An Introduction to Sequential Pattern Mining

Philippe Fournier-Viger

http://www.philippe-Fournier-viger.com


Source code and datasets available in the SPMF library
Data Mining: the goal is to discover or extract useful knowledge from data.

Many types of data can be analyzed: graphs, relational databases, time series, sequences, etc.

In this presentation, we focus on analyzing a common type of data called discrete sequences to find interesting patterns in it.
What is a discrete sequence?

A **sequence** is an ordered list of symbols.

**Example 1:** a sequence can be the items that are purchased by a customer over time:
What is a discrete sequence?

A **sequence** is an ordered list of symbols.

**Example 2:** a sequence can be the list of words in a sentence:

I → go → back → home
What is a discrete sequence?

A **sequence** is an ordered list of symbols.

**Example 3:** a sequence can be the list of locations visited by a car in a city
Sequential Pattern Mining

- It is a popular data mining task, introduced in 1994 by Agrawal & Srikant.
- The goal is to find all subsequences that appear frequently in a set of discrete sequences.
- **For example:**
  - find sequences of items purchased by many customers over time,
  - find sequences of locations frequently visited by tourists in a city,
  - Find sequences of words that appear frequently in a text.
Definition: **Items**

Let there be a **set** of **items** (symbols) called $I$.

**Example:** $I = \{a, b, c, d, e\}$

- $a = \text{apple}$
- $b = \text{bread}$
- $c = \text{cake}$
- $d = \text{dattes}$
- $e = \text{eggs}$
Definition: Itemset

An itemset is a set of items that is a subset of $I$.

Example: $\{a, b, c\}$ is an itemset containing 3 items

$\{d, e\}$ is an itemset containing 2 items

- Note: an itemset cannot contain a same item twice.
- An itemset having $k$ items is called a $k$-itemset.
Definition: Sequence

A **discrete sequence** $S$ is a an ordered list of itemsets $S = \langle X_1, X_2, \ldots, X_n \rangle$ where $X_j \subseteq I$ for any $j \in \{1,2\ldots n\}$

**Example 1**: $\langle \{a, b\}, \{c\} \rangle$ is a sequence containing two itemsets.

It means that a customer purchased *apple* and *bread* at the same time and then purchased *cake*.

**Example 2**: $\langle \{a\}, \{a\}, \{c\} \rangle$
Definition: Subsequence \( (\subseteq) \)

Let there be two sequences: 

\[ S_A = \langle A_1, A_2, ..., A_r \rangle \] and \( S_B = \langle B_1, B_2, ..., B_t \rangle \).

The sequence \( S_A \) is a subsequence of \( S_B \) if and only if there exists \( r \) integers \( 1 \leq i_1 < i_2 < \cdots < i_r \leq t \) such that \( A_1 \subseteq B_{i_1}, A_2 \subseteq B_{i_2}, \ldots A_r \subseteq B_{i_r} \).

This is denoted as \( S_A \subseteq S_B \)

Examples:

\[ \langle \{a, c\} \rangle \subseteq \langle \{a, b, c\} \rangle \]
\[ \langle \{a, c\} \rangle \nsubseteq \langle \{a\}, \{c\} \rangle \]
\[ \langle \{a\}, \{c\} \rangle \subseteq \langle \{a, b\}, \{d\}, \{b, c\} \rangle \]
\[ \langle \{a\}, \{c\} \rangle \nsubseteq \langle \{a, c\}, \{d\} \rangle \]
Definition: **Sequence database**

A **sequence database** $D$ is a set of discrete sequences $D = \{S_1, S_2, \ldots, S_m\}$ where each sequence $S_j \in D$ has a unique identifier $j$.

