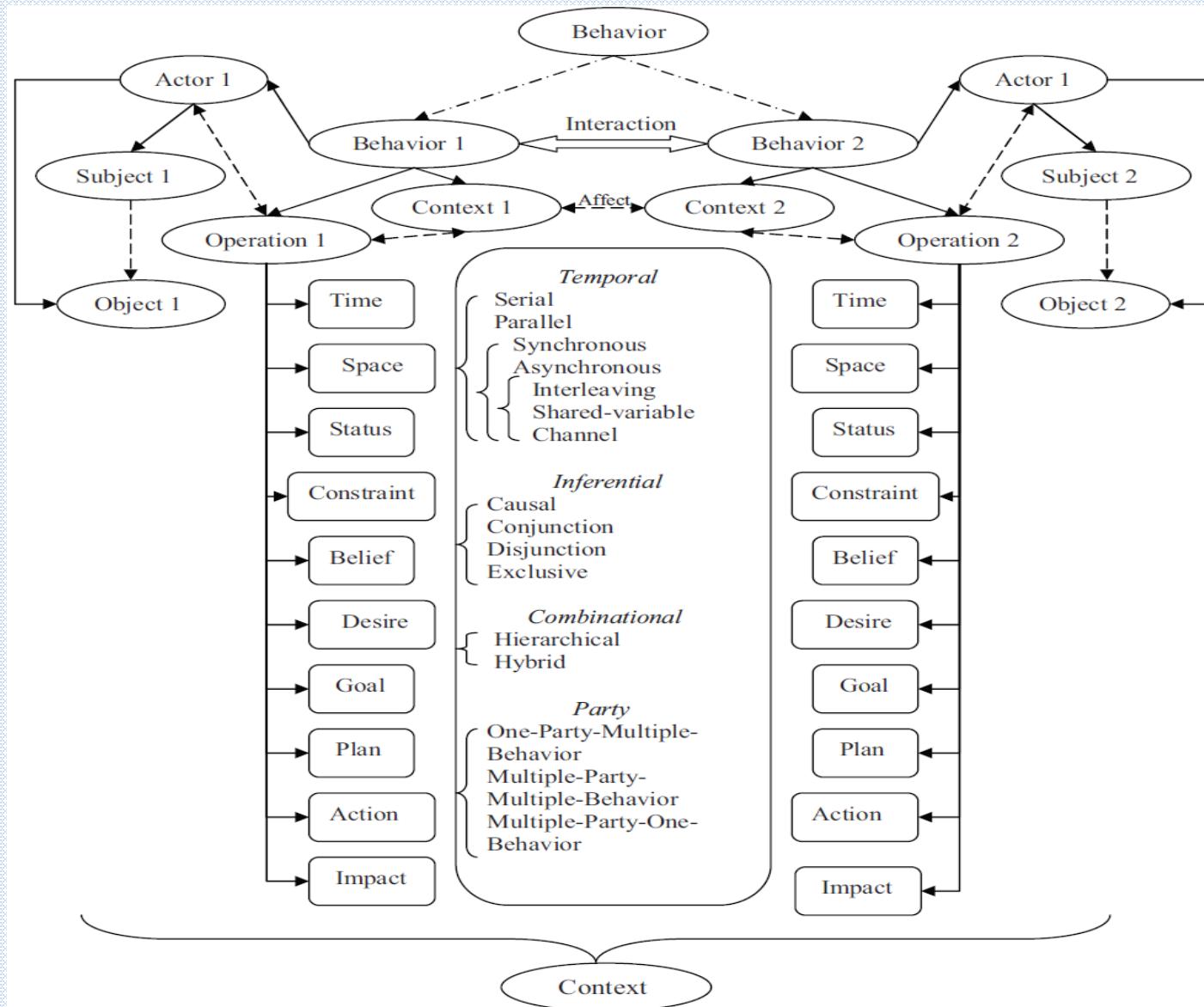


Relationship Sub-model



Relationship	<i>enable</i>	<i>disenable</i>	<i>or-split</i>	<i>and-split</i>	<i>or-join</i>	<i>and-join</i>
Logic Form	$a \rightarrow b$	$\neg(a \rightarrow b)$	$a \rightarrow (b \vee c)$	$a \rightarrow (b \wedge c)$	$(a \vee b) \rightarrow c$	$(a \wedge b) \rightarrow c$

Relationships between Agent Behaviors



Coupling Relationships

Coupling Relationships

Perspectives

Temporal

Inferential

Party-based

Serial Coupling

Parallel coupling

Synchronous relationship

Asynchronous coupling

Interleaving

Shared-variable

Channel system

Causal Coupling

Conjunction Coupling

Disjunction Coupling

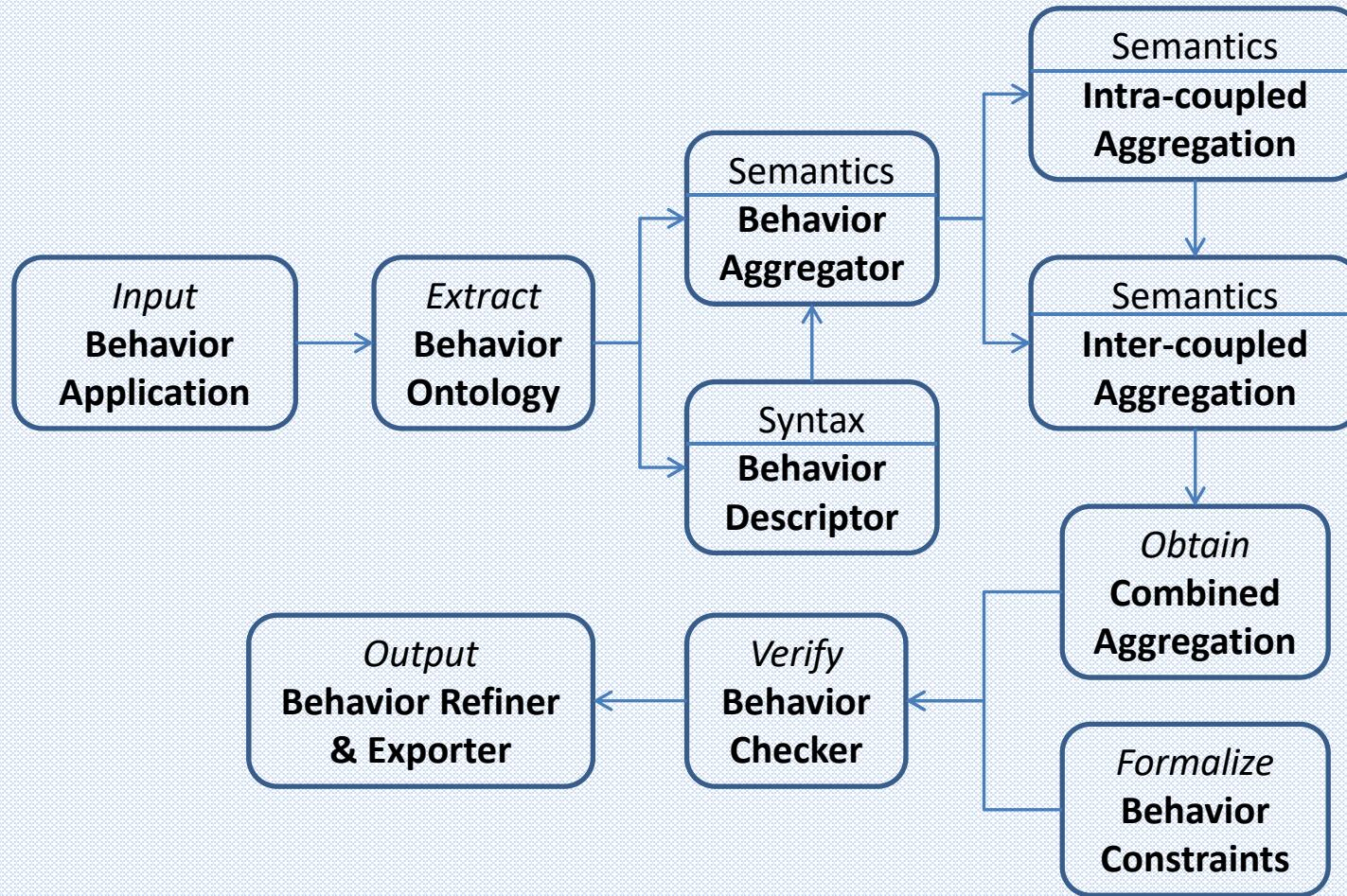
Exclusive Coupling

One-Party-Multiple-Operation

Multiple-Party-One-Operation

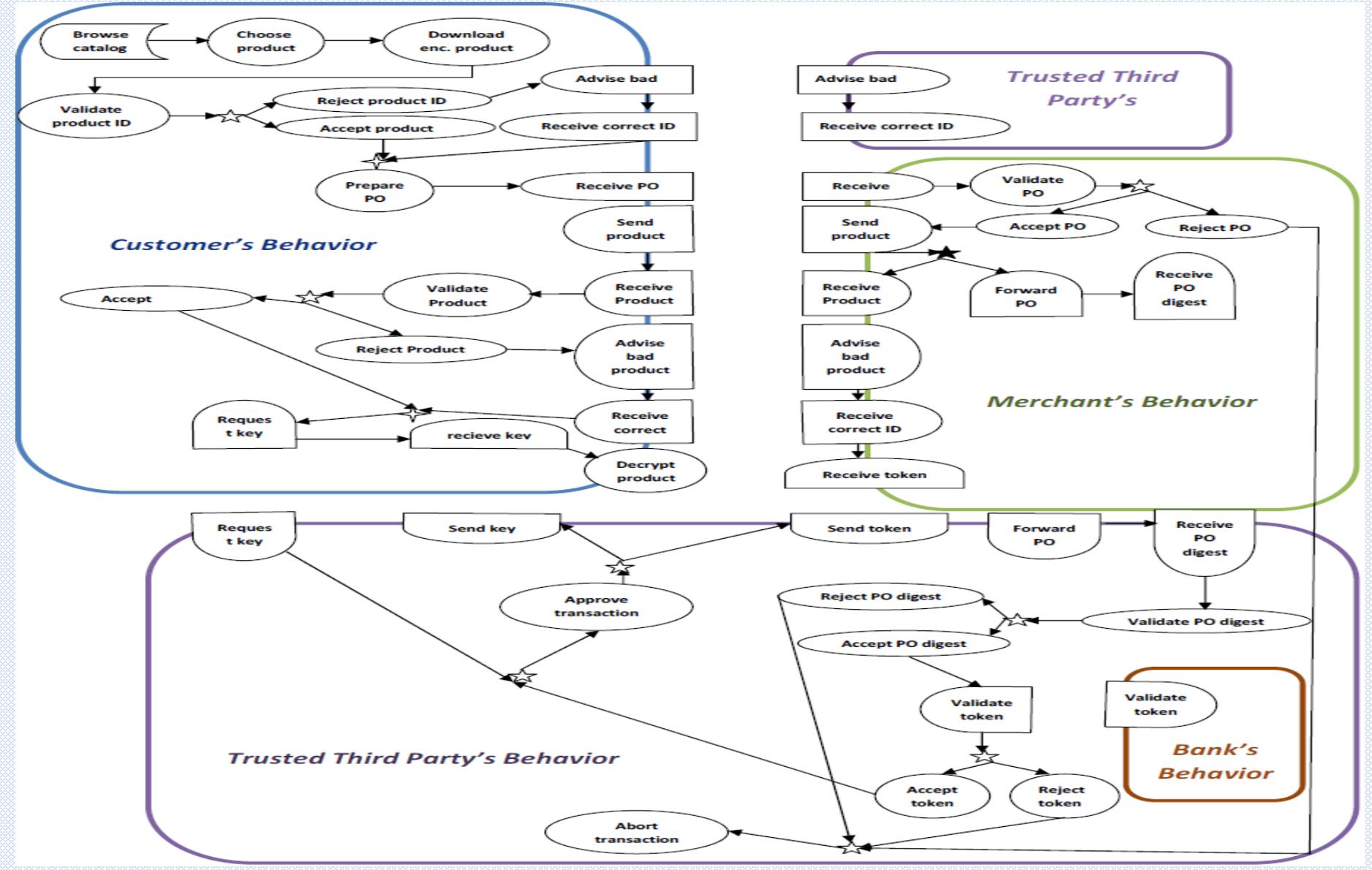
Multiple-Party-Multiple-Operation

Group Behavior Representation and Verification

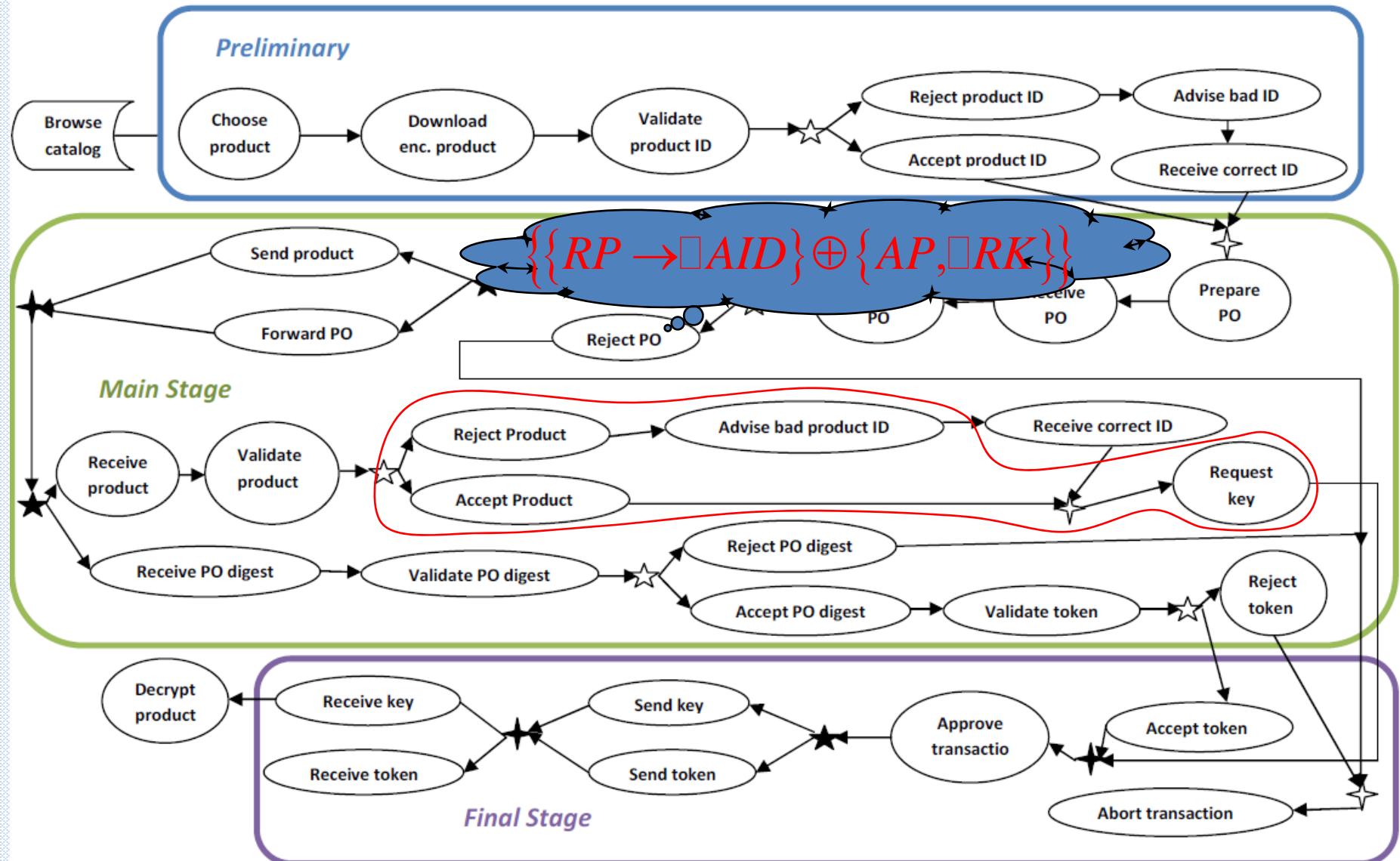


Case study of behavior representation

Graphical Action Sub-model of Online Shopping based on Actor's Roles



Graphical Action Sub-model of Online Shopping based on *Stages*



Exercise 3: Model behaviour in your business

Behaviour understanding

Organization: _____

Business problem: _____



Part II.

Behavior Pattern Analysis

Learning Objectives

- Behavior patterns
- High impact behavior patterns
- Behavior pattern combinations
- Combined behavior sequences associated with impact
- Manage impact of individual or group behaviors

Impact-oriented Behavior Analysis

Longbing Cao. Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, *IEEE Trans. on Knowledge and Data Engineering*, 20(8): 1053-1066, 2008.

Mining Impact-Targeted Activity Patterns in Imbalanced Data

Longbing Cao, Senior Member, IEEE, Yanchang Zhao, Member, IEEE, and Chengqi Zhang, Senior Member, IEEE

Abstract—Impact-targeted activities are rare but they may have a significant impact on the society. For example, isolated terrorism activities may lead to a disastrous event, threatening the national security. Similar issues can also be seen in many other areas. Therefore, it is important to identify such particular activities before they lead to having a significant impact to the world. However, it is challenging to mine impact-targeted activity patterns due to their imbalanced structure. This paper develops techniques for discovering such activity patterns. First, the complexities of mining imbalanced impact-targeted activities are analyzed. We then discuss strategies for constructing impact-targeted activity sequences. Algorithms are developed to mine frequent positive-impact-oriented ($P \rightarrow T$) and negative-impact-oriented ($P \rightarrow T'$) activity patterns, sequential impact-connected activity patterns (P is frequently associated with both patterns $P \rightarrow T$ and $P \rightarrow T'$ in separated data sets), and sequential impact-reversed activity patterns (both $P \rightarrow T$ and $P \rightarrow T'$ are frequent). Activity impact modeling is also studied to quantify the pattern impact on business outcomes. Social security debt-related activity data is used to test the proposed approaches. The outcomes show that they are promising for information and security informatics (ISI) applications to identify impact-targeted activity patterns in imbalanced data.

Index Terms—Clustering, classification, association rules, data mining.

1 INTRODUCTION

IN the emerging research on information and security informatics (ISI) [25], [26], [9], [10], [13], activity [5], [39] and event analysis [16], [11], [27], [30], [35], [24] have been the key research objects. Impact-targeted activities specifically refer to those activities associated with or leading to a specific impact of interest to the business world. The impact can be an event, a disaster, a government-customer debt, or any other interesting entities. For instance, a series of dispersed and isolated terrorism activities may finally result in a disastrous event [21], [23], [27]. In the social security network [6], [7], [39], [5], [40], a large volume of isolated fraudulent and criminal customer activities can result in a large amount of government-customer debt. For example, in the 5.4-billion government-customer activity transactions per year in Australia, the government social security agency Centrelink accumulates around one billion of customer debt from the delivery of a total of 64 billion payments to 6.5 million eligible customers in the financial year 2004–2005 [7]. Similar problems can be widely seen from other emerging areas such as distributed criminal activities, well-organized separated activities or events threatening the national security and homeland security, and self-organized computer network crimes [9], [12], [28]. Activities in traditional fields such as taxation, insurance services, telecommunication network malfunction, drug disease associations, customer contact centers, and healthcare services may also result in an impact on related organizations or business objectives.

Therefore, it is important to specifically analyze such impact-targeted activities to find out knowledge about what activity patterns are associated with certain types of the impact of interest to specific domain targets and what activity patterns are more likely to lead to the targeted impact. As a result, the findings may support related decision making by providing deep knowledge about the dynamics of impact-targeted activities, the causes of activities leading to certain types of impact, and possible solutions for preventing and minimizing the impact of activities on the society or business outcomes. For instance, in analyzing activities in the social security network, we identify those activities or activity sequences that are more likely to lead to government-customer debt. The resulting evidence and predictors can thus inform relevant officers of the risk of certain ongoing actions or activity sequences resulting in debt. As a result, a potential occurrence of debt can be prevented or minimized. Business decision making and processes, as well as governmental service and policy objectives, can thus be improved and enhanced.

However, impact-targeted activities present some special complexities, which cannot be well handled by existing information processing technologies, for instance, traditional event detection, event and process mining [11], [31], [16], and sequence analysis [18] in both ISI and data mining areas [19]. This is due to the following characteristics of impact-targeted activities. First, impact-targeted activities specifically focus on those activities that have resulted or will result in an impact on business situations. This is normally not concerned with traditional data mining such as sequence or event mining. Second, impact-targeted activities consist of only a very small fraction of the whole activity population. They are normally rare and dispersed in a large activity and customer populations. Nevertheless, it is them that lead to significant effects or even disasters to the society or related business. For instance, only around 4 percent of

* The authors are with the Department of Software Engineering, University of Technology, Sydney, PO Box 123 Broadway, New South Wales, Australia 2007. E-mail: {lba, yczhan, chengqi}@ut.edu.au

Manuscript received 8 May 2006; revised 11 June 2007; accepted 19 June 2007; published online 12 July 2007.

For information on obtaining reprints of this article, please send e-mail to: tskde@computer.org, and reference IEEECS Log Number TKDTS14-0258-0506. Digital Object Identifier no. 10.1109/TKDE.2007.190653.

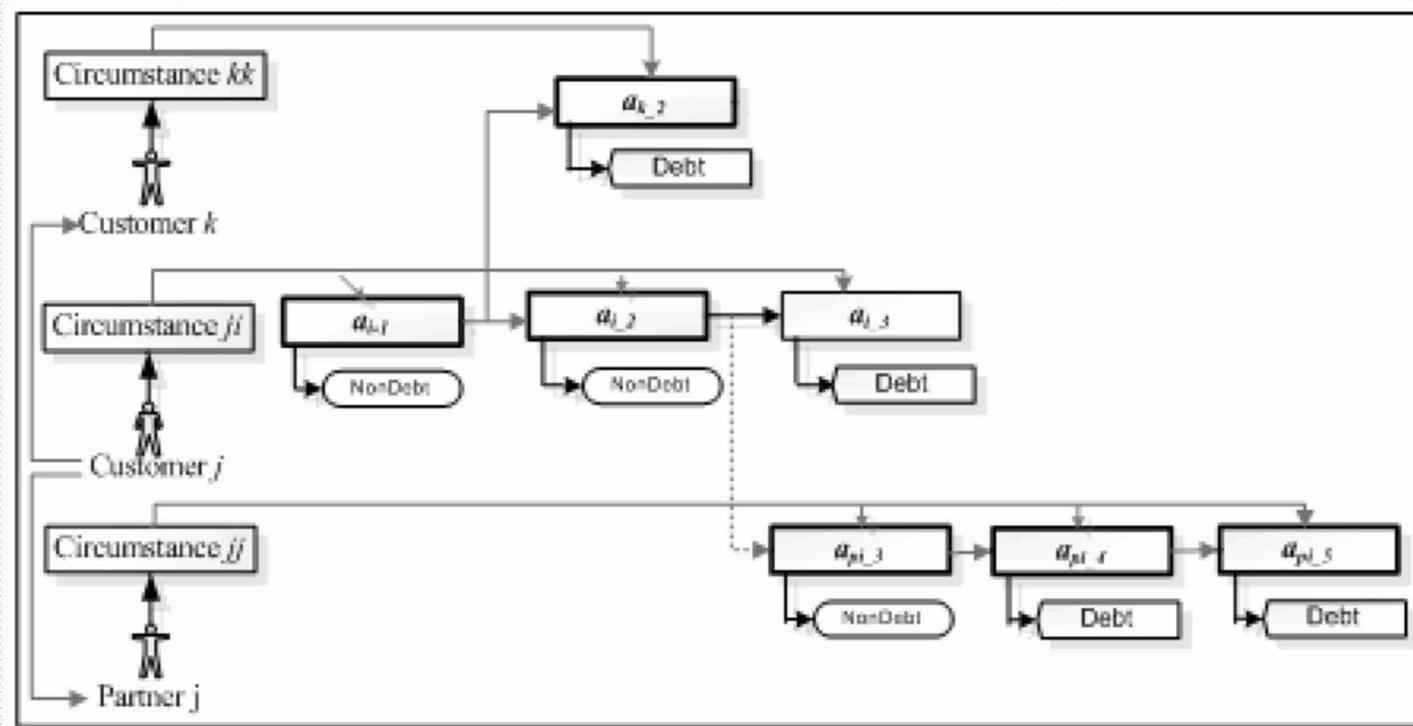
Behavior impact analysis

Longbing Cao, Zhao Y., Zhang,
C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008.

Issues Addressed

- What is impact of behavior?
- How to model behavior impact?
- How to construct impact-based behavior sequences?
- How to identify high impact behavior sequences?
- How to identify combined behavior sequences associated with impact?
- How to manage behavior patterns through combined impact-targeted behavior sequences?

Coupled impact-oriented behaviors



Risk/Impact Definition

- *Risk* is defined as a feasible detrimental outcome of an activity or action (e.g., launch or operation of a spacecraft) subject to hazard(s)
- (1) *magnitude (or severity)* of the adverse consequence(s) that can potentially result from the given activity or action, and
- (2) *likelihood* of occurrence of the given adverse consequence(s).

Impact

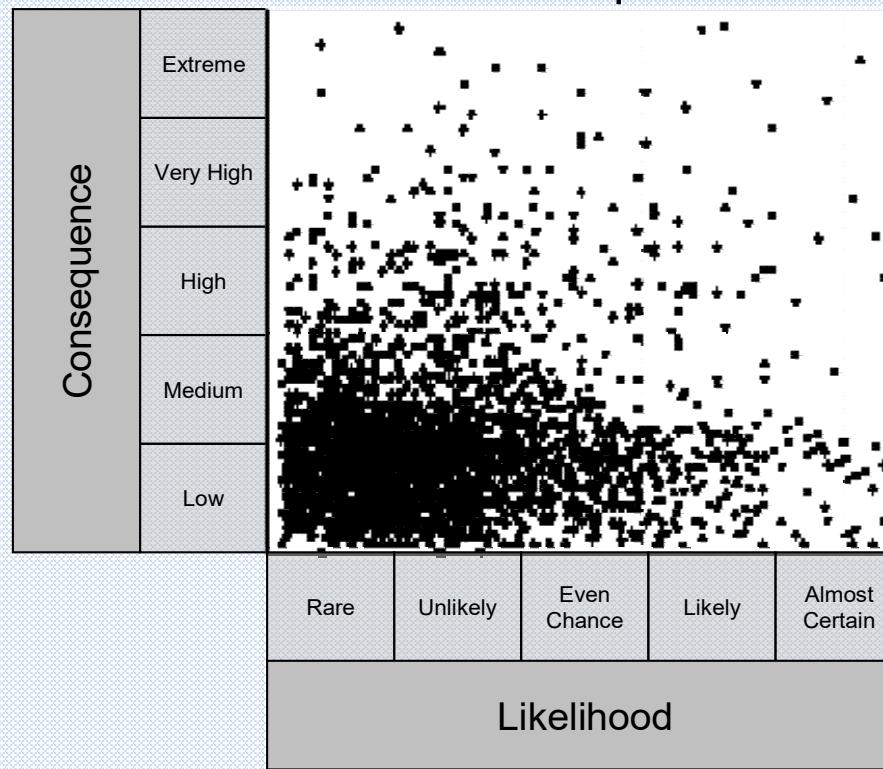
- Business impact of behavior
 - Consequence:
 - Fraud
 - Debt
 - Exception ...
 - Magnitude:
 - Positive/negative
 - Multi-level
 - Ratio
 - Probabilistic

Probabilistic Risk Assessment

- Causes/Initiators:
 - What can go wrong with the studied technological entity, or what are the *initiators or initiating events (undesirable starting events) that lead to adverse consequence(s)*?
- Effects/Consequences:
 - What and how severe are the potential detriments, or the adverse *consequences that the technological entity may be eventually subjected to as a result of the occurrence of the initiator*?
- Functions(cause, effect):
 - How likely to occur are these undesirable consequences, or what are their *probabilities or frequencies*?

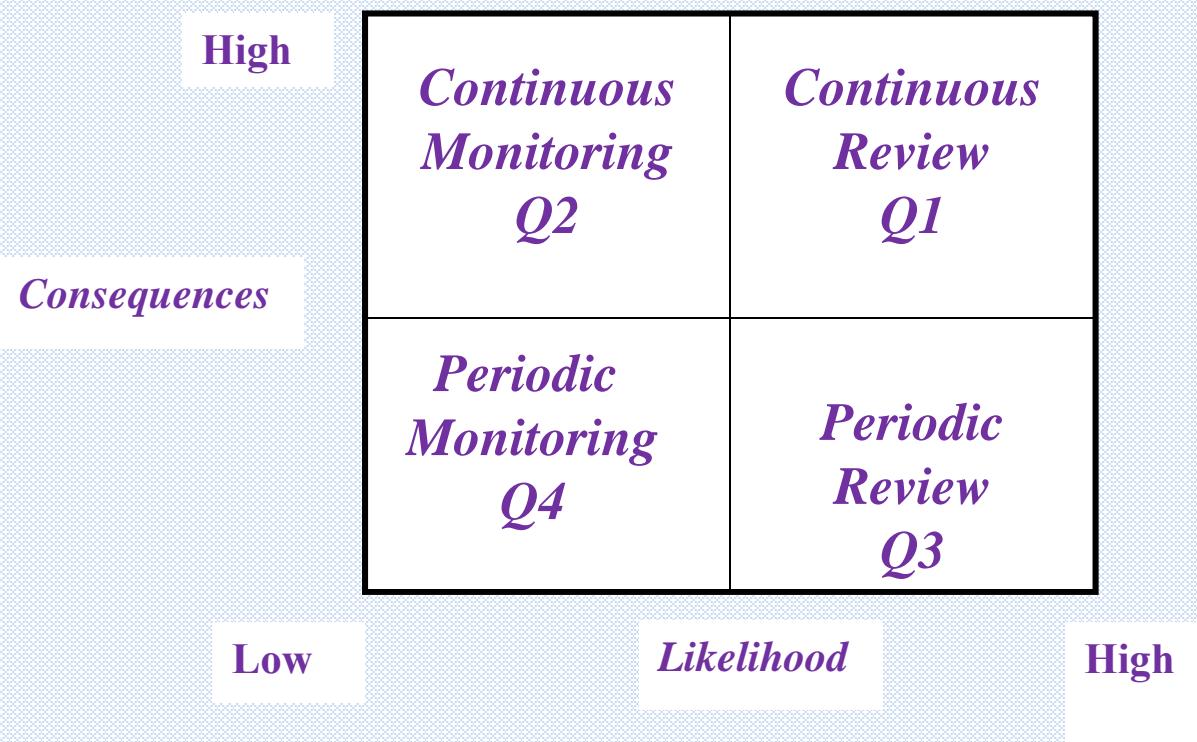
Expected Distribution of Clients with Risks

Most clients are relatively small.
Few have extreme consequences



Most clients are compliant.
Relatively few are deliberately non-compliant

Risk Differentiation Framework



Behavior impact modeling

- Impact measuring
 - Cost
 - Cost-sensitive
 - Profit
 - Cost-benefit
 - Risk score
 - ...
- Impact evolution
 - Positive → Negative
 - Negative → Positive

- Risk of a pattern, eg.

$$Risk(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{TotalCost(P)}$$

$$AvgCost(P \rightarrow T) = \frac{Cost(P \rightarrow T)}{Cnt(P \rightarrow T)}$$

Impact-Targeted Activity Mining

- Frequent **impact-oriented** activity patterns
- Frequent **impact-contrasted** activity patterns
- Sequential **impact-reversed** activity patterns

Here:

Impact → Debt, Fraud, Risk ...

Impact-Oriented Activity Patterns

$$\{P \rightarrow T\} \text{ or } \{P \rightarrow \bar{T}\} \quad (P \rightarrow \bar{T}, \text{ or } \bar{P} \rightarrow \bar{T})$$

- frequent *positive* impact-oriented (T) activity patterns
 - $P \rightarrow T$, or
 $\bar{P} \rightarrow T$
- frequent *negative* impact-oriented () activity patterns
 - $P \rightarrow \bar{T}$
 $\bar{P} \rightarrow \bar{T}$

P is an activity sequence, ($P = \{a_i, a_{i+1}, \dots\}, i=0, 1, \dots\}$).

Impact-Contrasted Activity Patterns

$$\{P \rightarrow T, P \rightarrow \bar{T}\} \quad \{P \rightarrow \bar{T}, P \rightarrow T\}$$

- **Pattern:** P is of high significance in positive impact dataset, and of low significance in negative impact dataset, or vice versa.
- *Positive impact-contrasted pattern*
 $P_{T\bar{T}}: \{P \rightarrow T, P \rightarrow \bar{T}\}$
- *Negative impact-contrasted pattern*
 $P_{\bar{T}T}: \{P \rightarrow \bar{T}, P \rightarrow T\}$

Impact-Reversed Activity Patterns

$\{P \rightarrow T\}$ $\{PQ \rightarrow \bar{T}\}$ $\{P \rightarrow \bar{T}\}$ $\{PQ \rightarrow T\}$

- *Sequential impact-reversed activity pattern pair*

— *underlying pattern:*

$\{P \rightarrow T\}$

$\{P \rightarrow \bar{T}\}$



— *derivative pattern:*

$\{PQ \rightarrow \bar{T}\}$

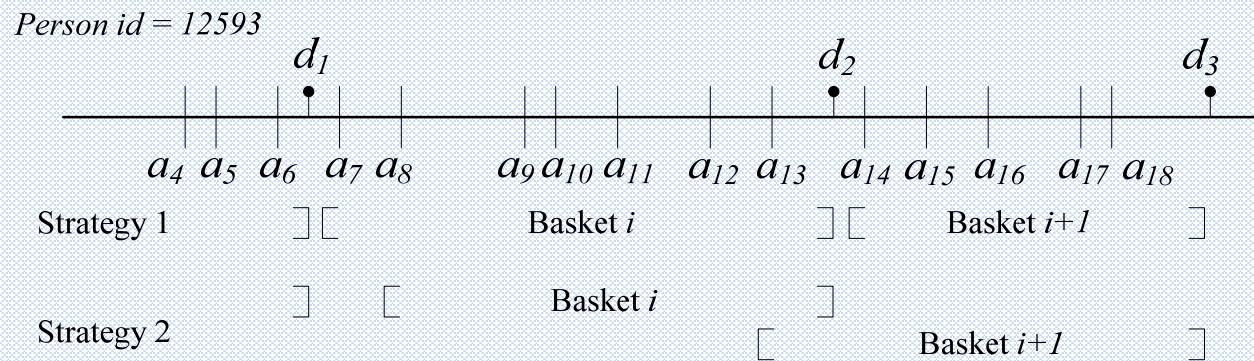
$\{PQ \rightarrow T\}$

Raw Data

- Data:
 - Time: [1/1/06, 31/3/06]
 - No. of activity transactions: 15,932,832
 - No. of customers: 495,891
 - No. of debts: 30,546

Constructing Activity Baskets and Sequences

- **Positive-impact** activity sequences: the activities before a debt are put in a basket. E.g., $\{a_8, a_9, a_{10}, a_{11}, a_{12}, a_{13}, d_2\}$, $\{a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, d_3\}$



- **Negative-impact** activity sequences
A virtual activity “NDT” is created for those customers have never had a debt.

Examples of Debt/Non-Debt Activity Sequences

Table 1. Example of an activity sequence associated with a debt from target dataset
a15, a9, a18, a19, a16, a9, DET

ACTIVITY CODE	START DATE	TIME
a_{15}	15/02/2006	13:34:05
a_9	16/02/2006	16:26:16
a_{18}	16/02/2006	16:26:17
a_{19}	20/02/2006	16:12:35
a_{16}	28/02/2006	11:27:50
a_9	1/03/2006	13:50:03
Debt	1/03/2006	23:59:59

Table 2. Example of an activity sequence related to non-debt from non-target dataset
a14, a16, a1, a20, a14, a21, a22, NDT

ACTIVITY CODE	START DATE	TIME
a_{14}	6/02/2006	2:19:37
a_{16}	6/02/2006	10:21:50
a_1	7/02/2006	3:51:07
a_{20}	7/02/2006	4:44:48
a_{14}	7/02/2006	9:48:59
a_{21}	8/02/2006	10:03:13
a_{22}	15/02/2006	13:55:39
No-Debt	15/02/2006	23:59:59

Frequent Debt-Targeted Activity Patterns

$$\{P \rightarrow T\} \text{ or } \{P \rightarrow \bar{T}\} \quad (P \rightarrow \bar{T}, \text{ or } \bar{P} \rightarrow \bar{T})$$

Patterns $P \rightarrow T$	$Supp_D(P)$	$Supp_D(T)$	$Supp_D(P \rightarrow T)$	Confidence	Lift	$AvgAmt$ (cents)	$AvgDur$ (days)	$risk_{amt}$	$risk_{dur}$
$a_1, a_2 \rightarrow T$	0.0015	0.0364	0.0011	0.7040	19.4	22074	1.7	0.034	0.007
$a_3, a_1 \rightarrow T$	0.0018	0.0364	0.0011	0.6222	17.1	22872	1.8	0.037	0.008
$a_1, a_4 \rightarrow T$	0.0200	0.0364	0.0125	0.6229	17.1	23784	1.2	0.424	0.058
$a_1 \rightarrow T$	0.0626	0.0364	0.0147	0.2347	6.5	23281	2.0	0.490	0.111
$a_6 \rightarrow T$	0.2613	0.0364	0.0133	0.0511	1.4	18947	7.2	0.362	0.370
$a_4 \rightarrow T$	0.1490	0.0364	0.0162	0.1089	3.0	21749	3.2	0.505	0.203
$a_5 \rightarrow T$	0.1854	0.0364	0.0139	0.0755	2.1	18290	6.2	0.363	0.334
$a_7 \rightarrow T$	0.1605	0.0364	0.0113	0.0706	1.9	19090	6.8	0.310	0.300

High impact behaviour analysis

TABLE 8
Common Frequent Sequential Patterns in Separate Data Sets

a_5	0.382	0.178	0.204	2.15	-0.204	0.47	18290	6.2	0.363	0.334	y_Time
a_7	0.312	0.154	0.157	2.02	-0.157	0.50	19090	6.8	0.310	0.300	24:13
a_6	0.367	0.257	0.110	1.43	-0.110	0.70	18947	7.2	0.362	0.370	
a_{14}	0.903	0.684	0.219	1.32	-0.219	0.76	19251	6.6	0.905	0.840	3:55

TABLE 9
Impact-Reversed Sequential Activity Patterns in Separate Data Sets

a_{14}, a_{15}	0.665	0.574	0.231	Underlying sequence (P)	Impact 1	Derivative activity Q	Impact 2	Cir	Cps	Local support of $P \rightarrow \text{Impact 1}$	Local support of $PQ \rightarrow \text{Impact 2}$	
a_{15}, a_{15}	0.539	0.373	0.167									
a_{16}, a_{14}	0.479	0.402	0.076									
a_{14}, a_{16}	0.441	0.393	0.049									
a_{16}, a_{16}	0.367	0.410	-0.043	a_{14}	\bar{T}	a_4	T	2.5	0.013	0.684	0.428	
a_{14}, a_{14}, a_{15}	0.477	0.257	0.220	a_{16}	\bar{T}	a_4	T	2.2	0.005	0.597	0.147	
a_{14}, a_{15}, a_{14}	0.435	0.255	0.179	a_{14}	\bar{T}	a_5	T	2.0	0.007	0.684	0.292	
a_{16}, a_{14}, a_{14}	0.361	0.267	0.093	a_{16}	\bar{T}	a_7	T	1.8	0.004	0.597	0.156	
a_{16}, a_{14}, a_{16}	0.265	0.255	0.010	a_{14}	\bar{T}	a_7	T	1.7	0.005	0.684	0.243	
				a_{15}	\bar{T}	a_5	T	1.7	0.007	0.567	0.262	
				*****	a_{14}, a_{14}	\bar{T}	a_4	T	2.3	0.016	0.474	0.367
					a_{16}, a_{14}	\bar{T}	a_5	T	2.0	0.006	0.402	0.133
					a_{14}, a_{16}	\bar{T}	a_5	T	2.0	0.005	0.393	0.118
					a_{16}, a_{15}	\bar{T}	a_5	T	1.8	0.006	0.339	0.128
					a_{15}, a_{14}	\bar{T}	a_5	T	1.7	0.007	0.381	0.179
					a_{16}, a_{14}	\bar{T}	a_7	T	1.6	0.004	0.402	0.108
					a_{14}, a_{16}, a_{14}	\bar{T}	a_{15}	T	1.2	0.005	0.248	0.188
					a_{16}, a_{14}, a_{14}	\bar{T}	a_{15}	T	1.2	0.005	0.267	0.220

Combined Behavior Pattern Analysis

Combined mining: Analyzing object and pattern relations for discovering and constructing complex yet actionable patterns

Longbing Cao*



Combined mining is a technique for analyzing object relations and pattern relations, and for extracting and constructing actionable knowledge (patterns or exceptions). Although combined patterns can be built within a single method, such as combined sequential patterns by aggregating relevant frequent sequences, this knowledge is composed of multiple constituent components (the left hand side) from multiple data sources, which are represented by different feature spaces, or identified by diverse modeling methods. In some cases, this knowledge is also associated with certain impacts (influence, action, or conclusion, on the right hand side). This paper presents an abstract high-level picture of combined mining and the combined patterns from the perspective of object and pattern relation analysis. Several fundamental aspects of combined pattern mining are discussed, including feature interaction, pattern interaction, pattern dynamics, pattern impact, pattern relation, pattern structure, pattern paradigm, pattern formation criteria, and pattern presentation (in terms of pattern ontology and pattern dynamic charts). We also briefly illustrate the concepts and discuss how they can be applied to mining complex data for complex knowledge in either a multifeature, multisource, or multimethod scenario. © 2013 Wiley Periodicals, Inc.

