


### Wzorce sekwencyjne - nieformalnie

- Wzorce sekwencyjne – wzorce występujące często w sekwencjach danych, w których istotny jest porządek występowania elementów.
- Przykład:** W przypadku zbioru zdarzeń, porządek ich zachodzenia jest określony przez czas wystąpienia; w przypadku dokumentu, porządek jest określony przez pozycję akapitów, zdań lub słów.

3




### Przykład: Wzorce sekwencyjne

- W kontekście danych sprzedażowych, wzorec sekwencyjny odzwierciedla typowe dla klientów zachowania zakupowe w czasie.
- Przykładowy zbiór danych D zawiera 4 sekwencje danych (klienckich). *CId* oznacza identyfikator klienta, a *TId* – czas dokonania transakcji.
- Seqwencja  $\langle (d)(bf)(a) \rangle$  składa się z 3 elementów: najwcześniej występującego elementu (*d*), występującego później elementu (*bf*) oraz występującego najpóźniej elementu (*a*).
- Dokonywanie sekwencji zakupów  $\langle (d)(bf)(a) \rangle$  jest charakterystyczne dla klienta 1 i 4, ponieważ  $\langle (d)(bf)(a) \rangle$  występuje w ich sekwencjach danych.

Przykładowy zbiór danych D

	<i>CId</i>	<i>TId</i>	<i>Pozycje</i>
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	


4



### Wsparcie sekwencji

- Wsparcie sekwencji S* jest oznaczane jako  $sup(S)$  i definiowane jako liczba sekwencji danych zawierających S.
- Własność.** Wsparcie podsekwencji S sekwencji S' jest nie mniejsze niż  $sup(S')$ .
- Własność.** Wsparcie nadsekwencji S sekwencji S' jest nie większe niż  $sup(S')$ .

5



### Przykład: Wsparcie sekwencji

- $sup(\langle (d)(bf)(a) \rangle)=2$
- $sup(\langle (b)(a) \rangle)=2 \geq sup(\langle (d)(bf)(a) \rangle)$
- $sup(\langle (cd)(bf)(a) \rangle)=1 \leq sup(\langle (d)(bf)(a) \rangle)$

Zbiór danych D

	<i>CId</i>	<i>TId</i>	<i>Pozycje</i>
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	

6

### Wzorce sekwencyjne - formalnie

- Sekwencja  $S$  jest definiowana jako *wzorzec sekwencyjny* (lub, alternatywnie, jako *sekwencja częsta*), jeżeli jej wsparcie przekracza wartość progową  $minCSup$ .

7

### Odkrywanie wzorców sekwencyjnych przy użyciu algorytmu SPADE

### SPADE: Tworzenie sekwencji kandydujących

- Sekwencje kandydujące o rozmiarze  $n$  są tworzone z par wzorców sekwencyjnych o rozmiarze  $n-1$ .

9

### SPADE: Tworzenie sekwencji kandydujących o rozmiarze 2

Wzorzec sekwencyjny	Wzorzec sekwencyjny	Sekwencje kandydujące dla $x \neq y$	Sekwencje kandydujące dla $x = y$
$\langle x \rangle$	$\langle y \rangle$	$\langle \textcolor{red}{x}y \rangle$ $\langle \textcolor{red}{x} \rangle \langle y \rangle$ $\langle y \rangle \langle \textcolor{red}{x} \rangle$	$\langle \textcolor{red}{x} \rangle \langle \textcolor{red}{x} \rangle$

10

### SPADE: Tworzenie sekwencji kandydujących o rozmiarze $> 2$

Wzorzec sekwencyjny	Wzorzec sekwencyjny	Sekwencje kandydujące dla $x \neq y$	Sekwencje kandydujące dla $x = y$
$\langle G(P)(x) \rangle$	$\langle G(P)(y) \rangle$	$\langle G(P)(xy) \rangle$ $\langle G(P)(x) \rangle \langle y \rangle$ $\langle G(P)(y) \rangle \langle x \rangle$	- $\langle G(P)(x)(x) \rangle$ -
$\langle G(Px) \rangle$	$\langle G(Py) \rangle$	$\langle G(Pxy) \rangle$	-
$\langle G(Px) \rangle$	$\langle G(P)(y) \rangle$	$\langle G(Px) \rangle \langle y \rangle$	$\langle G(Px)(x) \rangle$
$\langle G(P)(x) \rangle$	$\langle G(Py) \rangle$	$\langle G(Py) \rangle \langle x \rangle$	$\langle G(Py)(y) \rangle$

11

### Przykład: Rezultat wykonania SPADE

- $minCSup = 1$ .

- Rzadkie sekwencje kandydujące są pomijane.

12

**SPADE: Wyznaczanie wsparć sekwencji kandydujących**

- Wsparcie sekwencji każdej sekwencji kandydującej  $S$  jest wykonywane za pomocą list identyfikatorów transakcji (*tidlist*), oznaczanych jako  $t(S)$ ; a mianowicie:  
 $sup(S)$  jest równe liczbie sekwencji danych (klienckich) występujących w  $t(S)$ .

