UNIVERSITY OF TECHNOLOGY SYDNEY

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Behavior Informatics and Computing

Professor Longbing Cao

Advanced Analytics Institute, University of Technology Sydney, Australia

References Download

- http://www-staff.it.uts.edu.au/~lbcao/publication/behavior-informatics-tutorial-slidesx.pdf
- http://www-staff.it.uts.edu.au/~lbcao/publication/publications.htm
- www.behaviorinformatics.org

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Acknowledgement

- I appreciate all of my team members who have made contributions to this slide. The team member names can be found from the references.
- Appreciate Ms Can Wang's great efforts in creating many of the slides.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Outline

1. Why Behavior Informatics & Computing?
2. What is Behavior?
3. What is Behavior Informatics & Computing?
4. Related Work
5. Behavior Model/Representation
6. High Impact Behavior Analysis

7 **Impact-oriented Combined Behavior Analysis**

8 **High Utility Behavior Analysis**

9 **Negative Behavior Analysis**

10 **Coupled/Group Behavior Analysis**

11 **Challenges and Prospects of Complex Behavior Computing**

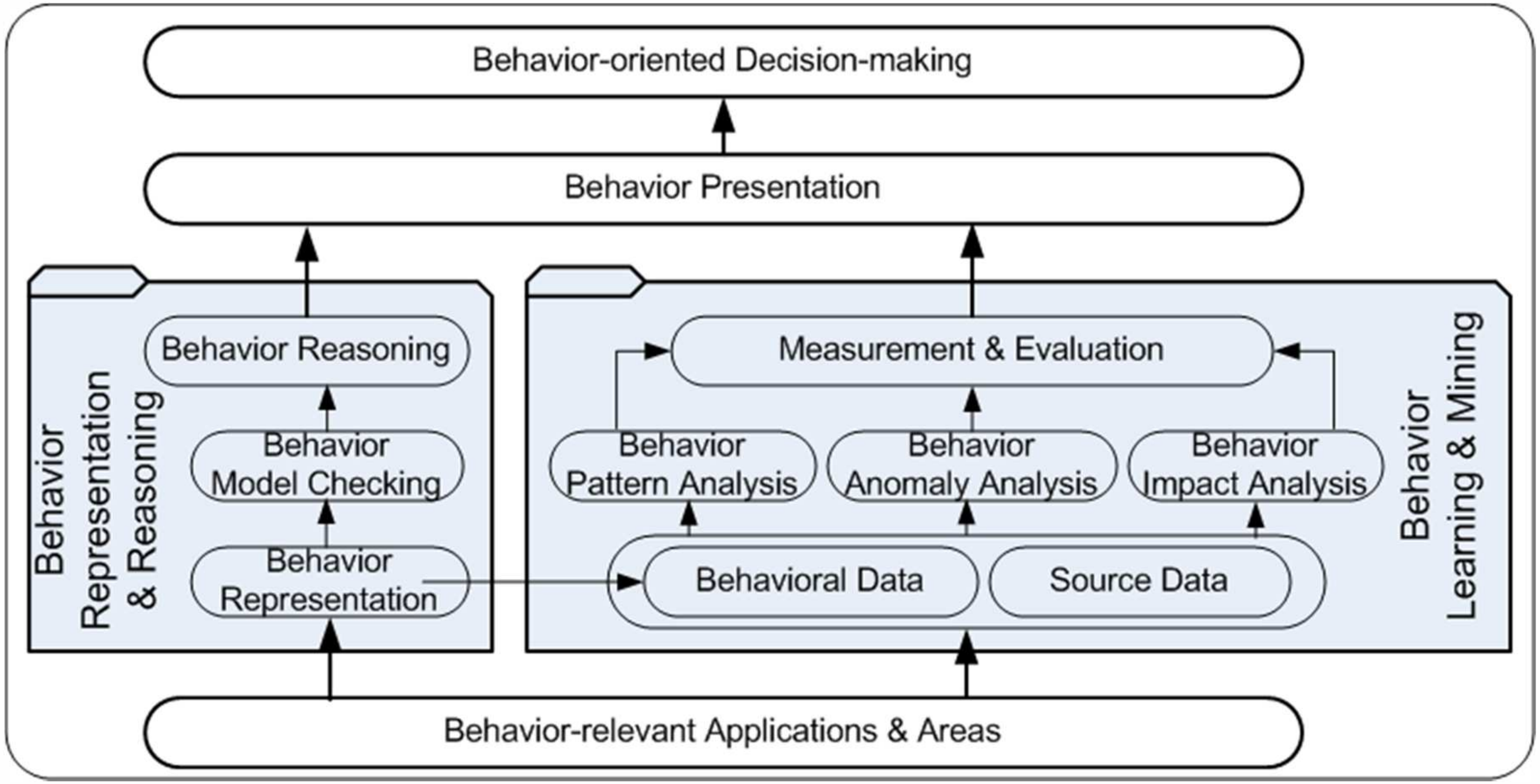
Behavior Informatics: Overview

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

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior informatics – Concept Map

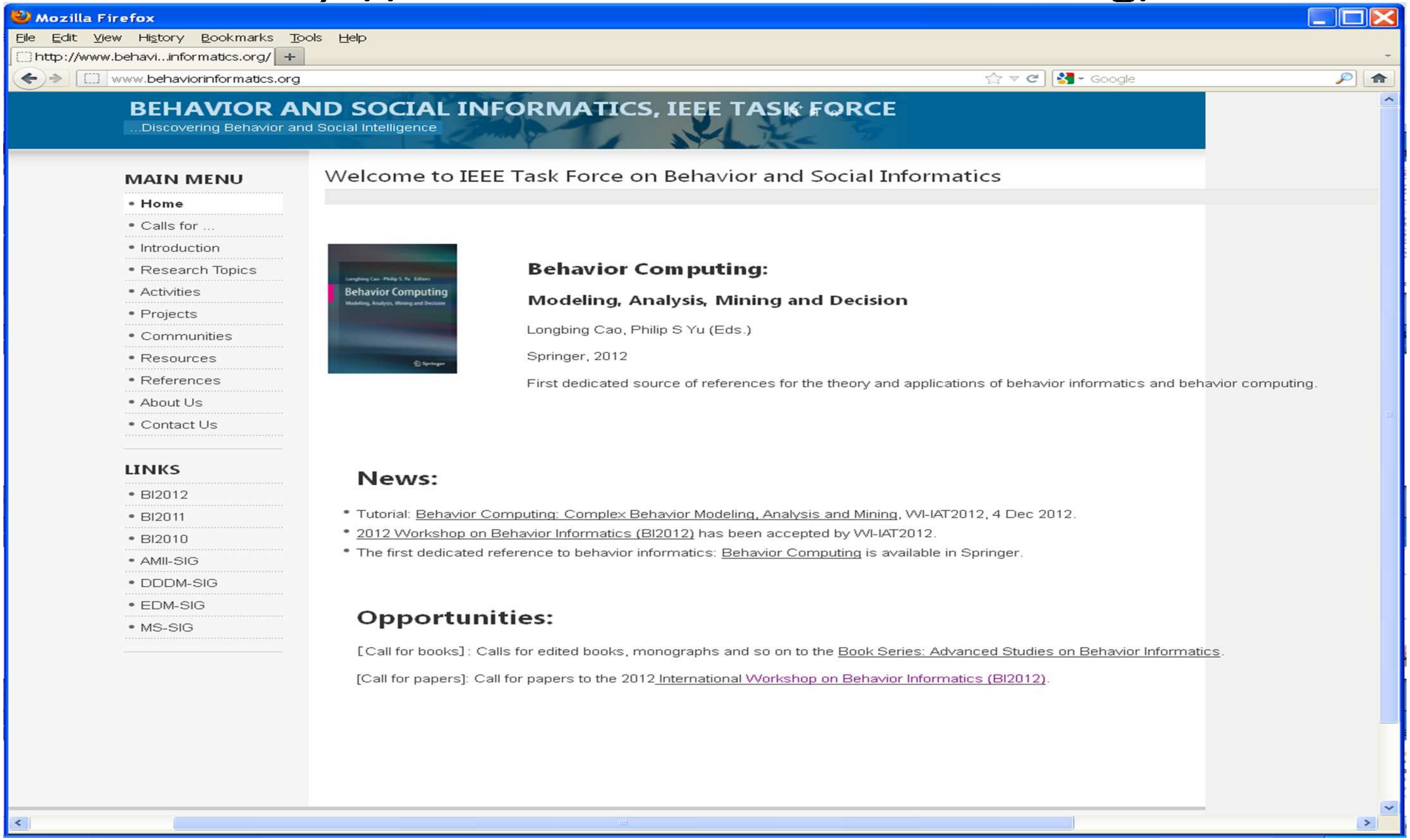


http://www.behaviorinformatics.org/

BEHAVIOR INFORMATICS
...Discovering Behavior Intelligence

Behavior Informatics-IEEE Task Force:

http://www.behaviorinformatics.org/



1. Why Behavior Informatics & Computing?

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

www.behaviorinformatics.org

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Argument 1: Behavior is ubiquitous

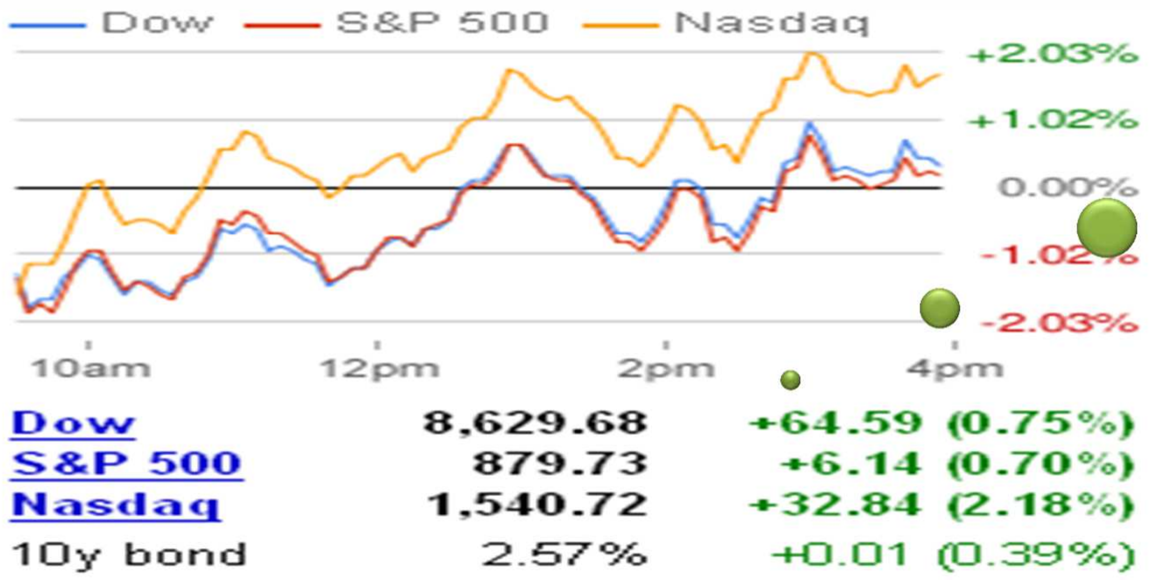
- Behavior is an important analysis object in
  - Consumer analysis
  - Marketing strategy design
  - Business intelligence
  - Customer relationship management
  - Social computing
  - Intrusion detection
  - Fraud detection
  - Event analysis
  - Risk analysis
  - Group decision-making, etc.

➢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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

- Example 1: Price movement as market behavior

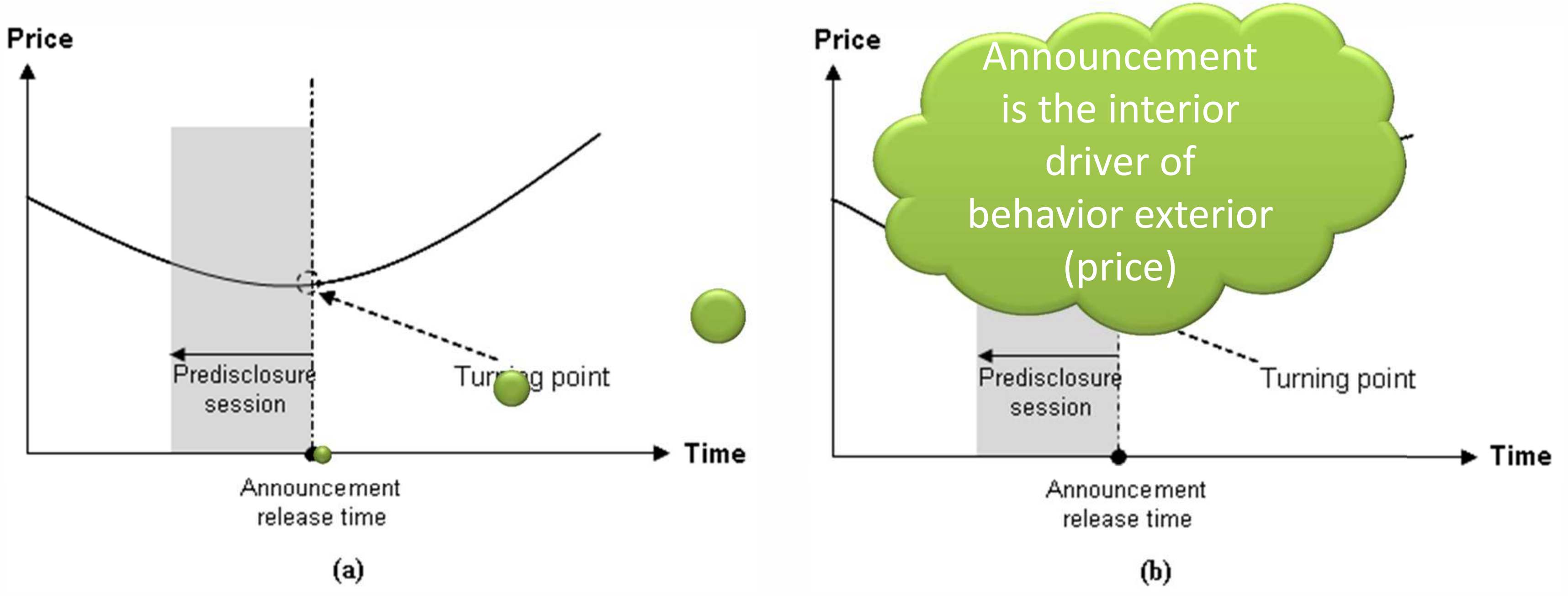


Price/index movement is the behavior exterior

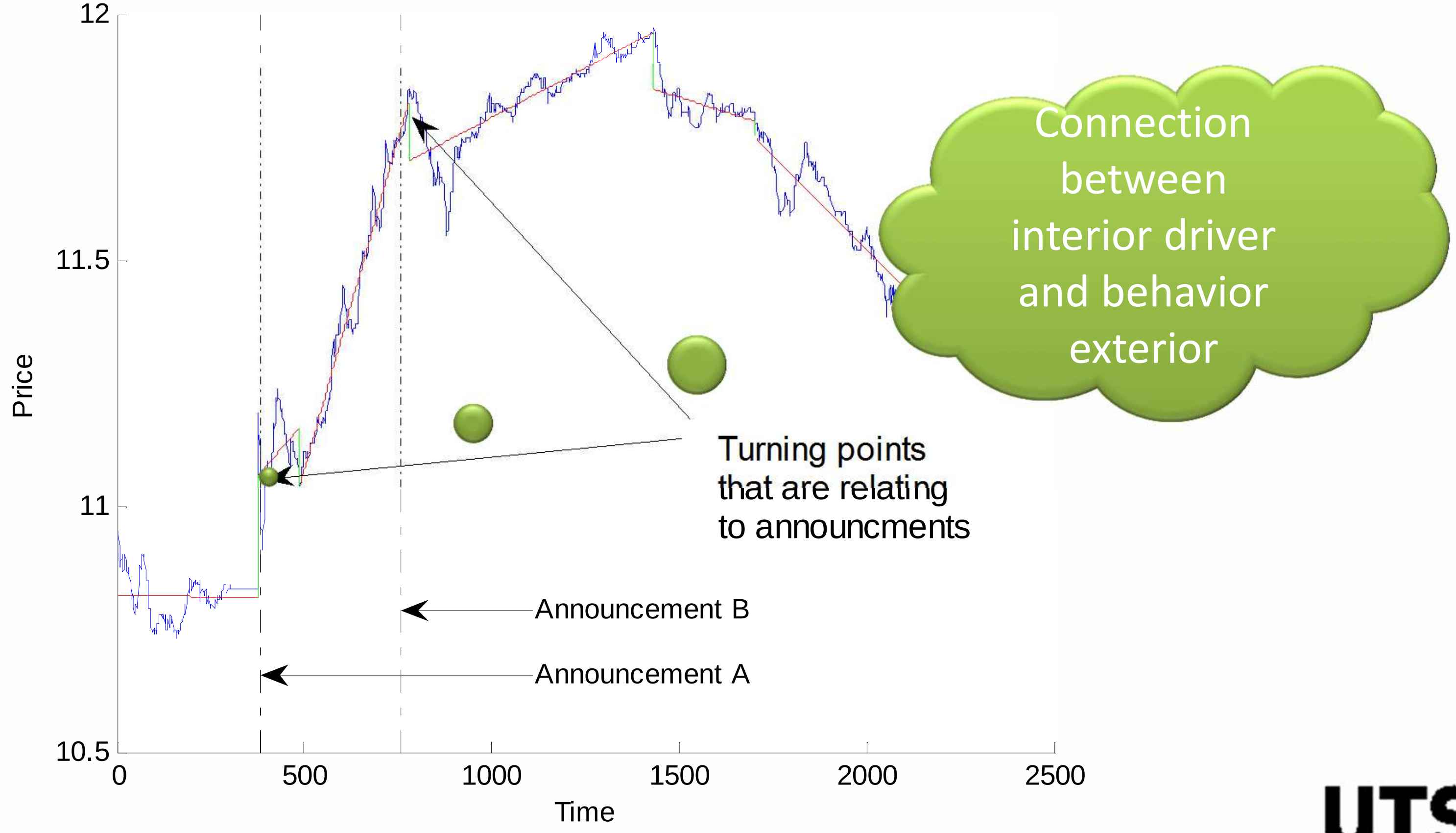
THE ADVANCED ANALYTICS INSTITUTE

Argument 3: Behavior interior-driven analysis can make difference

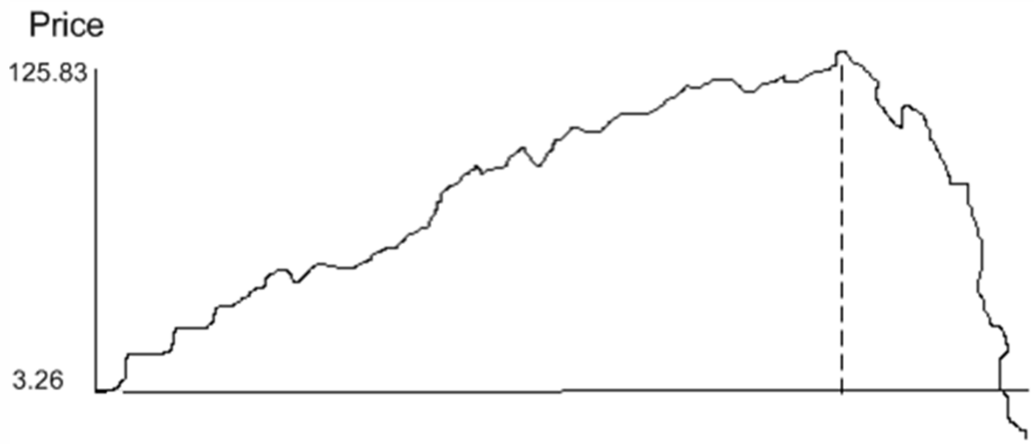
- Example 2: Announcement as market behavior driver



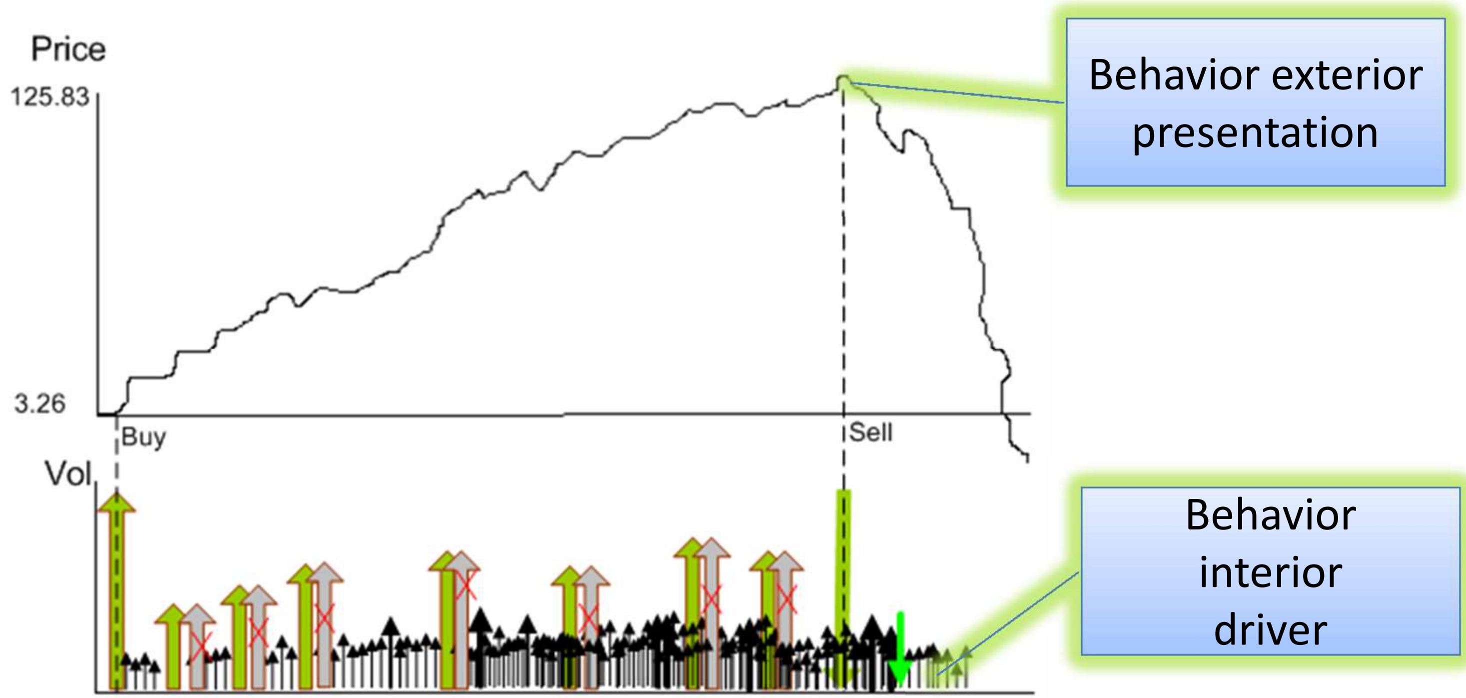
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE


Connection between interior driver and behavior exterior
Turning points that are relating to announcments
Announcement B
Announcement A
Price
Time
12
11.5
11
10.5
0
500
1000
1500
2000
2500

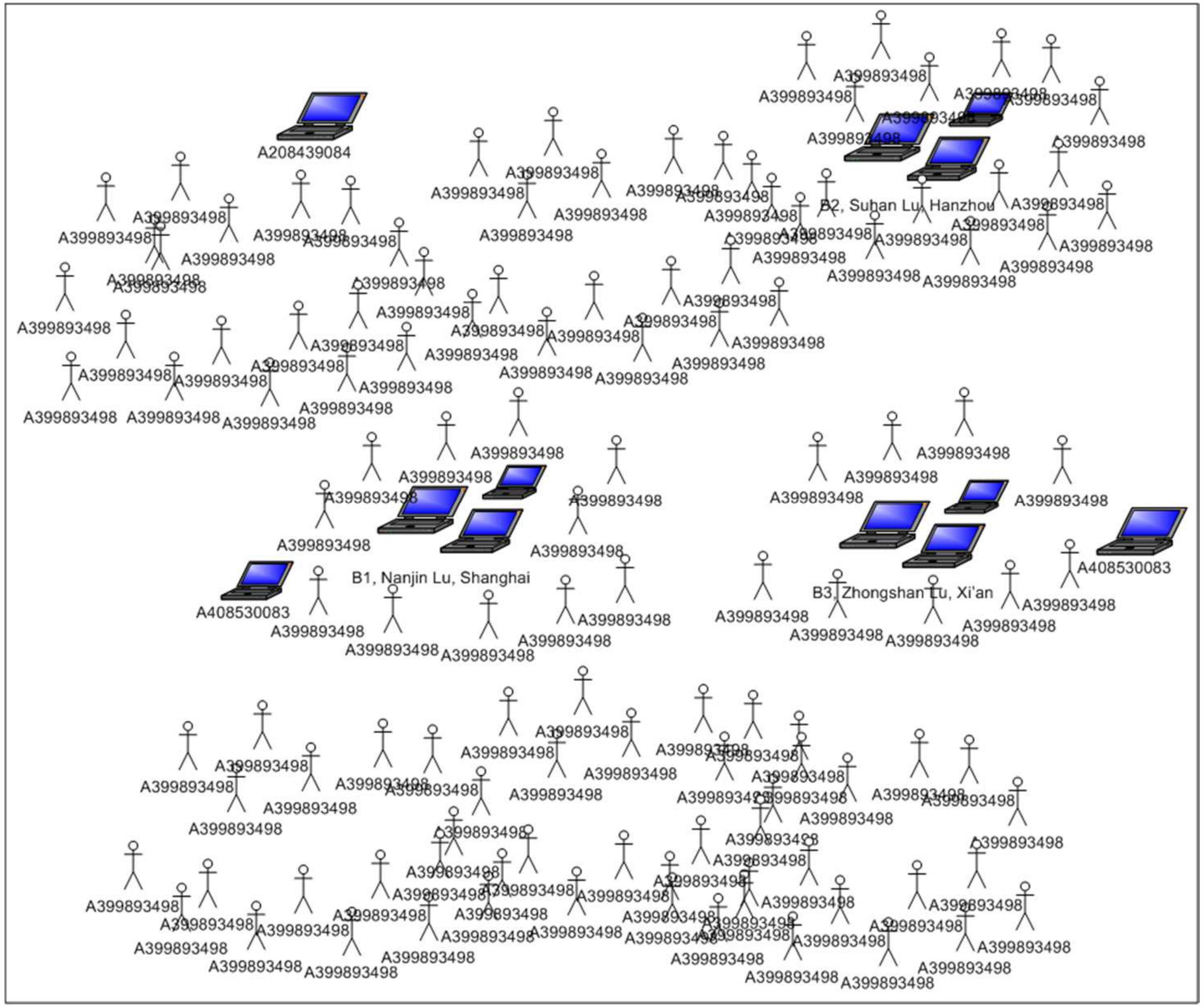
- Why does this stock go so crazily?

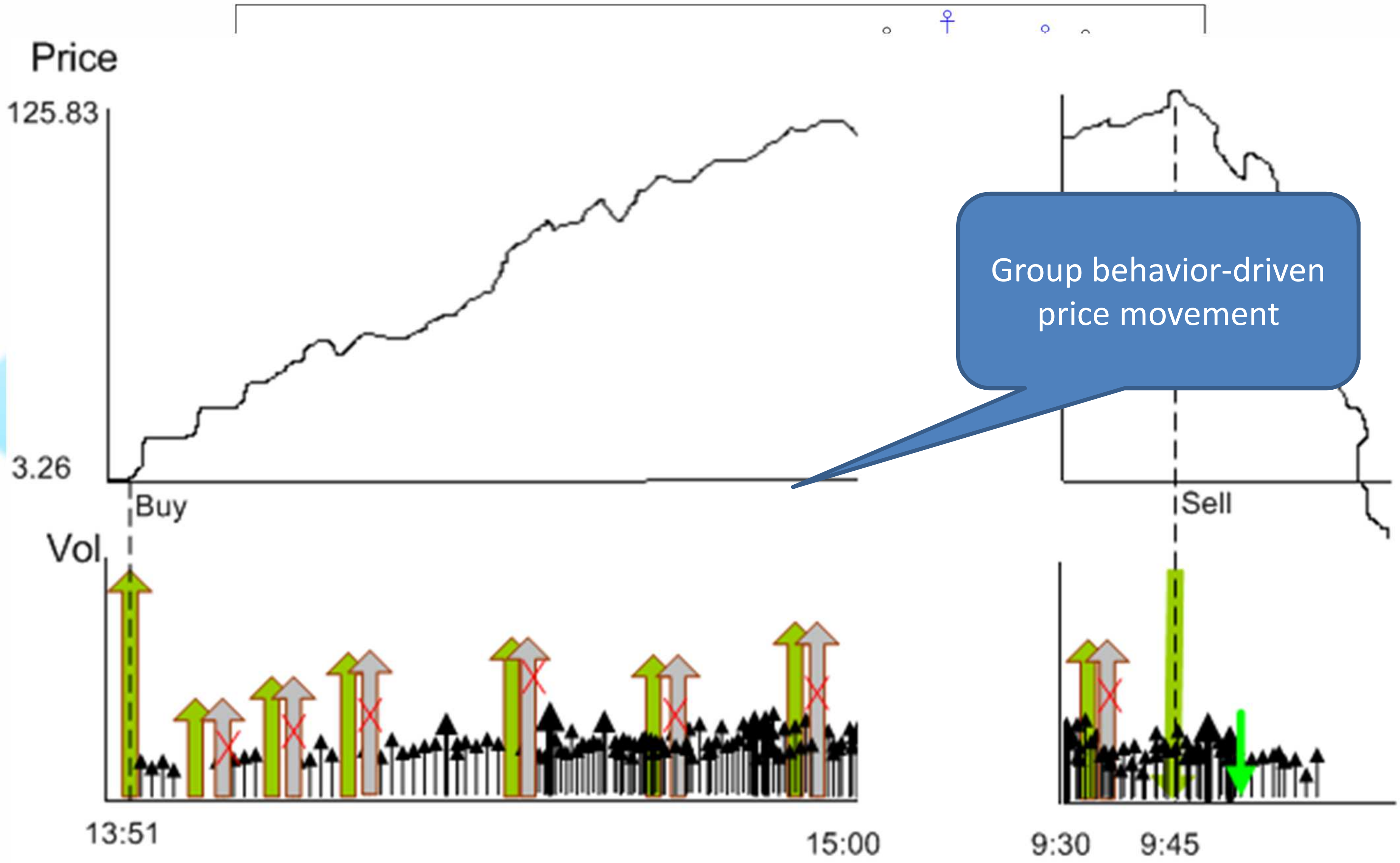


• Short-term manipulation behavior as cause



- Associated accounts





Price
125.83
3.26
Buy
Sell
Vol
13:51
15:00
9:30
9:45
Group behavior-driven price movement

Behavior is Ubiquitous

NSW Government Office of State Revenue

coateshire

HCF We're different.™

Australian Government | Centrelink giving you options

Australian Government Department of Human Services

Australian Government Australian Taxation Office

UNIVERSITY OF TECHNOLOGY SYDNEY

QANTAS Spirit of Australia

CREDIT SUISSE

Westpac

HUAWEI

AMP

炎黄集团 YAN HUANG GROUP

中国·洛阳 WWW.LY.GOV.CN

深圳证券交易所 SHENZHEN STOCK EXCHANGE

上海證券交易所 SHANGHAI STOCK EXCHANGE

citi

Microsoft Research

NOKIA Connecting People

IBM Research

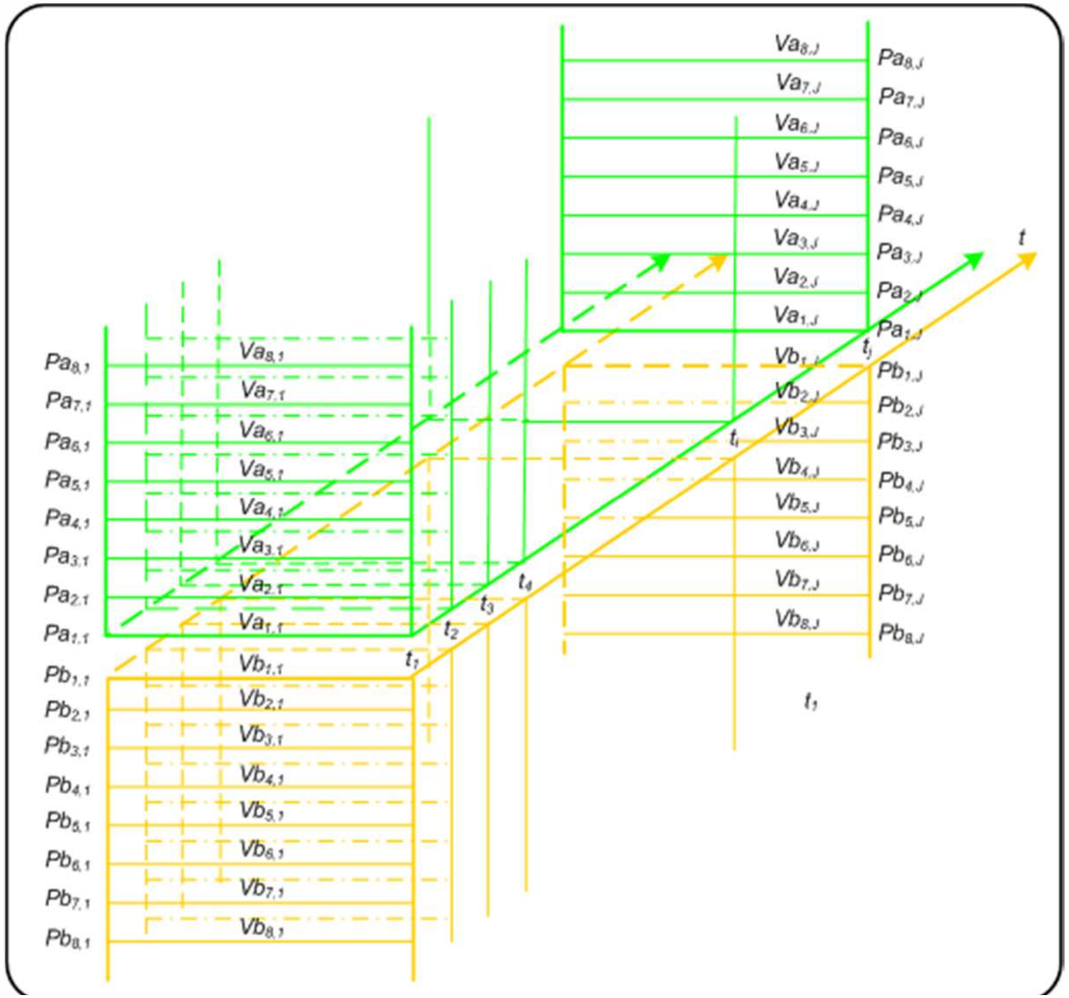
Capital Markets CRC Limited

sas THE POWER TO KNOW.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Argument 4: Need to consider behavior context

- Microstructure data



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

2. What is Behavior?

THE ADVANCED ANALYTICS INSTITUTE

What is Behavior ?

Activities

Human Beings

Actions
Operations
Events
Sequences

Context and Environment

What is behavior?

- An abstract behavior model
  - Demographics and circumstances of behavioral subjects and objects
  - Associates of a behavior may form into certain behavior sequences or network;
  - Social behavioral network consists of sequences of behaviors that are organized in terms of certain social relationships or norms.
  - Impact, costs, risk and trust of behavior/behavior network

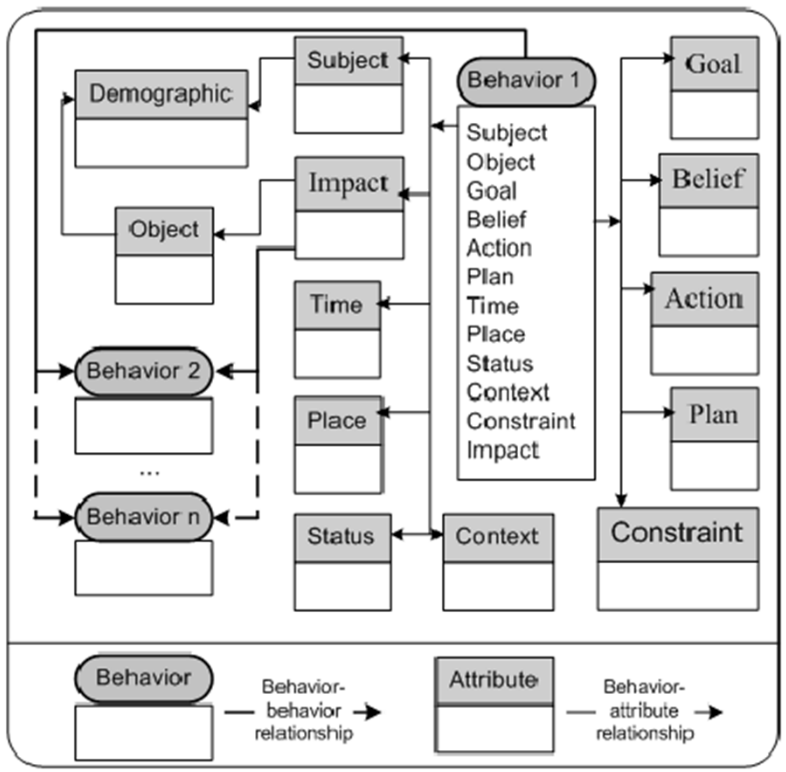


Figure 1. An Abstract Behavioral Model

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Abstract Behavior Model

Definition 1. *A* ***behavior*** *($\mathbb{B}$) is described as a four-ingredient tuple* $\mathbb{B} = (\mathscr{E}, \mathscr{O}, \mathscr{C}, \mathscr{R})$,

- *Actor* $\mathscr{E} = \langle \mathcal{SE}, \mathcal{OE} \rangle$ *is the entity that issues a behavior (subject, $\mathcal{SE}$) or on which a behavior is imposed (object, $\mathcal{OE}$).*

- *Operation* $\mathscr{O} = \langle \mathcal{OA}, \mathcal{SA} \rangle$ *is what an actor conducts in order to achieve certain goals; both objective ($\mathcal{OA}$) and subjective ($\mathcal{SA}$) attributes are associated with an operation. Objective attributes may include time, place, status and restraint; while subjective aspects may refer to action and its actor's belief and goal etc of the behavior and the behavior impact on business.*

- *Context* $\mathscr{C}$ *is the environment in which a behavior takes place.*

- *Relationship* $\mathscr{R} = \langle \theta(\cdot), \eta(\cdot) \rangle$ *is a tuple which reveals complex interactions within an actor's behaviors (named intra-coupled behaviors, represented by function $\theta(\cdot)$) and that between multiple behaviors of different actors (inter-coupled behaviors by relationship function $\eta(\cdot)$).*

- 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

$$\vec{\Gamma} = \{\vec{\gamma_1}, \vec{\gamma_2}, ..., \vec{\gamma_n}\}$$

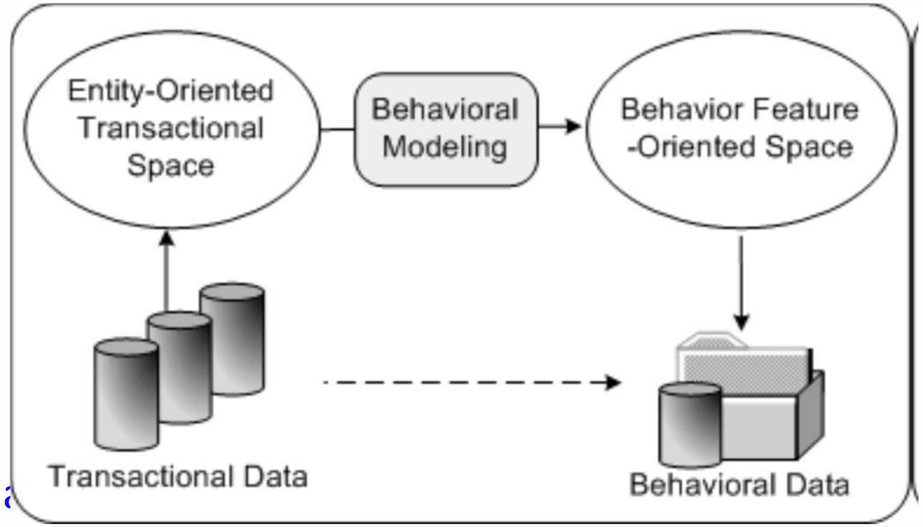
- Vector-oriented patterns

- 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 beha
- ➢ Behavioral data processing
- ➢ Behavioral data transformation

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

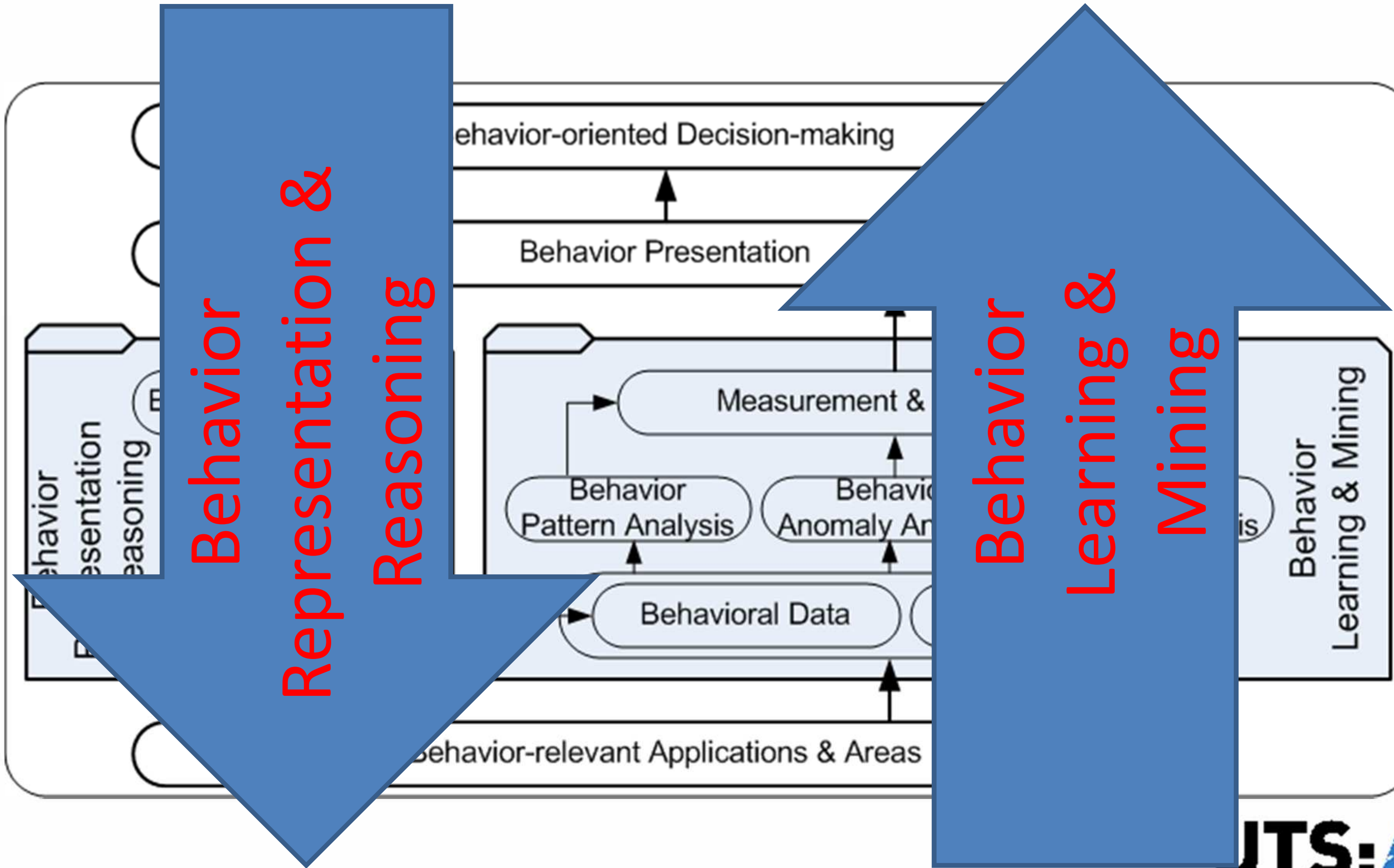
3. What is Behavior Informatics and Computing?

