EXTRACTING MULTI-ITEM SEQUENTIAL PATTERNS BY WAP-TREE BASED APPROACH

by Dilek Önal and Pınar Karagöz

Department of Computer Engineering, Middle East Technical University (METU), Turkey
THE PROBLEM:

EXTRACTING MULTI-ITEM SEQUENTIAL PATTERNS BY WAP-TREE BASED APPROACH
THE PROBLEM:

EXTRACTING MULTI-ITEM SEQUENTIAL PATTERNS BY WAP-TREE BASED APPROACH
PROPOSED APPROACH:

EXTRACTING MULTI-ITEM SEQUENTIAL PATTERNS BY WAP-TREE BASED APPROACH
Sequential Pattern Mining

- is a basic problem in data mining
- is the extraction of frequent sequences in a sequence database
  - Sequence data is common in web usage mining
    - User click patterns are used for many tasks such as recommendation and web site design
SINGLE-ITEM VS MULTI-ITEM SEQUENCES

- Two cases of the problem according to the transaction size.

**Multi-Item / General**

<table>
<thead>
<tr>
<th>Id</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(laptop) (mouse, printer, camera)</td>
</tr>
<tr>
<td>2</td>
<td>(laptop, bag) (camera, speakers)</td>
</tr>
</tbody>
</table>

*Transaction records of electronics shop*

**Single-Item**

<table>
<thead>
<tr>
<th>Id</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(mainpage) (callforpapers) (dates)</td>
</tr>
<tr>
<td>2</td>
<td>(mainpage) (paperlist) (program)</td>
</tr>
</tbody>
</table>

*User sessions from WEBIST web site*
CONTRIBUTIONS AND OUTLINE

- Extension of WAP-Tree to represent multi-item sequence databases
- A new multi-item sequential pattern mining algorithm MULTI-FOF-SP
A tree which encodes single-item sequence databases

<table>
<thead>
<tr>
<th>Id</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>aba</td>
</tr>
<tr>
<td>2</td>
<td>adcdb</td>
</tr>
<tr>
<td>3</td>
<td>beae</td>
</tr>
<tr>
<td>4</td>
<td>ac</td>
</tr>
</tbody>
</table>
CONTRIBUTION 1: MULTI-WAP-TREE

- An extension of WAP-Tree which can encode multi-item databases

- MULTI-WAP-TREE has two types of edges S-Edges and I-Edges differently from WAP-Tree
CONTRIBUTION 1: MULTI-WAP-TREE

<table>
<thead>
<tr>
<th>Id</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(ab)(c)</td>
</tr>
<tr>
<td>2</td>
<td>(a)(b)(c)</td>
</tr>
<tr>
<td>3</td>
<td>(abc)</td>
</tr>
</tbody>
</table>

R:3

a:3

b:2

c:1

b:1

c:1

c:1

I-Edge

S-Edge
CONTRIBUTION 1: MULTI-WAP-TREE

- b:2 encodes (ab)
- b:1 encodes (a)(b)
CONTRIBUTION 2: MULTI-FOF-SP

- Multi-FOF-SP: Multi-FOF-Sibling Principle
- A multi-item sequence mining algorithm
- Applies FOF mining strategy combined with Sibling Principle on MULTI-WAP-Tree
- Follows a hybrid search space traversal strategy in order to apply sibling principle
FOF (FIRST OCCURRENCES FOREST)

- WAP-Tree based single item sequence mining algorithm
- Performs depth first traversal of search space
- Performs support counting by summing counts of FOF nodes in the WAP-Tree.
**FOF: DEPTH FIRST TRAVERSAL**

- Pure depth first traversal on the lexicographic tree
- When a sequence is visited it's support is counted.
- If it's support is above threshold, it is labeled as frequent
- Backtrack whenever an infrequent pattern is found
FOF: SUPPORT COUNTING

- Whenever a sequence in the search tree is visited, FOF nodes for the last symbol of the pattern is found under the current roots of the WAP-Tree.

- Sum of the counts in FOF nodes gives the support of the sequence.

- FOF nodes are the First level Occurrences of a symbol under the current Root nodes.
FOF NODES ON WAP-TREE

‘a’ visited, support = 4

‘aa’ visited, support = 3
**THE SIBLING PRINCIPLE**

- FOF performs pure depth-first search
- The sibling principle provides early pruning

\[ \text{aa is frequent & ab is not frequent } \Rightarrow \text{aab can not frequent} \]
EXPERIMENTS

- Compared execution time and peak memory consumption of MULTI-FOF-SP with PrefixSpan and LAPIN-LCI
- Generated databases with IBM QUEST Data Generator [3]
- Data Generator Parameters:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Meaning</th>
<th>Value Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>Number of Sequences</td>
<td>{200K,800K}</td>
</tr>
<tr>
<td>N</td>
<td>Size of alphabet</td>
<td>{10,500}</td>
</tr>
<tr>
<td>C</td>
<td>Average Length of Sequences</td>
<td>{5,25}</td>
</tr>
<tr>
<td>T</td>
<td>Average Size of Transactions</td>
<td>{3,7}</td>
</tr>
</tbody>
</table>
EXPERIMENTS N=500
EXPERIMENTS $N=500$
EXPERIMENTS N=10
EXPERIMENTS $N=10$
CONCLUSION

- Introduced a new data structure MULTI-WAP-Tree and a new sequential pattern mining algorithm MULTI-FOF-SP
- MULTI-FOF-SP outperforms PrefixSpan and has a performance close to LAPIN-LCI in terms of execution time on dense multi-item databases with small alphabets.
- As a future work, other existing tree based algorithms can be investigated for multi-item sequence mining using the MULTI-WAP-Tree data structure.
REFERENCES


