Mining Precise-positioning Episode Rules from Event Sequences

Xiang Ao[1], Ping Luo[1], Jin Wang[2], Fuzhen Zhuang[1], Qing He[1]

Institute of Computing Technology, CAS, China[1] University of California at Los Angeles, USA[2]

MOTIVATION

Traditional Episode Rule

Given a frequent episode α, a **traditional episode rule** in the form of ***lhs* → *rhs*** is generated straightforwardly: The antecedent ***lhs*** is the prefix of α and the consequent ***rhs*** is the last event in α, if its confidence is larger than a user-specified threshold.

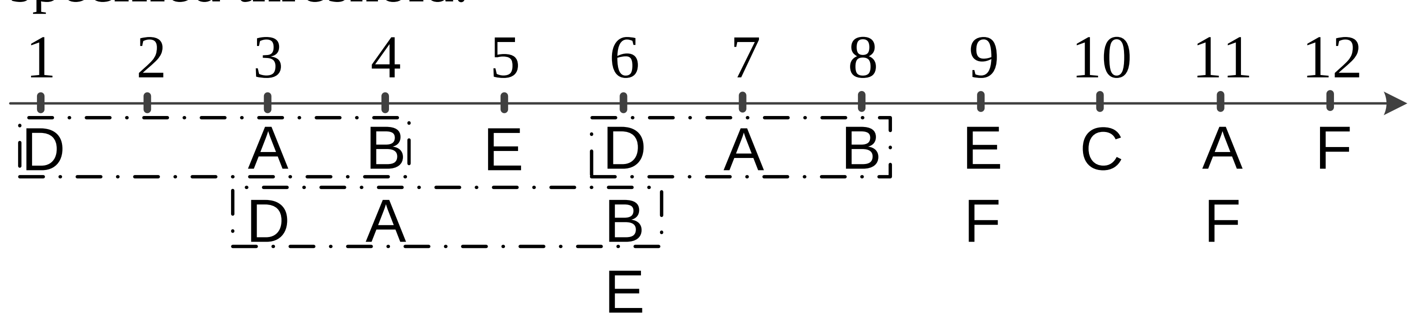


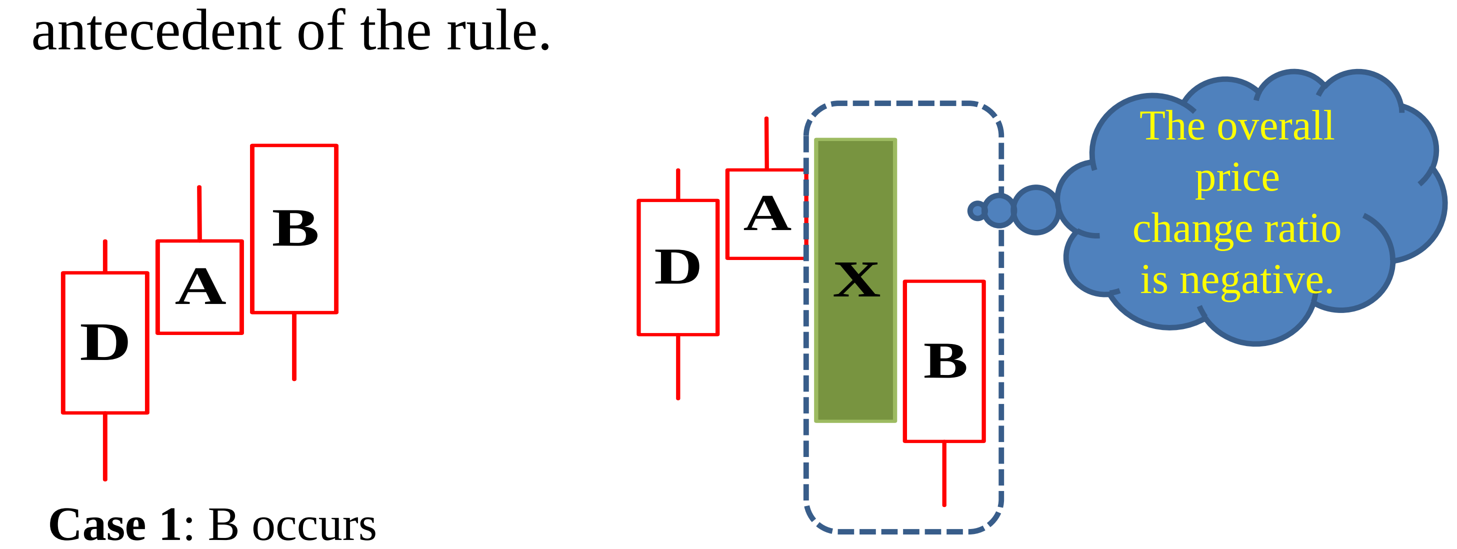
Fig.1 The running example event sequence.

