

Data Mining - Sequence Mining

By Dr Heng Tao SHEN
School of Information Technology and Electrical Engineering
The University Of Queensland
<http://www.itee.uq.edu.au/~shenht>

Introduction

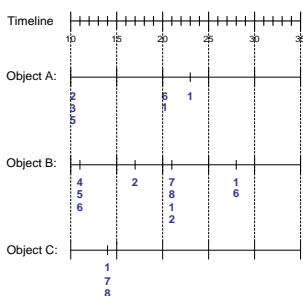
- What is sequence mining?
 - Well...

P. 2

Sequence Data

Sequence Database:

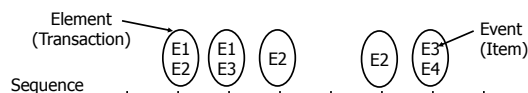
Object	Timestamp	Events
A	10	2, 3, 5
A	20	6, 1
A	23	1
B	11	4, 5, 6
B	17	2
B	21	7, 8, 1, 2
B	28	1, 6
C	14	1, 8, 7



P. 3

Examples of Sequence Data

Sequence Database	Sequence	Element (Transaction)	Event (Item)
Customer	Purchase history of a given customer	A set of items bought by a customer at time t	Books, diary products, CDs, etc
Web Data	Browsing activity of a particular Web visitor	A collection of files viewed by a Web visitor after a single mouse click	Home page, index page, contact info, etc
Event data	History of events generated by a given sensor	Events triggered by a sensor at time t	Types of alarms generated by sensors
Genome sequences	DNA sequence of a particular species	An element of the DNA sequence	Bases A, T, G, C



P. 4

Formal Definition of a Sequence

- A sequence is:
 - An ordered list of elements (transactions)
 - $s = \langle e_1, e_2, e_3, \dots \rangle$
 - Each element contains a collection of events (items)
 - $e_i = \{i_1, i_2, \dots, i_k\}$
 - Each element is attributed to a specific time or location
- Length of a sequence, $|s|$, is given by the number of elements of the sequence
- A k-sequence is a sequence that contains k events (items)

P. 5

Examples of Sequence

- Web purchasing sequence:
 - $\{\{\text{Homepage}\} \{\text{Electronics}\} \{\text{Digital Cameras}\} \{\text{Canon Digital Camera}\} \{\text{Shopping Cart}\} \{\text{Order Confirmation}\} \{\text{Return to Shopping}\}\}$
- Sequence of books checked out at a library:
 - $\{\{\text{Fellowship of the Ring}\} \{\text{The Two Towers}\} \{\text{Return of the King}\}\}$

P. 6

Candidate Generation

- Special case (k=2):
 - Merging two frequent 1-sequences $\langle \{i_1\} \rangle$ and $\langle \{i_2\} \rangle$ will produce two candidate 2-sequences: $\langle \{i_1\} \{i_2\} \rangle$ and $\langle \{i_1, i_2\} \rangle$
- General case (k>2):
 - A frequent (k-1)-sequence (w_1) is merged with another frequent (k-1)-sequence (w_2) to produce a candidate k-sequence if:
 - The subsequence obtained by removing the first item in w_1 is the same as the subsequence obtained by removing the last item in w_2
 - The resulting candidate after merging is given by the sequence w_1 extended with the last event of w_2 .
 - If the last two events in w_2 belong to the same element, then the last event in w_2 becomes part of the last element in w_1
 - Otherwise, the last event in w_2 becomes a separate element appended to the end of w_1

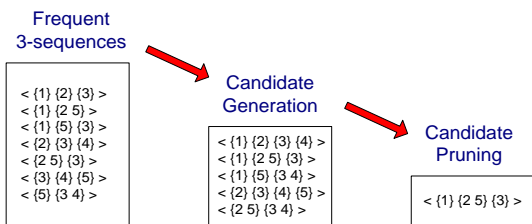
P. 13

Candidate Generation Examples

- Merging the sequences:
 - $w_1 = \langle \{1\} \{2,3\} \{4\} \rangle$ and $w_2 = \langle \{2,3\} \{4,5\} \rangle$ will produce the candidate sequence $\langle \{1\} \{2,3\} \{4,5\} \rangle$
- Merging the sequences:
 - $w_1 = \langle \{1\} \{2,3\} \{4\} \rangle$ and $w_2 = \langle \{2,3\} \{4\} \{5\} \rangle$ will produce the candidate sequence $\langle \{1\} \{2,3\} \{4\} \{5\} \rangle$
- We do not have to merge the sequences:
 - $w_1 = \langle \{1\} \{2,6\} \{4\} \rangle$ and $w_2 = \langle \{1\} \{2\} \{4,5\} \rangle$

