WARSAW UNIVERSITY OF TECHNOLOGY
DEVELOPMENT PROGRAMME

Discovering Sequential Patterns with SPADE

Marzena Kryszkiewicz

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

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

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	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh

3

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

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	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh

5

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

SPADE: Creation of Candidate Sequences

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

7

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

SPADE: Creation of Candidates of Size Greater than 2

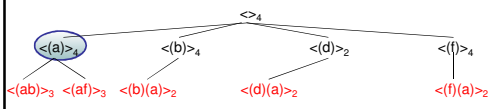
Sequential pattern 1	Sequential pattern 2	Candidate sequences for x ≠ y	Candidate sequences for x = y
<G(P)(x)>	<G(P)(y)>	<G(P)(xy)> <G(P)(x)(y)> <G(P)(y)(x)>	- <G(P)(x)(x)> -
<G(Px)>	<G(Py)>	<G(Pxy)>	-
<G(Px)>	<G(P)(y)>	<G(Px)(y)>	<G(Px)(x)>
<G(P)(x)>	<G(Py)>	<G(Py)(x)>	<G(Py)(y)>

9

Example: Result of SPADE…

- minSup = 1.

<(x)>	<(y)>	<(xy)> <(x)(y)> <(y)(x)>	<(x)(x)>



Candidates: <(a)(a)>1
<(ab)>3, <(a)(b)>1, <(b)(a)>2
<(ad)>1, <(a)(d)>1, <(d)(a)>2
<(af)>3, <(a)(f)>1, <(f)(a)>2

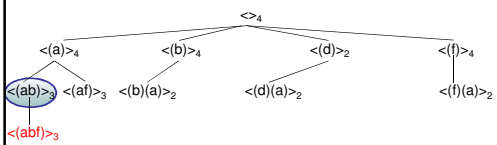
CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh

10

Example: Result of SPADE…

- minSup = 1

<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)>



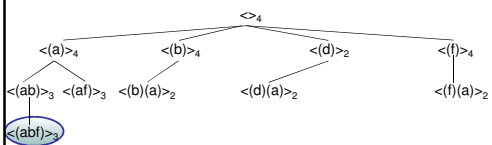
CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh

11

Example: Result of SPADE…

- minSup = 1

<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)>



CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh

12

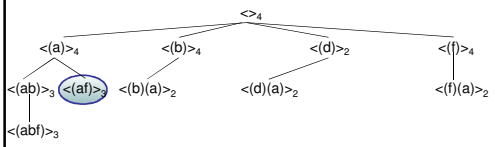
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

- minSup = 1

<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)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



13

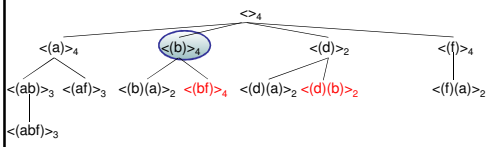
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

- minSup = 1.

<x>	<y>	<(xy)> <(x)(y)> <(y)(x)>	<(x)(x)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



14

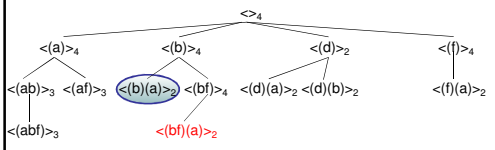
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

minSup = 1.

<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)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



15

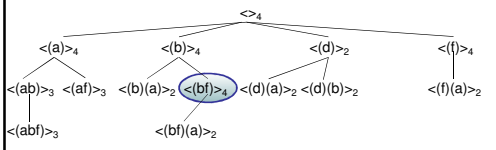
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

minSup = 1.

<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)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



16

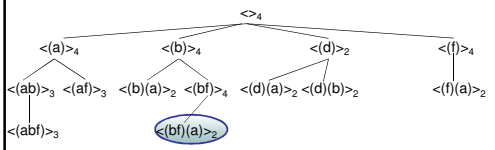
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

minSup = 1.

<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)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



17

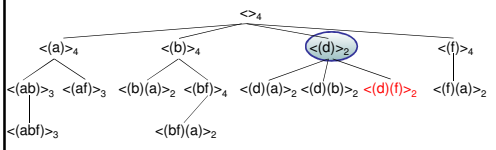
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

- minSup = 1.

<x>	<y>	<(xy)> <(x)(y)> <(y)(x)>	<(x)(x)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



18

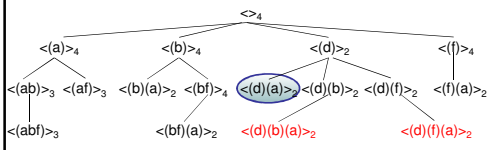
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



19

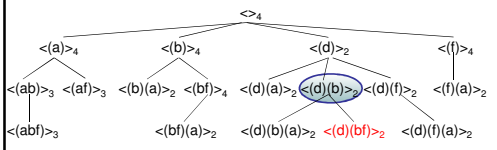
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