13

**SPADE: Tidlisty dla sekwencji o rozmiarze 1**

- Wyznaczone w oparciu o zbiór sekwencji danych (klienckich)  $D$ .

$t(<a>)$		$t(<b>)$		$t(<d>)$		$t(<f>)$	
Cld	Tld	Cld	Tld	Cld	Tld	Cld	Tld
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						

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

14

**SPADE: Tidlisty dla sekwencji o rozmiarze > 1...**

- Wyznaczone w oparciu o tidlisty sekwencji rodzicielskich.

$t(<d>)$		$t(<f>)$		$t(<df>)$		$t(<d(f)>)$	
Cld	Tld	Cld	Tld	Cld	Tld	Cld	Tld
1	10	1	20	1	25	1	20
1	25	1	25			1	25
4	10	2	15			4	20
		3	10				
		4	20				

$\Rightarrow$  d i f występują jednocześnie w sekwencji danych  
 f występuje później niż d w sekwencji danych

15

**SPADE: Tidlisty dla sekwencji o rozmiarze > 1**

- Wyznaczone w oparciu o tidlisty sekwencji rodzicielskich.

$t(<(d)(b)(a)>)$		$t(<(d)(bf)>)$		$t(<(d)(bf)(a)>)$	
Cld	Tld	Cld	Tld	Cld	Tld
1	20	1	20	1	25
1	25	4	20	4	25
4	25				

$\Rightarrow$  a występuje później niż  $<(d)(bf)>$  w sekwencji danych

16

**Przykład: Reguły sekwencyjne**

- $minCSup = 1$ .

```


graph TD
    Root(( )) --- A["<(a)>4"]
    Root --- B["<(b)>4"]
    Root --- D["<(d)>2"]
    Root --- F["<(f)>4"]
    A --- AB["<(ab)>3"]
    A --- AF["<(af)>3"]
    AB --- ABF["<(abf)>3"]
    B --- BA["<(b(a))>2"]
    B --- BF["<(bf)>4"]
    BF --- BFA["<(bf(a))>2"]
    D --- DA["<(d(a))>2"]
    D --- DB["<(d(b))>2"]
    D --- DF["<(d(f))>2"]
    DB --- DBA["<(d(b(a))>2"]
    DB --- DBF["<(d)(bf)>2"]
    DBF --- DBFA["<(d)(bf(a))>2"]
    DF --- DFA["<(d)(f(a))>2"]
    F --- FA["<(f(a))>2"]
  
```

Reguły sekwencyjne utworzone z wzorca sekwencyjnego  $<(d)(bf)(a)>$ :

- $<> \rightarrow <(d)(bf)(a)>$  [support: 2, confidence: 2/4]
- $<(d)> \rightarrow <(bf)(a)>$  [support: 2, confidence: 2/2]
- $<(d)(bf)> \rightarrow <(a)>$  [support: 2, confidence: 2/2]

17


**Uogólnione wzorce sekwencyjne**



### Transakcja → Okno transakcyjne

- Oknem transakcyjnym nazywamy dowolny podzbiór następujących po sobie transakcji w sekwencji danych.
- Rozmiar okna transakcyjnego =  $t_k - t_s$ , gdzie  $t_s$  i  $t_k$  są odpowiednio najmniejszą i największą wartością TId transakcji zawartych w tym oknie transakcyjnym.  $t_s$  jest nazywany czasem startowym tego okna, a  $t_k$  jego czasem końcowym.
- Minimalny odstęp pomiędzy oknami transakcyjnymi T i T' =  $t_s' - t_k$ .
- Maksymalny odstęp pomiędzy oknami transakcyjnymi T i T' =  $t_k' - t_s$ .

19




### Przykład: Okna transakcyjne

CId	TId	Pozycje
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	...	...
...	...	...

- Założenia:
  - T - okno transakcyjne o czasie startowym  $t_s = 10$  i czasie końcowym  $t_k = 25$ .
  - T' - okno transakcyjne o czasie startowym  $t_s' = 50$  i czasie końcowym  $t_k' = 65$ .
- Rozmiar okna T =  $t_k - t_s = 25 - 10 = 15$ .
- Minimalny odstęp pomiędzy oknami T i T' wynosi  $t_s' - t_k = 50 - 45 = 25$ .
- Maksymalny odstęp pomiędzy oknami T i T' wynosi  $t_k' - t_s = 65 - 10 = 55$ .

20




### Uogólnione wsparcie sekwencji

- Sekwencja danych wspiera sekwencję S w sposób uogólniony, jeżeli:
  - każdy element sekwencji S jest zawarty w pewnym oknie transakcyjnym o rozmiarze nie przekraczającym wartości progowej wS:
$$t_k - t_s \leq \text{windowSize},$$
  - okna transakcyjne zawierające bezpośrednio sąsiadujące elementy sekwencji S spełniają następujące warunki na minimalny i maksymalny odstęp:
$$t_s' - t_k > \text{minGap},$$

$$t_k' - t_s \leq \text{maxGap}.$$
- Uogólnione wsparcie sekwencji S jest definiowane jako liczba sekwencji danych wspierających S w sposób uogólniony.