**Example 1:** This is a sequence database with four sequences $D = \{S_1, S_2, S_3, S_4\}$:

<table>
<thead>
<tr>
<th>Sequence database</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle{a, b}, {c}, {a}\rangle$</td>
</tr>
<tr>
<td>$S_2 = \langle{a}, {b}, {c}\rangle$</td>
</tr>
<tr>
<td>$S_3 = \langle{b}, {c}, {d}\rangle$</td>
</tr>
<tr>
<td>$S_4 = \langle{b}, {a, b}, {c}\rangle$</td>
</tr>
</tbody>
</table>
Definition: Support of a sequence

The number of sequences in a sequence database $D$ that contain a sequence $S_A$ is called the support of $S_A$. It is defined as:

$$sup(S_A) = |\{S \mid S \in D \text{ and } S_A \subseteq S\}|$$

Example 1:

<table>
<thead>
<tr>
<th>Sequence database</th>
<th>$sup(\langle{a}\rangle) = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle{a, b}, {c}, {a}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_2 = \langle{a}, {b}, {c}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_3 = \langle{b}, {c}, {d}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_4 = \langle{b}, {a, b}, {c}\rangle$</td>
<td></td>
</tr>
</tbody>
</table>
Definition: Support of a sequence

The number of sequences in a sequence database $D$ that contain a sequence $S_A$ is called the support of $S_A$. It is defined as:

$$sup(S_A) = |\{S \mid S \in D \text{ and } S_A \subseteq S\}|$$

Example 2:

<table>
<thead>
<tr>
<th>Sequence database</th>
<th>$sup(\langle{b}\rangle) = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle{a, b}, {c}, {a}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_2 = \langle{a}, {b}, {c}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_3 = \langle{b}, {c}, {d}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_4 = \langle{b}, {a, b}, {c}\rangle$</td>
<td></td>
</tr>
</tbody>
</table>
**Definition: Support of a sequence**

The number of sequences in a **sequence database** $D$ that contain a sequence $S_A$ is called the support of $S_A$. It is defined as:

$$sup(S_A) = |\{S \mid S \in D \text{ and } S_A \subseteq S\}|$$

**Example 3:**

<table>
<thead>
<tr>
<th>Sequence database</th>
<th>$S_1$ $=$ $\langle{a, b}, {c}, {a}\rangle$</th>
<th>$S_2$ $=$ $\langle{a}, {b}, {c}\rangle$</th>
<th>$S_3$ $=$ $\langle{b}, {c}, {d}\rangle$</th>
<th>$S_4$ $=$ $\langle{b}, {a, b}, {c}\rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$sup(\langle{a}, {b}\rangle) = 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Definition: Support of a sequence

The number of sequences in a sequence database $D$ that contain a sequence $S_A$ is called the support of $S_A$. It is defined as:

$$sup(S_A) = |\{S \mid S \in D \text{ and } S_A \subseteq S\}|$$

Example 4:

<table>
<thead>
<tr>
<th>Sequence database</th>
<th>$sup(\langle{a, b}\rangle) = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle{a, b}, {c}, {a}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_2 = \langle{a}, {b}, {c}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_3 = \langle{b}, {c}, {d}\rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_4 = \langle{b}, {a, b}, {c}\rangle$</td>
<td></td>
</tr>
</tbody>
</table>
Definition: **Sequential pattern mining**

- **Input**: A sequence database $D$ and a minimum support threshold $\text{minsupt} > 0$.
- **Output**: All sequential patterns. A sequential pattern is a sequence $S$ where $\text{sup}(S) \geq \text{minsupt}$. 
Example 1

**INPUT:**

<table>
<thead>
<tr>
<th>Sequence database</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle {a, b}, {c}, {a} \rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_2 = \langle {a, b}, {b}, {c} \rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_3 = \langle {b}, {c}, {d} \rangle$</td>
<td></td>
</tr>
<tr>
<td>$S_4 = \langle {b}, {a, b}, {c} \rangle$</td>
<td></td>
</tr>
</tbody>
</table>

**OUTPUT:**

$\text{minsup} = 3$
Example 1

**INPUT:**

<table>
<thead>
<tr>
<th>Sequence database</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle {a, b}, {c}, {a} \rangle$</td>
</tr>
<tr>
<td>$S_2 = \langle {a, b}, {b}, {c} \rangle$</td>
</tr>
<tr>
<td>$S_3 = \langle {b}, {c}, {d} \rangle$</td>
</tr>
<tr>
<td>$S_4 = \langle {b}, {a, b}, {c} \rangle$</td>
</tr>
</tbody>
</table>