How to cite this article:
WIREs Data Mining Knowl Discov 2013, 3: 140–155 doi: 10.1002/widm.1080

INTRODUCTION

In this paper, we introduce the concept of combined (pattern) mining. Combined mining is mainly suitable for handling the complexity of employing multifeature sets, multi-information sources, constraints, multimethods, and multimodels in data mining, and for analyzing complex relations between objects or descriptors (attributes, sources, methods, constraints, labels, and impacts) or between identified patterns during the learning process. Combined patterns may be formed through analysis of the internal relations between objects or pattern constituents obtained by a single method on a single dataset; for instance, combined sequential patterns formed from analyzing the relations within a discovered sequential pattern space.

With the exception of object and pattern relation analysis, which is a very new topic in the data mining community, many approaches and algorithms are available in the literature on other aspects of the above combinations. The main contribution of combined mining is that it enables the extraction, discovery, construction, and induction of knowledge, which consists of not simply discriminant objects but also of interactions and relations between objects, as well as their impact. They are referred to as *complex but actionable patterns*, because they reflect pattern elements and relations, which form certain pattern structures and dynamics, and indicate decision-making actions.

Combined mining provides an overall solution for meeting the challenge of mining complex knowledge in complex data.¹ It also substantially builds upon other individual approaches such as conceptual inductive learning^{2,3} and inference, generalization, aggregation, and summarization,^{4,5} in order to

*Correspondence to: longbing.cao@uts.edu.au

Advanced Analytics Institute, University of Technology Sydney, Australia

DOI: 10.1002/widm.1080

Longbing Cao. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery

Pattern discovery process

$$\mathcal{P}_{n,m,l} : \mathcal{R}_l(\mathcal{F}_k) \rightarrow \mathcal{I}_{m,l} \quad (1)$$

Data set \mathcal{D} : $\mathcal{D} = \{\mathcal{D}_k; k = 1, \dots, K\}$

Feature set \mathcal{F} : $\mathcal{F} = \{\mathcal{F}_k; k = 1, \dots, K\}$

Method set \mathcal{R} : $\mathcal{R} = \{\mathcal{R}_l; l = 1, \dots, L\}$

Interestingness set \mathcal{I} : $\mathcal{I} = \{\mathcal{I}_{m,l}; m = 1, \dots, M; l = 1, \dots, L\}$

Impact set \mathcal{T} : $\mathcal{T} = \{\mathcal{T}_j; j = 1, \dots, J\}$

Pattern set \mathcal{P} : $\mathcal{P} = \{\mathcal{P}_{n,m,l}; n = 1, \dots, N; m = 1, \dots, M; l = 1, \dots, L\}$

Combined mining

Definition 1 (Combined Mining): Combined mining is a two-to-multistep data mining procedure, consisting of the following:

- 1) Mining atomic patterns $\mathcal{P}_{n,m,l}$ as described in (1).
- 2) Merging atomic pattern sets into combined pattern set $\mathcal{P}'_k = \mathcal{G}_k(\mathcal{P}_{n,m,l})$ for each data set D_k by pattern merging method \mathcal{G}_k ; $\mathcal{G}_k \in \mathcal{G}$, where \mathcal{G} includes a set of pattern-merging methods suitable for a particular business problem.
- 3) If multiple data sets are involved, combined patterns identified in specific data sets are then further merged into the combined pattern set $\mathcal{P} = \mathcal{G}(\mathcal{P}'_k)$.

From a high-level perspective, combined mining represents a generic framework for mining complex patterns in complex data as follows:

$$\mathcal{P} := \mathcal{G}(\mathcal{P}_{n,m,l}) \quad (2)$$

in which atomic patterns $\mathcal{P}_{n,m,l}$ from either individual sources D_k , individual methods \mathcal{R}_l , or particular feature sets \mathcal{F}_k are combined into groups with the members closely related to each other in terms of pattern similarity or difference.

The meaning of “combined”

- 1) The combination of multiple data sources (\mathcal{D}): The combined pattern set \mathcal{P} consists of multiple atomic patterns identified in several data sources, respectively, namely, $\mathcal{P} = \{\mathcal{P}'_k | \mathcal{P}'_k : \mathcal{I}'_k(X_j); X_j \in \mathcal{D}_k\}$; for example, demographic data and transactional data are two data sets involved in mining for demographic–transactional patterns.
- 2) The combination of multiple features (\mathcal{F}): The combined pattern set \mathcal{P} involves multiple features, namely, $\mathcal{P} = \{\mathcal{F}_k | \mathcal{F}_k \subset \mathcal{F}, \mathcal{F}_k \in \mathcal{D}_k, \mathcal{F}_{j+k} \in \mathcal{D}_{j+k}; j, k \neq 0\}$, e.g., features of customer demographics and behavior.
- 3) The combination of multiple methods (\mathcal{R}): The patterns in the combined set reflect the results mined by multiple data mining methods, namely, $\mathcal{P} = \{\mathcal{P}'_k | \mathcal{R}'_k \rightarrow \mathcal{P}'_k\}$, for instance, association mining and classification.
- 4) The combination of pattern impacts.

Basic paradigms

- Nonimpact-oriented combined patterns

$$\mathcal{P}_n : R_l(X_1 \wedge \cdots \wedge X_i) \rightarrow I_m \quad (3)$$

$$\mathcal{P} := \mathcal{G}(P_1 \wedge \cdots \wedge P_n) \rightarrow \mathcal{I} \quad (4)$$

- Impact-oriented combined patterns

$$P_n : \{R_l(X_1 \wedge \cdots \wedge X_i) \rightarrow I_m\} \rightarrow T_1 \quad (5)$$

$$\mathcal{P} := \mathcal{G}(P_1, \dots, P_n) \quad (6)$$

Number of constituent atoms

- Pair patterns

$$\mathcal{P} ::= \mathcal{G}(P_1, P_2)$$

- Cluster patterns

$$\mathcal{P} ::= \mathcal{G}(P_1, \dots, P_n) (n > 2)$$

Structural relations

- Peer-to-peer patterns

$$\mathcal{P} ::= P_1 \cup P_2$$

- Master-slave patterns

$$\{\mathcal{P} ::= P_1 \cup P_2, P_2 = f(P_1)\}$$

- Hierarchy patterns

$$\{\mathcal{P} ::= P_i \cup P'_i \cup P_j \cup P'_j, P_j = \mathcal{G}(P_i), \dots, P'_j = \mathcal{G}'(P_i)^j\}$$

Time frame

- Independent patterns

$$\{P_1 : P_2\}$$

- Sequential patterns

$$\{P_1; P_2\}$$

- Hybrid patterns

$$\{P_1 \otimes P_2 \cdots \otimes P_n; \otimes \in \{\cdot, \parallel, ;\}\}$$

Basic Process: an framework

- Multi-source combined pattern mining

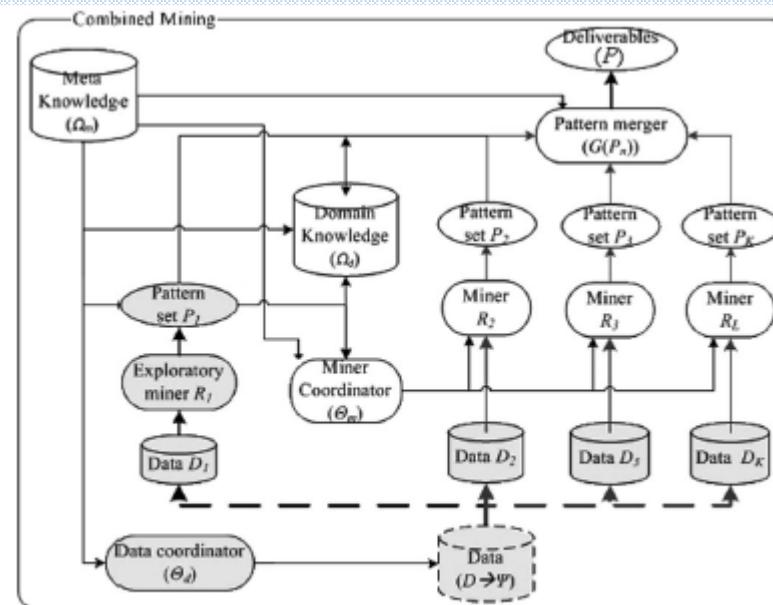


Fig. 1. Combined mining for actionable patterns.

$$CM := \underbrace{D_k [D \xrightarrow{\otimes} D_k]}_K \xrightarrow{\mathcal{I}_k, \mathcal{R}_k, \Omega_m} \{\mathcal{P}_k\} \longrightarrow \xrightarrow{g^N p_k, \Omega_d, \Omega_m} \mathcal{P}$$

PROCESS: Multisource Combined Mining

INPUT: target data sets \mathcal{D}_k ($k = 1, \dots, K$), business problem Ψ

OUTPUT: combined patterns \mathcal{P}

Step 1: Identify a suitable data set or data part, for example, \mathcal{D}_1 for initial mining exploration.

Step 2: Identify the next suitable data set for pattern mining, or partition whole source data into K data sets supervised by the findings in Step 1.

Step 3: *Data set-kmining*: Extract atomic patterns \mathcal{P}_k on data set/subset \mathcal{D}_k .

FOR $k = 1$ to K

 Develop modeling method \mathcal{R}_k with interestingness \mathcal{I}_k .

 Employ method \mathcal{R}_k on the environment e and data \mathcal{D}_k engaging metaknowledge Ω_m .

 Extract the atomic pattern set \mathcal{P}_k .

ENDFOR

Step 4: *Pattern merger*: Merge atomic patterns into combined pattern set \mathcal{P} .

FOR $k = 1$ to K

 Design the pattern merger functions \mathcal{G}_k to merge all relevant atomic patterns into \mathcal{P}_k by involving domain and metaknowledge Ω_d and Ω_m and interestingness \mathcal{I} .

 Employ the method $\mathcal{G}(\mathcal{P}_k)$ on the pattern set \mathcal{P}_k .

 Generate combined patterns into set $\mathcal{P} = \mathcal{G}_k(\mathcal{P}_k)$.

ENDFOR

Step 5: Enhance pattern actionability to generate deliverables \mathcal{P} .

Step 6: Output the deliverables \mathcal{P} .

- Multi-feature combined pattern mining

Definition 2 (MFCPs): Assuming that \mathcal{F}_k denotes the set of features in data set $\mathcal{D}_k \forall i \neq j$, $\mathcal{F}_{k,i} \cap \mathcal{F}_{k,j} = \emptyset$, based on the variables defined in Section IV-A, an MFCP P is in the form of

$$\begin{aligned}\mathcal{P}_k &: \mathcal{R}_l(\mathcal{F}_1, \dots, \mathcal{F}_k) \\ \mathcal{P} &:= \mathcal{G}_F(\mathcal{P}_k)\end{aligned}\tag{8}$$

where $\exists i, j, i \neq j, \mathcal{F}_i \neq \emptyset, \mathcal{F}_j \neq \emptyset$, and \mathcal{G}_F is the merging method for the feature combination.

$$F \wedge c_1 \wedge a_1 - a_2 \rightarrow N$$

- Multi-method combined pattern mining

Definition 10 (Multimethod Combined Mining): Assuming that there are l data mining methods \mathcal{R}_l ($l = 1, \dots, L$), their respective interestingness metrics are in the set \mathcal{I}_m ($m = 1, \dots, M$). The features available for mining the data set are denoted by \mathcal{F} , and *multimethod combined mining* is in the form of

$$\begin{aligned}\mathcal{P}_l : \mathcal{R}_l(\mathcal{F}) &\rightarrow \mathcal{I}_{m,l} \\ \mathcal{P} := \mathcal{G}_M(\mathcal{P}_l)\end{aligned}\tag{20}$$

where \mathcal{G}_M is the merging method integrating the patterns identified by multiple methods.

- Multi-method combined pattern mining
 - Parallel MMCM

$$\left\{ \begin{array}{l} D_1 \xrightarrow{e, \mathcal{I}_1, \mathcal{R}_1, \Omega_m} \mathcal{P}_1 \\ D_2 \xrightarrow{e, \mathcal{I}_2, \mathcal{R}_2, \Omega_m} \mathcal{P}_2 \\ \dots \\ D_K \xrightarrow{e, \mathcal{I}_K, \mathcal{R}_K, \Omega_m} \mathcal{P}_n \end{array} \right. \quad \mathcal{P} := \mathcal{G}(\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_n). \quad (22)$$

- Serial MMCM

$$D \xrightarrow{e, \mathcal{R}_1, \mathcal{F}_1, \mathcal{I}_1, \Omega_m} \mathcal{P}_1, \text{ or} \quad (23)$$

$$\{\mathcal{R}_1, \mathcal{F}_1, \mathcal{I}_1\} \xrightarrow{e, D, \Omega_m} \mathcal{P}_1. \quad (24)$$

$$\{\mathcal{R}_2, \mathcal{F}_2, \mathcal{I}_2\} \xrightarrow{e, D, \Omega_m, \mathcal{P}_1} \mathcal{P}_2. \quad (25)$$

$$\{\mathcal{R}_L, \mathcal{F}_L, \mathcal{I}_L\} \rightarrow \mathcal{P}. \quad (26)$$

Multi-Feature Combined Patterns

DEFINITION MULTI-FEATURE COMBINED PATTERNS. *Assume $\mathcal{F}_{k,i}$ to be the set of all features in dataset \mathcal{D}_k , and $\forall i \neq j$, $\mathcal{F}_{k,i} \cap \mathcal{F}_{k,j} = \emptyset$, based on the variables defined in Section 2.1, a Multi-Feature Combined Pattern (MFCP) P is in the form of*

$$\mathcal{R} : \mathcal{I}(\mathcal{F}_1, \dots, \mathcal{F}_k) \rightarrow T$$

$T \neq \emptyset$ is a target item or class and $\exists i, j, i \neq j, \mathcal{F}_i \neq \emptyset, \mathcal{F}_j \neq \emptyset$.

For example, A_1 can be a demographic itemset, A_2 can be a transactional itemset on marketing campaign, A_3 can be an itemset from a third-party dataset, and T can be the loyalty level of a customer.

Traditional Supports, Confidences & Lifts

- $\text{Supp}(A \rightarrow B) = \text{Prob}(A \wedge B)$
- $\text{Conf}(A \rightarrow B) = \text{Prob}(A \wedge B) / \text{Prob}(A)$
- $\text{Lift} = \text{Conf}(A \rightarrow B) / \text{Prob}(B)$

Table 6: Traditional Interestingness Measures for Rule
 $U + V \rightarrow C$

Supports	$\text{Supp}(U)$, $\text{Supp}(V)$, $\text{Supp}(UV)$, $\text{Supp}(C)$ $\text{Supp}(UC)$, $\text{Supp}(VC)$, $\text{Supp}(UVC)$
Confidences	$\text{Conf}(U \rightarrow C)$, $\text{Conf}(V \rightarrow C)$, $\text{Conf}(U + V \rightarrow C)$
Lifts	$\text{Lift}(U \rightarrow C)$, $\text{Lift}(V \rightarrow C)$, $\text{Lift}(U + V \rightarrow C)$

Contribution

DEFINITION CONTRIBUTION. *For a multi-feature combined pattern $P : X \rightarrow T$, where $X = X_p \wedge X_e$, the contribution of X_e to the occurrence of outcome T in rule P is*

$$\begin{aligned} Cont_e(P) &= \frac{Lift(X_p \wedge X_e \rightarrow T)}{Lift(X_p \rightarrow T)} \\ &= \frac{Conf(X_p \wedge X_e \rightarrow T)}{Conf(X_p \rightarrow T)} \end{aligned}$$

$Cont_e(P)$ is the lift of X_e with X_p as a precondition, which shows how much X_e contributes to the rule. *Contribution* can be taken as the increase of *lift* by appending additional items X_e to a rule. Its value falls in $[0, +\infty]$. A *contribution* greater than one means that the additional items in the rule contribute to the occurrence of the outcome, and a *contribution* less than one suggests that it incurs a reverse effect.

Interestingness of Combined Pattern

$$I_{\text{rule}}(X_p \wedge X_e \rightarrow T) = \frac{\text{Cont}_e(X_p \wedge X_e \rightarrow T)}{\text{Lift}(X_e \rightarrow T)}$$

I_{rule} indicates whether the *contribution* of X_p (or X_e) to the occurrence of T increases with X_e (or X_p) as a precondition. Therefore, “ $I_{\text{rule}} < 1$ ” suggests that $X_p \wedge X_e \rightarrow T$ is less interesting than $X_p \rightarrow T$ and $X_e \rightarrow T$. The value of I_{rule} falls in $[0, +\infty)$. When $I_{\text{rule}} > 1$, the higher I_{rule} is, the more interesting the rule is.

Combined Pattern Pairs

DEFINITION COMBINED PATTERN PAIRS. *For impact-oriented combined patterns, a Combined Pattern Pair (CPP) is in the form of*

$$\mathcal{P}: \left\{ \begin{array}{l} X_1 \rightarrow T_1 \\ X_2 \rightarrow T_2 \end{array} \right. ,$$

where 1) $X_1 \cap X_2 = X_p$ and X_p is called the prefix of pair \mathcal{P} ; $X_{1,e} = X_1 \setminus X_p$ and $X_{2,e} = X_2 \setminus X_p$; 2) X_1 and X_2 are different itemsets; and 3) T_1 and T_2 are contrary to each other, or T_1 and T_2 are same but there is a big difference in the interestingness (say confidences $conf$) of the two patterns.

- A combined rule pair is composed of two contrasting rules.
- Eg., for customers with the same characteristics U , different policies/campaigns, V_1 and V_2 , can result in different outcomes, T_1 and T_2 .

Interestingness of Pattern Pairs

$$I_{\text{pair}}(\mathcal{P}) = \begin{cases} |Conf(P_1) - Conf(P_2)|, & \text{if } T_1 = T_2; \\ \sqrt{Conf(P_1) \cdot Conf(P_2)}, & \text{if } T_1 \text{ and } T_2 \text{ are contrary;} \\ 0, & \text{otherwise;} \end{cases}$$

Combined Pattern Clusters

DEFINITION COMBINED PATTERN CLUSTERS. *Assume there are k local patterns $X_i \rightarrow T_i, (i = 1, \dots, k)$, $k \geq 3$ and $X_1 \cap X_2 \cap \dots \cap X_k = X_p$, a combined pattern cluster (CPC) is in the form of*

$$\mathcal{C}: \left\{ \begin{array}{l} X_1 \rightarrow T_1 \\ \dots \\ X_k \rightarrow T_k \end{array} \right. ,$$

where X_p is the prefix of cluster \mathcal{C} .

- Based on a combined rule pair, related combined rules can be organized into a cluster to supplement more information to the rule pair.
- The rules in cluster \mathcal{C} have the same U but different V , which makes them associated with various results T .

Interestingness of Pattern Clusters

$$I_{\text{cluster}}(\mathcal{C}) = \max_{P_i, P_j \in \mathcal{C}, i \neq j} I_{\text{pair}}(P_i, P_j)$$

Interestingness of Rule Pair/Cluster

$$I_{\text{pair}}(\mathcal{P}) = \text{Lift}_V(R_1) \text{ Lift}_V(R_2) \text{ dist}(T_1, T_2)$$

$$I_{\text{cluster}}(\mathcal{C}) = \max_{i \neq j, R_i, R_j \in \mathcal{C}, T_i \neq T_j} I_{\text{pair}}(R_i, R_j)$$

- `dist()`: the dissimilarity between the descendants of R_1 and R_2
- The interestingness of combined rule pair/cluster is decided by both the interestingness of rules and the most contrasting rules within the pair/cluster.
- A cluster made of contrasting confident rules is interesting, because it explains why different results occur and what can be done to produce an expected result or avoid an undesirable consequence.

Rule Pair vs Rule Cluster

$$\mathcal{P} : \begin{cases} U \wedge V_1 \rightarrow stay \\ U \wedge V_2 \rightarrow churn \end{cases}, \quad \mathcal{C} : \begin{cases} U \wedge V_1 \rightarrow stay \\ U \wedge V_2 \rightarrow churn \\ U \wedge V_3 \rightarrow stay \end{cases}.$$

- From P, we can see that V_1 is a preferable policy for customers with characteristics U.
- If, for some reason, policy V_1 is inapplicable to the specific customer group, P is no longer actionable.
- Rule cluster C suggests that another policy V_3 can be employed to retain those customers.

Extended Combined Pattern Pairs

DEFINITION EXTENDED COMBINED PATTERN PAIRS. *An Extended Combined Pattern Pair (ECPP) is a special combined pattern pair as follows*

$$\mathcal{E}: \left\{ \begin{array}{l} X_p \rightarrow T_1 \\ X_p \wedge X_e \rightarrow T_2 \end{array} \right. ,$$

where $X_p \neq \emptyset$, $X_e \neq \emptyset$ and $X_p \cap X_e = \emptyset$.

Conditional P-S ratio

DEFINITION *A metric for measuring the difference led by the occurrence of X_e in the above scenario is Conditional Piatetsky-Shapiro's (P-S) ratio Cps , which is defined as follows.*

$$Cps(X_e \rightarrow T | X_p) = Prob(X_e \rightarrow T | X_p) - Prob(X_e | X_p) \times Prob(T | X_p)$$

$$= \frac{Prob(X_p \wedge X_e \rightarrow T)}{Prob(X_p)} - \frac{Prob(X_p \wedge X_e)}{Prob(X_p)} \times \frac{Prob(X_p \rightarrow T)}{Prob(X_p)}$$

Extended Combined Pattern Clusters

DEFINITION EXTENDED COMBINED PATTERN SEQUENCES. *An Extended Combined Pattern Sequence (ECPC), or called Incremental Combined Pattern Sequence (ICPS), is a special combined pattern cluster with additional items appending to the adjacent local patterns incrementally.*

$$\mathcal{S}: \left\{ \begin{array}{l} X_p \rightarrow T_1 \\ X_p \wedge X_{e,1} \rightarrow T_2 \\ X_p \wedge X_{e,1} \wedge X_{e,2} \rightarrow T_3 \\ \dots \\ X_p \wedge X_{e,1} \wedge X_{e,2} \wedge \dots \wedge X_{e,k-1} \rightarrow T_k \end{array} \right.,$$

where $\forall i, 1 \leq i \leq k - 1, X_{i+1} \cap X_i = X_i$ and $X_{i+1} \setminus X_i = X_{e,i} \neq \emptyset$, i.e., X_{i+1} is an increment of X_i . The above cluster of rules actually makes a sequence of rules, which can show the impact of the increment of patterns on the outcomes.

Impact

DEFINITION IMPACT. *The impact of X_e on the outcome in the rule is*

$$impact_e(P) = \begin{cases} cont_e(P) - 1 & : \text{if } cont_e(P) \geq 1, \\ \frac{1}{cont_e(P)} - 1 & : \text{otherwise.} \end{cases}$$

Intervention Strategy 1

- Type A: Demographics differentiated combined pattern
 - Customers with the same actions but different demographics
→ different classes/business impact

$$\text{Type A: } \left\{ \begin{array}{l} A_1 + D_1 \rightarrow \text{quick payer} \\ A_1 + D_2 \rightarrow \text{moderate payer} \\ A_1 + D_3 \rightarrow \text{slow payer} \end{array} \right.$$

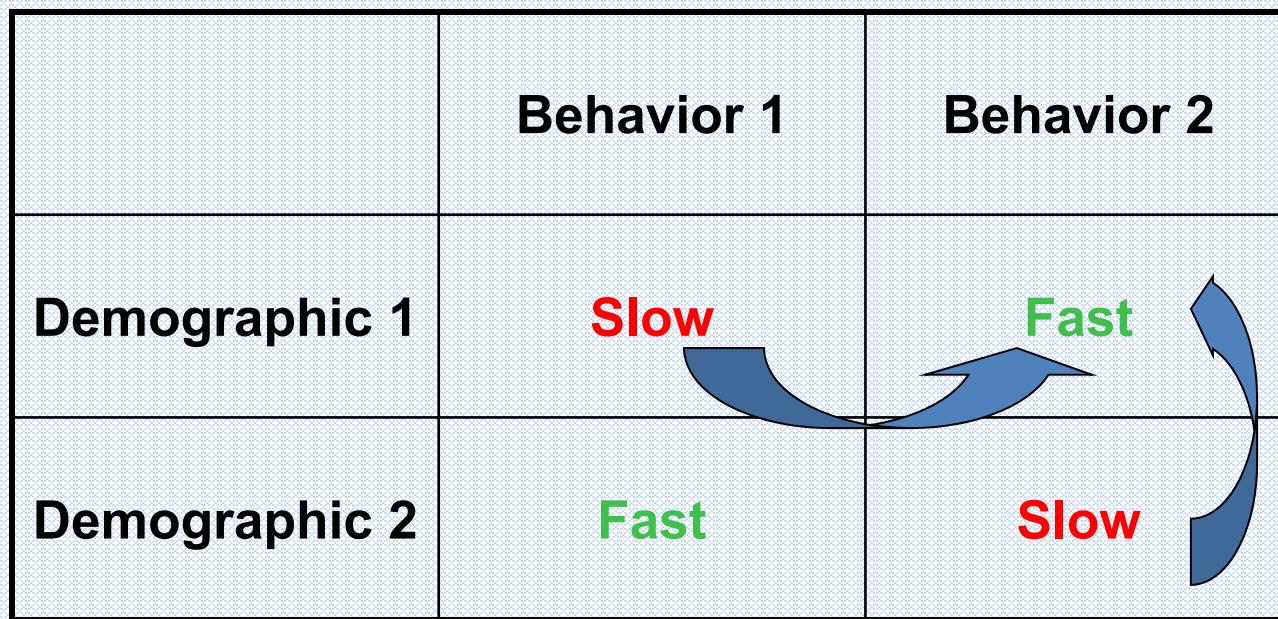
Intervention Strategy 2

- Type B: Action differentiated combined pattern
 - Customers with the same demographics but taking different actions
→ different classes/business impact

Type B: $\begin{cases} A_1 + D_1 & \rightarrow \text{quick payer} \\ A_2 + D_1 & \rightarrow \text{moderate payer} \\ A_3 + D_1 & \rightarrow \text{slow payer} \end{cases}$

Business Impact

- Able to move customers from one class to another class
- Useful for designing business policy



Business Problem

Case Study I: Debt Recovery

- To profile customers according to their capacity to pay off their debts in shortened timeframes.
- To target those customers with recovery and amount options suitable to their own circumstances, and increase the frequency and level of repayment.

Data (1)

- Customer demographic data
 - Customer ID, gender, age, marital status, number of children, declared wages, location, benefit type, ...
- Debt data
 - Debt amount, debt start/end date, ...
- Repayment data (transactional)
 - Repayment method, amount, time, date, ...
- Class ID: Quick/Moderate/Slow Payer

Data (2)

- The case study is on governmental social security data with debts raised in the calendar year 2006 and the corresponding customers and arrangement/repayment activities.
- The cleaned sample data contains 355,800 customers with their demographic attributes, arrangements and repayments.
- There are 7,711 traditional associations mined.

Results (1)

- There were 7,711 association rules before removing redundancy of combined rules.
- After removing redundancy of combined rules, 2,601 rules were left, which built up 734 combined rule clusters.
- After removing redundancy of combined rule clusters, 98 rule clusters with 235 rules remained, which was within the capability of human beings to read.

Results (2)

Traditional Association Rules

V			T	Conf(%)	Count	Lift
Arrangement	Repayment		Class			
irregular	cash or post office		A	82.4	4088	1.8
withholding	cash or post office		A	87.6	13354	1.9
withholding & irregular	cash or post office		A	72.4	894	1.6
withholding & irregular	cash or post office & withholding		B	60.4	1422	1.7

An Example of Combined Patterns

Rules	X_p		X_e		T	Cnt	Conf (%)	I_r	Lift	$Cont_p$	$Cont_e$	Lift of $X_p \rightarrow T$	Lift of $X_e \rightarrow T$
	Demographics	Arrangements	Repayments	Class									
P_1	age:65+	withholding & irregular	withholding	C	50	63.3	2.91	3.40	2.47	4.01	0.85	0.85	1.38
P_2	income:0 & remote:Y & marital:sep & gender:F	withholding	cash or post & withholding	B	20	69.0	1.47	1.95	1.34	2.15	0.91	0.91	1.46
P_3	income:0 & age:65+	withholding	cash or post & withholding	A	1123	62.3	1.38	1.35	1.72	1.09	1.24	1.24	0.79
P_4	income:0 & gender:F & benefit:P	withholding	cash or post	A	469	93.8	1.36	2.04	1.07	2.59	0.79	0.79	1.90

Results (3)

An Example of Combined Pattern Clusters

Clusters	Rules	X_p	X_e		T	Cnt	$Conf$ (%)	I_r	I_c	$Lift$	$Cont_p$	$Cont_e$	$Lift$ of $X_p \rightarrow T$	$Lift$ of $X_e \rightarrow T$
		demographics	arrangements	repayments										
\mathcal{P}_1	P_5	marital:sin &gender:F &benefit:N	irregular	cash or post	A	400	83.0	1.12	0.67	1.80	1.01	2.00	0.90	1.79
	P_6		withhold	cash or post	A	520	78.4	1.00		1.70	0.89	1.89	0.90	1.90
	P_7		withhold & irregular	cash or post & withhold	B	119	80.4	1.21		2.28	1.33	2.06	1.10	1.71
	P_8		withhold	cash or post & withhold	B	643	61.2	1.07		1.73	1.19	1.57	1.10	1.46
	P_9		withhold & vol. deduct	withhold & direct debit	B	237	60.6	0.97		1.72	1.07	1.55	1.10	1.60
	P_{10}		cash	agent	C	33	60.0	1.12		3.23	1.18	3.07	1.05	2.74
\mathcal{P}_2	P_{11}	age:65+	withhold	cash or post	A	1980	93.3	0.86	0.59	2.02	1.06	1.63	1.24	1.90
	P_{12}		irregular	cash or post	A	462	88.7	0.87		1.92	1.08	1.55	1.24	1.79
	P_{13}		withhold & irregular	cash or post	A	152	85.7	0.96		1.86	1.18	1.50	1.24	1.57
	P_{14}		withhold & irregular	withhold	C	50	63.3	2.91		3.40	2.47	4.01	0.85	1.38

Business Rule

BUSINESS RULES: Customer Demographic-Arrangement-Repayment combination business rules

For All *customer i* ($i \in I$ is the number of valid customers)

Condition:

satisfies *S/he is a debtor aged 65 or plus*;

relates

S/he is under arrangement of ‘withholding’ and ‘irregularly’,

and

His/her favorite Repayment method is ‘withholding’;

Operation:

Alert = “*S/he has ‘High’ risk of paying off debt in a very long timeframe.*”

Action = “*Try other arrangements and repayments in R_2 , such as trying to persuade her/him to repay under ‘irregular’ arrangement with ‘cash or post’.*”

End-All

Business Problem

Case Study II: Debt Prevention

- A case study of extend combined pattern pairs on Centrelink debt-related activity data is given as follows. More details can be found in [Cao et al. 2008], where they are called impact-reversed sequential activity patterns.
- The data involves four data sources, which are activity files recording activity details, debt files logging debt details, customer files enclosing customer circumstances, and earnings files storing earnings details.
- To analyse the relationship between activity and debt, the data from activity files and debt files are extracted.

Data (1)

- Customer demographic data
 - Customer ID, gender, age, marital status, number of children, declared wages, location, benefit type, ...
- Debt data
 - Debt amount, debt start/end date, ...
- Repayment data (transactional)
 - Repayment method, amount, time, date, ...
- Class ID: Quick/Moderate/Slow Payer

Date (2)

- The activity data for us to test the proposed approaches is Centrelink activity data from Jan. 1st to Mar. 31st 2006.
- We extract activity data including 15,932,832 activity records recording government-customer contacts with 495,891 customers, which lead to 30,546 debts in the first three months of 2006.
- After data preprocessing and transformation, there are 454,934 sequences: 16,540 (3.6%) activity sequences associated with debts and 438,394 (96.4%) sequences with nil debt.

Results (1)

Examples of Extended Combined Pattern Pairs

X_p	T_1	X_e	T_2	$Cont_e$	Cps	Local support of $X_p \rightarrow T_1$	Local support of $X_p \wedge X_e \rightarrow T_2$
a_{14}	\bar{T}	a_4	T	2.5	0.013	0.684	0.428
a_{16}	\bar{T}	a_4	T	2.2	0.005	0.597	0.147
a_{14}	\bar{T}	a_5	T	2.0	0.007	0.684	0.292
a_{16}	\bar{T}	a_7	T	1.8	0.004	0.597	0.156
a_{14}	\bar{T}	a_7	T	1.7	0.005	0.684	0.243
a_{15}	\bar{T}	a_5	T	1.7	0.007	0.567	0.262
a_{14}, a_{14}	\bar{T}	a_4	T	2.3	0.016	0.474	0.367
a_{14}, a_{16}	\bar{T}	a_5	T	2.0	0.005	0.393	0.118
a_{15}, a_{14}	\bar{T}	a_5	T	1.7	0.007	0.381	0.179
a_{14}, a_{16}, a_{14}	\bar{T}	a_{15}	T	1.2	0.005	0.248	0.188

An Example of Extended Combined Pattern Pair

$$\begin{cases} a_{14} \rightarrow \bar{T} \\ a_{14}, a_4 \rightarrow T \end{cases}$$

- The local supports of $a_{14} \rightarrow T$ and $a_{14} \rightarrow \bar{T}$ respectively 0.903 and 0.684, so the ratio of the two values is 1.3.
- The local supports of $a_{14}, a_4 \rightarrow T$ and $a_{14}, a_4 \rightarrow \bar{T}$ are 0.428 and 0.119 respectively, so the ratio of the two values is 3.6.
- When a14 occurs first, the appearance of a4 makes it more likely to become debtable.
- This kind of pattern pairs help to know what effect an additional activity will have on the impact of the patterns.

Case Study III

- Exploring the impact of behavior dynamics
- Identifying the most important behavior during the evolution

Combined pattern presentation

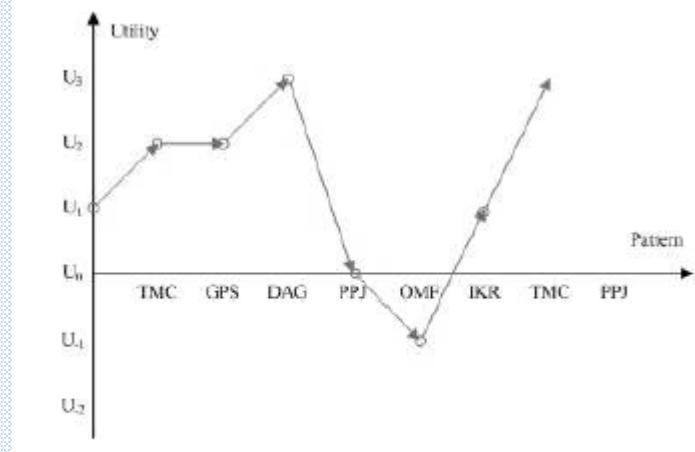


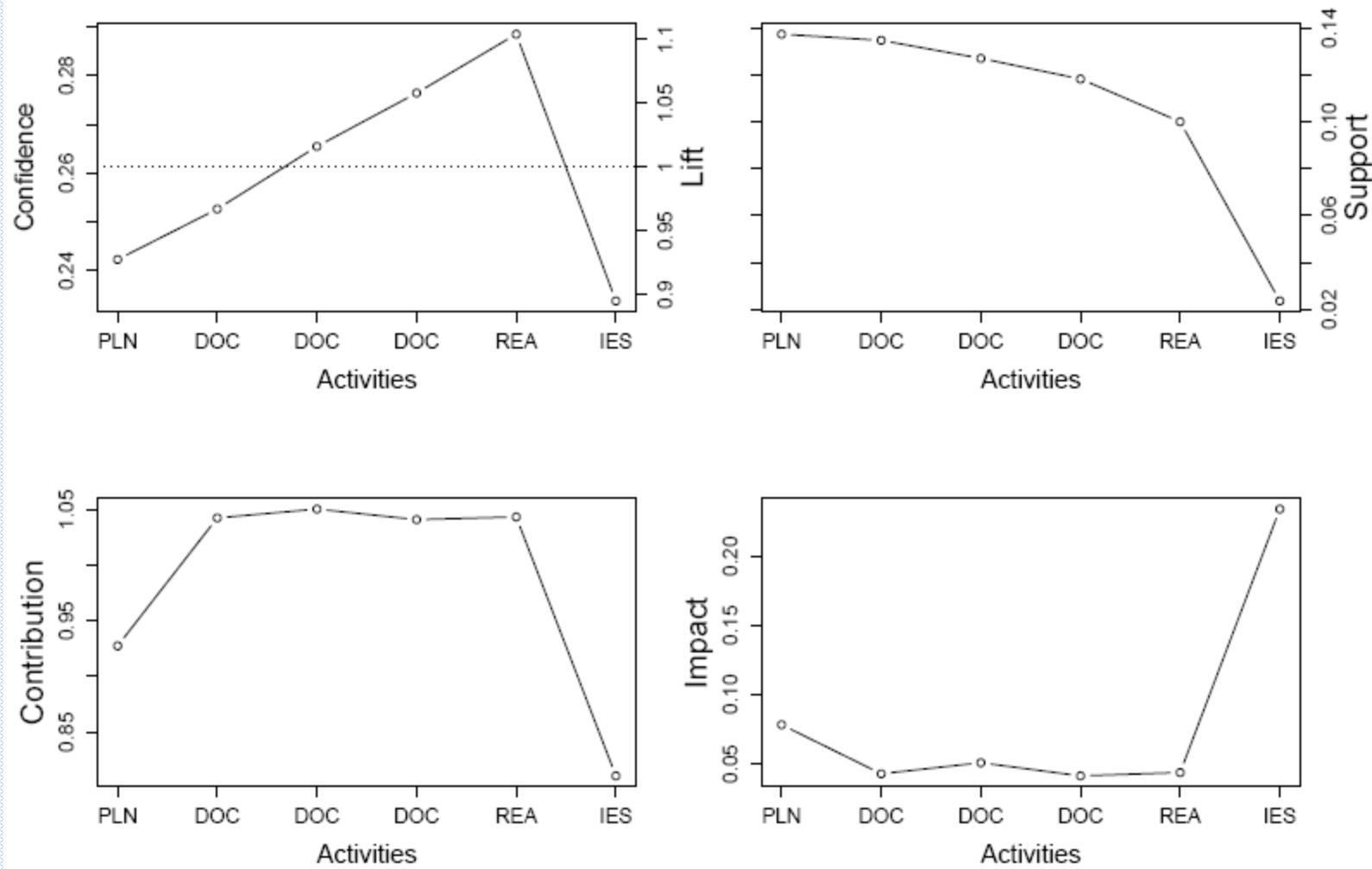
Figure 2: Pattern Evolution Chart

$$\left\{ \begin{array}{l} TMC \rightarrow U_1 \\ TMC, GPS \rightarrow U_2 \\ TMC, GPS, DAG \rightarrow U_2 \\ TMC, GPS, DAG, PPJ \rightarrow U_3 \\ TMC, GPS, DAG, PPJ, OMF \rightarrow U_0 \\ TMC, GPS, DAG, PPJ, OMF, IKR \rightarrow U_{-1} \\ TMC, GPS, DAG, PPJ, OMF, IKR, TMC \rightarrow U_1 \\ TMC, GPS, DAG, PPJ, OMF, IKR, TMC, PPJ \rightarrow U_3 \end{array} \right. , \quad (6)$$

An Example of Extended Combined Pattern Cluster

$$\left\{ \begin{array}{l} PLN \rightarrow T \\ PLN, DOC \rightarrow T \\ PLN, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC, REA \rightarrow T \\ PLN, DOC, DOC, DOC, REA, IES \rightarrow T \end{array} \right.$$

An Example of Extended Combined Pattern Cluster



Discussion 1: Behaviour in your organisation

- 1 List the business lines (drill down to specific business areas) in your organization where behaviour could be an important aspect/asset

- 2 Use a few keywords in a dot point format to describe behaviour analytics tasks conducted at your organization

Discussion 2: What is the behaviour in your organization

- 1 Write a few keywords (dimensions and aspects), or a diagram, to explain what is behaviour on your mind

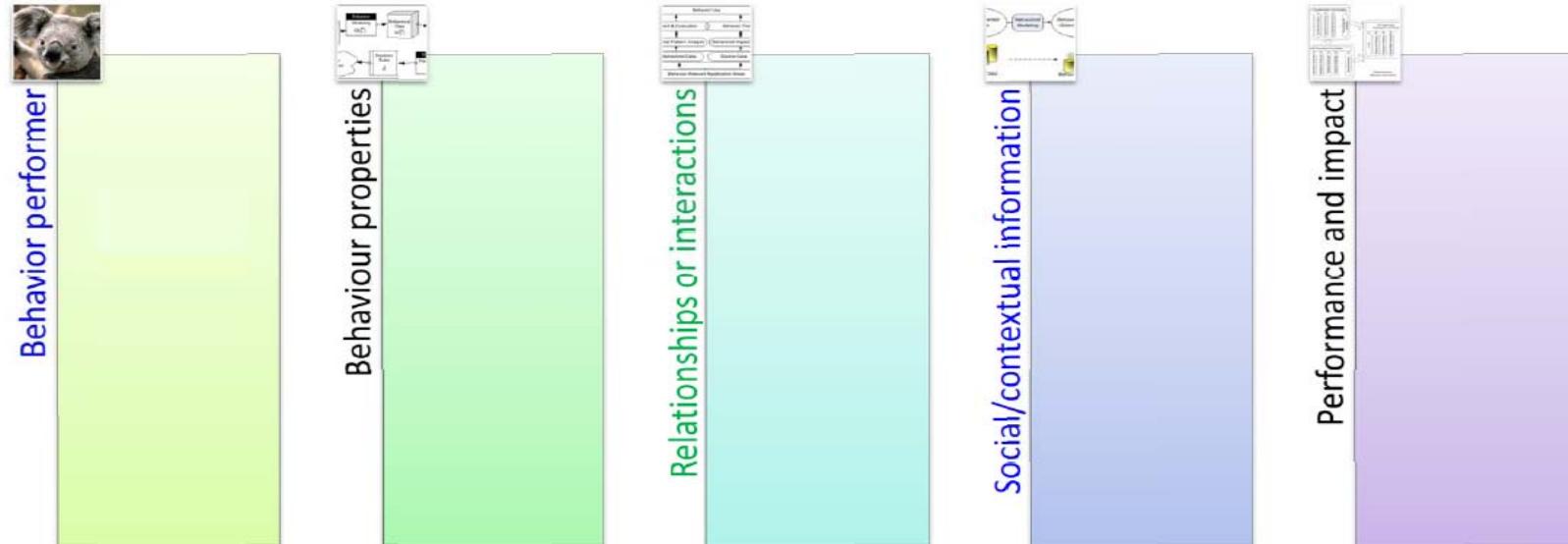
- 2 List three aspects that you believe are the most important in discussing behaviour

Exercise 3: Model behaviour in your business

Behaviour understanding

Organization: _____

Business problem: _____



Group discussion: behaviour impact

Session 3: In your business, how do you measure the impact or utility of behaviour?

