Mining Temporal Association Patterns under a Similarity Constraint

Jin Soung Yoo

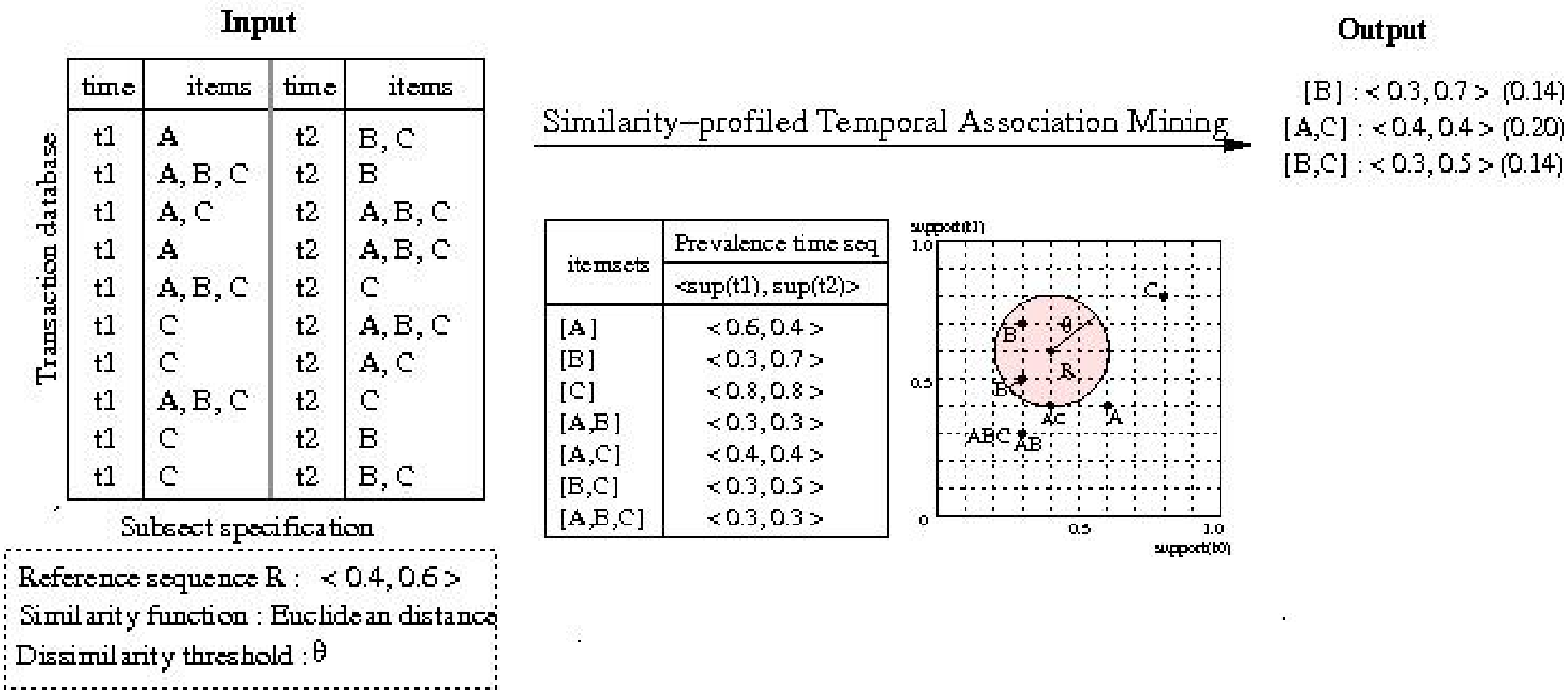
J. S. Yoo[1] and S. Shekhar[2]

[1]Indiana University-Purdue University

[2]University of Minnesota

Similarity-profiled Temporal Association

- A subset of items whose prevalence variation over time is similar to a reference sequence



Input

Transaction database

time	items	time	items
t1	A	t2	B, C
t1	A, B, C	t2	B
t1	A, C	t2	A, B, C
t1	A	t2	A, B, C
t1	A, B, C	t2	C
t1	C	t2	A, B, C
t1	C	t2	A, C
t1	A, B, C	t2	C
t1	C	t2	B
t1	C	t2	B, C

Subsect specification

Reference sequence R : < 0.4, 0.6 >
Similarity function : Euclidean distance
Dissimilarity threshold : θ

Similarity–profiled Temporal Association Mining

itemsets	Prevalence time seq
	<sup(t1), sup(t2)>
[A]	< 0.6, 0.4 >
[B]	< 0.3, 0.7 >
[C]	< 0.8, 0.8 >
[A,B]	< 0.3, 0.3 >
[A,C]	< 0.4, 0.4 >
[B,C]	< 0.3, 0.5 >
[A,B,C]	< 0.3, 0.3 >

Output

[B] : < 0.3, 0.7 > (0.14)
[A,C] : < 0.4, 0.4 > (0.20)
[B,C] : < 0.3, 0.5 > (0.14)

Motivation Examples

- Weather-to-Sales
  - Correlation between daily temperatures and merchandise sales – Walt Disney World [NOAAEconomics]
  - Popular sale items during hurricane in a region – Wal-Mart [FORTUNE Magazine]
    - Flashlights, generators and tarps with bottled water
    - Strawberry Pop-Tarts with bottled water
- Weather-to-Web Sites
  - Web sites depending on weather – [Weather.com]

