

---

# Mining Sequence Data



JERZY STEFANOWSKI

Inst. Informatyki PP

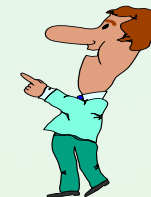
Wersja dla TPD 2009

„Zaawansowana eksploracja danych”

# Outline of the presentation

---

1. Relationships to mining frequent items
2. Motivations for sequence databases and their analysis
3. Applications
4. Approximate queries and basic techniques
5. Classification in data streams
6. Clustering
7. Conclusions



This lecture is partly based on the following resources - slides:  
J.Han (data mining book), slides Pinto, Pei, etc.  
and my other notes.

# What Is Frequent Pattern Analysis?

---

- ❑ **Frequent pattern**: a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- ❑ First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of **frequent itemsets** and **association rule mining**
- ❑ Motivation: Finding inherent regularities in data
  - What products were often purchased together? — Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- ❑ Applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

## Why is Frequent Pattern or Association Mining an Essential Task in Data Mining?

---

- ❑ Foundation for many essential data mining tasks
  - Association, correlation, causality
  - Sequential patterns, temporal or cyclic association, partial periodicity, spatial and multimedia association
  - Associative classification, cluster analysis, fascicles (semantic data compression)
- ❑ DB approach to efficient mining massive data
- ❑ Broad applications
  - Basket data analysis, cross-marketing, catalog design, sale campaign analysis
  - Web log (click stream) analysis, DNA sequence analysis, etc

# Sequence Databases and Sequential Pattern Analysis

---

- ❑ Transaction databases → sequence databases
- ❑ Frequent patterns vs. (frequent) sequential patterns
  - Finding time-related frequent patterns (frequent subsequences)
- ❑ Applications of sequential pattern mining
  - Customer shopping sequences:
    - First buy computer, then CD-ROM, and then digital camera, within 3 months.
  - Medical treatment, natural disasters (e.g., earthquakes), science & engineering processes, stocks and markets, etc.
  - Telephone calling patterns,
  - Weblog click streams
  - DNA sequences and gene structures

Seq. ID	Sequence
10	<(bd)cb(ac)>
20	<(bf)(ce)b(fg)>
30	<(ah)(bf)abf>
40	<(be)(ce)d>
50	<a(bd)bcb(ade)>

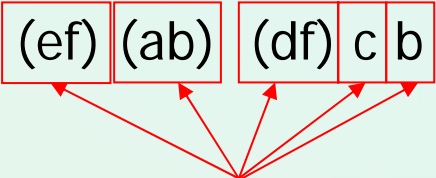
# What Is Sequential Pattern Mining?

- Given a set of sequences, find the complete set of *frequent* subsequences

## A *sequence database*

SID	sequence
10	<a(abc)(ac)d(cf)>
20	<(ad)c(bc)(ae)>
30	<(ef)(ab)(df)cb>
40	<eg(af)cbc>

A *sequence*: < (ef) (ab) (df) c b >



An element may contain a set of items. Items within an element are unordered and we list them alphabetically.

<a(bc)dc> is a *subsequence* of <a(abc)(ac)d(cf)>

Given *support threshold*  $min\_sup = 2$ , <(ab)c> is a *sequential pattern*

# Challenges on Sequential Pattern Mining

---

- ❑ A **huge** number of possible sequential patterns are hidden in databases
- ❑ A mining algorithm should
  - find the **complete set of patterns**, when possible, satisfying the minimum support (frequency) threshold
  - be highly **efficient, scalable**, involving only a small number of database scans
  - be able to incorporate various kinds of **user-specific constraints**



# Studies on Sequential Pattern Mining

---

- ❑ Concept introduction and an initial Apriori-like algorithm
  - R. Agrawal & R. Srikant. “Mining sequential patterns,” ICDE’95
- ❑ **GSP—An Apriori-based, influential mining method** (developed at IBM Almaden)
  - R. Srikant & R. Agrawal. “Mining sequential patterns: Generalizations and performance improvements,” EDBT’96
- ❑ FreeSpan and PrefixSpan (Han et al.@KDD’00; Pei, et al.@ICDE’01)
  - Projection-based
  - But only prefix-based projection: less projections and quickly shrinking sequences
- ❑ Vertical format-based mining: **SPADE** (Zaki00)

# A Basic Property of Sequential Patterns: Apriori

---

- A basic property: Apriori (Agrawal & Sirkant'94)
  - If a sequence  $S$  is not frequent
  - Then none of the super-sequences of  $S$  is frequent
  - E.g,  $\langle hb \rangle$  is infrequent  $\rightarrow$  so do  $\langle hab \rangle$  and  $\langle (ah)b \rangle$

