




PLATFORMA PRZEWIDZYWANIA 4.0

## Zbiory częste i reguły asocjacyjne


## Definicje i własności



### Reguły asocjacyjne - nieformalnie

- Załóżmy, że pozycja {ryba} występuje w 5% transakcji sprzedaży, a zbiór pozycji {ryba, białe wino} występuje w 4% transakcji. Ta informacja umożliwia wyprowadzenie *reguły asocjacyjnej*, mówiącej, że w 80% transakcji sprzedaży, klienci którzy kupili *rybę*, kupili także *białe wino*.
- Do wyprowadzania takich reguł potrzebna jest informacja o tym, ile transakcji wspiera stosowne *zbiory pozycji* (ang. *itemsets*).

3




### Wsparcie zbioru pozycji

- Niech  $D$  będzie zbiorem *transakcji*.
- *Wsparcie zbioru pozycji*  $X$ , oznaczane jako  $sup(X)$ , jest liczbą transakcji w  $D$ , które zawierają wszystkie pozycje z  $X$ , to jest:

$$sup(X) = |\{T \in D \mid X \subseteq T\}|.$$

4




### Przykład: Wsparcia zbiorów pozycji

Przykładowy zbiór transakcji  $D$

Id	Transakcja
$T_1$	ABCDEF
$T_2$	ABCDEF
$T_3$	ABCDEH
$T_4$	ABDE
$T_5$	ACDEH
$T_6$	BCE

- $sup(ABC) = 3, sup(EH) = 2$ .

5




### Względne wsparcie zbioru pozycji

- *Względne wsparcie zbioru pozycji*  $X$ , oznaczane jako  $rSup(X)$ , jest stosunkiem transakcji w  $D$ , które zawierają wszystkie pozycje z  $X$ , do liczby wszystkich transakcji w  $D$ :

$$rSup(X) = sup(X) / |D|.$$

- **Uwaga:**  $rSup(X)$  może być traktowane jako oszacowanie prawdopodobieństwa wystąpienia zbioru pozycji  $X$  w  $D$ .


6



### Częste zbiory pozycji

- $X$  jest definiowany jako *częsty zbiór pozycji* (ang. *frequent itemset*), jeżeli
 
$$\text{sup}(X) > \text{minSup},$$
 gdzie  $\text{minSup}$  jest wartością progową wsparcia, zadaną przez użytkownika.
- Podstawowa własność zbiorów pozycji:** Wsparcia nadzbiorów zbioru  $X$  nie są większe od  $\text{sup}(X)$ .

7




### Przykład: Wsparcia zbiorów pozycji

Przykładowy zbiór transakcji D

Id	Transakcja
$T_1$	ABCDEG
$T_2$	ABCDEF
$T_3$	ABCDEH
$T_4$	ABDE
$T_5$	ACDEH
$T_6$	BCE

- $\text{sup}(ABC) = 3, \text{sup}(EH) = 2$ .
- Niech  $\text{minSup} = 2$ . Wtedy:  $ABC$  jest częsty,  $EH$  nie jest częsty (czyli jest rzadki).
- Wsparcia wszystkich nadzbiorów zbioru  $EH$  także nie są większe niż 2, stąd nadzbiory zbioru  $EH$  nie są częste.
- Jednakże wsparcia podzbiorów zbioru  $EH$  mogą być większe niż 2. Zatem może się zdarzyć, że (niektóre) podzbiory zbioru  $EH$  są częste.


8



### Reguły asocjacyjne (ARs)

- Reguła asocjacyjna** jest wyrażeniem wiążącym dwa rozłączne zbiory pozycji:
 
$$X \rightarrow Y,$$
 gdzie  $\emptyset \neq Y \subseteq I \text{ i } X \subseteq I \setminus Y$ .
- O regule  $X \rightarrow Y$  mówi się, że jest *oparta na zbiorze pozycji*  $X \cup Y$ , przy czym:
  - zbiór  $X \cup Y$  jest nazywany *bazą reguły*  $X \rightarrow Y$ ,
  - $X$  – jej *poprzednikiem*,
  - $Y$  – jej *następnikiem*.


9



### Wsparcie reguły asocjacyjnej

- Wsparcie reguły**  $X \rightarrow Y$  jest definiowane jako liczba transakcji zawierających bazę tej reguły, czyli:
 
$$\text{sup}(X \rightarrow Y) = \text{sup}(X \cup Y).$$
- Względne wsparcie reguły**  $X \rightarrow Y$  jest definiowane jako względne wsparcie jej bazy:
 
$$r\text{Sup}(X \rightarrow Y) = r\text{Sup}(X \cup Y).$$

10




### Przykład: Reguły asocjacyjne

Zbiór transakcji D

Id	Transakcja
$T_1$	ABCDEG
$T_2$	ABCDEF
$T_3$	ABCDEH
$T_4$	ABDE
$T_5$	ACDEH
$T_6$	BCE

$\text{sup}(ABC) = 3,$   
 $\text{sup}(A) = 5.$   
 Stąd:  
 $\text{sup}(\{A\} \rightarrow \{BC\}) = \text{sup}(\{ABC\}) = 3.$

11



### Zaufanie reguły asocjacyjnej

- Zaufanie reguły**  $X \rightarrow Y$  jest definiowane jako stosunek liczby transakcji, które zawierają bazę  $X \cup Y$ , do liczby transakcji zawierających jej poprzednik  $X$ :
 
$$\text{conf}(X \rightarrow Y) = \text{sup}(X \rightarrow Y) / \text{sup}(X).$$
- Uwaga:**  $\text{conf}(X \rightarrow Y)$  może być traktowane jako oszacowanie prawdopodobieństwa warunkowego, że  $Y$  występuje w transakcji  $T$  pod warunkiem, że  $X$  występuje w  $T$ .