	Objective metrics	Subjective metrics
Business aspects		
Technical aspects		

Part III.

Negative Behavior Analysis

Learning Objectives

- What is negative behavior?
- Why care about negative behavior?
- How to represent/model behavior?
- How to check the behavior model?

Negative Behavior Analysis

Negative sequential pattern mining



EXPERT OPINION

Editor: Daniel Zeng, University of Arizona and Chinese Academy of Sciences, zengdaniel@gmail.com

Nonoccurring Behavior Analytics: A New Area

Longbing Cao, University of Technology Sydney
Philip S. Yu, University of Illinois at Chicago
Vipin Kumar, University of Minnesota

Behavior-related studies and applications, such as behavior analysis, data mining, machine learning, and behavioral science, have generally focused on behaviors that have occurred or will occur. Such behaviors are called *positive behaviors* (PBs) or *occurring behaviors* (OBs). Related work has focused on behavioral patterns, anomalies, impact, and dynamics. This constitutes the area of behavior analytics, which focuses on understanding, analyzing, learning, predicting, and managing past, present, and future behaviors. When behavior representation and modeling are also considered, we use the term *behavior informatics* or *behavior computing*¹ to describe the new perspective of modeling, reasoning about, verifying, analyzing, learning, and evaluating behaviors. This has emerged as an important and demanding area for comprehensively and deeply handling ubiquitous behaviors online, in business, government services, scientific activities, social activities, and economic and financial business.

Limited research has been conducted on analyzing, detecting, or predicting nonoccurring behaviors (NOBs), those that did not or will not occur. NOBs are also called *negative behaviors*, which are not straightforward, since they usually are hidden and difficult to understand, or one usually is not concerned with them. That NOBs are overlooked does not mean they are unimportant. For instance, if a patient misses an appointment with a specialist, and thus misses the opportunity to receive immediate and appropriate treatment for a health problem, the patient's health could worsen. Additionally, in many situations, failure to follow rules or policies could result in administrative or even legal obligations.

Therefore, it is important to build a theoretical foundation for NOB study.

Unfortunately, few research outcomes of NOB study can be identified in the literature. Relevant work includes event analysis; negative association rule mining,² which identifies patterns comprising nonoccurring items; and negative sequential patterns,³⁻⁸ which comprise sequential elements that do not appear in the business process. No systematic work has been conducted to understand, model, formalize, analyze, learn, detect, predict, intervene, and manage NOBs.

NOB is not a trivial problem. Some may argue that it is simple to treat an NOB as a special OB, and that all relevant techniques can then be used directly for NOB analytics. Unfortunately, this often does not work for reasons related to the different natures and complexities of occurring and nonoccurring behaviors. In this article, we outline the concept of NOBs and related complexities, draw a picture of NOB analytics, and present our view of NOB research directions and prospects.

What Is NOB?

We briefly discuss the essence, intrinsic characteristics, and complexities of NOBs, and the forms that NOBs can take, in order to understand the concept of NOB.

Intrinsic Characteristics

NOBs refer to those behaviors that should occur but do not for some reason. They are hidden but are widely seen in behavioral applications in business, economics, health, cyberspace, social and mobile networks, and natural and human systems. Many businesses, services, applications, and systems involve NOBs, including healthcare and

Nonoccurring behavior analysis

Longbing Cao, Philip S. Yu, Vipin Kumar. **Nonoccurring Behavior Analytics: A New Area.** IEEE Intelligent Systems 30(6): 4-11 (2015).

Xiangjun Dong, Zhigang Zhao, Longbing Cao, Yanchang Zhao, Chengqi Zhang, Jinjiu Li, Wei Wei, Yuming Ou. **e-NSP: Efficient Negative Sequential Pattern Mining Based on Identified Positive Patterns Without Database Rescanning**, CIKM 2011, 825-830.

Zhigang Zheng, Yanchang Zhao, Ziye Zu, Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Chengqi Zhang. **An Efficient GA-Based Algorithm for Mining Negative Sequential Patterns**, PAKDD2010, 262-273.

Yanchang Zhao, Huaifeng Zhang, Longbing Cao, Chengqi Zhang and Hans Bohlscheid. **Mining Both Positive and Negative Impact-Oriented Sequential Rules From Transactional Data**, PAKDD2009, pp.656-663.

Yanchang Zhao, Huaifeng Zhang, Shanshan Wu, Jian Pei, Longbing Cao, Chengqi Zhang and Hans Bohlscheid. **Debt Detection in Social Security by Sequence Classification Using Both Positive and Negative Patterns**, ECML/PKDD2009, 648-663, 2009.



e-NSP: Efficient Negative Sequential Pattern Mining[☆]

✉ Longbing Cao, ^{1,2} Xiangjun Dong and ^{1,2} Zhigang Zhao

¹ University of Technology Sydney, Australia

² QMUL University of Technology, China

³ University of Technology Sydney, Australia

Abstract

As an important tool for behavior informatics, negative sequential patterns (NSP) (such as missing medical treatments) are critical and sometimes much more informative than positive sequential patterns (PSP) (e.g. using a medical service) in many intelligent systems and applications such as intelligent transport systems, healthcare and risk management, as they often involve non-occurring but interesting behaviors. However, discovering NSP is much more difficult than identifying PSP due to the significant problem complexity caused by non-occurring elements, high computational cost and huge search space in calculating negative sequential candidates (NSC). So far, the problem has not been formulated well, and very few approaches have been proposed to mine for specific types of NSP, which rely on database re-scans after identifying PSP in order to calculate the NSC support. This has been shown to be very inefficient or even impractical, since the NSC search space is usually huge. This paper proposes a very innovative and efficient theoretical framework, set theory-based NSP mining (ST-NSP), and a corresponding algorithm, e-NSP, to efficiently identify NSP by involving only the identified PSP, without re-scanning the database. Accordingly, negative containment is first defined to determine whether a data sequence contains a negative sequence based on set theory. Second, an efficient approach is proposed to convert the negative containment problem in a positive containment problem. The NSC supports are then calculated based only on the corresponding PSP. This not only avoids the need for additional database scans, but also enables the use of existing PSP mining algorithms to mine for NSP. Finally, a simple but efficient strategy is proposed to generate NSC. Theoretical analyses show that e-NSP performs particularly well on datasets with a small number of elements in a sequence, a large number of itemsets and low minimum support. e-NSP is compared with two currently available NSP mining algorithms via intensive experiments on three synthetic and six real-life datasets from aspects including data characteristics, computational costs and scalability. e-NSP is tens to thousands of times faster than baseline approaches, and offers a sound and effective approach for efficient mining of NSP in large scale datasets by directly using existing PSP mining algorithms.

© 2015 Published by Elsevier Ltd.

Keywords:

Negative sequence analysis, sequence analysis, behavior analytics, non-occurring behavior, behavior informatics, behavior computing, pattern mining

1. Introduction

Behavior is widely seen in our daily study, work, living and entertainment [7]. A critical issue in understanding behavior from the informatics perspective, namely behavior informatics [6, 9], is to understand the complexities, dynamics and impact of non-occurring behaviors (NOB) [8]. Mining Negative sequential patterns (NSP) [43] is

[☆]The source codes of e-NSP are available from <http://www.staff.it.usyd.edu.au/~lbcnay/>.

[✉]Corresponding author. Email: longbing.cao@gmail.com

Longbing Cao, Xiangjun Dong,
Zhigang Zhao. **e-NSP: Efficient
Negative Sequential Pattern Mining.**
Artificial Intelligence, 2016.

Issues Addressed

- What is non-occurring behavior?
- Why do we care about non-occurring behaviors?
- What are issues in understanding non-occurring behaviors?
- What are the problems with existing behavior study in addressing non-occurring behaviors?
- Research opportunities and prospects of non-occurring behavior study

Problem description

- What is negative sequential patterns?
- *Focus on negative relationship between itemsets*
- *Absent items are taken into consideration*
- Example:
 $p_1 = \langle a \ b \ c \ d \rangle$ vs $p_2 = \langle a \ b \ \neg c \ e \rangle$
- *Each item, a, b, c, d and e, stands for a claim item of insurance.*
- *p1: an insurant usually claims for a, b, c and d in a claim.*
- *p2: does NOT claim c after a and b, then claim item e instead of d.*

Challenges for NSP

- *Apriori principle doesn't work for some situations*
- *Huge search space*
 - 10 distinct items
 - 3-item PSC: 10^3
 - 3-item NSC: 20^3

Non-occurrence behaviour analysis

(Negative sequence analysis)

Table 1. Supports, Confidences and Lifts of Four Types of Sequential Rules

	Rules	Support	Confidence	Lift
I	$A \rightarrow B$	$P(AB)$	$\frac{P(AB)}{P(A)}$	$\frac{P(AB)}{P(A)P(B)}$
II	$A \rightarrow \neg B$	$P(A) - P(AB)$	$\frac{P(A) - P(AB)}{P(A)}$	$\frac{P(A) - P(AB)}{P(A)(1 - P(B))}$
III	$\neg A \rightarrow B$	$P(B) - P(A \& B)$	$\frac{P(B) - P(A \& B)}{1 - P(A)}$	$\frac{P(B) - P(A \& B)}{P(B)(1 - P(A))}$
IV	$\neg A \rightarrow \neg B$	$1 - P(A) - P(B) + P(A \& B)$	$\frac{1 - P(A) - P(B) + P(A \& B)}{1 - P(A)}$	$\frac{1 - P(A) - P(B) + P(A \& B)}{(1 - P(A))(1 - P(B))}$

Table 4. Selected Positive and Negative Sequential Rules

Type	Rule	Support	Confidence	Lift
I	REA ADV ADV → DEB	0.103	0.53	2.02
	DOC DOC REA REA ANO → DEB	0.101	0.33	1.28
	RPR ANO → DEB	0.111	0.33	1.25
	RPR STM STM RPR → DEB	0.137	0.32	1.22
	MCV → DEB	0.104	0.31	1.19
	ANO → DEB	0.139	0.31	1.19
II	STM PYI → DEB	0.106	0.30	1.16
	STM PYR RPR REA RPT → ¬DEB	0.166	0.86	1.16
	MND → ¬DEB	0.116	0.85	1.15
	STM PYR RPR DOC RPT → ¬DEB	0.120	0.84	1.14
	STM PYR RPR REA PLN → ¬DEB	0.132	0.84	1.14
	REA PYR RPR RPT → ¬DEB	0.176	0.84	1.14
	REA DOC REA CPI → ¬DEB	0.083	0.83	1.12
	REA CRT DLY → ¬DEB	0.091	0.83	1.12
III	REA CPI → ¬DEB	0.109	0.83	1.12
	¬{PYR RPR REA STM} → DEB	0.169	0.33	1.26
	¬{PYR CCO} → DEB	0.165	0.32	1.24
	¬{STM RPR REA RPT} → DEB	0.184	0.29	1.13
	¬{RPT RPR REA RPT} → DEB	0.213	0.29	1.12
	¬{CCO RPT} → DEB	0.171	0.29	1.11
	¬{CCO PLN} → DEB	0.187	0.28	1.09
IV	¬{PLN RPT} → DEB	0.212	0.28	1.08
	¬{ADV REA ADV} → ¬DEB	0.648	0.80	1.08
	¬{STM EAN} → ¬DEB	0.651	0.79	1.07
	¬{REA EAN} → ¬DEB	0.650	0.79	1.07
	¬{DOC FRV} → ¬DEB	0.677	0.78	1.06
	¬{DOC DOC STM EAN} → ¬DEB	0.673	0.78	1.06
	¬{CCO EAN} → ¬DEB	0.681	0.78	1.05

Genetic-Algorithm based NSP approach: GA-NSP

- Find good (frequent) genes with good performance (supp), and optimize genes (FP) through crossover and mutation, m^* generations
- Improve gene quality (making more and more frequent)

Strengths:

- Treat candidates unequally
- Very low support threshold
- Find long-NSP at the beginning

GA-NSP

- *New generations*: good genes (freq patterns) through crossover and mutation operations.
- *Population evolution control*: fitness and dynamic fitness.
- *Performance improvement*: pruning method (check constraints of NSP)

Problem Statement

- Sequence (general)

$s = \langle e_1 \ e_2 \dots \ e_n \rangle$

i.e. $\langle a \ b \ (c,d) \ e \rangle, \langle a \ \neg b \ c \ e \rangle$

- Positive/Negative Sequence

$s_p = \langle e_1 \ e_2 \dots \ e_n \rangle$, all elements are positive

$s_n = \langle e_1 \ e_2 \dots \ e_n \rangle$, at least one element is negative

- Negative Sequential Pattern

- Its support is greater than minimum support threshold.
- Two or more continuous negative elements are not accepted.
- For each negative item, its corresponding positive item is required to be frequent.
- Items in an element should be all positive or all negative. i.e. $\langle a \ (a, \neg b) \ c \rangle$ is not allowed.

• Negative Matching

Negative Matching. A negative sequence $s_n = \langle e_1 \ e_2 \dots e_k \rangle$ matches a data sequence $s = \langle d_1 \ d_2 \dots d_m \rangle$, iff:

- 1) s contains the max positive subsequence of s_n
- 2) for each negative element $e_i (1 \leq i \leq k)$, there exist integers $p, q, r (1 \leq p \leq q \leq r \leq m)$ such that: $\exists e_{i-1} \subseteq d_p \wedge e_{i+1} \subseteq d_r$, and for $\forall d_q, e_i \not\subseteq d_q$

	Sequence	Matching	Data Sequence
S_1	$\langle b \ \neg c \ a \rangle$	No	$\langle b \ f \ d \ c \ a \rangle$
S_2	$\langle b \ \neg c \ d \ a \rangle$	Yes	$\langle b \ f \ d \ c \ a \rangle$

GA-NSP Algorithm

- Encoding

Sequence		Chromosome		
		gene ₁	gene ₂	gene ₃
$\langle a \ b \ \neg(c,d) \rangle$	\Rightarrow	+a	+b	$\neg(c,d)$

- Crossover

parent1	$b \ \neg c \uparrow a$	\Rightarrow	child1	$b \ \neg c \textcolor{red}{e}$
parent2	$d \downarrow e$	\Rightarrow	child2	$d \ a$

parent1	$b \ \neg c \ a \uparrow$	\Rightarrow	child1	$b \ \neg c \ a \textcolor{red}{d} \ e$
parent2	$\uparrow d \ e$	\Rightarrow	child2	$d \ e \ b \ \neg c \ a$

- Mutation

Select a random position and then replace all genes after that position with 1-item patterns

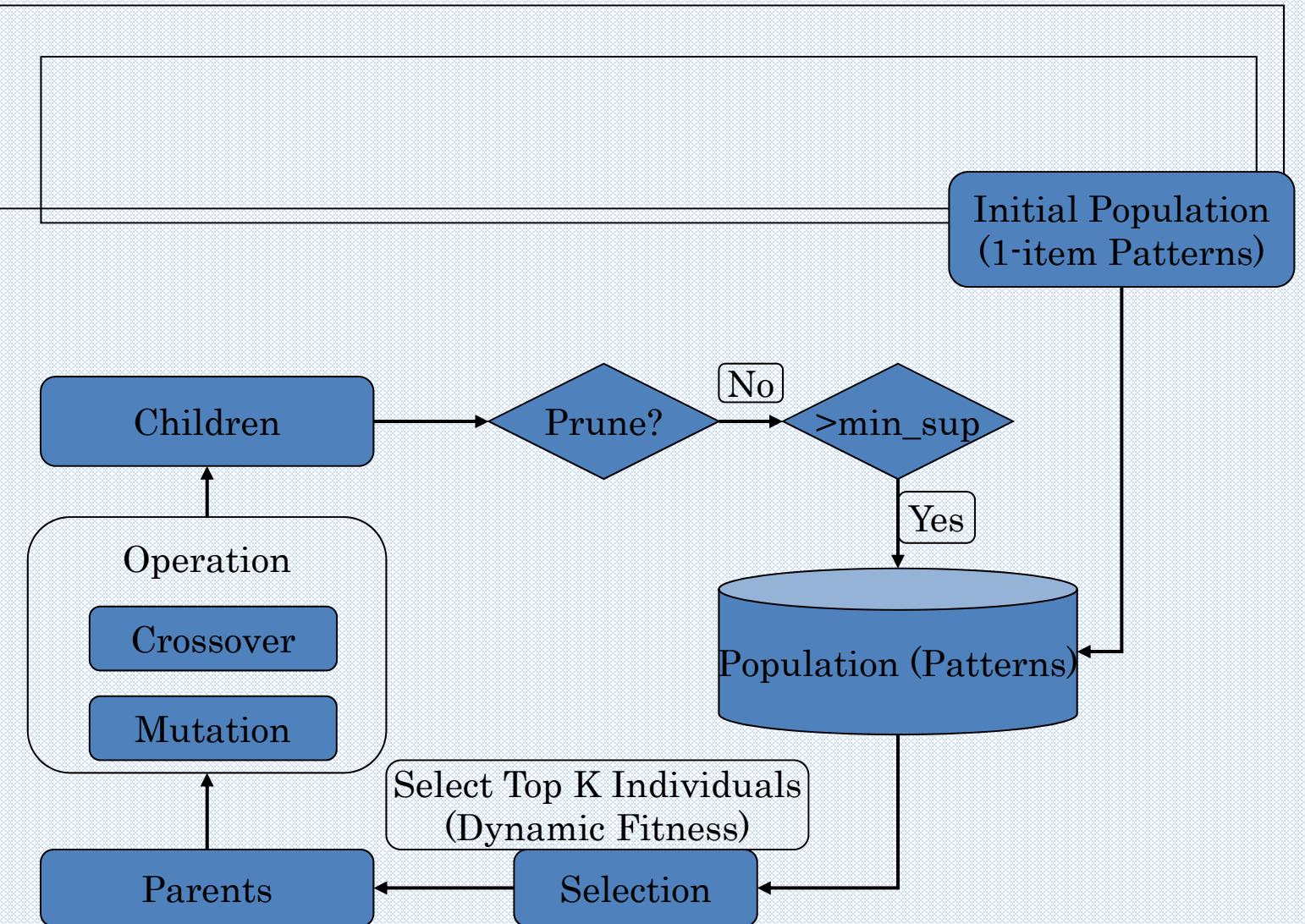
■ Fitness & Dynamic Fitness

$$ind.fitness = (ind.support - min_sup) \times DatasetSize. \quad (1)$$

$$ind.dfitness = \begin{cases} ind.fitness, & \text{initial set} \\ ind.dfitness \times (1 - \underline{DecayRate}), & \text{if } ind \text{ is selected} \end{cases} \quad (2)$$

■ Selection

```
Selection(pop){ //Subfunction for selecting top K individuals from population
    for (each ind with top K dfitness in pop){
        popK.add(ind);
        ind.dfitness = ind.dfitness * (1-decay_rate);
        if (ind.dfitness < 0.01) ind.dfitness = 0;
    }
    return popK;
}
```



$$ind.\text{fitness} = (ind.\text{support} - min_sup) \times DatasetSize. \quad (1)$$

$$ind.\text{dfitness} = \begin{cases} ind.\text{fitness}, & \text{initial set} \\ ind.\text{dfitness} \times (1 - DecayRate), & \text{if } ind \text{ is selected} \end{cases} \quad (2)$$

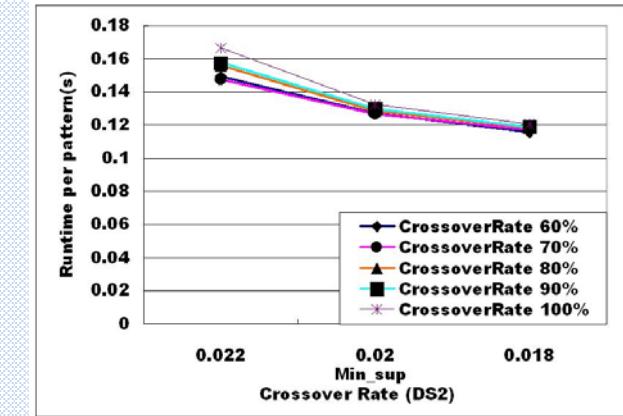
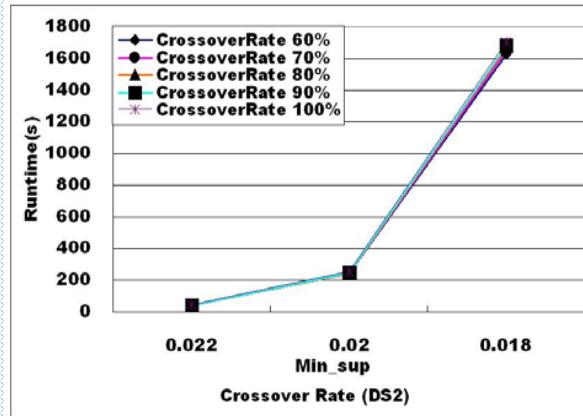
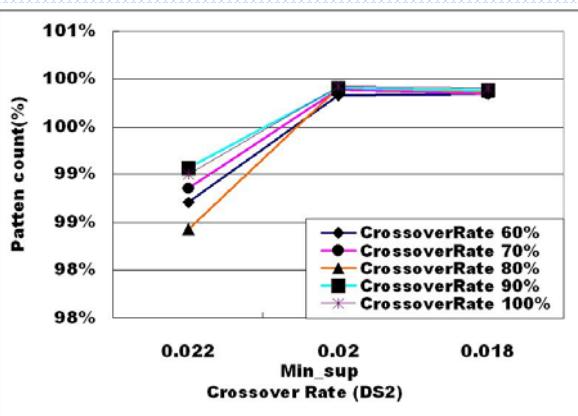
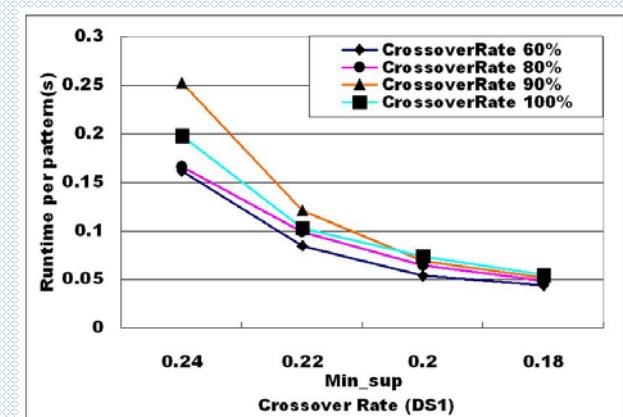
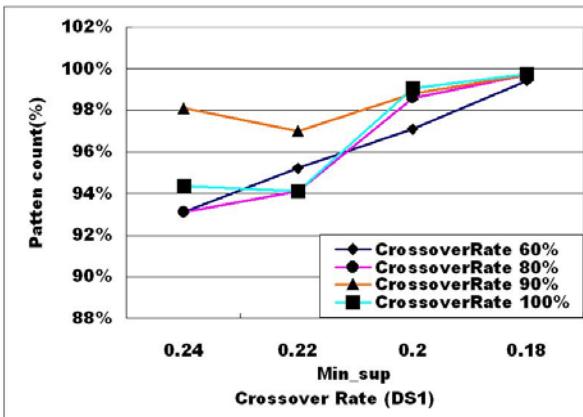
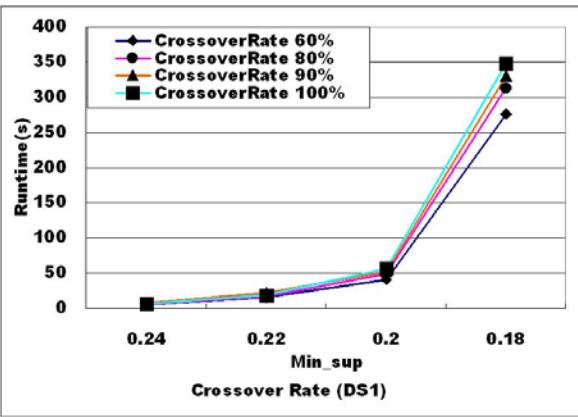
■ GA-NSP Pseudocode

```
RunGA(min-sup, decay-rate, crossover-rate, mutation-rate){
    pop = initialPopulation();
    for (each individual ind in pop){
        ind.fitness = calculateFitness(ind);
        ind.dfitness = ind.fitness
        pop.sum-dfitness = pop.sum-dfitness + ind.dfitness
    }
    while ( pop.sum-dfitness > 0 ){
        popK = Selection(pop);
        if (Random()<crossover-rate) Crossover(popK);
        if (Random()<mutation-rate) Mutation(popK);
        for (each individual ind in popK)
            if (Prune(ind) != true && ind.sup >= min-sup) pop.add(ind);
    }
    return pop;
}
```

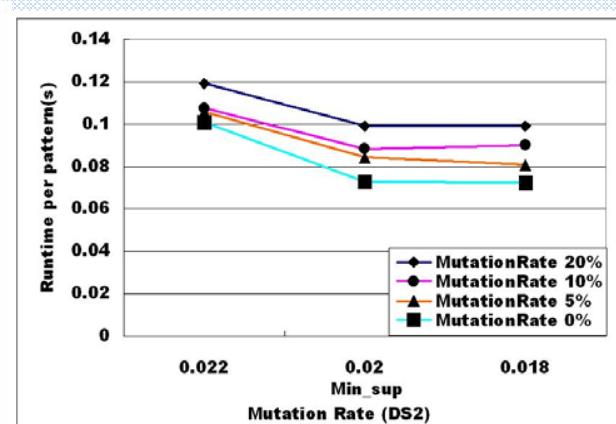
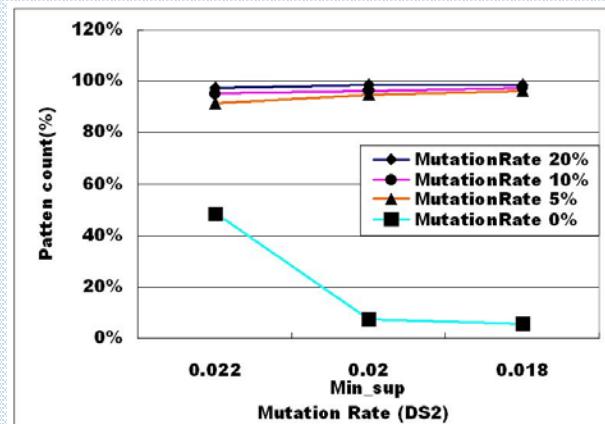
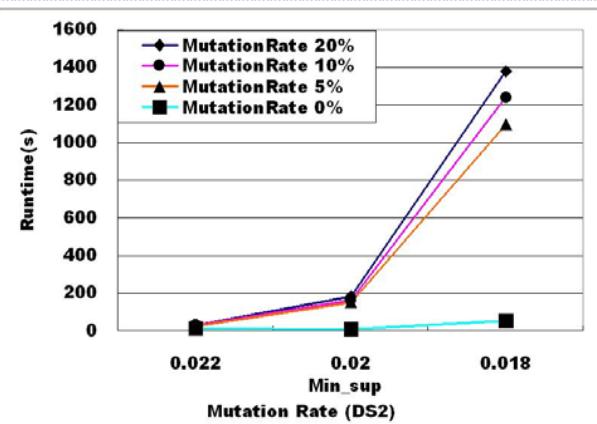
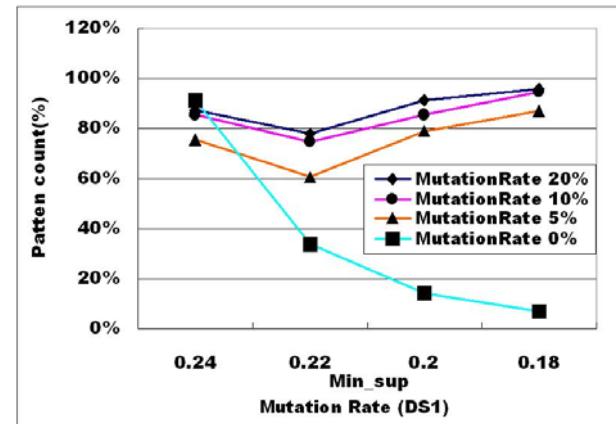
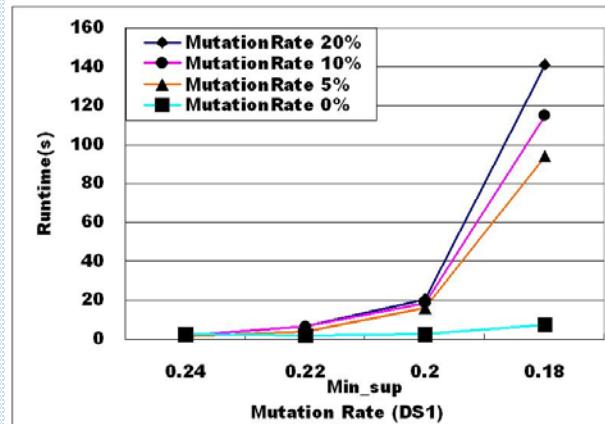
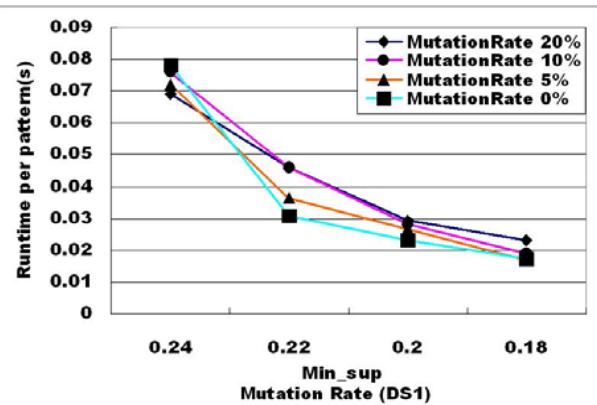
Experiments Result .1

- Datasets
- *Dataset1(DS1)* is C8.T8.S4.I8.DB10k.N1k, which means the average number of elements in a sequence is 8, the average number of items in an element is 8, the average length of a maximal pattern consists of 4 elements and each element is composed of 8 items average. The data set contains 10k sequences, the number of items is 1000.
- *Dataset2(DS2)* is C10.T2.5.S4.I2.5.DB100k.N10k.
- *Dataset3(DS3)* is C20.T4.S6.I8.DB10k.N2k.
- *Dataset4(DS4)* is real application data for insurance claims.

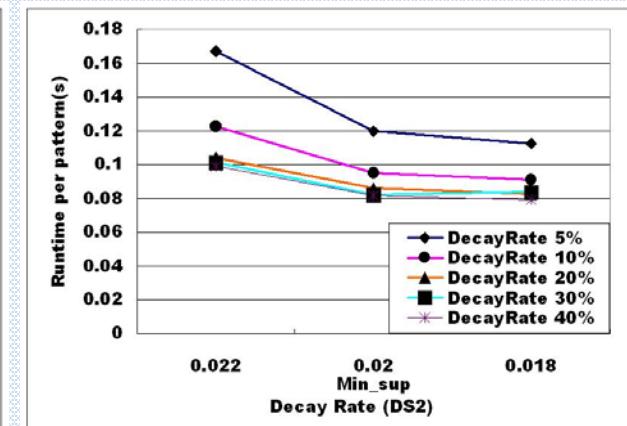
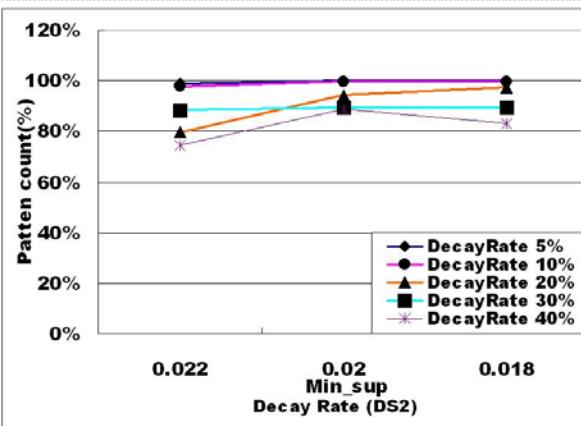
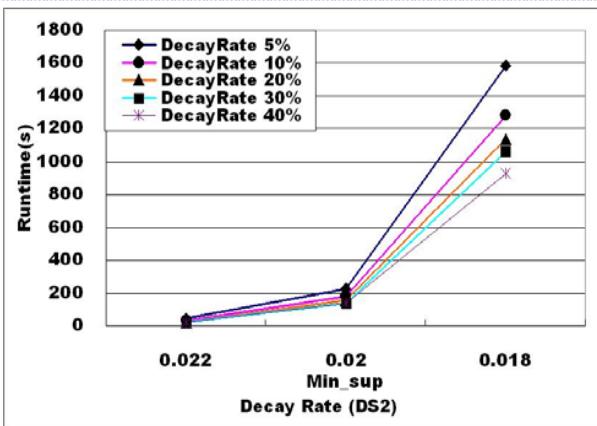
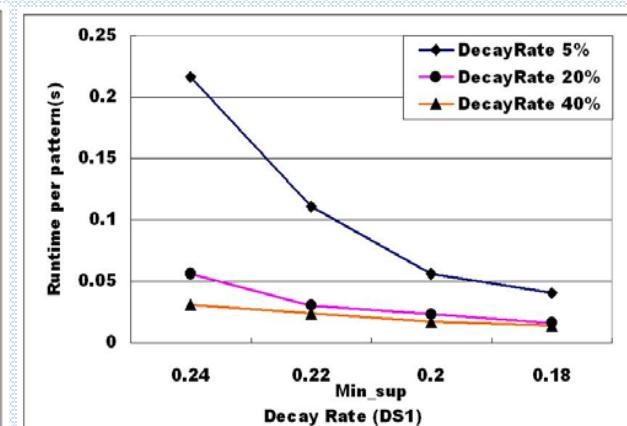
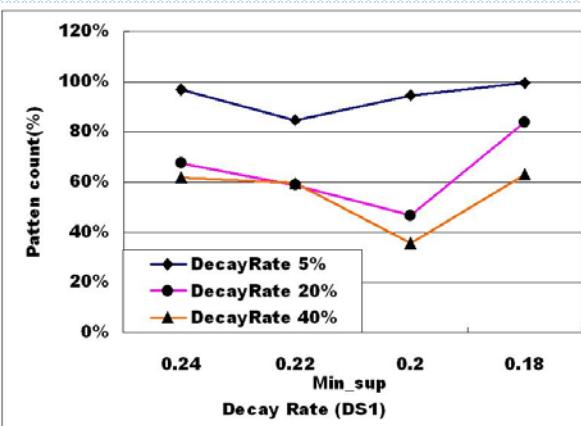
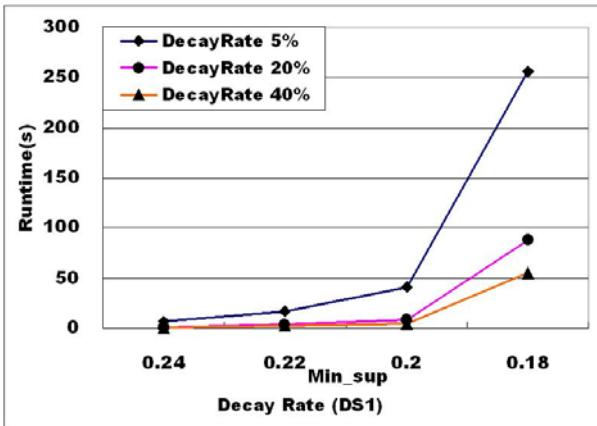
• Crossover Rate



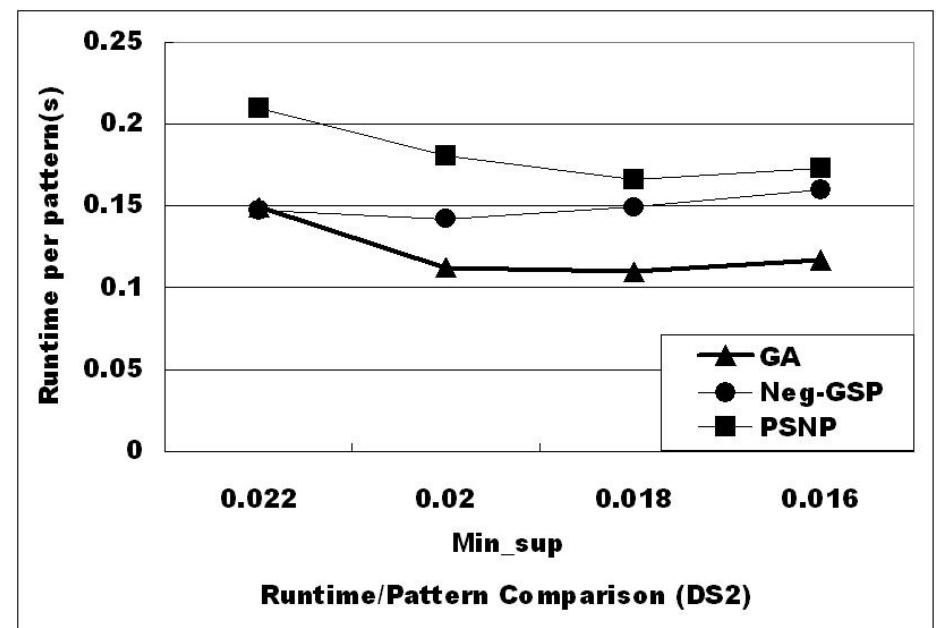
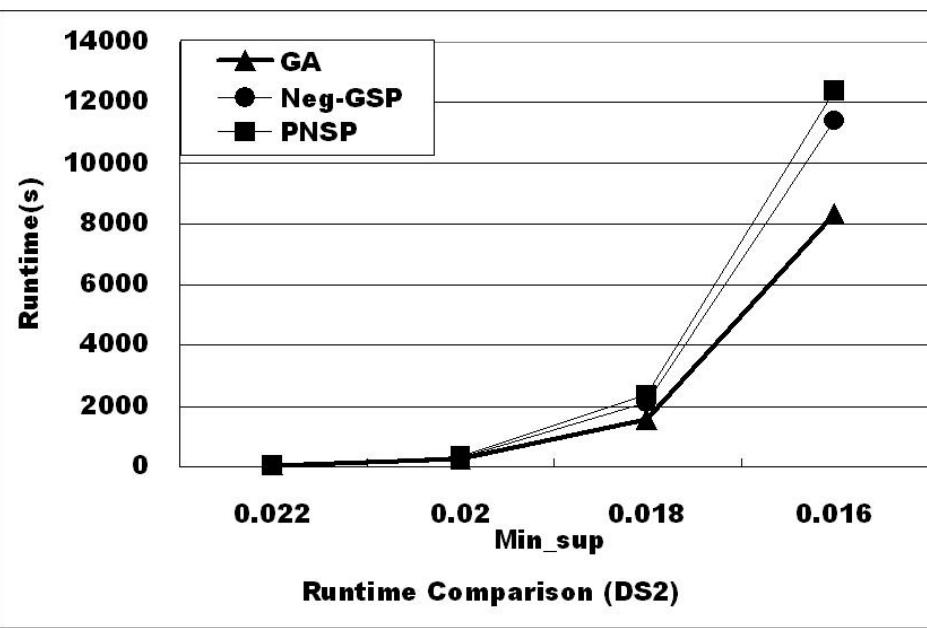
• Mutation Rate



• Decay Rate



- Comparison with PNSP, Neg-GSP



Classification of both positive and negative behavior patterns

- Huafeng Zhang, Yanchang Zhao, Longbing Cao, Chengqi Zhang and Hans Bohlscheid. Customer Activity Sequence Classification for Debt Prevention in Social Security, *Journal of Computer Science and Technology*, 24(6): 1000-1009 (2009).
- Yanchang Zhao, Huafeng Zhang, Shanshan Wu, Jian Pei, Longbing Cao, Chengqi Zhang and Hans Bohlscheid. Debt Detection in Social Security by Sequence Classification Using Both Positive and Negative Patterns, *ECML/PKDD2009*, 648-663.