Motivation Examples

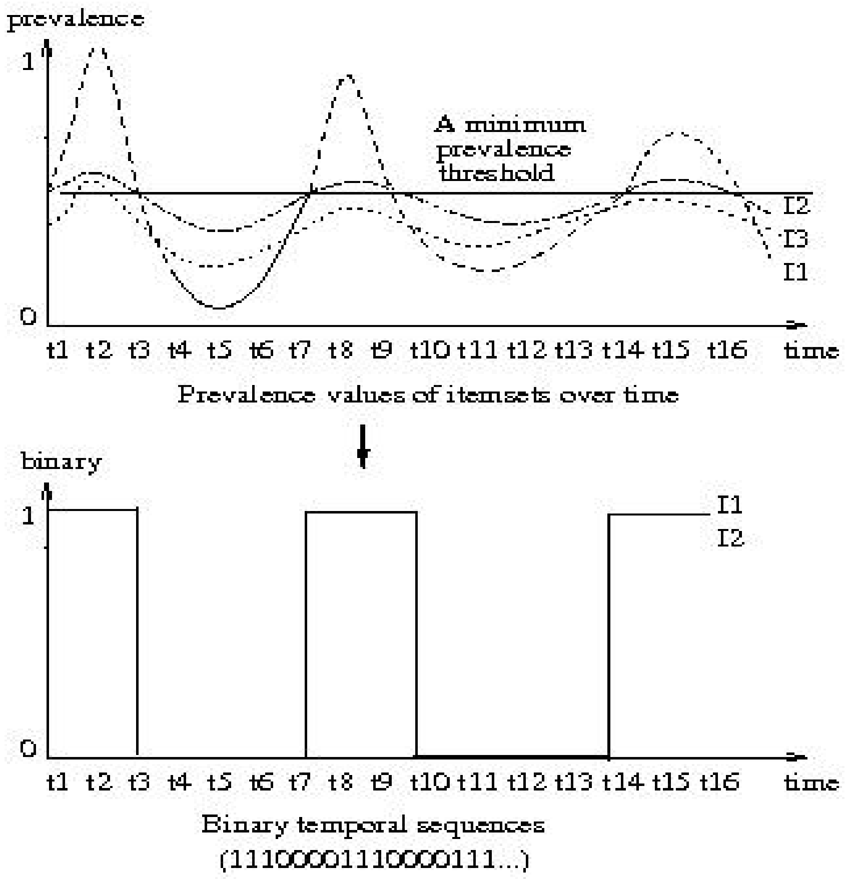
- Scientific Phenomena-to-Climates
  - Climate events correlated with El Nino
    - Low precipitation and low atmospheric carbon dioxide in Australia
- Scientific Phenomena-to-Agriculture
  - Agricultural products under the effect of El Nino
    - Wheat and other products in Australia

Related Work

- Cyclic associations [Ozden'98, Ramaswamy'94]
  - Periodically repetitive patterns for frequent itemsets
  - "Beer" and "chips" are sold together primarily between 6PM and 9PM
- Calendar based associations [Li'03]
  - Frequent itemsets on "15th day of a March", (*,3,15)
- User-defined temporal regulation patterns [Li'06, Bettini'98]
  - Frequent events "within 2 days" after "a rise of IBM stock"

Comparison with Related Work

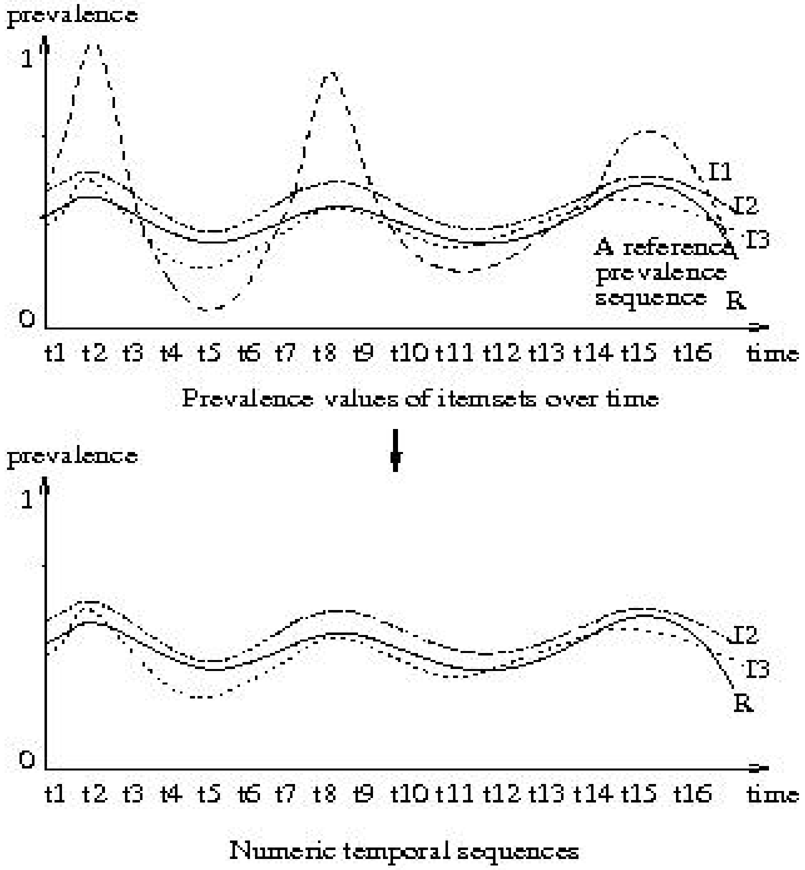
- Regulation patterns



Prevalence values of itemsets over time

Binary temporal sequences (11100001110000111...)

- Similarity patterns



Prevalence values of itemsets over time

Numeric temporal sequences

Contributions

- Formulate similarity-profiled association patterns
    - User-defined temporal similarity patterns using a subset specification (i.e., a reference sequence, a similarity function, and a dissimilarity threshold)
- Explore interesting properties for efficiently mining similarity-profiled associations
- Develop the mining algorithm.
- Experimentally evaluate it with synthetic and real data sets.

Problem Definition

- **Given**
  - A timestamped transaction database $\mathbf{D=D_1} \cup \ldots \cup \mathbf{D_n}$
    - $\mathbf{D_i}$ is a set of transactions included in time slot **i**
    - Each transaction $\mathbf{d} \in \mathbf{D}$ is a tuple < *timestamp*, *items*>
  - A subset specification
    - A reference time sequence $\mathbf{R=<r_1 , \ldots , r_n>}$
    - A similarity function $\mathbf{F_{similarity}(S_I , R)}$, where $\mathbf{S_I}$ is a support time sequence of itemset **I**
    - A dissimilarity threshold θ
- **Find:** A set of itemsets which satisfy the given subset specification, i.e., $\mathbf{F_{similarity}(S_I , R)} \leq \theta$
- **Objective:** A complete and correct result set while reducing the computation cost.

Background: Interest Measure

- Support
  - The support of itemset **I** in transaction dataset **D** is

    $support\,(I, \mathbf{D}) = |\{\mathbf{d} \in \mathbf{D}, I \subseteq \mathbf{d}\}| / |\{\mathbf{D}\}|$

Transaction database

tno	items
1	A
2	A, B, C
3	A, C
4	A
5	A, B, C
6	C
7	C
8	A, B, C
9	C
10	C

e.g., *support* ({A}, **D**)=6/10=0.6

Composite Interest Measure

- The support time sequence of itemset **I** in $\mathbf{D}=\mathbf{D}_1 \cup \ldots \cup \mathbf{D}_n$
  - $\mathbf{S}_\mathbf{I}$ = <*support* (*I*, $\mathbf{D}_1$), …., *support* (*I*, $\mathbf{D}_\mathbf{n}$)>
- Dissimilarity distance between a support sequence $\mathbf{S}_\mathbf{I}$ and a reference sequence **R**
  - $\mathbf{L}_\mathbf{p}$ *norm (p=1,2, …, ∞) based distance, e.g.,*
    - $\mathbf{L}_\mathbf{2}$ *norm (Euclidean distance)*
      - $D(\mathbf{R}, \mathbf{S}_\mathbf{I}) = (\Sigma_{t=1..n} |r_t - s_t|^2)$🗁
    - *Normalized L2 norm*
      - $D(\mathbf{R}, \mathbf{S}_\mathbf{I}) = ((\Sigma_{t=1..n} |r_t - s_t|^2) / \mathbf{n})$🗁