Seq. ID	Sequence
10	$\langle (bd)cb(ac) \rangle$
20	$\langle (bf)(ce)b(fg) \rangle$
30	$\langle (ah)(bf)abf \rangle$
40	$\langle (be)(ce)d \rangle$
50	$\langle a(bd)bcb(ade) \rangle$

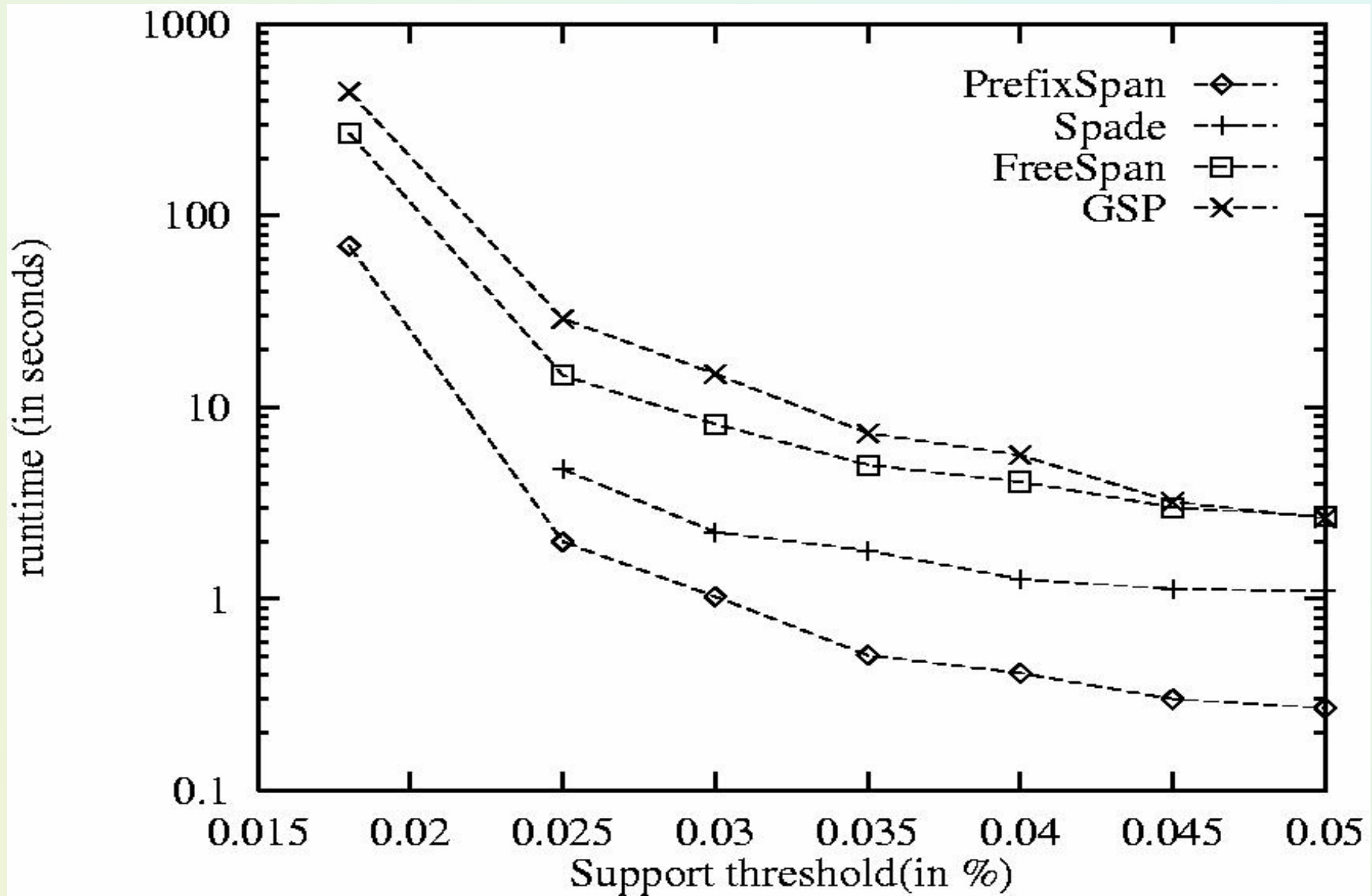
Given support threshold  
 $min\_sup = 2$

# GSP—A Generalized Sequential Pattern Mining Algorithm

---

- ❑ GSP (Generalized Sequential Pattern) mining algorithm
  - proposed by Agrawal and Srikant, EDBT'96
- ❑ Outline of the method
  - Initially, every item in DB is a candidate of length-1
  - for each level (i.e., sequences of length-k) do
    - scan database to collect support count for each candidate sequence
    - generate candidate length-(k+1) sequences from length-k frequent sequences using Apriori
  - repeat until no frequent sequence or no candidate can be found
- ❑ Major strength: Candidate pruning by Apriori