- Class correlation ratio

$$CCR(p_a \rightarrow \tau) = \frac{\hat{corr}(p_a \rightarrow \tau)}{\hat{corr}(p_a \rightarrow \neg\tau)} = \frac{a \cdot (c + d)}{c \cdot (a + b)},$$

$$\hat{corr}(p_a \rightarrow \tau) = \frac{sup(p_a \cup \tau)}{sup(p_a) \cdot sup(\tau)} = \frac{a \cdot n}{(a + c) \cdot (a + b)}.$$

Table 2. Feature-Class Contingency Table

	p_a	$\neg p_a$	\sum
τ	a	b	$a + b$
$\neg\tau$	c	d	$c + d$
\sum	$a + c$	$b + d$	$n = a + b + c + d$

Table 4. Selected Positive and Negative Sequential Rules

Type	Rule	Support	Confidence	Lift
I	REA ADV ADV → DEB	0.103	0.53	2.02
	DOC DOC REA REA ANO → DEB	0.101	0.33	1.28
	RPR ANO → DEB	0.111	0.33	1.25
	RPR STM STM RPR → DEB	0.137	0.32	1.22
	MCV → DEB	0.104	0.31	1.19
	ANO → DEB	0.139	0.31	1.19
	STM PYI → DEB	0.106	0.30	1.16
II	STM PYR RPR REA RPT → ¬DEB	0.166	0.86	1.16
	MND → ¬DEB	0.116	0.85	1.15
	STM PYR RPR DOC RPT → ¬DEB	0.120	0.84	1.14
	STM PYR RPR REA PLN → ¬DEB	0.132	0.84	1.14
	REA PYR RPR RPT → ¬DEB	0.176	0.84	1.14
	REA DOC REA CPI → ¬DEB	0.083	0.83	1.12
	REA CRT DLY → ¬DEB	0.091	0.83	1.12
	REA CPI → ¬DEB	0.109	0.83	1.12
III	¬{PYR RPR REA STM} → DEB	0.169	0.33	1.26
	¬{PYR CCO} → DEB	0.165	0.32	1.24
	¬{STM RPR REA RPT} → DEB	0.184	0.29	1.13
	¬{RPT RPR REA RPT} → DEB	0.213	0.29	1.12
	¬{CCO RPT} → DEB	0.171	0.29	1.11
	¬{CCO PLN} → DEB	0.187	0.28	1.09
	¬{PLN RPT} → DEB	0.212	0.28	1.08
IV	¬{ADV REA ADV} → ¬DEB	0.648	0.80	1.08
	¬{STM EAN} → ¬DEB	0.651	0.79	1.07
	¬{REA EAN} → ¬DEB	0.650	0.79	1.07
	¬{DOC FRV} → ¬DEB	0.677	0.78	1.06
	¬{DOC DOC STM EAN} → ¬DEB	0.673	0.78	1.06
	¬{CCO EAN} → ¬DEB	0.681	0.78	1.05

Table 5. The Number of Patterns in PS10 and PS05

	PS10 (<i>min_sup</i> = 0.1)		PS05 (<i>min_sup</i> = 0.05)	
	Number	Percent(%)	Number	Percent(%)
Type I	93,382	12.05	127,174	3.93
Type II	45,821	5.91	942,498	29.14
Type III	79,481	10.25	1,317,588	40.74
Type IV	556,491	71.79	846,611	26.18
Total	775,175	100	3,233,871	100

Table 6. Classification Results with Pattern Set PS05-4K

Pattern Number		40	60	80	100	150	200	300
Neg&Pos	Recall	.438	.416	.286	.281	.422	.492	.659
	Precision	.340	.352	.505	.520	.503	.474	.433
	Accuracy	.655	.670	.757	.761	.757	.742	.705
	Specificity	.726	.752	.909	.916	.865	.823	.720
Positive	Recall	.130	.124	.141	.135	.151	.400	.605
	Precision	.533	.523	.546	.472	.491	.490	.483
	Accuracy	.760	.758	.749	.752	.754	.752	.745
	Specificity	.963	.963	.946	.951	.949	.865	.790

e-NSP: Efficient Negative Sequential Pattern Mining Based on Identified Positive Patterns Without Database Rescanning

PSP & NSP

PSP: Positive Sequential Pattern

- Only contain occurring itemsets

E.g. $p1 = \langle a \ b \ c \ X \rangle$.

Existing Methos:

AprioriAll, GSP, FreeSpan, PrefixSpan, SPADE , SPAM

NSP: Negative Sequential Pattern

- Also contain non-occurring itemsets

E.g. $p1 = \langle a \ b \ \neg c \ X \rangle$.

Limited research:

Neg_GSP, PNSP

Difficulties in Mining NSP

- **High Computational Complexity.**

Additionally scanning database after identifying PSP.

- **Large NSC Search Space.**

k-size NSC by conducting a joining operation on (k-1)-size NSP. (NSC : Negative Sequential Candidates)

- **No Unified Definition about Negative Containment.**

How a data sequence contains a negative sequence?

$\langle a \rangle$ contains $\langle \neg a \neg a \rangle$? $\langle a \rangle$ contains $\langle \neg a \ a \neg a \rangle$?

Some Definitions

- **Negative Item/ Element :**
Non-occurring item / element
- **Negative Sequence**
A sequence includes at least one negative item
- **Positive-partner of a Negative Element /Sequence**
 $p(\neg e) = e.$
 $p(<\!a \neg(ab) c\!>) = <\!a(ab) c\!>.$
- **Max Positive Sub-sequence**
 $MPS(<\!a \neg(ab) c\!>) = <\!ac\!>.$

Constraints to Negative Sequence

Constraint 1. Frequency Constraint

This paper only focuses on the negative sequences ns whose positive partner is frequent, i.e., $\text{sup}(p(\text{ns})) \geq \text{min_sup}$.

Constraint 2. Format Constraint

Continuous negative elements in a NSC are not allowed.

- < $\neg(ab)$ $c \neg d$ > ✓
- < $\neg(ab)$ $\neg c$ d > X

Constraint 3. Element Negative Constraint

The minimum negative unit in a NSC is an element.

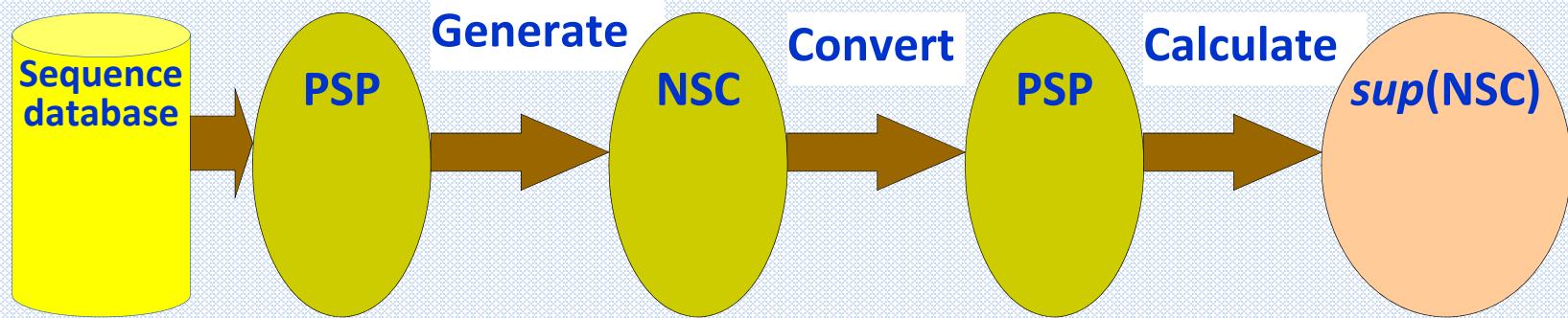
- < $\neg(ab)$ c d > ✓
- < $(\neg ab)$ c d > X

What does This Paper Do

E-NSP: Only use corresponding PSP information to calculate the support of negative sequence, without additionally database scanning.

- A definition about negative containment.
- Three constraints for negative sequence
- A smart method to generate negative sequence candidate (NSC).
- A conversion strategy to convert negative containment problems to positive containment problems.
- A method to calculate the support of NSC.

The framework of E-NSP



1. Mine all PSP by traditional PSP mining algorithms;
2. Generate NSC based on these PSP;
3. Convert these NSC to corresponding PSP;
4. Get supports of NSC by calculating support of corresponding PSP.

Negative Containment Definition

Definition 4. Negative Containment Definition

Let $ds = \langle d_1 \ d_2 \ \dots \ d_t \rangle$ be a data sequence, $ns = \langle s_1 \ s_2 \ \dots \ s_m \rangle$ be an *m-size* and *n-neg-size* negative sequence, (1) if $m > 2t+1$, then ds does not contain ns ; (2) if $m=1$ and $n=1$, then ds contains ns when $p(ns) \not\subseteq ds$; (3) otherwise, ds contains ns if, $\forall (s_i, id(s_i)) \in EidS_{ns}^-$ ($1 \leq i \leq m$), one of the following three holds:

- (a) ($lsb=1$) or ($lsb>1 \wedge p(s_1) \not\subseteq \langle d_1 \ \dots \ d_{lsb-1} \rangle$), when $i=1$,
- (b) ($fse=t$) or ($0 < fse < t \wedge p(s_m) \not\subseteq \langle d_{fse+1} \ \dots \ d_t \rangle$), when $i=m$,
- (c) ($fse>0 \wedge lsb=fse+1$) or ($fse>0 \wedge lsb>fse+1 \wedge p(s_i) \not\subseteq \langle d_{fse+1} \ \dots \ d_{lsb-1} \rangle$), when $1 < i < m$,

where $fse=FSE(MPS(\langle s_1 \ s_2 \ \dots \ s_{i-1} \rangle), ds)$, $lsb=LSB(MPS(\langle s_{i+1} \ \dots \ s_m \rangle), ds)$.

Negative Containment Definition

$$\begin{array}{ccc} ns = < ns_{left}, & \neg e, & ns_{right} > \\ MPS(ns_{left}) & e & MPS(ns_{right}) \\ \text{---} & \text{---} & \text{---} \\ \subseteq & \not\subseteq & \subseteq \\ \text{---} & \text{---} & \text{---} \\ ds = < s_1, \dots, s_i, s_{i+1}, \dots, s_{j-1}, s_j, \dots, s_t > \end{array}$$

ds contains ns if $< s_1, \dots, s_i >$ contain $MPS(ns_{left})$, $< s_j, \dots, s_t >$ contain $MPS(ns_{right})$, and $< s_{i+1}, \dots, s_{j-1}, \dots >$ doesn't contain $< e >$. (To EACH negative element $\neg e$ in ns)

Example: Negative Containment Definition

$ns = \langle a \neg b b(cde) \rangle.$ $ds = \langle a(bc)d(cde) \rangle.$

$$\begin{array}{ccc} < & a & \neg b & b(cde) \rangle \\ \text{\scriptsize } \sqsubseteq & \text{\scriptsize } \not\sqsubseteq & \text{\scriptsize } \sqsubseteq \\ \text{\scriptsize } \sqcup & \text{\scriptsize } \sqcup & \text{\scriptsize } \sqcup \\ ds = & \langle a & (bc)d(cde) \rangle. \end{array}$$

ds contains ns .

Definitions

1-neg-size Maximum Sub-sequence is a sequence that includes $MPS(ns)$ and one negative element e in original sequence order.

1-neg-size maximum sub-sequence set is a set that includes all 1-neg-size maximum sub-sequences of ns , denoted as 1-negMSS_{ns} .

Example $ns = \langle a \neg b c \neg d \rangle$,

$1\text{-negMSS}_{ns} = \{ \langle a \neg b c \rangle, \langle a c \neg d \rangle \}$

Negative Conversion Strategy

Given a data sequence $ds = \langle d_1 \ d_2 \ \dots \ d_t \rangle$, and $ns = \langle s_1 \ s_2 \ \dots \ s_m \rangle$, which is an m -size and n -neg-size negative sequence, the negative containment definition can be converted as follows: data sequence ds contains negative sequence ns if and only if the two conditions hold: (1) $MPS(ns) \subseteq ds$; and (2) $\forall 1\text{-}negMS \in 1\text{-}negMSS_{ns}, p(1\text{-}negMS) \not\subseteq ds$.

Example $ns = \langle a\neg bb\neg a(cde) \rangle$, $ds = \langle a(bc)d(cde) \rangle$.

$1\text{-}negMSS_{ns} = \{ \langle a\neg bb(cde) \rangle, \langle ab\neg a(cde) \rangle \}$

(1) $MPS(ns) = \langle ab(cde) \rangle \subseteq ds$;

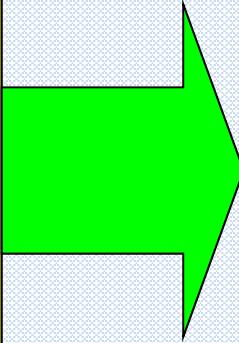
ds contains ns

(2) $p(\langle a\neg bb(cde) \rangle) = \langle abb(cde) \rangle \not\subset ds$;

$p(\langle ab\neg a(cde) \rangle) = \langle aba(cde) \rangle \not\subset ds$;

Negative Conversion Strategy

problem
whether a data
sequence contains
a negative
sequence



problem
whether the data
sequence does not
contain its
corresponding
positive sequences

Now we can calculate the support of NSC only
using the NSC's corresponding PSP.

Calculate the Support of NS

$$sup(ns) = |\{ns\}| = |\{MPS(ns)\} - \bigcup_{i=1}^n \{p(1-negMS_i)\}| \quad (1)$$

Because $\bigcup_{i=1}^n \{p(1-negMS_i)\} \subseteq \{MPS(ns)\}$, equation 1 can be rewritten as:

$$\begin{aligned} sup(ns) &= |\{MPS(ns)\}| - |\bigcup_{i=1}^n \{p(1-negMS_i)\}| \\ &= sup(MPS(ns)) - |\bigcup_{i=1}^n \{p(1-negMS_i)\}| \end{aligned} \quad (2)$$

Example 10 $sup(<\neg a \neg b c \neg d e>) = sup(<ace>) - |\{<abce>\} \cup \{<acde>\}|;$

$$sup(<\neg aa \neg a>) = sup(<a>) - |\{<aa>\} \cup \{<aa>\}| = sup(<a>) - sup(<aa>).$$

If ns only contains a negative element, the support of ns is:

$$sup(ns) = sup(MPS(ns)) - sup(p(ns)) \quad (3)$$

Example 11 $sup(<\neg a \neg b c e>) = sup(<ace>) - sup(<abce>)$

Specially, for negative sequence $<\neg e>$,

$$sup(<\neg e>) = |D| - sup(<e>). \quad (4)$$

Calculate the Support of NS

$$\begin{aligned} sup(ns) &= | \{MPS(ns)\} | - | \cup_{i=1}^n \{p(1-negMS_i)\} | \\ &= sup(MPS(ns)) - | \cup_{i=1}^n \{p(1 - negMS_i)\} | \quad (2) \end{aligned}$$

Known

PSP	Support	{sid}
$\langle a \rangle$	4	-
$\langle b \rangle$	3	-
$\langle c \rangle$	2	-
$\langle a \ a \rangle$	3	{20,30,40}
$\langle a \ b \rangle$	3	{10,20,30}
$\langle a \ c \rangle$	2	{10,30}
$\langle b \ c \rangle$	2	{10,30}
$\langle (ab) \rangle$	2	-
$\langle a \ b \ c \rangle$	2	{10,30}
$\langle a \ (ab) \rangle$	2	{20,30}

Calculate the union set of $\{p(1-negMS_i)\}$.
 $(p(1-negMS_i)$ are frequent.)

Negative Sequential Candidates Generation

e-NSP Candidate Generation

For a k -size PSP, its NSC are generated by changing any m non-contiguous element(s) to its (their) negative one(s), $m=1, 2, \dots, \lceil k/2 \rceil$, where $\lceil k/2 \rceil$ is a minimum integer that is not less than $k/2$.

Example. $s = <(ab) c d>$ include:

$m=1, <\neg(ab) c d>, <(ab) \neg cd>, <(ab) c \neg d>;$

$m=2, <\neg(ab) c \neg d>.$

An Example

Table 1: Example Data Set

Sid	Data Sequence
10	$\langle a \ b \ c \rangle$
20	$\langle a \ (ab) \rangle$
30	$\langle (ae) \ (ab) \ c \rangle$
40	$\langle a \ a \rangle$
50	$\langle d \rangle$

Table 2: Example Result - Positive Patterns

PSP	Support	{sid}
$\langle a \rangle$	4	-
$\langle b \rangle$	3	-
$\langle c \rangle$	2	-
$\langle a \ a \rangle$	3	{20,30,40}
$\langle a \ b \rangle$	3	{10,20,30}
$\langle a \ c \rangle$	2	{10,30}
$\langle b \ c \rangle$	2	{10,30}
$\langle (ab) \rangle$	2	-
$\langle a \ b \ c \rangle$	2	{10,30}
$\langle a \ (ab) \rangle$	2	{20,30}

An Example

Table 3: Example Result - NSC and Support (min_sup=2)

PSP	NSC	Related PSP	Sup
$\langle a \rangle$	$\langle \neg a \rangle$	$\langle a \rangle$	1
$\langle b \rangle$	$\langle \neg b \rangle$	$\langle b \rangle$	2
$\langle c \rangle$	$\langle \neg c \rangle$	$\langle c \rangle$	3
$\langle a\ a \rangle$	$\langle \neg a\ a \rangle$ $\langle a\ \neg a \rangle$	$\langle a \rangle, \langle a\ a \rangle$ $\langle a \rangle, \langle a\ \neg a \rangle$	1 1
$\langle a\ b \rangle$	$\langle \neg a\ b \rangle$ $\langle a\ \neg b \rangle$	$\langle b \rangle, \langle a\ b \rangle$ $\langle a \rangle, \langle a\ \neg b \rangle$	0 1
$\langle a\ c \rangle$	$\langle \neg a\ c \rangle$ $\langle a\ \neg c \rangle$	$\langle c \rangle, \langle a\ c \rangle$ $\langle a \rangle, \langle a\ \neg c \rangle$	0 2
$\langle b\ c \rangle$	$\langle \neg b\ c \rangle$ $\langle b\ \neg c \rangle$	$\langle c \rangle, \langle b\ c \rangle$ $\langle b \rangle, \langle b\ \neg c \rangle$	0 1
$\langle (ab) \rangle$	$\langle \neg (ab) \rangle$	$\langle (ab) \rangle$	3
$\langle a\ (ab) \rangle$	$\langle \neg a\ (ab) \rangle$ $\langle a\ \neg (ab) \rangle$	$\langle (ab) \rangle, \langle a\ (ab) \rangle$ $\langle a \rangle, \langle a\ \neg (ab) \rangle$	0 2
$\langle a\ b\ c \rangle$	$\langle \neg a\ b\ c \rangle$ $\langle a\ \neg b\ c \rangle$ $\langle a\ b\ \neg c \rangle$ $\langle \neg a\ b\ \neg c \rangle$	$\langle b\ c \rangle, \langle a\ b\ c \rangle$ $\langle a\ c \rangle, \langle a\ b\ c \rangle$ $\langle a\ b \rangle, \langle a\ b\ c \rangle$ $\langle b \rangle, \langle a\ b \rangle, \langle b\ c \rangle$	0 0 1 0

An Example

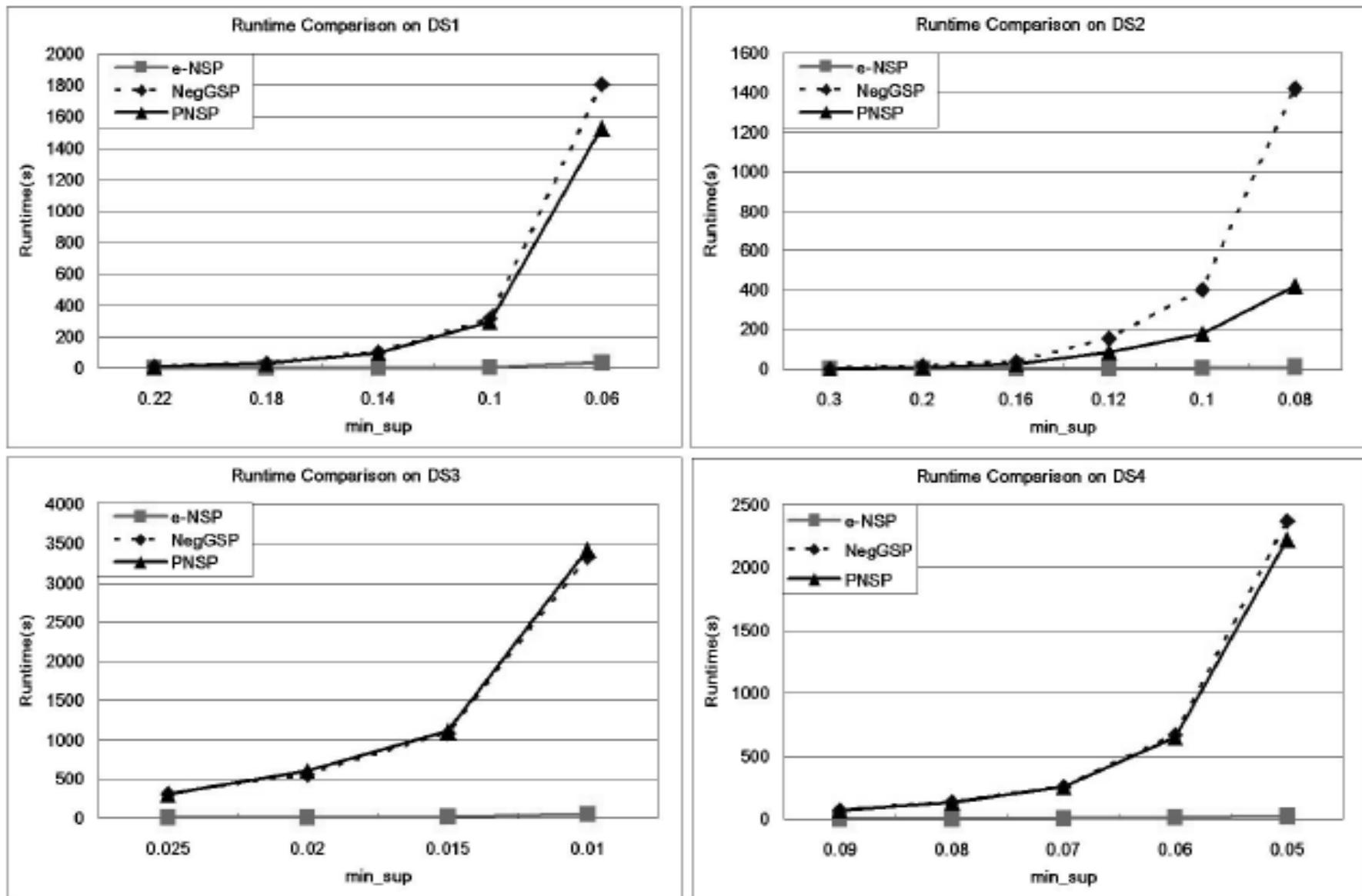
Experiment and Evaluation

Table 4: Dataset Characteristics Analysis Result

ID	Dataset Characteristics	min sup	NGSP (t_1, s)	PNSP (t_2, s)	eNSP (t_3, s)	t_3 / t_2
DS1	C8T4S6I6.DB10k.N100	0.04	1451.7	638.2	14.94	2.3%
		0.06	241.4	163.1	4.16	2.5%
		0.08	78.9	61.9	1.53	2.5%
DS1.1	<u>C4</u> T4S6I6.DB10k.N100	0.01	517.5	208.4	1.08	0.5%
		0.015	130.4	64.5	0.33	0.5%
		0.02	48.0	28.4	0.16	0.5%
DS1.2	<u>C12</u> T4S6I6.DB10k.N100	0.14	229.0	191.9	7.99	4.2%
		0.16	127.6	109.5	4.49	4.1%
		0.18	73.8	66.9	2.53	3.8%
DS1.3	C8 <u>T8</u> S6I6.DB10k.N100	0.22	130.8	118.5	5.22	4.4%
		0.24	83.7	76.5	3.19	4.2%
		0.26	55.9	52.8	2.14	4.1%
DS1.4	C8 <u>T12</u> S6I6.DB10k.N100	0.3	1205.2	969.3	57.55	5.9%
		0.4	133.2	123.5	6.75	5.5%
		0.5	23.6	23.0	1.06	4.6%
DS1.5	C8T4 <u>S12</u> I6.DB10k.N100	0.04	1130.0	478.6	12.22	2.6%
		0.06	187.0	124.7	3.39	2.7%
		0.08	61.2	47.5	1.23	2.6%
DS1.6	C8T4 <u>S18</u> I6.DB10k.N100	0.04	297.1	157.4	3.47	2.2%
		0.06	64.2	45.5	0.97	2.1%
		0.08	23.5	19.0	0.36	1.9%
DS1.7	C8T4S6 <u>I10</u> .DB10k.N100	0.06	690.2	395.1	7.33	1.9%
		0.07	334.7	227.5	4.23	1.9%
		0.08	188.1	138.0	2.63	1.9%
DS1.8	C8T4S6 <u>I14</u> .DB10k.N100	0.08	983.9	630.8	8.88	1.4%
		0.1	320.5	248.9	3.63	1.5%
		0.12	141.8	112.7	1.61	1.4%
DS1.9	C8T4S6I6.DB10k. <u>N200</u>	0.03	378.2	98.4	0.59	0.6%
		0.04	101.8	43.1	0.17	0.4%
		0.05	39.5	23.3	0.06	0.3%
DS1.10	C8T4S6I6.DB10k. <u>N400</u>	0.015	823.0	97.4	0.08	0.1%
		0.02	197.3	42.0	0.03	0.1%
		0.025	99.8	20.6	0.02	0.1%

Computational Cost

Experiment and Evaluation



Conclusions

We have proposed a simple but very efficient NSP mining algorithm: e-NSP. E-NSP includes:

- A formal definition, negative containment, to define how a data sequence contains a negative sequence.
- A negative conversion strategy to convert negative containing problems to positive containing problems.
- A method to calculate the supports of NSC only using the corresponding PSP.
- A simple but efficient approach to generate NSC.
- The experimental results and comparisons on 14 datasets from different data characteristics perspectives have clearly shown that e-NSP is much more efficient than existing approaches.

Group discussion: negative behaviour

Negative behaviour

Organization: _____

Business problem: _____

Business areas	Negative behaviour	Behaviour impact

Part IV.

Group Behavior Analysis

Learning Objectives

- What are group behaviors?
- How to formalize group behaviors?
- How to analyze group behavior?

Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors

Longbing Cao
Faculty of Engineering and IT
University of Technology
Sydney
lbciao@it.uts.edu.au

Yuming Ou
Faculty of Engineering and IT
University of Technology
Sydney
yuming@it.uts.edu.au

Philip S Yu
Department of Computer
Science
University of Illinois at Chicago
peyu@cs.uic.edu

Gang Wei
Department of Surveillance
Shanghai Stock Exchange

ABSTRACT

In capital market surveillance, an emerging trend is that a group of hidden manipulators collaborate with each other to manipulate three trading sequences: buy-orders, sell-orders and trades, through carefully arranging their prices, volumes and time, in order to mislead other investors, affect the instrument movement, and thus maximize personal benefits. If the focus is on only one of the above three sequences in attempting to analyze such hidden group based behavior, or if they are merged into one sequence as per an investor, the coupling relationships among them indicated through trading actions and their prices/volumes/times would be missing, and the resulting findings would have a high probability of mismatching the genuine fact in business. Therefore, typical sequence analysis approaches, which mainly identify patterns on a single sequence, cannot be used here. This paper addresses a novel topic, namely coupled behavior analysis in hidden groups. In particular, we propose a coupled Hidden Markov Models (HMM)-based approach to detect abnormal group-based trading behaviors. The resulting models cater for (1) multiple sequences from a group of people, (2) interactions among them, (3) sequence item properties, and (4) significant change among coupled sequences. We demonstrate our approach in detecting abnormal manipulative trading behaviors on orderbook-level stock data. The results are evaluated against alerts generated by the exchange's surveillance system from both technical and computational perspectives. It shows that the proposed coupled and adaptive HMMs outperform a standard HMM only modeling any single sequence, or the HMM combining multiple single sequences, without considering the coupling relationship. Further work on coupled behavior analysis, including coupled sequences/event analysis, hidden group analysis and behavior dynamics are very critical.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD'10, July 25–28, 2010, Washington, DC, USA.
Copyright 2010 ACM 978-1-4503-0055-1/10/07 ..\$10.00.

Categories and Subject Descriptors

H.2.8 [Information Systems]: Database applications—
Data Mining

General Terms

Algorithms, Economics, Security

Keywords

Coupled behavior analysis, coupled sequence analysis, sequence item property, sequence change, hidden group discovery, coupled hidden Markov model, abnormal behavior detection, market manipulation

1. INTRODUCTION

Abnormal behavior detection plays an important role in capital market surveillance [5] and risk management. The ongoing global financial crisis and recent urge regulation bodies to undertake a deep investigation of trading behaviors in capital markets. A emerging abnormal trading situation is that a group of experienced market manipulators collaborate with each other to manipulate an instrument by fine-tuning its prices/volumes and trading time, in order to mislead other investors. Once the instrument's market price reaches a comfortable level, these manipulators immediately take advantage of the market movement. It is very challenging to detect such hidden group-based manipulative behaviors. In fact, similar coupled behaviors (as well as sequences and events) can be found in many domains, including intrusion detection, crime and national security.

In stock markets, trading transactions consist of multiple streams, in which three typical trading behavioral sequences – buy orders, sell orders and trades from the manipulators - are coupled with each other in terms of timing, prices and volumes etc., according to a market's trading model and investor intention [6]. Often only an individual sequence in such multiple coupled sequences (e.g., trades) is focused on for pattern analysis, while the quotes-related actions and the action price and volume information associated with trades are missing. As a result, we cannot detect those group-based manipulative trading behaviors. Alternatively, if buys and sells are also combined with trades, we may identify more informative patterns disclosing the full trading process and rel-

Behavior sequence analysis

Longbing Cao, Yuming Ou,
Philip S Yu, Gang
Wei. Detecting Abnormal
Coupled Sequences and
Sequence Changes in Group-
based Manipulative Trading
Behaviors, KDD2010, 85-94.

Coupled Behavior Analysis with Applications

Longbing Cao, Senior Member, IEEE, Yuming Ou, and Philip S. Yu, Fellow, IEEE

Abstract—Coupled behaviors refer to the activities of one or many actors who are associated with each other in terms of certain relationships. With increasing network and community-based events and applications, such as geo-uploaded items and social network interactions, behavior coupling contributes to the causes of eventual business problems. Effective approaches for analyzing coupled behaviors are not available, since existing methods mainly focus on individual behavior analysis. This paper discusses the problem of Coupled Behavior Analysis (CBA) and its challenges. A Coupled Hidden Markov Model (CHMM)-based approach is illustrated to model and detect abnormal group-based trading behaviors. The CHMM model can handle 1) multiple behaviors from a group of people, 2) behavioral properties, 3) interactions among behaviors, customers, and behavioral properties, and 4) significant changes between coupled behaviors. We demonstrate and evaluate the model on order-book-level stock tick data from a major Asian exchange and demonstrate that the proposed CHMMs outperform HMM-only for modeling a single sequence or combining multiple single sequences without considering coupling relationships to detect anomalies. Finally, we discuss interaction relationships and models between coupled behaviors, which are worthy of substantial study.

Index Terms—Coupled behavior analysis, coupled sequence analysis, hidden group discovery, coupled hidden Markov model, abnormal behavior detection.

1 INTRODUCTION

BEHAVIOR analysis is an essential activity in many fields, from social and behavioral sciences to computer science [32], [33], [34], [35], [36], [37]. Although there is an emerging focus on deep behavior studies such as periodic behavior analysis [31] and social network analysis [30], previous research has mainly focused on individual behaviors. In practice, behaviors from either the same, or different actors are often coupled with each other. Coupled behaviors play a much more fundamental role than individuals in the cause, dynamics and effect of business problems [28], [7], [29], [30], [37].

1.1 Coupled Behavior Applications

While very limited research outcomes can be identified in the literature, coupled behavior is widely researched. As well as the example in Section 3.1, the following are typical coupled behavior applications:

- Group-based criminal behaviors. A group of criminals conduct series of activities in order to achieve their goal. The activities are associated with each other and aim for the same objective.
- Group-based insurance claims. A family or group of insureds lodge similar claims at the same time, or soon after. Another example is where a health care

provider may collaborate with multiple customers to overclaim health benefits by providing frequent visits by the customers for a variety of services. Such group claims may lead to over claims or overuse of services.

- Crossreference citation analysis. From the references cross-cited by relevant groups, we find either genuine collaboration or manipulation of citations.
- Crossmarket manipulation. Investors in an underlying market manipulate a security so that an accomplice can take arbitrage on the corresponding instrument listed in a derivative market.
- Car transport system. At a busy intersection, many cars from different localities compete/cooperate with each other to move in their respective directions.
- Social network interactions. A group of users interact with each other in a social network.
- Intrusion detection. A large number of hackers collaborate to interfere with a website by applying multiple intrusion techniques.

With the deepening and widening of networking, these coupled behaviors are increasing in a wide range of circumstances, in particular, complex networks, communities, organizations, and enterprise applications.

1.2 Challenges in Analyzing Coupled Behaviors

In the above applications, multiple traces of behaviors are often coupled in intrinsic and contextual relationships. The focus on any single trace of behaviors would not contribute to a full understanding of the underlying problem and its comprehensive solutions. It is very difficult to analyze such coupled behaviors.

- Behaviors refer not only to actions such as a buy quote, but also behavioral properties, for instance, the timing, price, and volume associated with a buy. The engagement of behavioral properties in behavior analysis may make the findings much more workable for problem-solving.

• L. Cao and Y. Ou are with the Faculty of Engineering and Information Technology, University of Technology, PO Box 123, Broadway, NSW 2007, Sydney, Australia.
E-mail: longbing.cao@uts.edu.au; yuming@uts.edu.au.

• P.S. Yu is with the Department of Computer Science, University of Illinois, Room 1138 SEO, 205 S. Morgan Street, Chicago, IL 60611.
E-mail: psyu@cs.illinois.edu.

Manuscript received 19 June 2010; revised 11 Sept. 2010; accepted 26 Mar. 2011; and the 7 July 2011.

Recommended for acceptance by S.C. Chen.
For information on obtaining reprints of this article, please send e-mail to: elibrary.ieee.org, and reference IEEECS Log Number TDKD-2010-06-0209.
Digital Object Identifier no. 10.1109/TKDE.2011.128

Longbing Cao, Yuming Ou, Philip S Yu,
[Coupled Behavior Analysis with Applications](#), IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).

Issues Addressed

- Behavior properties described by attribute vector
- How to construct behavior sequences?
- How to handle multiple sequences coupled with each other?
- How to model vector-based behavior sequences?
- How to map vector-based behavior sequences to Coupled Hidden Markov Model?
- How to detect pool manipulation by identifying abnormal coupled sequences?

What is Coupled Behavior?

Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, *Information Science*, 180(17); 3067-3085, 2010.

www.behaviorinformatics.org

Physical world



Intelligent Transport Systems

Virtual world

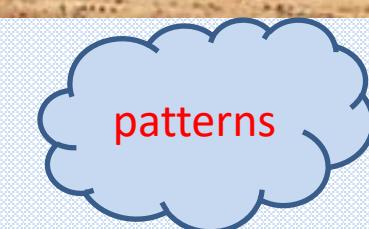
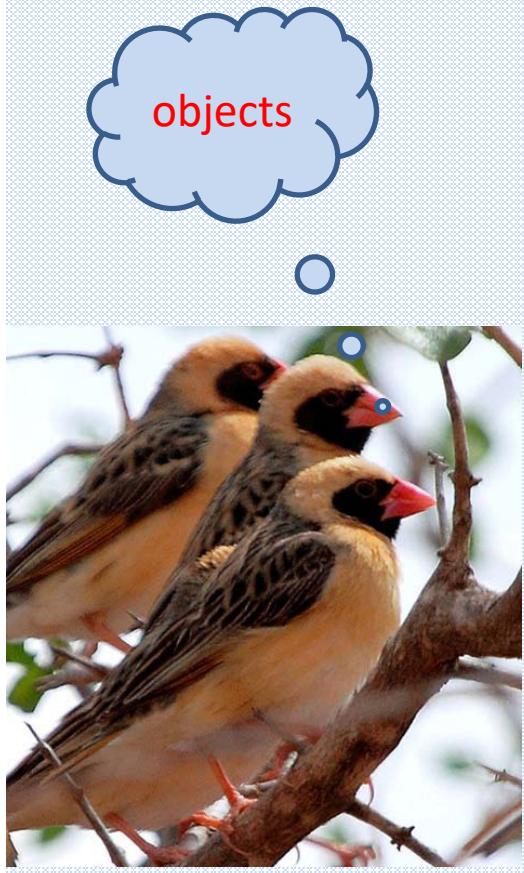


Problem-solving world

Self-organizing behaviors

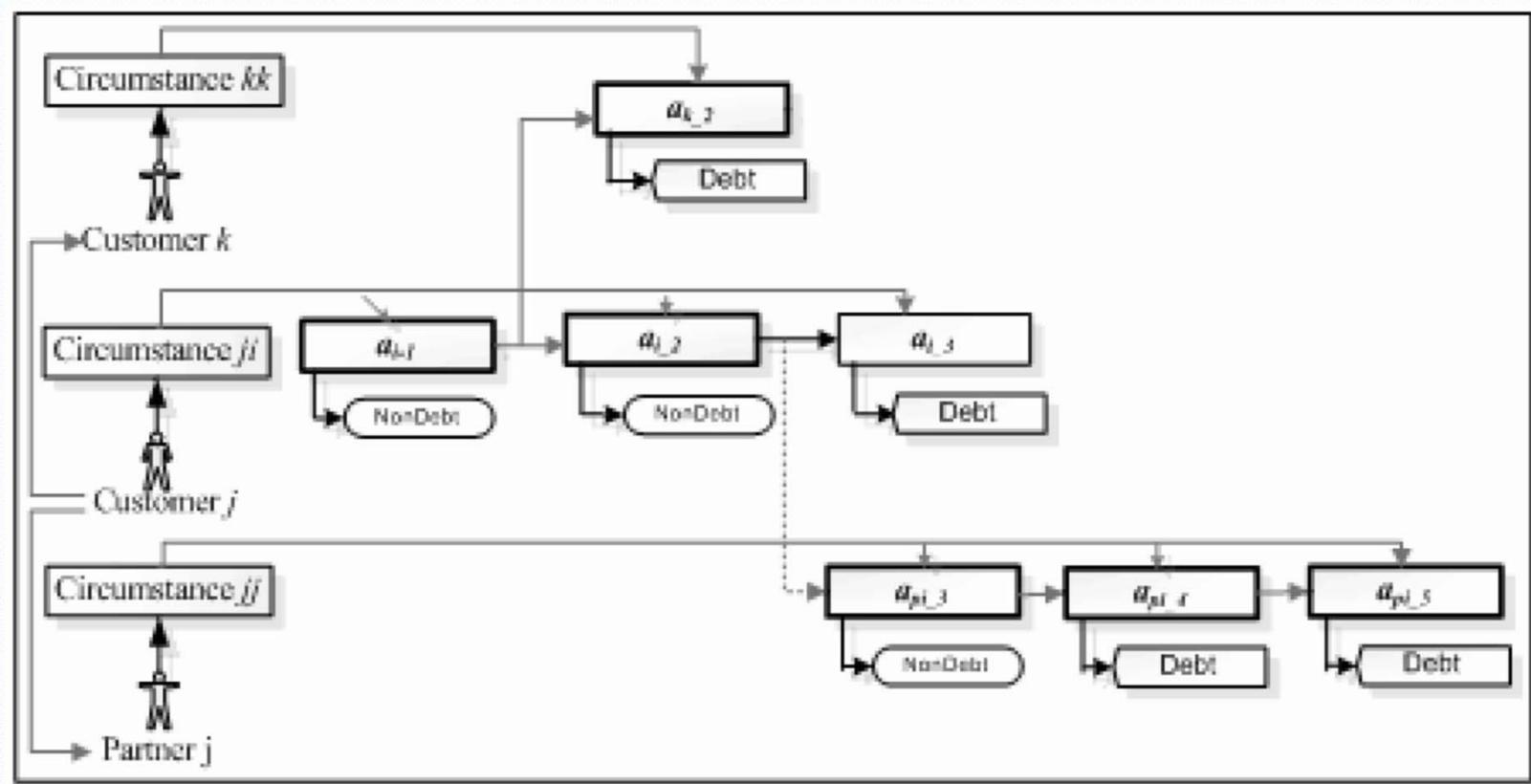


Quelea vs Elephant || Individual vs. Group



Coupled impact-oriented behaviors

- Social security



Relationship crossing behaviors

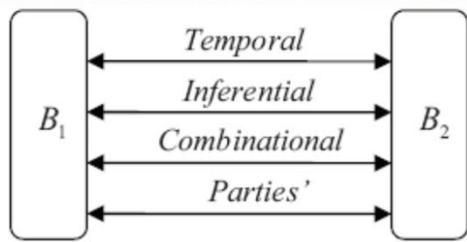


Figure 6: Relationships between Multiple Behaviors

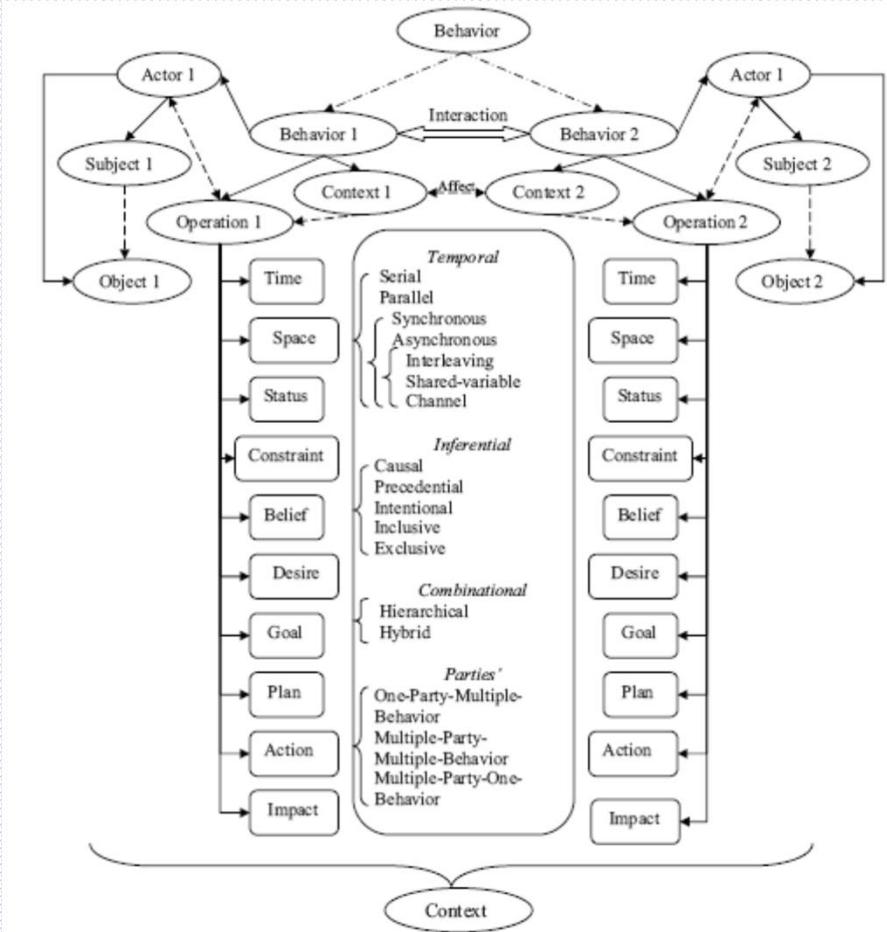


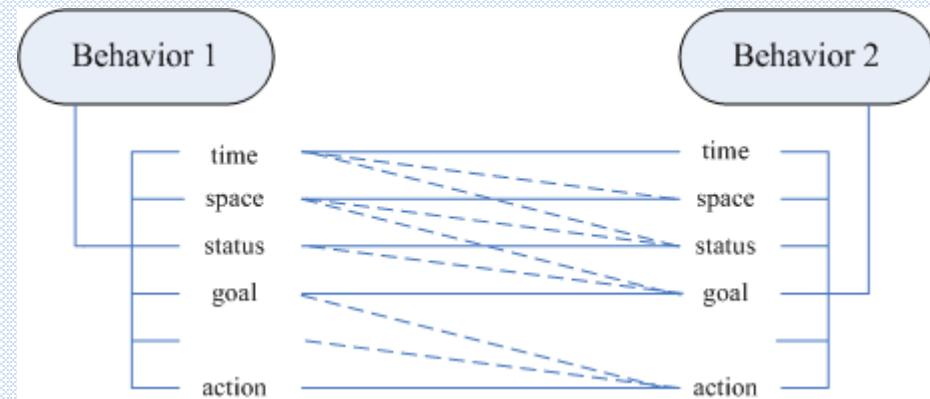
Figure 7: Relationships between behaviors

Relationship crossing objects/behaviors

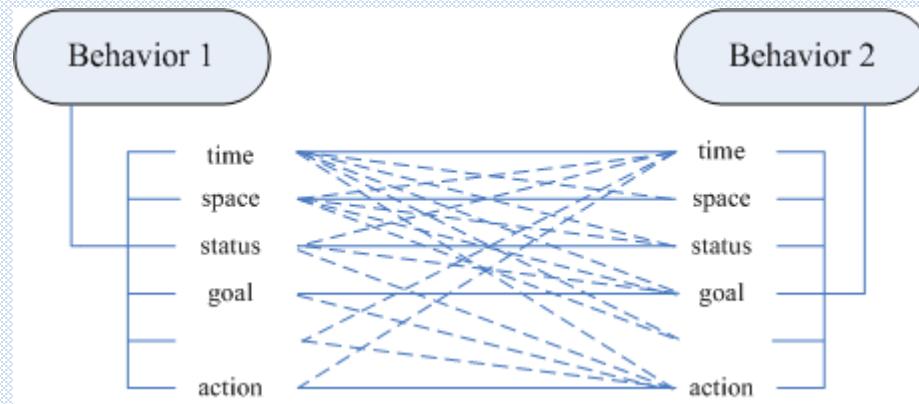
Homogeneous relation



Heterogeneous relation



Mixed relation/coupling relation



Coupling relationships

- From temporal aspect

- Serial Coupling: $TS_1; TS_2; \dots; TS_n$
- Interleaving Coupling: $TS_1 : TS_2 : \dots : TS_n$
- Shared-variable Coupling: $TS_1 || TS_2 || \dots || TS_n$
- Channel System Coupling: $TS_1 | TS_2 | \dots | TS_n$
- Synchronous Coupling: $TS_1 \parallel TS_2 \parallel \dots \parallel TS_n$

- From inferential aspect

- Causal Coupling: $TS_1 \rightarrow TS_2$
- Precedential Coupling: $TS_1 \Rightarrow TS_2$
- Intentional Coupling: $TS_1 \rightarrowtail TS_2$
- Inclusive Coupling: $TS_1 \leftrightarrow TS_2$
- Exclusive Coupling: $TS_1 \oplus TS_2$

- From combinational aspect

- Hierarchical Coupling: $f(g(TS_1, TS_2, \dots, TS_n))$
- Hybrid Coupling: $f(TS_1).g(TS_2)$, $f(TS_1)^*$, $(TS_1)^\omega$

- One-Party-Multiple-Behavior Coupling: $f(TS_1, TS_2, \dots, TS_n)^{[A_1]}$
- Multiple-Party-One-Behavior Coupling: $f(TS_1)^{[A_1 A_2 \dots A_n]}$
- Multiple-Party-Multiple-Behavior Coupling: $f(TS_1, TS_2, \dots, TS_n)^{[A_1 A_2 \dots A_n]}$

Basic Behavior Patterns

- Tracing: Different actions with sequential order.
 $\{a_1, a_2, \dots, a_n\}$
- Consequence: Different actions have causalities in occurrence.
 $\{a_i \rightarrow a_j\}$
- Synchronization: Different actions occur at the same time.
 $\{a_1 \leftrightarrow, \dots, \leftrightarrow a_n\}$
- Combination: Different actions occur in concurrency.
 $\{a_1 \| a_2 \|, \dots, \| a_n\}$

- Exclusion: Different actions occur mutually exclusively.

$$\{a_1 \oplus a_2 \oplus, \dots, \oplus a_n\}$$

- Precedence: Different actions have required precedence

$$\{a_i \Rightarrow a_j\}$$

And more to be explored...

- *Sequential Combination* $\longrightarrow A \times B \times C \times \dots$
- *Parallel Combination* $\longrightarrow A \otimes B \otimes C \otimes \dots$
- *Nested Combination*
- *Fuzzy or probabilistic Combination*

What is the Coupled Behavior Analysis (CBA) problem?

Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Application, *IEEE Trans. Knowledge and Data Engineering*.

Longbing Cao, In-depth Behavior Understanding and Use: the Behavior Informatics Approach, *Information Science*, 180(17); 3067-3085, 2010.

www.behaviorinformatics.org

Individual Objects' behaviors

- Customer a_i 's N behaviors

$$B_i: \{b_{i1}, b_{i2}, \dots, b_{in}\}$$



Remark:

- Individual objects
- Objects are independent
- Individual behaviors
- Behaviors of the same object are somehow dependent

Individual vs Group



If they behave in the same/different way dependently?



If they behave in the same/different way independently?

Group Objects' behaviors

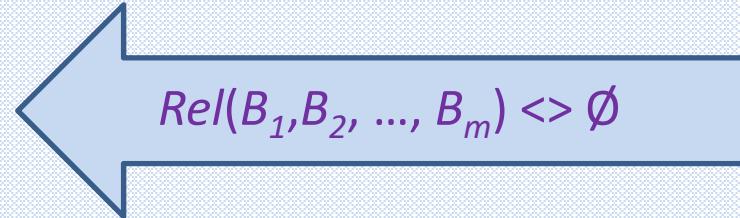
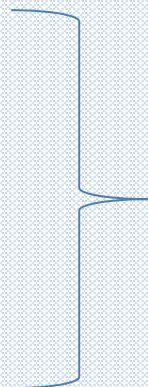
- M customers' behaviors

$B_1: \{b_{11}, b_{12}, \dots, b_{1n}\}$

$B_2: \{b_{21}, b_{22}, \dots, b_{2n}\}$

.....

$B_m: \{b_{m1}, b_{m2}, \dots, b_{mn}\}$



$$Rel(B_1, B_2, \dots, B_m) \neq \emptyset$$



Interactions
between
brains



Individual
dependence

How to
-Team frequent
behaviors?
-Team member
grouping?
- Team member opinion
consensus
(Agree/Reject)?

Behavior Feature Matrix

I actors (customers): $\{\mathcal{E}_1, \mathcal{E}_2, \dots, \mathcal{E}_I\}$

J_i behaviors for an actor \mathcal{E}_i : $\{\mathbb{B}_{i1}, \mathbb{B}_{i2}, \dots, \mathbb{B}_{iJ_i}\}$

Behavior \mathbb{B}_{ij} : $\overrightarrow{\mathbb{B}}_{ij} = ([p_{ij}]_1, [p_{ij}]_2, \dots, [p_{ij}]_K)$

Behavior Feature Matrix:

$$FM(\mathbb{B}) = \begin{pmatrix} \mathbb{B}_{11} & \mathbb{B}_{12} & \dots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \dots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \dots & \mathbb{B}_{IJ_{max}} \end{pmatrix}$$

An Example of Stock Market

Transactional Data



Behavior Feature
Matrix

	Investor	Time	Direction	Price	Volume
B1	(1)	09:59:52	Sell	12.0	155
B2	(2)	10:00:35	Buy	11.8	2000
B3	(3)	10:00:56	Buy	11.8	150
B4	(2)	10:01:23	Sell	11.9	200
B5	(1)	10:01:38	Buy	11.8	200
B6	(4)	10:01:47	Buy	11.9	200
B7	(5)	10:02:02	Buy	11.9	250
B8	(2)	10:02:04	Sell	11.9	500



$$FM(\mathbb{B}) = \begin{pmatrix} B_1 & B_5 & \emptyset \\ B_2 & B_4 & B_8 \\ B_3 & \emptyset & \emptyset \\ B_6 & \emptyset & \emptyset \\ B_7 & \emptyset & \emptyset \end{pmatrix}$$

Behavior Intra-relationship

Definition 2. (*Intra-Coupled Behaviors*) Actor \mathcal{E}_i 's behaviors \mathbb{B}_{ij} ($1 \leq j \leq J_{max}$) are intra-coupled in terms of coupling function $\theta_j(\cdot)$,

$$\mathbb{B}_{i\cdot}^\theta ::= \mathbb{B}_{i\cdot}(\mathcal{E}, \mathcal{O}, \mathcal{C}, \theta) \mid \sum_{j=1}^{J_{max}} \theta_j(\cdot) \odot \mathbb{B}_{ij} \quad (1)$$

$$|\theta_j(\cdot)| \geq \theta_0 \quad (2)$$

where θ_0 is the intra-coupling threshold, $\sum_{j=1}^{J_{max}} \odot$ means the subsequent behavior of \mathbb{B}_i is \mathbb{B}_{ij} intra-coupled with $\theta_j(\cdot)$, and so on, with nondeterminism.

$$FM(\mathbb{B}) = \begin{array}{c} \hline (\mathbb{B}_{11} \quad \mathbb{B}_{12} \quad \dots \quad \mathbb{B}_{1J_{max}}) \\ \hline \mathbb{B}_{21} \quad \mathbb{B}_{22} \quad \dots \quad \mathbb{B}_{2J_{max}} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ \mathbb{B}_{I1} \quad \mathbb{B}_{I2} \quad \dots \quad \mathbb{B}_{IJ_{max}} \end{array}$$

Behavior Inter-relationship

Definition 3. (*Inter-Coupled Behaviors*) Actor \mathcal{E}_i 's behaviors \mathbb{B}_{ij} ($1 \leq i \leq I$) are inter-coupled with each other in terms of coupling function $\eta_i(\cdot)$,

$$\mathbb{B}_{\cdot j}^\eta := \mathbb{B}_{\cdot j}(\mathcal{E}, \mathcal{O}, \mathcal{C}, \eta) | \sum_{i=1}^I \eta_i(\cdot) \odot \mathbb{B}_{ij} \quad (3)$$

$$|\eta_i(\cdot)| \geq \eta_0 \quad (4)$$

where η_0 is the inter-coupling threshold, $\sum_i^I \odot$ means the subsequent behavior of \mathbb{B}_i is \mathbb{B}_{ij} inter-coupled with $\eta_i(\cdot)$, and so on, with nondeterminism.

$$FM(\mathbb{B}) = \left(\begin{array}{c|cccc} \mathbb{B}_{11} & \mathbb{B}_{12} & \dots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \dots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \dots & \mathbb{B}_{IJ_{max}} \end{array} \right)$$

Behavior Relationship

Definition 4 (Coupled Behaviors) Coupled behaviors \mathbb{B}_c refer to behaviors $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$ that are coupled in terms of relationships $f(\theta(\cdot), \eta(\cdot))$, where $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$

$$\mathbb{B}_c = (\mathbb{B}_{i_1 j_1}^\theta)^\eta * (\mathbb{B}_{i_2 j_2}^\theta)^\eta := \mathbb{B}_{ij}(\mathcal{E}, \mathcal{O}, \mathcal{C}, \mathcal{R}) | \sum_{i_1, i_2=1}^I \sum_{j_1, j_2=1}^{J_{max}} f(\theta_{j_1 j_2}(\cdot), \eta_{i_1 i_2}(\cdot)) \odot (\mathbb{B}_{i_1 j_1} \mathbb{B}_{i_2 j_2}) \quad (5)$$

$$FM(\mathbb{B}) = \left(\begin{array}{cccc|cc} \mathbb{B}_{11} & \mathbb{B}_{12} & \dots & \mathbb{B}_{1J_{max}} \\ \mathbb{B}_{21} & \mathbb{B}_{22} & \dots & \mathbb{B}_{2J_{max}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{B}_{I1} & \mathbb{B}_{I2} & \dots & \mathbb{B}_{IJ_{max}} \end{array} \right)$$

Coupled Behavior Analysis

Theorem 1. (*Coupled Behavior Analysis (CBA)*) The analysis of coupled behaviors (CBA Problem for short) is to build the objective function $g(\cdot)$ under the condition that behaviors are coupled with each other by coupling function $f(\cdot)$, and satisfy the following conditions.

$$f(\cdot) := f(\theta(\cdot), \eta(\cdot)), \quad (9)$$

$$g(\cdot)|(f(\cdot) \geq f_0) \geq g_0 \quad (10)$$

Not an easy job to find

$\Theta()$, $\eta()$, $f()$, $g()$

C1	Beer, Diaper, Banana, Harry Potter, iPhone
C2	Apple, Cherry, Blackberry, Plum
C3	Pencil case, Rubber, Lego toy, Scooter
C4	Pear, Cherry, Peach, Plum, Melon, Apple
C5	Beer, iPhone, Fish, Meat
C6	Scooter, Pen, Notebooks

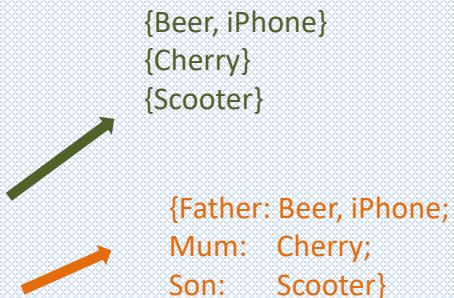
If

C1 – Father; C2 – Mum, C3 – Son;
C4 – Mum, C5 – Father, C6 – Son;
C1.Address = C2.Address = C3.Address;
C4.Address = C5.Address = C6.Address

Then

What will be the difference between

- Outcomes from classic Association Rule or Frequent Pattern Mining
- Outcomes by considering the above coupling relationship?



How to handle CBA

Combined mining

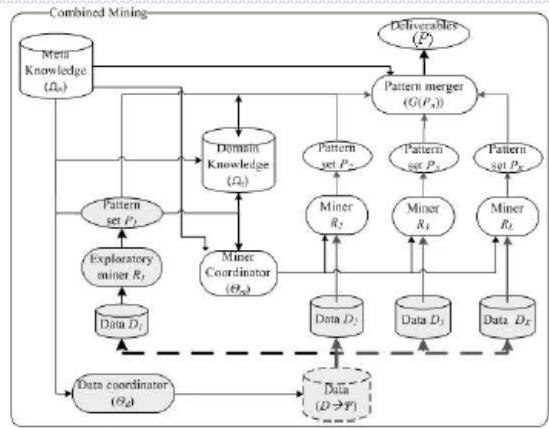
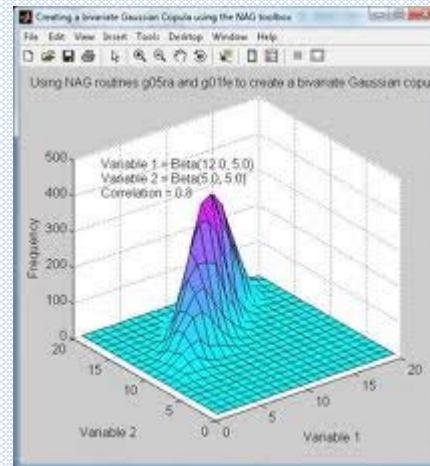
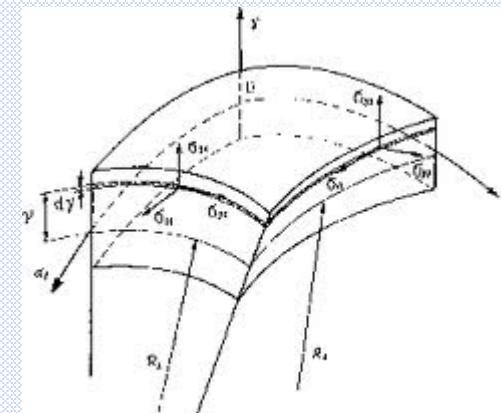


Fig. 1. Combined Mining for Actionable Patterns

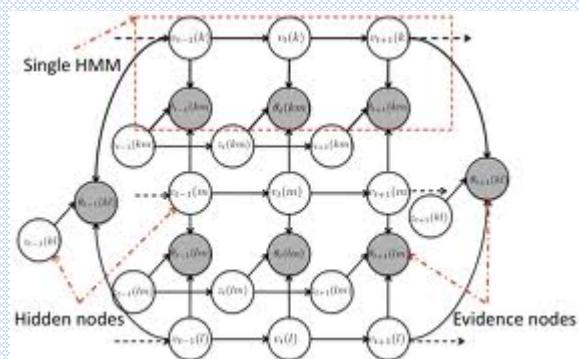
Copula



Tensor theory



Coupled hidden markov model



Combined Pattern Mining in Coupled Objects for High Impact Behavior Analysis

Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, accepted by *IEEE Trans. SMC Part B*

Longbing Cao, Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, *IEEE Trans. on Knowledge and Data Engineering*, 20(8): 1053-1066, 2008.

Combined Pattern Pairs

DEFINITION COMBINED PATTERN PAIRS. *For impact-oriented combined patterns, a Combined Pattern Pair (CPP) is in the form of*

$$\mathcal{P}: \begin{cases} X_1 \rightarrow T_1 \\ X_2 \rightarrow T_2 \end{cases},$$

where 1) $X_1 \cap X_2 = X_p$ and X_p is called the prefix of pair \mathcal{P} ; $X_{1,e} = X_1 \setminus X_p$ and $X_{2,e} = X_2 \setminus X_p$; 2) X_1 and X_2 are different itemsets; and 3) T_1 and T_2 are contrary to each other, or T_1 and T_2 are same but there is a big difference in the interestingness (say confidences $conf$) of the two patterns.

- A combined rule pair is composed of two contrasting rules.
- For customers with same characteristics U , different policies/campaigns, V_1 and V_2 , can result in different outcomes, T_1 and T_2 .

Extended Combined Pattern Pairs

DEFINITION EXTENDED COMBINED PATTERN PAIRS. *An Extended Combined Pattern Pair (ECPP) is a special combined pattern pair as follows*

$$\mathcal{E}: \left\{ \begin{array}{l} X_p \rightarrow T_1 \\ X_p \wedge X_e \rightarrow T_2 \end{array} \right. ,$$

where $X_p \neq \emptyset$, $X_e \neq \emptyset$ and $X_p \cap X_e = \emptyset$.

Extended Combined Pattern Clusters

DEFINITION EXTENDED COMBINED PATTERN SEQUENCES. *An Extended Combined Pattern Sequence (ECPC), or called Incremental Combined Pattern Sequence (ICPS), is a special combined pattern cluster with additional items appending to the adjacent local patterns incrementally.*

$$\mathcal{S}: \left\{ \begin{array}{l} X_p \rightarrow T_1 \\ X_p \wedge X_{e,1} \rightarrow T_2 \\ X_p \wedge X_{e,1} \wedge X_{e,2} \rightarrow T_3 \\ \dots \\ X_p \wedge X_{e,1} \wedge X_{e,2} \wedge \dots \wedge X_{e,k-1} \rightarrow T_k \end{array} \right.,$$

where $\forall i, 1 \leq i \leq k - 1, X_{i+1} \cap X_i = X_i$ and $X_{i+1} \setminus X_i = X_{e,i} \neq \emptyset$, i.e., X_{i+1} is an increment of X_i . The above cluster of rules actually makes a sequence of rules, which can show the impact of the increment of patterns on the outcomes.

Coupled Object Analysis: Combined Demographics + Behavior Analysis

- Longbing Cao, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, *IEEE Trans. SMC Part B*.
- Longbing Cao. Zhao Y., Zhang, C. Mining Impact-Targeted Activity Patterns in Imbalanced Data, *IEEE Trans. on Knowledge and Data Engineering*, 20(8): 1053-1066, 2008.
- Yanchang Zhao, Huaifeng Zhang, Longbing Cao Chengqi Zhang. Combined Pattern Mining: from Learned Rules to Actionable Knowledge, *Australian AI2008*.

Combined Pattern Mining

- Type A: Demographics differentiated combined pattern
 - Customers with the same actions but different demographics
→ different classes/business impact

$$\text{Type A: } \left\{ \begin{array}{l} A_1 + D_1 \rightarrow \text{quick payer} \\ A_1 + D_2 \rightarrow \text{moderate payer} \\ A_1 + D_3 \rightarrow \text{slow payer} \end{array} \right.$$

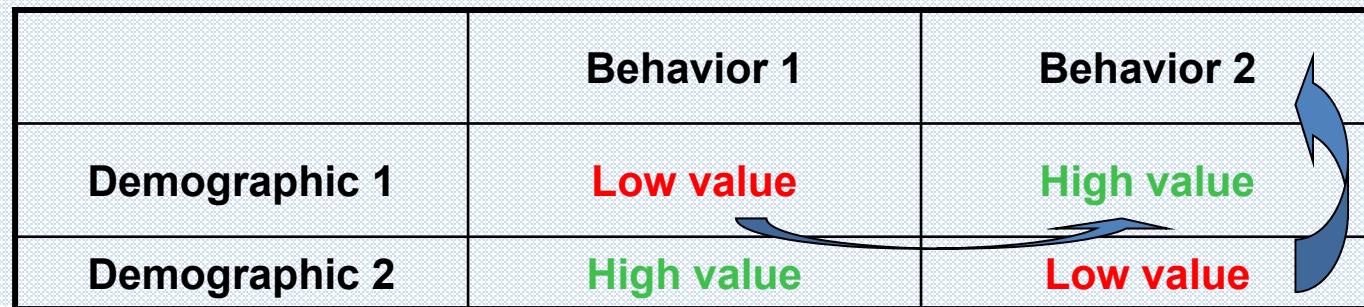
Combined Pattern Mining

- Type B: Action differentiated combined pattern
 - Customers with the same demographics but taking different actions
→ different classes/business impact

Type B: $\begin{cases} A_1 + D_1 & \rightarrow \text{quick payer} \\ A_2 + D_1 & \rightarrow \text{moderate payer} \\ A_3 + D_1 & \rightarrow \text{slow payer} \end{cases}$

An Example of Combined Pattern Clusters

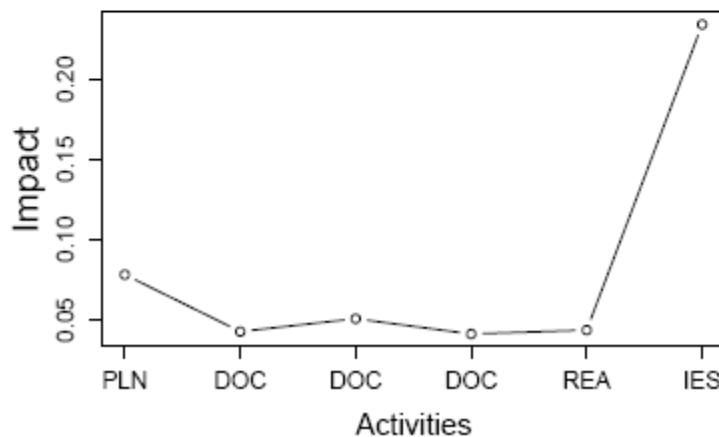
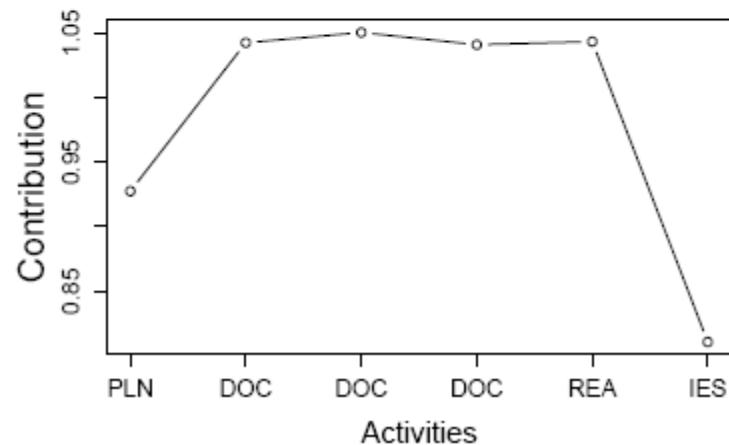
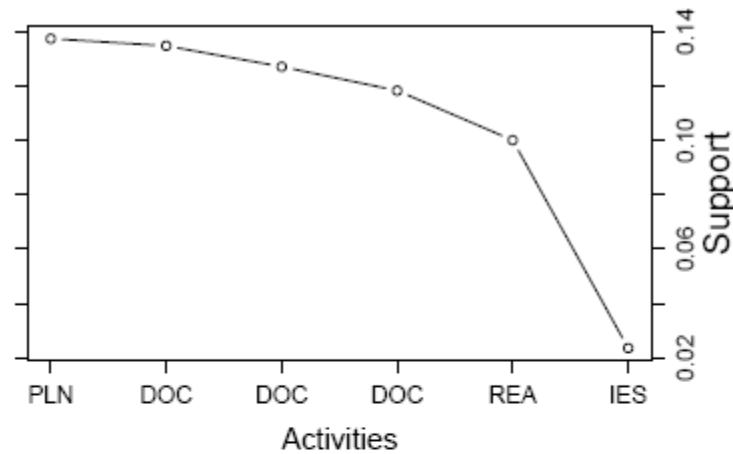
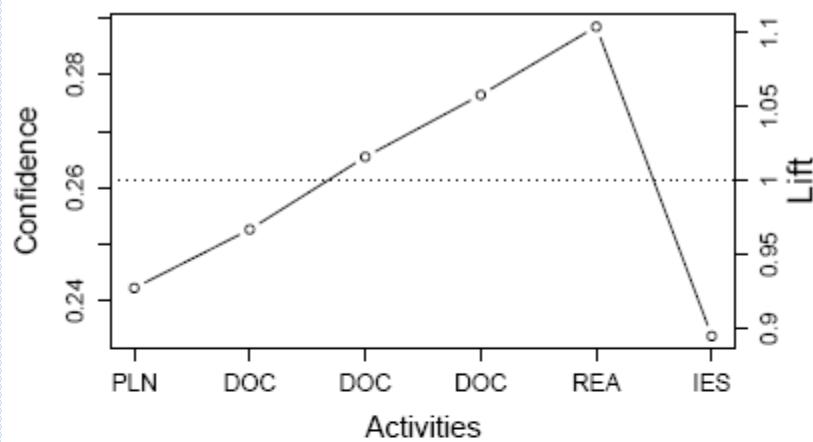
Clusters	Rules	X_p	X_e			T	Cnt	Conf (%)	I_r	I_c	Lift	$Cont_p$	$Cont_e$	Lift of $X_p \rightarrow T$	Lift of $X_e \rightarrow T$
			demographics	arrangements	repayments										
\mathcal{P}_1	P_5	marital:sin &gender:F &benefit:N	irregular	cash or post	A	400	83.0	1.12	0.67	1.80	1.01	2.00	0.90	1.79	
	P_6		withdraw	cash or post	A	520	78.4	1.00		1.70	0.89	1.89	0.90	1.90	
	P_7		withdraw & irregular	cash or post & withhold	B	119	80.4	1.21		2.28	1.33	2.06	1.10	1.71	
	P_8		withdraw	cash or post & withhold	B	643	61.2	1.07		1.73	1.19	1.57	1.10	1.46	
	P_9		withdraw & vol. deduct	withdraw & direct debit	B	237	60.6	0.97		1.72	1.07	1.55	1.10	1.60	
	P_{10}		cash	agent	C	33	60.0	1.12		3.23	1.18	3.07	1.05	2.74	
\mathcal{P}_2	P_{11}	age:65+	withdraw	cash or post	A	1980	93.3	0.86	0.59	2.02	1.06	1.63	1.24	1.90	
	P_{12}		irregular	cash or post	A	462	88.7	0.87		1.92	1.08	1.55	1.24	1.79	
	P_{13}		withdraw & irregular	cash or post	A	132	85.7	0.96		1.86	1.18	1.50	1.24	1.57	
	P_{14}		withdraw & irregular	withdraw	C	50	63.3	2.91		3.40	2.47	4.01	0.85	1.38	



An Example of Extended Combined Pattern Cluster

$$\left\{ \begin{array}{l} PLN \rightarrow T \\ PLN, DOC \rightarrow T \\ PLN, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC, REA \rightarrow T \\ PLN, DOC, DOC, DOC, REA, IES \rightarrow T \end{array} \right.$$

Identifying high impact behavior in behavior evolution



Coupled Hidden Markov Model-based Abnormal Coupled Behavior Analysis

Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Application, *IEEE Trans. Knowledge and Data Engineering*.

Cao, L., Ou Y, Yu PS, Wei G. Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors, *KDD2010*.

Pool manipulation

TABLE 1
An example of buy and sell orders

Investor	Time	Direction	Price	Volume
(1)	09:59:52	Sell	12.0	155
(2)	10:00:35	Buy	11.8	2000
(3)	10:00:56	Buy	11.8	150
(2)	10:01:23	Sell	11.9	200
(1)	10:01:38	Buy	11.8	200
(4)	10:01:47	Buy	11.9	200
(5)	10:02:02	Buy	11.9	250
(2)	10:02:04	Sell	11.9	500

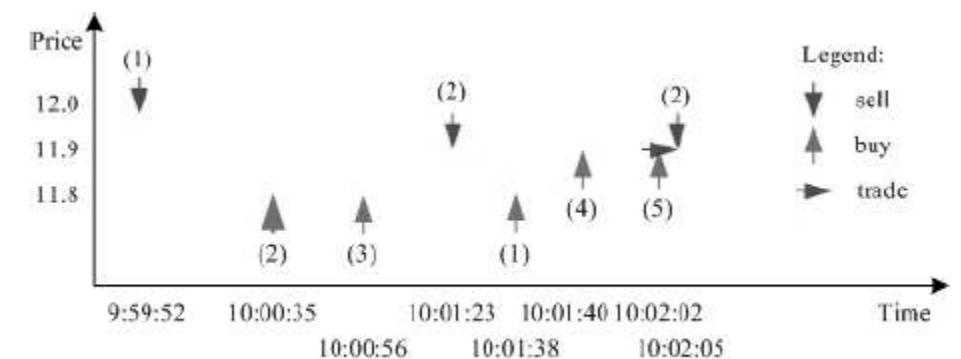


Fig. 1. Coupled Trading Behaviors

Construct behavior sequences

- Behavioral data structure 1:

$$\left\{ \frac{\text{Actor}_i - \text{Operation}_i}{\text{Attributes}_i} \xrightarrow{\eta} \frac{\text{Actor}_j - \text{Operation}_j}{\text{Attributes}_j} \right\}_{i,j=1; \text{win size}}^{I,J} \quad (12)$$

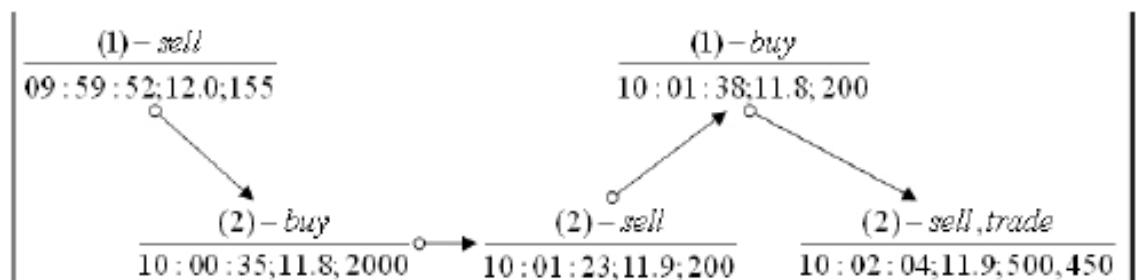


Fig. 2. Behavior sequences - Data Structure 1

- Data structure 2:

$$Category : \left\{ \frac{Actor_i - Operation_i}{Attributes_i} \xrightarrow{\eta} \frac{Actor_j - Operation_j}{Attributes_j} \right\}_{i,j=1; winsize}^{I,J} \quad (14)$$

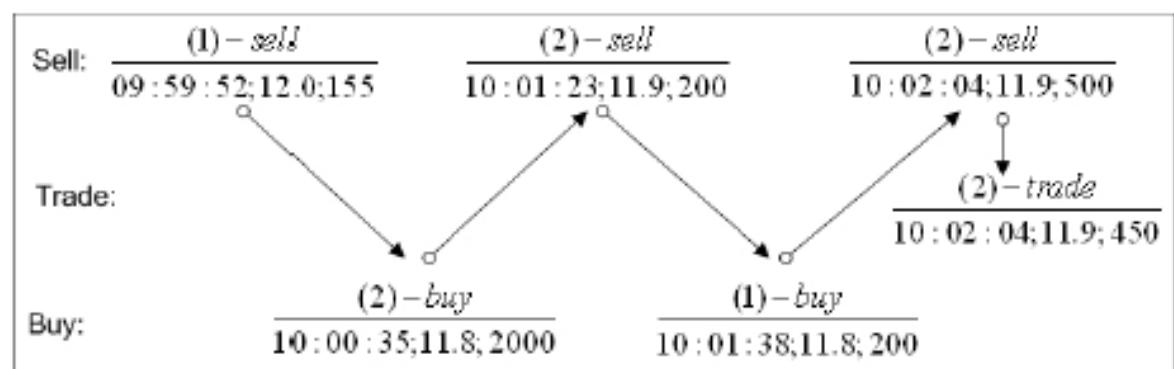


Fig. 3. Behavior sequences - Data Structure 2

CHMM Based Coupled Sequence Modeling

- Coupled behavior sequences

- Multiple sequences

$$\Phi_1 = \{\phi_{11}, \dots, \phi_{1T}\}$$

$$\Phi_2 = \{\phi_{21}, \dots, \phi_{2F}\}$$

$$\Phi_C = \{\phi_{C1}, \dots, \phi_{CG}\}$$

- Coupling relationship

$$R_{ij}(\Phi_i, \Phi_j)$$

$$R_{ij} \subset R, R_{ij}(\Phi_i, \Phi_j) = \emptyset$$

- Behavior properties

$$\phi_{ik}(p_{ik,1}, \dots, p_{ik,L})$$

CBA - CHMM

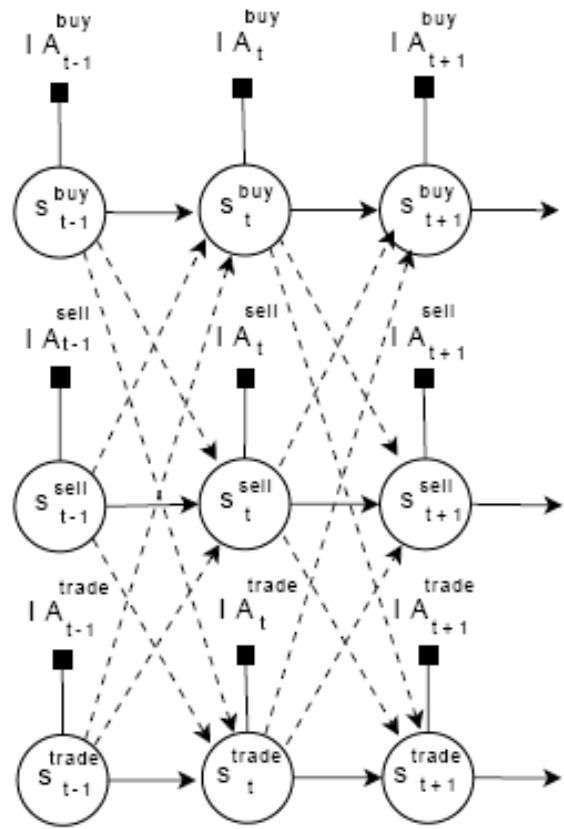


Figure 2: Architecture of CHMM

CBA problem \rightarrow CHMM model (15)

$\Phi(\mathbb{B}_c)|category \rightarrow X$ (16)

$M(\Phi(\mathbb{B}_c))|\phi_{ik}([p_{ij}]_1, \dots, [p_{ij}]_K) \rightarrow Y$ (17)

$f(\theta(\cdot), \eta(\cdot)) \rightarrow Z$ (18)

Initial distribution of $\Phi(\mathbb{B}_c)|category \rightarrow \pi$ (19)

Framework: abnormal CBA

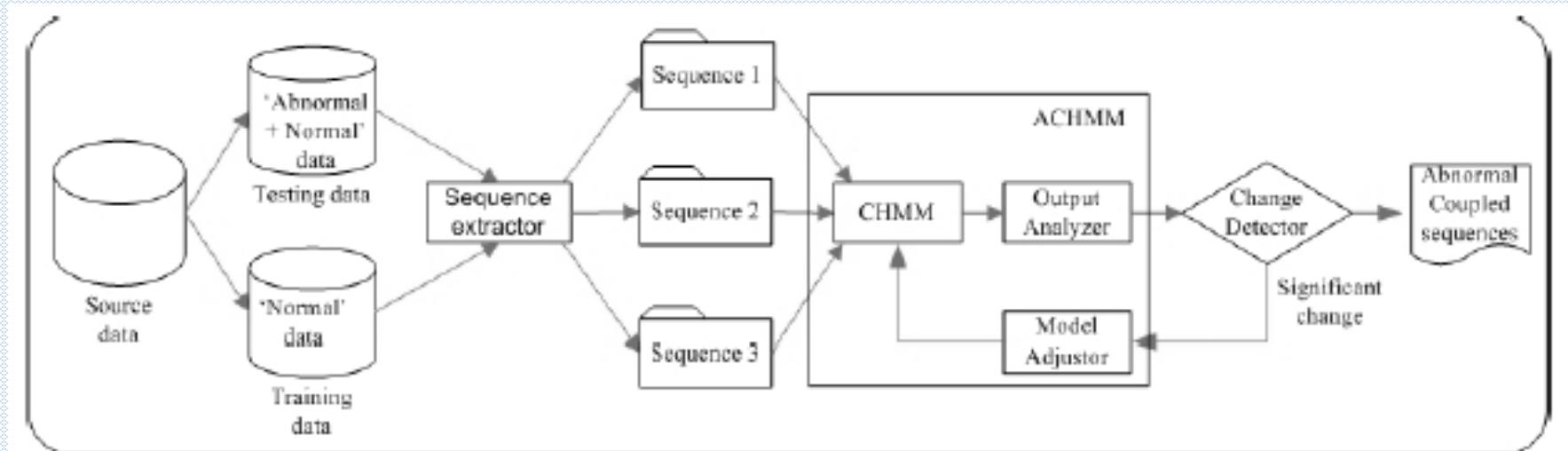


Fig. 5. Framework of abnormal coupled behavior detection

Hidden States

$$S^{buy} = \{Positive\ Buy, Neutral\ Buy, Negative\ Buy\}$$
$$S^{sell} = \{Positive\ Sell, Neutral\ Sell, Negative\ Sell\}$$
$$S^{trade} = \{Market\ Up, Market\ Down\}$$

Observation Sequences

Activity (A)

$$A = \{a_1, a_2, \dots, \}$$

$$a_i = (a(t_i), p(t_i), v(t_i))$$

$$\begin{aligned} a(t_i) &= \{buy \mid sell \mid trade \\ &\quad trade\ price\} \end{aligned}$$

$$p(t_i) = \{buy\ price \mid sell\ price\}$$

trade volume}

$$v(t_i) = \{buy\ volume \mid sell\ volume\}$$

Interval Activity (IA)

$$\mathcal{A} = \{A_1, A_2, \dots, A_n\}$$

$$A_i(a) = A_j(a)$$

$$\bar{p} = \frac{\sum_{i=1}^n p_i}{f} \quad f = |\mathcal{A}| = n \quad \bar{v} = \frac{\sum_{i=1}^n v_i}{f}$$

$$IA(\mathcal{A}, \bar{p}, \bar{v}, f) \xrightarrow{\text{quantization}} IA'(\bar{p}', \bar{v}', f')$$

- CHMM model:

$$\lambda^{CHMM} = (X, Y, Z, \pi)$$

$$\bar{x}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) x_{ij} y_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i) \beta_t(j)}, 1 \leq i, j \leq N \quad (20)$$

$$\bar{y}_j(k) = \frac{\sum_{t=1, o_t=O_k}^T \alpha_t(j) \beta_t(j)}{\sum_{t=1}^T \alpha_t(j) \beta_t(j)}, 1 \leq j \leq N \quad (21)$$

$$\bar{\pi}_i = \frac{\alpha_1(i) \beta_1(i)}{\sum_{j=1}^N \alpha_1(j) \beta_1(j)}, 1 \leq i \leq N \quad (22)$$

$$Pr(Q|\lambda^{HMM}) = \sum_{i=1}^N \alpha_T(i) \quad (23)$$

$$Z = \{z_{ij'}\} \quad z_{ij'} = Pr(s'_{t+1} = S_{j'} | s_t = S_i)$$

Adaptive CHMM for Detecting Sequence Changes



Figure 3: Update Point of ACHMM

$$x_{ij}^{update} = (1 - w)x_{ij}^{old} + w * x_{ij}^{new} \quad (15)$$

$$y_{ij}^{update} = (1 - w)y_{ij}^{old} + w * y_{ij}^{new} \quad (16)$$

$$z_{ij'}^{update} = (1 - w)z_{ij'}^{old} + w * z_{ij'}^{new} \quad (17)$$

$$\pi_i^{update} = (1 - w)\pi_i^{old} + w * \pi_i^{new} \quad (18)$$

The Algorithms

Algorithm 1 Constructing observation sequences

Step 1: Segment the whole trading day into L intervals by a time window with the length $winsize$.

