



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

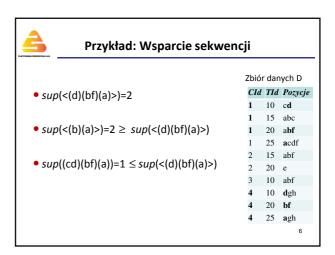


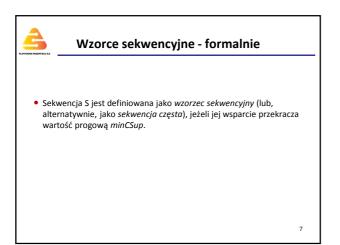


Wsparcie sekwencji

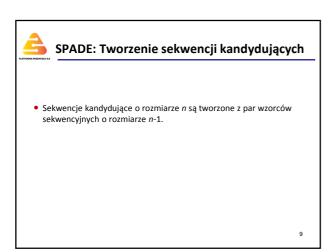
- *Wsparcie sekwencji* S jest oznaczane jako *sup*(S) i definiowane jako liczba sekwencji danych zawierających S.
- Własność. Wsparcie podsekwencji S sekwencji S' jest nie mniejsze niż sup(S').
- Własność. Wsparcie nadsekwencji S sekwencji S' jest nie większe niż sup(S').

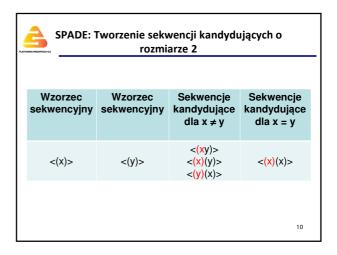
5

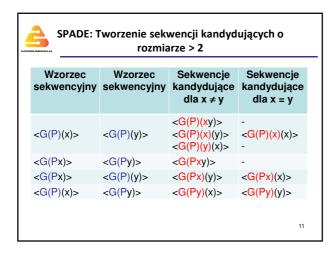


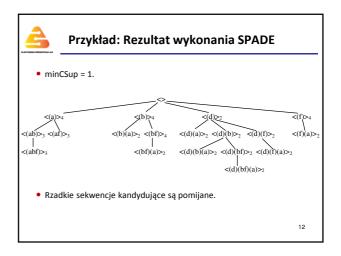


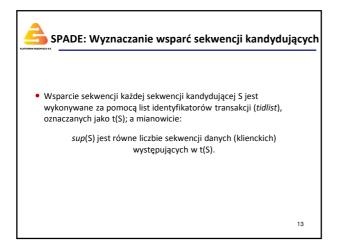


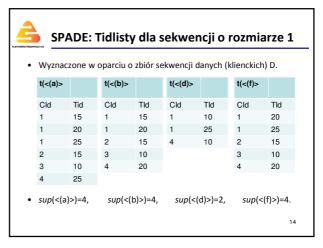


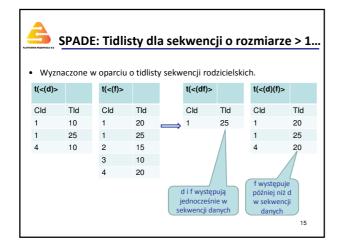


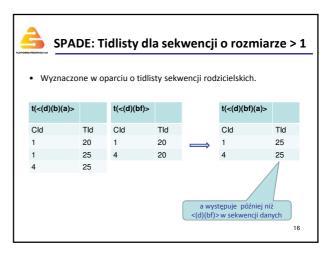


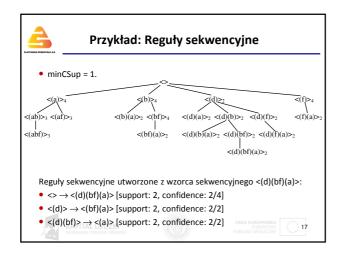


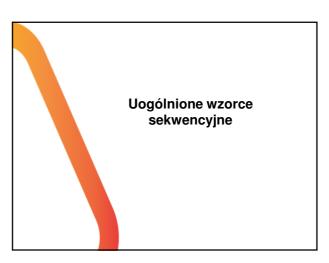








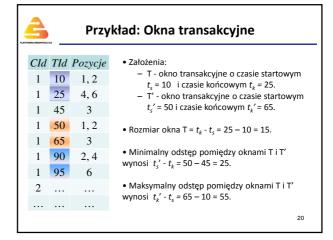






Transakcja → Okno transakcyjne

- 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 Tld 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' = $t_k' t_s$.





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

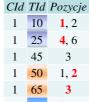
 $t_k - t_s \le windowSize$,

okna transakcyjne zawierające bezpośrednio sąsiadujące elementy sekwencji S spełniają następujące warunki na minimalny i maksymalny odstęp:

$$t_s' - t_k > minGap$$
,
 $t_k' - t_s \le maxGap$.

Uogólnione wsparcie sekwencji S jest definiowane jako liczba sekwencji danych wspierających S w sposób uogólniony.

Przykład: uogólnione wspieranie...



elementów sekwencji: $t_s' - t_k > minGap$ $t_k' - t_s \leq maxGap$

> t_k - t_s ≤ windowSize

• warunek dla każdego elementu sekwencji:

• warunki dla bezpośrednio sąsiadujących

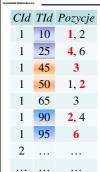
. . .

- 90 **2**, 4 95
- 2
- windowSize = 20
- minGap = 19 • 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

22



Przykład: uogólnione wspieranie



- warunek dla każdego elementu sekwencji:
 - $\succ t_k t_s \le windowSize$
- warunki dla bezpośrednio sąsiadujących elementów sekwencji:

 - $\succ t_s' t_k > minGap$ $\succ t_k' t_s \leq maxGap$
- windowSize = 20
- minGap = 19
- 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

23

21



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





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. <u>IEEE Trans. Knowl. Data</u> <u>Eng. 19</u>(8): 1042-1056 (2007)
- Mohammed Javeed Zaki: SPADE: An Efficient Algorithm for Mining Frequent Sequences. <u>Machine Learning</u> 42(1/2): 31-60 (2001)

26



Ćwiczenia...

- Czy sekwencja danych składająca się z 2 transakcji może wspierać sekwencję o 3 elementach?
- Korzystając z sekwencji danych na slajdzie 4, określ wsparcie sekwencji <(d)(bf)(a)> i sekwencji <(c)(bc)>.
- Korzystając z tidlist przedstawionych na slajdzie 14, wyznacz tidlisty dla następujących sekwencji: <(f)(d)>, <(bd)>, <(b)(d)> i <(d)(b)>.
- 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?
- 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 = ∞ ?