20

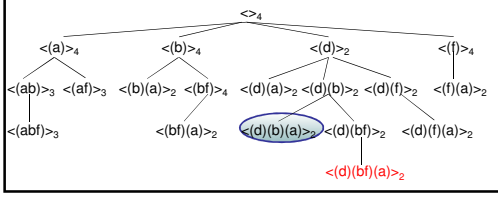
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



21

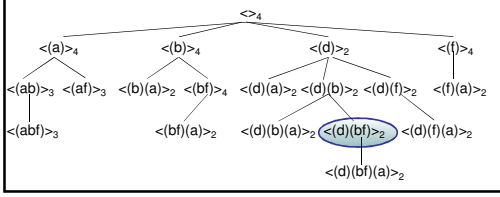
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



22

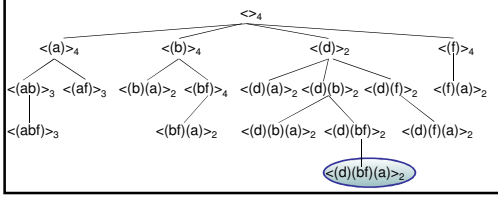
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



23

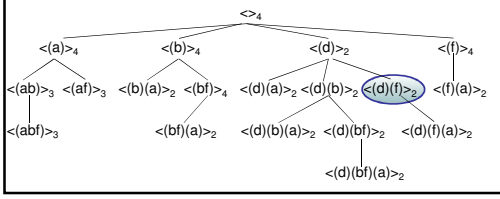
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



24

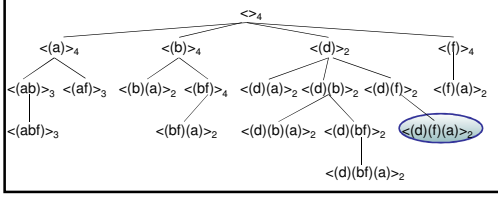
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



25

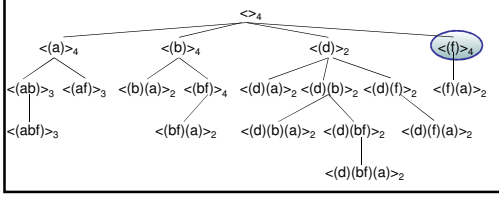
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE…

- minSup = 1.

<(x)>	<(y)>	<(xy)> <(x)(y)> <(y)(x)>	<(x)(x)>

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



26

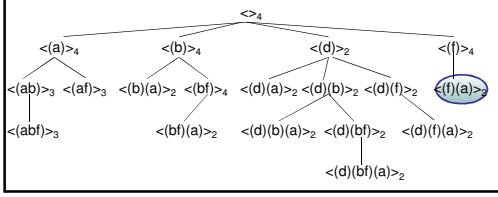
WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Result of SPADE

<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)>

minSup = 1.

CId	Tid	Items
1	10	cd
1	15	abc
1	20	abf
1	25	acdf
2	15	abf
2	20	e
3	10	abf
4	10	dgh
4	20	bf
4	25	agh



27

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).

28

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Tidlists of Sequences of Size 1

- Determined based on given dataset D.

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

t(<(b)>)	
CId	Tid
1	15
1	20
2	15
3	10
4	20

t(<(d)>)	
CId	Tid
1	10
1	25
4	10

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

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

29

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

SPADE: Tidlists of Sequences of Size Greater than 1

- Determined based on tidlists of parents.

t(<(d)>)	
CId	Tid
1	10
1	25
4	10

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

⟹

t(<(df)>)	
CId	Tid
1	25

t(<(d)(f)>)	
CId	Tid
1	20
1	25
4	20

d and f occur simultaneously in a data sequence

f occurs later than d in a data sequence

30

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)>)	
Cld	Tld
1	20
1	25
4	25

t(<(d)(bf)>)	
Cld	Tld
1	20
4	20

⇒

t(<(d)(bf)(a)>)	
Cld	Tld
1	25
4	25

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

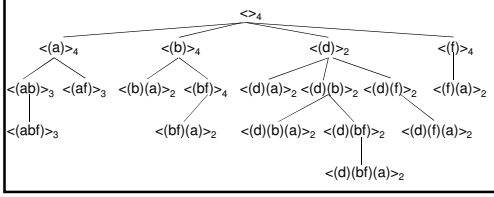
31

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Sequential Rules

- Sequential rules created from <(d)(bf)(a)>:
  - <> → <(d)(bf)(a)> [sup.: 2, conf.: 2/4]
  - <(d)> → <(bf)(a)> [sup.: 2, conf.: 2/2]
  - <(d)(bf)> → <(a)> [sup.: 2, conf.: 2/2]

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

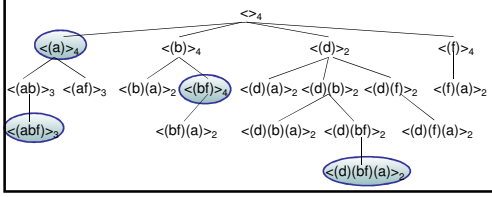


32

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Closed Sequential Patterns

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



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

33

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References

- 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)

34