Step 2: Calculate IA for buy-order, sell-order and trade activities respectively in each window. They are denoted as IA_l^{buy} , IA_l^{sell} and IA_l^{trade} , respectively.

Step 3: Obtain $IA_l'^{buy}$, $IA_l'^{sell}$ and $IA_l'^{trade}$ by quantizing IA_l^{buy} , IA_l^{sell} and IA_l^{trade} .

Step 4: Obtain the trading activity sequence IA^{buy} for buy-order by putting all $IA_l'^{buy}$ in a trading day together. Obtain IA^{sell} and IA^{trade} in the same way. We obtain

$$IA^{type} = IA_1'^{type}, IA_2'^{type}, \dots, IA_L'^{type} \quad (19)$$

where $type \in \{buy, sell, trade\}$. IA^{buy} , IA^{sell} and IA^{trade} are the observation sequences of CHMM in the day.

Step 5: Repeat Step 1-4 for each trading day

Algorithm 2 Detecting abnormal trading sequences

Step 1: Construct trading sequences including training sequences $Seq_1, Seq_2, \dots, Seq_K$ and test sequences $Seq'_1, Seq'_2, \dots, Seq'_{K'}$.

Step 2: Train the ACHMM model on the training sequences;

Step 3: Compute the mean (μ) and standard deviation (σ) of probability of training sequences according to the following formulas:

$$\mu = \frac{\sum_{i=1}^K Pr(Seq_i | ACHMM)}{K} \quad (20)$$

$$\sigma = \sqrt{\frac{1}{K} \sum_{i=1}^K Pr(Seq_i | ACHMM) - \mu} \quad (21)$$

where K is the total number of training sequences, mean μ represents the centroid of model ACHMM, and the standard deviation σ represents the radius of model ACHMM.

Step 4: For each test sequence Seq'_i , calculate its distance D_i to the centroid of model by

$$D_i = \frac{\mu - Pr(Seq'_i | \mathcal{M})}{\sigma} \quad (22)$$

Consequently, Seq'_i is an exceptional pattern, if it satisfies:

$$D_i > \psi_0 \quad (23)$$

where ψ_0 is a given threshold.

- Benchmark Models
 - HMM-B: Buy-based HMM
 - HMM-S: Sell-based HMM
 - HMM-T: Trade-based HMM
 - IHMM: HMM-B + HMM-S + HMM-T
 - CHMM: CHMM(buy, sell, trade)
 - ACHMM: Adaptive CHMM(buy, sell, trade)

Evaluation

- Technical performance

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (43)$$

$$Precision = \frac{TP}{TP + FP} \quad (44)$$

$$Recall = \frac{TP}{TP + FN} \quad (45)$$

$$Specificity = \frac{TN}{FP + TN} \quad (46)$$

- Business performance

$$Return = \ln \frac{p_t}{p_{t-1}} \quad (48)$$

$$Abnormal\ Return = Return - (\gamma + \xi Return^{market}) \quad (49)$$

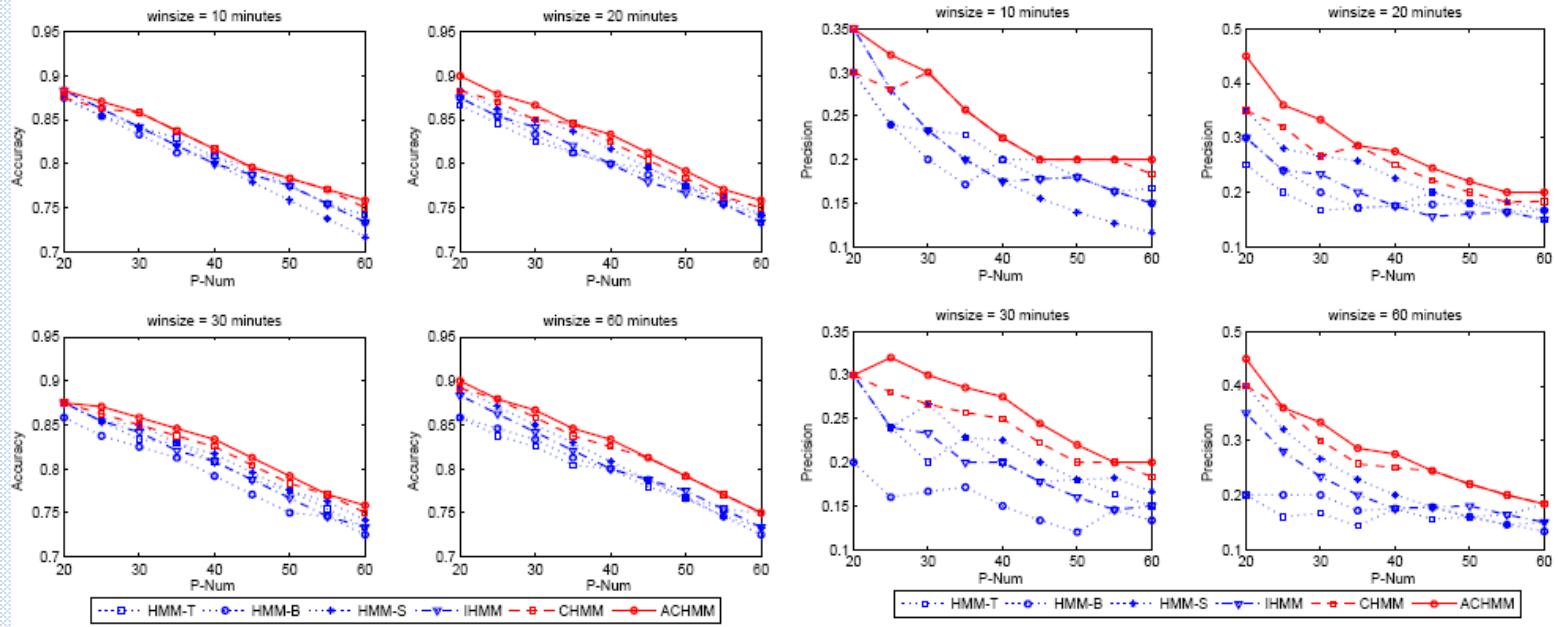


Figure 4: Accuracy of Six Models

Figure 5: Precision of Six Models

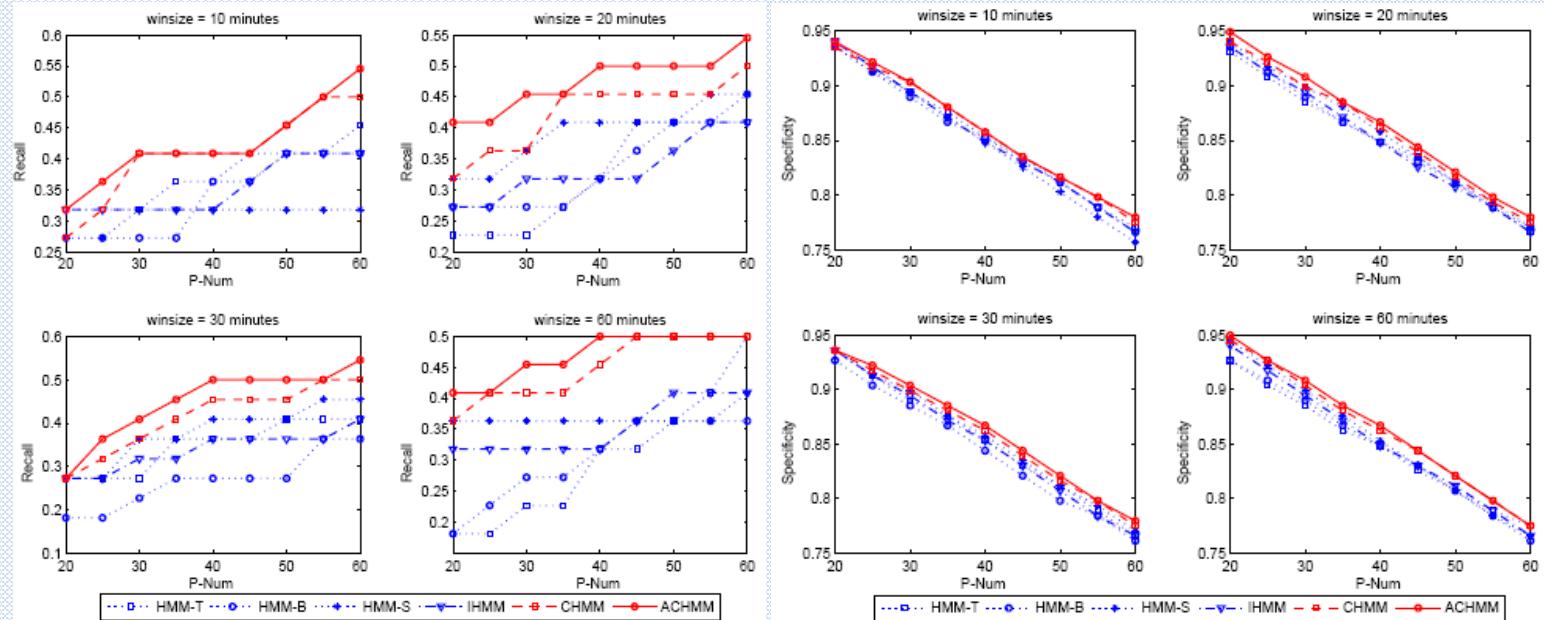


Figure 6: Recall of Six Models

Figure 7: Specificity of Six Models

- Business Performance

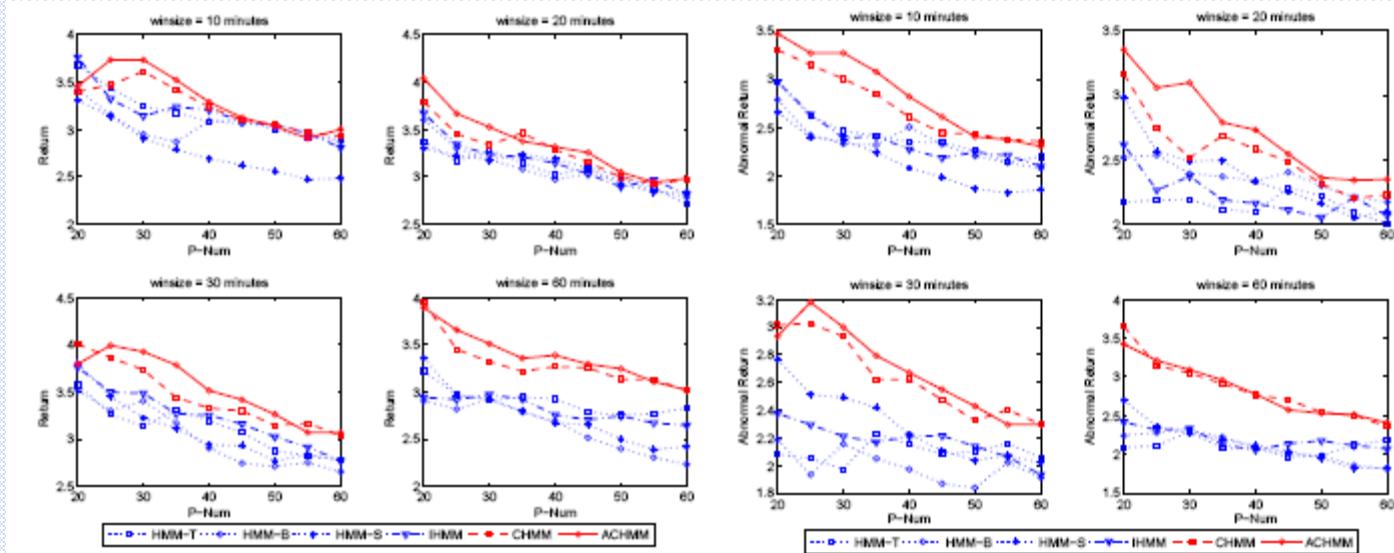


Fig. 9. Return of Six Models

Fig. 10. Abnormal Return of Six Models

- Computational cost

$$(O(TN^6)) \xrightarrow{\text{N-heads dynamic programming}} O(T(3N)^{\tilde{j}})$$

TABLE 5
Computational performance

		IHMM	CHMM	ACHMM
winsize =10 (m)	Training time (s)	0.574	11.978	11.988
	Test time (s)	0.056	1.296	3.576
winsize =20 (m)	Training time (s)	0.256	4.929	4.933
	Test time (s)	0.047	0.655	3.486
winsize =30 (m)	Training time (s)	0.206	4.121	4.119
	Test time (s)	0.042	0.447	2.429
winsize =60 (m)	Training time (s)	0.109	2.003	2.004
	Test time (s)	0.036	0.221	1.206

Hierarchical CHMM-based & Relational Learning-based Group Behavior Learning

- Yin Song and Longbing Cao. Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets, IJCNN 2012.
- Yin Song, Longbing Cao, et al. Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation, KDD 2012.

Framework

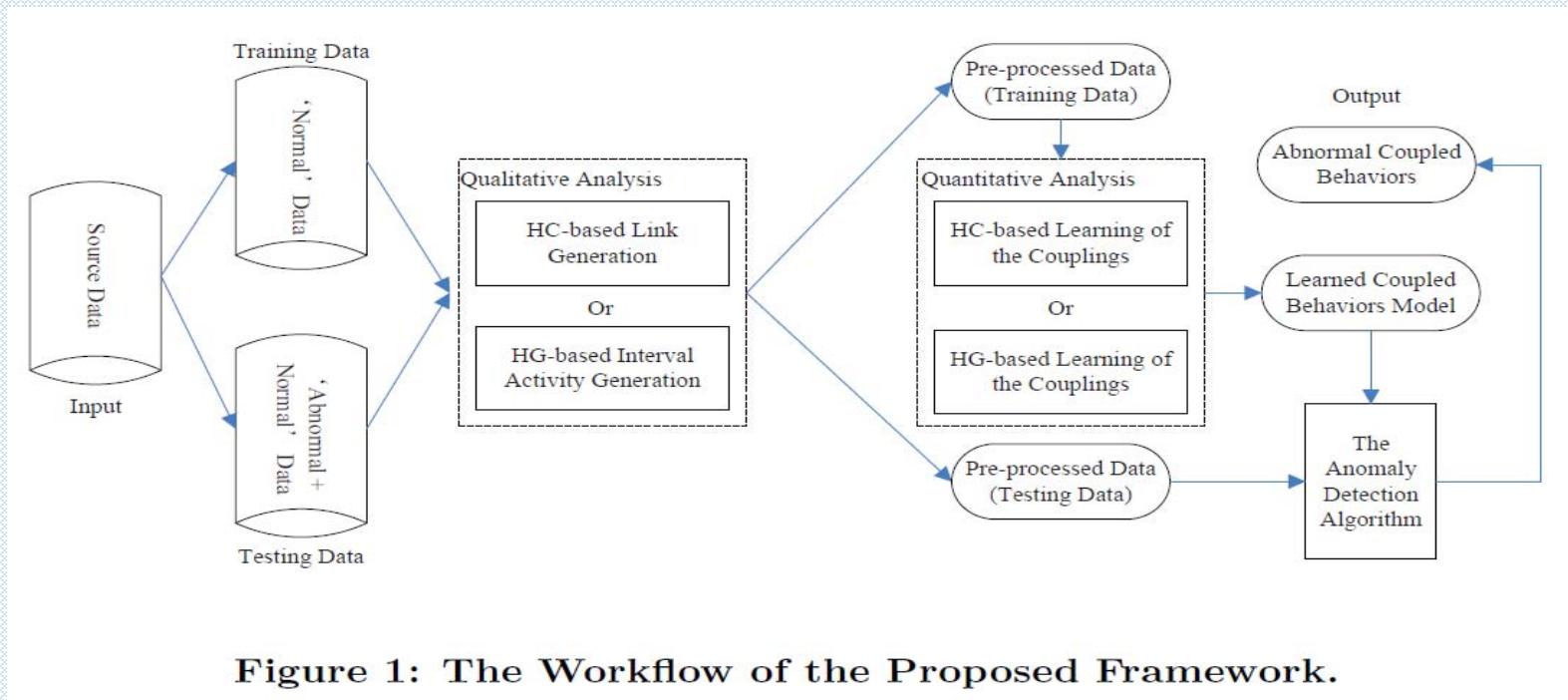


Figure 1: The Workflow of the Proposed Framework.

- *HC: Hybrid coupling; HG: Hierarchical grouping*
- *1st qualitative analysis*, generates possible qualitative coupling relationships between behaviors with or without domain knowledge;
- *2nd quantitative representation* of coupled behaviors is learned via proper methods;
- *3rd anomaly detection* algorithms are proposed to cater for different application scenarios.

Hybrid coupling-based analysis

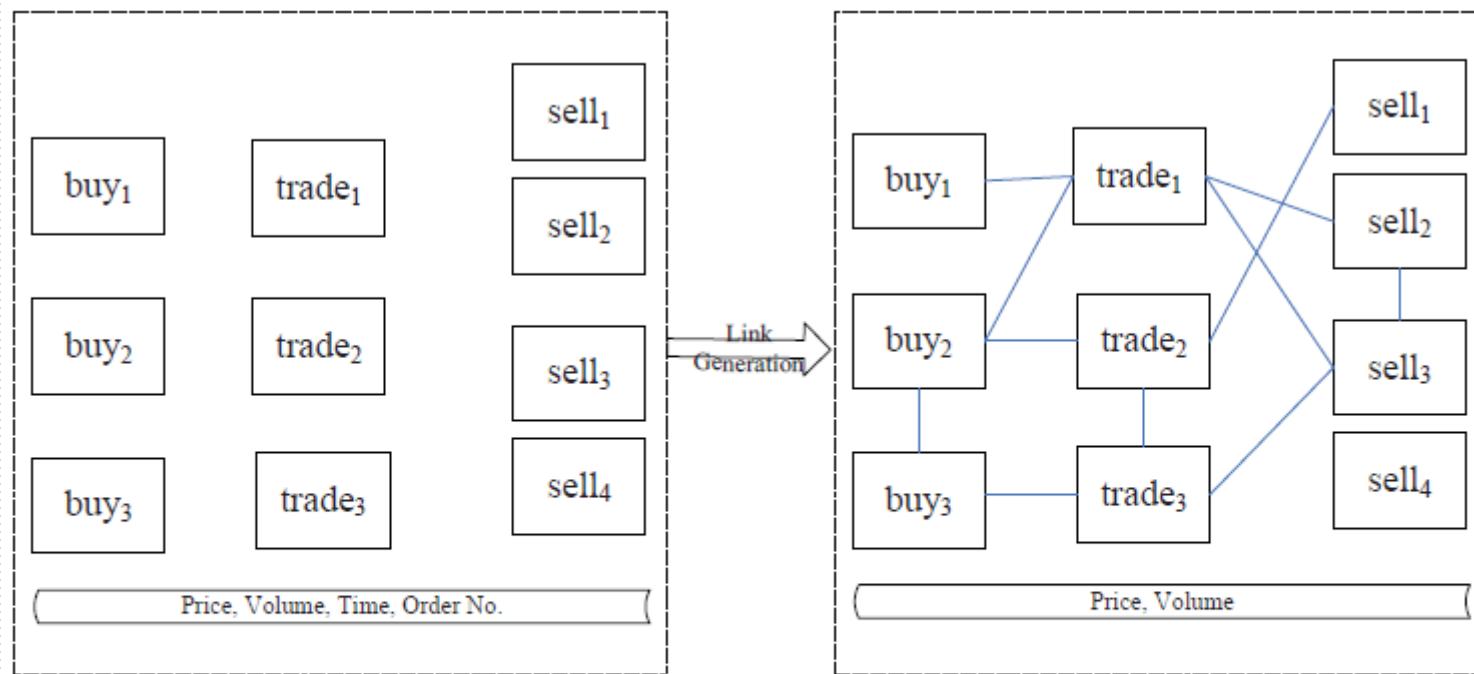
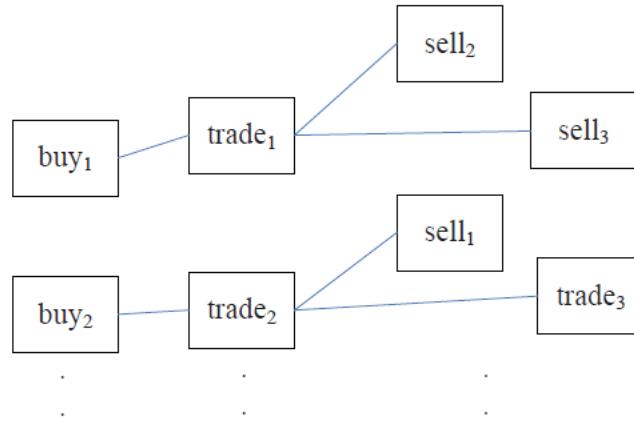


Figure 2: An Example of Qualitative Analysis.

From qualitative to quantitative



(a) An Example of the Subgraphs for Coupled Behaviors

	A	RF_1	RF_2	\dots	RF_n
$trade_1$	x_1	rf_{11}	rf_{21}	\dots	rf_{n1}
$trade_2$	x_2	rf_{12}	rf_{22}	\dots	rf_{n2}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

(b) An Example of the Relational Features for Coupled Behaviors

Figure 3: An Example of the “Flattened” Propositional Coupled Behavioral Data

Hierarchical Grouping-based analysis

- Qualitative Analysis: Domain Knowledge driven Initial Grouping

DEFINITION 4 (PARTICLE GROUPS). *The particle groups, which are represented by $\{PG_j\}$ ($1 \leq j \leq N$) are the partitioning result of actors $\{A_i\}$ ($1 \leq i \leq I$) by the rule $R(\cdot)$ made by domain experts:*

$$R(\cdot) | \{A_1, A_2, \dots, A_I\} \rightarrow \{PG_1, PG_2, \dots, PG_N\}. \quad (3)$$

- For each group, both corresponding CHMM
- The similarity between two CHMMs (two groups)

- Symmetric distance between the coupled behaviors of two particle groups for the given CHMM

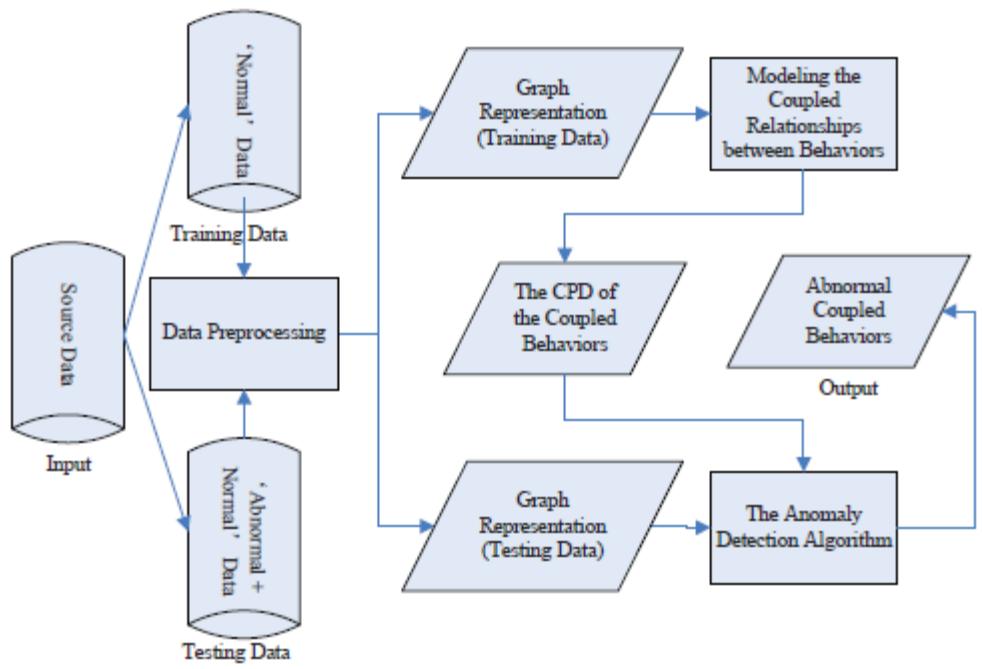


Figure 2: The Work Flow of the Proposed Framework.

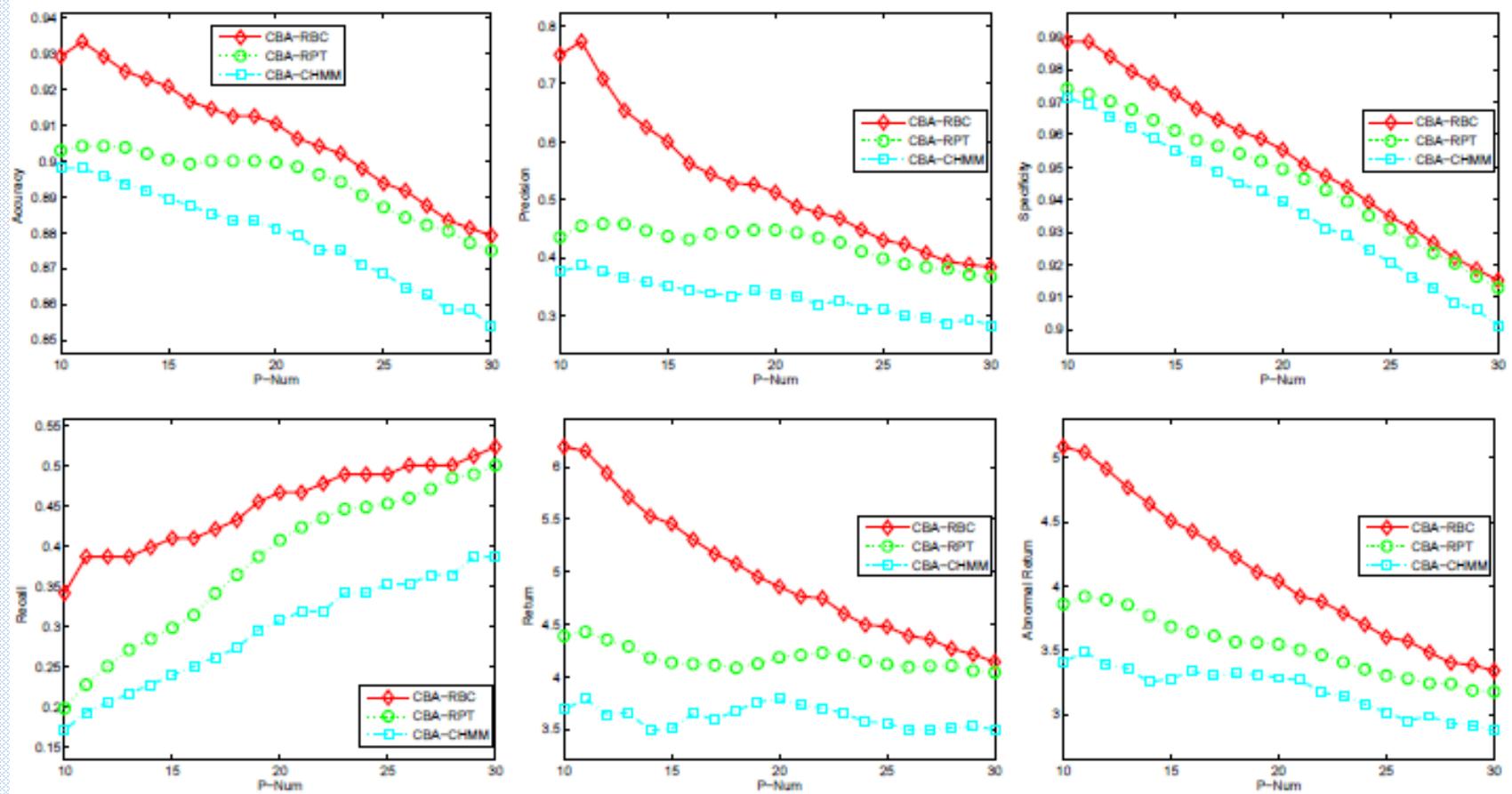
- Relational features for behaviors

$$p(RF_1|X^{(t)}) \quad p(RF_2|X^{(t)}) \cdots, p(RF_n|X^{(t)})$$

- Conditional probability distribution

$$\text{CPD } p(X^{(t)}|RF_1, \dots, RF_n)$$

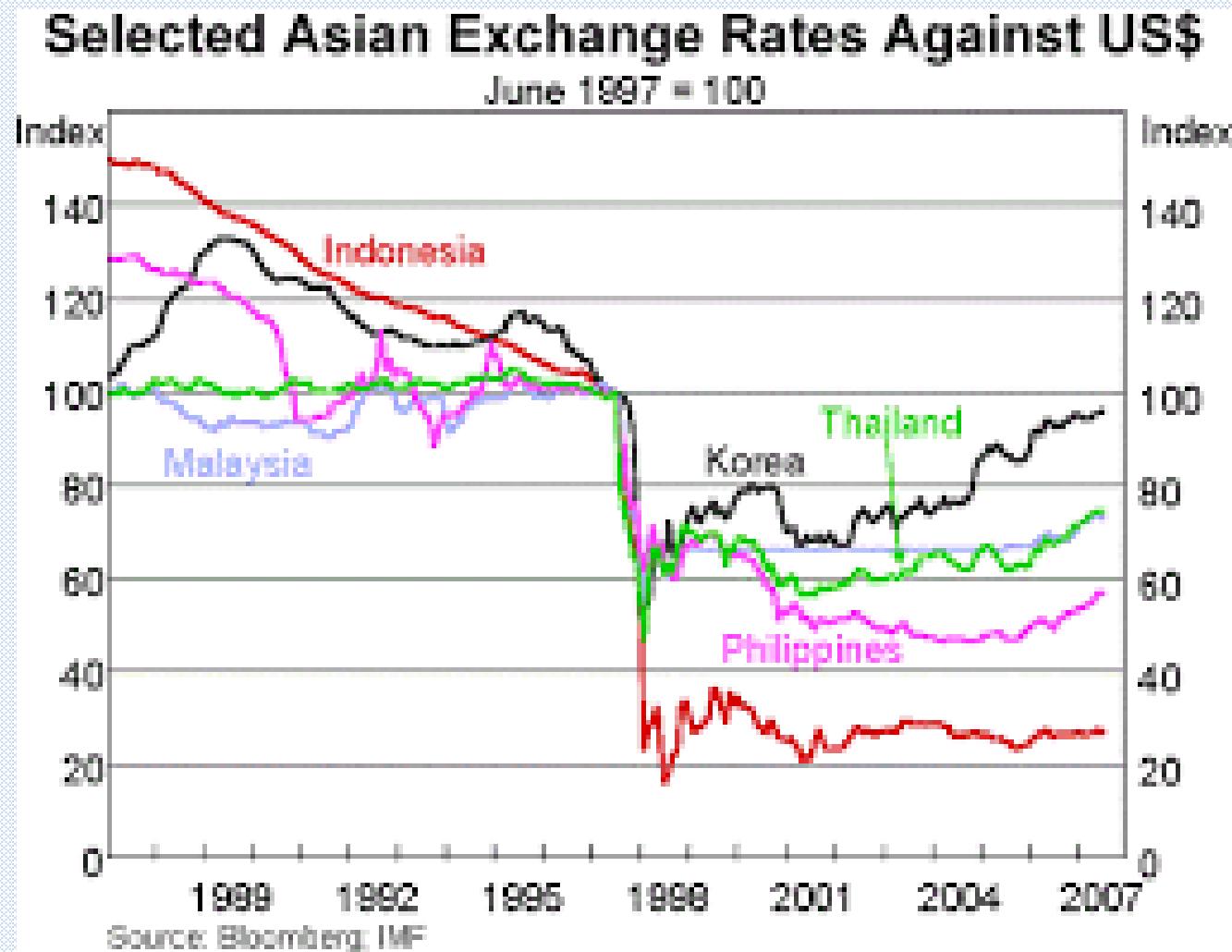
- Relational Bayesian Classifier (RBC)
- Conditional Probability Tree (CPT)



Cross-market Behavior Analysis for Financial Crisis Contagion

- Wei Cao, Liang Hu, Longbing Cao: Deep Modeling Complex Couplings within Financial Markets. AAAI 2015: 2518-2524
- Wei Cao, Longbing Cao. Financial Crisis Forecasting via Coupled Market State Analysis, IEEE Intelligent Systems, 30(2): 18-25 (2015).

Example: 1998 financial crisis



Motivation: gaps

- Most of the existing literature simply tests the existence of market contagion.
- Multiple markets are affected by financial crisis.
- Limited research pays attention to the unknown underlying market couplings which are the ``fundamental'' reasons for the market contagion.

Motivation: challenges

- To properly understand financial crisis:
 - (1) **Selection** of global markets and discriminative market indicators;
 - (2) **Market couplings** are very complicated to capture, which includes intra-market coupling (couplings within a market) and inter-market coupling (couplings between different types of markets);
 - (3) **Hidden couplings** behind the market indicators need to be captured
 - (4) **Measurement** of effect/relationship of market couplings on understanding crisis, namely how the couplings contagions reflect the crisis.

Solution: Design

- **Model cross-market couplings** via building a CHMM-LR framework:
 - (1) couplings from Equity market and Commodity market ($C(E,C)$);
 - (2) couplings from Equity market and Interest market ($C(E,I)$);
 - (3) couplings from Commodity market and Interest market ($C(C,I)$).
- **Estimate global crisis** by the crisis forecasting capability of integrating the different pairwise market couplings

Technical method: CHMM

Market 1:

Market 2:

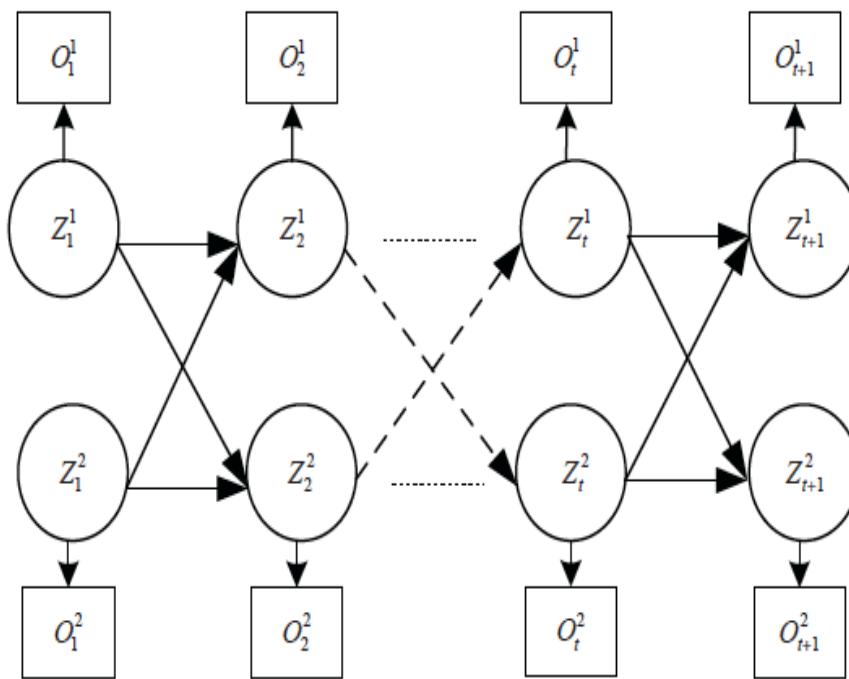


Fig. 1: A CHMM with Two Chains

$$CHMM = \lambda(A, B, R, \pi)$$

Technical method: LR

- The probability of financial crisis at time t

$$P_t = P(Y_t = 1 | X = x) = E(Y_t | X = x) = \frac{1}{1 + e^{-b_0 + b_1 x_1 + \dots + b_n x_n + \varepsilon}} \quad (1)$$

- The likelihood of financial crisis at time period T

$$L(\theta) = \prod_{t=1}^T P_t^{Y_t} (1 - P_t)^{1-Y_t} \quad (2)$$

Modeling cross-market couplings

Definition 1. Intra-market Coupling. This is the interaction between the behaviors from the same market. Formally, the representation of intra-market coupling w.r.t market i is given by:

$$\theta_i = \{m_i \otimes m_i\}_{t=1}^T \quad (3)$$

where m_i denotes the observations from market i , \otimes represents the coupled interactions among market i 's observations from time 1 to T . In this paper, there are three global financial markets, so $i \in \{E, C, I\}$.

Definition 2. Inter-market Coupling. This is the interaction between the behaviors from pairwise markets. Formally, the representation of inter-market coupling w.r.t market i and j is given by:

$$\eta_i = \{m_i \circledast m_j\}_{t=1}^T \quad (4)$$

where \circledast represents the coupled interactions between market i 's observations and market j 's observations from time 1 to T , $i, j \in \{E, C, I\}$.