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

www.behaviorinformatics.org

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior informatics – Concept Map



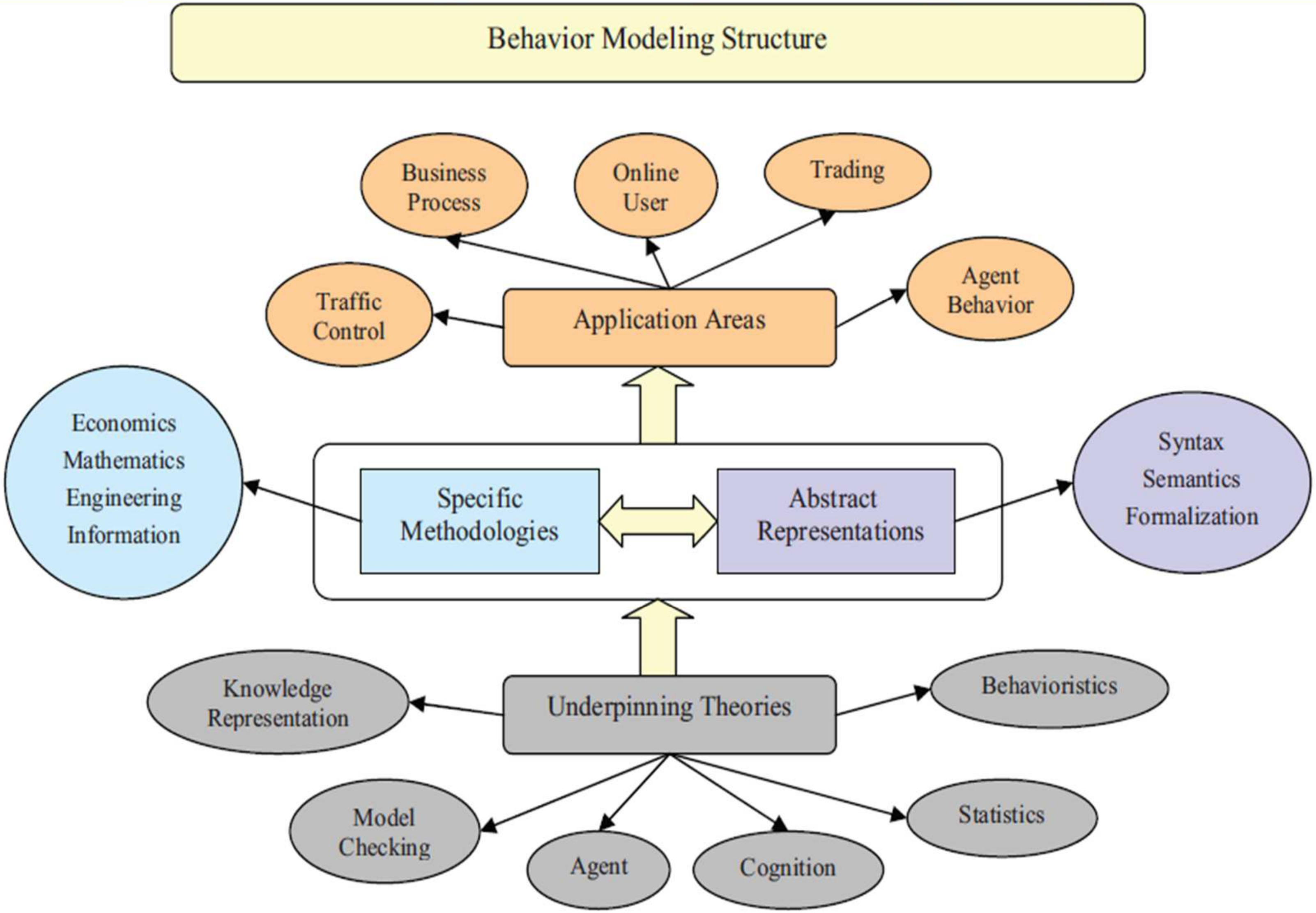
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

What is Behavior Computing

Behavior Modeling	Behavior Analysis	Behavior Mining
Develop modeling and representation methods to capture behavior characteristics and dynamics.	Propose effective techniques and tools for emergent areas and domains in analyzing behaviors.	Identify patterns in behavior entities and networks, such as detection, prediction and prevention of critical behavior.

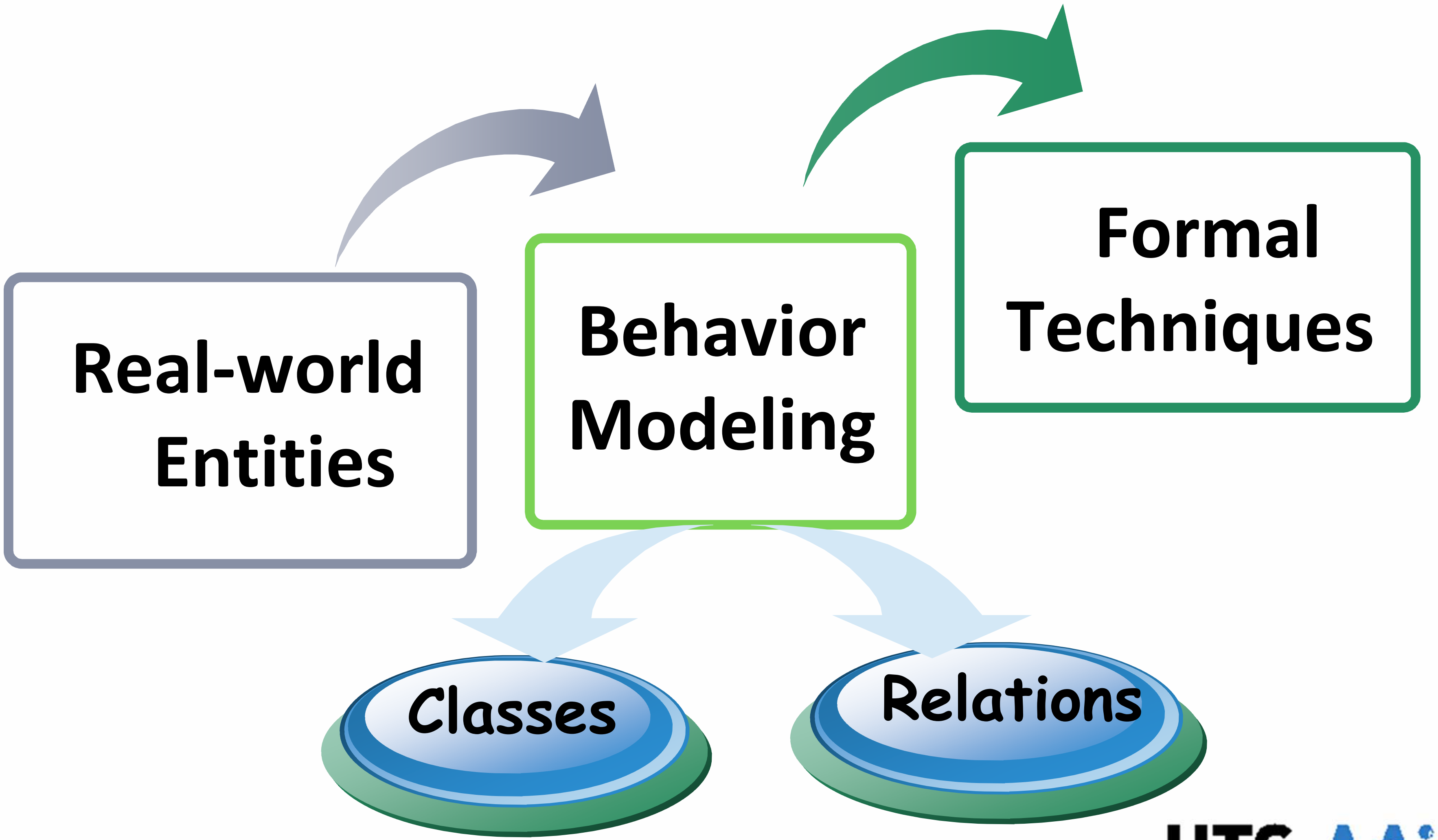
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Comprehensive Review



1. What is Behavior and Behavior Computing

Behavior Modeling

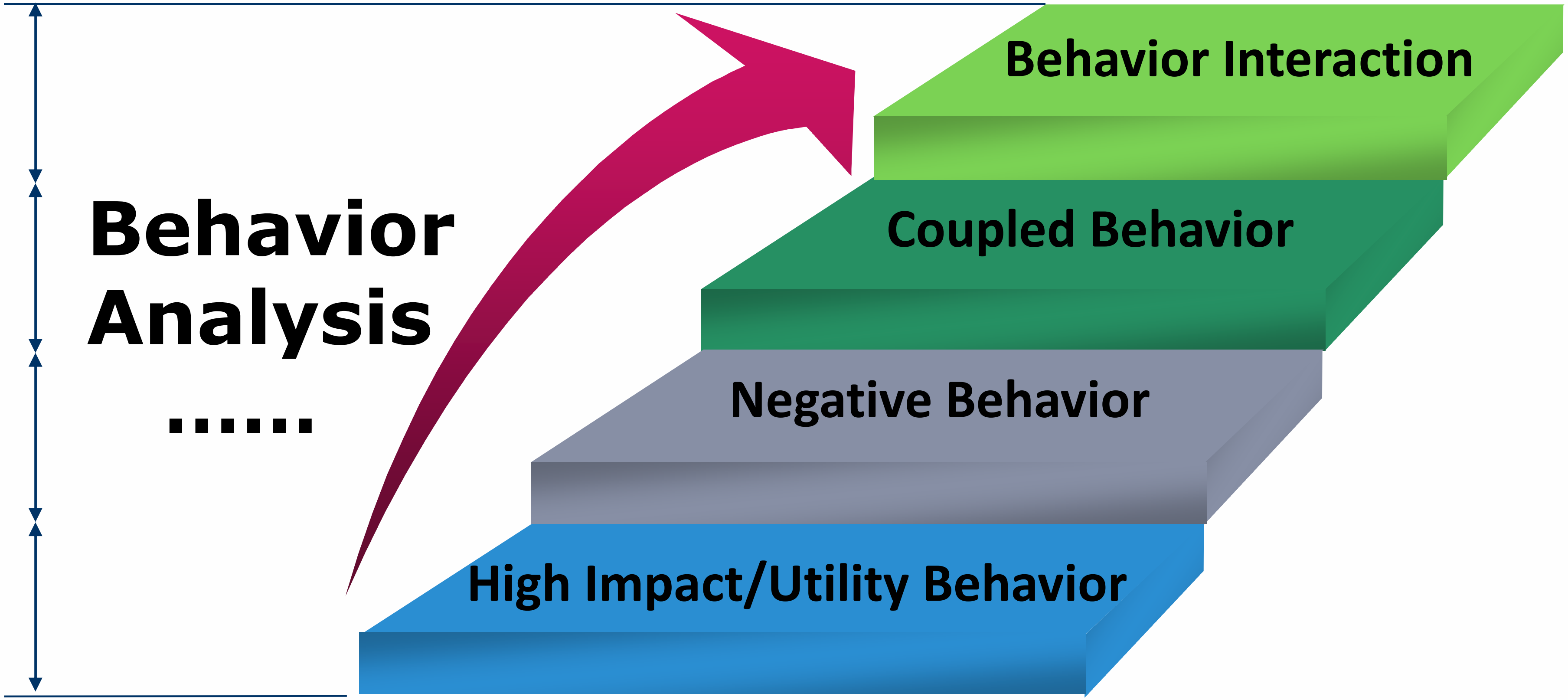


UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavioral representation

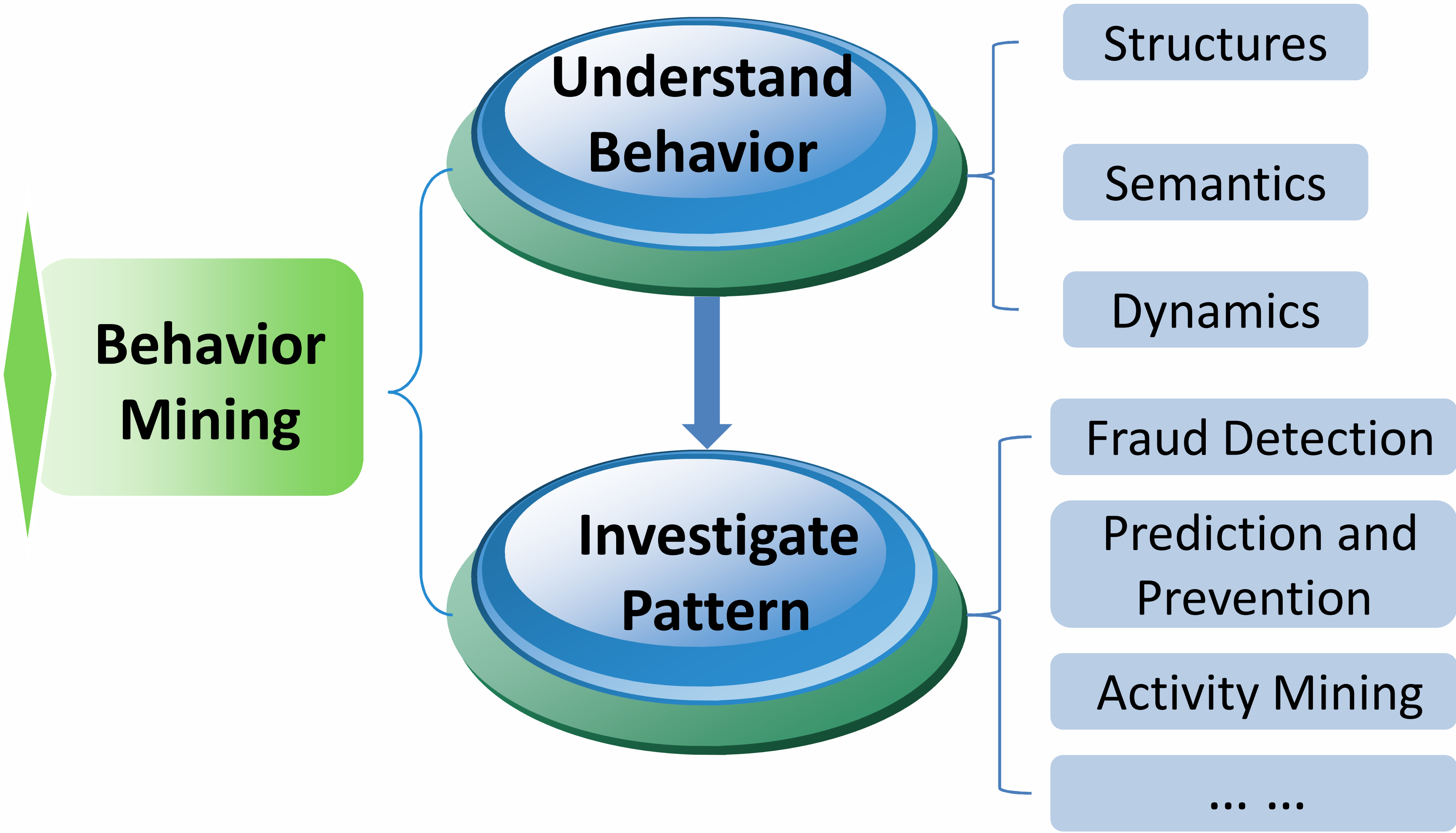
- (Behavior modeling)
  - describing behavioral elements
  - extracting syntactic and semantic relationships amongst the elements
  - presentation and construction of behavioral sequences and properties
  - unified mechanism for describing and presenting behavioral elements, properties, behavioral impact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

Behavior Analysis



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Mining



UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

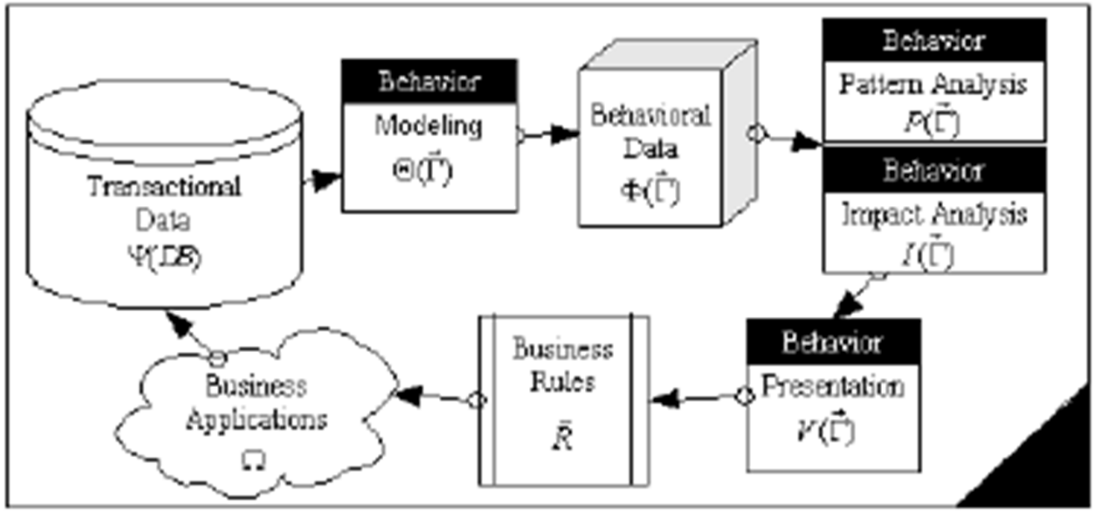
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

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

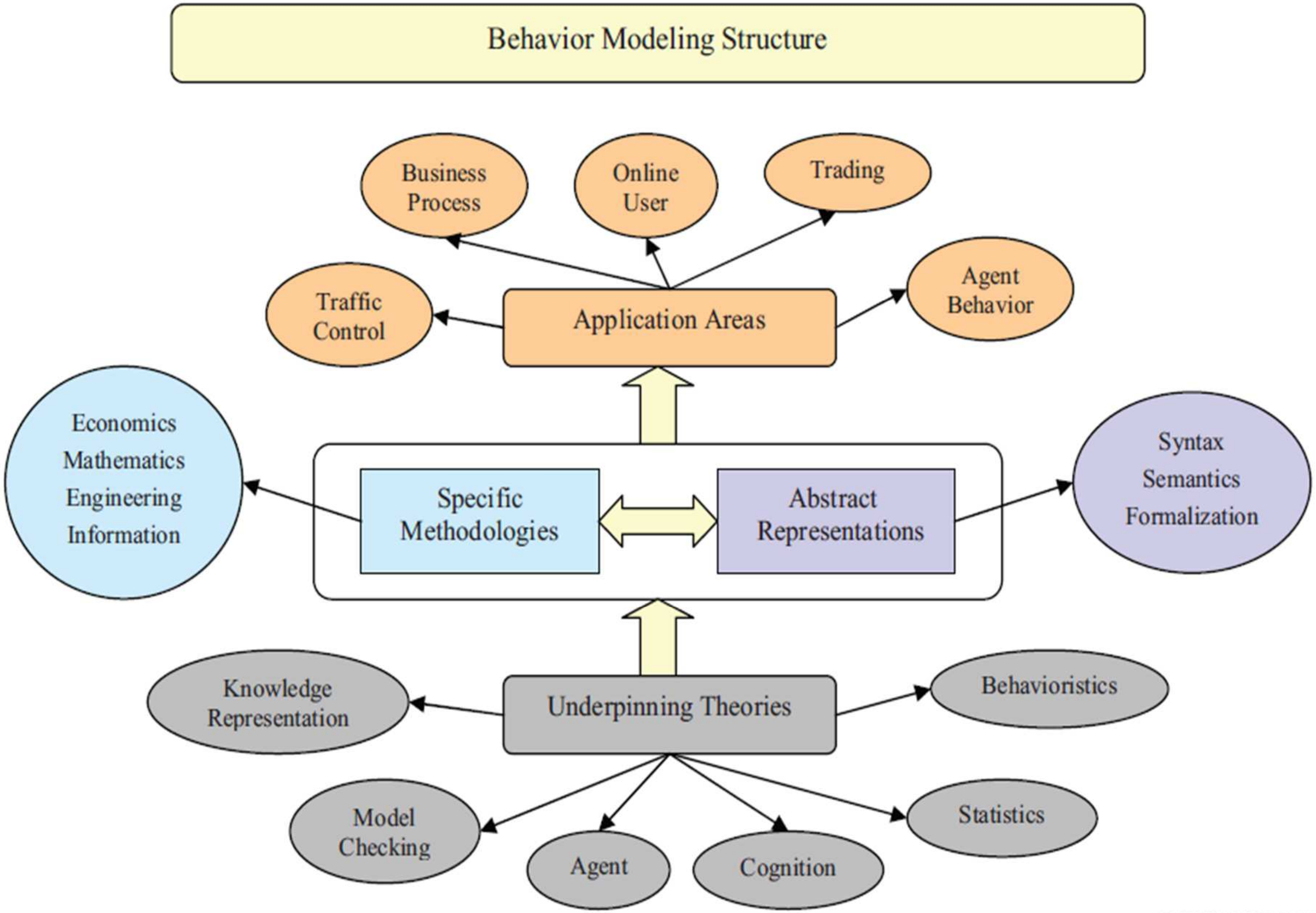
Step 3: Analyzing behavioral patterns $P\vec{\Gamma}$;
- Given behavior data $(\Phi(\vec{\Gamma})$;
- Design pattern mining method $\omega \in \Omega$;
- Employ the method ω on dataset $\Phi\vec{\Gamma}$;
- Extract behavior pattern set $\tilde{P}$;

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 $\tilde{R}$;

4. Related Work

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Related Work



THE ADVANCED ANALYTICS INSTITUTE

Related Work

Several **qualitative models** have been abstracted:

- belief-desire-intention model
- situation calculus
- human-machine interaction
- reasoning about action
- behavior composition
- action recognition and simulation
- action coordination and planning
- modeling systems rather than behaviors
- …

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Related Work

Several **quantitative models** have been proposed:

- user behavior modeling
- activity monitoring
- customer and consumer behavior analysis
- ontological engineering and semantic web
- sequence analysis
- reality mining
- activity mining
- multivariate time series
- coupled hidden Markov model
- …

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Research Limitations

1 Traditional behavior modeling that mainly relies on qualitative methods from behavior and social sciences often leads to ineffective and limited analysis in understanding social activities deeply and accurately.

2 Traditional behavior modeling approaches have too many styles and forms according to distinct situations. There is very limited research on formalizing the concept of behavior and its elements. There are no formal behavior representation models stated from a general perspective and providing a comprehensive understanding of behavior constitution.

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Research Limitations

3 Traditional behavior expressiveness is too weak to reveal that behavior plays the key role of an internal driving force for social activities.

4 The existing work often overlooks the checking of behavior modeling, which weakens the soundness and robustness of models built for complex behavior applications.

5 Complex coupling relationships between group behaviors are often ignored or only weakly addressed; few building blocks are available to explicitly model complex interactions between group behaviors.

Research Issues

- **Qualitative Reasoning and Verification**

  With the formal representation of coupled behaviors, the qualitative analytics to address the task of behavior reasoning and verification is in great demand.

- **Quantitative Leaning and Evaluation**

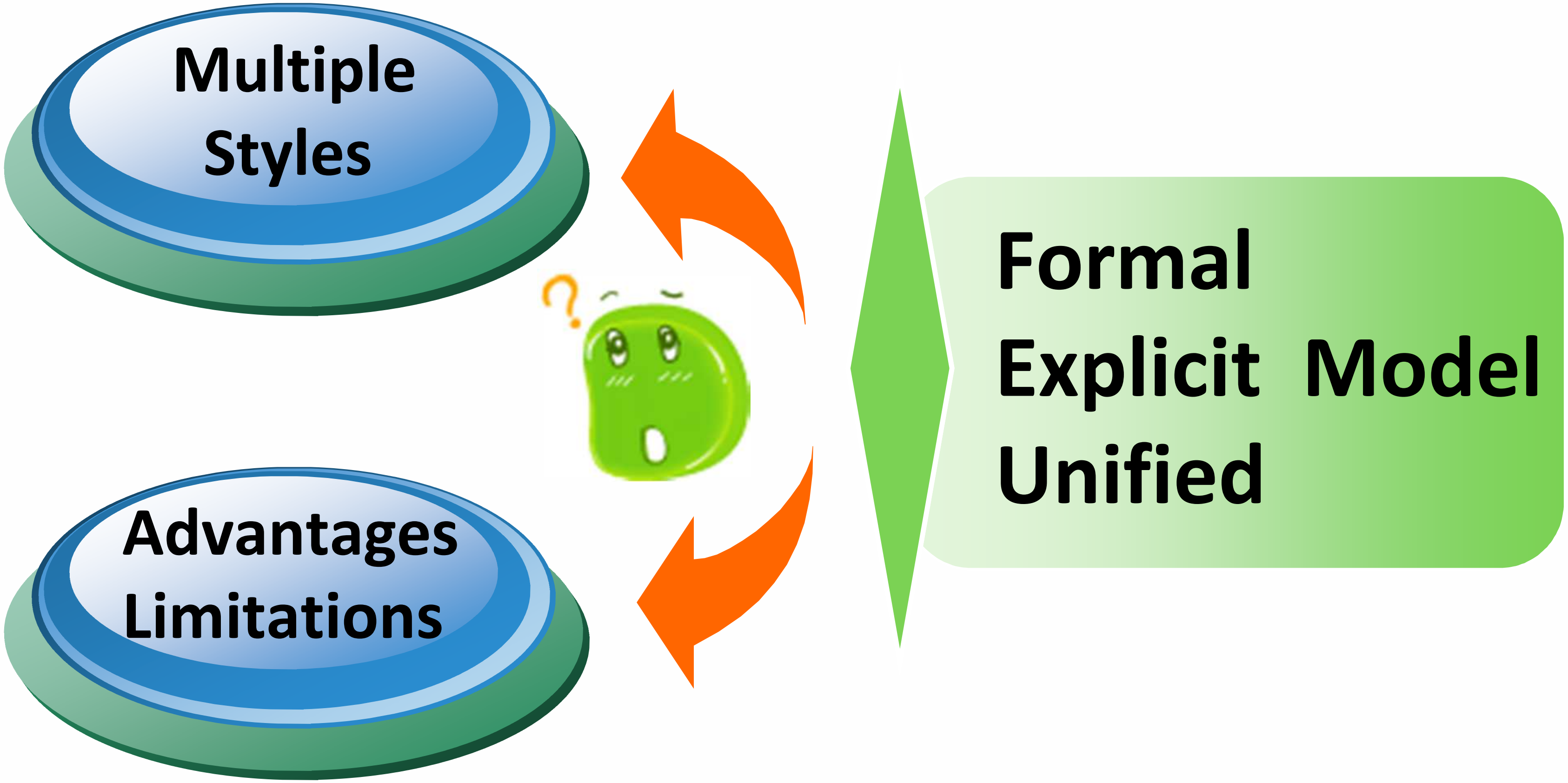
  The quantitative research to target behavior learning and evaluation must be focused on.

- **Integrated Understanding of Behavior Algebra**

  An appropriate way could be chosen to integrate these two studies to obtain an integrated understanding of the implicit complex behaviors

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Research Question



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

5. Behavior Modeling and Representation

UTS/AAI Technique Report 2011

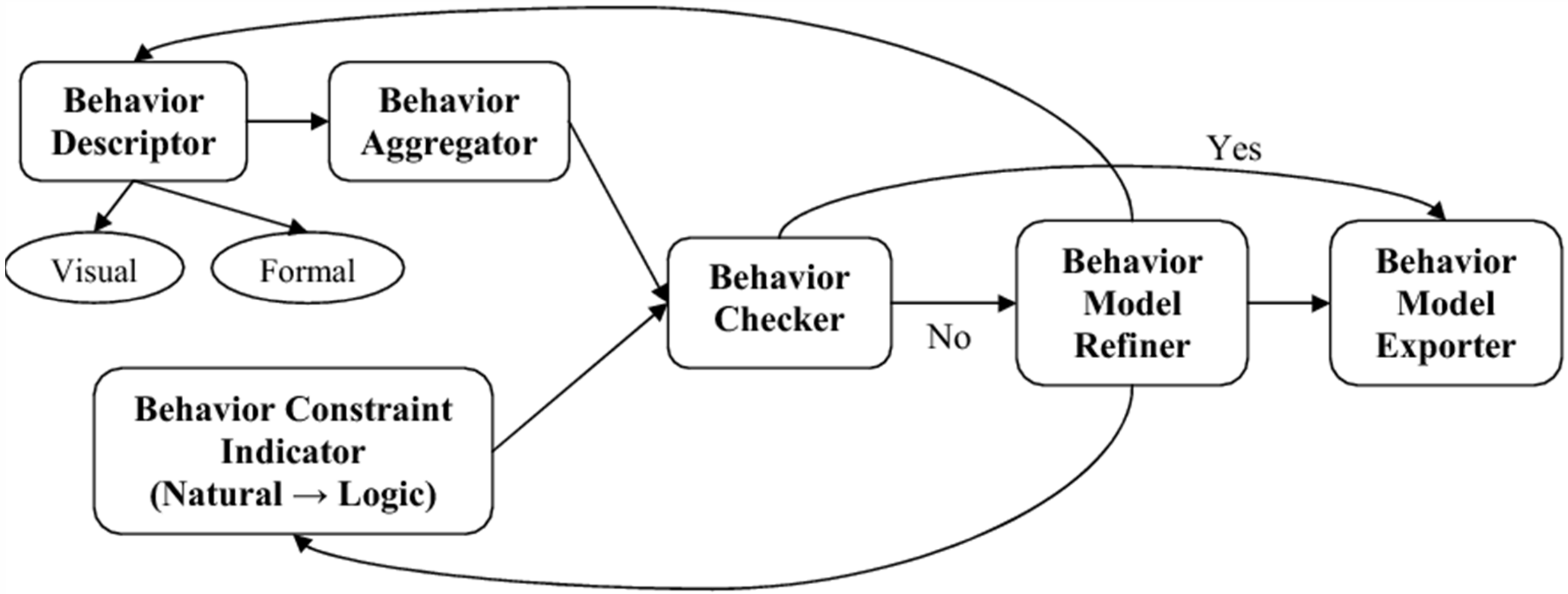
Formalization and Verification of Group Behavior Interactions

Can Wang, Longbing Cao

University of Technology, Sydney, Australia

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

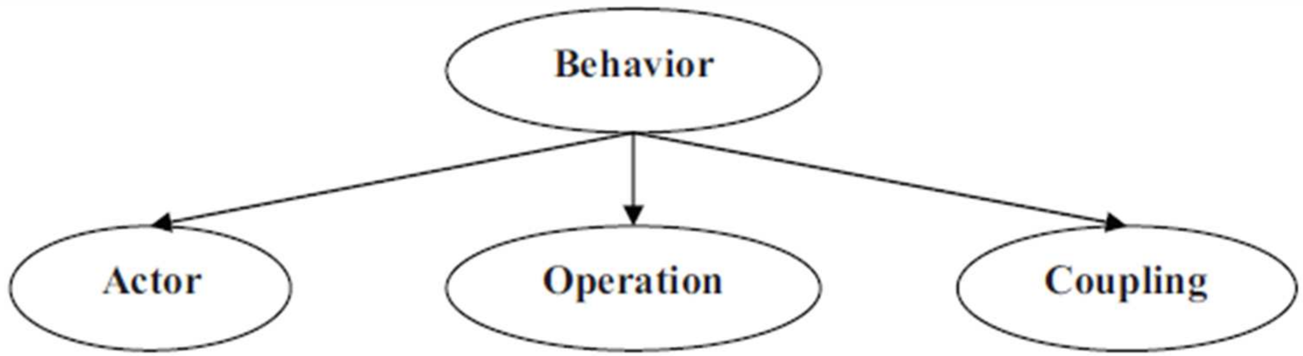
Behavior Modeling and Checking Framework



Ontology-based Behavior Modeling and Checking

Behavior Visual Descriptor

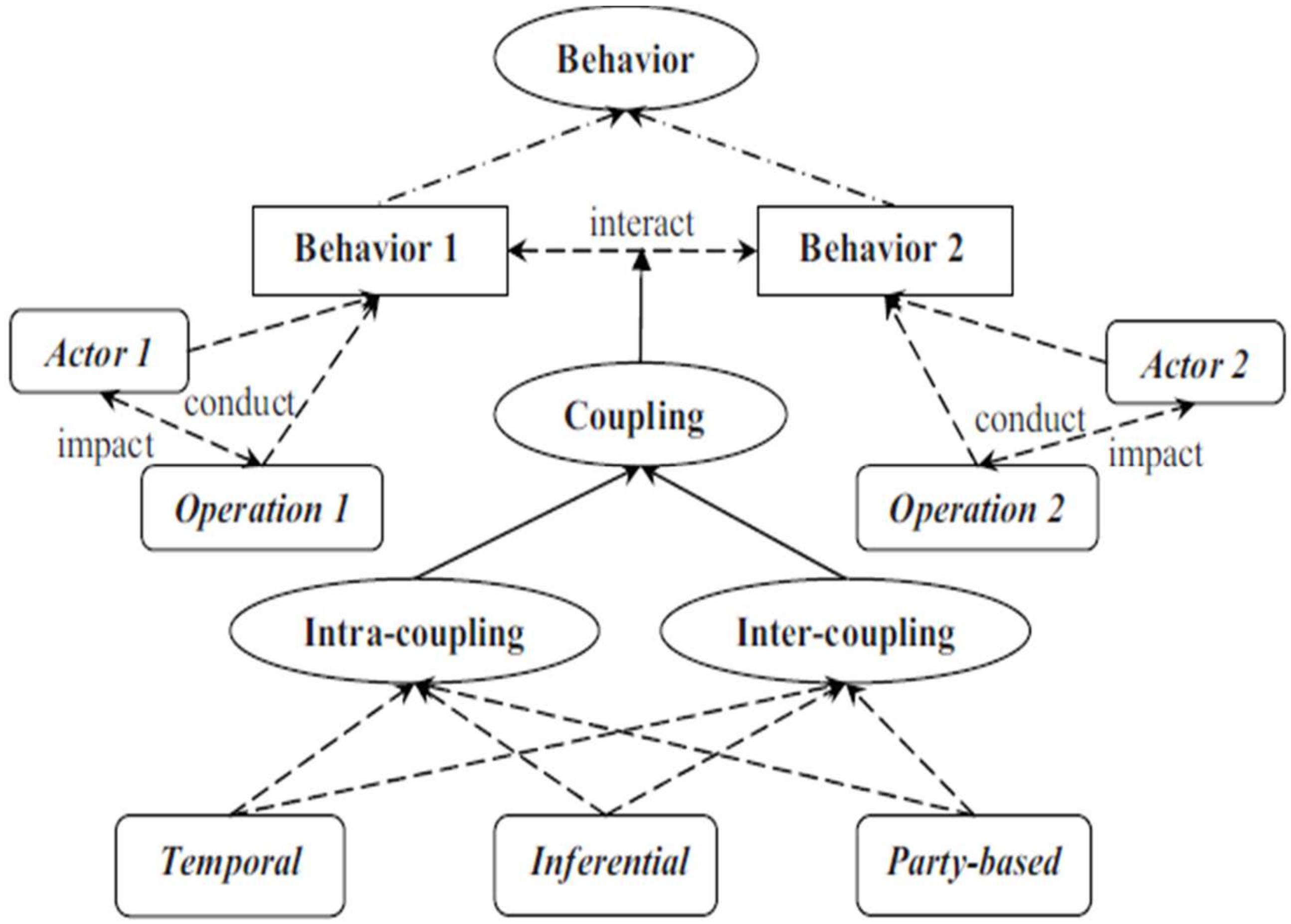
- **Actor**: refers to the subject(s) or object(s) of a behavior, for example, organizations, departments, systems, agents and people involved in an activity or activity sequence.
- **Operation**: represents activities, actions or events in a behavior or behavior sequence.
- **Coupling**: refers to the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



UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

3. Behavior Model/Representation

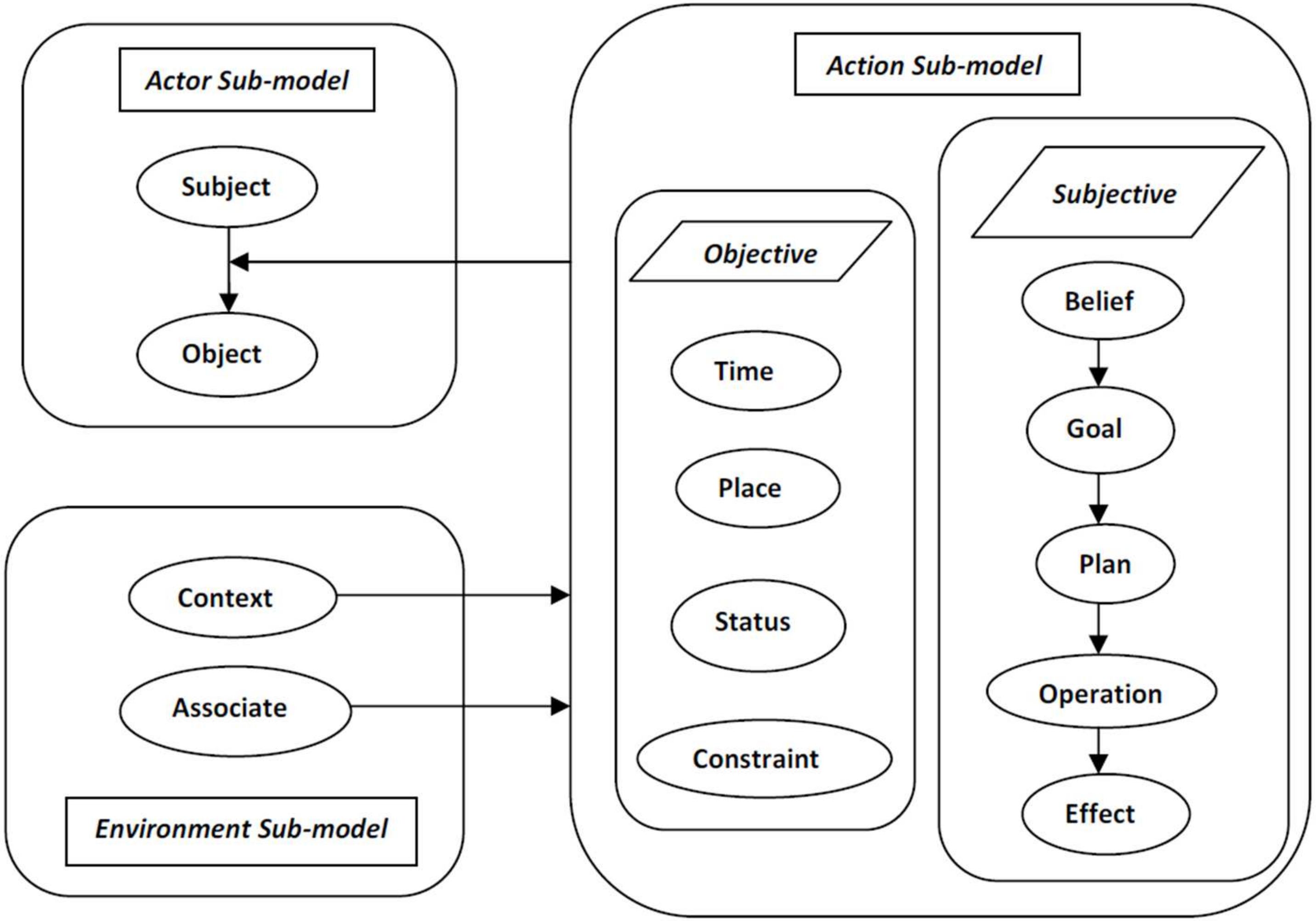
Behavior Visual Descriptor



- **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.

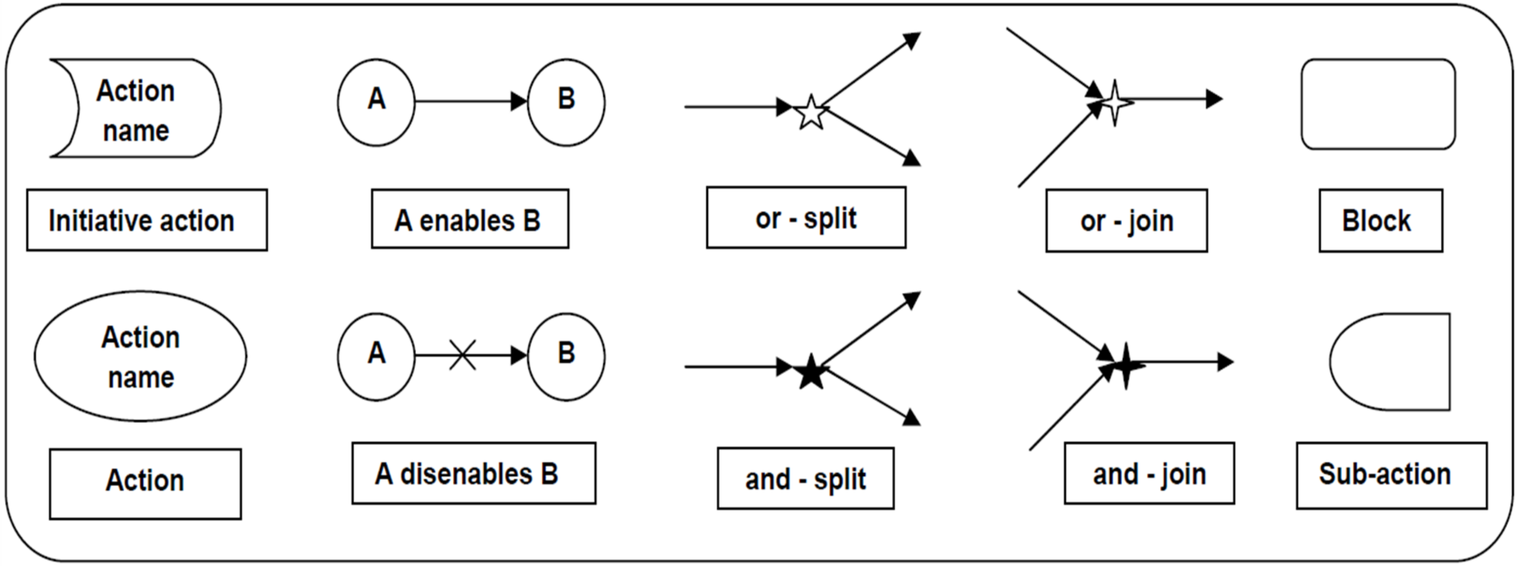
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Overall Single Behavior Model



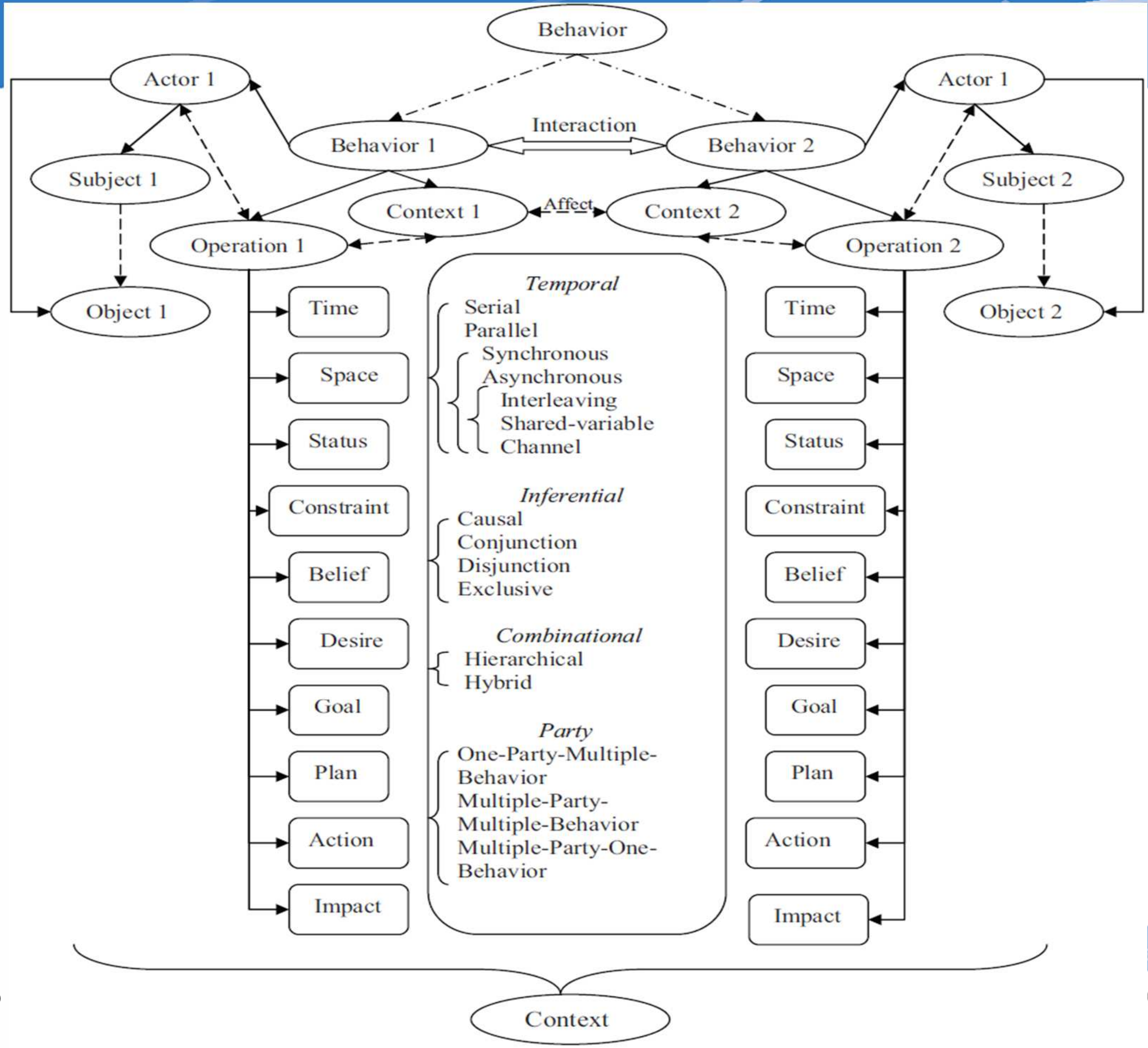
THE ADVANCED ANALYTICS INSTITUTE

Relationship Sub-model



Relationship	*enable*	*disenable*	*or-split*	*and-split*	*or-join*	*and-join*
Logic Form	$a \to b$	$\neg(a \to b)$	$a \to (b \vee c)$	$a \to (b \wedge c)$	$(a \vee b) \to c$	$(a \wedge b) \to c$

Relationships between Agent Behaviors



Coupling Relationships

Coupling Relationships

Perspectives

Temporal

- *Serial Coupling*
- *Parallel coupling*
  - *Synchronous relationship*
  - *Asynchronous coupling*
    - *Interleaving*
    - *Shared-variable*
    - *Channel system*

Inferential

- *Causal Coupling*
- *Conjunction Coupling*
- *Disjunction Coupling*
- *Exclusive Coupling*

Party-based

- *One-Party-Multiple-Operation*
- *Multiple-Party-One-Operation*
- *Multiple-Party-Multiple-Operation*

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Temporal Coupling

- **Serial coupling**, denoted by $\{B_1;B_2\}$, showing the situation in which behavior B_2 follows behavior B_1.

- **Parallel coupling**, by which behaviors happen in varying concurrent manners, including synchronous coupling and asynchronous coupling.

  – **Synchronous relationship**, denoted by $\{B_1 \| B_2\}$, indicating that B1 and B2 present at the same time based on certain communication protocols.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Temporal Coupling

– **Asynchronous coupling**, showing that two behaviors B_1 and B_2 interact with each other at different time points.

∗ **Interleaving**, denoted by $\{B_1 : B_2\}$, representing the involvement of independent complex behaviors by nondeterministic choice (independently).

∗ **Shared-variable**, denoted by $\{B_1 \mid\mid\mid B_2\}$, signifying that the relevant behaviors have variables in common.

∗ **Channel system**, denoted by $\{B_1 \mid B_2\}$, is a parallel system in which complex behaviors communicate via a channel, for instance, first-in and first-out buffers.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Inferential Coupling

- **Causal coupling**, represented as $\{B_1 \rightarrow B_2\}$, meaning that behavior B_1 causes behavior B_2.

IMPLY

- **Conjunction coupling**, $\{B_1 \wedge B_2\}$, specifying that B_1 and B_2 take place together.

AND

- **Disjunction coupling**, $\{B_1 \vee B_2\}$, by which at least one of the associated behaviors must happen.

OR

- **Exclusive coupling**, $\{B_1 \oplus B_2\}$, indicating that if B_1 happens, B_2 will not happen, and vice versa.

XOR

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Party-based Coupling

- **One-Party-Multiple-Operation,** represented as $\{(B_1,B_2)^{[A_1]}\}$, depicts that distinct behaviors B_1 and B_2 are performed by the same actor A_1.
- **Multiple-Party-One-Operation,** shown as $\{(B_1)^{[A_1A_2]}\}$, represents that multiple actors A_1 and A_2 implement the same behavior B_1 to achieve their own intentions.
- **Multiple-Party-Multiple-Operation,** presented as $\{(B_1,B_2)^{[A_1A_2]}\}$, describes that different behaviors B_1 and B_2 are carried out by distinct actors A_1 and A_2.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Formal Descriptor

Definition 1 (Behavior): A behavior $\mathbb{B}$ is described as a three-ingredient tuple $\mathbb{B} = (\mathscr{A}, \mathscr{O}, \mathscr{C})$, where:

- Actor $\mathscr{A}$ is the entity that issues a behavior or on which a behavior is imposed.
- Operation $\mathscr{O}$ is what an actor conducts in order to achieve certain goals.
- Coupling $\mathscr{C} =< \theta(\cdot), \eta(\cdot) >$ is a tuple that reveals complex interactions including intra-coupling ($\theta(\cdot)$) and inter-coupling ($\eta(\cdot)$).

For instance, in a stock market, a behavior can be represented as "an investor places a buy order". The involved actor is the "investor" himself or herself, the operation is the transaction of "buy". The third component coupling exposes the intra-relationship between this behavior and this investor's sell order on the other day, together with the inter-relationship between this behavior and another investor's buy order on the same day.

Behavior Formal Descriptor

We tackle the coupled behaviors from either one or different actors, denoted as intra-coupling and inter-coupling, respectively.

Behavior Feature Matrix

intra-coupling

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

inter-coupling

An actor $\mathscr{A}_i$ undertakes J_i operations $\{\mathscr{O}_{i1}, \mathscr{O}_{i2}, \ldots, \mathscr{O}_{iJ_i}\}$

I actors: $\{\mathscr{A}_1, \mathscr{A}_2, \ldots, \mathscr{A}_I\}$

Intra-Coupling

The intra-coupling reveals the complex couplings within an actor's distinct behaviors.

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

$$\mathbb{B}_{i\cdot}^{\theta} ::= \mathbb{B}_{i\cdot}(\mathscr{A}, \mathscr{O}, \theta)| \sum_{j=1}^{J_{max}} \theta_j(\mathbb{B}) \odot \mathbb{B}_{ij}, \qquad \text{(IV.2)}$$

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

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

For instance, in the stock market, the investor will place a sell order at some time after buying his or her desired instrument due to a great rise in the trading price. This is, to some extent, one way to express how these two behaviors are intra-coupled with each other.

Inter-Coupling

The inter-coupling embodies the way multiple behaviors of different actors interact.

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

$$\mathbb{B}^{\eta}_{.j} ::= \mathbb{B}_{.j}(\mathscr{A}, \mathscr{O}, \eta) | \sum_{i=1}^{I} \eta_i(\mathbb{B}) \odot \mathbb{B}_{ij}, \qquad \text{(IV.3)}$$

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

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

For instance, a trading happens successfully only when an investor sells the instrument at the same price as the other investor buys this instrument. This is another example of how to trigger the interactions between inter-coupled behaviors.

Coupling

In practice, behaviors may interact with one another in both ways of intra-coupling and inter-coupling.

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 $h(\theta(\mathbb{B}), \eta(\mathbb{B}))$, where $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$

$$\mathbb{B}_c = (\mathbb{B}^{\theta}_{i_1 j_1})^{\eta} * (\mathbb{B}^{\theta}_{i_2 j_2})^{\eta} ::= \mathbb{B}_{ij}(\mathscr{A}, \mathscr{O}, \mathscr{C}) | \sum_{i_1, i_2 = 1}^{I} \sum_{j_1, j_2 = 1}^{J_{max}} h(\theta_{j_1 j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B})) \odot (\mathbb{B}_{i_1 j_1} \mathbb{B}_{i_2 j_2}), \quad \text{(IV.4)}$$

where $h(\theta_{j_1, j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B}))$ is the coupling function denoting the corresponding relationships between $\mathbb{B}_{i_1 j_1}$ and $\mathbb{B}_{i_2 j_2}$, $\sum_{i_1, i_2 = 1}^{I} \sum_{j_1, j_2 = 1}^{J_{max}} \odot$ means the subsequent behaviors of $\mathbb{B}$ are $\mathbb{B}_{i_1 j_1}$ coupled with $h(\theta_{j_1}(\mathbb{B}), \eta_{i_1}(\mathbb{B}))$, $\mathbb{B}_{i_2 j_2}$ with $h(\theta_{j_2}(\mathbb{B}), \eta_{i_2}(\mathbb{B}))$, and so on.

For instance, we consider both the successful trading between investor A_1 (buy) and investor A_2 (sell), and then the selling behavior conducted by A_1 after he or she has bought the instrument at a relative low price.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Aggregator

We conduct behavior aggregations to interpret the interactions of intra-coupled and inter-coupled behaviors. The outcomes of the behavior aggregations form the basis of behavior verification.

Three types of aggregations

- Intra-Coupled Aggregation — function θ_j
- Inter-Coupled Aggregation — function η_i
- Combined Aggregation — function $h(\theta_{j_1 j_2}, \eta_{i_1 i_2})$

Intra-Coupled Aggregation

For the behaviors conducted by the same actor, we interpret the behavior dynamics in terms of a transition system (TS).

TS: Directed Graphs

Nodes: System States

A state describes the behavior status at a certain moment of system dynamics.

Edges: Model Transitions

State changes of a system.

TS is often used in computer science for modeling the behavior dynamics of a system.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Intra-Coupled Aggregation

In particular, the TS interpretation of the intra-coupled behaviors $\mathbb{B}_{i.}^{\theta}$ for actor $\mathscr{A}_i$ is the tuple (St; Act;→; In), where θ_j is the intra-coupling function.

- $\mathrm{St} = \{\theta_j(\mathbb{B})\}$ is a set of states.
- $\mathrm{Act} = \{\mathscr{O}\}$ is a set of actions or operations.
- $\theta_{\mathrm{j}}(\mathbb{B}) \xrightarrow{\mathscr{O}} \theta_{\mathrm{j}+1}(\mathbb{B})$ is a transition relation.
- $\mathrm{In} = \{\theta_0(\mathbb{B})\}$ is a set of initial states.

Every actor is interpreted by an independent transition system, we regard an operation as a corresponding action in TS; and the intra-coupling function θ_j, which links intra-coupled behaviors, represents the associated states in TS to connect all the involved operations.

Inter-coupled Aggregation

Apart from the intra-coupled behaviors, inter-coupling $\mathbb{B}^{\eta}_{\cdot j}$ refers to interactions between operations by different actors.

Definition 5 (Inter-coupling Operators): The behavior inter-couplings are essentially the various interactions among multiple behaviors. Let $\mathbb{B}_1$ and $\mathbb{B}_2$ be two behaviors, then the inter-coupling function $\eta_i(\mathbb{B})$ is defined as:

$$\eta_i(\mathbb{B}) ::= \mathbb{B}_1;\mathbb{B}_2 \mid \mathbb{B}_1 \| \mathbb{B}_2 \mid \mathbb{B}_1 : \mathbb{B}_2 \mid \mathbb{B}_1 ||| \mathbb{B}_2 \mid \mathbb{B}_1 | \mathbb{B}_2 \mid \mathbb{B}_1 \rightarrow \mathbb{B}_2 | \ \mathbb{B}_1 \wedge \mathbb{B}_2 \mid \mathbb{B}_1 \vee \mathbb{B}_2 \mid \mathbb{B}_1 \oplus \mathbb{B}_2 \mid f(\mathbb{B}_1)^{[\mathscr{A}_1]}. \qquad \text{(V.1)}$$

Temporal Operators

Inferential Operators

Part-based Operators

Combined Aggregation

With the intra-coupled and inter-coupled interactions defined, we develop the combined aggregation of coupled behaviors to model complex behavior-oriented applications.

1. *Behavior Combination*
2. *Rule Reduction*
3. *TS Conversion*

$$h(\theta_{j_1 j_2}(\mathbb{B}), \eta_{i_1 i_2}(\mathbb{B}))$$

Behavior Combination

First, we consider the extension of behavior sequences towards hierarchical and hybrid combinations, in which behaviors are associated in a hierarchical structure that consists of different relationships.

$f(\mathbb{B}_1, g(\mathbb{B}_2,\ \mathbb{B}_3)) = \{\mathbb{B}_1; (\mathbb{B}_2 \| \mathbb{B}_3)\}$ → Behavior $\mathbb{B}_1$ is followed by the handshaking (g(·)) of $\mathbb{B}_2$ and $\mathbb{B}_3$

$\{f(\mathbb{B}_1).g(\mathbb{B}_2)\}$ → The concatenation of $\mathbb{B}_1$ and $\mathbb{B}_2$

$\{f\ (\mathbb{B}_1)^*\}$ → Finite repetition of $\mathbb{B}_1$

$\{f(\mathbb{B}_1)^\omega\}$ → Infinite iteration of $\mathbb{B}_1$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Rule Reduction

Second, interaction rules (IR) are induced to support appropriate combinational reduction of multiple coupling relationships.

Definition 6 (Interaction Rule): An interaction rule

$$IR : \mathbb{B}_1 \times \cdots \times \mathbb{B}_n \to \frac{f(\mathbb{B}_1, \cdots, \mathbb{B}_n)}{g(\mathbb{B}_1, \cdots, \mathbb{B}_n)} \qquad \text{(V.3)}$$

is the combinational equivalence and reduction about the coupling relationships among behaviors $\mathbb{B}_i (1 \leq i \leq n)$, where $f(\cdot)$ and $g(\cdot)$ are two coupling expressions for the involved behaviors.

In the above SOS-notation based interaction rule, if the numerator formula holds, then the denominator part holds as well. With interaction rules, we can perform reasoning about behaviors to simplify and conclude critical rules.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Rule Reduction

For instance, four interaction rules are induced as follows (where $*$; $*1$; $*2$ are the coupling operators):

Basic Equivalence Rules

$$IR_1 : \frac{(\mathbb{B}_1 * \mathbb{B}_2) * \mathbb{B}_3}{\mathbb{B}_1 * (\mathbb{B}_2 * \mathbb{B}_3)},$$

Associative Law

$$IR_2 : \frac{(\mathbb{B}_1 *_1 \mathbb{B}_2) *_2 (\mathbb{B}_3 *_1 \mathbb{B}_2)}{(\mathbb{B}_1 *_2 \mathbb{B}_3) *_1 \mathbb{B}_2},$$

Distributive Law

$$IR_3 : \frac{(\mathbb{B}_1 * \mathbb{B}_2) * (\mathbb{B}_2 * \mathbb{B}_3)}{\mathbb{B}_1 * \mathbb{B}_2 * \mathbb{B}_3},$$

Absorption Law

$$IR_4 : \frac{(\mathbb{B}_1 *_1 \mathbb{B}_2) *_2 (\mathbb{B}_1 *_1 \mathbb{B}_3)}{\mathbb{B}_1 *_1 (\mathbb{B}_2 *_2 \mathbb{B}_3)}.$$

Distributive Law

Induction Reduction

Behavior Aggregation and Reasoning

TS Conversion

Finally, concurrent transition systems (TSs) are constructed to specify complex interactions by utilizing temporal, inferential, and party-based couplings to describe, combine and aggregate the coupling relationships.

The relationships among TSs are concerned since complex behaviors are represented as TSs. Assume that there are n complex behaviors (TSs) associated with one another in terms of different coupling relationships.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

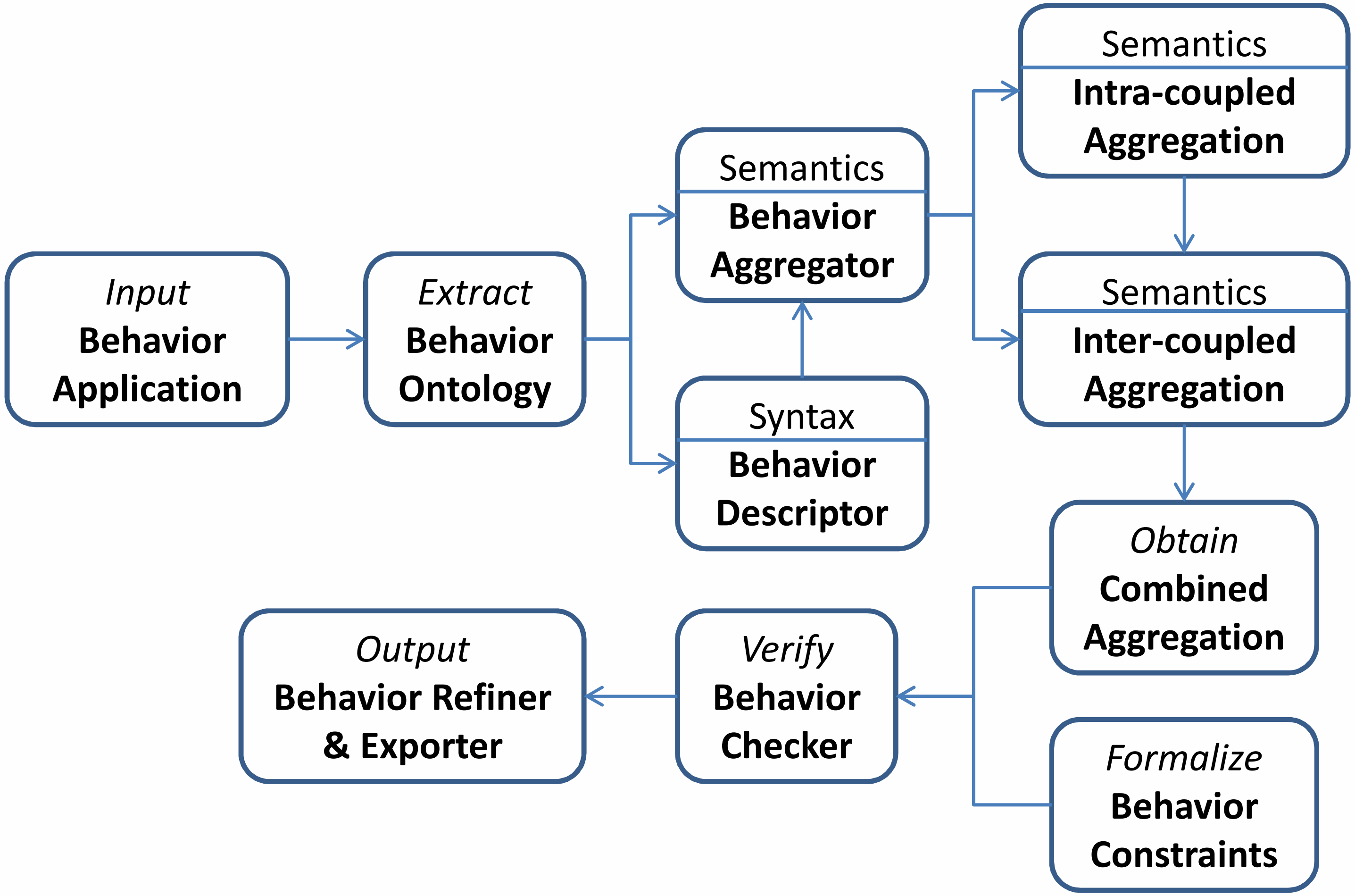
TS Conversion

- *Serial Coupling*: $TS_1; TS_2; \cdots; TS_n$
- *Synchronous Coupling*: $TS_1 \parallel TS_2 \parallel \cdots \parallel TS_n$
- *Interleaving Coupling*: $TS_1 : TS_2 : \cdots : TS_n$
- *Shared-variable Coupling*: $TS_1 ||| TS_2 ||| \cdots ||| TS_n$
- *Channel System Coupling*: $TS_1 \mid TS_2 \mid \cdots \mid TS_n$
- *Causal Coupling*: $TS_1 \rightarrow TS_2$
- *Conjunction Coupling*: $TS_1 \wedge TS_2$
- *Disjunction Coupling*: $TS_1 \vee TS_2$
- *Exclusive Coupling*: $TS_1 \oplus TS_2$
- *Hierarchical Coupling*: $f(g(TS_1, TS_2, \cdots, TS_n))$
- *Hybrid Coupling*: $f(TS_1).g(TS_2)$, $f(TS_1)^*$, $(TS_1)^\omega$
- *OPMO Coupling*: $f(TS_1, TS_2, \cdots, TS_n)^{[A_1]}$
- *MPOO Coupling*: $f(TS_1)^{[A_1 A_2 \cdots, A_n]}$
- *MPMO Coupling*: $f(TS_1, TS_2, \cdots, TS_n)^{[A_1 A_2 \cdots A_n]}$

The combined aggregation of coupled behaviors reflects the semantics of behavior coupling and interaction.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Group Behavior Representation and Verification



UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Behavior Constraint Indicator

In order to improve the quality of the behavior model, a simulation can be conducted prior to the behavior checking. For verification purposes, the behavior model under consideration needs to be accompanied by a relevant constraint specification that is to be verified.

Constraints, i.e., prior simulations, can be used effectively to get rid of the simpler categories of modeling errors. To make a rigorous verification possible, constraints should be described in a precise and unambiguous manner. This is done through a constraint specification language.

For instance, a business constraint in stock markets is that investors are not allowed to make transactions after trading hours.

Behavior Constraint Indicator

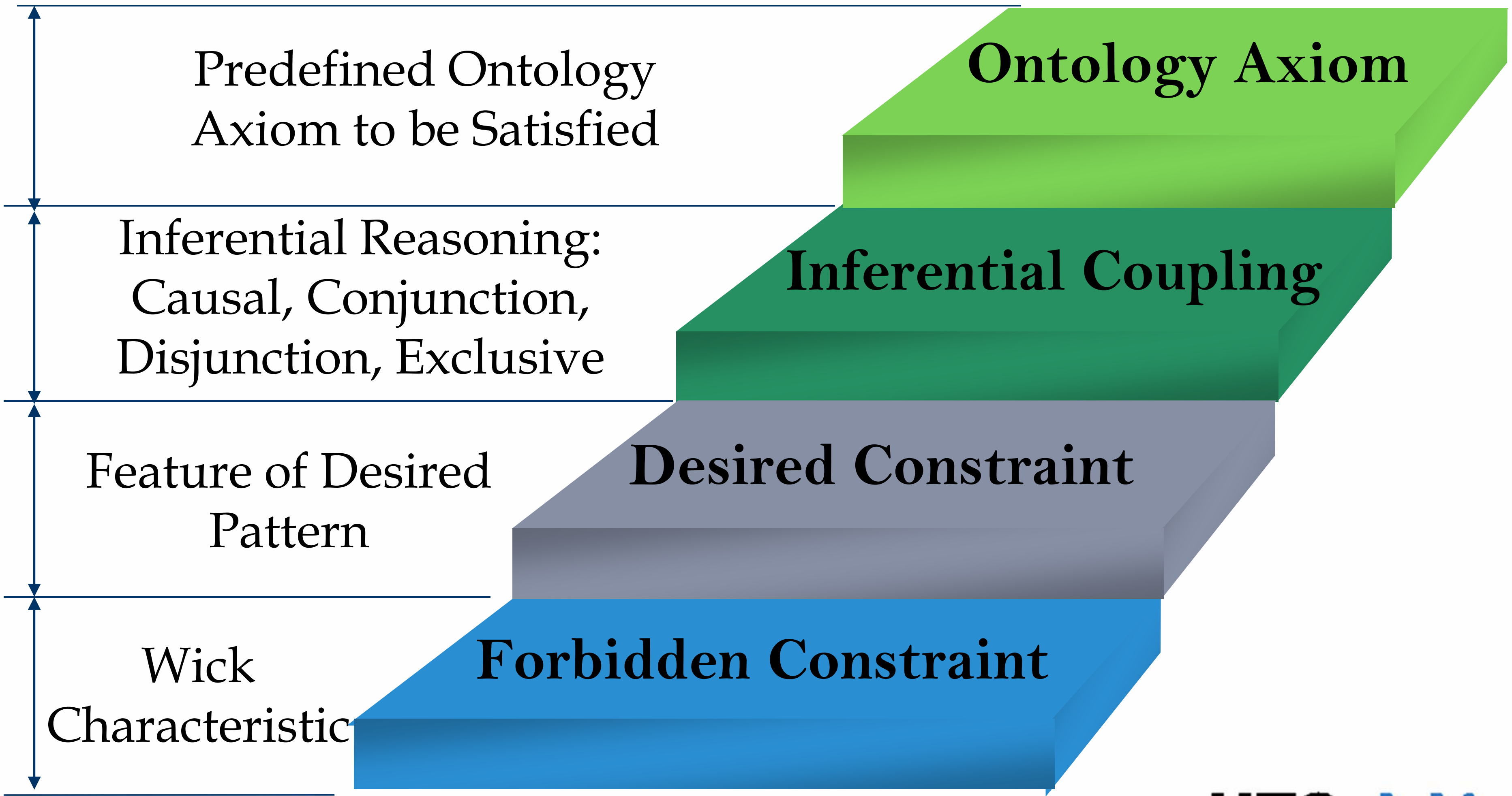
We take advantage of the propositional logic and temporal logic to express the constraints of the desired model.

synchronous and asynchronous behavior couplings

- **Temporal Logic**
  - **Linear Temporal Logic (LTL)**
    - Next (○)
    - Until (u)
    - Eventually (◊)
    - Always (□)
  - **Computation Tree Logic (CTL)**

3. Behavior Model/Representation

Behavior Constraint Indicator



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Checker

Different types of formal verification:

Manual Proof of Mathematical Arguments

- Time-consuming
- Error-prone
- Often not economically viable

✗

Interactive Computer Aided Theorem Proof

- Require significant expert knowledge

✗

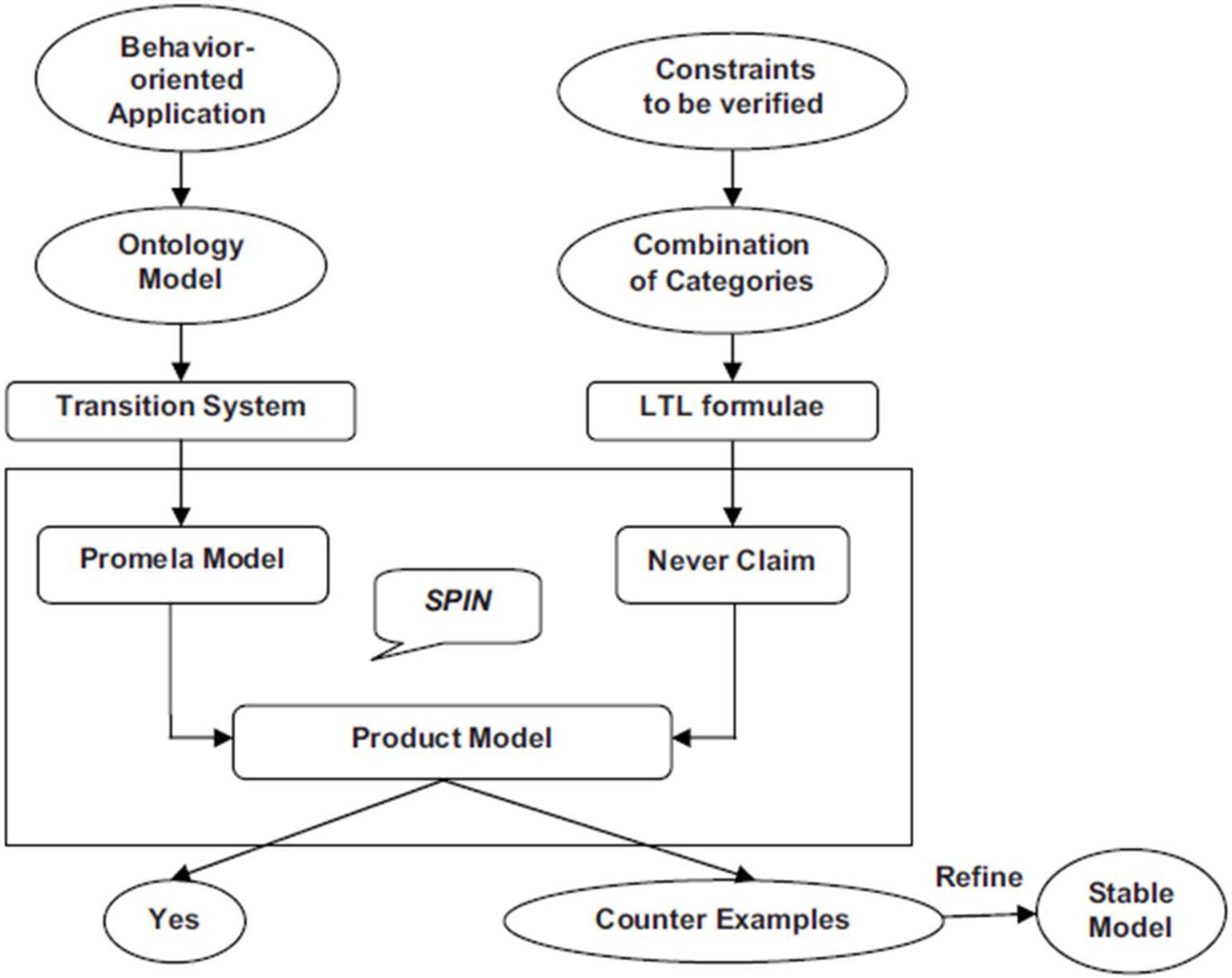
Automated Model Checking

An automated technique that, given a finite-state model of a system and a formal property, can systematically check whether or not this property holds for that model. If not, model checkers can help to identify the input sequence that triggers the failure.

√

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Checker

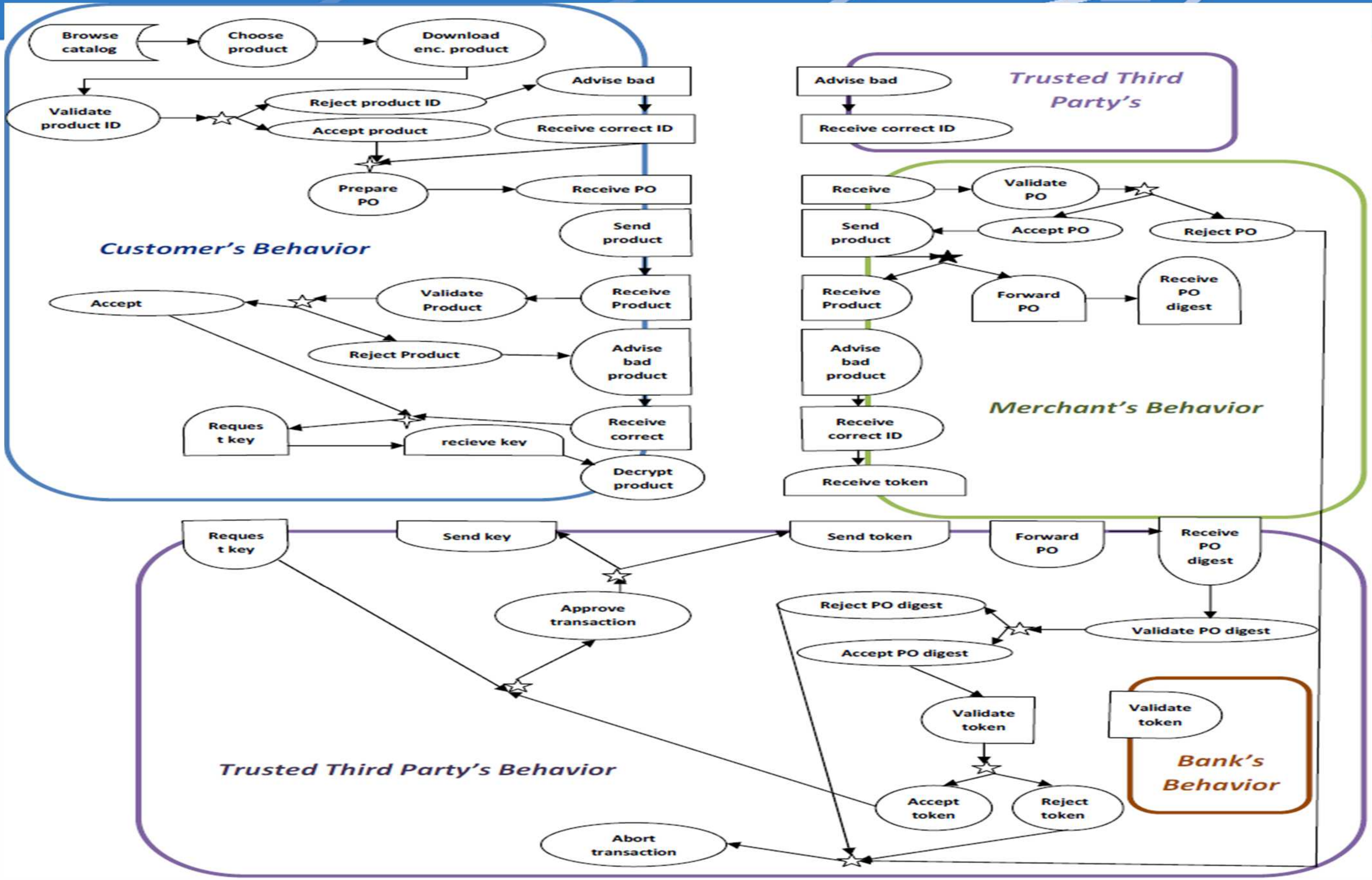


UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

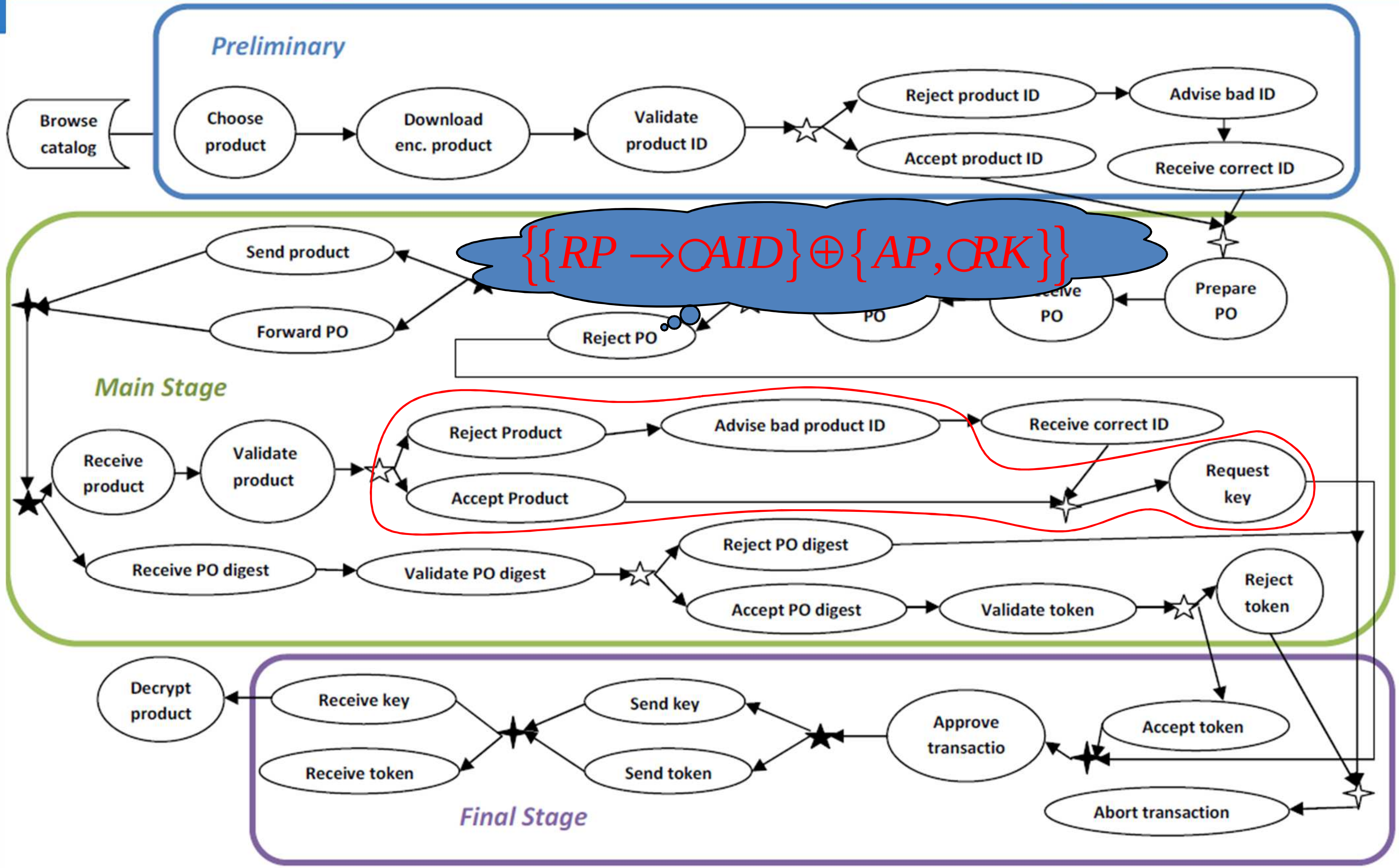
Case study of behavior representation

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

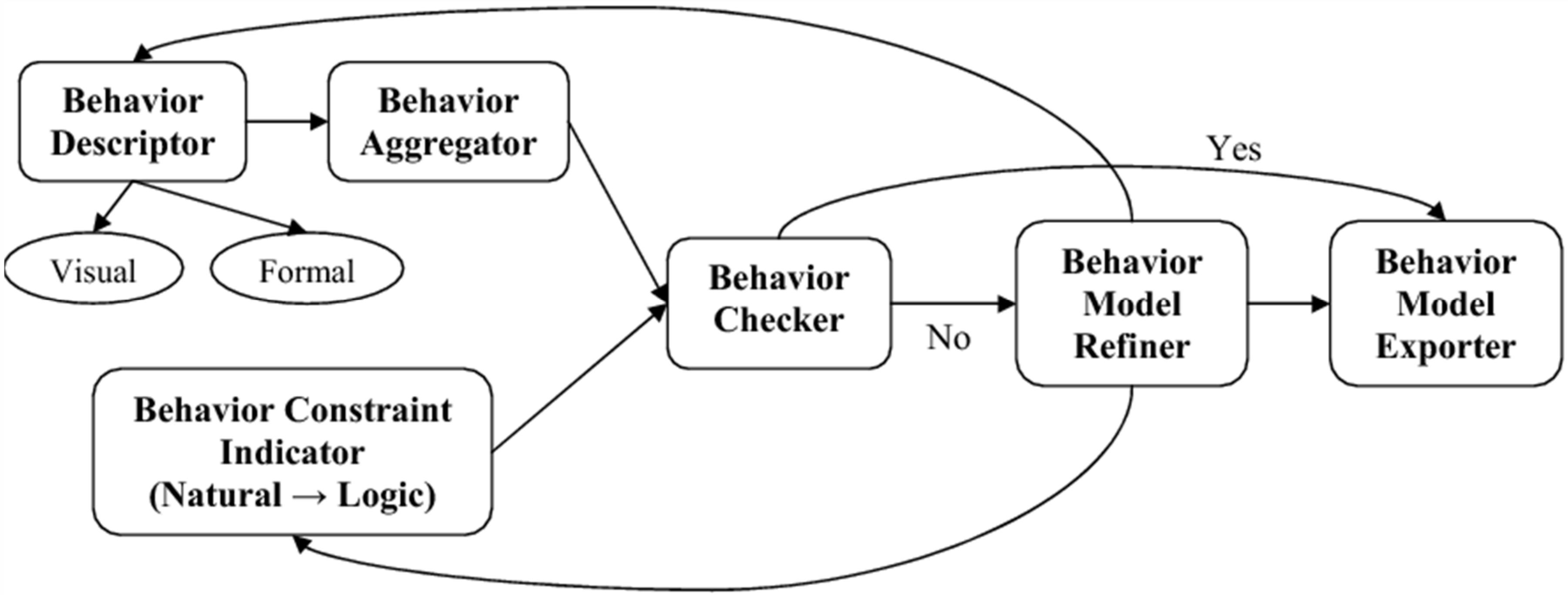
Graphical Action Sub-model of Online Shopping based on *Actor's Roles*



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



Behavior Modeling and Checking Framework

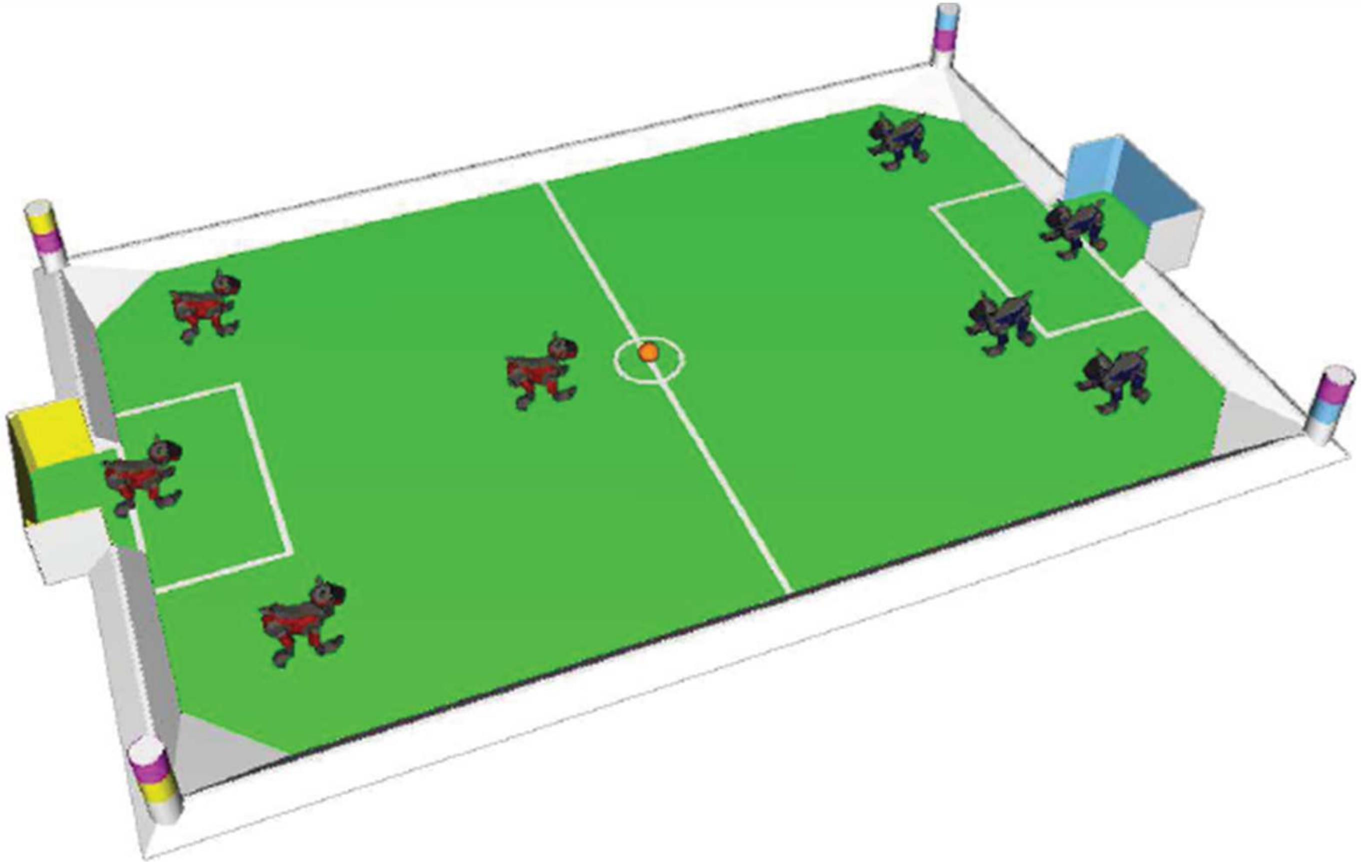


Ontology-based Behavior Modeling and Checking

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Case Study: Robot Soccer Game

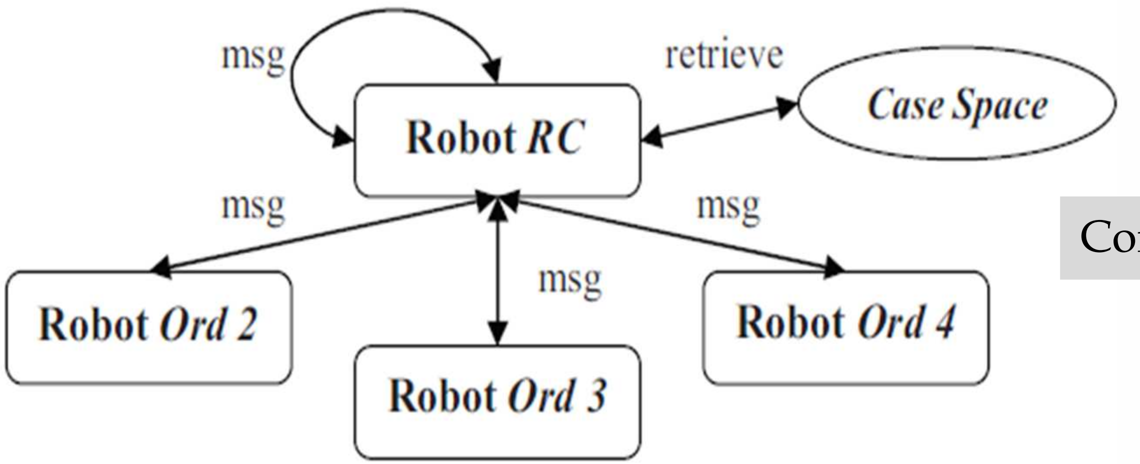
Snapshot of the four-legged league in the Robocup soccer competition: two teams participate in a Robocup soccer competition with four Sony AIBO robots in each group.

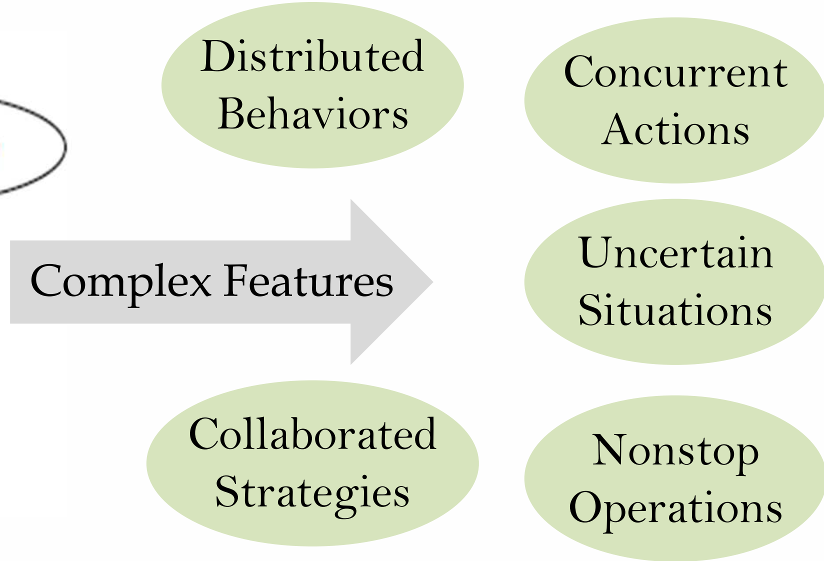


3. Behavior Model/Representation

Case Study: Behavior Descriptor

A case-based multi-robot architecture with n robots and k retrievers:



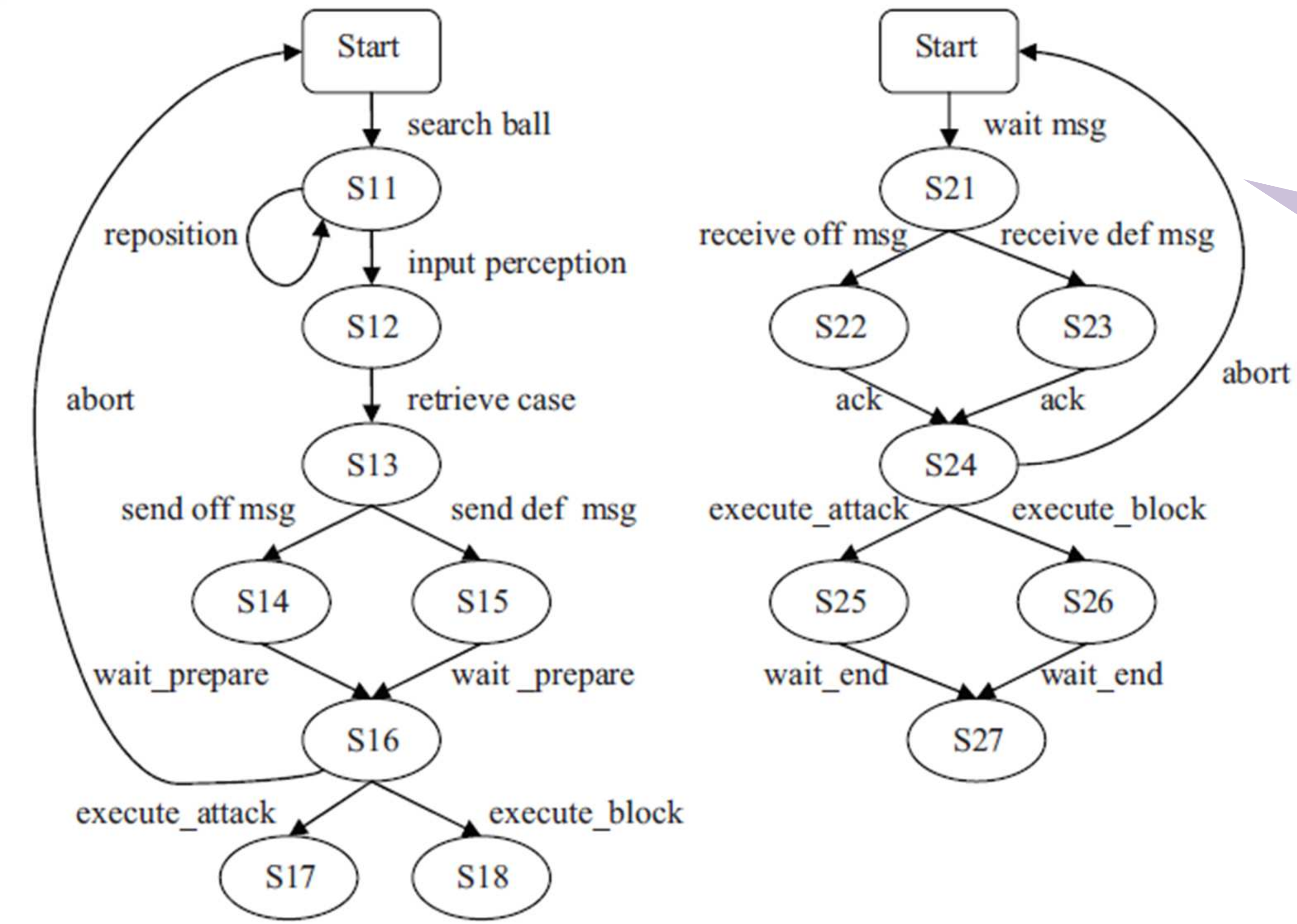


Robot RC firstly retrieves a case from the case space and then informs the rest of the Ords robot players. Once the Ords successfully receive the messages from RC, they send acknowledgments back to the retriever RC for confirmation. the RC also coordinates all the other players including itself to defeat the opponent. All the robots, no matter RC or Ord, could abort the executions at any moment if timeout expires, or messages or cases are lost in the interactions.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Aggregator

Transition system models $TS(\mathbb{B}(RC_p))$ **and** $TS(\mathbb{B}(Ord_q))$



Intra-coupled Aggregation $\theta_j^{(Ord_q)}$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Aggregator

Inter-coupled Aggregation

$\eta_i^{(RC,Ords)}$

$$(\mathbb{B}(RC)|\mathbb{B}(Ord_2)) : (\mathbb{B}(RC)|\mathbb{B}(Ord_3)) : (\mathbb{B}(RC)|\mathbb{B}(Ord_4))$$

The syntax of coupled behaviors between retriever RC and players Ords:

$$\mathbb{B}(RC, Ords) = (\mathbb{B}^{\theta^{(RC)}})^{\eta^{(RC,Ords)}} * (\mathbb{B}^{\theta^{(Ords)}})^{\eta^{(RC,Ords)}}$$

Combined Aggregation

$h^{(RC,Ord)}$

$$TS(\mathbb{B}(RC))|(TS(\mathbb{B}(Ord_2)) : TS(\mathbb{B}(Ord_3)) : TS(\mathbb{B}(Ord_4)))$$

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Constraint Indicator

Ontology Axiom

$$\Box(\neg(execute\_attack^{[Ord_i]} \wedge execute\_block^{[Ord_i]}))$$

It is never the case that any Ord can both implement the executions of attack and block opponent players

Inferential Coupling

$$\Box((TS(retrieve\ case)^{[CR]} \wedge \bigcirc TS(send\ msg)^{[CR]}) \rightarrow \Diamond(TS(receive\ msg)^{[Ord_i]} \wedge \bigcirc TS(ack)^{[Ord_i]}))$$

If the case is successfully retrieved by CR, then eventually the message sent is received and the acknowledgment is sent by Ord.

Desired Constraint

$$\Box(wait\_end^{[Ord_i]} \cup (\wedge_{j \neq i})wait\_end^{[Ord_j]}))$$

The execution of a case will not be done until all Ords have completed their actions.

Forbidden Constraint

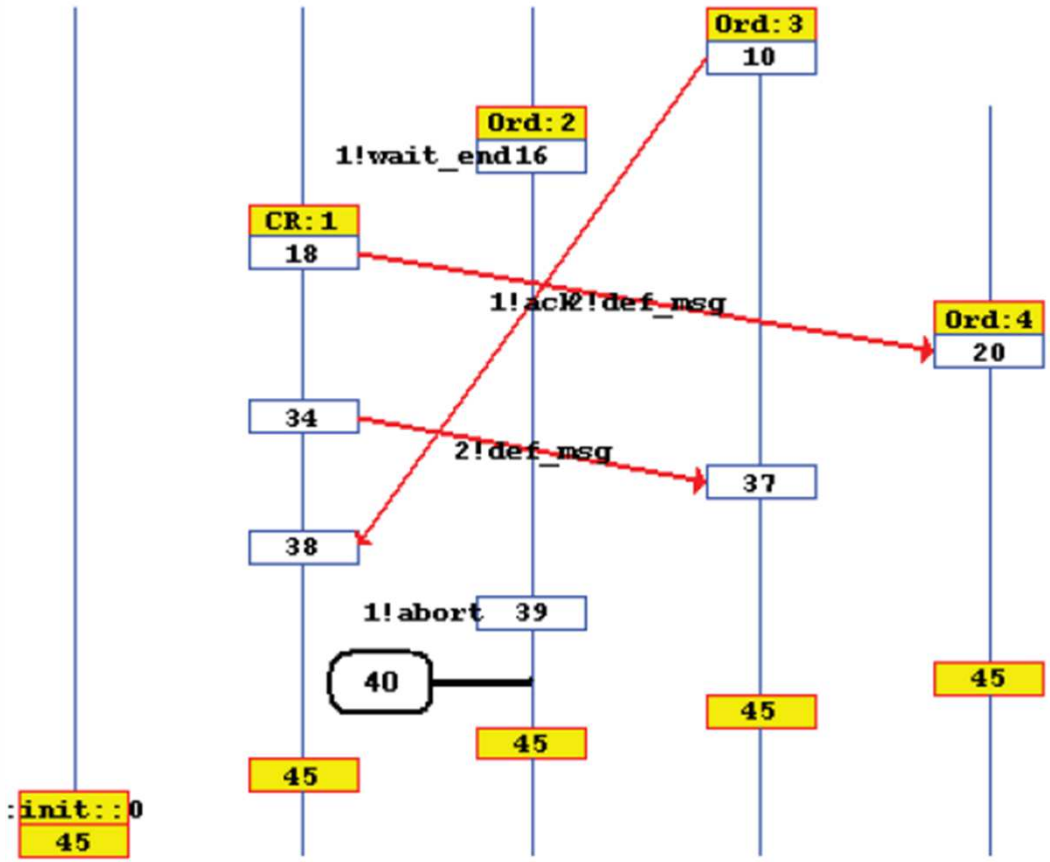
$$\Box\Diamond(\vee_i abort^{[Ord_i]})$$

Ord will infinitely often abort the execution.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Checker

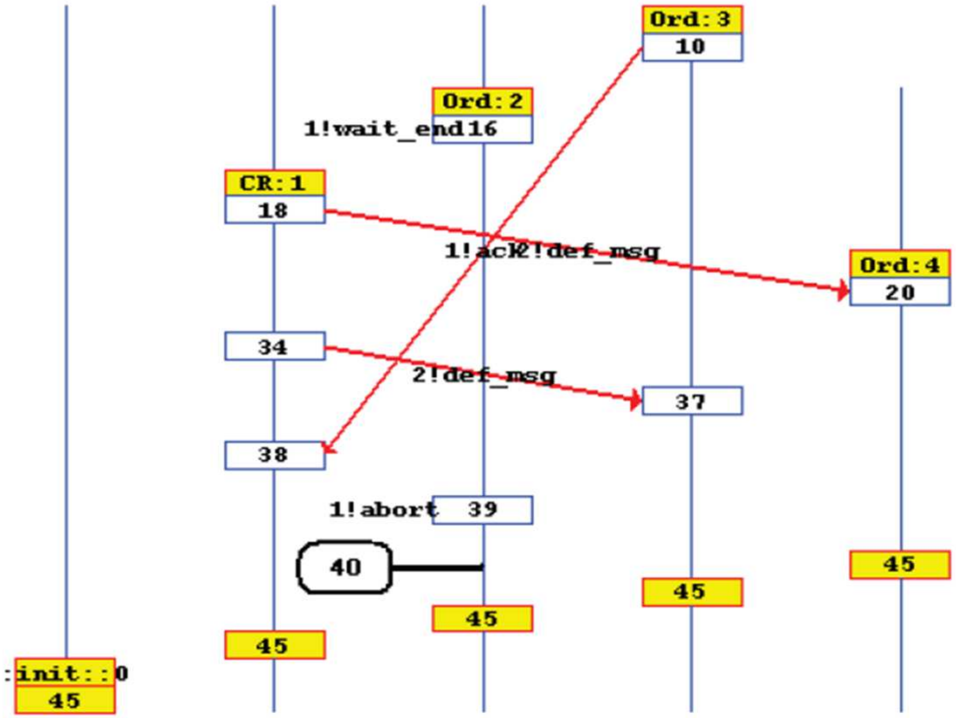
SPIN is used to perform checking of the corresponding $TS(\mathbb{B})$ and constraints.



The graphical interface of the counter example process with XSPIN is shown on the left, which is based on a Message Sequence Chart window of XSPIN. The vertical lines represent robot behaviors, boxes represent states, and arrows represent messages sent.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Checker



- State 10: $ack^{[Ord_3]} \rightarrow wait\_prepare^{[RC]}$
- State 18: $send\ def\ msg^{[RC]} \rightarrow wait\ msg^{[Ord_4]}$
- State 34: $send\ def\ msg^{[RC]} \rightarrow wait\ msg^{[Ord_3]}$
- State 39: $\Box(receive\ msg^{[Ord_2]} \rightarrow abort^{[Ord_2]})$
- State 45: $\Box(\wedge_i wait\_end^{[Ord_i]} \wedge wait\_prepare^{[RC]})$

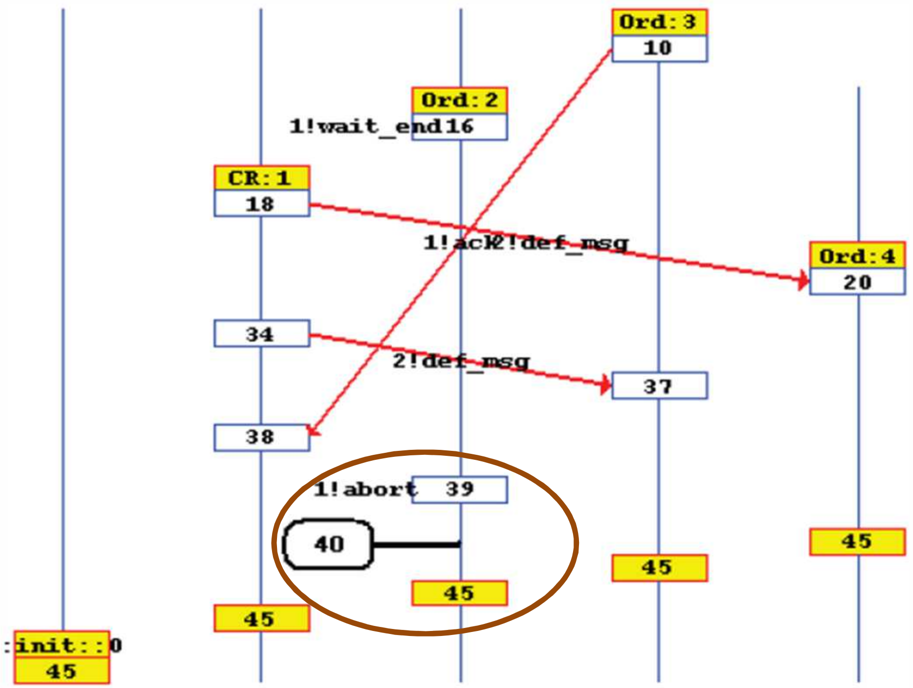
At State 39, the robot player Ord2 aborts the execution whenever it receives messages from RC. **Therefore, at State 45, Ord2 and RC wait for each other, resulting in an infinite wait loop while the executions of other robots are interrupted simultaneously, which is the so-called deadlock.** A typical deadlock scenario occurs when components mutually wait for each other to progress.

UTS:AAI
THE ADVANCED ANALYTICS INSTITUTE

Case Study: Behavior Model Refiner and Exporter

After analyzing the deadlock scenario, we introduce an additional state called "hold on" to break the loop.

- State 40: State 39$\rightarrow$ $hold\_on^{[Ord_i]\vee[RC]}$



When such a deadlock happens, the next state will be 'hold on', which means that the other two robot players Ord_3 and Ord_4 will continue their execution as usual. RC continues to retrieve cases and send messages without receiving ack from Ord_2 until the behaviors of Ord_2 become normal. If this does not occur, there must be design flaws in Ord_2, which should be explored by robot experts. In fact, "State 40" serves as a Behavior Model Refiner.

Finally, a refined system (in addition with State 40) will be provided by the Behavior Model Exporter

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Model Refiner

An additional state called “hold_on” to break the loop.

Deadlock $\rightarrow$ hold_on

- Two robot players Ord3 and Ord4 will continue their executions as usual.
- CR continues to retrieve cases and send messages without receiving acknowledgment from Ord1 until the behaviors of Ord1 become normal.
- Else, there must be some design flaws in Ord1, which should be further explored by robot experts.

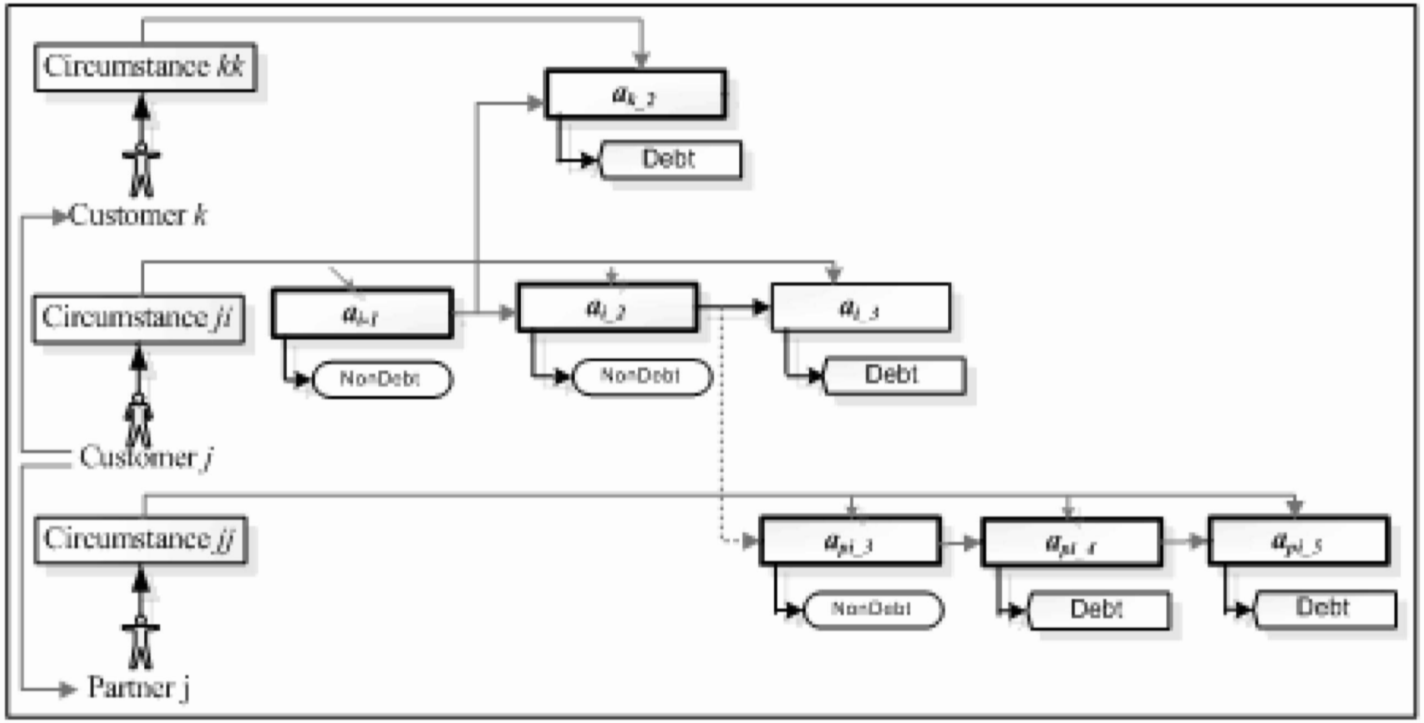
UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

6. High Impact 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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Coupled impact-oriented behaviors



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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).

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Impact

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

- *qualitative risk assessment:*
  - severity and likelihood are both expressed qualitatively (e.g., high, medium, or low)
- *quantitative risk assessment/probabilistic risk assessment:*
  - Consequences are expressed numerically
  - Their likelihoods of occurrence are expressed as *probabilities or frequencies*

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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?*

Cause/initiator modeling

- Factor analysis
- Rule-based methods
- Cause-effect analysis
- Failure Modes and Effects Analyses
- Sensitivity analysis
- Statistics techniques
- …

Effects/Consequences Modeling

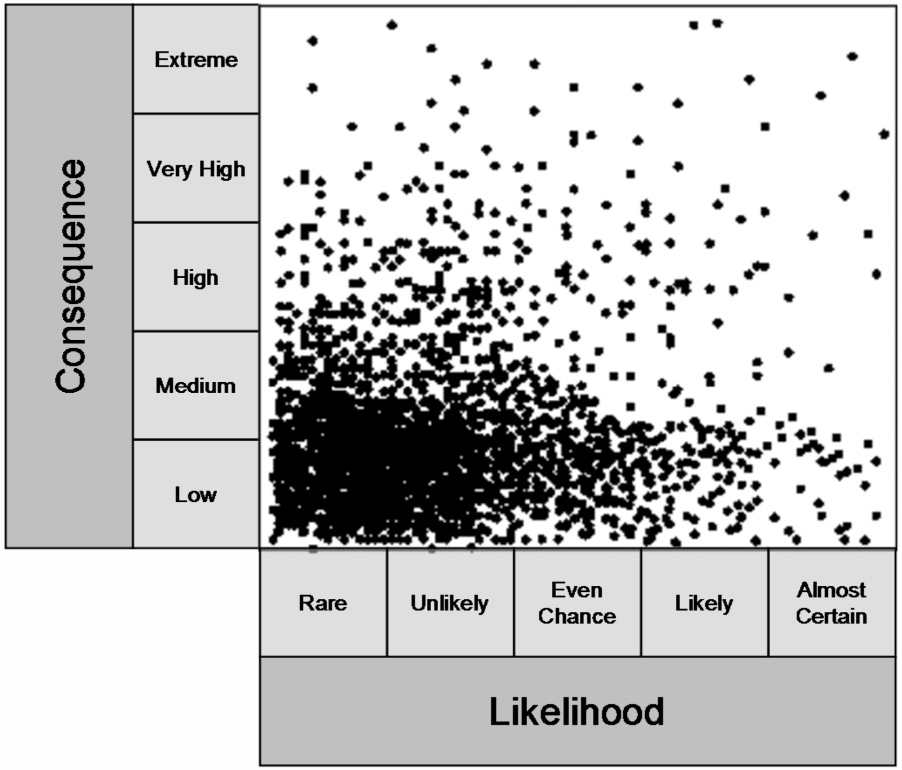
- Quantifying *accident (or mishap) scenarios*
  - chains of events that link the initiator to the end-point detrimental consequences
- *Deterministic analysis*
- *Probabilistic analysis*

Function(Cause, Effect)

- Probabilistic or statistical methods
- *Inductive* logic methods like *event tree analysis or event sequence diagrams*
- *Deductive methods like fault tree analysis*

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Risk Differentiation Framework

High	***Continuous Monitoring Q2***	***Continuous Review Q1***
Consequences	***Periodic Monitoring Q4***	***Periodic Review Q3***
	Low	***Likelihood*** **High**

Behavior impact modeling

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- Risk of a pattern, eg.

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

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Impact-Targeted Activity Mining

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

Here:
Impact → Debt, Fraud, Risk …

$\bar{T}$

Impact-Oriented Activity Patterns

$\{P \dashrightarrow T\}$ or $\{P \dashrightarrow \bar{T}\}$ $(P \dashrightarrow \bar{T}$, or $\bar{P} \dashrightarrow \bar{T}$)

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Impact-Contrasted Activity Patterns

$\{P \rightarrow T, P \rightarrow \overline{T}\}$ $\{P \rightarrow \overline{T}, P \rightarrow T$

- **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*

$$P_{T\overline{T}}: \{P \rightarrow T, P \rightarrow \overline{T}\}$$

- *Negative impact-contrasted pattern*

$$P_{\overline{T}T}: \{P \rightarrow \overline{T}, P \rightarrow T\}$$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Impact-Reversed Activity Patterns

{P --> T} {PQ --> $\overline{T}$ } {P-->$\overline{T}$ } {PQ --> T}

- *Sequential impact-reversed activity pattern pair*
  - *underlying pattern*: {P --> T} {P-->$\overline{T}$ }
  - *derivative pattern:* {PQ --> $\overline{T}$ } {PQ --> T}

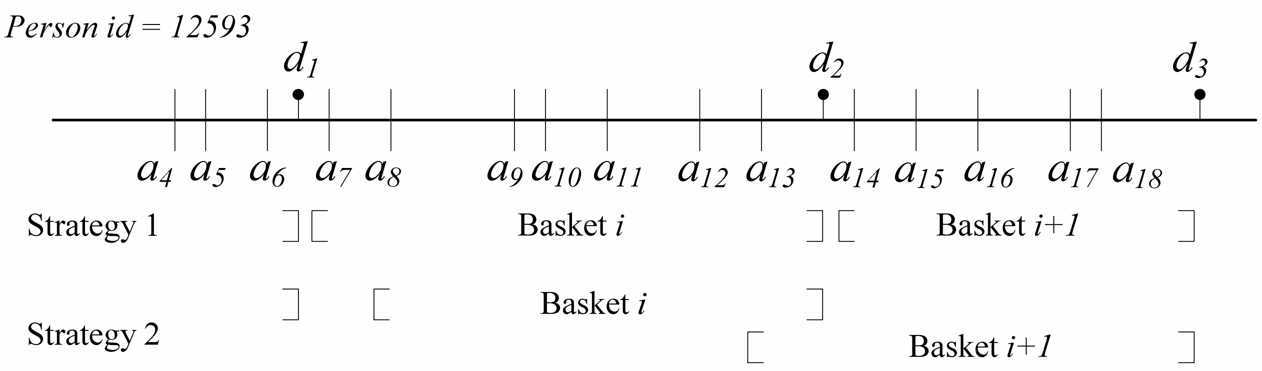
UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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., {*a8*, *a9*, *a10*, *a11*, *a12*, *a13*, ***d2***}, {*a13*, *a14*, *a15*, *a16*, *a17*, *a18*, ***d3***}



- 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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Frequent Debt-Targeted Activity Patterns

$\{P \dashrightarrow T\}$ or $\{P \dashrightarrow \overline{T}\}$ $(P \dashrightarrow \overline{T}$, or $\overline{P} \dashrightarrow \overline{T}$)

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

High impact behaviour analysis

(Impact-targeted behavior pattern mining)

TABLE 8
Common Frequent Sequential Patterns in Separate Data Sets

Patterns (P)	$Supp_{D_T}(P)$	$Supp_{D_{\bar{T}}}(P)$	$Cd_{T,\bar{T}}(P)$	$Cdr_{T,\bar{T}}(P)$	$Cd_{\bar{T},T}(P)$	$Cdr_{\bar{T},T}(P)$	$AvgAmt$ (cents)	$AvgDur$ (days)	$risk_{amt}$	$risk_{dur}$
a_5	0.382	0.178	0.204	2.15	-0.204	0.47	18290	6.2	0.363	0.334
a_7	0.312	0.154	0.157	2.02	-0.157	0.50	19090	6.8	0.310	0.300
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
a_{15}	0.746	0.567	0.179							
a_{16}	0.604	0.597	0.007							
a_{14}, a_{15}	0.605	0.374	0.231							
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}, a_{14}, a_{15}	0.477	0.257	0.220							
a_{14}, a_{15}, a_{14}	0.435	0.255	0.179							
a_{16}, a_{14}, a_{14}	0.361	0.267	0.093							
a_{16}, a_{14}, a_{16}	0.265	0.255	0.010							

ity Data

y_Time

24:13

33:55

*****

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

Underlying sequence (P)	Impact 1	Derivative activity Q	Impact 2	Cir	Cps	Local support of $P \rightarrow$ Impact 1	Local support of $PQ \rightarrow$ Impact 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_{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

7. Impact-oriented Behavior Combined Pattern Analysis

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

References

- **Longbing Cao**. Combined Mining: Analyzing Object and Pattern Relations for Discovering and Constructing Complex but Actionable Patterns, WIREs Data Mining and Knowledge Discovery.
- **Longbing Cao**, Huaifeng Zhang, Yanchang Zhao, Dan Luo, Chengqi Zhang. Combined Mining: Discovering Informative Knowledge in Complex Data, IEEE Trans. SMC Part B, 41(3): 699 - 712, 2011.
- Yanchang Zhao, Huaifeng Zhang, **Longbing Cao,** Chengqi Zhang. Combined Pattern Mining: from Learned Rules to Actionable Knowledge, LNCS 5360/2008, 393-403, 2008.
- Huaifeng Zhang, Yanchang Zhao, **Longbing Cao** and Chengqi Zhang. Combined Association Rule Mining, PAKDD2008.
- Yanchang Zhao, Huaifeng Zhang, Fernando Figueiredo, **Longbing Cao** Chengqi Zhang, Mining for Combined Association Rules on Multiple Datasets, Proc. of 2007 ACM SIGKDD Workshop on Domain Driven Data Mining (DDDM 07), 2007, pp. 18-23.

UTS:AAI THE ADVANCED ANALYTICS INSTITUTE

Pattern discovery process

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

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

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

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

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

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

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

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 $\mathcal{D}_k$ 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:

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

in which atomic patterns $\mathcal{P}_{n,m,l}$ from either individual sources $\mathcal{D}_k$, 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, $\mathcal{P} = \{\mathcal{P}'_k | \mathcal{R}'_k \to \mathcal{P}'_k\}$, 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

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Number of constituent atoms

- Pair patterns

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

- Cluster patterns

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

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

$$\{\mathcal{P} ::= P_i \cup P_i' \cup P_j \cup P_j', P_j = \mathcal{G}(P_i), \ldots, P_j' = \mathcal{G}'(P_i)'\}$$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Time frame

- Independent patterns

$$\{P_1 : P_2\}$$

- Sequential patterns

$$\{P_1 ; P_2\}$$

- Hybrid patterns

$$\{P_1 \otimes P_2 \cdots \otimes P_n ; \otimes \in \{:, \|, ;\}\}$$

Basic Process: an framework

- Multi-source combined pattern mining

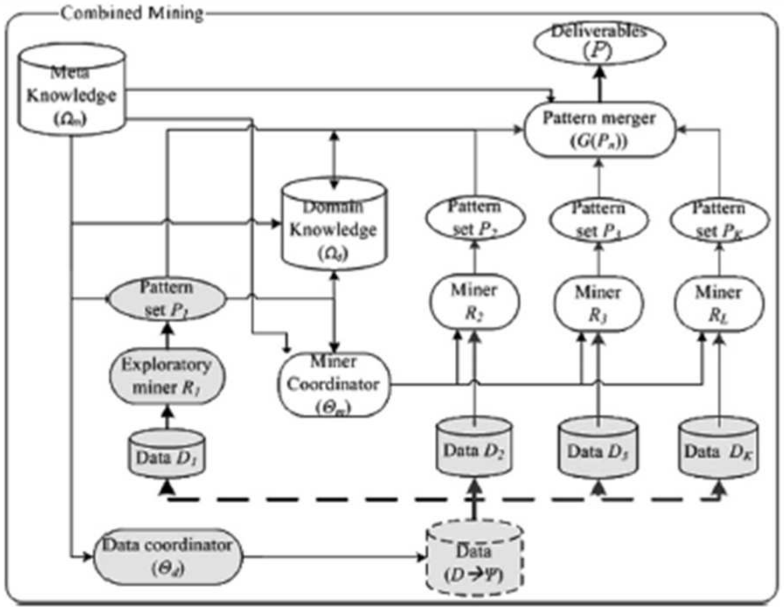


Fig. 1. Combined mining for actionable patterns.

$$CM ::= \underbrace{\mathcal{D}_k[\mathcal{D} \xrightarrow{\otimes} \mathcal{D}_k]^{\mathcal{I}_k,\mathcal{R}_k,\Omega_m} \{\mathcal{P}_k\}}_{K} \longrightarrow {}^{\mathcal{G}^N} \xrightarrow{\mathcal{P}_k,\Omega_d,\Omega_m} \mathcal{P}$$

PROCESS: Multisource Combined Mining
INPUT: target data sets $\mathcal{D}_k$ $(k = 1, \ldots, K)$, 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 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

$$\mathcal{P}_k : \mathcal{R}_l(\mathcal{F}_1, \cdots, \mathcal{F}_k)$$
$$\mathcal{P} := \mathcal{G}_F(\mathcal{P}_k) \tag{8}$$

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.

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

- Multi-method combined pattern mining

Definition 10 (Multimethod Combined Mining): Assuming that there are l data mining methods $\mathcal{R}_l\,(l = 1, \ldots, L)$, their respective interestingness metrics are in the set $\mathcal{I}_m\,(m = 1, \ldots, M)$. 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

$$\begin{cases} \mathcal{D}_1 \stackrel{e,\mathcal{I}_1,\mathcal{R}_1,\Omega_m}{\longrightarrow} \mathcal{P}_1 \\ \mathcal{D}_2 \stackrel{e,\mathcal{I}_2,\mathcal{R}_2,\Omega_m}{\longrightarrow} \mathcal{P}_2 \\ \ldots \\ \mathcal{D}_K \stackrel{e,\mathcal{I}_l,\mathcal{R}_l,\Omega_m}{\longrightarrow} \mathcal{P}_n \end{cases} \qquad \mathcal{P} := \mathcal{G}(\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n). \qquad (22)$$

  - Serial MMCM

$$\mathcal{D} \stackrel{e,\mathcal{R}_l,\mathcal{F}_l,\mathcal{I}_l,\Omega_m}{\longrightarrow} \mathcal{P}_1, \ or \qquad (23)$$

$$\{\mathcal{R}_1, \mathcal{F}_1, \mathcal{I}_1\} \stackrel{e,\mathcal{D},\Omega_m}{\longrightarrow} \mathcal{P}_1. \qquad (24)$$

$$\{\mathcal{R}_2, \mathcal{F}_2, \mathcal{I}_2\} \stackrel{e,\mathcal{D},\Omega_m,\mathcal{P}_1}{\longrightarrow} \mathcal{P}_2. \qquad (25)$$

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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*

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

$T \neq \emptyset$ 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

- Supp(A->B) = Prob(A^B)
- Conf(A->B) = Prob(A^B) / Prob(A)
- Lift = Conf(A->B) / Prob(B)

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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

$$I_{\text{rule}}(X_{\text{p}} \wedge X_{\text{e}} \to T) = \frac{Cont_{\text{e}}(X_{\text{p}} \wedge X_{\text{e}} \to T)}{Lift(X_{\text{e}} \to T)}$$

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 $X_{\text{p}} \wedge X_{\text{e}} \to T$ is less interesting than $X_{\text{p}} \to T$ and $X_{\text{e}} \to T$. 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.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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) $X_1 \cap X_2 = X_{\mathrm{p}}$ *and* X_{p} *is called the* prefix *of pair* $\mathcal{P}$*;* $X_{1,e} = X_1 \setminus X_{\mathrm{p}}$ *and* $X_{2,e} = X_2 \setminus X_{\mathrm{p}}$*; 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

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

Combined Pattern Clusters

DEFINITION COMBINED PATTERN CLUSTERS. *Assume there are* k *local patterns* $X_i \to T_i, (i = 1, \ldots, k)$, $k \geq 3$ *and* $X_1 \cap X_2 \cap \cdots \cap X_k = X_{\mathrm{p}}$, *a combined pattern cluster (CPC) is in the form of*

$$\mathcal{C}: \begin{cases} X_1 \to T_1 \\ \cdots \\ X_k \to T_k \end{cases},$$

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 C have the same U but different V , which makes them associated with various results T.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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

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

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

- 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

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

- 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.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Extended Combined Pattern Pairs

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

$$\mathcal{E}: \begin{cases} X_{\mathrm{p}} \rightarrow T_1 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e}} \rightarrow T_2 \end{cases},$$

where $X_{\mathrm{p}} \neq \emptyset$, $X_{\mathrm{e}} \neq \emptyset$ *and* $X_{\mathrm{p}} \cap X_{\mathrm{e}} = \emptyset$.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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.*

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

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

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.*

$$\mathcal{S}: \begin{cases} X_{\mathrm{p}} \to T_1 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \to T_2 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \to T_3 \\ \cdots \\ X_{\mathrm{p}} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \wedge \cdots \wedge X_{\mathrm{e,k-1}} \to T_k \end{cases},$$

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*

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

Intervention Strategy 1

- Type A: Demographics differentiated combined pattern
  - Customers with the same actions but different demographics

    → different classes/business impact

$$\text{Type A:} \begin{cases} A_1 + D_1 & \rightarrow \quad \text{quick payer} \\ A_1 + D_2 & \rightarrow \quad \text{moderate payer} \\ A_1 + D_3 & \rightarrow \quad \text{slow payer} \end{cases}$$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Intervention Strategy 2

- Type B: Action differentiated combined pattern
  - Customers with the same demographics but taking different actions

    → different classes/business impact

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

Business Impact

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

	Behavior 1	Behavior 2
Demographic 1	Slow	Fast
Demographic 2	Fast	Slow

Case Study I

- Mining Combined Patterns and Patterns Clusters for Debt Recovery

Business Problem

- 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.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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	$Lift$ of
	Demographics	Arrangements	Repayments	Class		(%)					$X_p \to T$	$X_e \to T$
P_1	age:65+	withholding & irregular	withholding	C	50	63.3	2.91	3.40	2.47	4.01	0.85	1.38
P_2	income:0 & remote:Y & marrital:sep & gender:F	withholding	cash or post & withholding	B	20	69.0	1.47	1.95	1.34	2.15	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	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	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 $X_p \to T$	$Lift$ of $X_e \to T$
		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	132	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

UTS:AAI THE ADVANCED ANALYTICS INSTITUTE

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

Case Study II

- Mining Extended Combined Pattern Pairs for Debt Prevention

Business Problem

- 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.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Results (1)

Examples of Extended Combined Pattern Pairs

X_p	T_1	X_e	T_2	$Cont_e$	Cps	Local support of $X_p \to T_1$	Local support of $X_p \wedge X_e \to T_2$
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 $a_{14} \rightarrow \bar{T}$ are 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

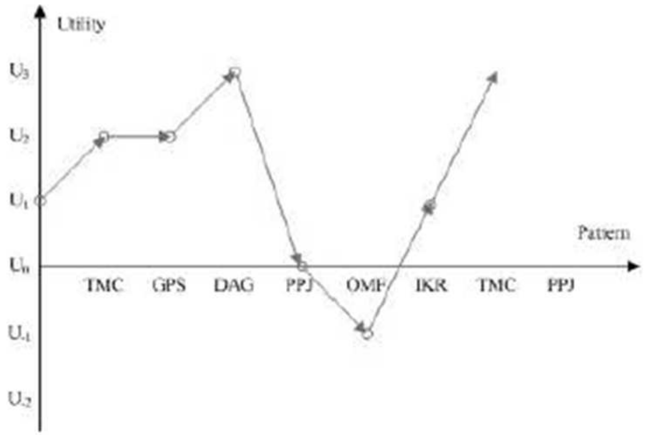


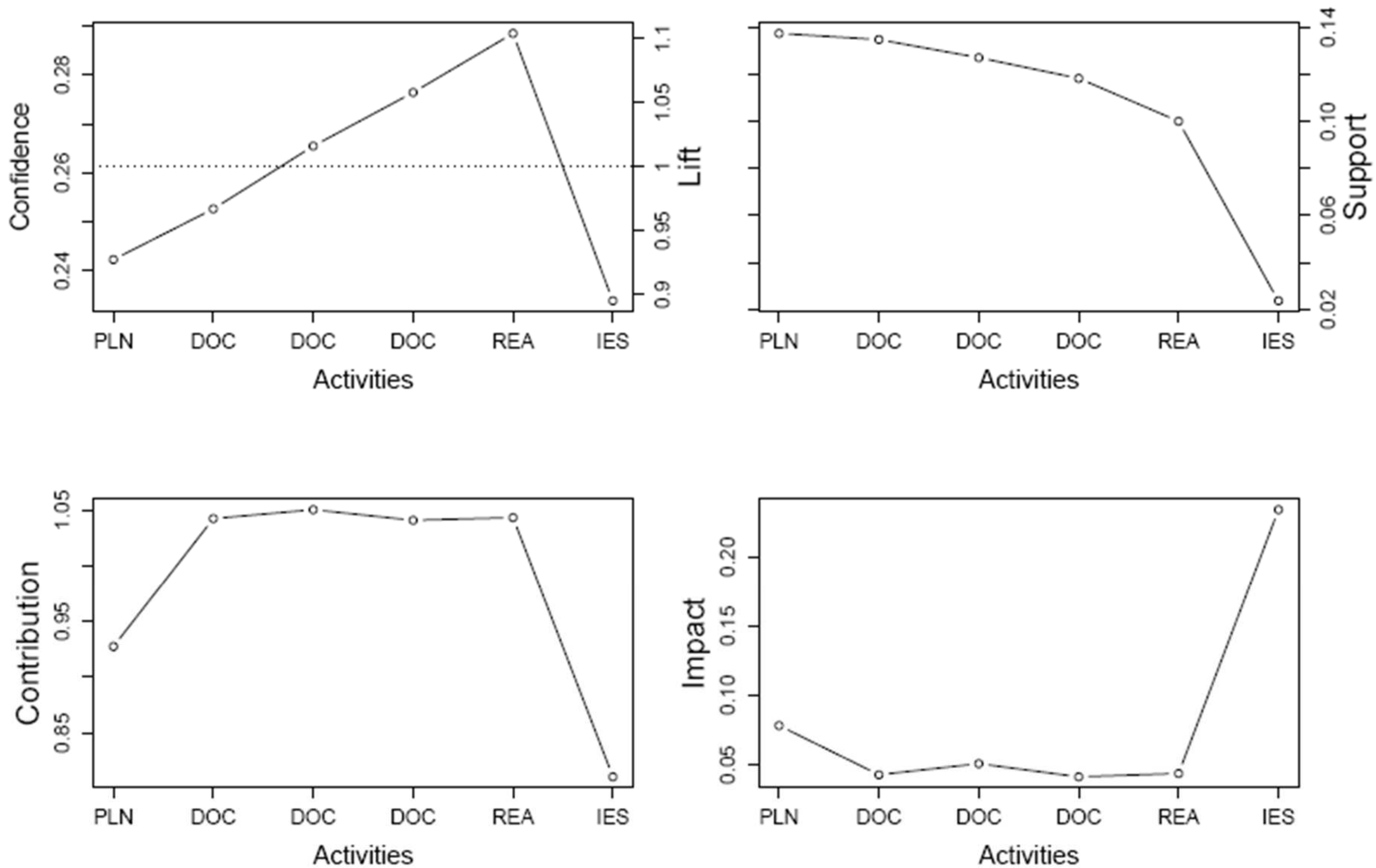
Figure 2: Pattern Evolution Chart

$$\begin{cases} TMC \rightarrow U_1 \\ TMC, GPS \rightarrow U_2 \\ TMC, GPS, DAG \rightarrow U_2 \\ TMC, GPS, DAG, PPJ \rightarrow U_3 \\ TMC, GPS, DAG, PPJ, OMF \rightarrow U_0 \\ TMC, GPS, DAG, PPJ, OMF, IKR \rightarrow U_{-1} \\ TMC, GPS, DAG, PPJ, OMF, IKR, TMC \rightarrow U_1 \\ TMC, GPS, DAG, PPJ, OMF, IKR, TMC, PPJ \rightarrow U_3 \end{cases}, \qquad (6)$$

An Example of Extended Combined Pattern Cluster

$$
\begin{cases}
PLN \rightarrow T \\
PLN, DOC \rightarrow T \\
PLN, DOC, DOC \rightarrow T \\
PLN, DOC, DOC, DOC \rightarrow T \\
PLN, DOC, DOC, DOC, REA \rightarrow T \\
PLN, DOC, DOC, DOC, REA, IES \rightarrow T
\end{cases}
$$

An Example of Extended Combined Pattern Cluster



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

8. High Utility Behavior Analysis

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

High Utility Sequential Pattern Mining

The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2012)

USpan: An Efficient Algorithm for High Utility Sequential Pattern Mining

Junfu Yin, Zhigang Zheng and Longbing Cao

Advanced Analytics Institute
University of Technology, Sydney, Australia

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Outline

1. Introduction
2. Related Work
3. Problem Statement
4. USpan Algorithm
5. Experiments
6. Conclusions

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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 quanity 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

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)] >

The utility of <*e*> in (*e*, 6) is 6×1 = 6

The utility of <*ea*> in s_2 is
{ ((6×1) + (1×2)), ((6×1) + (1×2)) }
= {8, 10}

The utility of <*ea*> is the database is
{{}, {8, 10}, {}, {16, 10}, {15, 7}}.

Add the highest utility in each sequence to represent the utility of <*ea*>:
10 + 16 + 15 = 41

If the minimum utility threshold ξ = 40 then <*ea*> is a high utility pattern.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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)].

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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 <*ea*> **matches:**

<(*e*, 6)(*a*, 1)> and <(*e*, 6)(*a*, 2)> in s_2 ;
<(*e*, 2)(*a*, 7)> and <(*e*, 2)(*a*, 4)> in s_4 ;
<(*e*, 3)(*a*, 6)> and <(*e*, 3)(*a*, 2)> in s_5 ;

Denote as <(*e*, 6)(*a*, 1)> ~ <*ea*>

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Problem Statement: Utilities

The Sequence Utility Framework

The q-item utility:

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

The q-itemset utility:

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

The q-sequence utility:

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

The q-sequence database utility:

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

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)}

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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:

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

The q-sequence utility:

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

The q-sequence database utility:

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

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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

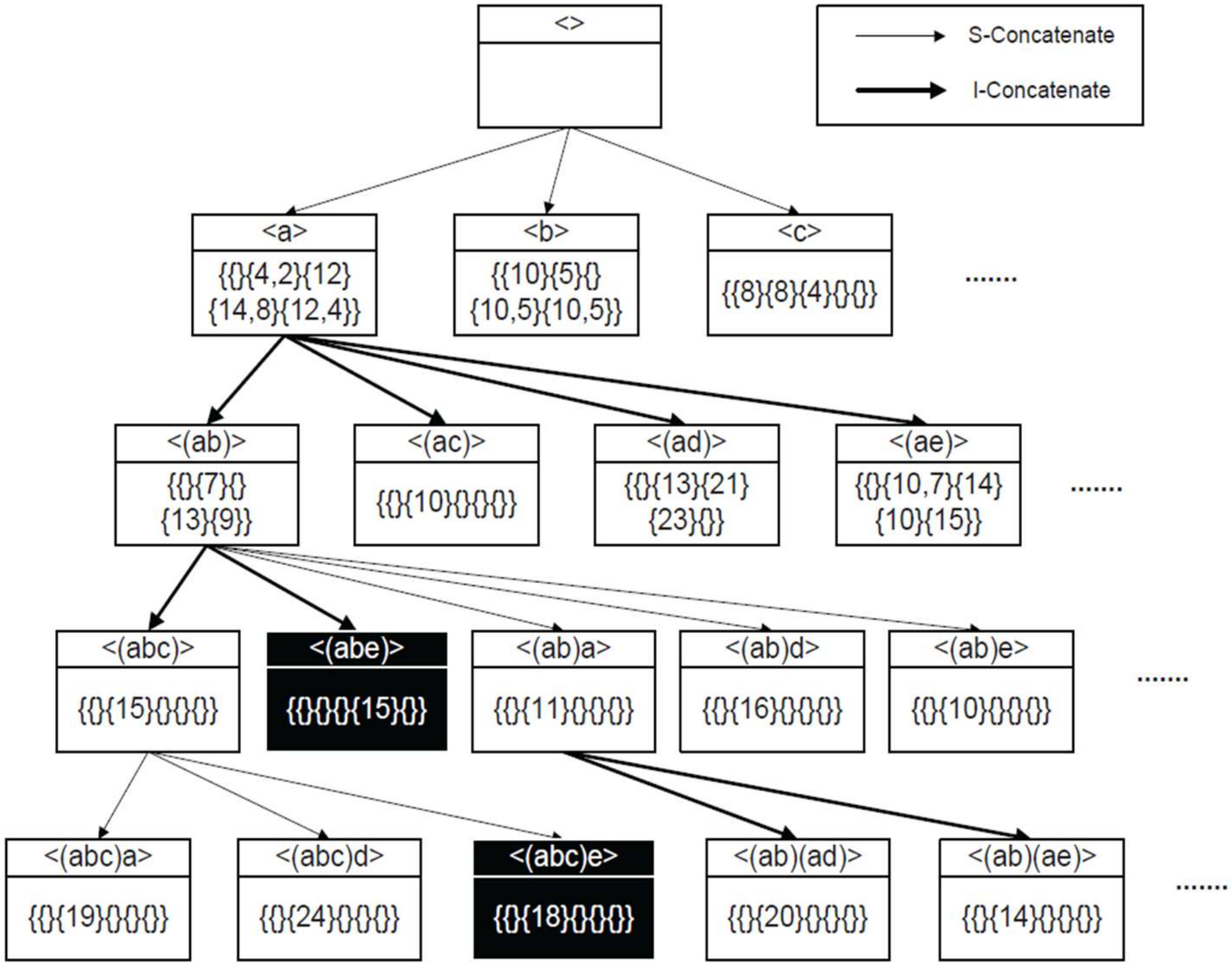
u_{max} (<$(abc)a$>)= 19

No Downward Closure Property

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

USpan Algorithm

Lexicographic Q-sequence Tree



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

4. High Impact/Utility Behavior Analysis

USpan Algorithm

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)] >

Items	Itemset 1	Itemset 2	Itemset 3
a		14	8
b	10		5
d		9	
e	2		2

v(<*b*>) = {10, 5}

Items	*I 1*	*I 2*	*I 3*
a		14	8
b	**10**●		**5**■
d		9	
e	2		2

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

Items	*I 1*	*I 2*	*I 3*
a		14	8
b	10		5
d		9	
e	2●		2■

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

Items	*I 1*	*I 2*	*I 3*
a		14●	8■
b	10		5
d		9	
e	2		2

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

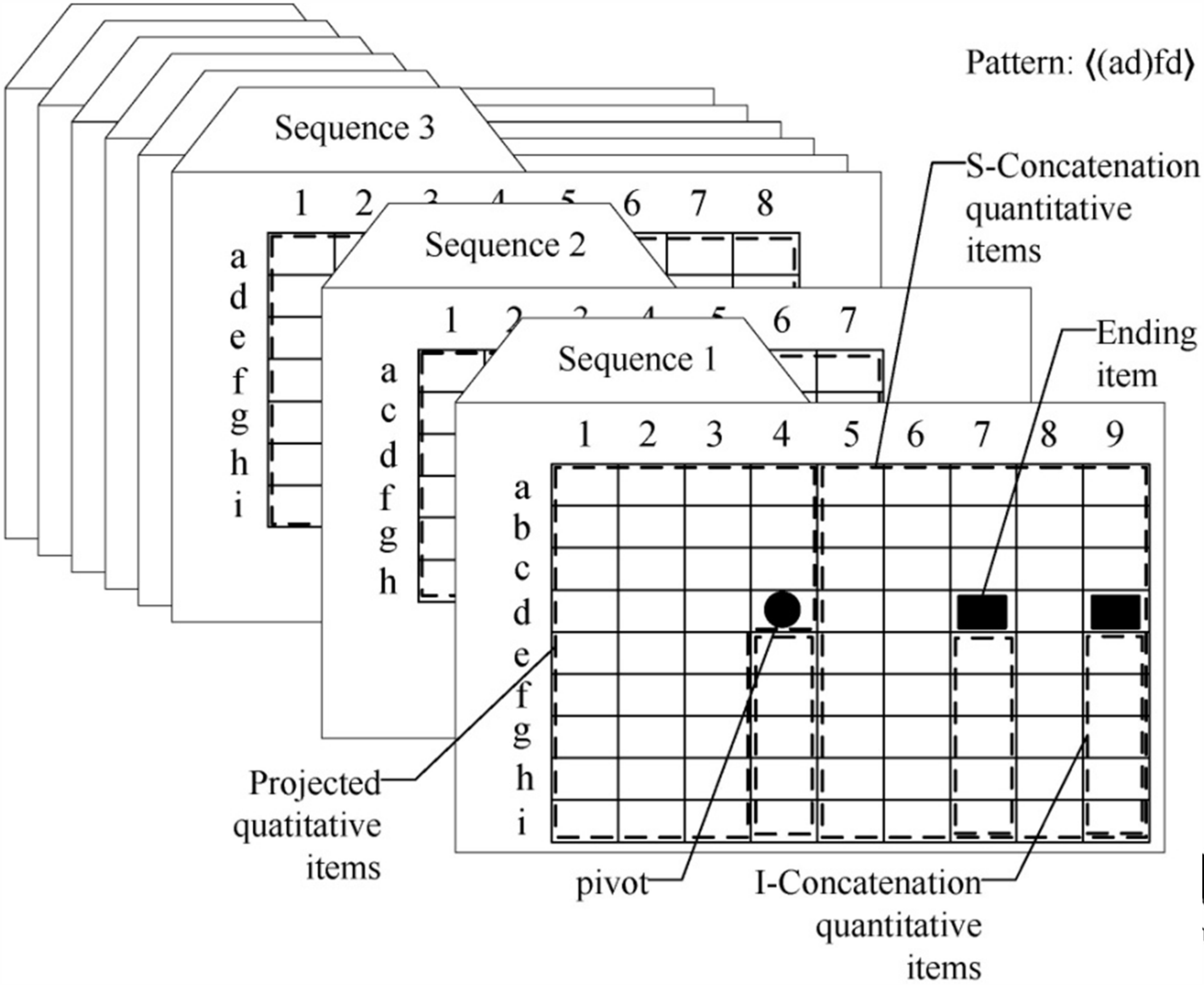
Items	*I 1*	*I 2*	*I 3*
a		14	8
b	10		5
d		9●	
e	2		2

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

Items	*I 1*	*I 2*	*I 3*
a		14	8●
b	10		5
d		9	
e	2		2

USpan Algorithm: Concatenation

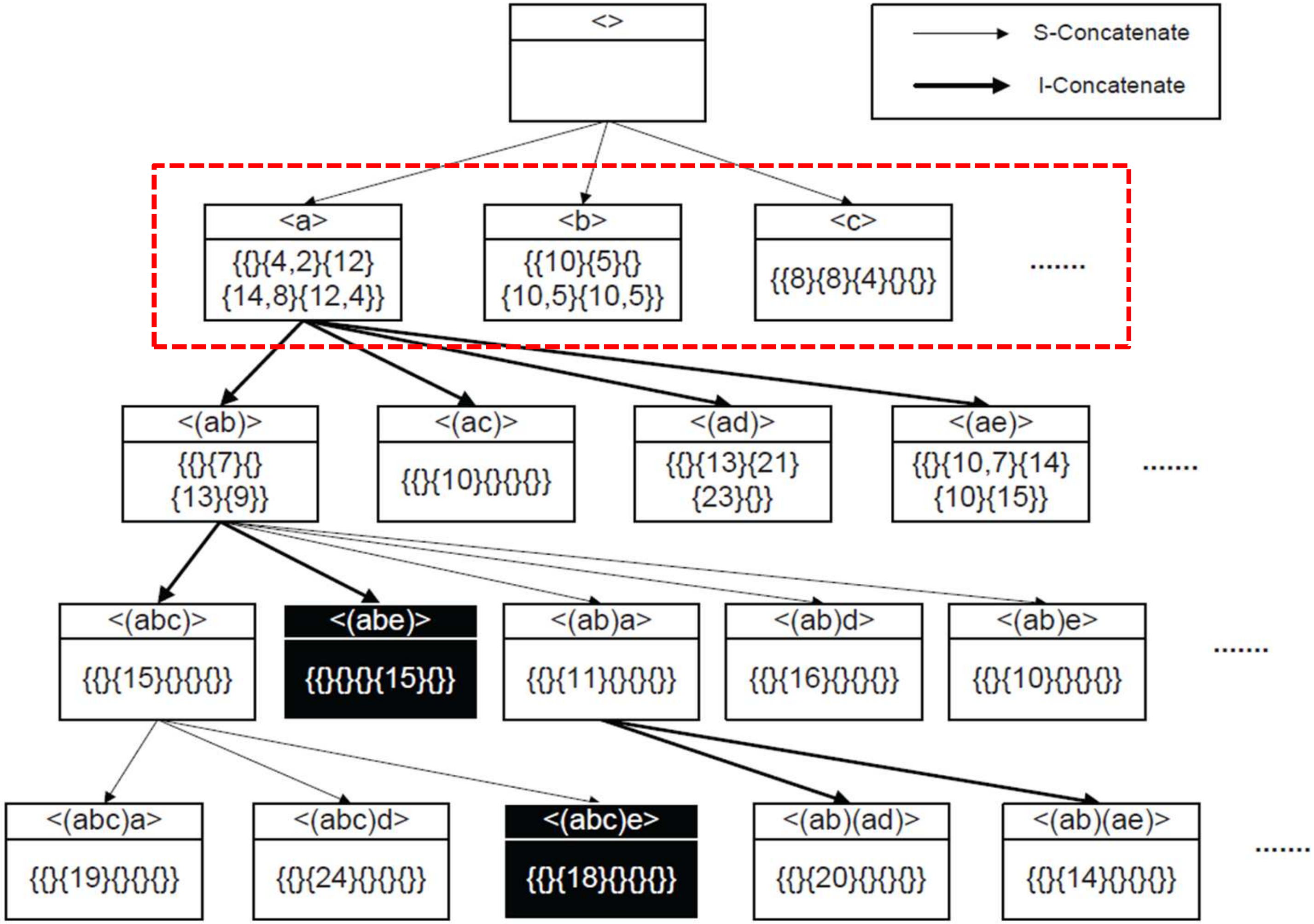
Data Representation



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

USpan Algorithm: Width Pruning

What is Width Pruning



AAi
YTICS INSTITUTE

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

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**

$$SWU(<ea>) = u(s_2) + u(s_4) + u(s_5) = 41 + 50 + 24 = 115$$

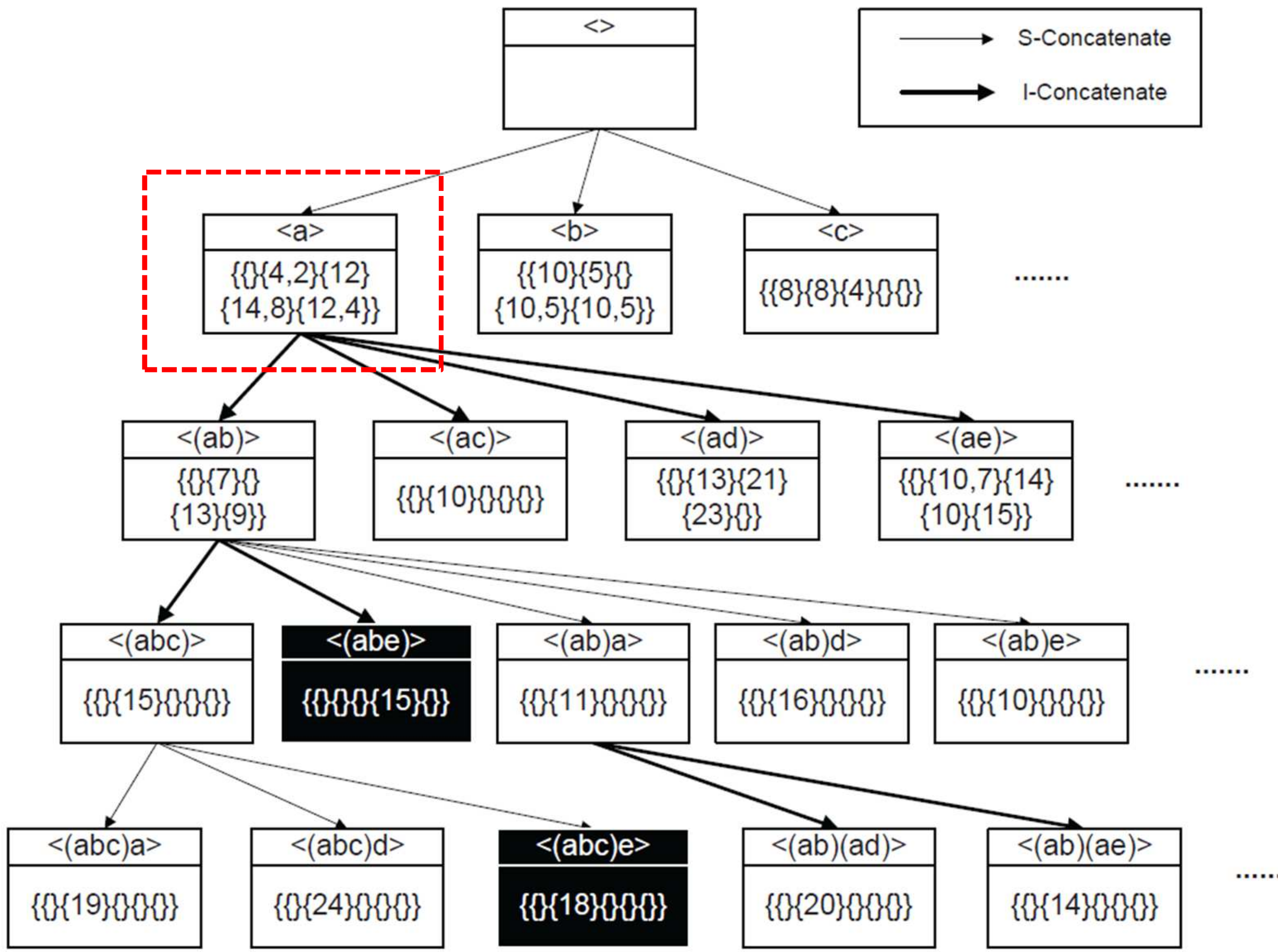
SID	Quantitative Sequence	SU
1	**< (*e*, 5) [(*c*, 2)(*f*, 1)] (*b*, 2) >**	**24**
2	< [(*a*, 2)(*e*, 6)] [(*a*, 1)(*b*, 1)(*c*, 2)] [(*a*, 2)(*d*, 3)(*e*, 3)] >	41
3	< (*c*, 1) [(*a*, 6)(*d*, 3)(*e*, 2)] >	27
4	< [(*b*, 2)(*e*, 2)] [(*a*, 7)(*d*, 3)] [(*a*, 4)(*b*, 1)(*e*, 2)] >	50
5	< [(*b*, 2)(*e*,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

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

$$SWU(<f>) = u(s_1) = 24$$

USpan Algorithm: Depth Pruning

What is Depth Pruning



ALYTICS INSTITUTE

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

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

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

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

$$u_{rest}(\langle ea \rangle) = (8+29) + (16+24) + (15+17)$$
$$= 37 + 40 + 32$$
$$= 109$$

$$u_{rest}(\langle e(ae) \rangle) = (18 + 9)$$
$$= 27$$

THE ADVANCED ANALYTICS INSTITUTE

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

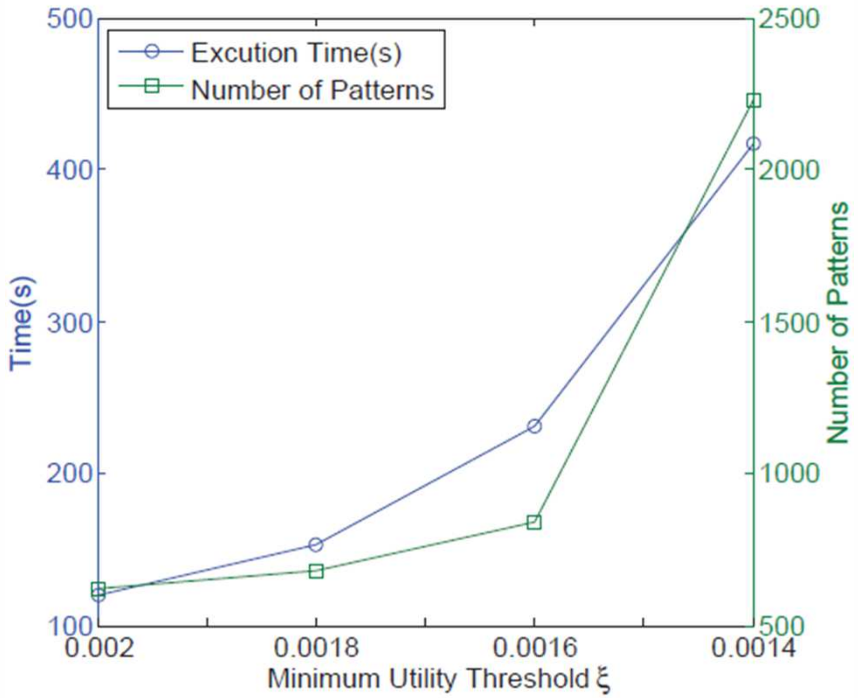
DS3 is a dataset consisting of online shopping transactions which contains 350,241 transactions and 59,477 customers.

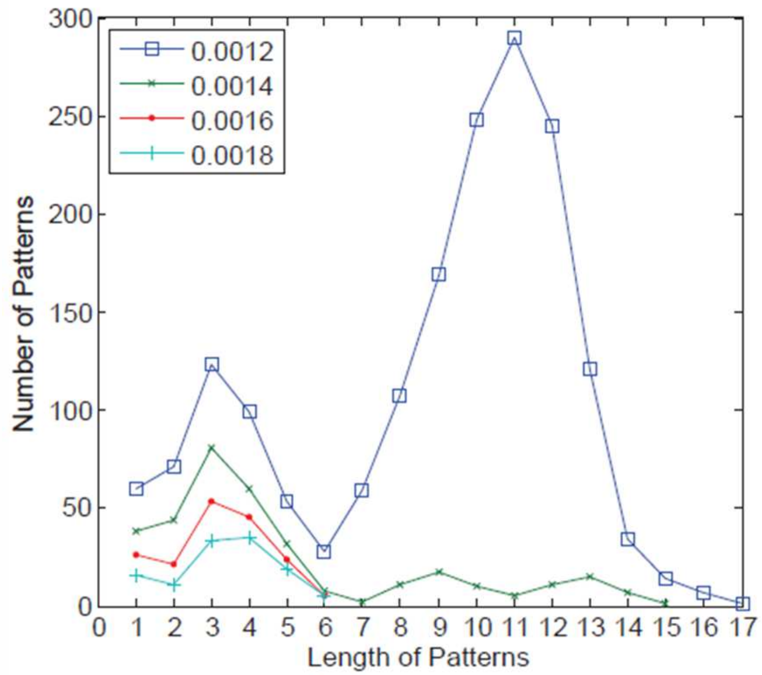
DS4 is a real dataset that includes mobile communication transactions. The dataset is a 100,000 mobile call history from a specific day. There are 67,420 customers in the dataset.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Experiments

Performance and distributions (DS2)

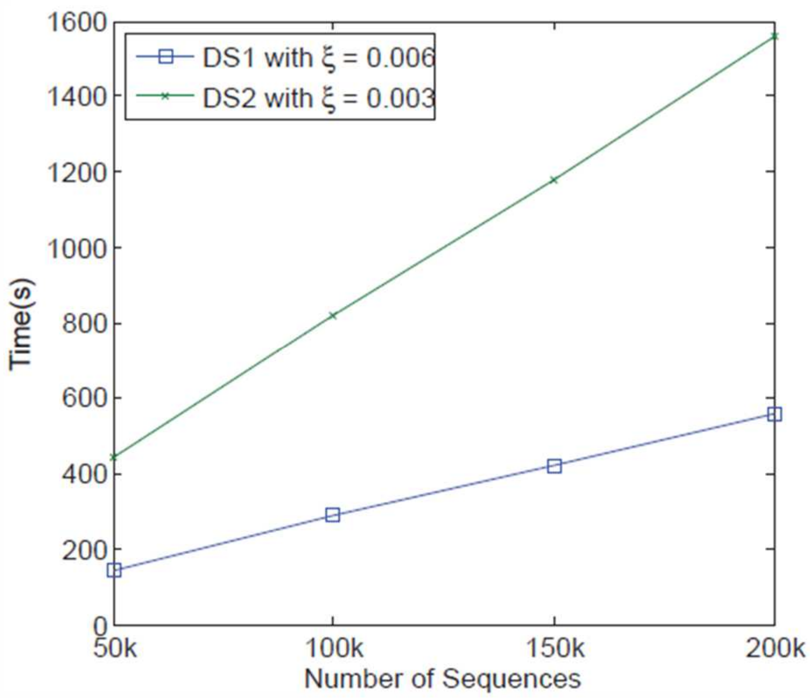


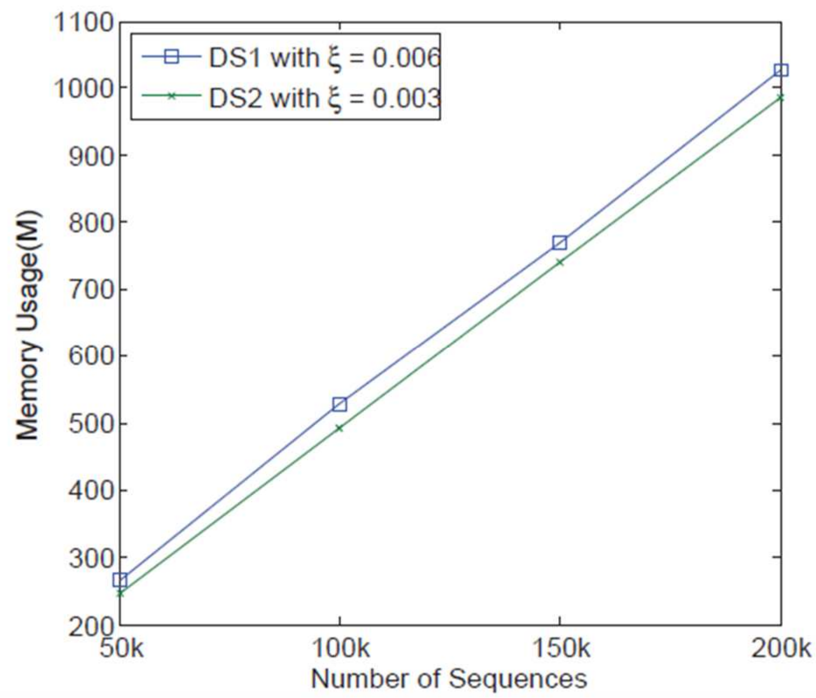


- The running time and the number of patterns grow exponentially with respect to ξ.
- The high utility sequential patterns are mid-long patterns.

Experiments

Scalability Test (DS1 & DS2)



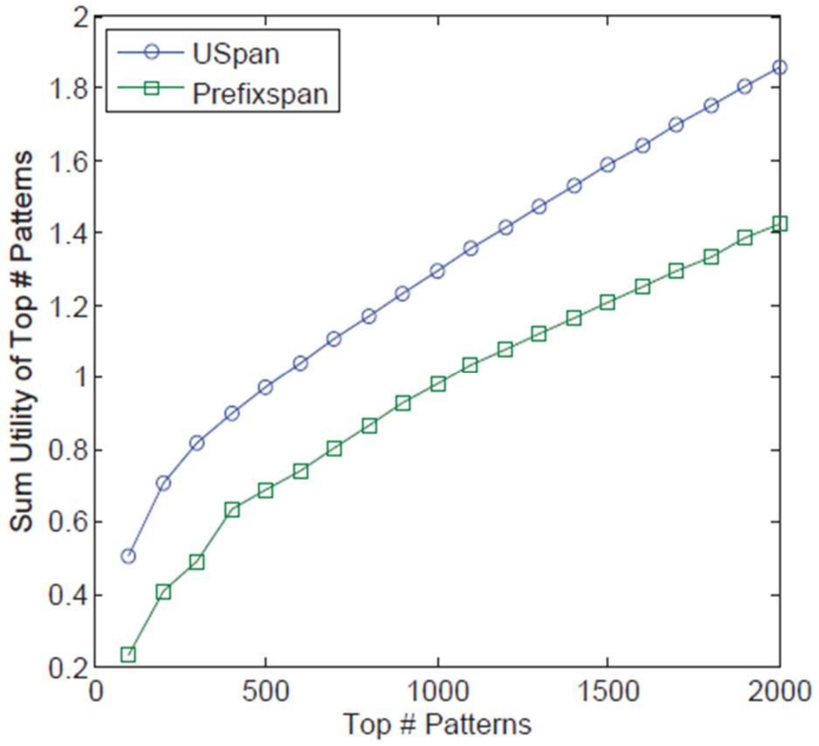


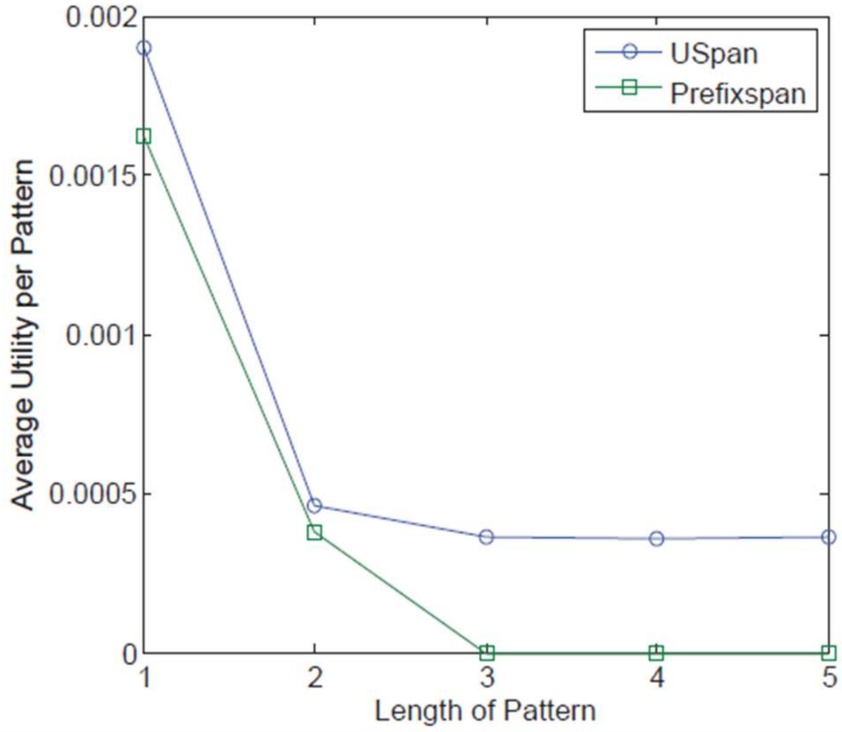
- Both the time and memory usage grow linearly with respect to the size of the DB.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Experiments

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





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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

9. Negative Behavior Analysis

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Negative sequential pattern mining

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

References

- 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 Zuo, Longbing Cao. An Efficient GA-Based Algorithm for Mining Negative Sequential Patterns, *PAKDD2010*, 262-273.
- Zhigang Zheng, Yanchang Zhao, Ziye Zuo, Longbing Cao. Negative-GSP: An Efficient Method for Mining Negative Sequential Patterns, *AusDM 2009*: 63-67.
- 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.
- 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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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.*

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

PSP & NSP

PSP: Positive Sequential Pattern

- **Only contain occurring itemsets**

E.g. p1=<a b c X>.