# Performance on Data Set Gazelle



# Motivating Example

---

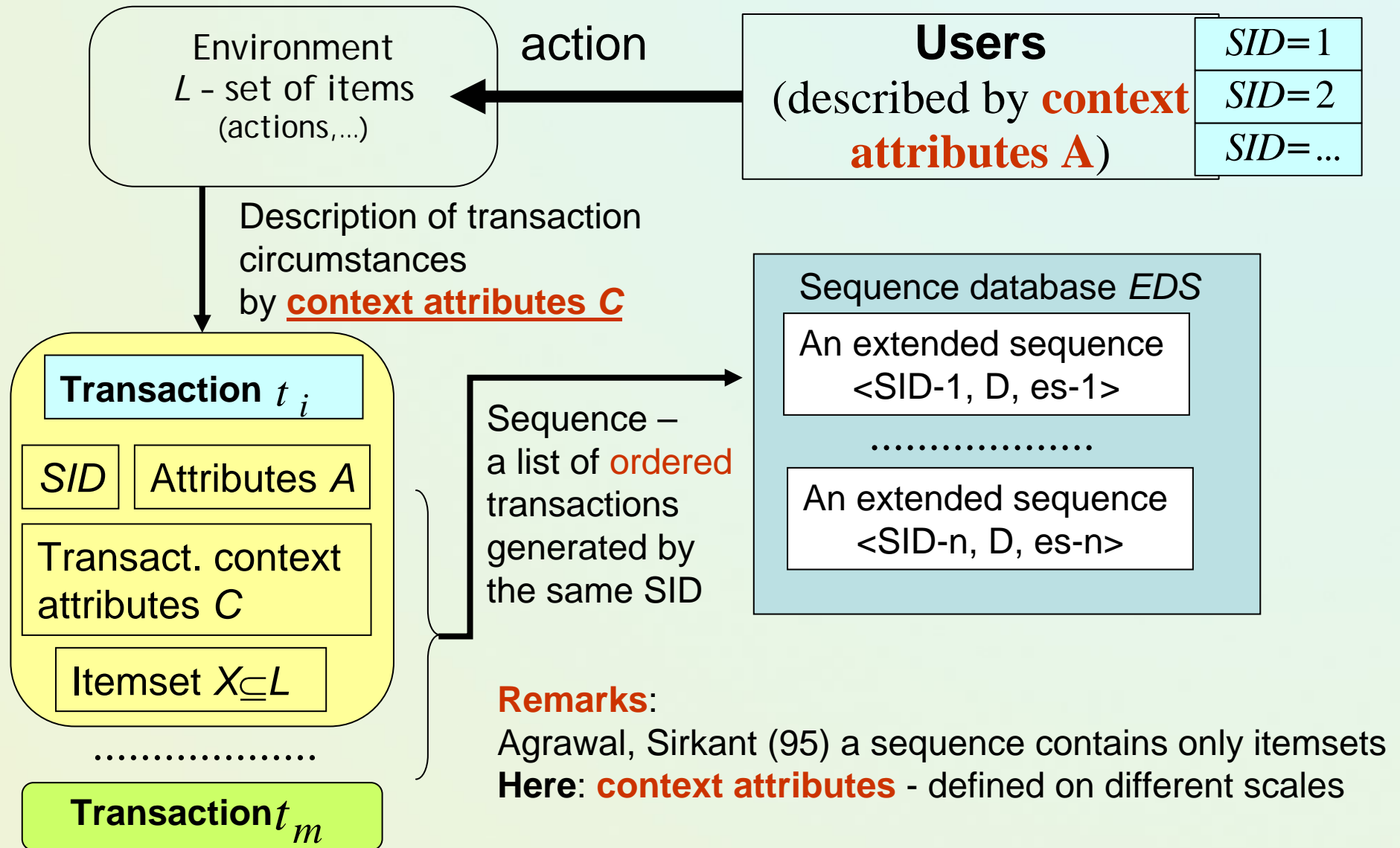
- ❑ Sequential patterns are useful
  - “free internet access → buy package 1 → upgrade to package 2”
  - Marketing, product design & development
- ❑ Problems: lack of focus
  - Various groups of customers may have different patterns
- ❑ MD-sequential pattern mining: integrate multi-dimensional analysis and sequential pattern mining

