Aula 13 – Mineração de Dados
Regras de Associação - Parte2

Profa. Elaine Faria

UFU

Material

- Este material foi construído por meio de traduções dos slides do do prof. Tan, disponíveis em:
  - https://www-users.cse.umn.edu/~kumar001/dmbook/index.php
  - Todas as figuras usadas foram retiradas dos slides do prof. Tan

- O material é baseado no livro
  - Tan P., SteinBack M. e Kumar V. Introduction to Data Mining, Pearson, 2006.
  - Faceli, K., Lorena, A. C., Gama, J., Carvalho, A. C. P. L. F., Inteligência Artificial: Uma abordagem de Aprendizado de Máquina, LTC, 2011.

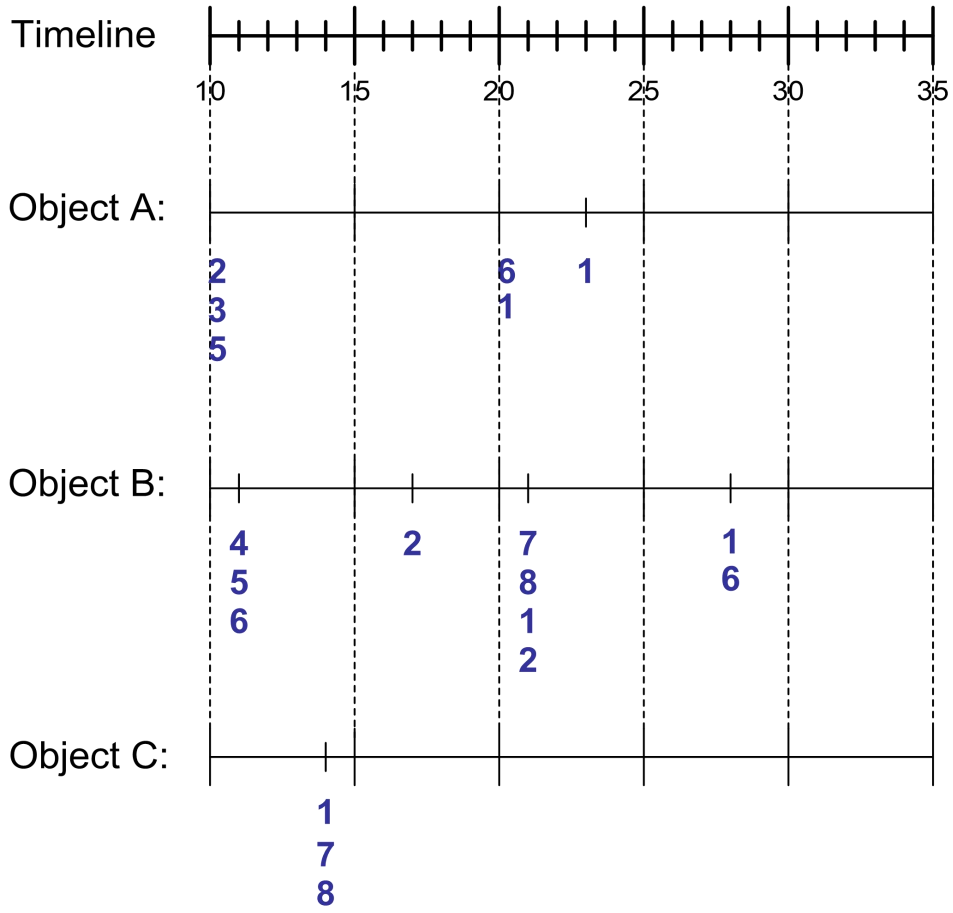
Padrões Sequenciais

- As cesta de compras do supermercado contém uma informação temporal sobre quando um item foi comprado pelo cliente
  - Pode ser usado para colocar junto uma sequência de transações feitas por um cliente em um determinado período de tempo
  - Conhecer a evolução das compras de seus clientes
    - EX: “que sequência de produtos são comprados por um mesmo cliente em momentos consecutivos”?
    - Se você pudesse descobrir que uma sequência de produtos< p1, p2, p3> é frequentemente comprada nesta ordem pelos clientes, você poderia enviar folhetos promocionais envolvendo os produtos p2 ou p3 para aqueles clientes que compraram o produto p1.

