Generalized Sequential Patterns (GSP) Mining

- **Proposed by R. Srikant and R. Agrawal in IBM Almaden, 1996**

- **Drawbacks of existing mining methods**
  - **absence of time constraints**
    - **Users often want to specify maximum or minimum time gaps between adjacent elements of the sequential pattern**

  <e.g.>

    - **A bookstore may not care if someone bought “Gone with Wind”, followed by “Titanic” three years later**
    - **A sequence is meaningful only if adjacent elements occur within a specified time interval, say two months**

GSP Mining (cont.)

- Rigid definition of a transaction
  - sliding time window
  - <e.g.>

    If the bookstore specifies a time window of a week

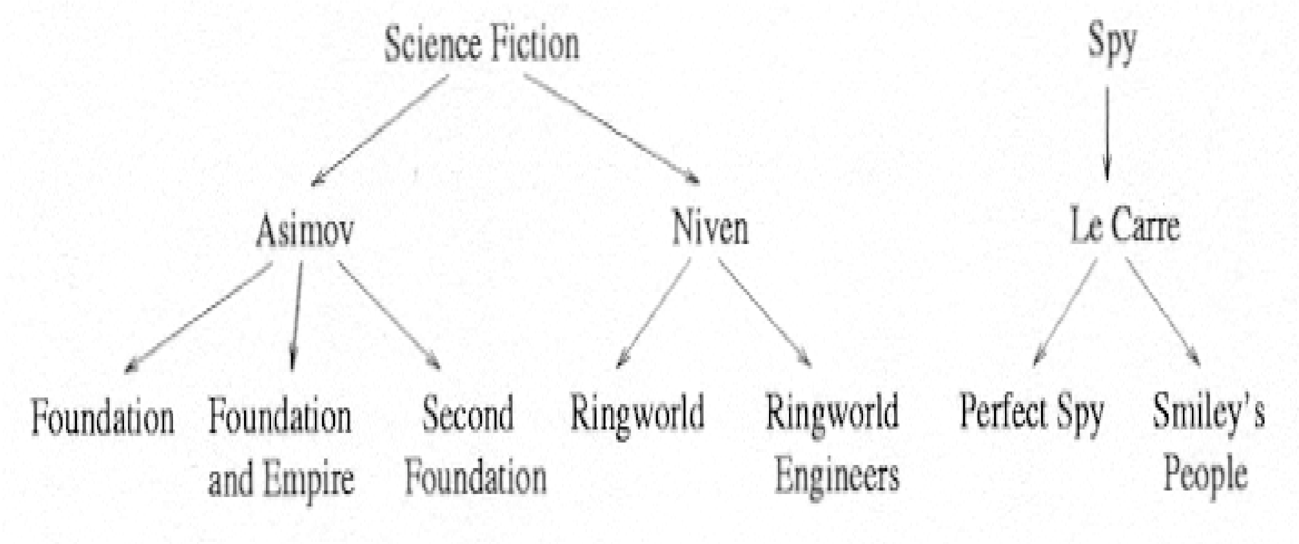
    Then a customer who bought "Foundation" on Monday
    "Ringworld" on Saturday
    "Last Empire" a few weeks later

    Will support the sequence

    <(Foundation, Ringworld), (Last Empire)>

GSP Mining (cont.)

- **Absence of taxonomy**



If a customer bought “Foundation” followed by “Perfect Spy”

Supported Sequence:

<(Foundation), (Perfect Spy)>

<(Asimov), (Perfect Spy)>

<(Science Fiction), (Le Carre)>

GSP Mining (cont.)