Definition 3. Market Coupling. The representation of market coupling w.r.t market i and j is given by:

$$\mathbb{C}(i, j) = \{\theta_i, \eta_{ij}\} \quad (5)$$

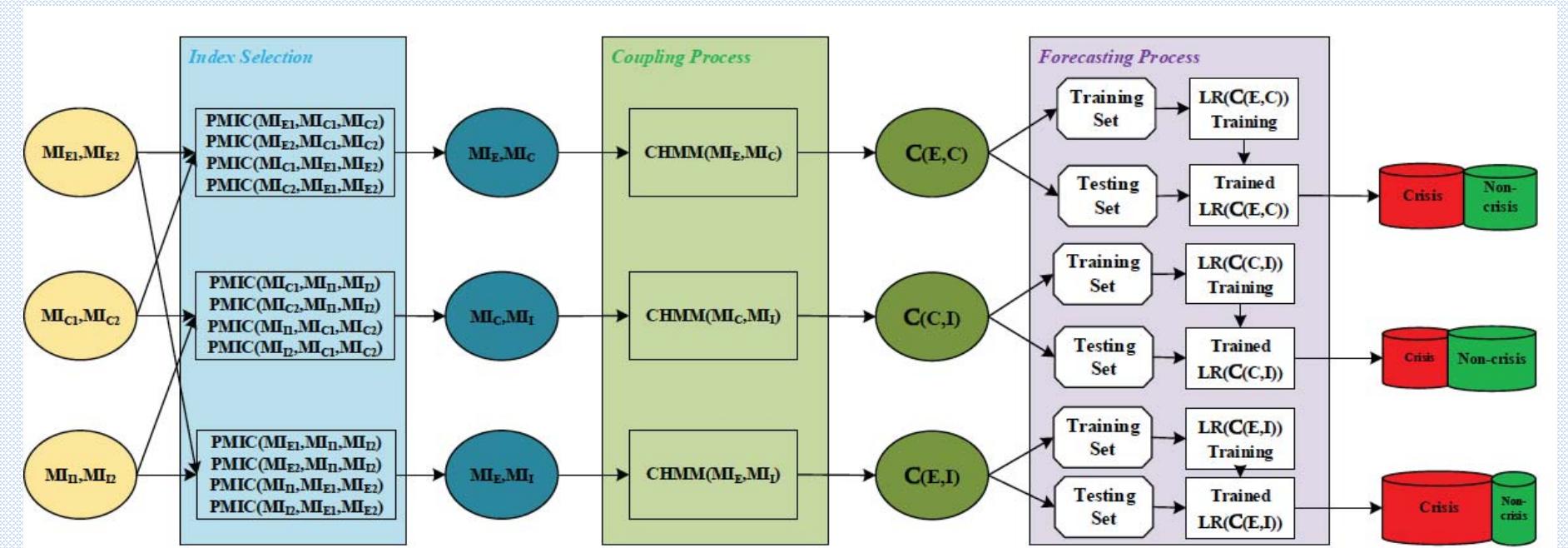
where θ_i denotes the intra-market coupling in market i , and η_{ij} represents the inter-market coupling between markets i and j .

Objective function

$$\operatorname{argmax}_{(i,j)} R(\text{crisis}, \mathbb{C}(i, j))$$

The identification (estimation) of financial crisis is to build a proper model to determine the specific pairwised cross-market couplings $\mathbb{C}(i, j)$ and the corresponding objective function $R(\cdot)$.

Modeling Framework



Indicator selection

Definition 4. Pairwise Market Indicator Correlation (PMIC).

This is the correlation of one indicator in a market (MI_{ik}) with indicators in another market ($\{MI_{jl}\}$), where $(i \neq j) \wedge (i, j \in \{E, C, I\}) \wedge (k, l \in \{1, 2\})$.

Detrended cross-correlation coefficient

$$PMIC(MI_{ik}, \{MI_{jl}\}) = \sum_l | \rho_{DCCA}(MI_{ik}, MI_{jl}) | \quad (7)$$

where $\rho_{DCCA}(\cdot)$ is the cross-correlation coefficient of the two market indicators.

$$\operatorname{argmax}_k PMIC(MI_{ik}, \{MI_{jl}\})$$

Coupling process

Pairwise Market Couplings \rightarrow CHMM modeling

$$\Phi(MI^i) | \text{observation} \rightarrow B(P(o_t^i = X_v | z_t^i = Z_h)) \quad (9)$$

$$\begin{aligned} \Phi(MI^i) | \text{intra-transition}(\theta_i) \rightarrow \\ A | \text{intra}(P(z_{t+1}^i = Z_{h'} | z_t^i = Z_h)) \end{aligned} \quad (10)$$

$$\begin{aligned} \Phi(MI^i), \Phi(MI^j) | \text{inter-transition}(\eta_{ij}) \rightarrow \\ A | \text{inter}(P(z_{t+1}^i = Z_{h'} | z_t^j = Z_h)) \end{aligned} \quad (11)$$

$$\mathbb{C}(i, j) \rightarrow \{z^i, z^j\} \quad (12)$$

Forecasting process

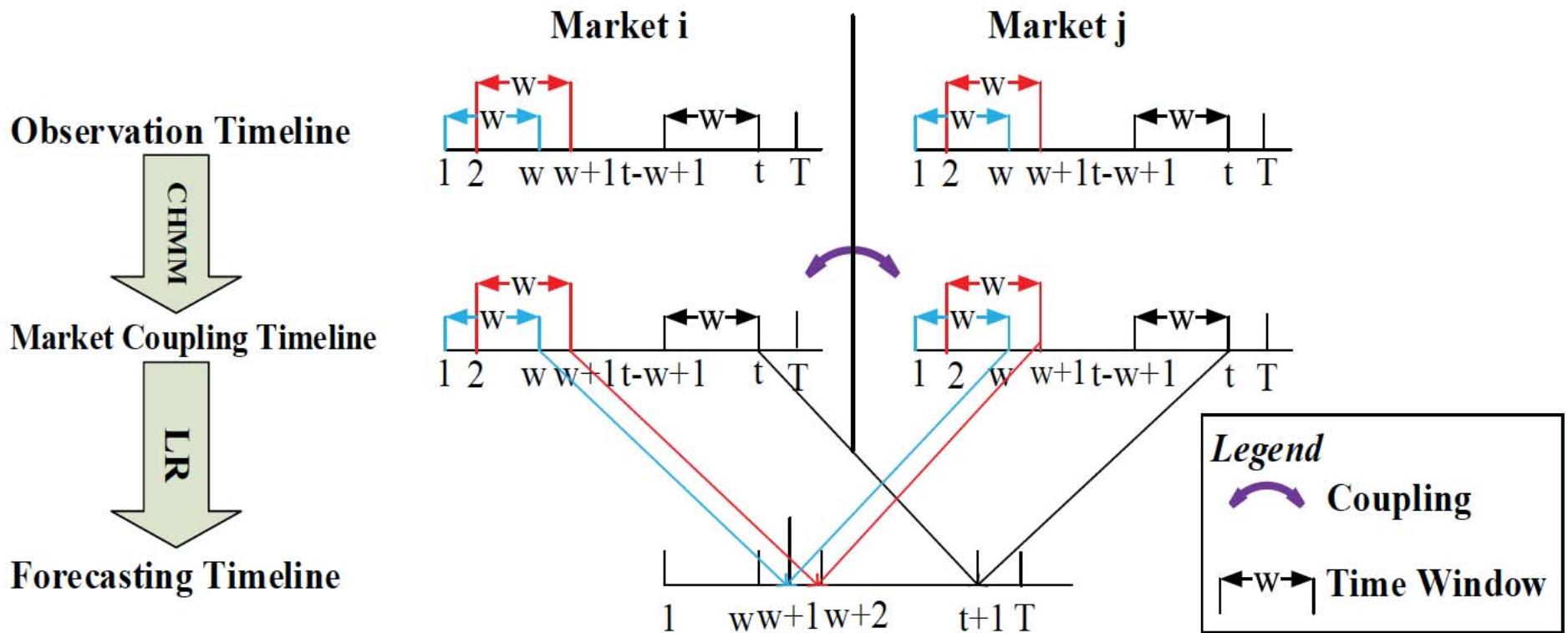


Fig. 3: Forecasting Process

Forecasting process

Algorithm 1: Financial Crisis Forecasting via Market Couplings

Input: A training set Tr ; A testing set $Te = \{\{z_1^i, z_1^j\}, \{z_2^i, z_2^j\}, \dots, \{z_t^i, z_t^j\}, \dots, \{z_T^i, z_T^j\}\}$

Output: A predicted financial crisis set CS ; A predicted non-financial crisis set NS

- 1 Train the Logistic Regression model Ω on the training set Tr , obtained trained model Ω^{Tr} ;
 - 2 **forall the** $\{z_\tau^i, z_\tau^j\}_{\tau=t-w+1}^t$ and $t \in [w, T]$ *in the Testing set* **do**
 - 3 Compute the probability of crisis given the trained model Ω^{Tr} and couplings $\{z^i, z^j\}_{\tau=t-w+1}^t$):
 $P_{t+1}(\text{crisis} = 1 | \{z_\tau^i, z_\tau^j\});$
 - 4 **if** $P_{t+1}(\text{crisis} = 1 | \{z_\tau^i, z_\tau^j\}) > 0.5$ **then**
 - 5 time $t + 1 \rightarrow CS^{(i,j)}$;
 - 6 **else**
 - 7 time $t + 1 \rightarrow NS^{(i,j)}$;
 - 8 **end**
 - 9 **end**
 - 10 **end**
-

Experiments: Data

- Data:
 - Markets: Equity market, Commodity market and Interest market.
 - Time period: January 1990 to December 2010.

TABLE II: Selected Indicators

Pairwise Coupling	Market Indicator
$C(E, C)$	E : DJIA / C : WTI Oil Price
$C(C, I)$	C : Gold Price / C : TED Spread
$C(E, I)$	E : DJIA / I : BAA Spread

Cross-market indicator trends

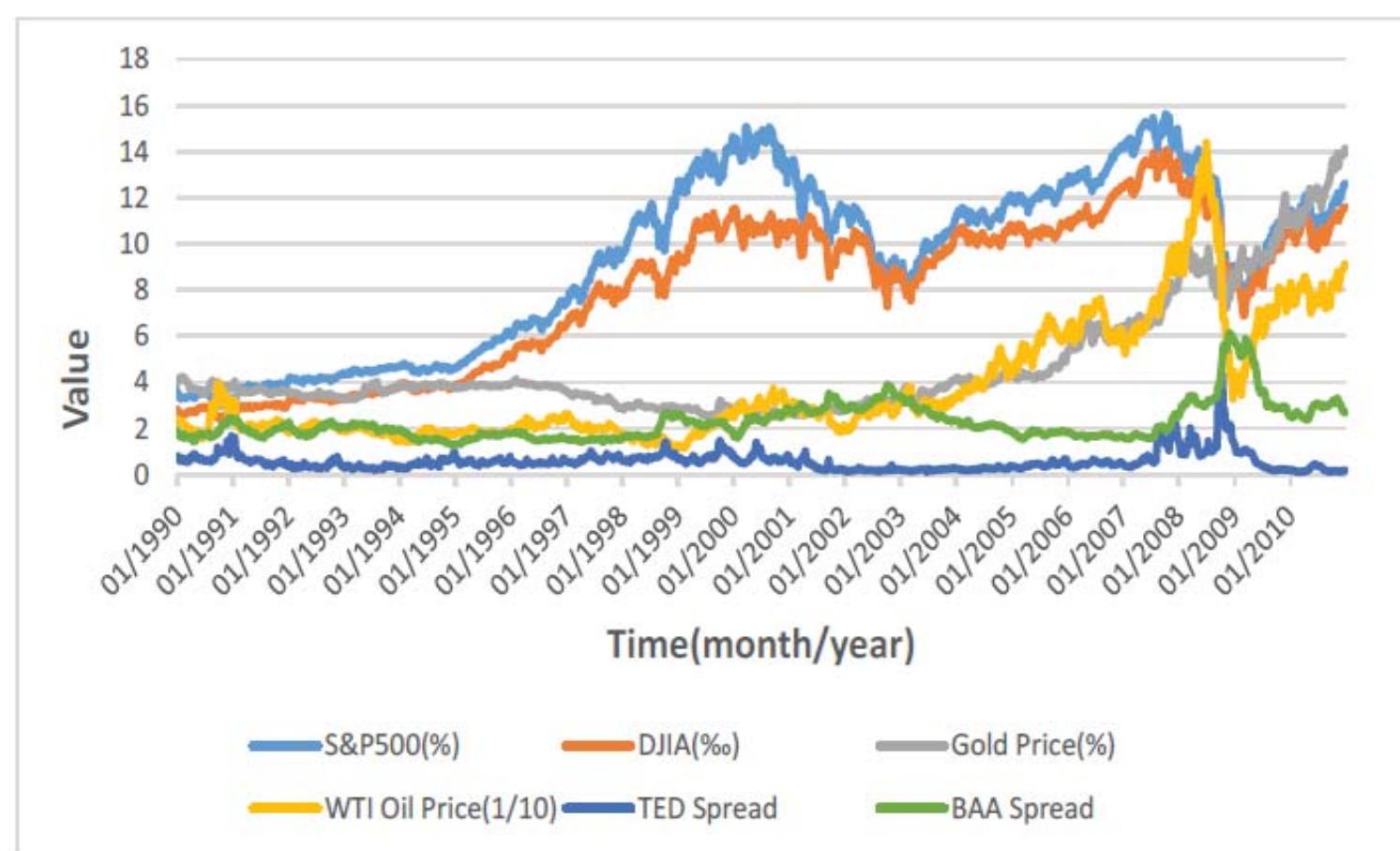


Fig. 4: Indicator behavior During the Period 1990-2010

Evaluation Models

Comparative Methods:

A. IID models

LR-(E, C): This model forecasts crisis based on selected indicators of Equity market and Commodity market directly, without considering the hidden complex market couplings.

LR-(C, I): This model forecasts crisis based on selected indicators of Commodity market and Interest market directly, without considering the hidden complex market couplings.

LR-(E, I): This model forecasts crisis based on selected indicators of Equity market and Interest market directly, without considering the hidden complex market couplings.

B: Non-IID models

LR- $\mathbb{C}(E, C)$: This model forecasts crisis based on market couplings from Equity market and Commodity market.

LR- $\mathbb{C}(C, I)$: This model forecasts crisis based on market couplings from Commodity market and Interest market.

LR- $\mathbb{C}(E, I)$: This model forecasts crisis based on market couplings from Equity market and Interest market.

Experiments: Results

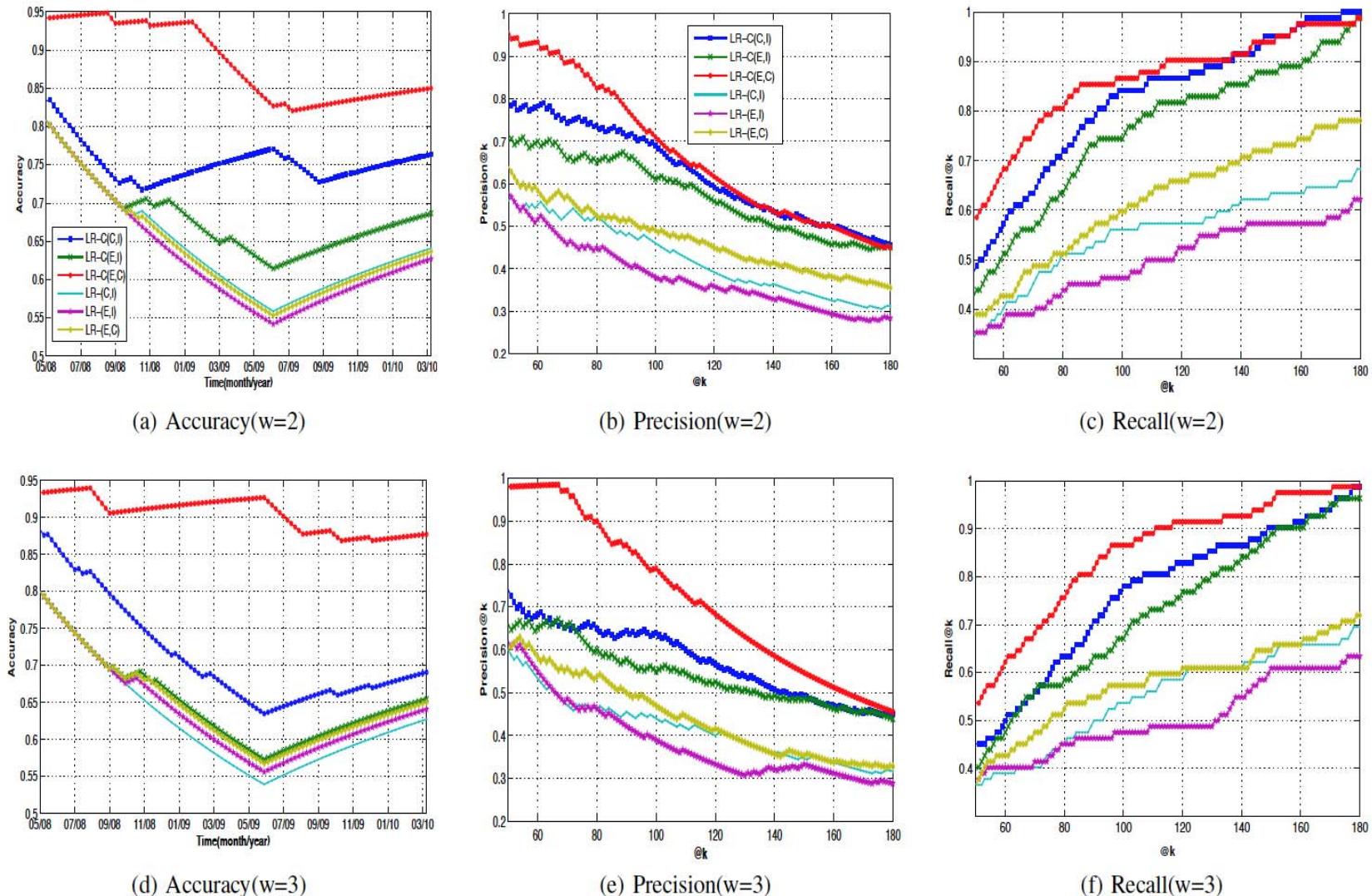


Fig. 7: Technical Performance of Various Approaches

Experiments: Explanation

- These illustrate that the pairwise market couplings have higher relations with financial crisis when compared with those simple indicators. The reasons can be interpreted as following:
 - 1) the pairwise couplings is the ``essence'' of market contagion, which means that the pairwise couplings can better reflect the financial crisis;
 - 2) two different types of couplings (intra-market couplings and inter-market couplings) which can represent the pairwise market couplings well;
 - 3) the CHMM is demonstrated as a useful tool to capture the complex hidden couplings between pairwise markets.

Experiments: Explanation

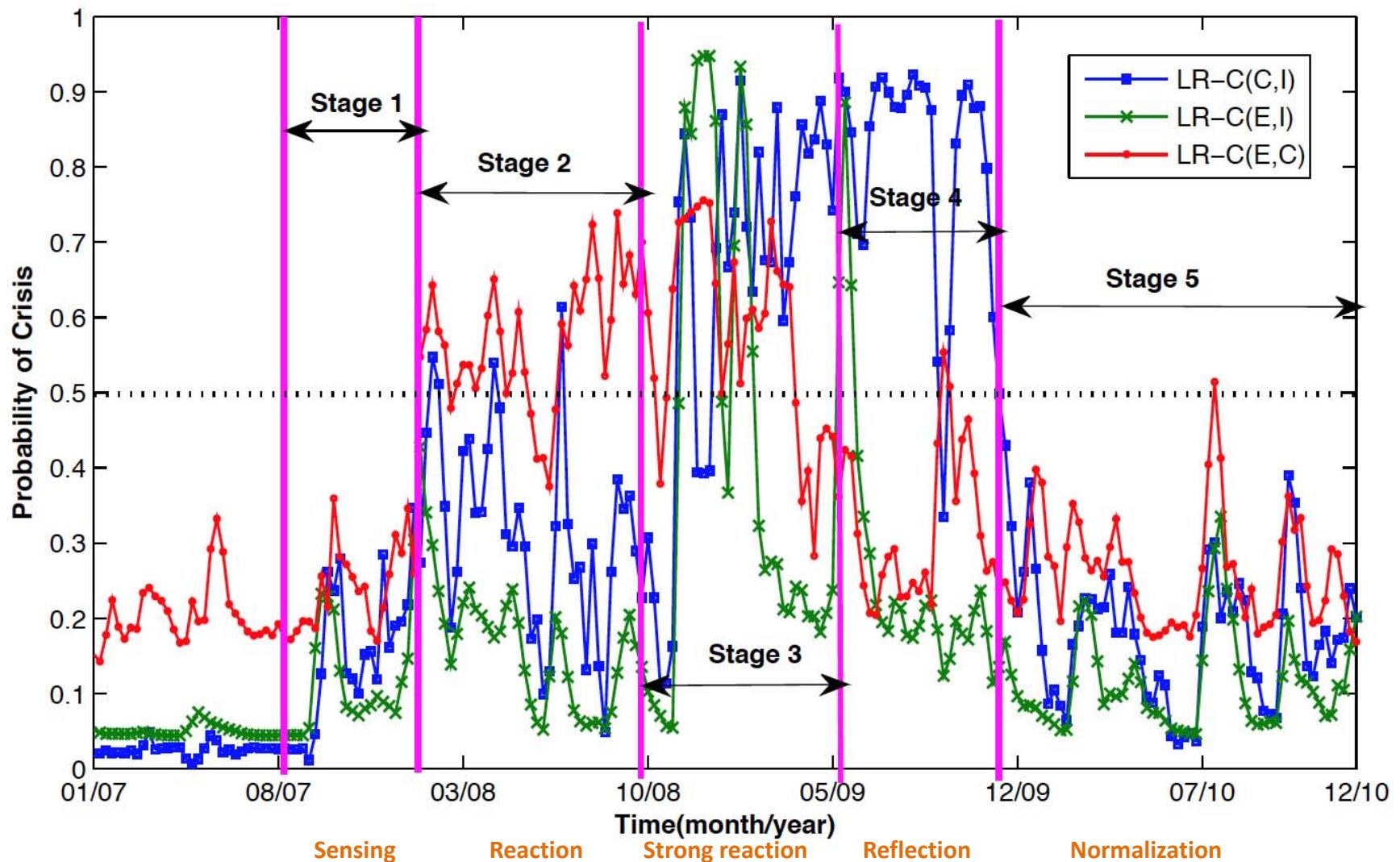


Fig. 8: Market Couplings Behavior during 2008 Financial Crisis ($w = 2$)

Explanation: Crisis Evolution

- Stage 1 is a stage of ``crisis launch'' and spans from August 2007 to December 2007, in this period the probabilities of crisis forecasted by all three pairwise couplings begin to grow.
- Stage 2 is defined as $\mathbf{C}(E,C)$ stage, where the couplings from Equity market and Commodity market has a sharp increase in this stage (December 2007 to September 2008). A possible explanation is that crisis always first revealed by Equity market and Commodity market, the Equity market is always considered as risky market while the Commodity market is the opposite.
- Stage 3 is described as ``sharp fluctuation'' stage, where the all pairwise market couplings reveal high financial crisis probabilities (September 2008 to April 2009). This maybe caused by the spread news of crisis and shifts in investors' common but changing appetite of risk.
- Stage 4 is a $\mathbf{C}(C,I)$ stage spans from April 2009 to November 2009. An explanation is at this stage the macro-control measures (e.g. cutting rate) begin to take effect.
- Stage 5 is described as ``post-crisis'' while the behaviors from all the pairwise couplings become stable (after November 2009).

Conclusions

- Cross-market coupling learning:
 - Equity market, Commodity market and Interest market
- CHMM-LR for financial crisis analysis:
 - CHMM captures the complex hidden pairwise market couplings,
 - LR is applied to evaluate the crisis forecasting capability based on the couplings

Part V.

High Utility Behavior Analysis

Junfu Yin, Zhigang Zheng, Longbing Cao. [USpan: An Efficient Algorithm for Mining High Utility Sequential Patterns](#), KDD 2012, 660-668.

Learning Objectives

- Why care about behavior utility?
- How to measure behavior utility?
- How to identify high utility behavior?

Introduction

- **Sequential pattern mining**
 - Very essential for handling order-based critical business problems.
 - Interesting and significant sequential patterns are generally selected by frequency.
- **Insufficient of frequency/support framework**
 - They do not show the business value and impact.
 - Some truly interesting sequences may be filtered because of their low frequencies.

Example: Retail business

Introduction

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >

In sequence s_2 , there are three transactions:

$[(a, 2)(e, 6)]$,
 $[(a, 1)(b, 1)(c, 2)]$ and
 $[(a, 2)(d, 3)(e, 3)]$.

Transaction $[(a, 2)(e, 6)]$ means the customer buys two items, namely a and e . $(a, 2)$ means the quantity of item a is 2.

The square brackets omitted when there is only one item in the transaction. For example: $(e, 5)$, $(b, 2)$ in s_1 and $(c, 1)$ in s_3 .

Introduction

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

The utility of $\langle e \rangle$ in $(e, 6)$ is $6 \times 1 = 6$

The utility of $\langle ea \rangle$ in s_2 is

$$\begin{aligned} & \{ ((6 \times 1) + (1 \times 2)), ((6 \times 1) + (1 \times 2)) \} \\ & = \{8, 10\} \end{aligned}$$

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence
1	$\langle (e, 5) [(c, 2)(f, 1)] (b, 2) \rangle$
2	$\langle [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] \rangle$
3	$\langle (c, 1) [(a, 6)(d, 3)(e, 2)] \rangle$
4	$\langle [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] \rangle$
5	$\langle [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] \rangle$

The utility of $\langle ea \rangle$ is the database is
 $\{\}, \{8, 10\}, \{\}, \{16, 10\}, \{15, 7\}\}.$

Add the highest utility in each sequence to represent the utility of $\langle ea \rangle$:

$$10 + 16 + 15 = 41$$

If the minimum utility threshold $\xi = 40$ then $\langle ea \rangle$ is a high utility pattern.

Introduction

Contributions:

1. We define the problem of mining high utility sequential patterns systematically.
2. USpan as a novel algorithm for mining high utility sequential patterns.
3. Two pruning strategies, namely width and depth pruning, are proposed to reduce the search space substantially.

Related Work

- **High utility pattern mining**
 - Two-Phase Algorithm (Liu et al., UBDM' 2005)
 - IHUP Algorithm (Ahmed et al., IEEE Trans. TKDE' 2009)
 - UP-Growth (Tseng et al., SIGKDD' 2010)
- **High utility sequential pattern mining**
 - UMSP (Shie et al., DASFAA' 2011) Designed for mining high utility mobile sequential patterns.
 - UWAS-tree / IUWAS-tree (Ahmed et al., SNPD' 2010) Designed for mining the high utility weblog data. IUWAS-tree is for incremental environment.
 - UI / US (Ahmed et al., ETRI Journal' 2010) Uses two measurements of utilities of sequences. No generic framework is proposed.

Problem Statement: Containing

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence		
1	< (e, 5)	[(c, 2)(f, 1)]	(b, 2) >
2	< [(a, 2)(e, 6)]	[(a, 1)(b, 1)(c, 2)]	[(a, 2)(d, 3)(e, 3)] >
3	< (c, 1)	[(a, 6)(d, 3)(e, 2)]	>
4	< [(b, 2)(e, 2)]	[(a, 7)(d, 3)]	[(a, 4)(b, 1)(e, 2)] >
5	< [(b, 2)(e, 3)]	[(a, 6)(e, 3)]	[(a, 2)(b, 1)] >

$(a, 2)$: **Q-item**

$[(a, 2)(e, 6)]$: **Q-itemset**

$s_1 - s_5$: **Q-sequence**

- Q-itemset containing

$[(a, 4)(b, 1)(e, 2)]$ contains q-itemsets $(a, 4)$, $[(a, 4)(e, 2)]$ and $[(a, 4)(b, 1)(e, 2)]$ but not $[(a, 2)(e, 2)]$ and $[(a, 4)(c, 1)]$.

- Q-sequence containing

$<[(b, 2)(e, 3)][(a, 6)(e, 3)][(a, 2)(b, 1)]>$ contains q-sequences $<(b, 2)>$, $<[(b, 2)(e, 3)]>$ and $<[(b, 2)][(e, 3)](a, 2)>$ but not $[(a, 2)(e, 2)]$ and $[(a, 4)(c, 1)]$.

Problem Statement: Matching

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence		
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >		
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >		
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >		
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >		
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >		

Sequence $\langle ea \rangle$ matches:

$\langle (e, 6)(a, 1) \rangle$ and $\langle (e, 6)(a, 2) \rangle$ in s_2 ;
 $\langle (e, 2)(a, 7) \rangle$ and $\langle (e, 2)(a, 4) \rangle$ in s_4 ;
 $\langle (e, 3)(a, 6) \rangle$ and $\langle (e, 3)(a, 2) \rangle$ in s_5 ;

Denote as $\langle (e, 6)(a, 1) \rangle \sim \langle ea \rangle$

Problem Statement: Utilities

The Sequence Utility Framework

The q-item utility:

$$u(i, q) = f_{u_i}(p(i), q)$$

The q-itemset utility:

$$u(l) = f_{u_{ls}}\left(\bigcup_{j=1}^n u(i_j, q_j)\right)$$

The q-sequence utility:

$$u(s) = f_{u_s}\left(\bigcup_{j=1}^m u(l_j)\right)$$

The q-sequence database utility:

$$u(S) = f_{u_{ab}}\left(\bigcup_{j=1}^r u(s_j)\right)$$

The sequence utility in a q-sequence:

$$v(t, s) = \bigcup_{s' \sim t \cap s' \subseteq s} u(s')$$

The sequence utility in a database:

$$v(t) = \bigcup_{s \in S} v(t, s)$$

For example:

$$v(<ea>, s_4) = \{u(<(e, 2)(a, 7)>), u(<(e, 2)(a, 4)>)\}$$

$$v(<ea>) = \{v(<ea>, s_2), v(<ea>, s_4), v(<ea>, s_5)\}$$

Problem Statement: Utilities

High Utility Sequential Pattern Mining

The q-item utility:

$$f_{u_i}(p(i), q) = p(i) \times q$$

The q-itemset utility:

$$f_{u_{is}}\left(\bigcup_{j=1}^n u(i_j)\right) = \sum_{j=1}^n u(i_j, q_j)$$

The q-sequence utility:

$$f_{u_s}\left(\bigcup_{j=1}^m u(l_j)\right) = \sum_{j=1}^m u(l_j)$$

The q-sequence database utility:

$$f_{u_{ab}}\left(\bigcup_{j=1}^r u(s_j)\right) = \sum_{j=1}^r u(s_j)$$

The sequence utility in a database:

$$v(t) = u_{max}(t) = \sum \max\{u(s') | s' \sim t \cap s' \subseteq s \cap s \in S\}$$

For example:

$$V(<ea>, s_4) = \{16, 10\}$$

$$V(<ea>) = \{ \{8, 10\}, \{16, 10\}, \{15, 7\} \}$$

Sequence t is a high utility sequential pattern if and only if $u_{max} \geq \xi$
where ξ is a user-specified minimum utility.

Target: Extracting all high utility sequential patterns in S satisfying ξ .

USpan Algorithm

Challenges of mining for high utility patterns

$$u_{max}(<a>) = 4 + 12 + 14 + 12 = 42$$

$$u_{max}(<ab>) = 7 + 13 + 9 = 29$$

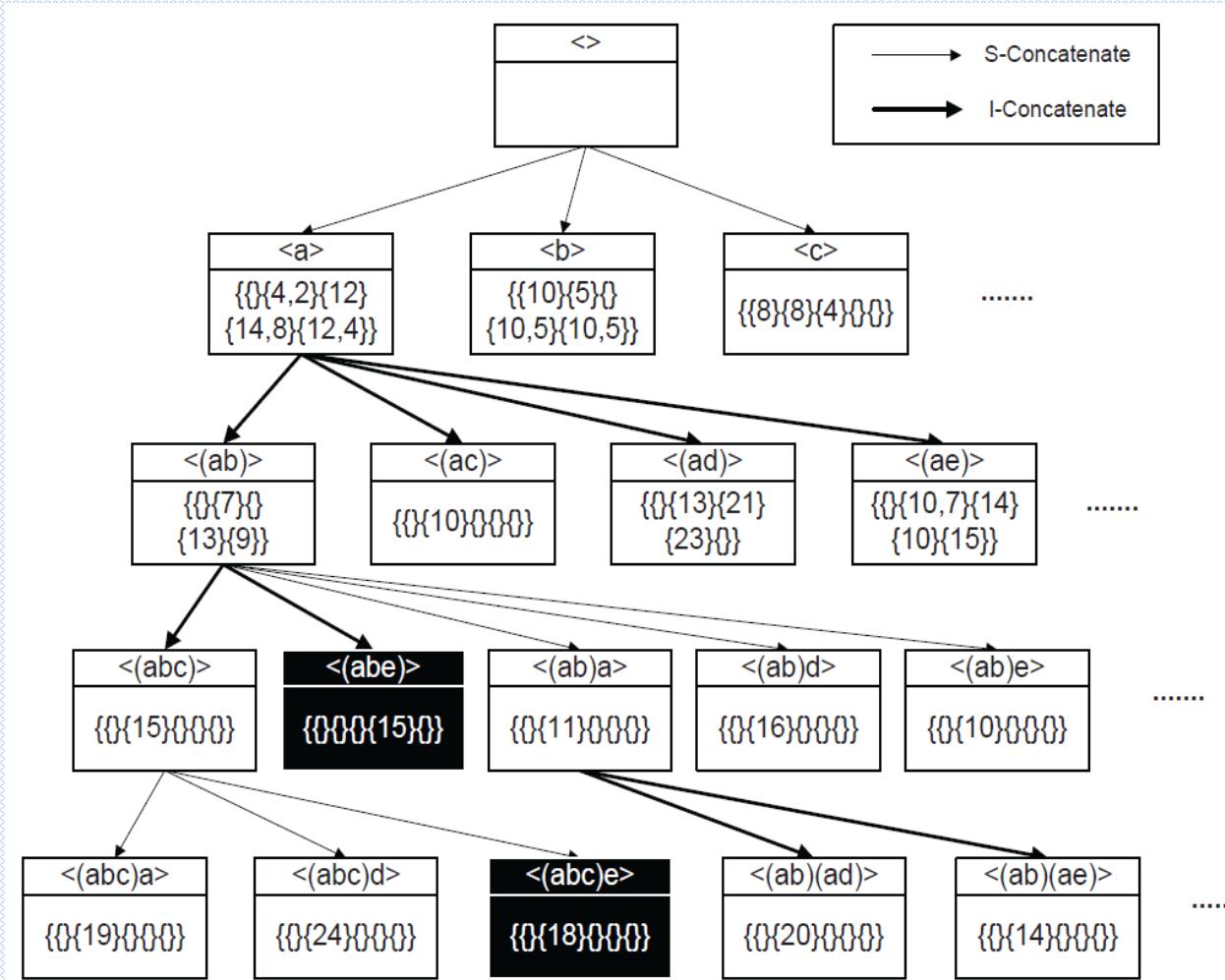
$$u_{max}(<abc>) = 15$$

$$u_{max}(<(abc)a>) = 19$$

No Downward Closure Property

USpan Algorithm

Lexicographic Q-sequence Tree



USpan Algorithm

Table 1: Quality Table

Items	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	<i>e</i>	<i>f</i>
Quality	2	5	4	3	1	1

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence		
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >		
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >		
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >		
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >		
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >		

Items	Itemset 1	Itemset 2	Itemset 3
<i>a</i>		14	8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

$v(<(be)>) = \{10 + 2, 5 + 2\} = \{12, 7\}$

Items	<i>I 1</i>	<i>I 2</i>	<i>I 3</i>
<i>a</i>		14	8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

$$v(<(be)>) = \{10 + 2, 5 + 2\} = \{12, 7\}$$

Items	<i>I 1</i>	<i>I 2</i>	<i>I 3</i>
<i>a</i>		14	8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

$$v(<(be)a>) = \{12 + 14, 12 + 8\} = \{26, 20\}$$

Items	<i>I 1</i>	<i>I 2</i>	<i>I 3</i>
<i>a</i>	14		8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

$$v(<(be)(ad)a>) = \{35 + 8\}$$

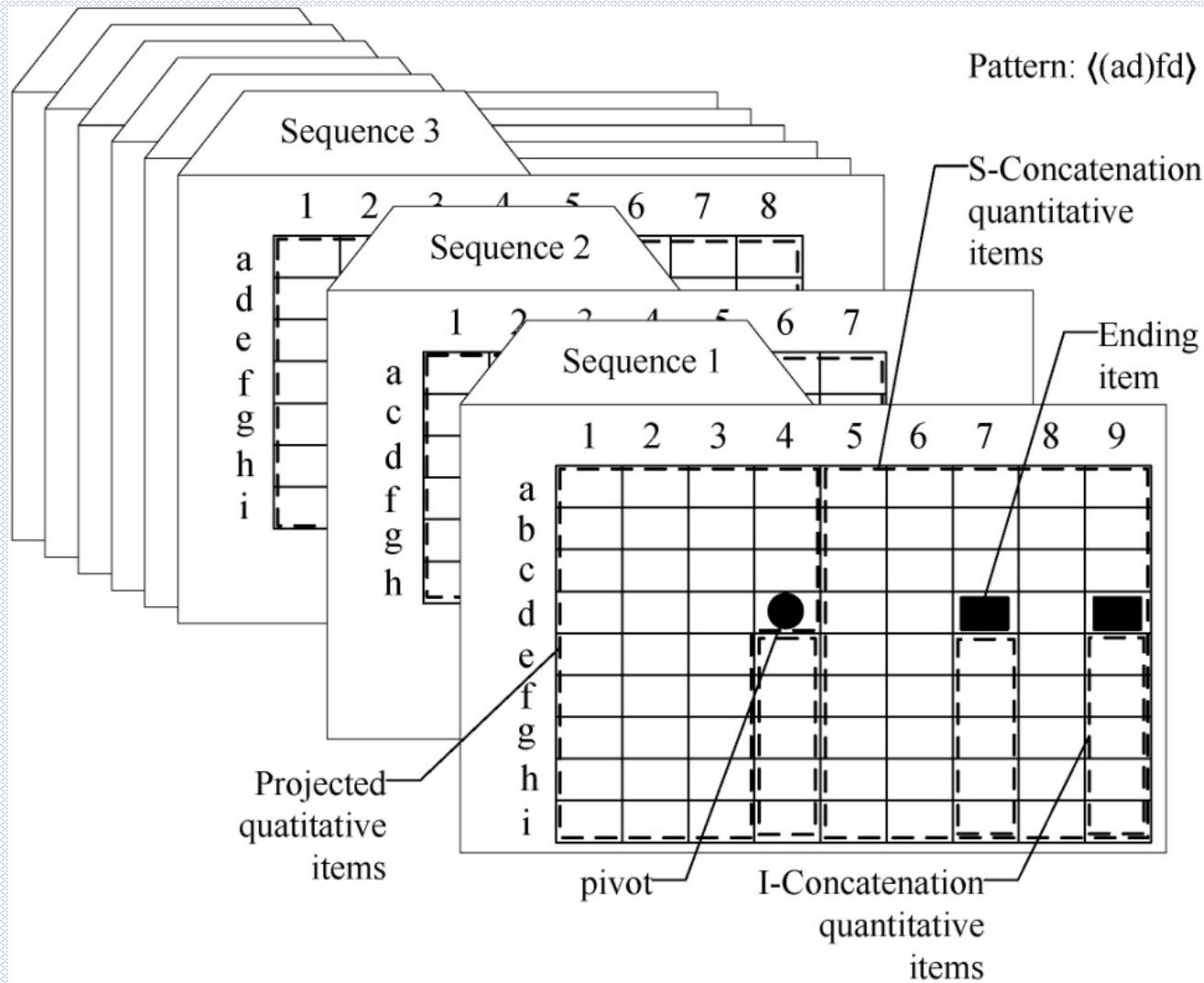
Items	<i>I 1</i>	<i>I 2</i>	<i>I 3</i>
<i>a</i>		14	8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

$$v(<(be)(ad)>) = \{26 + 9\} = \{35\}$$

Items	<i>I 1</i>	<i>I 2</i>	<i>I 3</i>
<i>a</i>		14	8
<i>b</i>	10		5
<i>d</i>		9	
<i>e</i>	2		2

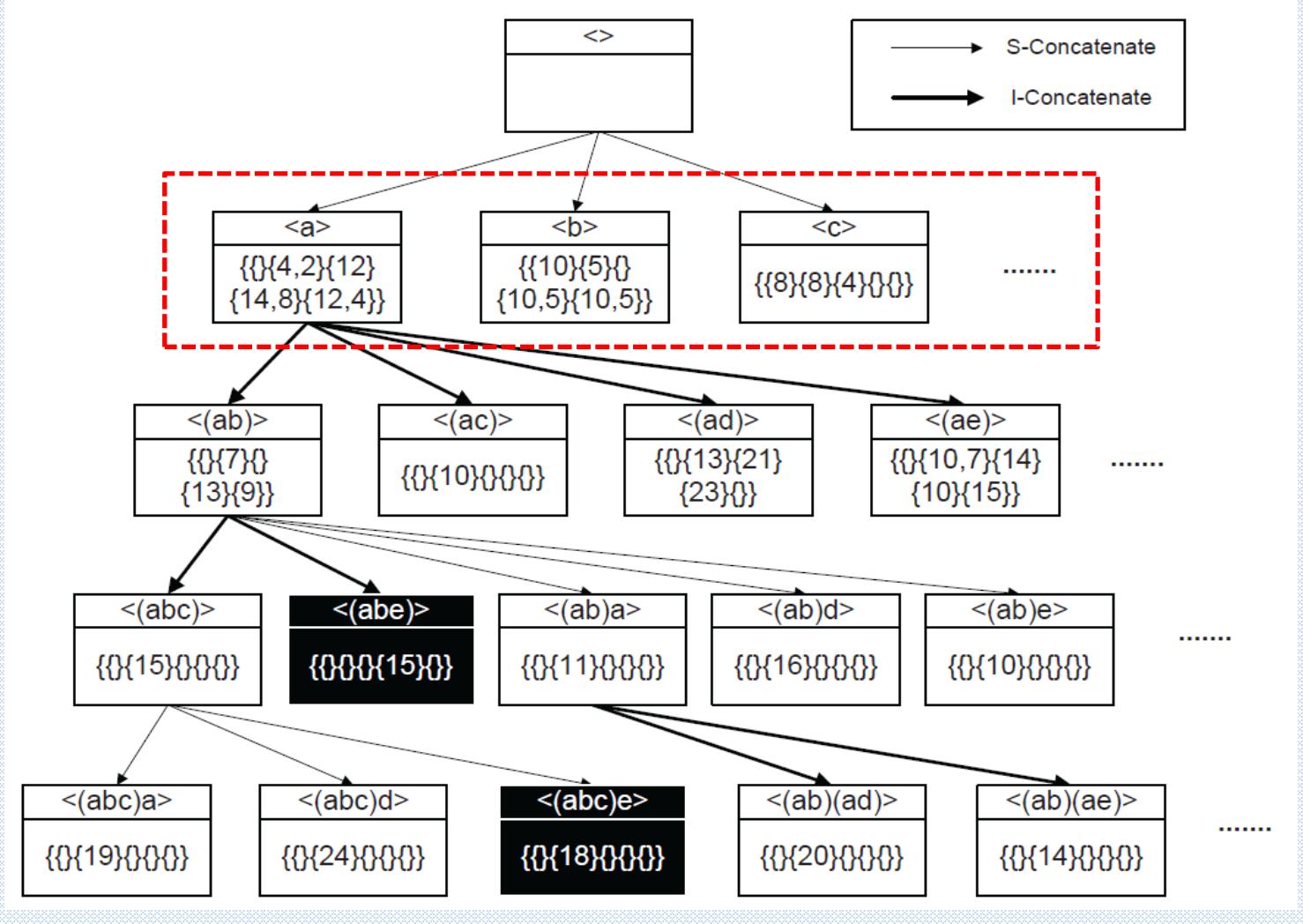
USpan Algorithm: Concatenation

Data Representation



USpan Algorithm: Width Pruning

What is Width Pruning



USpan Algorithm: Width Pruning

What to Width Prune

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

<f> should be **width-pruned**

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence	SU
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >	24
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >	41
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >	27
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >	50
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >	42