Outline

- Introduction
- Problem Definition
- Related Work

☞ Algorithmic Design Concept

- Algorithm
- Experimental Results
- Conclusion

Computational Challenge

- Naïve Approach
  - Two separate phrases
    - Compute the support values of all possible itemsets at each time point, and generate their prevalence sequences
    - Compare the support sequences with a reference sequence, and find similar itemsets.
  - Computationally expensive
    - Exponential number of itemsets with number of item types, 2^|*n* -1|

Our Questions

- Can we reduce the search space for only interesting patterns?
- How can we estimate the similarity distance of an itemset without the generation of the support sequence?

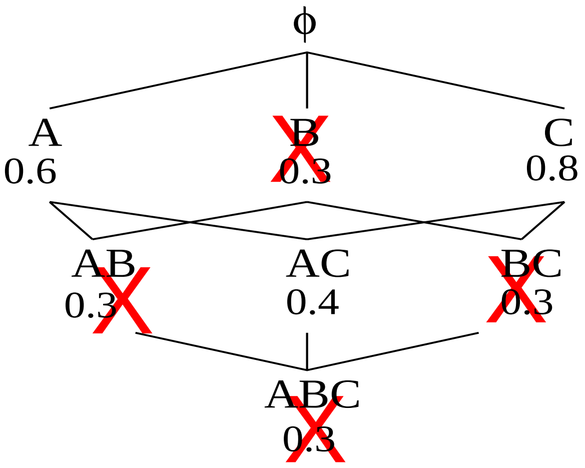
Background: Frequent Itemset Pruning

- Using the monotonicity of support
  - Support is monotonically non-increasing with the size of itemset,

    i.e., $J \subseteq I$, *then support* $(J, \mathbf{D}) \geq$ *support* $(I, \mathbf{D})$

Transaction database

tno	items
1	A
2	A, B, C
3	A, C
4	A
5	A, B, C
6	C
7	C
8	A, B, C
9	C
10	C



Support value

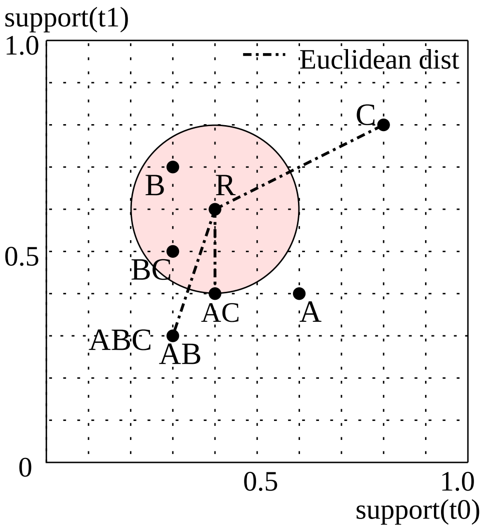
Frequent threshold: 0.3

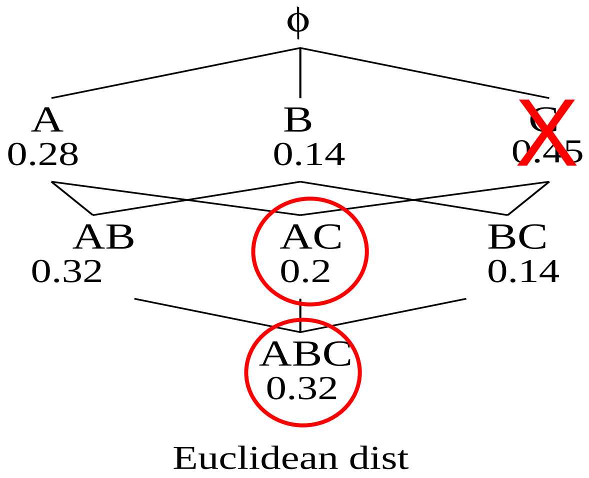
Which one is greater than 0.3?

Observation

- It is not easy to reduce our search space.
  - **L_p** *norm based distance does not show any monotonic.*

    e.g., $\boldsymbol{D}(\mathbf{R}, \mathbf{S}_{\{ABC\}}) > \boldsymbol{D}(\mathbf{R}, \mathbf{S}_{\{AC\}})$ but $\boldsymbol{D}(\mathbf{R}, \mathbf{S}_{\{AC\}}) < \boldsymbol{D}(\mathbf{R}, \mathbf{S}_{\{C\}})$

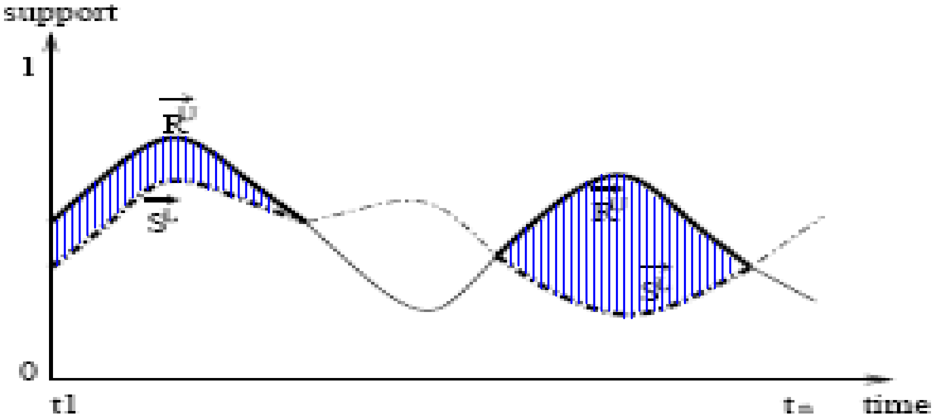




Let dissimilarity threshold: 0.3
Can we prune a super set of C ??

