Sequential Pattern Mining and Frequent Subgraph Mining

Rina Singh

Tennessee Tech. University

April 12, 2017

Outline

1 Sequential Pattern Mining

2 Frequent Subgraph Mining

Sequential Pattern Mining

Sequential Pattern Mining

Sequential pattern mining focuses on identifying statistically relevant patterns in data represented by sequences of discrete items or events. It is closely related to time series mining, where data is represented by sequences of real (numerical) values.

Sequence

Sequence

A *sequence* is simply on ordered list. Depending on the mining task, a sequence can consist of items or sets of items.

Sequence of Items

- $\langle$breakfast, lunch, dinner$\rangle$
- $\langle$breakfast, second breakfast, elevenses, luncheon, afternoon tea, dinner, supper$\rangle$

Sequence of Sets of Items

- $\langle\{\texttt{cereal}\}, \{\texttt{apple,banana}\}, \{\texttt{bread}, \texttt{fish}\}\rangle$

Subsequence

Subsequence

A sequences S' is said to be a *subseqence* of a sequence S if S' can be derived from S by deleting some elements without changing the order of the remaining elements.

Example

- $\langle g, f, b\rangle \prec \langle c, g, f, a, b\rangle$
- $\langle \{a\}, \{f, g\}\rangle \prec \langle \{a, b\}, \{c\}, \{f, g\}\rangle$

Nonexample

- $\langle g, a, f, b\rangle \not\prec \langle c, g, f, a, b\rangle$
- $\langle \{b, c\}, \{f, g\}\rangle \not\prec \langle \{a, b\}, \{c\}, \{f, g\}\rangle$

Support

Support

Let D be a sequential database. The *support* of a sequence S is the number of sequences in D for which S appears as a subsequence. If the support of a sequence meets a user-defined minimum value, then the sequence is called *frequent*.

Example

Sample Database

$\langle c, a, a, b, c \rangle$
$\langle a, b, c, b \rangle$
$\langle c, a, b, c \rangle$
$\langle a, b, b, c, a \rangle$

Subsequences

Support	Sequence	Support	Sequence
1	$\langle a, a, b \rangle$	2	$\langle a, a \rangle$
1	$\langle a, a, b, c \rangle$	2	$\langle a, b, b \rangle$
1	$\langle a, b, a \rangle$	2	$\langle c, a, b \rangle$
1	$\langle a, b, b, a \rangle$	2	$\langle c, a, b, c \rangle$
1	$\langle a, b, b, c \rangle$	3	$\langle c, a \rangle$
1	$\langle a, b, b, c, a \rangle$	3	$\langle c, b \rangle$
1	$\langle a, b, c, b \rangle$	4	$\langle a \rangle$
1	$\langle c, a, a \rangle$	4	$\langle a, b \rangle$
1	$\langle c, a, a, b \rangle$	4	$\langle a, b, c \rangle$
1	$\langle c, a, a, b, c \rangle$		

The Sequential Pattern Mining Problem

Sequential Pattern Mining

Let D be a sequential database and n a user-defined minimum support. The *sequential pattern mining problem* asks to find the complete set of frequent subsequences of D.

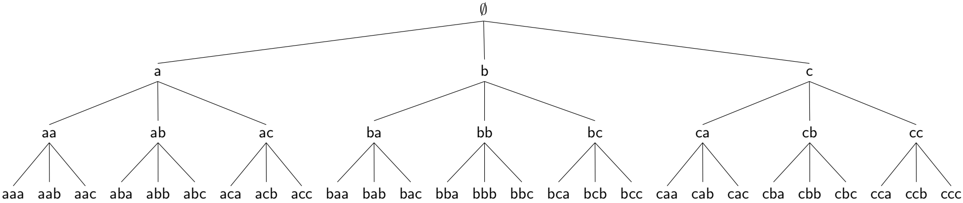
Mining Sequences of Items

Mining Sequences of Items

```
Require: $S$: a sequence
Require: $\mathcal{I}$: a set of items
Require: $D$: a sequential database
Require: $n$: a user-defined minimum support
procedure SequentialPatternMining($S$, $\mathcal{I}$, $D$, $n$)
    if support($S, D$) $\geq n$ then
        yield $S$
        for $item \in \mathcal{I}$ do
            $S' \leftarrow S + \langle item \rangle$
            SequentialPatternMining($S'$, $\mathcal{I}$, $D$, $n$)
        end for
    end if
end procedure
```