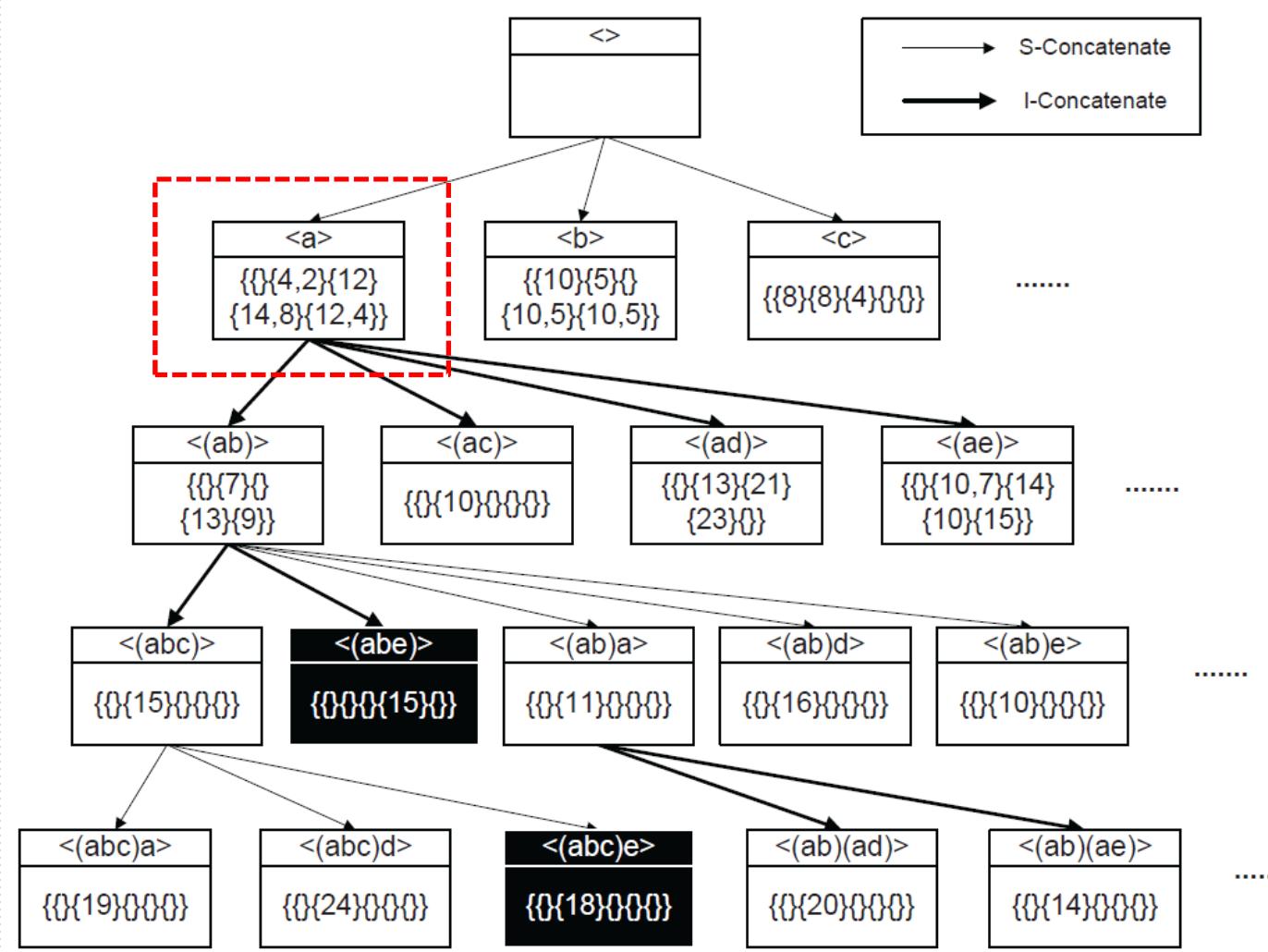
$$\begin{aligned}
 \text{SWU}(<ea>) &= u(s_2) + u(s_4) + u(s_5) \\
 &= 41 + 50 + 24 \\
 &= 115
 \end{aligned}$$

SID	Quantitative Sequence	SU
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >	24
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >	41
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >	27
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >	50
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >	42

$$\text{SWU}(<f>) = u(s_1) = 24$$

USpan Algorithm: Depth Pruning

What is Depth Pruning



USpan Algorithm: Depth Pruning

What to Depth Prune

Table 1: Quality Table

Items	a	b	c	d	e	f
Quality	2	5	4	3	1	1

<e(ae)> should be **depth-pruned**

Table 2: Quantitative Sequence Database

SID	Quantitative Sequence	SU
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >	24
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >	41
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >	27
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >	50
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >	42

$$\begin{aligned}
 u_{rest}(<ea>) &= (8+29) + (16+24) + (15+17) \\
 &= 37 + 40 + 32 \\
 &= 109
 \end{aligned}$$

SID	Quantitative Sequence	SU
1	< (e, 5) [(c, 2)(f, 1)] (b, 2) >	24
2	< [(a, 2)(e, 6)] [(a, 1)(b, 1)(c, 2)] [(a, 2)(d, 3)(e, 3)] >	41
3	< (c, 1) [(a, 6)(d, 3)(e, 2)] >	27
4	< [(b, 2)(e, 2)] [(a, 7)(d, 3)] [(a, 4)(b, 1)(e, 2)] >	50
5	< [(b, 2)(e, 3)] [(a, 6)(e, 3)] [(a, 2)(b, 1)] >	42

$$\begin{aligned}
 u_{rest}(<e(ae)>) &= (18 + 9) \\
 &= 27
 \end{aligned}$$

Experiments

Datasets

Synthetic Datasets

Parameters	DS1	DS2
that the average number of elements	10	8
the average number of items in an element	2.5	2.5
the average length of a maximal pattern	4	6
the average number of items per element	2.5	2.5
Number of sequences	10k	10k
Number of items	1k	10k

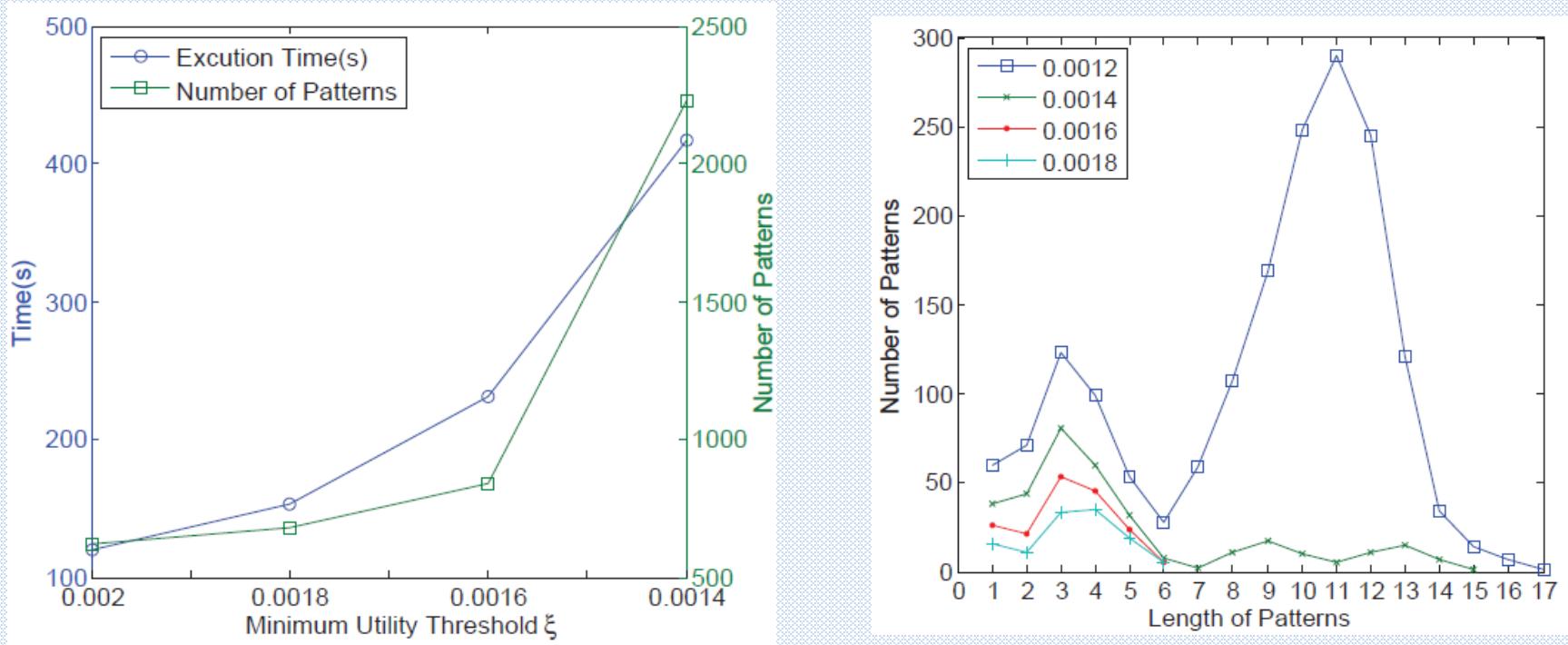
Real Datasets

DS3 is a dataset consisting of online shopping transactions which contains 350,241 transactions and 59,477 customers.

DS4 is a real dataset that includes mobile communication transactions. The dataset is a 100,000 mobile call history from a specific day. There are 67,420 customers in the dataset.

Experiments

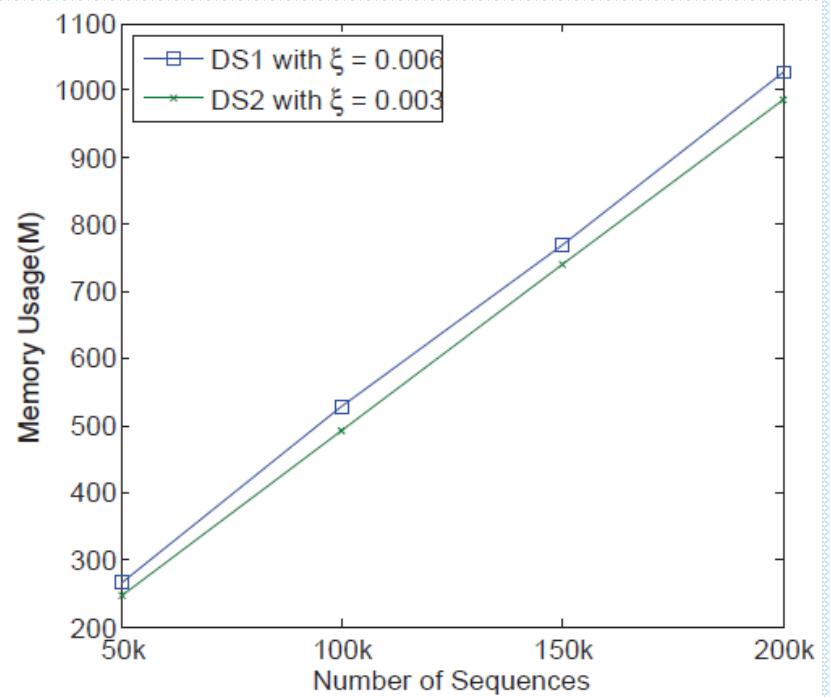
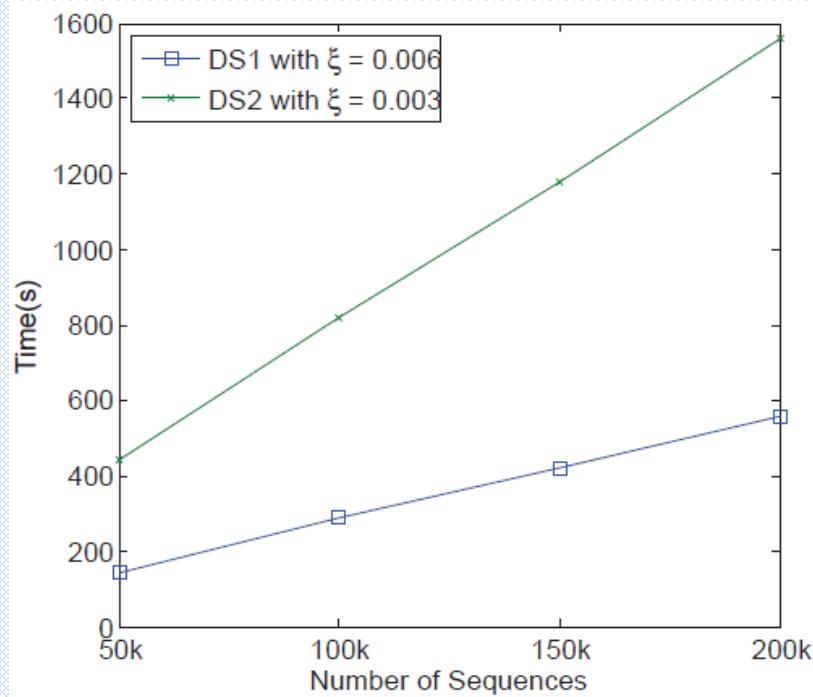
Performance and distributions (DS2)



- The running time and the number of patterns grow exponentially with respect to ξ .
- The high utility sequential patterns are mid-long patterns.

Experiments

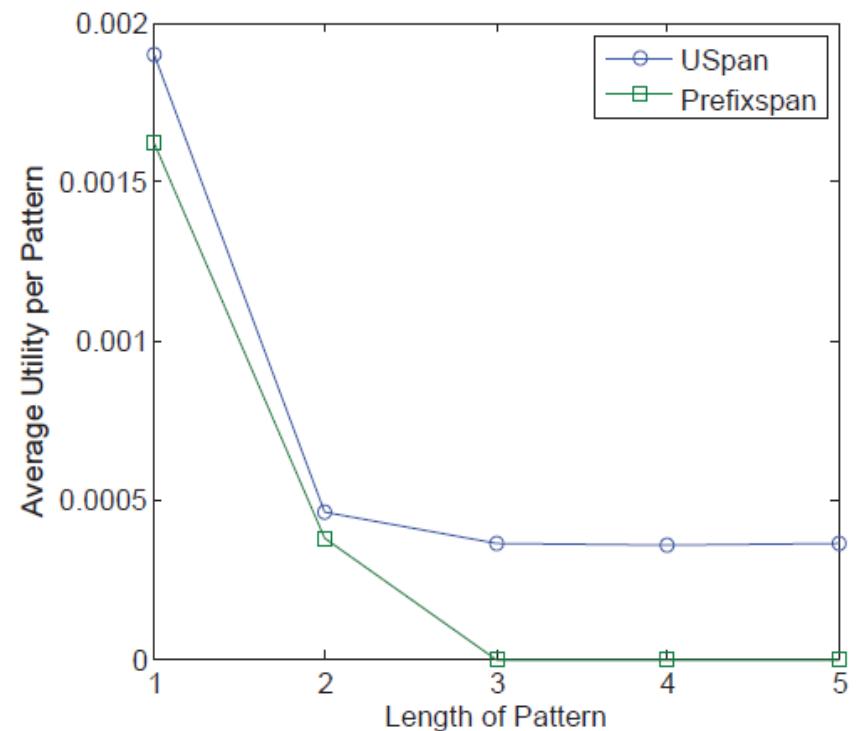
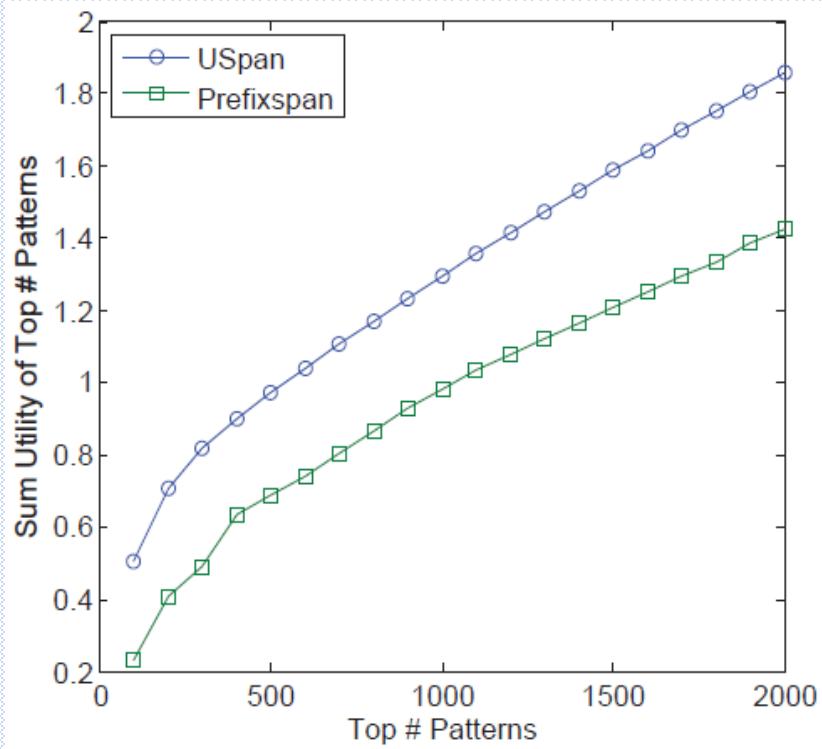
Scalability Test (DS1 & DS2)



- Both the time and memory usage grow linearly with respect to the size of the DB.

Experiments

High Utility Sequential Pattern vs. Frequent Sequential Patterns (DS3)



- USpan out performs Prefixspan with respect to the utilities of the patterns.

Conclusions

1. We define the problem of mining high utility sequential patterns.
2. We propose the USpan to efficiently mine for mining high utility sequential patterns.
3. Two pruning strategies are proposed to substantially reduce the search space.
4. Experiments on both synthetic and real datasets show that USpan can discover the high utility sequential patterns efficiently.

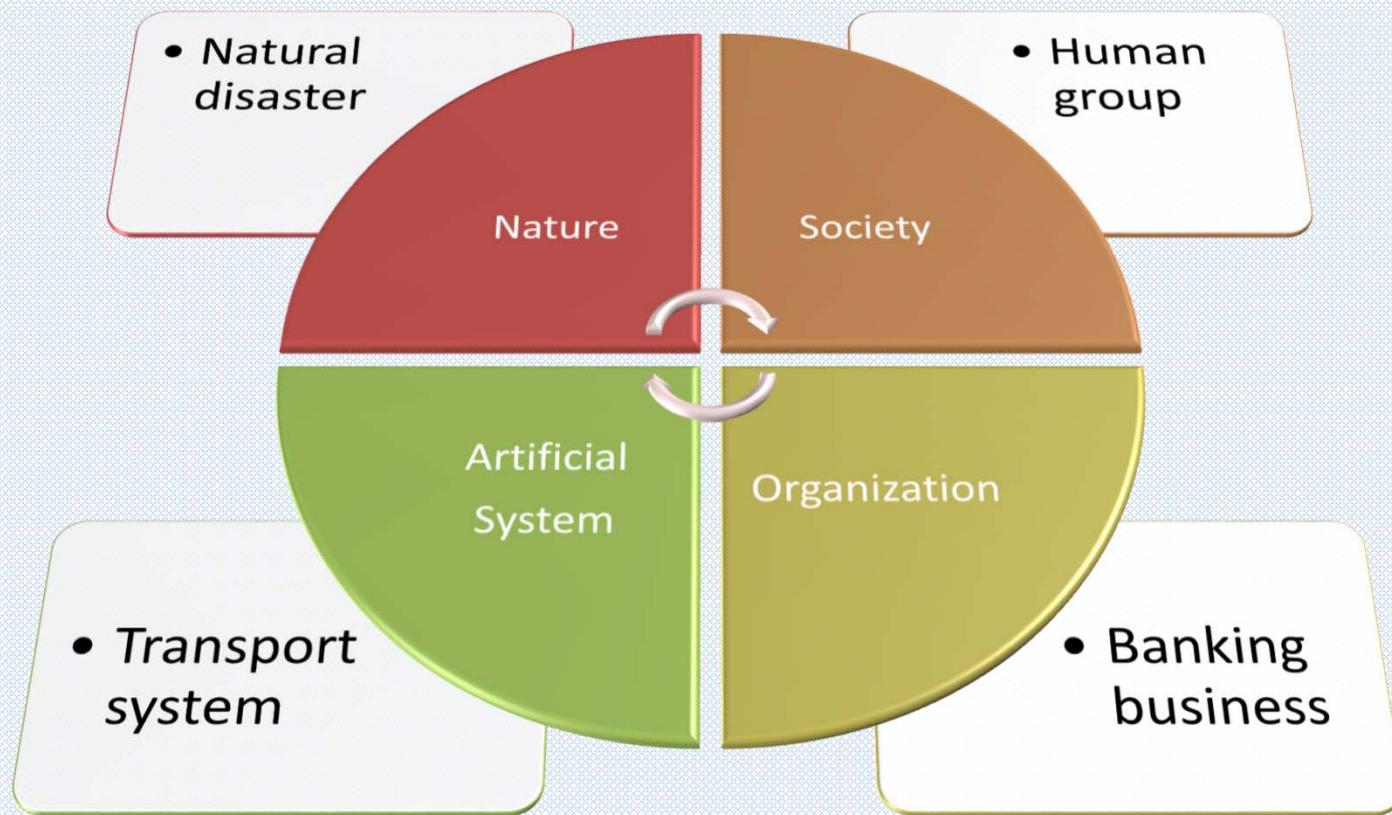
Part VI.

Behavior Analysis Applications

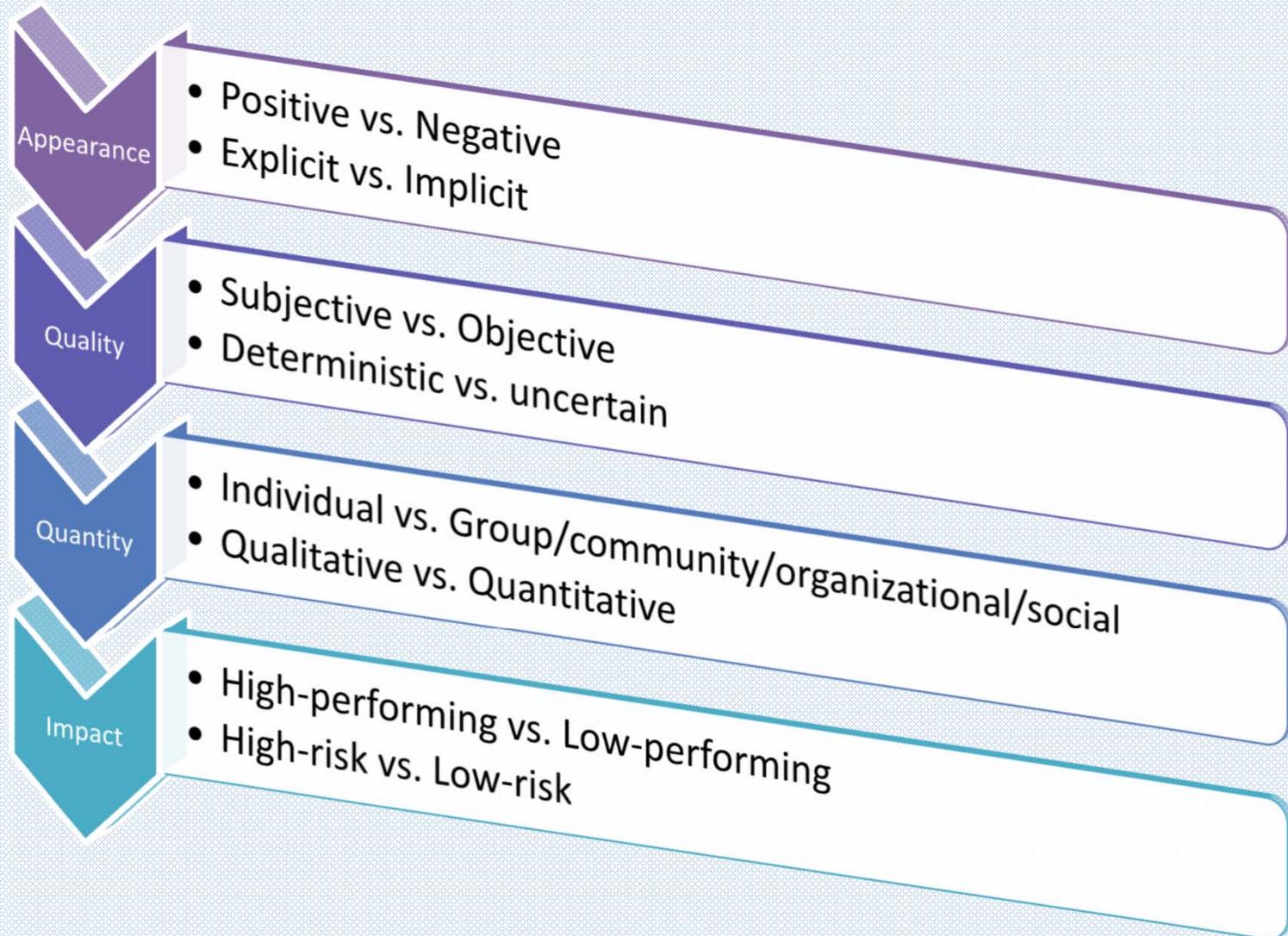
Part VI.

Prospects & Summary

Behaviour is ubiquitous



Behaviour is a valuable asset



So-called behavior analysis vs. behavior informatics

Aspects	Traditional behavior analysis	Behavior computing
Objective	<i>Behavior exterior Explicit behaviors</i>	<i>Behavior interior & exterior; Implicit behaviors</i>
Means	Empirical, qualitative, psychological understanding	Quantitative understanding
Data	Observations and appearance including customer demographic, service usage	Actors, actions, couplings, context
Management	Transactions with entity relationships	Behavior feature matrix
Expect	Appearance, observations of behaviors	Behavioral actor's belief, desire, intention and impact for internal driving forces or causes

Prospects

-Individual behavior
-- individual customer

Sequence analysis

Frequent Pattern mining

Event detection

- Group behavior
-- Group customer

Coupled behavior analysis

Non-IID Behaviors

Impact-oriented behaviors

- 1) Impact-oriented:
 - Positive
 - Negative
 - Multi-level
 - Mixed
 - Evolution
- 2) Non-IID behaviors
- 3) Group behaviors

Non-IIDness Learning in Behavioral and Social Data

LONGBING CAO*

Advanced Analytics Institute, University of Technology, Sydney, Australia

*Corresponding author: longbing.cao@uts.edu.au

Most of the classic theoretical systems and tools in statistics, data mining and machine learning are built on the fundamental assumption of IIDness, which assumes the independence and identical distribution of underlying objects, attributes and/or values. However, complex behavioral and social problems often exhibit strong couplings and heterogeneity between values, attributes and objects (i.e., non-IIDness). This fundamentally challenges the IIDness-based learning methodologies and techniques. This paper presents a high-level overview of the needs, challenges and opportunities of non-IIDness learning for handling complex behavioral and social problems. By reviewing the nature and issues of classic IIDness-based algorithms in frequent pattern mining, clustering and classification to complex behavioral and social applications, concepts, structures, frameworks and exemplar techniques are discussed for non-IIDness learning. Case studies, related work and prospects of non-IIDness learning are presented. Non-IIDness learning is also a fundamental issue in big data analytics.

Keywords: non-IIDness learning; IIDness; IID data; non-IID data; coupling; behavior informatics; social informatics

Received 13 February 2013; revised 5 July 2013
Handling editor: Guandong Xu

1. INTRODUCTION

Behavioral and social applications are ubiquitous, ranging from business and online applications to social and organizational applications and domains. With the increasing and continuous development of such applications, an emerging need is to develop an in-depth understanding of the underlying working mechanism driving force, dynamics and evolution of a behavioral and/or social system, as well as the impact on business and context. To this end, building on the classic theories and tools available in behavioral science and social science, behavior informatics [1, 2] and social informatics [3]¹ have recently been studied to 'formalize', 'quantify' and 'compute' complex behavioral and social applications.

As an emerging area of research, behavior and social informatics is in its earliest stage and features many challenges and opportunities. A canonical trend is to develop theories, tools and algorithms based on the classic outcomes available in extant disciplines including statistics, data mining and machine learning. Typically, frequent pattern mining, clustering

and classification of behavioral and social applications are conducted by expanding the corresponding existing theories and algorithms. In this paper, we discuss the potential issues and risk in pursuing this path for complex behavioral and social applications by explicitly or implicitly taking the IIDness assumption, and thus reveal the need for developing non-IIDness learning for behavior and social informatics.

Arguably, most of the existing theories, tools and systems in statistics, data mining and machine learning are built on the IIDness assumption, which assumes the independence and identical distribution of the underlying objects, attributes and/or values. Based on a high-level abstraction, it is assumed that objects, attributes and values are independent and identically distributed, with most of existing learning theories, models and algorithms proposed on the basis of this assumption. This works well in simple business applications and abstract problems with weakened and avoidable relations and heterogeneity, and serves as the foundation of classic analytics, mining and learning theories, algorithms, systems and tools.

Complex behavioral and social applications often exhibit strong coupling relations (which are beyond the usual dependency relation) and heterogeneity between objects, object

¹See more from the IEEE Task Force on Behavior and Social Informatics and Computing: www.bsic.info

Behavior non-IIDness analysis

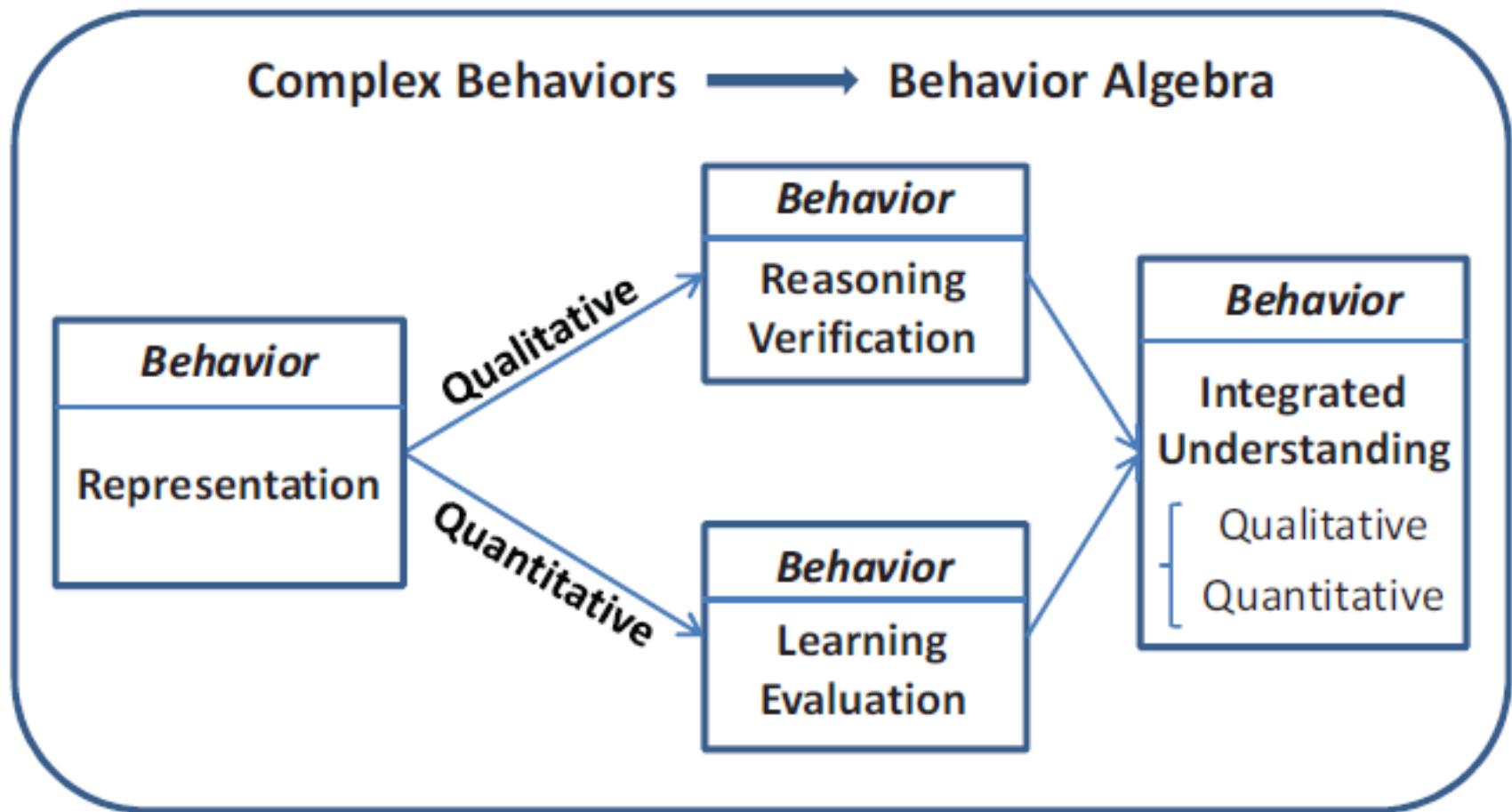
Longbing Cao. Non-IIDness Learning in Behavioral and Social Data, The Computer Journal, 57(9): 1358-1370 (2014).

Longbing Cao. Coupling Learning of Complex Interactions, Journal of Information Processing and Management, 51(2): 167-186 (2015).

Issues Addressed

- Behaviors are non-IID, namely not independent and identically distributed (IID)
- What is the non-IIDness of behavior-related problems?
- What are coupling relationships of non-IID behaviors?
- What are heterogeneity of non-IID behaviors?
- What are opportunities and prospects of non-IID behavior and social problem study?

Modeling and Analysis of Complex Behaviors



Part VII.

Checklist

Checklist

- What is behavior
- Why behavior informatics
- How to represent a behavioral application/system
- How to verify a behavior model
- Individual's behavior pattern
- Group behavior pattern
- How to measure behavior impact

Checklist

- Impact-oriented behavior pattern
- Non-occurring behavior pattern
- Group behavior analysis
- Behavior informatics conceptual map
- Application of BI

Behaviour analytics matrix

Behaviour analytics matrix

Organization: _____

Business problem: _____

	Actor model	Behaviour model	Relationship model	Analytical goal	Impact & evaluation metrics	Analytical methods	Actionable deliverable
Individual							
Group							
Context							

Maximize the behaviour value

Behaviour analytics for smart business



References

- **Longbing Cao**, Xiangjun Dong and Zhigang Zheng. [e-NSP: Efficient Negative Sequential Pattern Mining](#), Artificial Intelligence, 235: 156-182, <http://dx.doi.org/10.1016/j.artint.2016.03.001>, 2016. Download [e-NSP codes and example](#); Download [NegSeq codes and example](#)
- Yin Song, **Longbing Cao**, et al. [Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulation](#), KDD 2012, 976-984.

Yin Song and **Longbing Cao**. [Graph-based Coupled Behavior Analysis: A Case Study on Detecting Collaborative Manipulations in Stock Markets](#), IJCNN 2012, 1-8.

Longbing Cao, Yuming Ou, Philip S Yu. [Coupled Behavior Analysis with Applications](#), IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).

Zhong She, Can Wang, and **Longbing Cao**. A Coupled Framework of Clustering Ensembles, AAAI2012 (poster)

Can Wang, and **Longbing Cao**. [Modeling and Analysis of Social Activity Process](#), in Longbing Cao and Philip S Yu (eds) Behavior Computing, 21-35, Springer, 2012

Can Wang, Mingchun Wang, Zhong She, **Longbing Cao**. [CD: A Coupled Discretization Algorithm](#), PAKDD2012, 407-418

Can Wang, **Longbing Cao**, Minchun Wang, Jinjiu Li, Wei Wei, Yuming Ou. [Coupled Nominal Similarity in Unsupervised Learning](#), CIKM 2011, 973-978.

Xiangjun Dong, Zhigang Zhao, **Longbing Cao**, Yanchang Zhao, Chengqi Zhang, Jinjiu Li, Wei Wei, Yuming Ou. [e-NSP: Efficient Negative Sequential Pattern Mining Based on Identified Positive Patterns Without Database Rescanning](#), CIKM 2011, 825-830.

- **Longbing Cao**, [In-depth Behavior Understanding and Use: the Behavior Informatics Approach](#), Information Science, 180(17); 3067-3085, 2010.
Longbing Cao, Yuming Ou, Philip S YU, Gang Wei. [Detecting Abnormal Coupled Sequences and Sequence Changes in Group-based Manipulative Trading Behaviors](#), KDD2010, 85-94.
Zhigang Zheng, Yanchang Zhao, Ziye Zuo**Longbing Cao**, Huafeng Zhang, Yanchang Zhao, Chengqi Zhang. [An Efficient GA-Based Algorithm for Mining Negative Sequential Patterns](#), PAKDD2010, 262-273
- **Longbing Cao**, Philip S Yu, Behavior Informatics: An Informatics Perspective for Behavior Studies, The Intelligent Informatics Bulletin, 10(1): 6-11, 2009.
Zhigang Zheng, Yanchang Zhao, Ziye Zuo, **Longbing Cao**. [Negative-GSP: An Efficient Method for Mining Negative Sequential Patterns](#), AusDM 2009: 63-67.
Shanshan Wu, Yanchang Zhao, Huafeng Zhang, Chengqi Zhang, **Longbing Cao**, Hans Bohlscheid. [Debt Detection in Social Security by Adaptive Sequence Classification](#), KSEM 2009: 192-203.
- Yanchang Zhao, Huafeng Zhang, Shanshan Wu, Jian Pei, **Longbing Cao**, Chengqi Zhang and Hans Bohlscheid. [Debt Detection in Social Security by Sequence Classification Using Both Positive and Negative Patterns](#), ECML/PKDD2009, 648-663, 2009.
- Yanchang Zhao, Huafeng Zhang, **Longbing Cao**, Chengqi Zhang and Hans Bohlscheid. [Mining Both Positive and Negative Impact-Oriented Sequential Rules From Transactional Data](#), PAKDD2009, pp.656-663.

- **Longbing Cao**, [Behavior Informatics and Analytics: Let Behavior Talk](#), DDDM2008 joint with ICDM2008, 87 - 96.
Longbing Cao Yuming Ou. [Market Microstructure Patterns Powering Trading and Surveillance Agents](#). Journal of Universal Computer Sciences, 14(14): 2288-2308, 2008.
Yanchang Zhao, Huaifeng Zhang, **Longbing Cao**, Chengqi Zhang and Hans Bohlscheid. [Efficient Mining of Event-Oriented Negative Sequential Rules](#), WI 08, pp. 336-342.
Longbing Cao. Zhao Y., Zhang, C. [Mining Impact-Targeted Activity Patterns in Imbalanced Data](#), IEEE Trans. on Knowledge and Data Engineering, 20(8): 1053-1066, 2008.
Longbing Cao, Yanchang Zhao, Chengqi Zhang, Huaifeng Zhang. [Activity Mining: from Activities to Actions](#), International Journal of Information Technology & Decision Making, 7(2): 259-273, 2008
Longbing Cao, [Behavior Informatics and Analytics: Let Behavior Talk](#), DDDM2008 joint with ICDM2008.
Chengqi Zhang, **Longbing Cao**. Keynote: Activity Mining to Strengthen Debt Prevention, Pacific Asia Conf. on Intelligence and Security Informatics (PAISI), 2007.
Longbing Cao, Yanchang Zhao, Fernando Figueiredo, Yuming Ou, Dan Luo. [Mining High Impact Exceptional Behavior Patterns](#), PAKDD2007 industry track, LNCS4819, 56-63, 2007.
Longbing Cao.[Activity mining: challenges and prospects](#). ADMA2006, LNAI4093, 582-593.

Your feedback is appreciated.

- Longbing Cao
longbing.cao@gmail.com

www-staff.it.uts.edu.au/~lbcao

www.behaviorinformatics.org