From Fig.1, **<D, A> → <B> is a traditional episode rule** which indicates it is **within 2 time intervals** after the occurrence of <D, A> that B will occur (with 100% confidence).

Limitation of Traditional Episode Rule

Example: In stock investment application, we can map price change ratios to events and use candlestick charts to represent events. Red bars denote price increase of a stock, and green bars denote prices decrease.

- The episode rule <D, A> → <B> **predicts correct** in the following two cases, however **we will lose money in Case 2** if we long the stock after we observed the antecedent of the rule.



Case 1: B occurs right behind <D, A>.

Case 2: B occurs after <D, A> appears within two days but right behind a significant decrease.

Precise-positioning Episode Rule (PER)

We define **precise-positioning episode rule** in the form of:

$$\Gamma = \alpha \xrightarrow{\Delta t} \beta$$

α: a **traditional episode**, as the antecedent;
β: a **fixed-gap episode**, as the consequent;
Δ*t*: **the time constraint** between the antecedent and the consequent.

Fixed-gap episode:

$$\beta = (\langle e_{\beta_1}, \cdots, e_{\beta_k} \rangle, \langle \Delta t_1, \cdots, \Delta t_{k-1} \rangle)$$

time constraints between two consecutive events

- The traditional episode rule <D, A> → <B> in Fig.1 becomes two PERs: **<D, A> $\xrightarrow{1}$ <B>** and **<D, A> $\xrightarrow{2}$ <B>**.

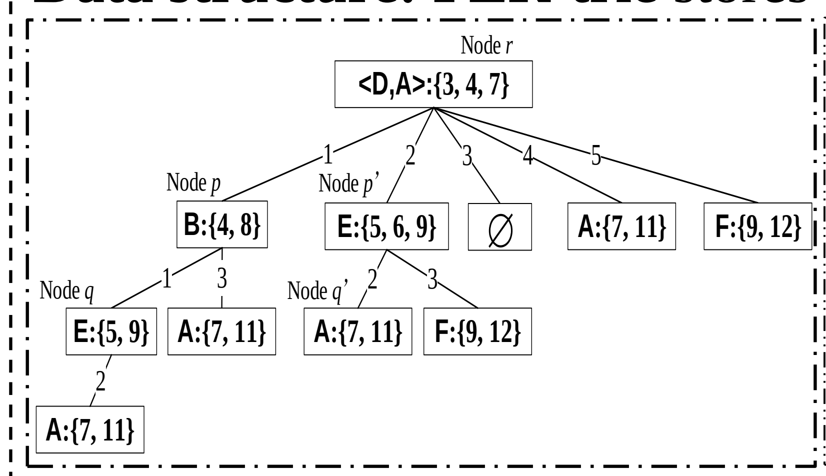
Mining ALGORITHM & EFFICIENCY

1. MIP-ENUM Algorithm

The basic idea of MIP-ENUM is to enumerate PER candidates by concatenating discovered traditional episode with fixed-gap episode and subsequently filter the invalid ones according their confidence values.

2. MIP-TRIE Algorithm

Data structure: PER-trie stores valid PER compactly.



Root node: a traditional episode with end time of these minimal occurrence;
Non-root node: event with its occurrence time after a distance to its parent.
Edge: time distance between parent and child.

Algorithm: MIP-TRIE(DFS) and MIP-TRIE(PRU).

We use PER-trie to store all valid PER given an antecedent α and propose two algorithms to build complete PER-trie.

- **MIP-TRIE(DFS)** expands the PER-trie by a recursively depth first search manner.
- **MIP-TRIE(PRU)** adopts an improved traverse strategy with pruning technique.