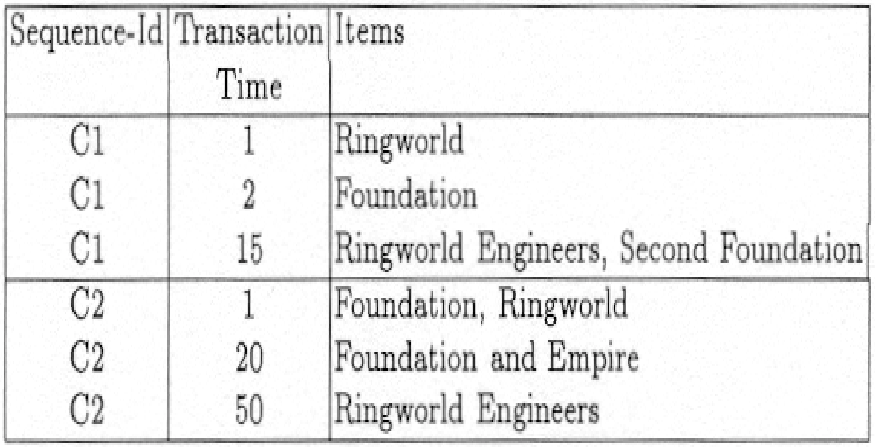
- **New Definitions**
  - plus taxonomy
    - a transaction T contains an item x if x is in T or x is an ancestor of some item in T
  - plus sliding window
    - a data-sequence d = $<d_1 \ldots d_m>$ contains a sequence s= $<s_1 \ldots s_n>$ if there exist integers $l_1 \leq u_1 \leq l_2 \leq u_2 \leq \ldots \leq l_n \leq u_n$ such that

      1. s_i is contained in union of d_k (from $_{ui}$ to $_{li}$) $1 \leq i \leq n$, and
      2. Transaction-time(d_{ui}) - transaction-time(d_{li}) $\leq$ window-size, $1 \leq i \leq n$
  - plus time constraints
    - a data-sequence d = $<d_1 \ldots d_m>$ contains a sequence s= $<s_1 \ldots s_n>$ if there exist integers $l_1 \leq u_1 \leq l_2 \leq u_2 \leq \ldots \leq l_n \leq u_n$ such that

      1. s_i is contained in union of d_k (from $_{ui}$ to $_{li}$) $1 \leq i \leq n$,
      2. Transaction-time(d_{ui}) - transaction-time(d_{li}) $\leq$ window-size, $1 \leq i \leq n$
      3. Transaction-time(d_{li}) - transaction-time(d_{li-1}) $\leq$ window-size, $2 \leq i \leq n$
      4. Transaction-time(d_{ui}) - transaction-time(d_{li-1}) $\leq$ window-size, $2 \leq i \leq n$

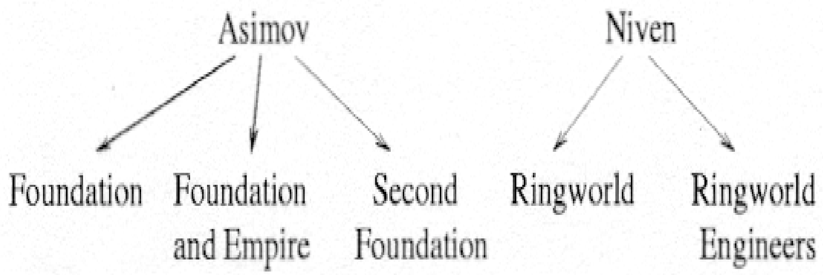
GSP Mining (cont.)

Database $\mathcal{D}$



Sequence-Id	Transaction Time	Items
C1	1	Ringworld
C1	2	Foundation
C1	15	Ringworld Engineers, Second Foundation
C2	1	Foundation, Ringworld
C2	20	Foundation and Empire
C2	50	Ringworld Engineers

Taxonomy $\mathcal{T}$



Let minimum support = 2 sequences

SP1 = <(Ringworld) (Ringworld Engineers)>

Setting sliding-window of 7 days will add

SP2 = < (Foundation, Ringworld) (Ringworld Engineers)>

Setting max-gap of 30 days will drop both SP1 and SP2

Add taxonomy only will add

SP3 = <(Foundation) (Asimov)>

GSP Mining (cont.)

- The Method
  - Candidate generation
    - Join Phase
    - Prune Phase
  - Counting candidates
    - reduction
    - checking whether a data-sequence contains a specific sequence
      - forward phase
      - backward phase
  - Taxonomies

- Performance
  - GSP is 2 to 20 times faster than AprioriAll