Existing Methods:

AprioriAll, GSP, FreeSpan, PrefixSpan, SPADE , SPAM

NSP: Negative Sequential Pattern

- **Also contain non-occurring itemsets**

E.g. p1=<a b ¬c X>.

Limited research:

Neg_GSP, PNSP

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

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?

<a> contains < a¬a >?   <a> contains < ¬a a¬a >?

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Non-occurrence behaviour analysis

(Negative sequence analysis)

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

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

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

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)

  $s = <e_1\ e_2 \ldots e_n>$

  i.e. $<a\ b\ (c,d)\ e>$, $<a\ \neg b\ c\ e>$

- Positive/Negative Sequence

  $s_p = <e_1\ e_2 \ldots e_n>$, *all elements are positive*

  $s_n = <e_1\ e_2 \ldots e_n>$, *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.* $<a\ (a,\neg b)\ c>$ *is not allowed.*

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

- # Negative Matching

Negative Matching. A negative sequence s_n=<e_1 e_2 ... e_k> *matches* a data sequence s=<d_1 d_2 ... d_m>, 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	<b $\neg c$ a>	No	<b f d c a>
S_2	<b $\neg c$ d a>	Yes	<b f d c a>

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

GA-NSP Algorithm

- Encoding

Sequence		Chromosome		
		$gene_1$	$gene_2$	$gene_3$
$<a\ b\ \neg(c,d)>$	$\Rightarrow$	$+a$	$+b$	$\neg(c,d)$

- Crossover

$parent1$	$b\ \neg c \updownarrow a$	$\Rightarrow$	$child1$	$b\ \neg c\ \mathbf{e}$
$parent2$	$\mathbf{d} \updownarrow \mathbf{e}$	$\Rightarrow$	$child2$	$\mathbf{d}\ a$

$parent1$	$b\ \neg c\ a \updownarrow$	$\Rightarrow$	$child1$	$b\ \neg c\ a\ \mathbf{d}\ \mathbf{e}$
$parent2$	$\updownarrow \mathbf{d}\ \mathbf{e}$	$\Rightarrow$	$child2$	$\mathbf{d}\ \mathbf{e}\ b\ \neg c\ a$

- Mutation

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

- Fitness & Dynamic Fitness

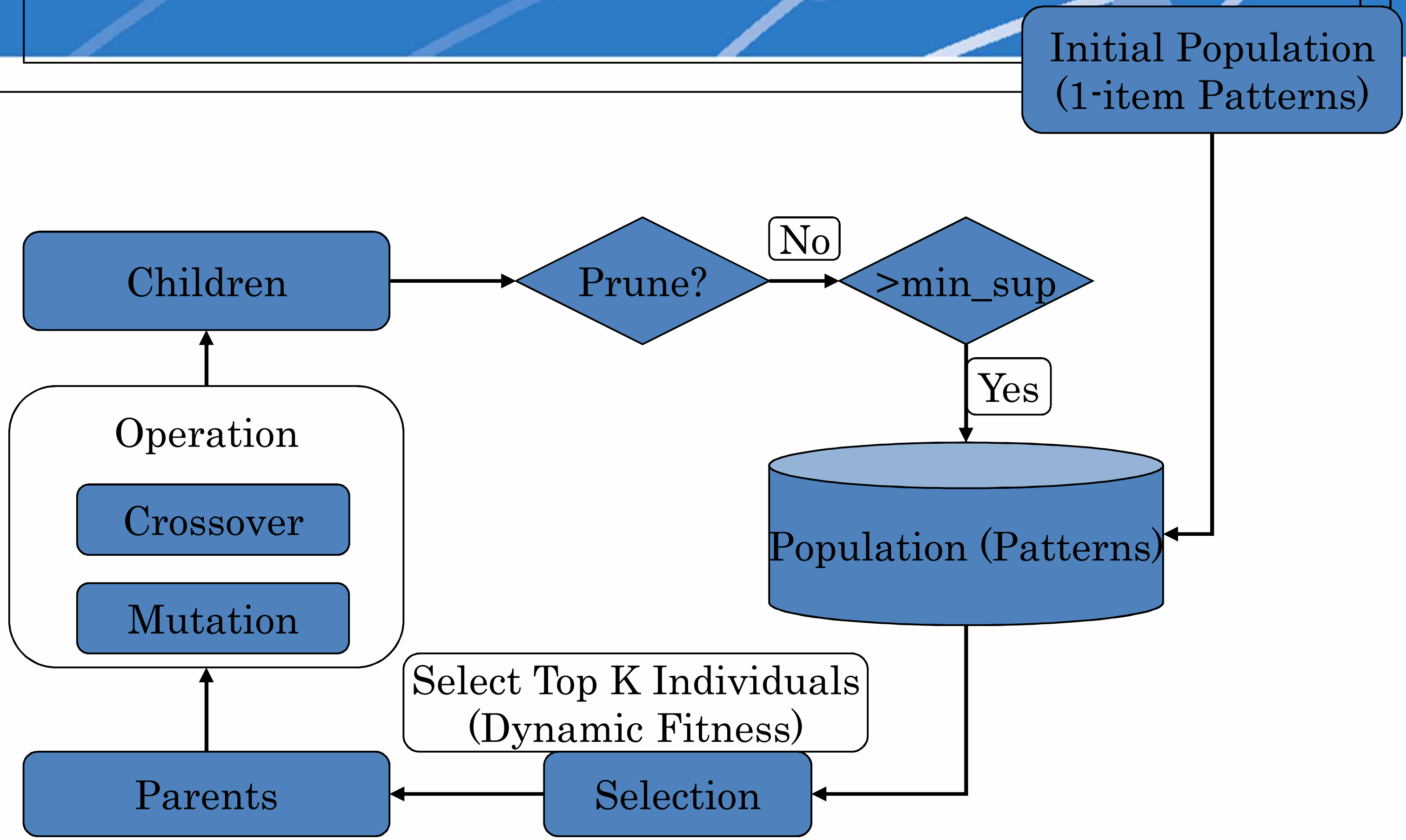
$$ind.fitness = (ind.support - min\_sup) \times DatasetSize. \quad (1)$$