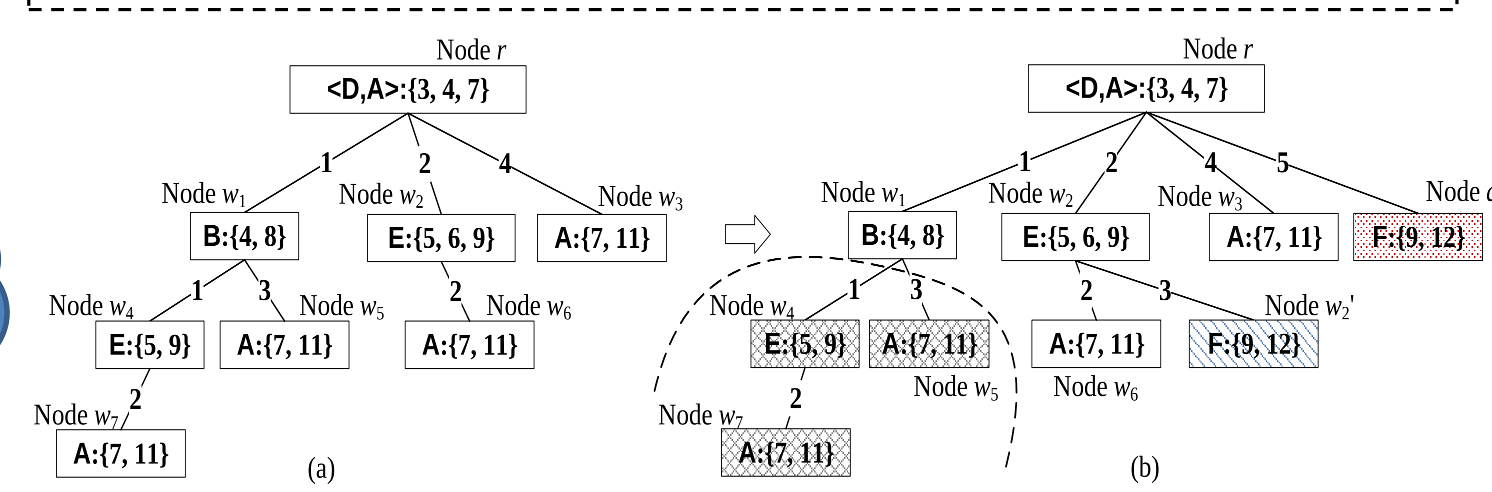
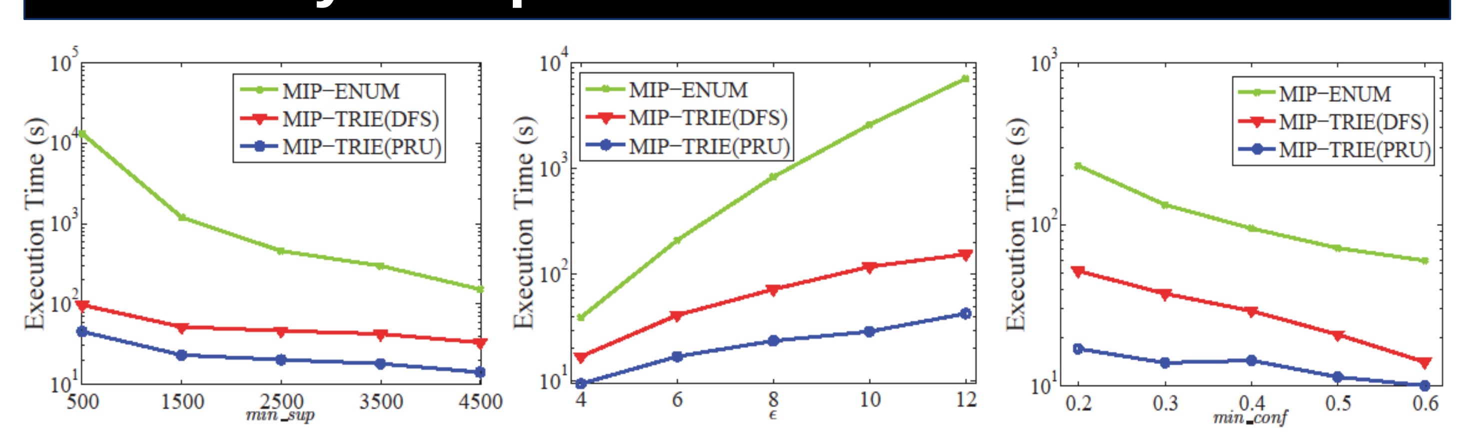


Fig.2 The expansion process of MIP-TRIE(PRU).

- *q*' is expanded first as child or *r*, and we traverse w_1 - w_3 and pruning w_4, w_5 and w_7 and finally traverse w_6 with *q*'.

Efficiency Comparisons



Dataset: Retail -- http://fimi.cs.helsinki.fi/data/

Observations: 1. MIP-TRIE(PRU) outperforms MIP-TRIE(DFS) and MIP-ENUM algorithm; 2. MIP-TRIE algorithms significantly outperform MIP-ENUM.

EFFECTIVENESS of PER

DATASET: 150 related industry sector pairs of China stock market from Jan. 1, 2010 to Aug. 29, 2014.

EVT SEQ. CONSTRUCTION: **UP** (if the price increases) and **DN** (otherwise) for each industry sector.