21




### Przykład: uogólnione wspieranie...

CId	TId	Pozycje
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	...	...
...	...	...

- warunek dla każdego elementu sekwencji:
  - $t_k - t_s \leq \text{windowSize}$
- warunki dla bezpośrednio sąsiadujących elementów sekwencji:
  - $t_s' - t_k > \text{minGap}$
  - $t_k' - t_s \leq \text{maxGap}$
- windowSize = 20
- minGap = 19
- maxGap = 50
- $\text{sup}(<(1,4)(2,3)(2,6)>) = ?$

$\begin{matrix} <(1,4) & (2,3) & (2,6)> \\ \longleftrightarrow & \longleftrightarrow & \longleftrightarrow \\ t_s & t_k & t_s' & t_k' & t_s'' & t_k'' \\ 10 & 25 & 45 & 50 & 90 & 95 \end{matrix}$

22




### Przykład: uogólnione wspieranie

CId	TId	Pozycje
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	...	...
...	...	...

- warunek dla każdego elementu sekwencji:
  - $t_k - t_s \leq \text{windowSize}$
- warunki dla bezpośrednio sąsiadujących elementów sekwencji:
  - $t_s' - t_k > \text{minGap}$
  - $t_k' - t_s \leq \text{maxGap}$
- windowSize = 20
- minGap = 19
- maxGap = 50
- $\text{sup}(<(1,4)(2,3)(2,6)>) = ?$

$\begin{matrix} <(1,4) & (2,3) & (2,6)> \\ \longleftrightarrow & \longleftrightarrow & \longleftrightarrow \\ t_s & t_k & t_s' & t_k' & t_s'' & t_k'' \\ 10 & 25 & 45 & 50 & 90 & 95 \end{matrix}$

23



### Uogólniony wzorzec sekwencyjny

- Sekwencja S jest definiowana jako uogólniony wzorzec sekwencyjny, jeżeli jej uogólnione wsparcie przekracza wartość progową minCSup.
- Własność. Jeżeli windowSize = 0, minGap = 0 i maxGap = ∞, to uogólnione wzorce sekwencyjne są zwykłymi wzorcami sekwencyjnymi.

24

**GSP: Tworzenie sekwencji kandydujących**

<i>CId</i>	<i>Tid</i>	<i>Items</i>	$GSP_3$	$C_3$ po złączeniu	$C_3$ po weryfikacji
1	10	1, 2			
1	25	4, 6	<(1,2) (3)>	<(1,2) (3,4)>	<(1,2) (3,4)>
1	45	3	<(1,2) (4)>	<(1,2) (3) (5)>	
1	50	1, 2	<(1) (3,4)>		
1	65	3	<(1,3) (5)>		
1	90	2, 4	<(2) (3,4)>		
1	95	6	<(2) (3) (5)>		
2	...	...			
...	...	...			

KAPITAŁ LUDZKI  
NARODOWA STRATEGIA OPRAWDNI

UNIA EUROPEJSKA  
EUROPEJSKI FUNDUSZ SPOŁECZNY

25

**Literatura**

- Marzena Kryszkiewicz, Łukasz Skonieczny, Fast Discovery of Generalized Sequential Patterns, Intelligent Methods and Big Data in Industrial Applications, 155-170
- Tadeusz Morzy, Eksploracja danych: Metody i algorytmy, Wydawnictwo Naukowe PWN (2013)
- 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)

26

**Ćwiczenia...**

1. Czy sekwencja danych składająca się z 2 transakcji może wspierać sekwencję o 3 elementach?
2. Korzystając z sekwencji danych na slajdzie 4, określ wsparcie sekwencji <(d)(bf)(a)> i sekwencji <(c)(bc)>.
3. Korzystając z tidlist przedstawionych na slajdzie 14, wyznacz tidlisty dla następujących sekwencji: <(f)(d)>, <(bd)>, <(b)(d)> i <(d)(b)>.
4. Zakładając, że skorzystano z algorytmu SPADE, określ sekwencje rodzicielskie dla następujących sekwencji kandydujących: <(abf)>, <(d)(a)(bf)>, <(abf)(d)>?

**Ćwiczenia**

5. Niech  $windowSize = 15$ ,  $minGap = 10$ ,  $maxGap = 52$ . Czy sekwencja danych na slajdzie 19 wspiera sekwencję <(4)(2,3)(2)> w sposób uogólniony?
6. Korzystając z sekwencji danych na slajdzie 4, określ uogólnione wsparcie sekwencji <(c)(bc)>, jeśli:
  - $windowSize = 6$ ,  $minGap = 2$ ,  $maxGap = 20$ ?
  - $windowSize = 6$ ,  $minGap = 6$ ,  $maxGap = 10$ ?
  - $windowSize = 0$ ,  $minGap = 0$ ,  $maxGap = \infty$ ?