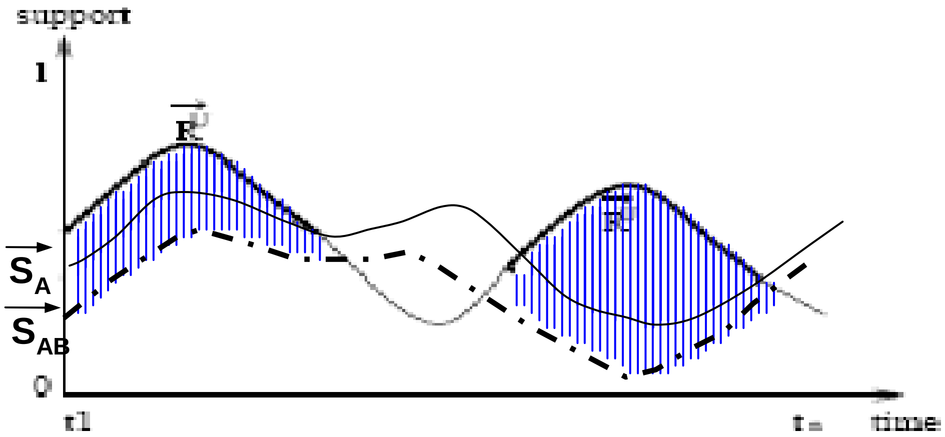
Our Approach: Upper Lower Bounding Distance

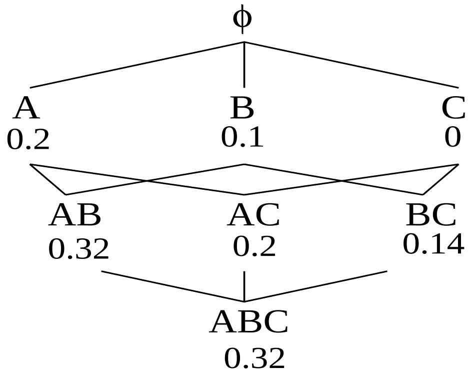
- Let
  - $\mathbf{R^U=<r_1 , \ldots , r_k>}$ be a subsequence of a reference sequence **R and** $\mathbf{S^L=<s_1 , \ldots , s_k>}$ be a subsequence of a support sequence **S,** where $\mathbf{r_i} > \mathbf{s_i}$
- The upper lower-bounding distance between **R** and **S**, $\boldsymbol{D_{Ulb}}(\mathbf{R}, \mathbf{S_l})$, is $\boldsymbol{D}(\mathbf{R^U}, \mathbf{S^L})$.



Monotonicity of Upper LBD

- The upper lower-bounding distance is monotonically non-decreasing with the size of the itemset.
- Proof: The support values of an itemset are monotonically non-increasing with the size of itemset at each time slot.





Upper lower bounding dist

E.g., $D_{Ulb}(R, S_A) \le D_{Ulb}(R, S_{AB})$,

$D_{Ulb}(R, S_B) \le D_{Ulb}(R, S_{AB})$,

Prune by Upper LBD

- Let Itemsets $J \subseteq I$
- If $\boldsymbol{D_{Ulb}}(\mathbf{R}, \mathbf{S_J}) > \theta$, always $\boldsymbol{D_{Ulb}}(\mathbf{R}, \mathbf{S_I}) > \theta$
- Prune all superset of J

Our Questions

- Can we reduce the search space for only interesting patterns?

☞ How can we estimate the similarity distance of an itemset without the generation of the support sequence?

Upper Bound of Support Sequence

- Let
  - **$D=D_1$** ∪ … ∪ **D_n** be a set of disjoint transactions.
  - **J**={**J_1** , …, **J_k**} be a set of all size *k*-1 subsets of a size *k* itemset **I**.
- Upper bound support time sequence of itemset **I**, **U_1** =< **u_1**, … , **u_n** > is defined as
  - **u_1** = *min* {*support* (**J_1**, **D_1**), …., *support* (**J_k** , **D_1**)}
  - **u_n** = *min* {*support* (**J_1**, **D_n**), …., *support* (**J_k** , **D_n**)}

itemsets	Prevalence time seq
	<sup(t1), sup(t2)>
{A}	< 0.6, 0.4 >
{B}	< 0.3, 0.7 >
{C}	< 0.8, 0.8 >
{A,B}	< 0.3, 0.3 >
{A,C}	< 0.4, 0.4 >
{B,C}	< 0.3, 0.5 >
{A,B,C}	?

- E.g., **U_{ABC}** =< **u_1**, **u_2** >=< 0.3, 0.3>

u_1 = *min*{*supp*(**AB**, **D_1**), *supp*(**AC**, **D_1**), *supp*(**BC**, **D_1**)}

u_2 = *min*{*supp*(**AB**, **D_2**), *supp*(**AC**, **D_2**), *supp*(**BC**, **D_2**)}

Upper Bound of Support Sequence

- Let
  - $\mathbf{D}=\mathbf{D}_1 \cup \ldots \cup \mathbf{D}_n$ be a set of disjoint transactions.
  - $\mathbf{J}=\{\mathbf{J}_1, \ldots, \mathbf{J}_k\}$ be a set of all size k-1 subsets of a size k itemset $\mathbf{I}$.
- Lower bound support time sequence of itemset $\mathbf{I}$, $\mathbf{L}_1 =< \mathbf{l}_1, \ldots , \mathbf{l}_n >$ is defined as
  - $\mathbf{l}_1 = max\ \{(support\ (\mathbf{J}_1, \mathbf{D}_1)+support\ (\mathbf{I}\text{-}\mathbf{J}_1, \mathbf{D}_1)-1), \ldots, (support\ (\mathbf{J}_k, \mathbf{D}_1)+support\ (\mathbf{I}\text{-}\mathbf{J}_k, \mathbf{D}_1)-1),0\ \}$
  - $\mathbf{l}_n = max\ \{(support\ (\mathbf{J}_1, \mathbf{D}_n)+support\ (\mathbf{I}\text{-}\mathbf{J}_1, \mathbf{D}_n)-1), \ldots, (support\ (\mathbf{J}_k, \mathbf{D}_n)+support\ (\mathbf{I}\text{-}\mathbf{J}_k, \mathbf{D}_n)-1),0\ \}$

itemsets	Prevalence time seq
	<sup(t1), sup(t2)>
{A}	< 0.6, 0.4 >
{B}	< 0.3, 0.7 >
{C}	< 0.8, 0.8 >
{A,B}	< 0.3, 0.3 >
{A,C}	< 0.4, 0.4 >
{B,C}	< 0.3, 0.5 >
{A,B,C}	?

- E.g., $\mathbf{L}_{ABC} =< \mathbf{l}_1, \mathbf{l}_2 >=< 0.1, 0.1>$

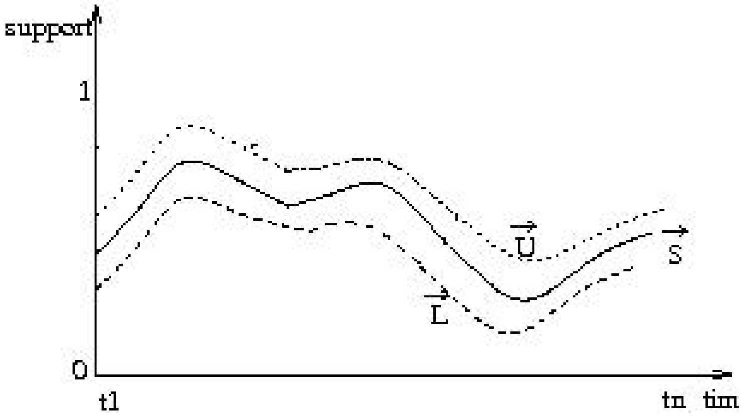
$\mathbf{u}_1 = max\{(supp(\mathbf{AB}, \mathbf{D}_1) +supp(\mathbf{C}, \mathbf{D}_1)-1),$

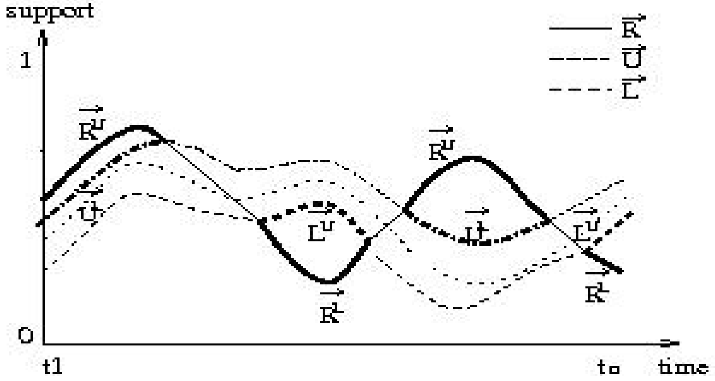
$(supp(\mathbf{AC}, \mathbf{D}_1) +supp(\mathbf{B}, \mathbf{D}_1)-1),$

$(supp(\mathbf{BC}, \mathbf{D}_1) +supp(\mathbf{A}, \mathbf{D}_1)-1),$

$0\}$

Subsequences for Lower Bounding Distance

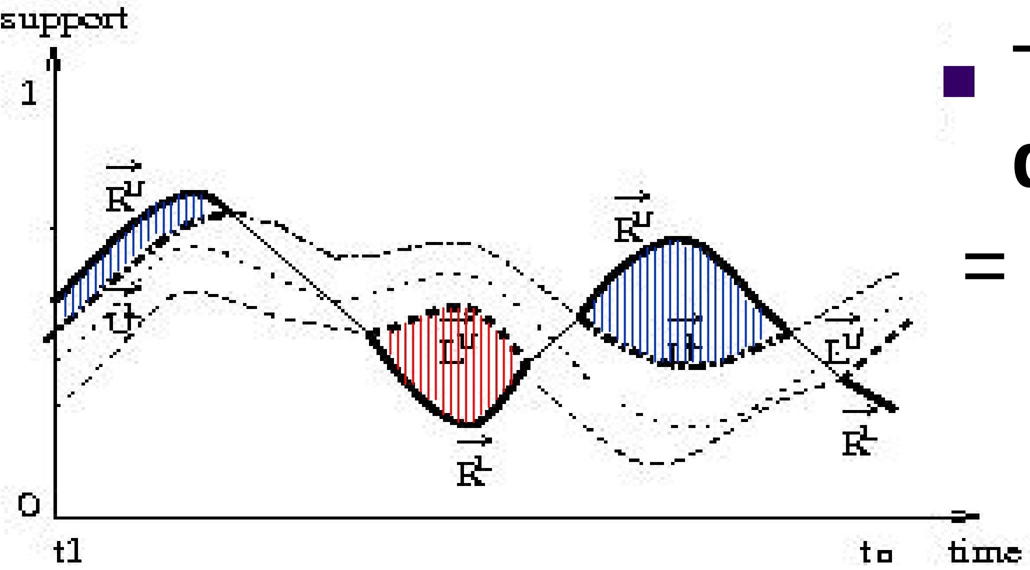




- $\mathbf{R^U=<r_1 , \ldots , r_k>}$ be a subsequence of **R and** $\mathbf{U^L=<u_1 , \ldots , u_k>}$ be a subsequence of **U,** where $\mathbf{r_i} > \mathbf{u_i}$
- $\mathbf{R^L=<r_1 , \ldots , r_k>}$ be a subsequence of **R and** $\mathbf{L^U=<l_1 , \ldots , l_k>}$ be a subsequence of **L,** where $\mathbf{r_i} > \mathbf{l_i}$

Lower Bounding Distance

- The upper lower-bounding distance between **R** and **U**, $D_{Ulb}(\mathbf{R}, \mathbf{U})$ is defined to $D(\mathbf{R}^U, \mathbf{U}^L)$.
- The lower lower-bounding distance between **R** and **L**, $D_{Llb}(\mathbf{R}, \mathbf{L})$ defined to $D(\mathbf{R}^L, \mathbf{L}^U)$.



- The **lower-bounding distance**, $D_{lb}(\mathbf{R}, U, \mathbf{L})$

$$= D_{Ulb}(\mathbf{R}, \mathbf{U}) + D_{Llb}(\mathbf{R}, \mathbf{L})$$

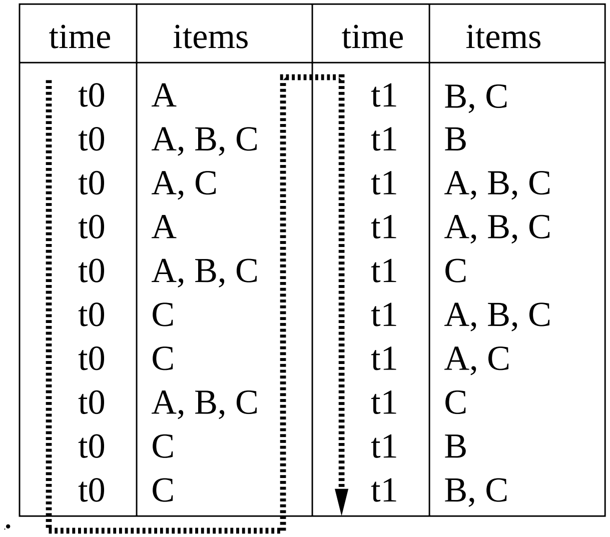
Prune by Lower Bounding Distance

- The lower bounding distance $\boldsymbol{D_{lb}}(\mathbf{R}, U_{\mathbf{I}}, \mathbf{L}_{\mathbf{I}})$ is always not greater than true distance $\boldsymbol{D}(\mathbf{R}, S_{\mathbf{I}})$.
- So, If $\boldsymbol{D_{lb}}(\mathbf{R}, U_{\mathbf{I}}, \mathbf{L}_{\mathbf{I}}) > \theta$, $\boldsymbol{D}(\mathbf{R}, S_{\mathbf{I}}) > \theta$
- Prune itemset **I**

Database Scan Strategy

- Lattice-dominant scan

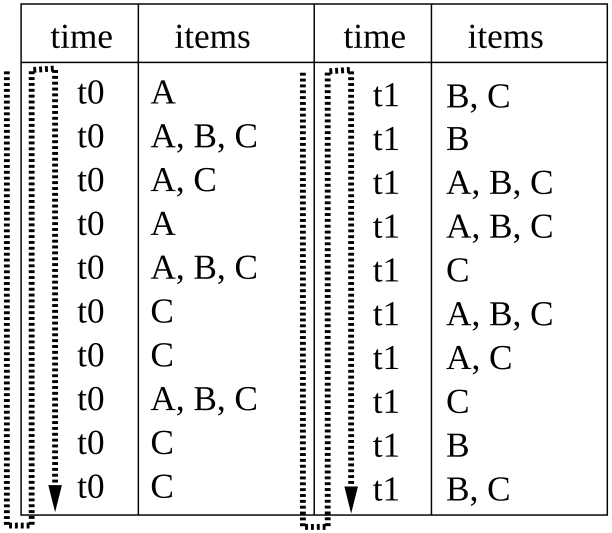
Time−stamped transaction database



time	items	time	items
t0	A	t1	B, C
t0	A, B, C	t1	B
t0	A, C	t1	A, B, C
t0	A	t1	A, B, C
t0	A, B, C	t1	C
t0	C	t1	A, B, C
t0	C	t1	A, C
t0	A, B, C	t1	C
t0	C	t1	B
t0	C	t1	B, C

- Snapshot-dominant scan

Time−stamped transaction database



time	items	time	items
t0	A	t1	B, C
t0	A, B, C	t1	B
t0	A, C	t1	A, B, C
t0	A	t1	A, B, C
t0	A, B, C	t1	C
t0	C	t1	A, B, C
t0	C	t1	A, C
t0	A, B, C	t1	C
t0	C	t1	B
t0	C	t1	B, C

Outline

- Introduction
- Problem Definition
- Related Work
- Algorithmic Design Concept

☞ Algorithm

- Experimental Results
- Conclusion

Similarity-Profiled temporal Association MINing methods

- Two algorithms by different database scan methods
  - L-SPAMINE: Lattice-dominant SPAMINE
  - S-SPAMINE: Snapshot-dominant SPAMINE

Algorithm (L-SPAMINE)

- **Input**
  - A time-stamped dataset
  - A reference sequence, A similarity function, and A threshold
- **Procedure**
  - Generate size K candidate itemsets
    - Prune if **any subset's** $\boldsymbol{D_{Ulb}}$< threshold
  - Estimate **upper and lower bound sequences** of candidates
  - Filter candidates using $\boldsymbol{D_{lb}}$ (=D_{Ulb} + D_{Llb})
  - Scan database and generate true support sequences
  - Find similar itemsets having $\boldsymbol{D}$ < threshold
  - Keep size K itemsets having $\boldsymbol{D_{Ulb}}$ < threshold
  - K=K+1

L-SPAMINE Trace

Transaction database

time	items	time	items
t1	A	t2	B, C
t1	A, B, C	t2	B
t1	A, C	t2	A, B, C
t1	A	t2	A, B, C
t1	A, B, C	t2	C
t1	C	t2	A, B, C
t1	C	t2	A, C
t1	A, B, C	t2	C
t1	C	t2	B
t1	C	t2	B, C

Dissimilarity threshold : 0.2

Similarity Function: Euclidean

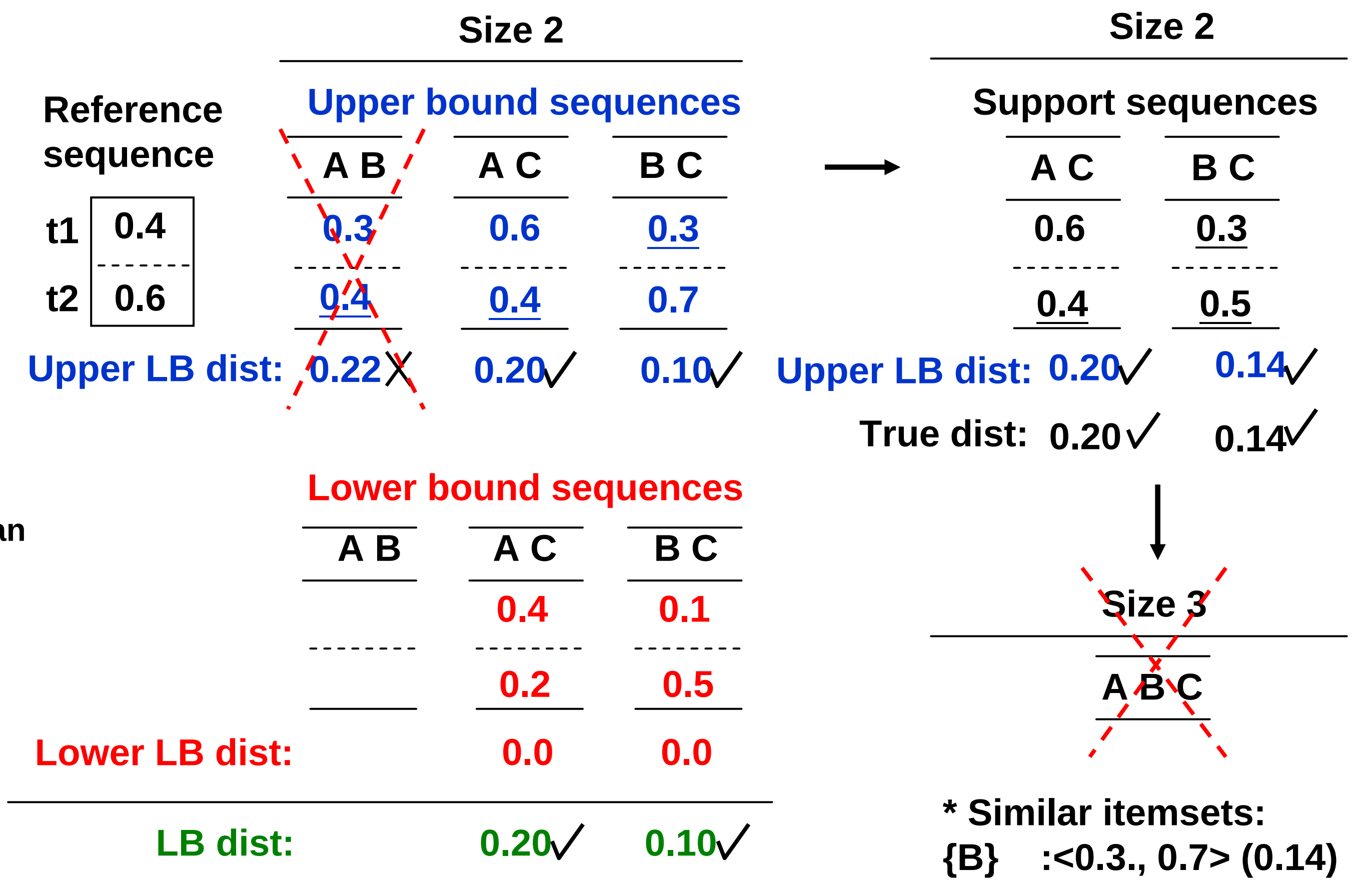
		Size 1		
	Reference sequence	Support sequences		
		A	B	C
t1	0.4	0.6	0.3	0.8
t2	0.6	0.4	0.7	0.8
Upper LB dist:		0.20 ✓	0.10 ✓	0 ✓
True dist:		0.28 ✗	0.14 ✓	0.45 ✗

L-SPAMINE Trace

Transaction database

time	items	time	items
t1	A	t2	B, C
t1	A, B, C	t2	B
t1	A, C	t2	A, B, C
t1	A	t2	A, B, C
t1	A, B, C	t2	C
t1	C	t2	A, B, C
t1	C	t2	A, C
t1	A, B, C	t2	C
t1	C	t2	B
t1	C	t2	B, C

Dissimilarity threshold : 0.2
Similarity Function: Euclidean



Reference sequence

t1	0.4
t2	0.6

Size 2

Upper bound sequences

~~A B~~	A C	B C
~~0.3~~	0.6	0.3
~~0.4~~	0.4	0.7

Upper LB dist: ~~0.22~~ ✗ 0.20 ✓ 0.10 ✓

Lower bound sequences

A B	A C	B C
	0.4	0.1
	0.2	0.5

Lower LB dist: 0.0 0.0

LB dist: 0.20 ✓ 0.10 ✓

Size 2

Support sequences

A C	B C
0.6	0.3
0.4	0.5

Upper LB dist: 0.20 ✓ 0.14 ✓

True dist: 0.20 ✓ 0.14 ✓

~~Size 3~~

~~A B C~~

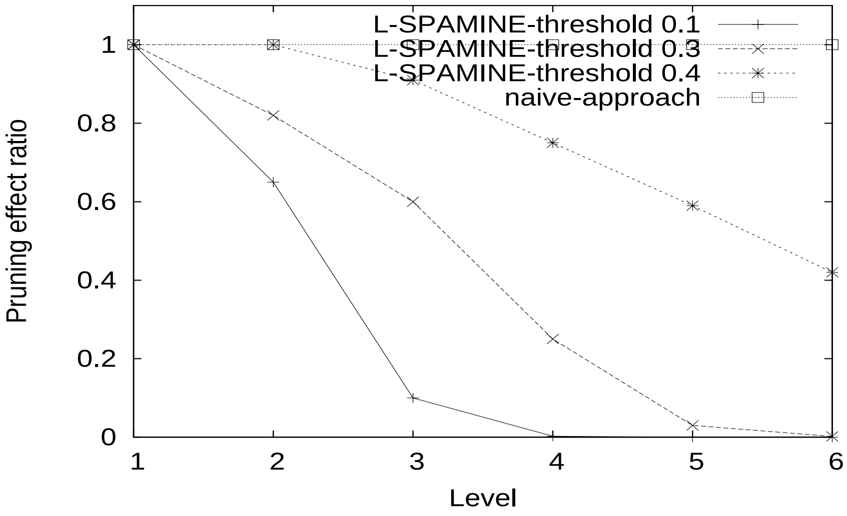
*** Similar itemsets:**
{B} :<0.3., 0.7> (0.14)
{A,C}:<0.6., 0.4> (0.2)
{B,C}:<0.3., 0.5> (0.14)

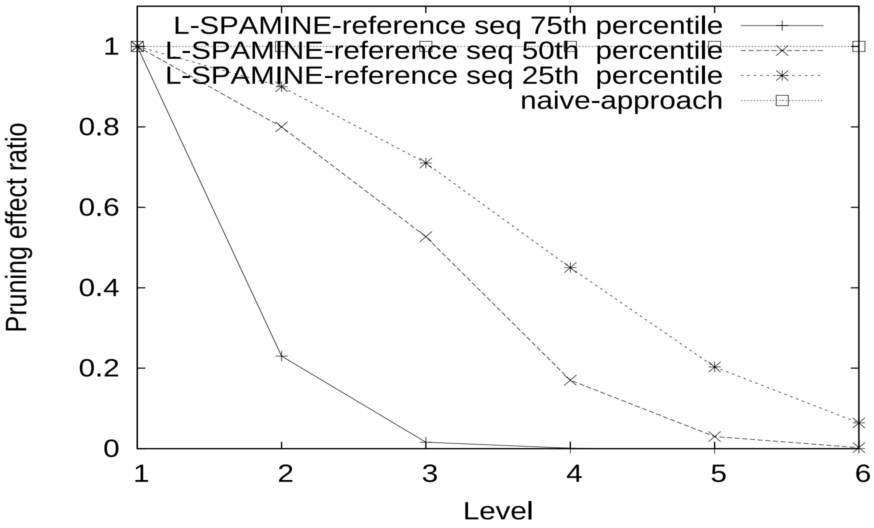
Experiment

- Datasets
    - Synthetic datasets: a modified IBM data generator
    - Real dataset: Earth Climate
    - Query sequences: randomly chosen in different quintiles of supports
- Test cases
    - Effect of lower bounding distance
    - Effect of database scanning method
    - Effect of number of items
    - Effect of number of time slots
    - Experiment with a real dataset

Experiment Results

- Effect of lower bounding distance pruning
  - TD100-D1-L10-I20-T100
  - Pruning effect ratio : the number of candidate itemsets which need database scan over the total number of possible itemsets per level

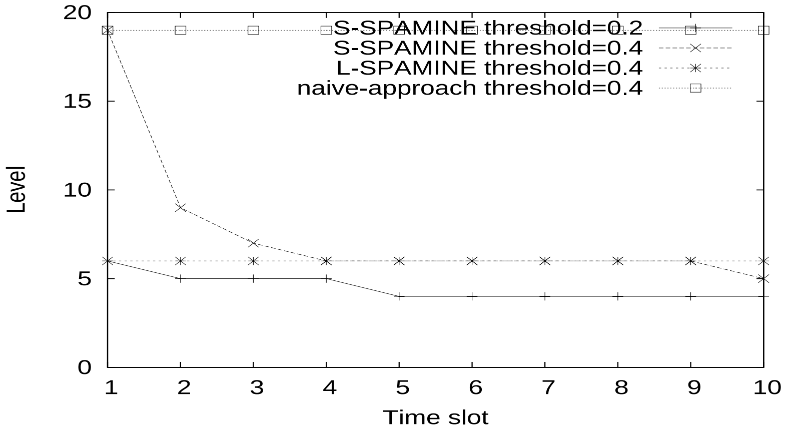


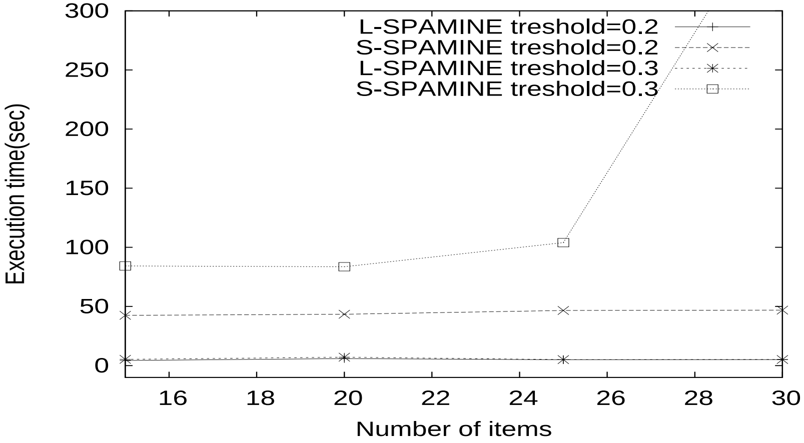


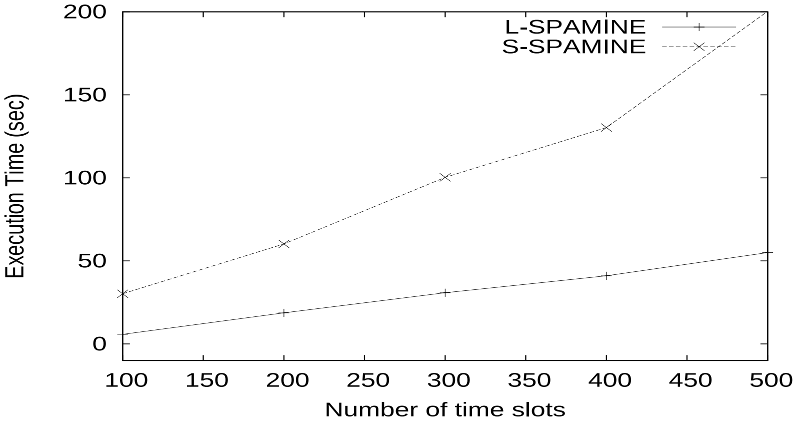
Experiment Results

- Effect of different scanning

  (TD100-D1-L10-I20-T100)







- Effect of number of items

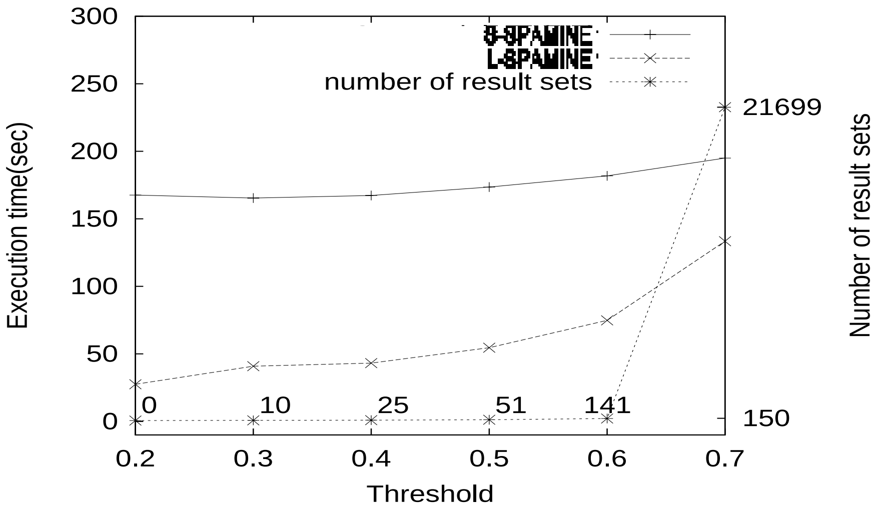
  (TD100-D1-L10-I*-T10)

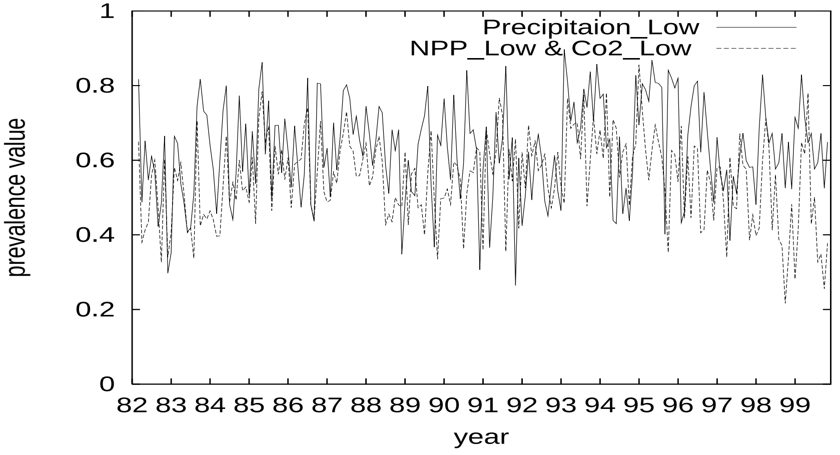
- Effect of number of time slots

  (TD*-D1-L6-I20-T*)

Experiment with a real dataset

- Dataset: Earth Climate
  - # of items: 50,
  - # of time slots: 214
  - # of transaction per time slot: 2827,
  - Total # of transaction: 64,978
- Reference sequences
  - SOI index
    - Normalization to 0 to 1 range.
  - Prevalence sequence of low participation





Conclusion

- Summary
  - Formulate the problem of mining similarity-profiled temporal association patterns
  - Propose a novel algorithm
    - Substantially reduce the search space by pruning candidate itemsets using lower bounding distance and the monotonicity of upper lower bounding distance
  - Experimentally evaluate the algorithm
- Future Work
  - Explore different similarity measures with different similarity models.