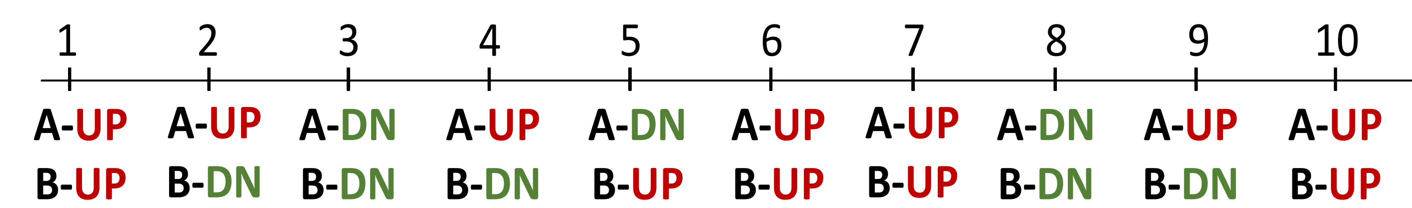


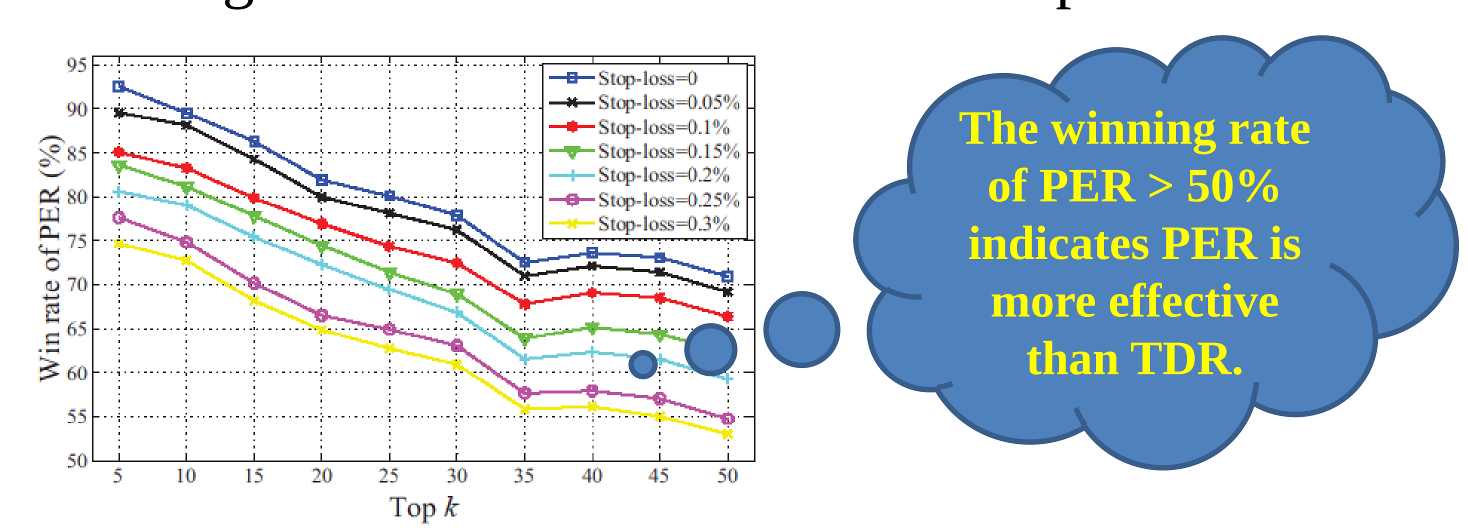
Fig.3 The example stock industry sector event sequence.

- A and B denote stock industry sectors.

SETTINGS: We use first 4-year sequence as the training set to mine PER on each sequence and degrade PER whose Δ*t* = 5 to traditional episode rule (denoted as TDR), then test prediction ability of them on the rest.

COMPARISON: For PER, we trade strictly according to the rule; for TDR, we trade after antecedent occurs and close out either consequent appears or the maximal occurrence window for consequent reaches.

MEASURE: We close out when the float loss exceeds a stop-loss threshold during the holdings by TDR. We compute the return of holdings and visualize the winning rate of PER under different stop-loss threshold.



VENUE & CONTACT INFORMATION

The 33rd IEEE International Conference on Data Engineering, San Diego, California, USA, April 19-22, 2017.

Email: {aoxiang, luop}@ict.ac.cn, jinwang@cs.ucla.edu, {zhuangfz, heq}@ics.ict.ac.cn

Homepage of MLDM Group, ICT, CAS: http://mldm.ict.ac.cn

Xiang's personal homepage: http://mldm.ict.ac.cn/MLDM/~aox

Public account on Wechat: