WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Sequential Patterns

HUMAN CAPITAL
HUMAN – BEST INVESTMENT!
EUROPEAN UNION
EUROPEAN SOCIAL FUND
Project is co-financed by European Union within European Social Fund

WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Sequential Patterns - Informally

- Sequential patterns – patterns occurring frequently in data sequences in which the order of elements is important.
- **Example:** In the case of a set of events, the order is determined by timestamps, while in the case of a document, the order is determined by positions of paragraphs, sentences or words.

2

WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Example: Sequential Patterns

- In the context of market basket data, a sequential pattern is a typical purchase behavior of customers.
- Example dataset D contains 4 data (customer) sequences.
- Purchase sequence <(d)(bf)(a)> occurs for customers: 1 and 4 in D.

Sample dataset D

CId	*Tid*	*Items*
1	10	**cd**
1	15	abc
1	20	a**bf**
1	25	**a**cdf
2	15	abf
2	20	e
3	10	abf
4	10	**d**gh
4	20	**bf**
4	25	**a**gh

3

WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Support of a Sequence

- *Support of a sequence* S is denoted as *sup*(S) and defined as the number of data sequences containing S.
- **Property.** Support of a subsequence S *of a sequence* S' is not less than *sup*(S').
- **Property.** Support of a supersequence S *of a sequence* S' is not greater than *sup*(S').

4

WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Example: Sequence's Support

- *sup*(<(d)(bf)(a)>)=2
- *sup*(<(b)(a)>)=2 ≥ *sup*(<(d)(bf)(a)>)
- *sup*((cd)(bf)(a))=1 ≤ *sup*(<(d)(bf)(a)>)

Dataset D

CId	*Tid*	*Items*
1	10	**cd**
1	15	abc
1	20	a**bf**
1	25	**a**cdf
2	15	abf
2	20	e
3	10	abf
4	10	**d**gh
4	20	**bf**
4	25	**a**gh

5

WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Sequential Patterns - Formally

- Sequence S is defined as *a sequential pattern* (or alternatively, as a *frequent sequence*) if its support is above a threshold *minSup*.

6

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Creation of Candidate Sequences

- Candidate sequences of size *n* are created from pairs of sequential patterns of size *n*-1.

7

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Creation of Candidates Sequences of Size 2

Sequential pattern	Sequential pattern	Candidate sequences for x ≠ y	Candidate sequences for x = y
<(x)>	<(y)>	<(xy)> <(x)(y)> <(y)(x)>	<(x)(x)>

8

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Creation of Candidates of Size Greater than 2

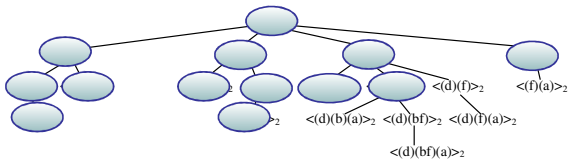
Sequential pattern 1	Sequential pattern 2	Candidate sequences for x ≠ y	Candidate sequences for x = y
<G(P)(x)>	<G(P)(y)>	<G(P)(xy)> <G(P)(x)(y)> <G(P)(y)(x)>	- <G(P)(x)(x)> -
<G(Px)>	<G(Py)>	<G(Pxy)>	-
<G(Px)>	<G(P)(y)>	<G(Px)(y)>	<G(Px)(x)>
<G(P)(x)>	<G(Py)>	<G(Py)(x)>	<G(Py)(y)>

9

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE

- minSup = 1.



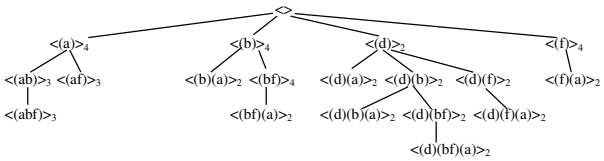
- Infrequent candidate sequences are skipped.

10

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE

- minSup = 1.



- Infrequent candidate sequences are skipped.

11

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Evaluation of Candidate Sequences

- Evaluation of each candidate sequence S is carried out by means of its list of transaction identifiers (shortly, *tidlists*; denoted as t(S)); namely,

sup(S) is equal to the number of candidate sequences registered in t(S).

12

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Tidlists of Sequences of Size 1

- Determined based on given dataset D.

t(<(a)>)		t(<(b)>)		t(<(d)>)		t(<(f)>)	
CId	Tid	CId	Tid	CId	Tid	CId	Tid
1	15	1	15	1	10	1	20
1	20	1	20	1	25	1	25
1	25	2	15	4	10	2	15
2	15	3	10			3	10
3	10	4	20			4	20
4	25						

- *sup*(<(a)>)=4, *sup*(<(b)>)=4, *sup*(<(d)>)=2, *sup*(<(f)>)=4.

13

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Tidlists of Sequences of Size Greater than 1

- Determined based on tidlists of parents.

t(<(d)>)		t(<(f)>)		t(<(df)>)		t(<(d)(f)>)	
CId	Tid	CId	Tid	CId	Tid	CId	Tid
1	10	1	20	1	25	1	20
1	25	1	25			1	25
4	10	2	15			4	20
		3	10				
		4	20				

d and f occur simultaneously in a data sequence

f occurs later than d in a data sequence

14

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Tidlists of Sequences of Size Greater than 1

- Determined based on tidlists of parents.

t(<(d)(b)(a)>)		t(<(d)(bf)>)		t(<(d)(bf)(a)>)	
CId	Tid	CId	Tid	CId	Tid
1	20	1	20	1	25
1	25	4	20	4	25
4	25				

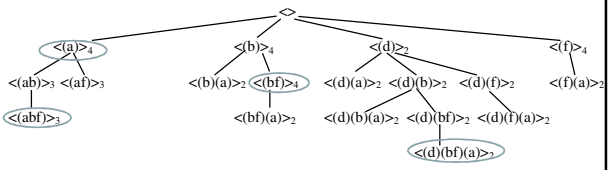
a occurs later than <(d)(bf)> in a data sequence

15

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Closed Sequential Patterns

- minSup = 1.



- *A sequential pattern S is closed* if all its proper supersequences have supports different from (less than) *sup*(*S*).

16

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

GSP: Constraints on Elements of Generalized Sequential Patterns

CId	*Tid*	*Items*
1	10	1, 2
1	25	4, 6
1	45	3
1	50	1, 2
1	65	3
1	90	2, 4
1	95	6
2	…	…
…	…	…

- for each element:
  - $t_k - t_s \leq wS$
- for each 2 neighboring elements:
  - $t_s' - t_k > minGap$
  - $t_k' - t_s \leq maxGap$

- *wS* = 20 (0)
- *minGap* = 19 (0)
- *maxGap* = 50 (∞)

- *sup*(<(1,4)(2,3)(2,6)>) = ?

<(1,4) (2,3) (2,6)>

t_s t_k t_s' t_k' t_s'' t_k''

17

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

GSP: Example…

CId	*Tid*	*Items*
1	10	1, 2
1	25	4, 6
1	45	3
1	50	1, 2
1	65	3
1	90	2, 4
1	95	6
2	…	…
…	…	…

- constraint for each element:
  - $t_k - t_s \leq wS$
- constraints for each 2 neighboring elements:
  - $t_s' - t_k > minGap$
  - $t_k' - t_s \leq maxGap$

- *wS* = 20 (0)
- *minGap* = 19 (0)
- *maxGap* = 50 (∞)

- *sup*(<(1,4)(2,3)(2,6)>) = ?

<(1,4) (2,3) (2,6)>

t_s t_k t_s' t_k' t_s'' t_k''

10 25 50 65 90 95

18

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

GSP: Example

CId	*Tid*	*Items*
1	10	**1**, 2
1	25	**4**, 6
1	45	**3**
1	50	1, **2**
1	65	3
1	90	**2**, 4
1	95	**6**
2	…	…
…	…	…

- constraint for each element:
  - $t_k - t_s \leq wS$
- constraints for each 2 neighboring elements:
  - $t_s' - t_k > minGap$
  - $t_k' - t_s \leq maxGap$

- wS = 20 (0)
- $minGap$ = 19 (0)
- $maxGap$ = 50 (∞)

- sup(<(1,4)(2,3)(2,6)>) = ?

<(1,4) (2,3) (2,6)>

t_s t_k t_s' t_k' t_s'' t_k''

10 25 45 50 90 95

19

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Creating candidates for GSP

CId	*Tid*	*Items*
1	10	1, 2
1	25	4, 6
1	45	3
1	50	1, 2
1	65	3
1	90	2, 4
1	95	6
2	…	…
…	…	…

GSP_3	C_3 after merging	C_3 after pruning
<(1,2) (3)>	<(1,2) (3,4)>	<(1,2) (3,4)>
<(1,2) (4)>	<(1,2) (3) (5)>	
<(1) (3,4)>		
<(1,3) (5)>		
<(2) (3,4)>		
<(2) (3) (5)>		

20

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References

- Marzena Kryszkiewicz, Łukasz Skonieczny: Fast Discovery of Generalized Sequential Patterns, Intelligent Methods and Big Data in Industrial Applications, 155-170, online 2018
- Ramakrishnan Srikant, Rakesh Agrawal: Mining Sequential Patterns: Generalizations and Performance Improvements. EDBT 1996: 3-17
- Jianyong Wang, Jiawei Han, Chun Li: Frequent Closed Sequence Mining without Candidate Maintenance. IEEE Trans. Knowl. Data Eng. 19(8): 1042-1056 (2007)
- Mohammed Javeed Zaki: SPADE: An Efficient Algorithm for Mining Frequent Sequences. Machine Learning 42(1/2): 31-60 (2001)

21