# Extensions of mining sequence patterns

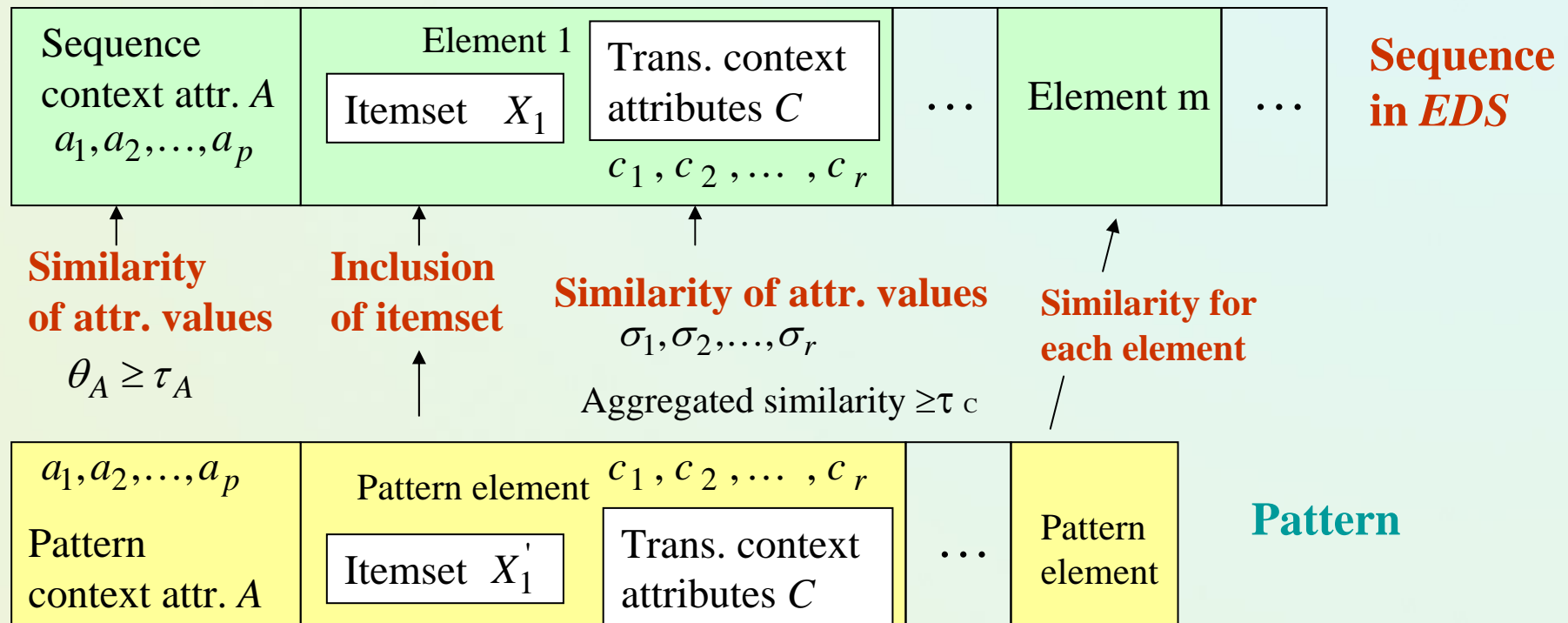
- ❑ Mining sequential patterns in a database of users' activities
  - Given a sequence database, where each sequence  $s$  is an ordered list of transactions  $t$  containing sets of items  $X \subseteq L$ , find all sequential patterns with a minimum support.
- ❑ An important task for Web usage mining
  - „20% users which access A page, then go to C page and finally select F page”
- ❑ Our contribution (R.Ziembinski, J.Stefanowski):
  - Extending a problem into a problem of mining **context based sequential patterns**.



# Extended sequence database



# Sequences supporting patterns



- Given a database *EDS* with sets of context attributes  $A$ ,  $C$  and their similarity functions, find all maximal sequential patterns among all sequences, which are similar to at least *min\_support* sequences in *EDS*.



# An example of e-bank customers

## Sequence /customer context:

*Monthly earnings, Martial status,  
Profession, Age*

## Transaction context:

*Time from money supply,  
Day of the week when action done*

## User actions:

SD –receive money, TM – transfer  
WM – withdraw money, CD – create  
time deposit, RD – cancel this deposit

## Sequences:

SID1	(2,Friday)	{TM,CD}
(4200,married,tech,24)	(4,Sunday)	{WM}
	(20,Saturday)	{RD,WM,TM}
SID2	(3,Tuesday)	{TM,CD,WM}
(4000,married,tech,22)	(7,Sunday)	{WM,CD}
	(20,Saturday)	{RD,WM}
	(1,Tuesday)	{TM,CD}
SID3	(3,Monday)	{CD,TM,WM}
(1500,single,retired,70)	(10,Monday)	{CD,TM,WM}
	(16,Sunday)	{WM}

...

...

## Examples of patterns:

- Traditional sequential pattern:

$\langle \{TM,CD\}, \{WM\}, \{WM,RD\} \rangle$

- Extended context sequential pattern:

$(4000,married,*,*) \langle (3,*)\{TM,CD\}, (*,Sunday)\{WM\}, (20,*)\{WM,RD\} \rangle$