$\text{minsup} = 3$

**OUTPUT:**

**all sequential patterns:**

- $\langle \{a\} \rangle$ support = 3
- $\langle \{b\} \rangle$ support = 4
- $\langle \{c\} \rangle$ support = 4
- $\langle \{a\}, \{c\} \rangle$ support = 3
- $\langle \{a, b\} \rangle$ support = 2
- $\langle \{b\}, \{c\} \rangle$ support = 4
- $\langle \{a, b\}, \{c\} \rangle$ support = 3

What will happen if we change the threshold? →
Example 2

INPUT:

Sequence database

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$\langle{a, b}, {c}, {a}\rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$\langle{a, b}, {b}, {c}\rangle$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$\langle{b}, {c}, {d}\rangle$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$\langle{b}, {a, b}, {c}\rangle$</td>
</tr>
</tbody>
</table>

$\text{minsups} = 4$

OUTPUT:

Observation: If we increase the $\text{minsups}$ threshold, less patterns may be found
Example 2

INPUT:

<table>
<thead>
<tr>
<th>Sequence database</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = \langle {a, b}, {c}, {a} \rangle$</td>
</tr>
<tr>
<td>$S_2 = \langle {a, b}, {b}, {c} \rangle$</td>
</tr>
<tr>
<td>$S_3 = \langle {b}, {c}, {d} \rangle$</td>
</tr>
<tr>
<td>$S_4 = \langle {b}, {a, b}, {c} \rangle$</td>
</tr>
</tbody>
</table>

OUTPUT:

all sequential patterns:
- $\langle \{b\} \rangle$ support = 4
- $\langle \{c\} \rangle$ support = 4
- $\langle \{b\}, \{c\} \rangle$ support = 4

Observation: If we increase the $\text{minsup}$ threshold, less patterns may be found.

$\text{minsup} = 4$
It is a difficult problem!

• **A naïve algorithm** would read the database and count the support (frequency) of all possible patterns.

• **Inefficient** because there can be a very large number of sequential patterns.

• For example:

  \[
  \langle\{a\}\rangle, \langle\{b\}\rangle, \langle\{c\}\rangle, \ldots \]

  \[
  \ldots
  \]

  \[
  \langle\{a, b\}\rangle, \langle\{a, c\}\rangle, \langle\{a, d\}\rangle, \ldots
  \]

  \[
  \ldots
  \]

  \[
  \langle\{a\}, \{a\}\rangle, \langle\{a\}, \{a\}, \{a\}, \langle\{a\}, \{a\}, \{a\}, \{a\}\rangle, \ldots \langle\{a, b\}\{a\}\rangle, \ldots
  \]

  \[
  \langle\{a\}, \{b\}\{a\}\rangle, \ldots
  \]

  \[
  \ldots
  \]

• **An efficient algorithm** must find the frequent sequential patterns, without checking all possibilities.
Some popular algorithms


They all have the same input and output. The difference is performance due to optimizations, search strategies and data structures!

Fast implementations available in the [SPMF library](https://en.wikipedia.org/wiki/SPMF)
A performance comparison

Four benchmark datasets are used

- Kosarak
- BMS
- Leviathan
- Snake
The “Apriori” property

**Property (anti-monotonicity).**

Let be two subsequences $X$ and $Y$. If $X \subseteq Y$, then the support of $Y$ is less than or equal to the support of $X$.

**Example**

**Sequence database**

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$\langle {a, b}, {c}, {a} \rangle$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_2$</td>
<td>$\langle {a, b}, {b}, {c} \rangle$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$\langle {b}, {c}, {d} \rangle$</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$\langle {b}, {a, b}, {c} \rangle$</td>
</tr>
</tbody>
</table>

The support of $\langle \{b\} \rangle$ is 4
The support of $\langle \{b\}, \{c\} \rangle$ is 4
The support of $\langle \{b\}, \{c\}, \{d\} \rangle$ is 1