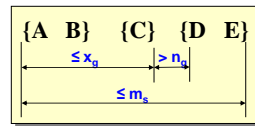
P. 14

GSP Example



P. 15

Timing Constraints I



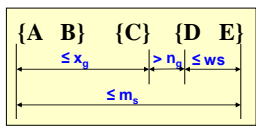
x_g : max-gap
 n_g : min-gap
 m_s : maximum span

Constraint: $x_g = 2, n_g = 0, m_s = 4$

Data sequence	Subsequence	Contain?
$\langle \{2,4\} \{3,5,6\} \{4,7\} \{4,5\} \{8\} \rangle$	$\langle \{6\} \{5\} \rangle$	Yes
$\langle \{1\} \{2\} \{3\} \{4\} \{5\} \rangle$	$\langle \{1\} \{4\} \rangle$	No
$\langle \{1\} \{2,3\} \{3,4\} \{4,5\} \rangle$	$\langle \{2\} \{3\} \{5\} \rangle$	Yes
$\langle \{1,2\} \{3\} \{2,3\} \{3,4\} \{2,4\} \{4,5\} \rangle$	$\langle \{1,2\} \{5\} \rangle$	No

P. 16

Timing Constraints II



x_g : max-gap
 n_g : min-gap
 ws : window size
 m_s : maximum span

Constraint: $x_g = 2, n_g = 0, ws = 1, m_s = 5$

Data sequence	Subsequence	Contain?
$\langle \{2,4\} \{3,5,6\} \{4,7\} \{4,6\} \{8\} \rangle$	$\langle \{3\} \{5\} \rangle$	No
$\langle \{1\} \{2\} \{3\} \{4\} \{5\} \rangle$	$\langle \{1,2\} \{3\} \rangle$	Yes
$\langle \{1,2\} \{2,3\} \{3,4\} \{4,5\} \rangle$	$\langle \{1,2\} \{3,4\} \rangle$	Yes

P. 17

Note that Apriori Principle Not Holds...

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

Suppose:
 $x_g = 1$ (max-gap)
 $n_g = 0$ (min-gap)
 $m_s = 5$ (maximum span)
 $minsup = 60\%$
 $\langle \{2\} \{5\} \rangle$ support = 40%
 but
 $\langle \{2\} \{3\} \{5\} \rangle$ support = 60%

Problem exists because of max-gap constraint

No such problem if max-gap is infinite

P. 18

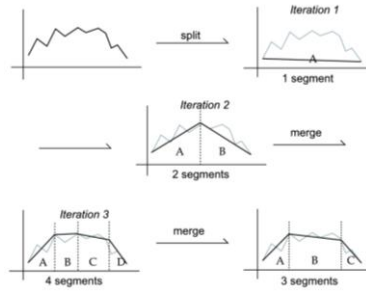
Time Series Segmentation

- What is time series segmentation?
 - Approximate a time series of length N by K straight lines, where $K \ll N$
 - Example:
 - Why it is useful?
 - Well...

P. 19

Offline Algorithm

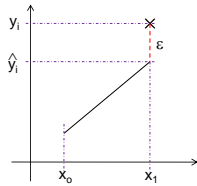
- General Idea:
 - Split and merge



P. 20

Online Algorithm

- General Idea:
 - When a new point comes, extend the current segment to see whether the error (ϵ) between the new point and the projected point is larger than a specific threshold
 - If No:
 - Continue the segment
 - If Yes:
 - Break the segment and continue at the new point



P. 21

Online and Offline Algorithm

- Major question:
 - How to determine when to split (or merge)?
 - i.e. how to set the thresholds?

Equation of a straight line: $\frac{y - y_1}{x - x_1} = \frac{y_1 - y_2}{x_1 - x_2}$

$$SSE(y_i, \hat{y}_i) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

P. 22