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

- Selection

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

THE ADVANCED ANALYTICS INSTITUTE



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

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

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;
}
```

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

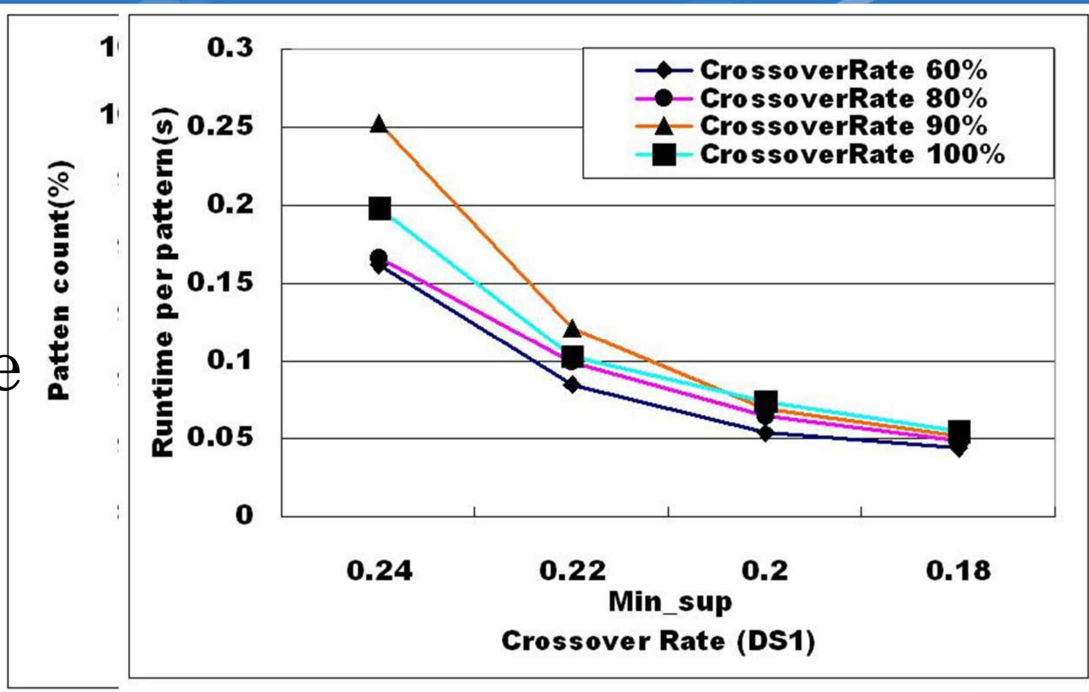
Experiments Result .1

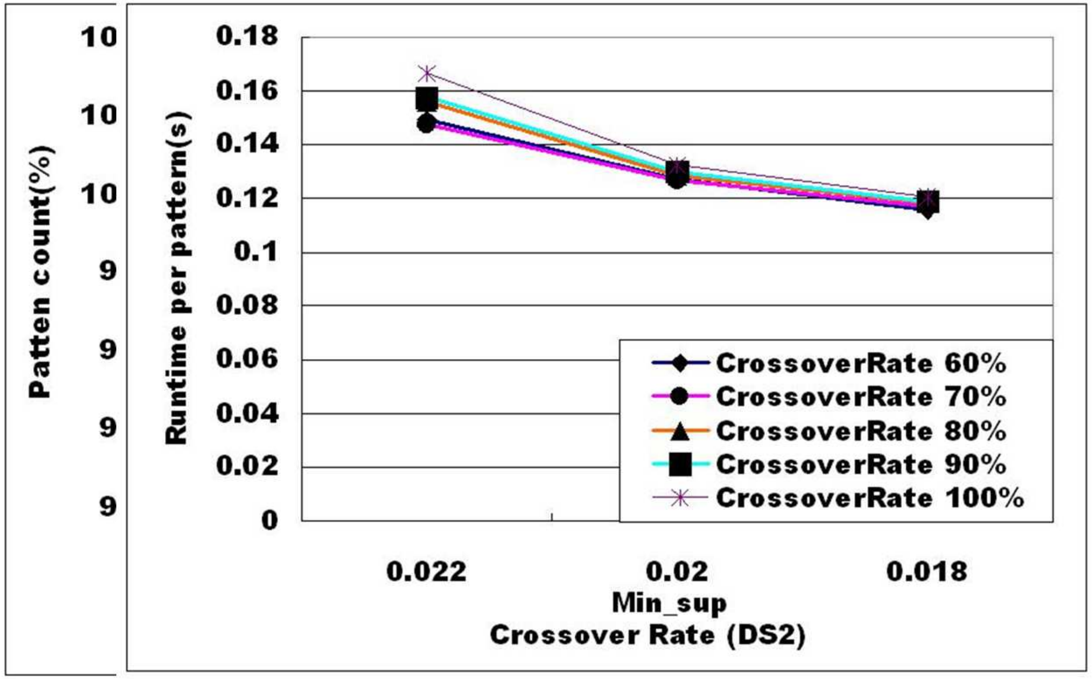
- Datasets

- *Dataset*1(*DS*1) 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.
- *Dataset*2(*DS*2) is C10.T2.5.S4.I2.5.DB100k.N10k.
- *Dataset*3(*DS*3) is C20.T4.S6.I8.DB10k.N2k.
- *Dataset*4(*DS*4) is real application data for insurance claims.

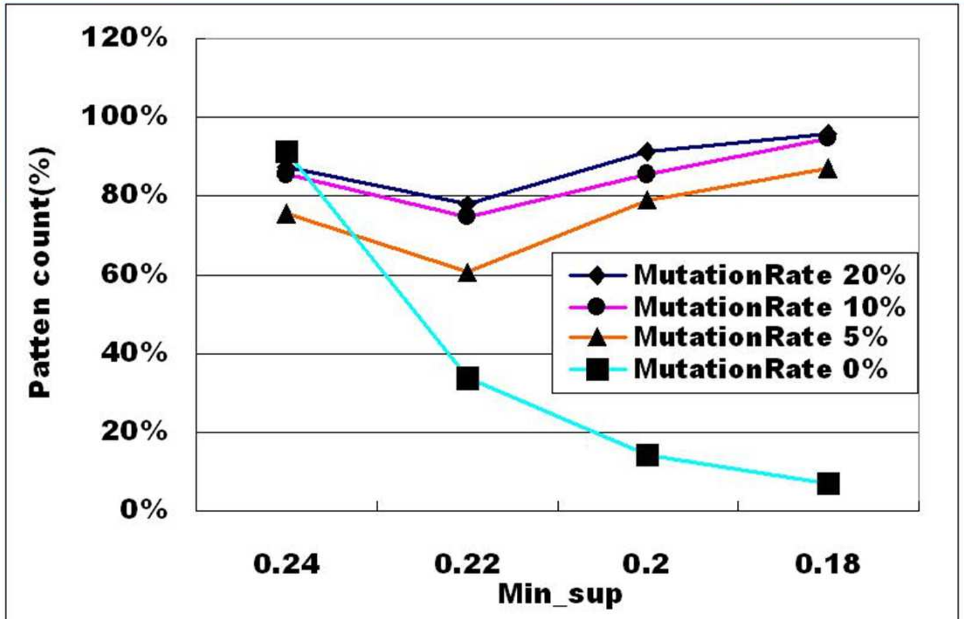
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- Crossover Rate

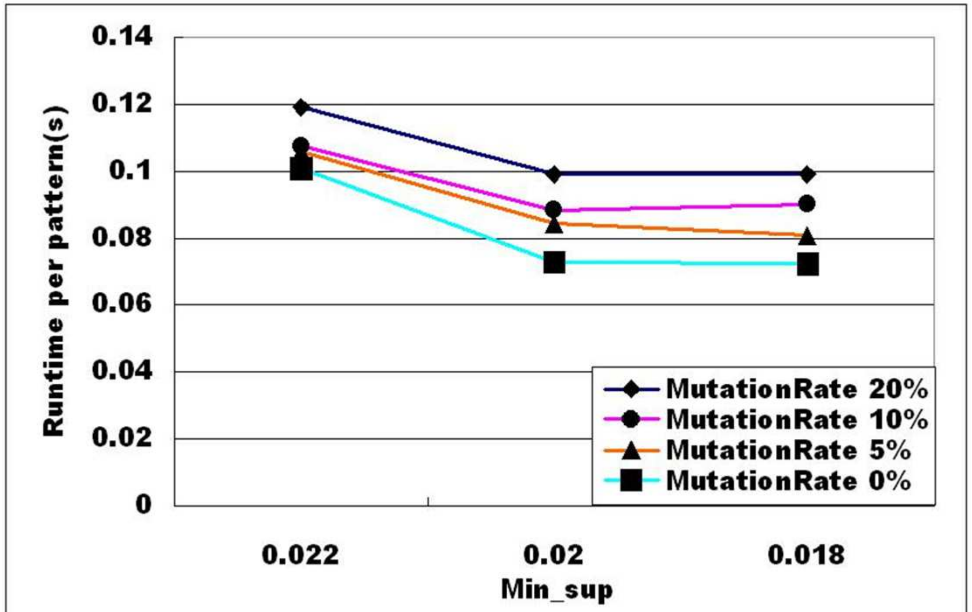




- Mutation Rate

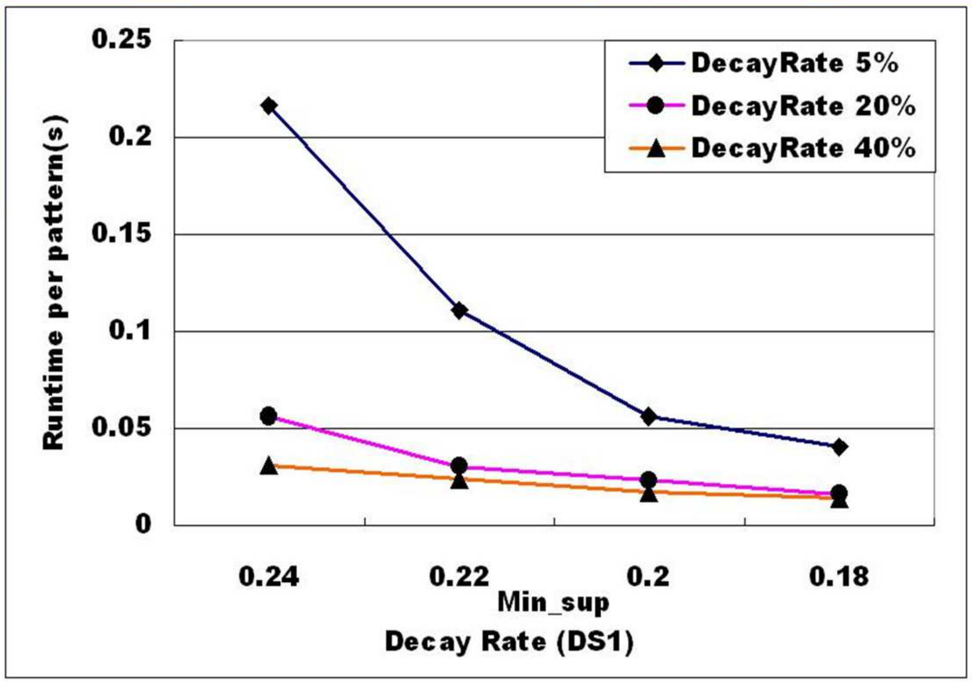


Mutation Rate (DS1)

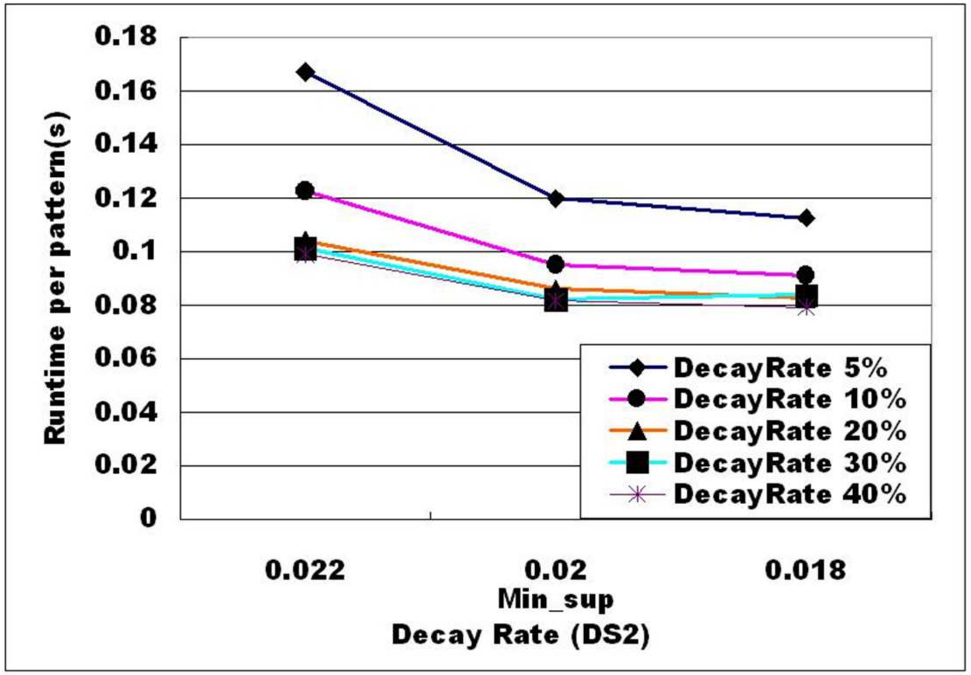


Mutation Rate (DS2)

• Decay Rate

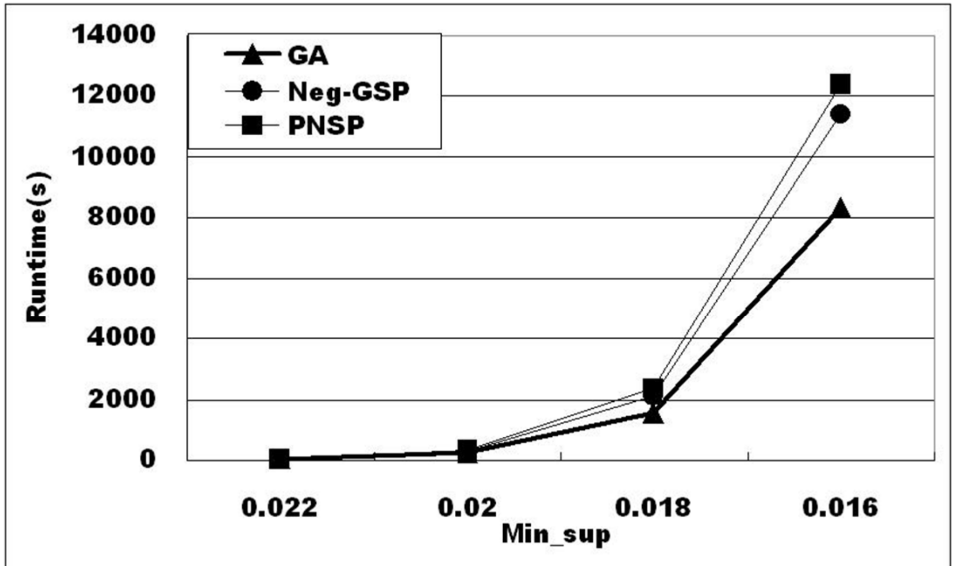


0.25
0.2
0.15
0.1
0.05
0
Runtime per pattern(s)
DecayRate 5%
DecayRate 20%
DecayRate 40%
0.24
0.22
0.2
0.18
Min_sup
Decay Rate (DS1)

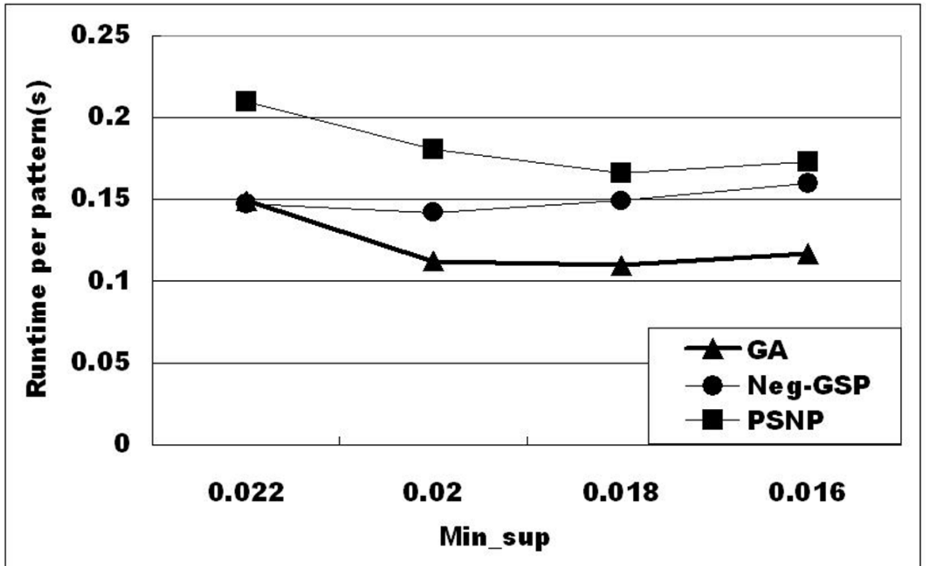


0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Runtime per pattern(s)
DecayRate 5%
DecayRate 10%
DecayRate 20%
DecayRate 30%
DecayRate 40%
0.022
0.02
0.018
Min_sup
Decay Rate (DS2)

- Comparison with PNSP, Neg-GSP



Runtime Comparison (DS2)



Runtime/Pattern Comparison (DS2)

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Classification of both positive and negative behavior patterns

•Huaifeng 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, 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.

Sequence classification

Let $\mathcal{T}$ be a finite set of *class labels*. A *sequential classifier* is a function

$$\mathcal{F} : \mathcal{S} \rightarrow \mathcal{T}. \tag{1}$$

In sequence classification, the classifier $\mathcal{F}$ is built on the base of frequent *classifiable sequential patterns* $\mathcal{P}$.

Definition 3.1 (Classifiable Sequential Pattern). *Classifiable Sequential Patterns (CSP) are frequent sequential patterns for the sequential classifier in the form of $p_a \Rightarrow \tau$, where p_a is a frequent pattern in the sequence database $\mathcal{S}$.*

Based on the mined classifiable sequential patterns, a sequential classifier can be formulised as

$$\mathcal{F} : s \xrightarrow{\mathcal{P}} \tau. \tag{2}$$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- 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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Table 5. The Number of Patterns in PS10 and PS05

	PS10 ($min\_sup = 0.1$)		PS05 ($min\_sup = 0.05$)	
	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

Negative Sequential Pattern Mining

The 20th ACM Conference on Information and Knowledge Management (CIKM 2011)

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

Xiangjun Dong
School of Information
Shandong Polytechnic
University, Jinan, China
dxj@spu.edu.cn

Zhigang Zheng, Longbing Cao
QCIS, AAI
University of Technology,
Sydney, Australia
{zgzheng,lbcao}@it.uts.edu.au

Yanchang Zhao
Centrelink
Australia
yanchang.zhao@centrelink.gov.au

Chengqi Zhang
QCIS
University of Technology,
Sydney, Australia
chengqi@it.uts.edu.au

Jinjiu Li
QCIS, AAI
University of Technology,
Sydney, Australia
jinjiu.li@eng.uts.edu.au

Wei Wei, Yuming Ou
QCIS, AAI
University of Technology,
Sydney, Australia
{wwei,yuming}@it.uts.edu.au

THE ADVANCED ANALYTICS INSTITUTE

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**

  $p(\neg e) = e$.

  $p(\langle a \neg (ab)\ c \rangle) = \langle a (ab)\ c \rangle$.

- **Max Positive Sub-sequence**

  $MPS(\langle a \neg (ab)\ c \rangle) = \langle ac \rangle$.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Constraints to Negative Sequence

Constraint 1. Frequency Constraint

This paper only focuses on the negative sequences ns whose positive partner is frequent, i.e., sup(p(ns))>=min_ sup.

Constraint 2. Format Constraint

Continuous negative elements in a NSC are not allowed.

< ¬(ab) c ¬ d> ✔

< ¬(ab) ¬ c d> ✗

Constraint 3. Element Negative Constraint

The minimum negative unit in a NSC is an element.

< ¬(ab) c d> ✔

<(¬ab) c d> ✗

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

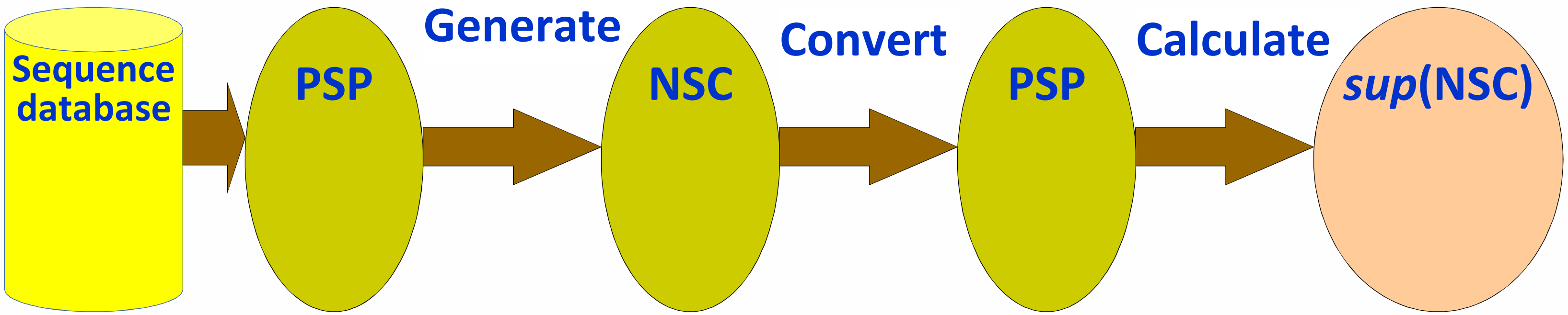
What does This Paper Do

E-NSP: **Only use corresponding PSP information to calculate the support of negative sequence, without additional 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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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=<d_1 d_2 ... d_t> be a data sequence, ns=<s_1 s_2 ... s_m> 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, $\forall(s_i, id(s_i)) \in EidS^-_{ns}$ $(1 \leqslant i \leqslant m)$, one of the following three holds:

(a) $(lsb=1)$ or $(lsb>1) \wedge p(s_1) \not\subseteq <d_1 \ldots d_{lsb-1}>$, when $i=1$,

(b) $(fse=t)$ or $(0<fse<t) \wedge p(s_m) \not\subseteq <d_{fse+1} \ldots d_t>$, when $i=m$,

(c) $(fse>0 \wedge lsb=fse+1)$ or $(fse>0 \wedge lsb>fse+1) \wedge p(s_i) \not\subseteq <d_{fse+1} \ldots d_{lsb-1}>$, when $1<i<m$,

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Negative Containment Definition

$$ns = <\ ns_{left},\quad \neg e,\quad ns_{right}\ >$$

$$MPS(ns_{left}) \qquad e \qquad MPS(ns_{right})$$

$$\subseteq \qquad \not\subset \qquad \subseteq$$

$$ds = <s_1, \ldots\ldots, s_i,\ s_{i+1}, \ldots s_{j-1},\ s_j, \ldots\ldots\ldots s_t>$$

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

Example: Negative Containment Definition

$ns = <a\neg bb(cde)>$. $ds = <a(bc)d(cde)>$.

$$< \underbrace{a} \quad \underbrace{\neg b} \quad \underbrace{b(cde)}>$$

$$\subseteq \qquad \not\subseteq \qquad \subseteq$$

$$ds = \ <\overbrace{a} \quad \overbrace{} \quad \overbrace{(bc)d(cde)}>.$$

***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 $1\text{-}negMSS_{ns}$.

Example *ns*=<*a¬bc¬d*>,

1-negMSSns ={<*a¬bc*>, <*ac¬d*>}

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Negative Conversion Strategy

Given a data sequence ds=<d_1 d_2 ... d_t>, and ns=<s_1 s_2 ... s_m>, 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-negMS* $\in$ *1-negMSS*$_{ns}$, *p(1-negMS)* $\not\subseteq$ *ds*.

Example ***ns*** **=<*a*¬*bb*¬*a*(*cde*)>, *ds*=<*a*(*bc*)*d*(*cde*)>.**

1-*negMSSns*={ <*a*¬*bb*(*cde*)> , <*ab*¬*a*(*cde*)> }

(1)MPS(ns)=< *ab*(*cde*)>$\subseteq$ds;

ds* contains *ns

(2)p(<*a*¬*bb*(*cde*)>)= <*abb*(*cde*)> $\not\subset$ ds;

p(<*ab*¬*a*(*cde*)>)= <*aba*(*cde*)> $\not\subset$ 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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Calculate the Support of NS

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

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

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

$$= sup(MPS(ns)) - |\bigcup_{i=1}^{n}\{p(1\text{-}negMS_i)\}| \qquad (2)$$

Example 10 $sup(<a\neg bc\neg de>)=sup(<ace>\})-|\{<abce>\}\cup\{<acde>\}|;$

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

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

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

Example 11 $sup(<a\neg bce>) = sup(<ace>) - sup(<abce>)$

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Calculate the Support of NS

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

Known

Calculate the union set of {*p*(1-*negMSi*)}. (*p*(1-*negMSi*) are frequent.)

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Negative Sequential Candidates Generation

Definition . 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, …,*$\lceil k/2 \rceil$, where $\lceil k/2 \rceil$ is a minimum integer that is not less than *k/*2.

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

m=1, <¬(ab) c d>,<(ab) ¬cd>,<(ab) c¬d>;

m=2, <¬(ab) c ¬d>.

An Example

Table 1: Example Data Set

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

Table 2: Example Result - Positive Patterns

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

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**
$\langle a\ a \rangle$	$\langle \neg a\ a \rangle$	$\langle a \rangle$, $\langle a\ a \rangle$	1
	$\langle a\ \neg a \rangle$	$\langle a \rangle$, $\langle a\ a \rangle$	1
$\langle a\ b \rangle$	$\langle \neg a\ b \rangle$	$\langle b \rangle$, $\langle a\ b \rangle$	0
	$\langle a\ \neg b \rangle$	$\langle a \rangle$, $\langle a\ b \rangle$	1
$\langle a\ c \rangle$	$\langle \neg a\ c \rangle$	$\langle c \rangle$, $\langle a\ c \rangle$	0
	$\langle a\ \neg c \rangle$	$\langle a \rangle$, $\langle a\ c \rangle$	**2**
$\langle b\ c \rangle$	$\langle \neg b\ c \rangle$	$\langle c \rangle$, $\langle b\ c \rangle$	0
	$\langle b\ \neg c \rangle$	$\langle b \rangle$, $\langle b\ c \rangle$	1
$\langle (ab) \rangle$	$\langle \neg (ab) \rangle$	$\langle (ab) \rangle$	**3**
$\langle a\ (ab) \rangle$	$\langle \neg a\ (ab) \rangle$	$\langle (ab) \rangle$, $\langle a\ (ab) \rangle$	0
	$\langle a\ \neg (ab) \rangle$	$\langle a \rangle$, $\langle a\ (ab) \rangle$	**2**
$\langle a\ b\ c \rangle$	$\langle \neg a\ b\ c \rangle$	$\langle b\ c \rangle$, $\langle a\ b\ c \rangle$	0
	$\langle a\ \neg b\ c \rangle$	$\langle a\ c \rangle$, $\langle a\ b\ c \rangle$	0
	$\langle a\ b\ \neg c \rangle$	$\langle a\ b \rangle$, $\langle a\ b\ c \rangle$	1
	$\langle \neg a\ b\ \neg c \rangle$	$\langle b \rangle$, $\langle a\ b \rangle$, $\langle b\ c \rangle$	0

Experiment and Evaluation

Data Sets

Four source datasets including both real data and synthetic datasets generated by IBM data generator. Partition these datasets to 14 datasets according to different data factors.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

5. Negative Behavior Analysis

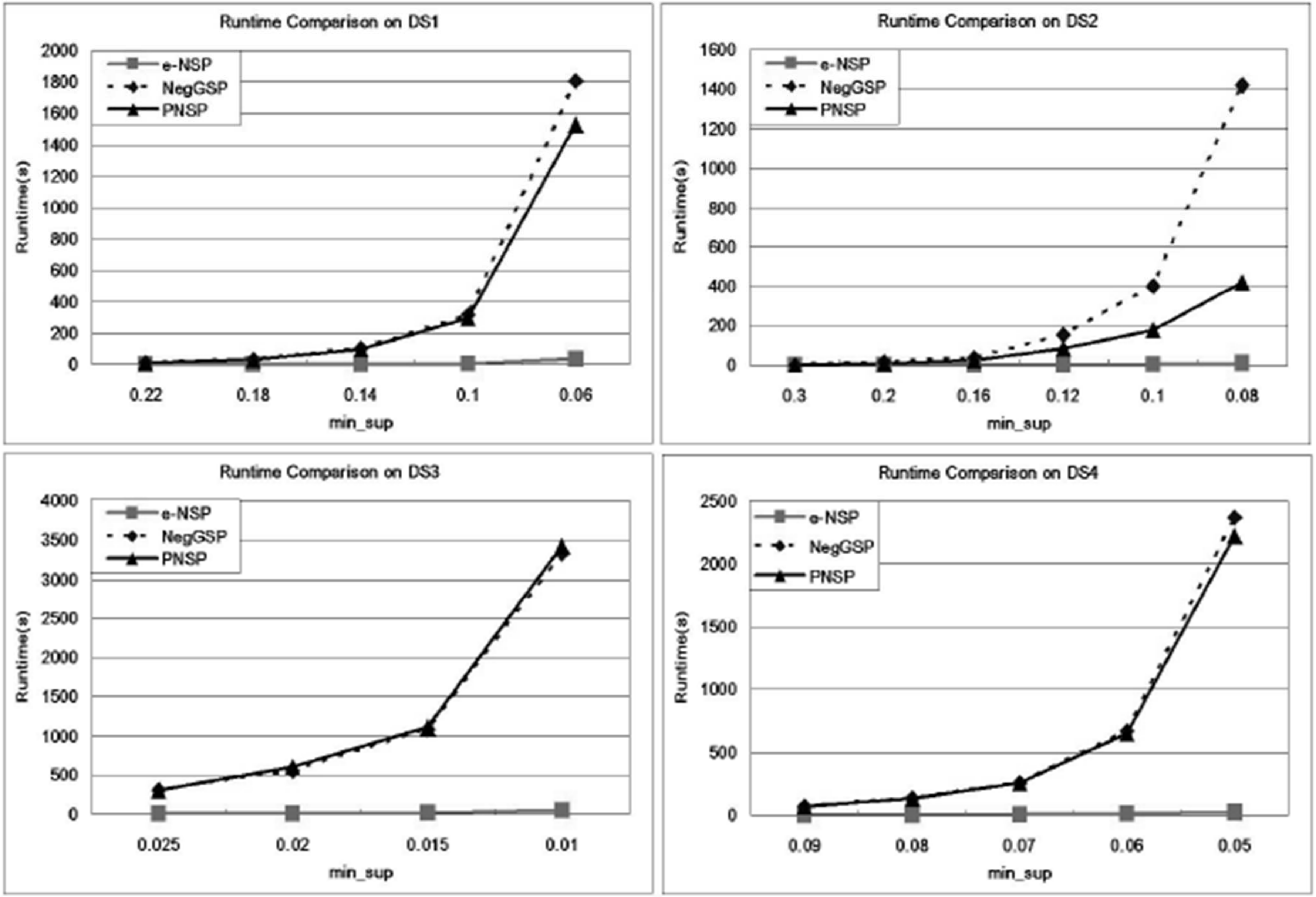
An Example

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	**C4**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	**C12**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**T8**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**T12**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**S12**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**S18**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**I10**.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**I14**.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.**N200**	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.**N400**	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%

Experiment and Evaluation

Computational Cost



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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

10. Coupled/Group Behavior Analysis

THE ADVANCED ANALYTICS INSTITUTE

References

- Can Wang, Zhong She, Longbing Cao. Coupled Clustering Ensemble: Incorporating Coupling Relationships Both between Base Clusterings and Objects, ICDE2013.
- Longbing Cao, Yuming Ou, Philip S Yu. Coupled Behavior Analysis with Applications, IEEE Trans. on Knowledge and Data Engineering, 24(8): 1378-1392 (2012).
- 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.
- 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.
- Can Wang, Mingchun Wang, Zhong She, Longbing Cao. CD: A Coupled Discretization Algorithm, PAKDD2012, 407-418

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

What is Coupled Behavior?

Intelligent Transport Systems

Physical world

Virtual world

Problem-solving world

Relationship crossing behaviors

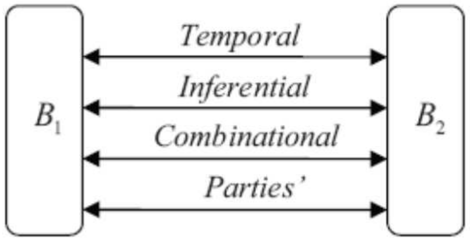


Figure 6: Relationships between Multiple Behaviors

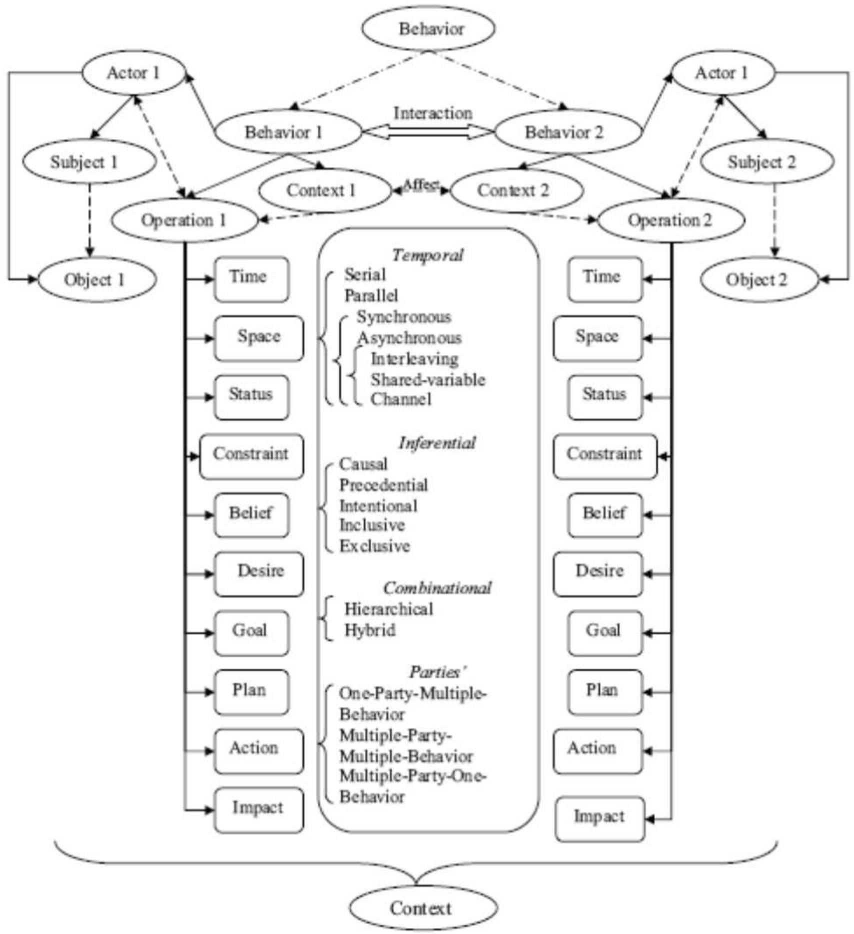


Figure 7: Relationships between behaviors

Behavior Coupling Types

- Logic/semantic relation based behavior coupling
- Statistical/Probabilistic relation based behavior coupling

Logic/Semantic Relation based Group Behavior Analysis

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

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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Coupling relationships

- From temporal aspect
  - Serial Coupling: $TS_1; TS_2; \cdots ; TS_n$
  - Interleaving Coupling: $TS_1 : TS_2 : \cdots : TS_n$
  - Shared-variable Coupling: $TS_1 ||| TS_2 ||| \cdots ||| TS_n$
  - Channel System Coupling: $TS_1 \mid TS_2 \mid \cdots \mid TS_n$
  - Synchronous Coupling: $TS_1 \parallel TS_2 \parallel \cdots \parallel TS_n$
- From inferential aspect
  - Causal Coupling: $TS_1 \rightarrow TS_2$
  - Precedential Coupling: $TS_1 \Rightarrow TS_2$
  - Intentional Coupling: $TS_1 \twoheadrightarrow TS_2$
  - Inclusive Coupling: $TS_1 \mapsto TS_2$
  - Exclusive Coupling: $TS_1 \oplus TS_2$
- From combinational aspect
  - Hierarchical Coupling: $f(g(TS_1, TS_2, \cdots , TS_n))$
  - Hybrid Coupling: $f(TS_1).g(TS_2)$, $f(TS_1)^*$, $(TS_1)^\omega$
  - One-Party-Multiple-Behavior Coupling: $f(TS_1, TS_2, \cdots , TS_n)^{[A_1]}$
  - Multiple-Party-One-Behavior Coupling: $f(TS_1)^{[A_1 A_2 \cdots , A_n]}$
  - Multiple-Party-Multiple-Behavior Coupling: $f(TS_1, TS_2 , \cdots , TS_n)^{[A_1 A_2 \cdots A_n]}$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Basic Behavior Patterns

- Tracing: Different actions with sequential order.

$$\{a_1, a_2, \cdots, a_n\}$$

- Consequence: Different actions have causalities in occurrence.

$$\{a_i \to a_j\}$$

- Synchronization: Different actions occur at the same time.

$$\{a_1 \leftrightarrow, \cdots, \leftrightarrow a_n\}$$

- Combination: Different actions occur in concurrency.

$$\{a_1 \| a_2 \|, \cdots, \| a_n\}$$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

- Exclusion: Different actions occur mutually exclusively.

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

- Precedence: Different actions have required precedence

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

And more to be explored…

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Logic Coupling Based Combined Pattern Pairs

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

$$\mathcal{E}: \begin{cases} X_{\mathrm{p}} \rightarrow T_1 \\ X_{\mathrm{p}} \wedge X_{\mathrm{e}} \rightarrow T_2 \end{cases},$$

Group 1 behavior

Group 2 behavior

where $X_{\mathrm{p}} \neq \emptyset$, $X_{\mathrm{e}} \neq \emptyset$ *and* $X_{\mathrm{p}} \cap X_{\mathrm{e}} = \emptyset$.

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Logic Coupling Based 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.*

Group 1 behavior

$$\mathcal{S}: \begin{cases} X_\mathrm{p} \rightarrow T_1 \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \rightarrow T_2 \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \rightarrow T_3 \\ \cdots \\ X_\mathrm{p} \wedge X_{\mathrm{e},1} \wedge X_{\mathrm{e},2} \wedge \cdots \wedge X_{\mathrm{e},\mathrm{k}-1} \rightarrow T_k \end{cases},$$

Group K behavior

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.*

Multi-group Pattern Relation

- Type A: Demographics differentiated combined pattern
  - Customers with the same actions but different demographics

    → different classes/business impact

$$\text{Type A:} \begin{cases} A_1 + D_1 & \rightarrow & \text{quick payer} \\ A_1 + D_2 & \rightarrow & \text{moderate payer} \\ A_1 + D_3 & \rightarrow & \text{slow payer} \end{cases}$$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Multi-group Pattern Relation

- Type B: Action differentiated combined pattern
  - Customers with the same demographics but taking different actions

    → different classes/business impact

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Multiple Group Pattern Relations

An Example of Combined Pattern Clusters

Clusters	Rules	X_p demographics	X_e arrangements	X_e repayments	T	Cnt	$Conf$ (%)	I_r	I_c	$Lift$	$Cont_p$	$Cont_e$	$Lift$ of $X_p \to T$	$Lift$ of $X_e \to T$
$\mathcal{P}_1$	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	132	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

Multi-Group Combined Patterns

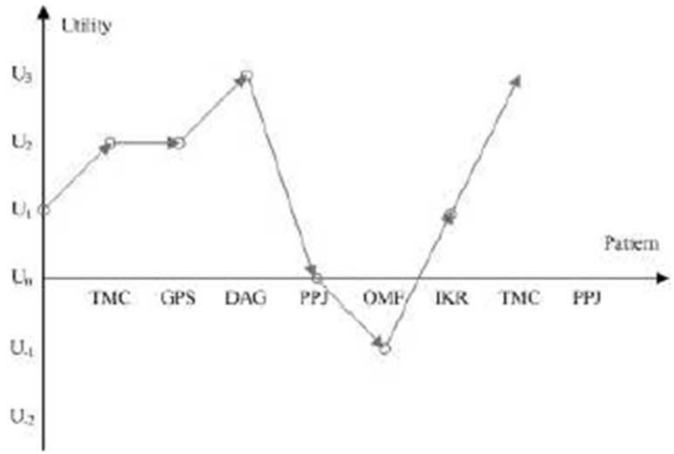


Figure 2: Pattern Evolution Chart

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Multi-Group Combined Patterns

$$\begin{cases} PLN \rightarrow T \\ PLN, DOC \rightarrow T \\ PLN, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC \rightarrow T \\ PLN, DOC, DOC, DOC, REA \rightarrow T \\ PLN, DOC, DOC, DOC, REA, IES \rightarrow T \end{cases}$$

- Divergence vs. convergence of group behaviors

Statistical/Probabilistic Behavior Coupling Analysis

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).

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Feature Matrix

I actors (customers): $\{\mathscr{E}_1, \mathscr{E}_2, \ldots, \mathscr{E}_I\}$

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

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

Behavior Feature Matrix:

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Behavior Intra-relationship

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

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

$$|\theta_j(\cdot)| \geq \theta_0 \tag{2}$$

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.

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Behavior Inter-relationship

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

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

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

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.*

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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* $(i_1 \neq i_2) \vee (j_1 \neq j_2) \wedge (1 \leq i_1, i_2 \leq I) \wedge (1 \leq j_1, j_2 \leq J_{max})$

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

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

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Behavior 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.*

$$f(\cdot) ::= f(\theta(\cdot), \eta(\cdot)), \tag{9}$$

$$g(\cdot)|(f(\cdot) \geq f_0) \geq g_0 \tag{10}$$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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*.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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

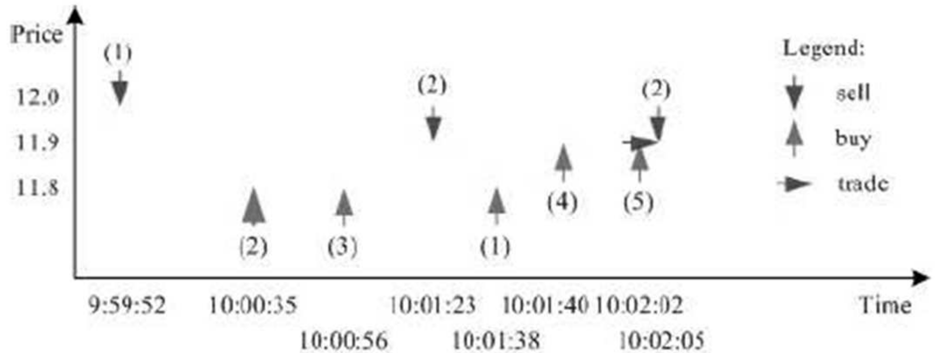
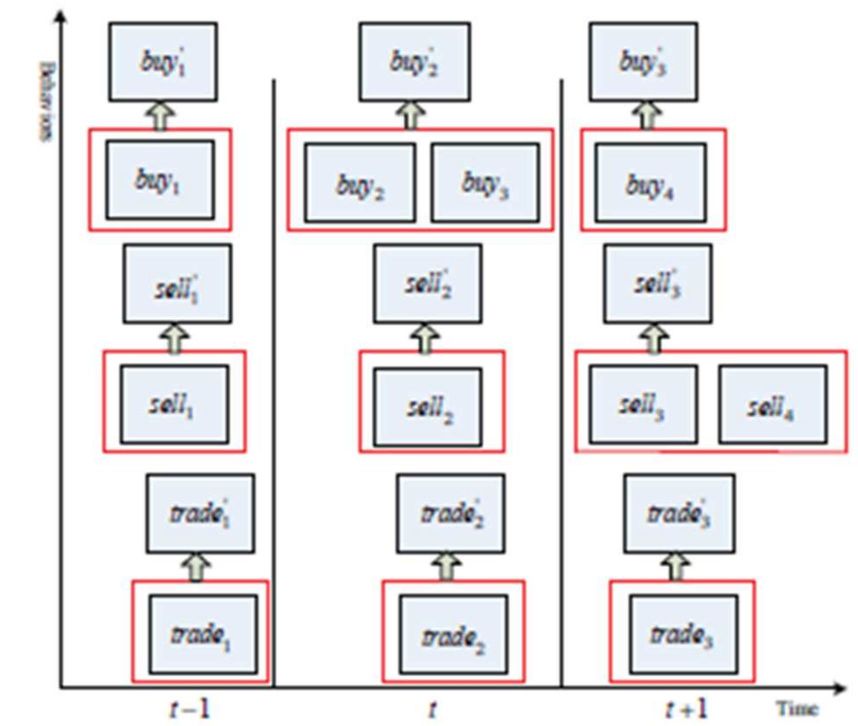


Fig. 1. Coupled Trading Behaviors

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE



(a) An Example of Coupled Trading Behaviors in Stock Markets

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Construct behavior sequences

$$\left\{ \frac{Actor_i - Operation_i}{Attributes_i} \xrightarrow{\eta} \frac{Actor_j - Operation_j}{Attributes_j} \right\}_{i,j=1;winsize}^{I,J} \qquad (12)$$

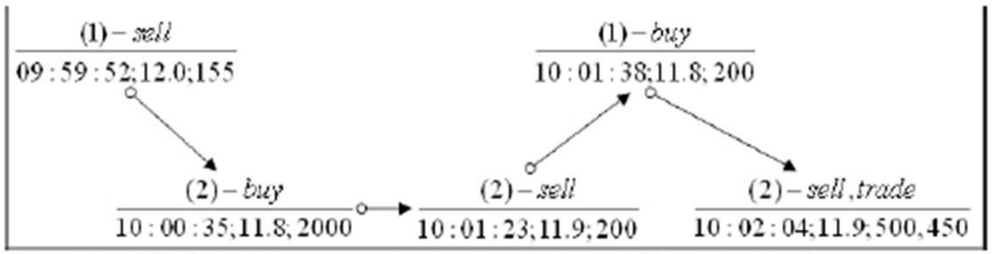


Fig. 2. Behavior sequences - Data Structure 1

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

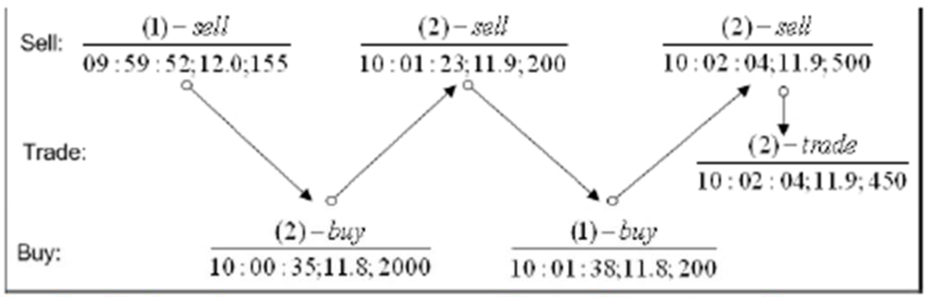