Sequential Pattern Mining Search Space

$\mathcal{I} = \{a, b, c\}$



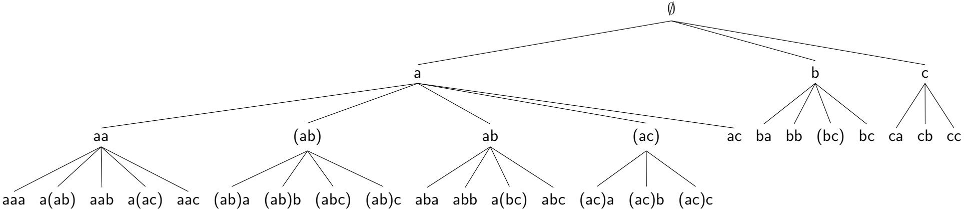
Mining Sequences of Sets of Items

Mining Sequences of Sets of Items

```
Require: $S$: a sequence
Require: $\mathcal{I}$: a set of items
Require: $D$: a sequential database
Require: $n$: a user-defined minimum support
procedure SequentialPatternMining($S$, $\mathcal{I}$, $D$, $n$)
    if support$(S, D) \geq n$ then
        yield $S$
        for $item \in \mathcal{I}$ do
            if $item > item', \forall item' \in S'.last$ then
                $S' \leftarrow S$
                $S'.last \leftarrow S'.last \cup item$
                SequentialPatternMining($S'$, $\mathcal{I}$, $D$, $n$)
            end if
            $S' \leftarrow S + \langle\{item\}\rangle$
            SequentialPatternMining($S'$, $\mathcal{I}$, $D$, $n$)
        end for
    end if
end procedure
```

Sequential Pattern Mining Search Space

$\mathcal{I} = \{a, b, c\}$



Sequential Pattern Mining Issues

Sequential Pattern Mining Issues

- Large Result Sets
- Computationally Expensive

Addressing Large Result Sets

Reducing Result Size

- Apriori Property
  - Every subsequence of a frequent sequence is also frequent.
- Mine Frequent Closed Patterns
  - A sequence is closed if it is not a subsequence of another sequence having the same support.
- Mine Frequent Maximal Patterns
  - A sequence is maximal if it is not a subsequence of another sequence having positive support.

Closed Patterns

Sample Database

$\langle c, a, a, b, c \rangle$
$\langle a, b, c, b \rangle$
$\langle c, a, b, c \rangle$
$\langle a, b, b, c, a \rangle$

Subsequences

1	$\langle a, a, b \rangle$	2	$\langle a, a \rangle$
1	$\langle a, a, b, c \rangle$	2	$\langle a, b, b \rangle$
1	$\langle a, b, a \rangle$	2	$\langle c, a, b \rangle$
1	$\langle a, b, b, a \rangle$	2	$\langle c, a, b, c \rangle$
1	$\langle a, b, b, c \rangle$	3	$\langle c, a \rangle$
1	$\langle a, b, b, c, a \rangle$	3	$\langle c, b \rangle$
1	$\langle a, b, c, b \rangle$	4	$\langle a \rangle$
1	$\langle c, a, a \rangle$	4	$\langle a, b \rangle$
1	$\langle c, a, a, b \rangle$	4	$\langle a, b, c \rangle$
1	$\langle c, a, a, b, c \rangle$		

Maximal Patterns

Sample Database

$\langle c, a, a, b, c \rangle$
$\langle a, b, c, b \rangle$
$\langle c, a, b, c \rangle$
$\langle a, b, b, c, a \rangle$

Subsequences

1	$\langle a, a, b \rangle$	2	$\langle a, a \rangle$
1	$\langle a, a, b, c \rangle$	2	$\langle a, b, b \rangle$
1	$\langle a, b, a \rangle$	2	$\langle c, a, b \rangle$
1	$\langle a, b, b, a \rangle$	2	$\langle c, a, b, c \rangle$
1	$\langle a, b, b, c \rangle$	3	$\langle c, a \rangle$
1	$\langle a, b, b, c, a \rangle$	3	$\langle c, b \rangle$
1	$\langle a, b, c, b \rangle$	4	$\langle a \rangle$
1	$\langle c, a, a \rangle$	4	$\langle a, b \rangle$
1	$\langle c, a, a, b \rangle$	4	$\langle a, b, c \rangle$
1	$\langle c, a, a, b, c \rangle$		