Exemplo de uma base de dados sequências

Base de Dados de Sequência

Sequence ID	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7



Exemplo de sequências

- Sequência de diferentes transações de um cliente em uma loja online:

  < {Camera Digital,iPad} {cartão de memória} {headphone,capa de iPad} >

- Sequência de livros retirados em uma biblioteca:

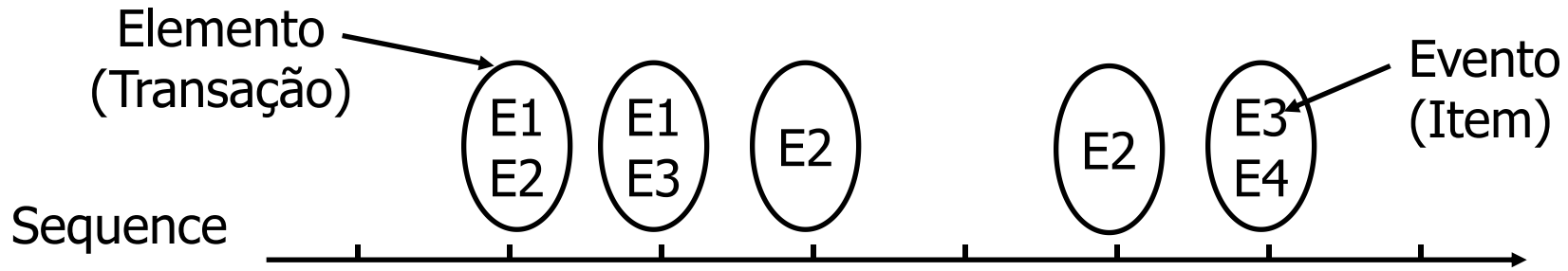
  <{Sociedade do Anel} {As Duas Torres} {Retorno do Rei}>

Descoberta de padrões sequenciais: Exemplo

- Sequências de transações de ponto de vendas
  - Livraria de informática:
    (Intro_To_Visual_C) (C++_Primer) -->
    (Perl_for_dummies,Tcl_Tk)
  - Loja de roupas esportivas:
    (Shoes) (Racket, Racketball) --> (Sports_Jacket)

Exemplo de elementos e eventos em uma base de dados sequencial

Bases de Dados	Sequência	Elemento (Transação)	Evento (Item)
Cliente	Histórico de compras de um determinado cliente	Um conjunto de itens comprados por um cliente no momento t	Livros, produtos diários, CDs, etc.
Dados da Web	Atividade de navegação de um determinado visitante da web	Uma coleção de arquivos visualizados por um visitante da web após um único clique do mouse	Página inicial, página de índice, informações de contato,
Dados de Evento	Histórico de eventos gerados por um determinado sensor	Eventos disparados por um sensor no tempo t	Tipos de alarmes gerados por sensores



Dados de sequência vs. Dados de cesta de compras

Base de Dados de Sequência

Customer	Date	Items bought
A	10	2, 3, 5
A	20	1,6
A	23	1
B	11	4, 5, 6
B	17	2
B	21	1,2,7,8
B	28	1, 6
C	14	1,7,8

Dados de cesta de compras

Events
2, 3, 5
1,6
1
4,5,6
2
1,2,7,8
1,6
1,7,8

Definição formal de uma sequência

- Uma sequência (ou padrão sequencial) é uma lista ordenada de elementos (itemsets)

  $s = < e_1, e_2, e_3 \ldots >$

  - Cada elemento contém uma coleção de eventos (items)

    $e_i = \{i_1, i_2, \ldots, i_k\}$

- Comprimento da sequência, $|s|$, é dada pelo nro de elementos na sequência
- Uma k-sequência é uma sequência que contem k eventos (items)
  - <{a,b} {a}> tem o comprimento 2 e é uma 3-sequência
    - Essa definição é usada pelo algoritmo GSP, mas é ligeiramente diferente no algoritmo AprioriAll

Definição formal de uma sub-sequência

- Uma sequência t: $<a_1\ a_2 \ldots a_n>$ **está contida** em outra sequência s: $<b_1\ b_2 \ldots b_m>$ $(m \geq n)$ se existe inteiros : $i_1 < i_2 < \ldots < i_n$ tal que $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, …, $a_n \subseteq b_{in}$
- Exemplo ilustrativo:

s: b_1 b_2 b_3 b_4 b_5

t: a_1 a_2 a_3

t **é uma subsequência** de s se $a_1 \subseteq b_2$, $a_2 \subseteq b_3$, $a_3 \subseteq b_5$.