Fig. 3. Behavior sequences - Data Structure 2

CHMM Based Coupled Sequence Modeling

- Coupled behavior sequences
  - Multiple sequences

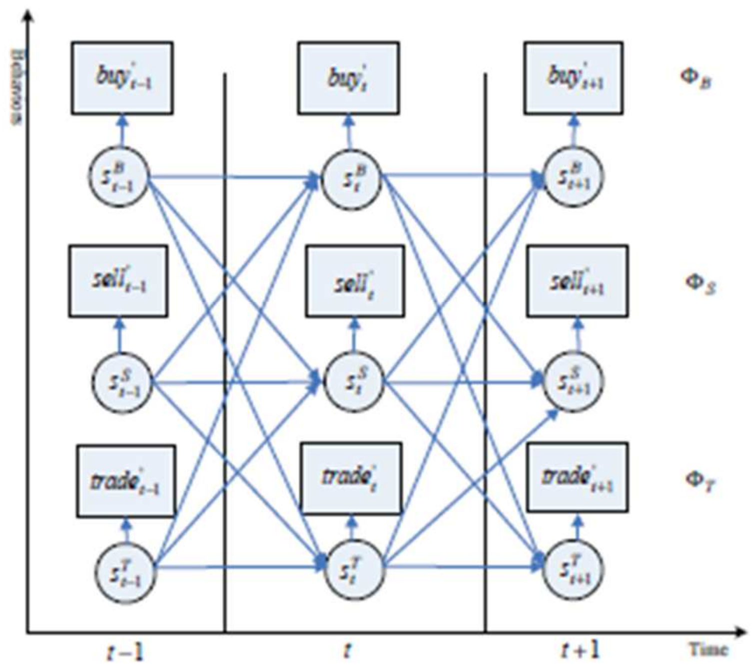
$$\Phi_1 = \{\phi_{11}, \ldots, \phi_{1T}\}$$
$$\Phi_2 = \{\phi_{21}, \ldots, \phi_{2F}\}$$
$$\Phi_C = \{\phi_{C1}, \ldots, \phi_{CG}\}$$

  - Coupling relationship

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

  - Behavior properties

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



(b) The Structure of the CHMM

CBA - CHMM

$$CBA\ problem \rightarrow CHMM\ model \tag{15}$$

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

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

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Framework: abnormal CBA

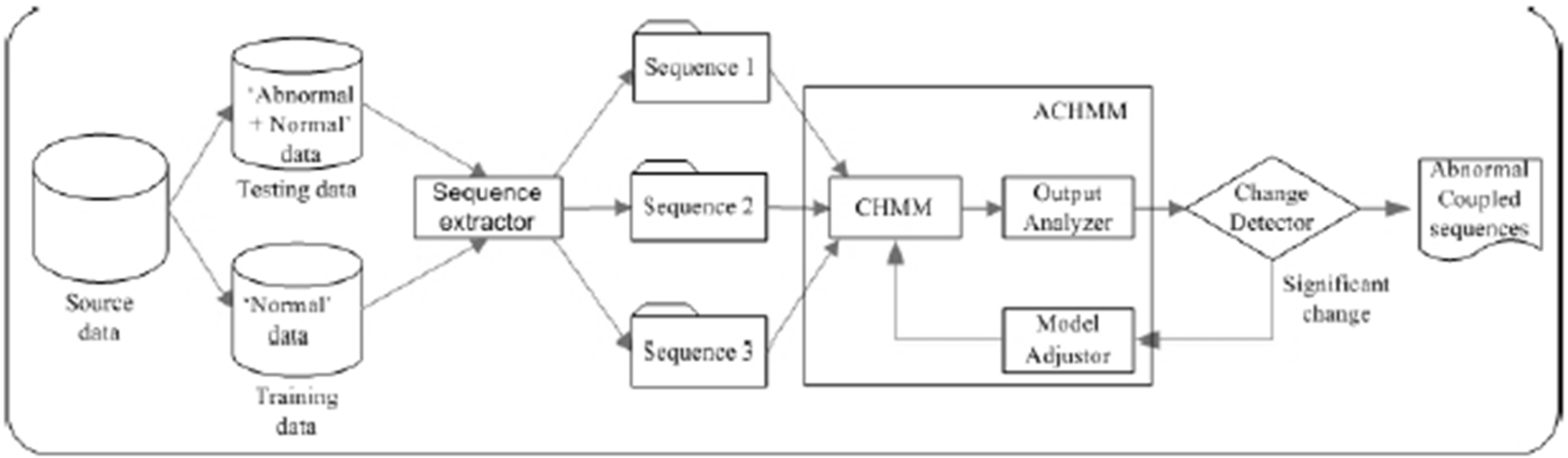


Fig. 5. Framework of abnormal coupled behavior detection

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

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)

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

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

$$a(t_i) = \{buy \mid sell \mid trade\}$$

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

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

Interval Activity (IA)

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

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

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

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Adaptive CHMM for Detecting Sequence Changes

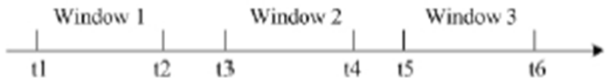


Figure 3: Update Point of ACHMM

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

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

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

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

The Algorithm

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 sequnce 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

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

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Algorithm 2 Detecting abnormal trading sequences

Step 1: Construct trading sequences including training sequences $Seq_1, Seq_2, \cdots, Seq_K$ and test sequences $Seq_1^{'}, Seq_2^{'}, \cdots, Seq_{K'}^{'}$.

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:

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

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

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

$$D_i = \frac{\mu - Pr(Seq_i^{'}|\mathcal{M})}{\sigma} \tag{22}$$

Consequently, $Seq_i^{'}$ is an exceptional pattern, if it satisfies:

$$D_i > \psi_0 \tag{23}$$

where ψ_0 is a given threshold.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- Benchmark Models
  - HMM-B
  - HMM-S
  - HMM-T
  - IHMM
  - CHMM
  - ACHMM

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

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

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

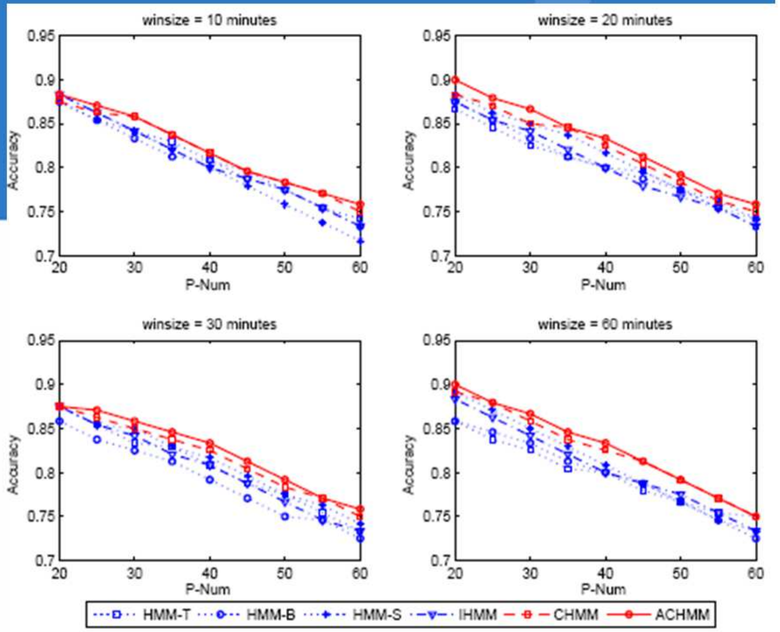


Figure 4: Accuracy of Six Models

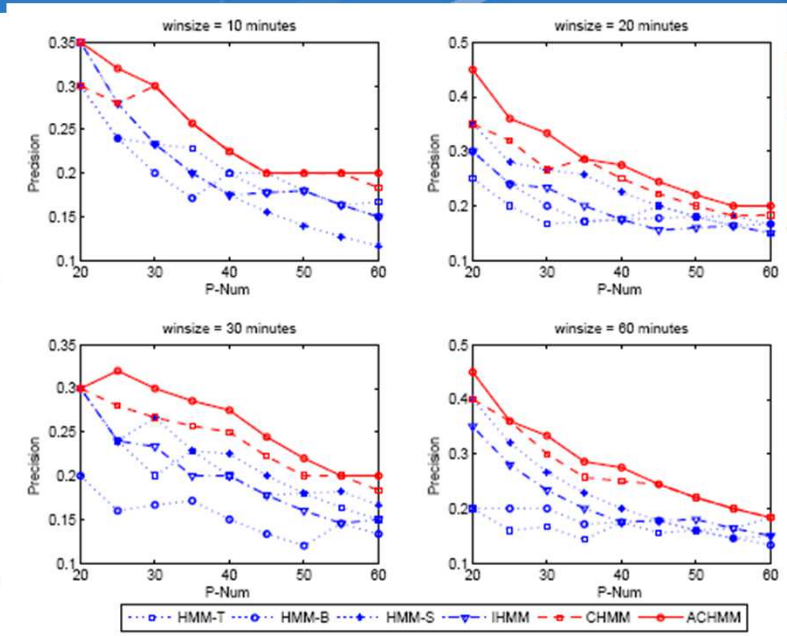


Figure 5: Precision of Six Models

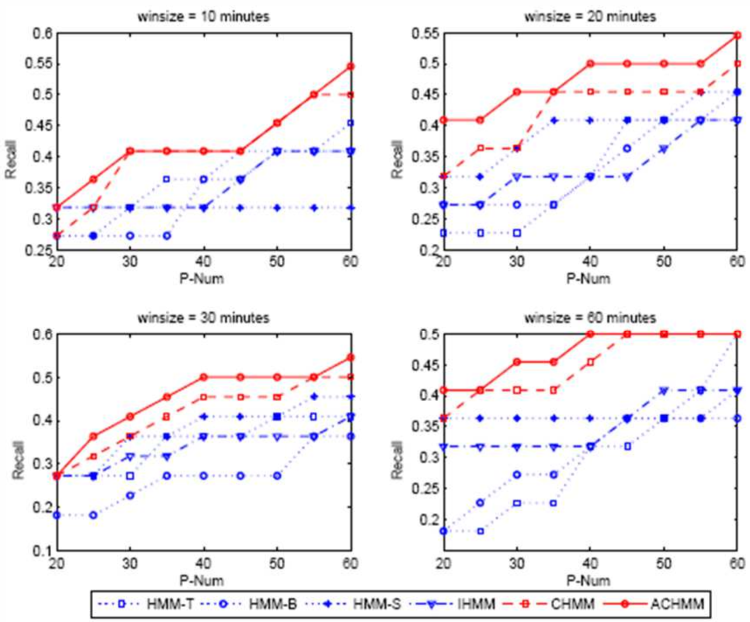


Figure 6: Recall of Six Models

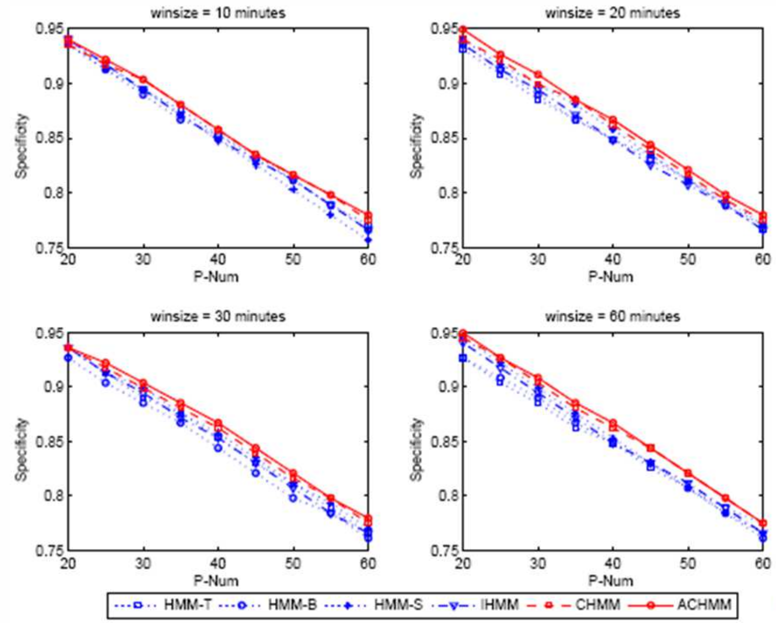


Figure 7: Specificity of Six Models

- Business Performance

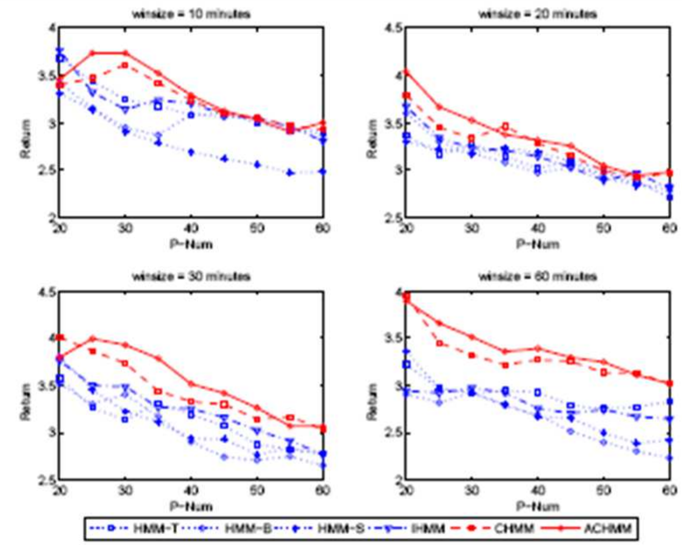


Fig. 9. Return of Six Models

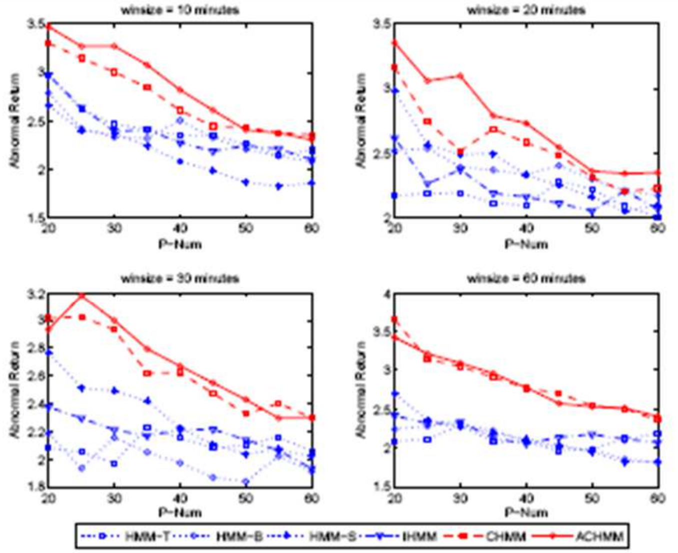


Fig. 10. Abnormal Return of Six Models

• Computational cost

TABLE 5
Computational performance

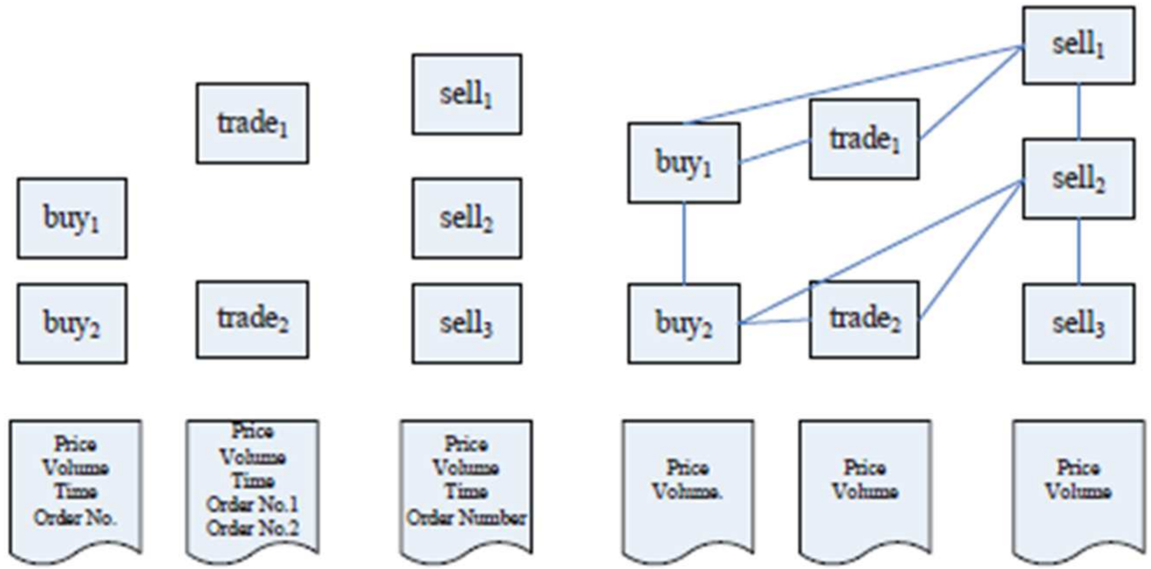
		IHMM	CHMM	ACHMM
winsize =10 (m)	Training time (s)	0.574	11.978	11.988
	Test time (s)	0.056	1.296	3.576
winsize =20 (m)	Training time (s)	0.256	4.929	4.933
	Test time (s)	0.047	0.655	3.486
winsize =30 (m)	Training time (s)	0.206	4.121	4.119
	Test time (s)	0.042	0.447	2.429
winsize =60 (m)	Training time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

Conditional Probability Distribution-based Coupled Behavior Analysis

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.

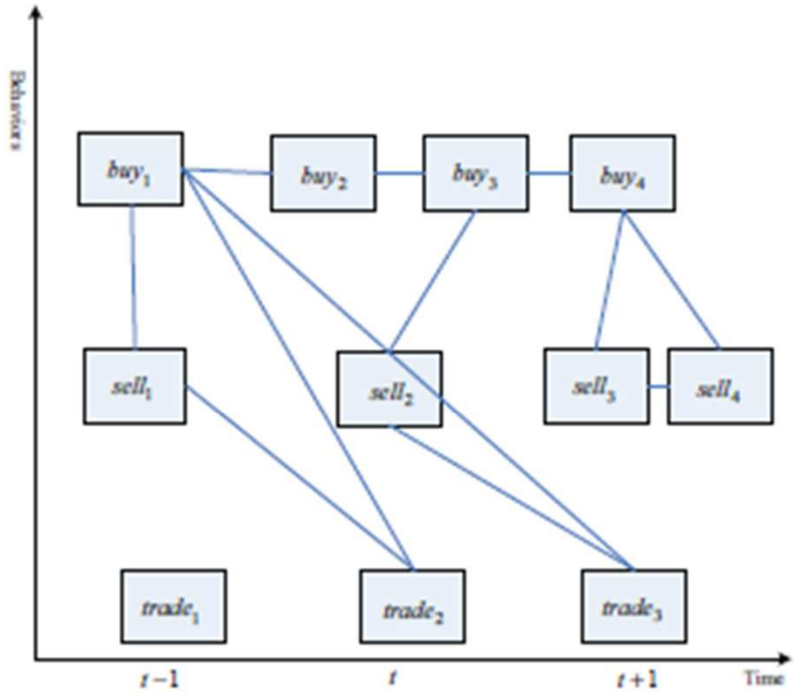
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE



(a) The Coupled Behaviors with Reference and Analysis Properties.

(b) Link Generation Using Reference Properties.

Graph-based Coupled Behavior Presentation



(c) The Structure of Graph-based Coupled Behavior Model

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

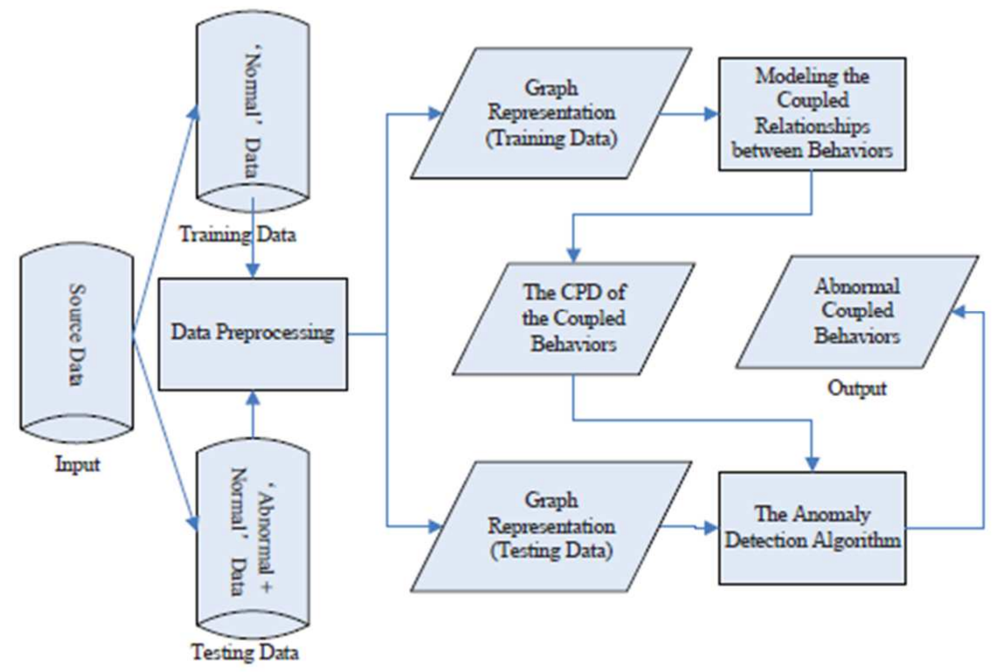
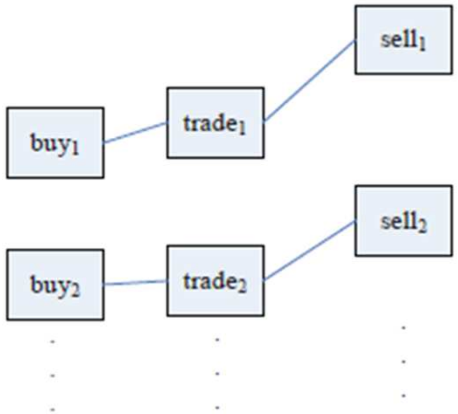


Figure 2: The Work Flow of the Proposed Framework.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Propositional Coupled Behavior



(a) An Example of the Subgraphs for Each Target Behavior

	$X^{(t)}$	RF_1	RF_2	$\cdots$	RF_n
$trade_1$	x_1	rf_{11}	rf_{21}	$\cdots$	rf_{n1}
$trade_2$	x_2	rf_{12}	rf_{22}	$\cdots$	rf_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(b) An Example of the Relational Features for Each Target Behavior

- CPD

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

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

- Estimate *p(RF|X)*

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

- Estimate CPD $p(X^{(t)}|RF_1, \cdots, RF_n)$

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

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- CBA problem → CPD problem

$$
\begin{aligned}
CBA\ problem &\rightarrow SRL\ Modeling \qquad (5)\\
f(\theta(\cdot), \eta(\cdot)) &\rightarrow the\ CPD\ p(X^{(t)}|RF_1, \cdots, RF_n) \qquad (6)
\end{aligned}
$$

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Relational Bayesian Classifiers (RBCs)

The CPD $p(X^{(t)}|RF_1, \cdots, RF_n)$ can be estimated as

$$\alpha p(X^{(t)})p(RF_1|X^{(t)})p(RF_2|X^{(t)}) \cdots p(RF_n|X^{(t)}) \quad (8)$$

where α is the normalized constant.

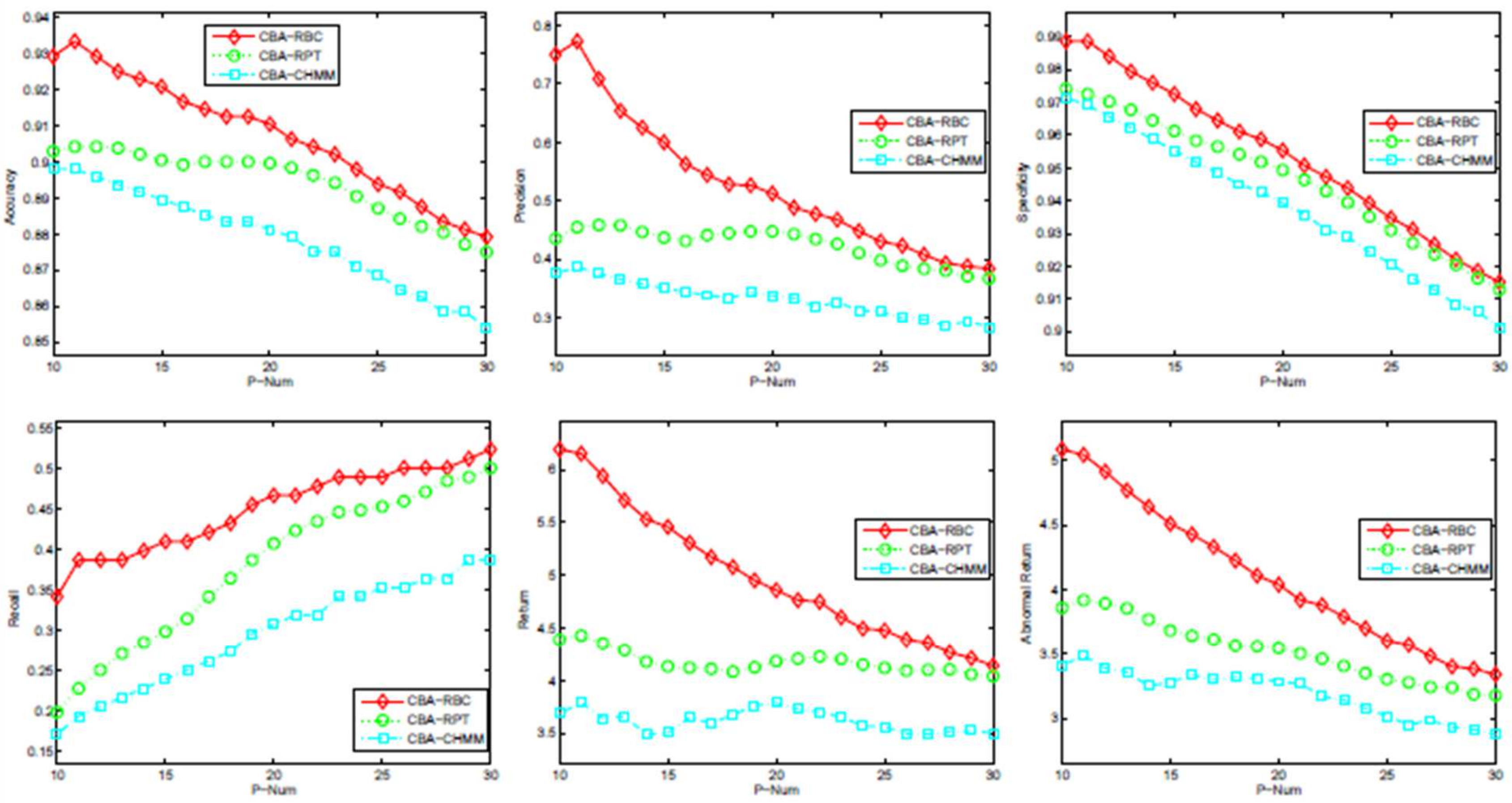
- Conditional likelihood:

$$CL(\mathbf{b}^{\mathbf{k}}) = \prod_{\mathbf{b}_{\mathbf{i}}^{(\mathbf{t})} \in \mathbf{b}^{\mathbf{k}}} p(X^{(t)} = x_{b_i^{(t)}} | rf_{1i}, rf_{2i}, \cdots, rf_{ni}; M)$$

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Relational Probability Trees (RPTs)

The RPT algorithm uses aggregation functions (e.g, mode, count, proportion and degree) to transform the relational features of subgraphs to propositional features and use these features to construct probability trees.



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

The 20th ACM Conference on Information and Knowledge Management (CIKM 2011)

Coupled Nominal Similarity in Unsupervised Learning

Can Wang, Longbing Cao, Mingchun Wang,
Jinjiu Li, Wei Wei, Yuming Ou

University of Technology, Sydney, Australia

Wednesday, 26 Oct. 2011, Glasgow,UK

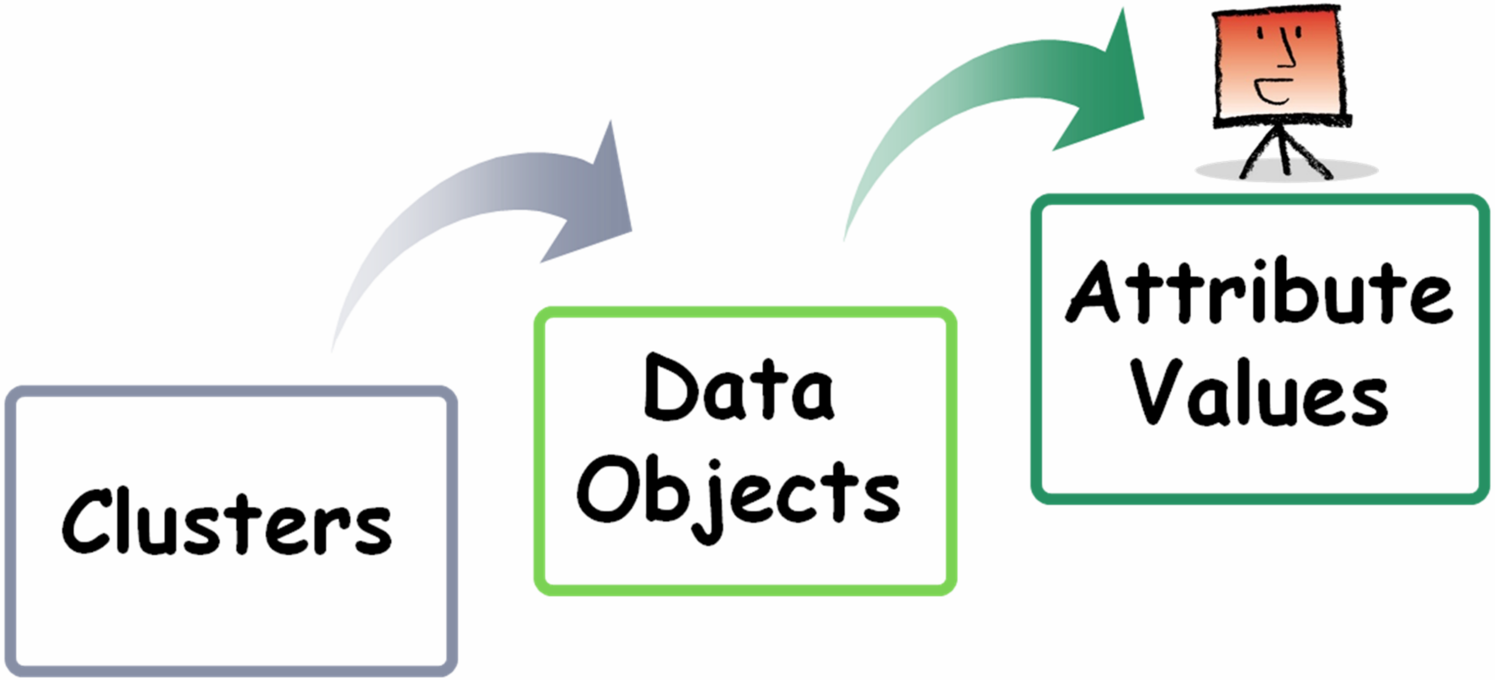
UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Coupled Nominal Similarity

- **Similarity Analysis**
- **Related Work**
- **Motivation: Example**
- **Coupled Nominal Similarity**
  - Intra-coupled Interaction
  - Inter-coupled Interaction
- **Theoretical Analysis**
- **Back to Example**
- **Experiment and Evaluation**
- **Conclusion**

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Similarity Analysis



The **more** two objects **resemble**

The **larger** the **similarity**

Similarity Analysis

Attribute Values

Continuous

$$\delta(x, y) = | x - y |$$

L^P Distance
- Manhattan
- Euclidean
- Minkowski
- Chebyshev

Nominal

- **Supervised**
  - Value Distance Matrix
  - Modified Value Distance Matrix
  - Heterogeneous
- **Unsupervised**
  …… (Our Focus)

UTS:AAi

THE ADVANCED ANALYTICS INSTITUTE

Related Work

Unsupervised Nominal Similarity

- **Matching Coefficient**
  - ✓ Simple Matching (SM)
  - ✓ Jaccard
  - ✓ Russell and Rao
- **Value Frequency Distribution**
  - ✓ Occurrence Frequency
  - ✓ Goodall
  - ✓ Anderberg
- **Feature Dependency Aggregation**
  - ✓ Iterated Contextual
  - ✓ **Ahmad and Dey (AD)**

Motivation

Movie	Director	Actor	Genre	Class
Godfather II	Scorsese	De Niro	Crime	G_1
Good Fellas	Coppola	De Niro	Crime	G_1
Vertigo	Hitchcock	Stewart	Thriller	G_2
N by NW	Hitchcock	Grant	Thriller	G_2
Bishop's Wife	Koster	Grant	Comedy	G_2
Harvey	Koster	Stewart	Comedy	G_2

Matching Coefficient:

Similar directors

Sim(*Scorsese, Coppola*) = 0;

Sim(*Koster, Hitchcock*) = Sim(*Koster, Coppola*).

Former, Larger

Value Frequency Distribution:

Sim (*Scorsese, Coppola*) < Sim(*Koster, Hitchcock*)

Former, Larger

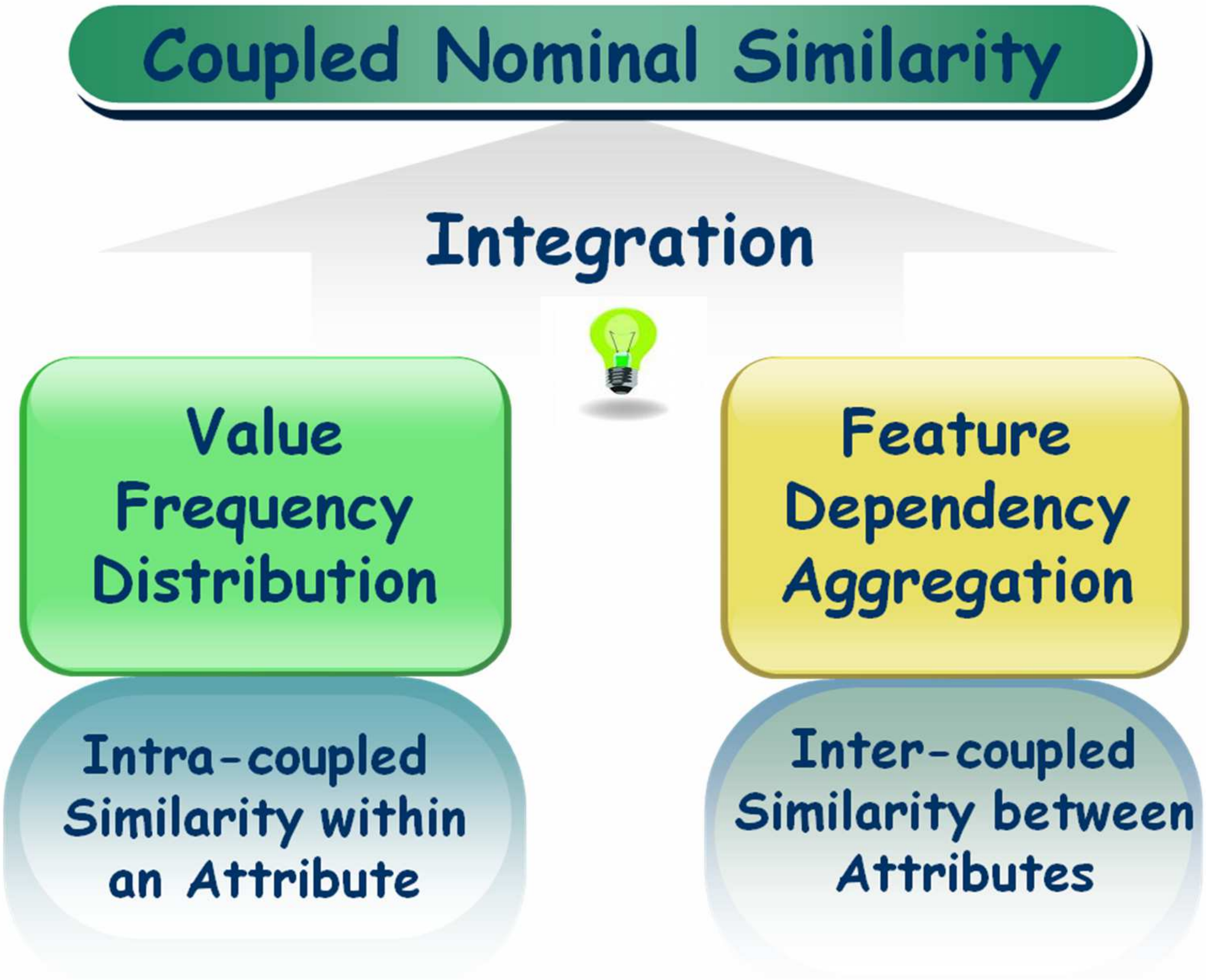
Feature Dependency Aggregation:

Sim (*Koster, Koster*) = Sim(*Scorsese, Coppola*)

Former, Larger

UTS:AAi
…LYTICS INSTITUTE

Coupled Nominal Similarity



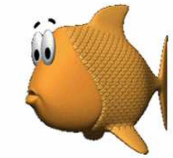
Coupled Nominal Similarity

U \ A	a_1	a_2	a_3
u_1	A_1	B_1	C_1
u_2	A_2	B_1	C_1
u_3	A_2	B_2	C_2
u_4	A_3	B_3	C_2
u_5	A_4	B_3	C_3
u_6	A_4	B_2	C_3

DEFINITION 4.1. *Given an information table S, the* ***Coupled Attribute Value Similarity (CAVS)*** *between attribute values x and y of feature a_j is:*

$$\delta_j^A(x,y) = \delta_j^{Ia}(x,y) \cdot \delta_j^{Ie}(x,y) \tag{4.1}$$

where δ_j^{Ia} and δ_j^{Ie} are IaAVS and IeAVS, respectively.



Intra-coupled Interaction: $\delta_j^{Ia}(x,y)$

Inter-coupled Interaction: $\delta_j^{Ie}(x,y)$

UTS:AAi THE ADVANCED ANALYTICS INSTITUTE

Intra-coupled Interaction

DEFINITION 4.2. *Given an information table S, the **Intra-coupled Attribute Value Similarity (IaAVS)** between attribute values x and y of feature a_j is:*

$$\delta_j^{Ia}(x, y) = \frac{|g_j(x)| \cdot |g_j(y)|}{|g_j(x)| + |g_j(y)| + |g_j(x)| \cdot |g_j(y)|}. \qquad (4.2)$$

Rationale:

- The Greater similarity is assigned to the attribute value pair which owns approximately equal frequencies.
- The higher these frequencies are, the closer such two values are.

***IaAVS* has been captured to characterize the value similarity in terms of attribute value occurrence times.**

Inter-coupled Interaction

Modified Value Distance Matrix:

$$D_{j|c}(x,y) = \sum_{g \in L} |P_{c|j}(\{g\}|x) - P_{c|j}(\{g\}|y)|$$

Object Co-occurrence Probability

Inter-coupled Relative Similarity based on Power Set (IRSP), Universal Set (IRSU), Join Set (IRSJ), and Intersection Set (IRSI).

IRSP: $$\delta^P_{j|k}(x,y) = \min_{W \subseteq V_k} \{2 - P_{k|j}(W|x) - P_{k|j}(\overline{W}|y)\}$$

IRSU: $$\delta^U_{j|k}(x,y) = 2 - \sum_{w \in V_k} \max\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}$$

IRSJ: $$\delta^J_{j|k}(x,y) = 2 - \sum_{w \in \varphi_{j \to k}(x) \bigcup \varphi_{j \to k}(y)} \max\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}$$

IRSI: $$\delta^I_{j|k}(x,y) = \sum_{w \in \varphi_{j \to k}(x) \bigcap \varphi_{j \to k}(y)} \min\{P_{k|j}(\{w\}|x), P_{k|j}(\{w\}|y)\}$$

Inter-coupled Interaction

DEFINITION 4.5. *Given an information table S, the **Inter-coupled Attribute Value Similarity (IeAVS)** between attribute values x and y of feature a_j is:*

$$\delta_j^{Ie}(x,y) = \sum_{k=1,k\neq j}^{n} \alpha_k \delta_{j|k}(x,y), \qquad (4.7)$$

where α_k is the weight parameter for feature a_k, $\sum_{k=1}^{n} \alpha_k = 1$, $\alpha_k \in [0,1]$, and $\delta_{j|k}(x,y)$ is one of the inter-coupled relative similarity candidates.

IeAVS **focuses on the object co-occurrence comparisons with four inter-coupled relative similarity options.**

Coupled Object Similarity (COS) between objects:

$$COS(u_{i_1}, u_{i_2}) = \sum_{j=1}^{n} \delta_j^A(x_{i_1 j}, x_{i_2 j}) \text{ where } \delta_j^A(x,y) = \delta_j^{Ia}(x,y) \cdot \delta_j^{Ie}(x,y)$$

Theoretical Analysis

– Computational Accuracy Equivalence:

THEOREM 5.1. *IRSP, IRSU, IRSJ and IRSI are all equivalent to one another.*[2]

Inter-coupled Relative Similarity

IRSP $\Leftrightarrow$ IRSU $\Leftrightarrow$ IRSJ $\Leftrightarrow$ IRSI

– Computational Complexity Comparison:

AD

Metric	Calculation Steps	Flops per Step	Complexity
$IRSP$	$nR(R-1)/2$	$2(n-1)2^R$	$O(n^2R^22^R)$
$IRSU$	$nR(R-1)/2$	$2(n-1)R$	$O(n^2R^2R)$
$IRSJ$	$nR(R-1)/2$	$2(n-1)P$	$O(n^2R^2R)$
$IRSI$	$nR(R-1)/2$	$2(n-1)Q$	$O(n^2R^2R)$

$$2^R > R \geq P \geq Q$$

R: The maximal number of attribute values.

IRSP ≥ IRSU ≥ IRSJ ≥ IRSI

THE ADVANCED ANALYTICS INSTITUTE

Back to Example

Movie	Director	Actor	Genre	Class
Godfather II	Scorsese	De Niro	Crime	G_1
Good Fellas	Coppola	De Niro	Crime	G_1
Vertigo	Hitchcock	Stewart	Thriller	G_2
N by NW	Hitchcock	Grant	Thriller	G_2
Bishop's Wife	Koster	Grant	Comedy	G_2
Harvey	Koster	Stewart	Comedy	G_2

Coupled Nominal Similarity:

Sim(*Scorsese, Coppola*) = Sim(*Coppola, Coppola*) = 0.33

Sim(*Koster, Hitchcock*) = 0.25   Sim(*Koster, Coppola*) = 0

Sim(*Koster, Koster*) = Sim(*Hitchcock, Hitchcock*) = 0.5

Scorsese and *Coppola* are very similar directors

Sim(*Koster*, *Hitchcock*) > Sim(*Koster*, *Coppola*)

Sim (*Scorsese*, *Coppola*) > Sim(*Koster*, *Hitchcock*)

Sim (*Koster*, *Koster*) > Sim(*Scorsese*, *Coppola*)

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Experiment and Evaluation

Several experiments are performed on extensive UCI data sets to show the **effectiveness** and **efficiency.**

- **Coupled Similarity Comparison**

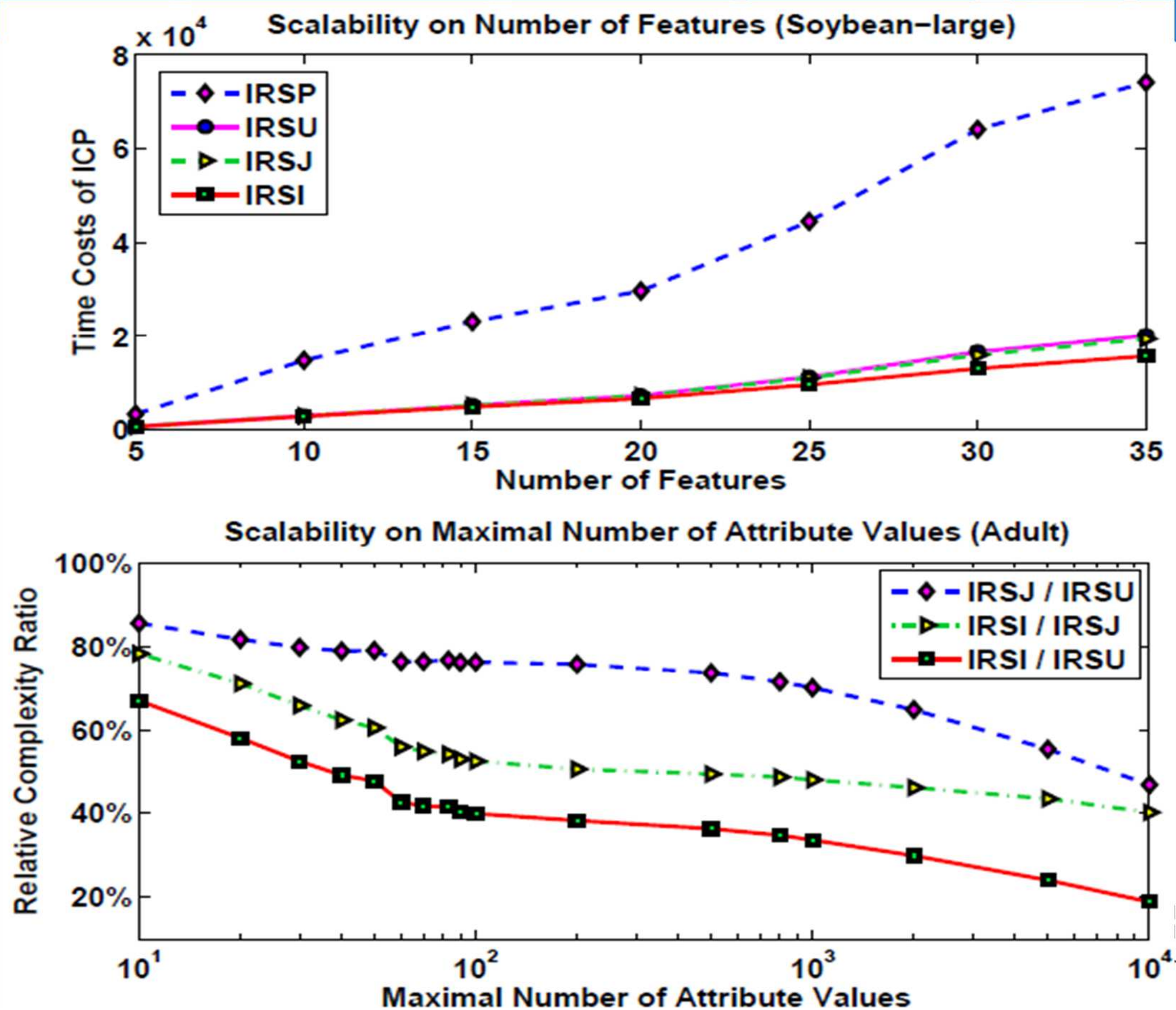
The goal is to show the obvious superiority of *IRSI, compared with the* most time-consuming one *IRSP.*

- ***COS* Application (*COD*)**

Four groups of experiments are conducted on the same data sets by k-modes(*KM*) *with ADD* (existing methods)*, KM with COD,* spectral clustering(*SC*) *with ADD, and SC* with *COD.*

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Coupled Similarity Comparison

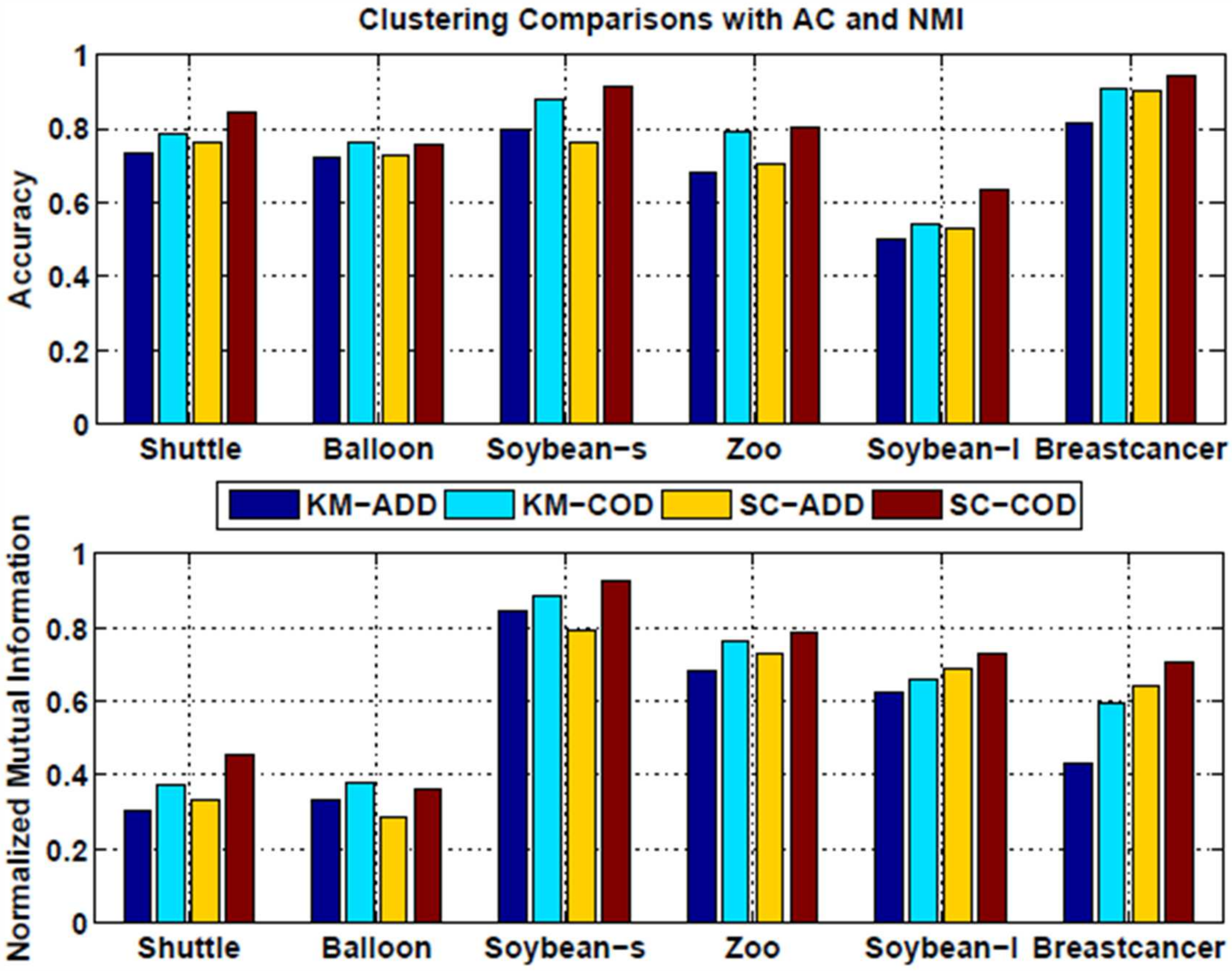


Coupled Similarity Comparison

In summary, all of the above experiment results clearly show that *IRSI outperforms IRSU, IRSJ, and IRSP in terms* of the computational complexity, no matter how small or large, simple or complicated a data set is.

In particular, with the increasing numbers of either features or attribute values, *IRSI demonstrates superior efficiency compared to* the others. *IRSJ and IRSU follow, with IRSP being the* most time-consuming, especially for the large-scale data set.

Application



Experiment and Evaluation

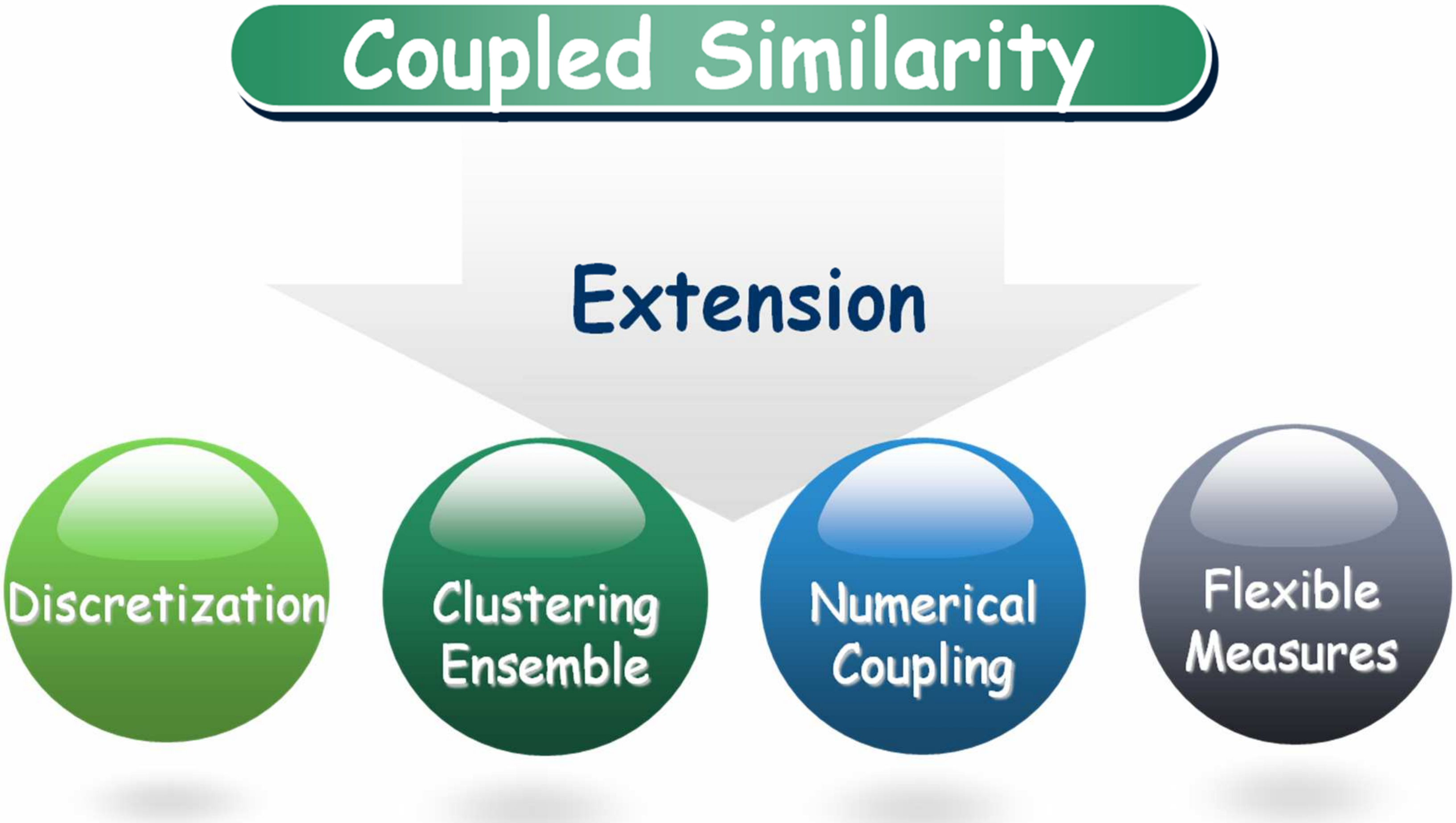
We draw the following two conclusions:

- Intra-coupled relative similarity IRSI is the most efficient one when compared with IRSP, IRSU and IRSJ, especially for large-scale data.

- Our proposed object dissimilarity metric COD is better than others, such as dependency aggregation only ADD, for categorical data in terms of clustering qualities.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

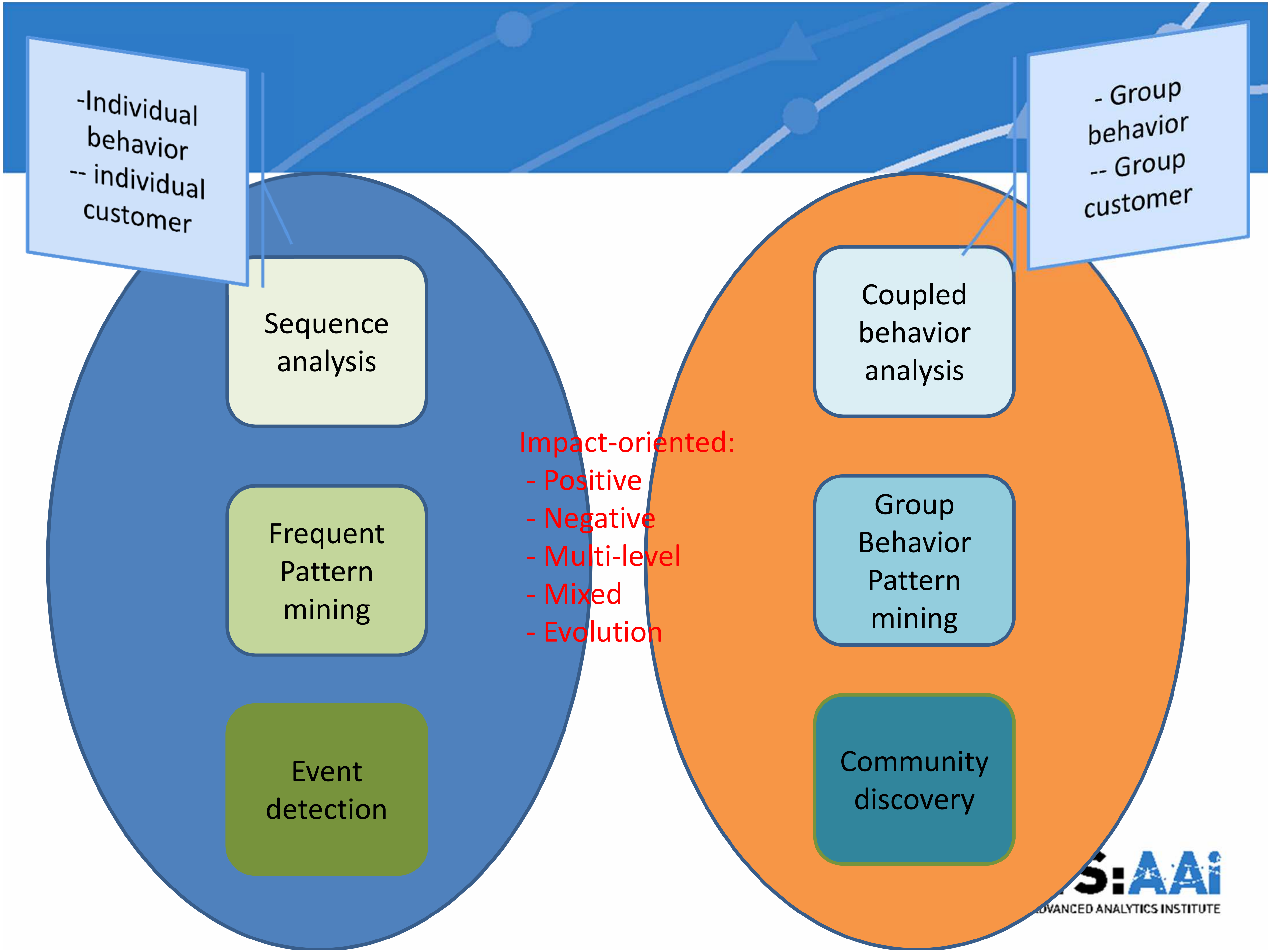
Conclusion


Coupled Similarity
Extension
Discretization
Clustering Ensemble
Numerical Coupling
Flexible Measures

……

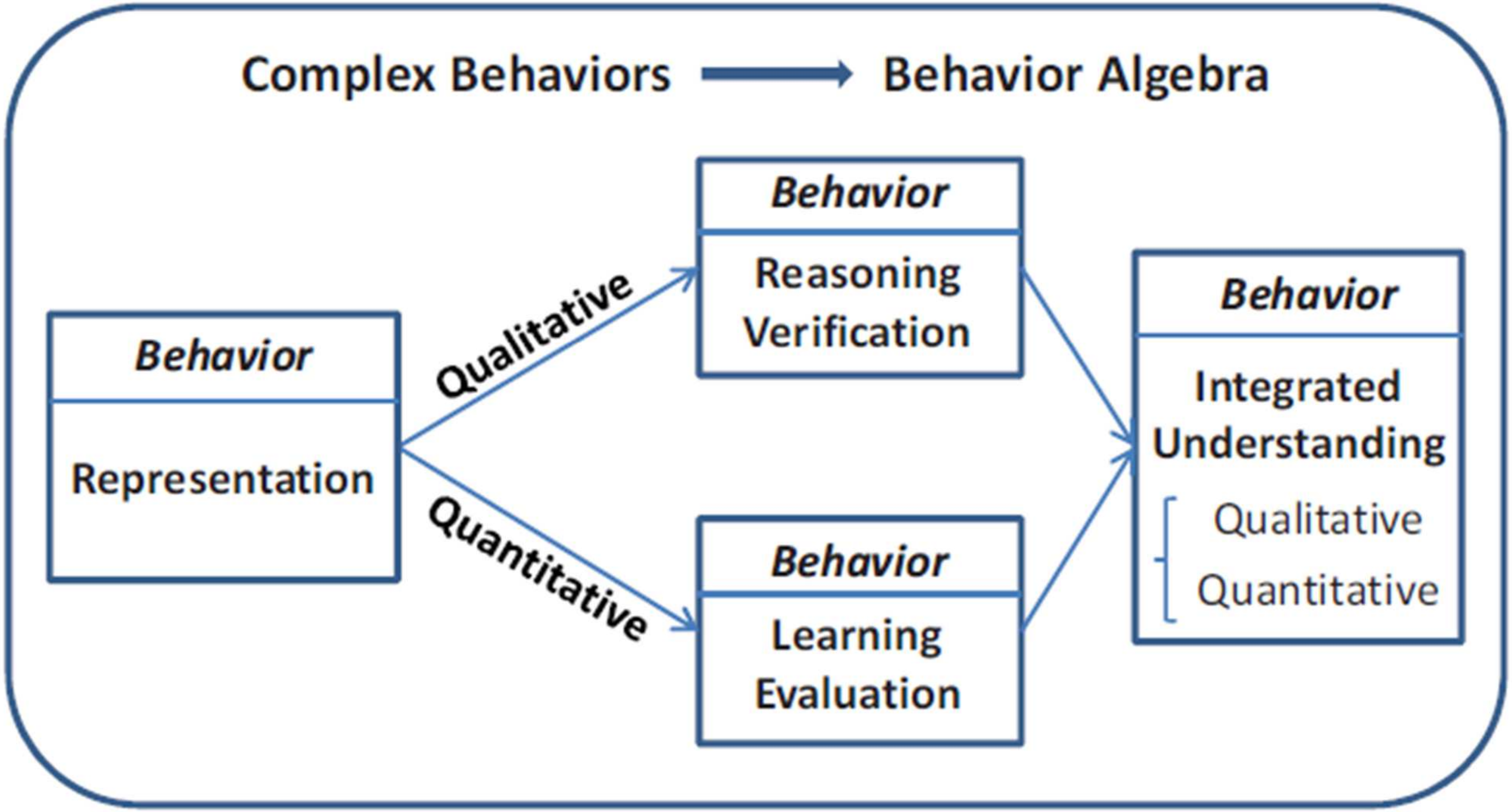
11. Challenges and Prospects

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE


-Individual behavior
-- individual customer
Sequence analysis
Frequent Pattern mining
Event detection
Impact-oriented:
- Positive
- Negative
- Multi-level
- Mixed
- Evolution
- Group behavior
-- Group customer
Coupled behavior analysis
Group Behavior Pattern mining
Community discovery
ADVANCED ANALYTICS INSTITUTE

9. Challenges and Prospects of Complex Behavior Computing

Modeling and Analysis of Complex Behaviors



UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Modeling and Analysis of Complex Behaviors

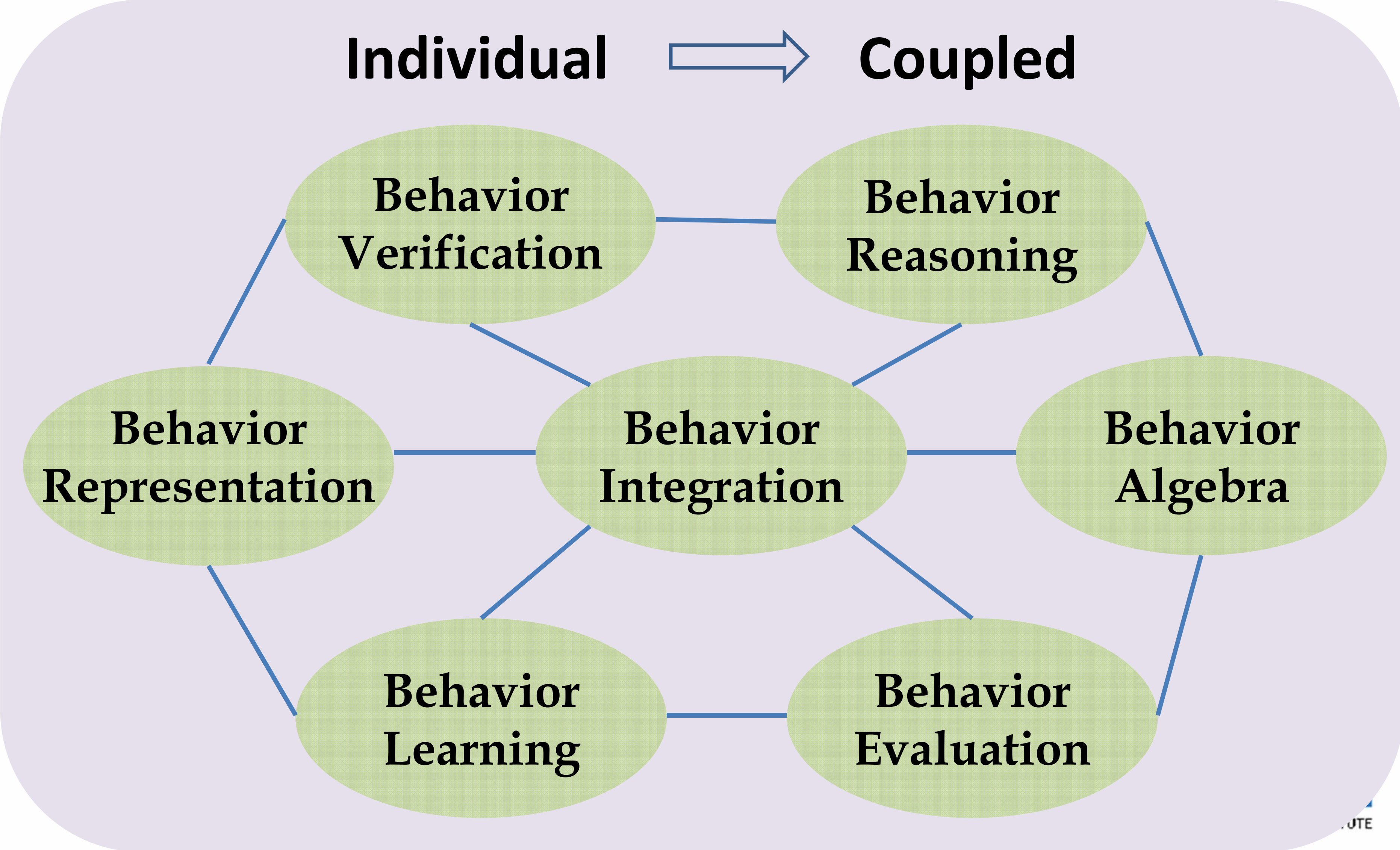
We could develop two directions to explicate complex behaviors: **qualitative and quantitative behavior analytics**

With the formal representation of coupled behaviors, the **qualitative analytics** addresses the task of behavior reasoning and verification, while the **quantitative research** targets behavior learning and evaluation. Finally, an appropriate way could be chosen to integrate these two studies to obtain an **integrated understanding** of the implicit complex behaviors from both qualitative and quantitative aspects.

During this process, many open issues are worth systematic investigation along with case studies from aspects such as ***behavior reasoning, behavior learning, behavior evaluation, behavior integration*** at individual but more on group levels.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Modeling and Analysis of Complex Behaviors



Fundamental

- Formal methods
- Reasoning
- Modelling check
- Quantitative representation and learning

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Individual Behaviour Learning

- Intention learning
- Negative sequence/behaviour analysis
- Complex behaviour/sequence analysis
- Behaviour impact learning
- Behaviour utility learning
- Early prediction of high impact/utility behaviours
- …

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Group-oriented Coupled Learning

- Group intent learning
- Coupled sequence modelling and analysis
- Coupling relationship learning
- Heterogeneous behaviour learning
- Social influence analysis
- Contrast group analysis
- Divergence vs. convergence of group behaviors

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Noniidness learning

$O \backslash A$	A_1	A_2	...	A_J	M_1	...	M_Q
O_1	$\mathcal{V}_{11}$	$\mathcal{V}_{12}$	...	$\mathcal{V}_{1J}$	C_{11}	...	C_{1Q}
O_2	$\mathcal{V}_{21}$	$\mathcal{V}_{22}$	...	$\mathcal{V}_{2J}$	C_{21}	...	C_{2Q}
...	...	...	...	...	...	...	...
O_n	$\mathcal{V}_{n1}$	$\mathcal{V}_{n2}$	...	$\mathcal{V}_{nJ}$	C_{n1}	...	C_{nQ}
...	...	...	...	...	...	...	...
O_N	$\mathcal{V}_{N1}$	$\mathcal{V}_{N2}$	...	$\mathcal{V}_{NJ}$	C_{N1}	...	C_{NQ}

FIGURE 3. Information table and couplings for noniidness learning.

References

- 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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

- **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**, Huaifeng 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, Huaifeng Zhang, Chengqi Zhang, **Longbing Cao**, Hans Bohlscheid. Debt Detection in Social Security by Adaptive Sequence Classification, KSEM 2009: 192-203.

  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.

  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.

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

-

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.

IEEE Task Force

- IEEE Task Force on Behavior and Social Informatics and Computing (BSIC)
- www.behaviorinformatics.org

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Call for Papers

Behavior and Social Informatics Workshops:

- PAKDD - BSI 2013 Australia

http://datamining.it.uts.edu.au/bsi/bsi2013/

- IJCAI – BSIC 2013 China

http://datamining.it.uts.edu.au/bsi/bsic2013/

- IJCAI 2013 Tutorial
  - Behavior Informatics
- Special Issue with World Wide Web Journal

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE

Your feedback is appreciated.

- Longbing Cao
  longbing.cao@uts.edu.au

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

  www.behaviorinformatics.org

UTS:AAi
THE ADVANCED ANALYTICS INSTITUTE