Addressing Computation Requirements

Computational Complexity

Sequential Pattern Mining has Exponential Time Complexity

- Sequences of Items: $\sum_{i=1}^{k} |\mathcal{I}|^k = \frac{|\mathcal{I}|^{k+1} - |\mathcal{I}|}{|\mathcal{I}| - 1}$
- Sequences of Itemsets: $\sum_{i=1}^{k} (2^{|\mathcal{I}|} - 1)^k = \frac{(2^{|\mathcal{I}|} - 1)^{k+1} - (2^{|\mathcal{I}|} - 1)}{(2^{|\mathcal{I}|} - 1) - 1}$

Efficient Mining

Efficient Mining Strategies

- Horizontal Databases - Sequences Stored in Arrays
    - Projected Databases
    - Pseudo-Projected Databases
- Vertical Databases - Sequences Stored in an Inverted Index

Search Space Reduction

Search Space Reduction

Numerous sequential pattern mining algorithms exist and center around more aggressive pruning strategies than that of apriori.

- SPADE
- SPAM
- PrefixSpan
- BIDE
- CloSpan

The majority of these algorithms involve closed/maximal pattern mining and early termination techniques.

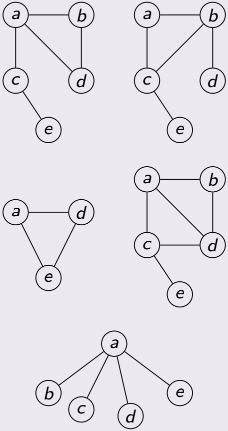
The Frequent Subgraph Mining Problem

Frequent Subgraph Mining

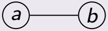
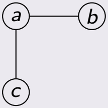
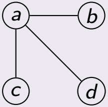
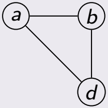
Let D be a graph database and n a user-defined minimum support. The *frequent subgraph mining problem* asks to find the complete set of frequent **connected** subgraphs of D.

Example

Sample Database



Subgraphs

Support	Subgraph
4	
4	
3	
2	

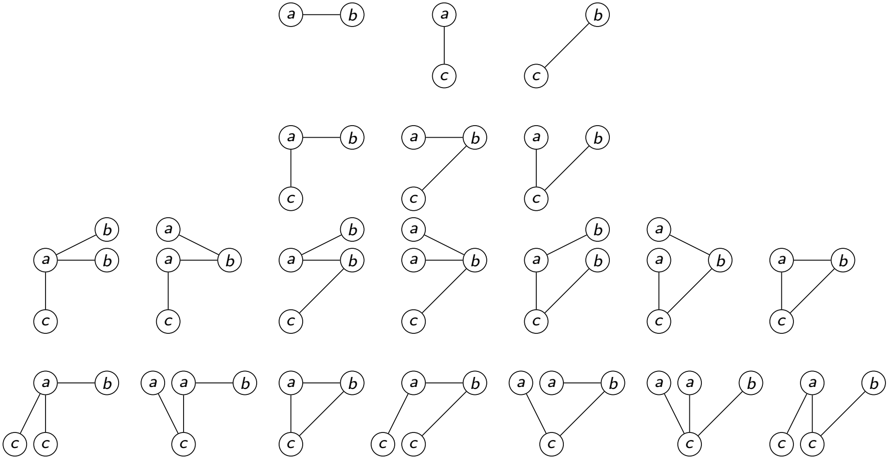
Frequent Subgraph Mining

Frequent Subgraph Mining

The basic frequent subgraph mining algorithm is similar to that of sequential pattern mining. However, graphs introduce a level of complexity not present in sequential pattern mining or frequent itemset mining.

- Generating Subgraphs
- Identifying Duplicate Subgraphs (Graph Isomorphism)
- Subgraph Isomorphism (NP-Complete)

Example



Questions?