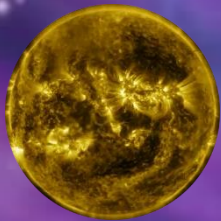




METODY INŻYNIERII WIEDZY

KNOWLEDGE ENGINEERING AND DATA MINING



EKSPLORACJA DANYCH **Ćwiczenia**



Adrian Horzyk

Akademia Górniczo-Hutnicza

***Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej
Katedra Automatyki i Inżynierii Biomedycznej, Laboratorium Biocybernetyki***

30-059 Kraków, al. Mickiewicza 30, paw. C3/205

horzyk@agh.edu.pl, Google: Adrian Horzyk



POLICZ WSPARCIE POSZCZEGÓLNYCH WZORCÓW I POSORTUJ JE WZGLĘDEM CZĘSTOŚCI



1. Wczytaj zbiór transakcji ze strony:
<http://home.agh.edu.pl/~horzyk/lectures/ahdydmiw.php>
2. Oblicz wsparcie (*support*) s – to częstotliwość lub ilość wystąpień wzorca lub zbioru elementów X w analizowanej encji lub transakcji.
3. Wyznacz próg na poziomie $\sigma = 50\%$ i określ, które wzorce są *częste* (*frequent*), tzn. gdy jego wsparcie (*support*) jest nie mniejszy niż ustalony próg σ (*minsup*)

Przykład:

✓	CZĘSTE > 50%	✓	Mleko (40%)
✓	Cukier (80%)	✓	Orzeszki (40%)
✓	Kawa (60%)	✓	Masło (40%)
✓	Jajka (60%)	✓	Chleb (20%)
		✓	Miód (20%)

ID TRANSAKCJI	ELEMENTY TRANSAKCJI
1	kawa, mleko, cukier, orzeszki
2	kawa, cukier, jajka
3	kawa, chleb, cukier, masło
4	orzeszki, cukier, miód, jajka
5	masło, mleko, jajka



OKREŚL REGUŁY ASOCJACYJNE (wsparcie i pewność) DLA TRANSAKCJI Z POPRZEDNIEGO ĆWICZENIA



- ✓ Reguły asocjacyjne (*association rules*) elementów transakcji/wzorców: $X \rightarrow Y (s, c)$.
- ✓ Wsparcie (*support*) s to prawdopodobieństwo, że określona transakcja zawiera $X \cup Y$ liczone względem wszystkich możliwych transakcji.
- ✓ Pewność (*confidence*) c – to prawdopodobieństwo warunkowe, że transakcja zawierająca X zawiera również Y .
- ✓ Eksploracja reguł asocjacyjnych polega na odnalezieniu wszystkich reguł $X \rightarrow Y$ o określonym minimalnym **wsparciu** s oraz o określonej minimalnej **pewności** c : np. $s \geq 50\%$, $c \geq 50\%$.
- ✓ Wielowymiarowe reguły asocjacyjne, np.:
wiek (X , „18-24”) $\wedge$ zawód (X , „student”) $\Rightarrow$ kupuje (X , „cola”)
wiek (X , „18-24”) $\wedge$ kupuje (X , „pop-corn”) $\Rightarrow$ kupuje (X , „cola”)

PRZYKŁADY REGUŁ ASOCJACYJNYCH:

- ✓ Kawa $\rightarrow$ Cukier (80%, 100%)
- ✓ Cukier $\rightarrow$ Kawa (80%, 75%)
- ✓ Cukier $\rightarrow$ Jajka (100%, 50%)
- ✓ Jajka $\rightarrow$ Cukier (100%, 67%)

NIE SĄ NIMI dla $s \geq 50\%$, $c \geq 50\%$:

- ❖ Kawa $\rightarrow$ Jajka (100%, 33%)
- ❖ Jajka $\rightarrow$ Kawa (100%, 33%)

ID TRANSAKCJI	ELEMENTY TRANSAKCJI
1	kawa, mleko, cukier, orzeszki
2	kawa, cukier, jajka
3	kawa, chleb, cukier, masło
4	orzeszki, cukier, miód, jajka
5	masło, mleko, jajka



ZASTOSUJ REGUŁĘ OCZYSZCZANIA APRIORI



Zastosuj regułę oczyszczania Apriori (*pruning principle*) do usunięcia rzadkich podzbiorów i odfiltrowania częstych.

Reguła Apriori:

Każdy podzbiór **zbioru częstego** (*frequent itemset*) jest częsty (frequent).

Wniosek:

Jeśli jakikolwiek podzbiór zbioru S jest **rzadki** (*infrequent*), wtedy S również jest rzadki (*infrequent*).

Powyższy wniosek umożliwia **odfiltrowanie** wszystkich **większych wzorców** (*super-patterns*), które zawierają **rzadkie** (*infrequent*) podzbiory (*itemsubsets*), w celu podniesienia efektywności przeszukiwania wzorców w trakcie ich eksploracji.

Reguła oczyszczania Apriori (*pruning principle*) mówi, iż jeśli istnieje jakikolwiek podzbiór (*itemsubset*), który jest rzadki (*infrequent*), wtedy jego dowolny **zawierający go zbiór** (*superset*) nie powinien być uwzględniany/generowany w procesie eksploracji.



DOKONAJ EKWIWALENTNEJ TRANSFORMACJI KLAS



Ekwiwalentna Transformacja Klas ECLAT (*Equivalence Class Transformation*)

to algorytm przeszukiwania w głąb (depth-first search) wykorzystujący przecięcie zbiorów. Służy do eksploracji częstych wzorców poprzez badanie ich wertykalnego (kolumnowego) formatu:

$$t(B) = \{T_2, T_3\}; t(C) = \{T_1, T_3\} \Rightarrow t(BC) = \{T_3\}$$

$$t(E) = \{T_1, T_2, T_3\} \Rightarrow \text{diffset}(BE, E) = \{T_1\} - \text{zbiór różnic}$$

HORYZONTALNY FORMAT DANYCH	
TRANSAKCJE	ELEMENTY ZBIORU
1	A, C, D, E
2	A, B, E
3	B, C, E

TRANSFORMACJA



WERTYKALNY FORMAT DANYCH	
ELEMENT	LISTA TRANSAKCJI
A	1, 2
B	2, 3
C	1, 3
D	1
E	1, 2, 3

tablica asocjacji

Częsty wzorec to taki podzbiór elementów, który często występuje w transakcjach. Należy więc w tablicy asocjacji odszukać takie elementy, które równocześnie występują w kilku transakcjach, a więc policzyć przecięcie zbiorów transakcji dla poszczególnych elementów, np. dla C i E otrzymamy podzbiór transakcji {1, 3}.