Data sequence	Subsequence	Contain?
< {2,4} {3,5,6} {8} >	< {2} {8} >	Yes
< {1,2} {3,4} >	< {1} {2} >	No
< {2,4} {2,4} {2,5} >	< {2} {4} >	Yes
<{2,4} {2,5} {4,5}>	< {2} {4} {5} >	No
<{2,4} {2,5} {4,5}>	< {2} {5} {5} >	Yes
<{2,4} {2,5} {4,5}>	< {2, 4, 5} >	No

Mineração de Padrão Sequencial: Definição

- **Dados:**
  - uma base de sequências
  - um limite mínimo de suporte especificado pelo usuário, *minsup*

- **Tarefa:**
  - Encontrar todas as subsequências com suporte ≥ *minsup*

- O suporte de uma subsequência *w* é definido como a fração de todas as sequências de dados que contém *w*
- Um *padrão sequencial* é uma subsequência frequente (i.e., uma subsequência cujo suporte é ≥ *minsup*)

Mineração de Padrão Sequencial: Exemplo

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

Minsup = 50%

Examplos de Subsequências Frequentes:

< {1,2} >	s=60%
< {2,3} >	s=60%
< {2,4}>	s=80%
< {3} {5}>	s=80%
< {1} {2} >	s=80%
< {2} {2} >	s=60%
< {1} {2,3} >	s=60%
< {2} {2,3} >	s=60%
< {1,2} {2,3} >	s=60%

Base de dados sequencial vs. Cesta de compra

Base de Dados de Sequências

Customer	Date	Items bought
A	10	2, 3, 5
A	20	1,6
A	23	1
B	11	4, 5, 6
B	17	2
B	21	1,2,7,8
B	28	1, 6
C	14	1,7,8

{2} -> {1}

$$conf(\{2\} \rightarrow \{1\}) = \frac{\sigma(\{2\}\,\{1\})}{\sigma(\{2\})}$$

Base de Dados de cesta de compras

Eventos
2, 3, 5
1,6
1
4,5,6
2
1,2,7,8
1,6
1,7,8

(1,8) -> (7)

$$conf(1,8) \rightarrow (7)) = \frac{\sigma(1,7,8)}{\sigma(\{1,8\})}$$

Extraindo padrões sequenciais

- Dados n eventos: $i_1, i_2, i_3, \ldots, i_n$
- 1-subsequência candidata

  $\langle\{i_1\}\rangle, \langle\{i_2\}\rangle, \langle\{i_3\}\rangle, \ldots, \langle\{i_n\}\rangle$
- 2-subsequências candidatas:

  $\langle\{i_1, i_2\}\rangle, \langle\{i_1, i_3\}\rangle, \ldots,$

  $\langle\{i_1\}\ \{i_1\}\rangle, \langle\{i_1\}\ \{i_2\}\rangle, \ldots, \langle\{i_n\}\ \{i_n\}\rangle$
- 3-subsequências candidatas:

  $\langle\{i_1, i_2, i_3\}\rangle, \langle\{i_1, i_2, i_4\}\rangle, \ldots,$

  $\langle\{i_1, i_2\}\ \{i_1\}\rangle, \langle\{i_1, i_2\}\ \{i_2\}\rangle, \ldots,$

  $\langle\{i_1\}\ \{i_1, i_2\}\rangle, \langle\{i_1\}\ \{i_1, i_3\}\rangle, \ldots,$

  $\langle\{i_1\}\ \{i_1\}\ \{i_1\}\rangle, \langle\{i_1\}\ \{i_1\}\ \{i_2\}\rangle, \ldots$

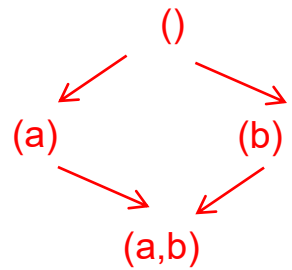
Extraindo padrões sequenciais: Exemplo simples

- Dados 2 eventos: a, b
- 1-subsequência candidata

  <{a}>, <{b}>
- 2-subsequências candidatas

  <{a} {a}>, <{a} {b}>, <{b} {a}>, <{b} {b}>, <{a, b}>.
- 3-subsequências candidatas

  <{a} {a} {a}>, <{a} {a} {b}>, <{a} {b} {a}>, <{a} {b} {b}>,
  <{b} {b} {b}>, <{b} {b} {a}>, <{b} {a} {b}>, <{b} {a} {a}>
  <{a, b} {a}>, <{a, b} {b}>, <{a} {a, b}>, <{b} {a, b}>



Item-set patterns