12

**Przykład: Reguły asocjacyjne**

Zbiór transakcji D

Id	Transakcja
T <sub>1</sub>	ABCDEG
T <sub>2</sub>	ABCDEF
T <sub>3</sub>	ABCDEH
T <sub>4</sub>	ABDE
T <sub>5</sub>	ACDEH
T <sub>6</sub>	BCE

$sup(ABC) = 3,$   
 $sup(A) = 5.$

Stąd:

- $sup(\{A\} \rightarrow \{BC\}) = sup(\{ABC\}) = 3,$
- $conf(\{A\} \rightarrow \{BC\}) = sup(\{ABC\}) / sup(\{A\}) = 3/5.$

13

**Silne reguły asocjacyjne**

- **Silne reguły asocjacyjne (AR)** są definiowane jako te reguły asocjacyjne, których **wsparcie przekracza minSup**, a **zaufanie przekracza minConf**, czyli:

$$AR = \{r \in AR \mid sup(r) > minSup \wedge conf(r) > minConf\},$$

gdzie  $minSup \in [0, |D|)$  i  $minConf \in [0, 1)$ .

- **Własność.** AR to te reguły asocjacyjne, których **bazy są zbiorami częstymi**, a **zaufanie przekracza minConf**.

14

**Odkrywanie silnych reguł asocjacyjnych**

AR jest odkrywane w dwóch etapach:

- Znajdź częste (ang. *frequent*) zbiory pozycji **F** i ich wsparcia w zbiorze transakcji D.
- Generuj **AR** wyłącznie z **F**: Niech  $Z \in F$ ,  $Z \neq \emptyset$  i  $Y \subseteq Z$ . Wtedy, kandydująca reguła  $Z \setminus Y \rightarrow Y$  jest silną regułą asocjacyjną, jeśli:

$$sup(Z) / sup(Z \setminus Y) > minConf.$$

15

**Odkrywanie zbiorów częstych przy użyciu algorytmu Apriori**

**Odkrywanie zbiorów częstych wg algorytmu Apriori**

- W ramach każdej iteracji  $i$ :
- Wyznacz wsparcia kandydujących zbiorów pozycji o długości  $i$ .
- Z kandydatów o długości  $i$ , którzy okazali się częstymi zbiorami, twórz kandydatów o długości  $i + 1$ .

17

**Przykład: Częste 1-zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acth
6	bef
7	h
8	af

- Niech  $minSup = 1$
- Iteracja 1:  
 $C_0 \rightarrow F_0: \emptyset$   
 $C_1 \rightarrow F_1: a_6, b_5, c_4, e_4, f_4, h_3$

18

**Przykład: Częste 2-zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

- Niech minSup = 1
- Po iteracji 1:  
 $F_1: a_6 b_5 c_4 e_4 f_4 h_3$
- Iteracja 2:  
 $C_2 \rightarrow F_2: ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 bh_1$   
 $ce_2 cf_2 ch_2 ef_2 eh_0 fh_1$

Zbiory uznane za rzadkie po wyznaczeniu wsparcia.

19

**Przykład: Częste 3-zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

- Niech minSup = 1
- Po iteracji 2:  
 $F_2: ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2$   
 $ce_2 cf_2 ch_2 ef_2$
- Iteracja 3:  
 $C_3 \rightarrow F_3: abc_3 abe_3 abf_1 abh ace_2 acf_2 ach_2$   
 $ae_1 aeh afh bce_2 bcf_1 bef_2 cef_1 ceh$   
 $cfh$

Zbiory pozycji uznane za rzadkie jako nadzbiory rzadkich zbiorów.

20

**Przykład: Częste 4-zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

- Niech minSup = 1
- Po iteracji 3:  
 $F_3: abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$
- Iteracja 4:  
 $C_4 \rightarrow F_4: abce_2 acefaceh acfh$

21

**Przykład: Częste 5-zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

- Niech minSup = 1
- Po iteracji 4:  
 $F_4: abce_2$
- Iteracja 5:  
 $C_5: -$

22

**Przykład: Znalezione częste zbiory pozycji**

Tid	Pozycje
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

$\emptyset_8$   
 $a_6 b_5 c_4 e_4 f_4 h_3$   
 $ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 ce_2 cf_2 ch_2 ef_2$   
 $abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$   
 $abce_2$

- Uwaga:** Niech  $n$  będzie długością najdłuższego częstego zbioru pozycji. Apriori znajduje wszystkie zbiory częste w  $n$  albo  $n+1$  iteracjach.

23

**Wyznaczanie reguł asocjacyjnych na podstawie zbiorów częstych**

24

**Odkrywanie reguł asocjacyjnych z użyciem algorytmu AprioriRuleGen...**

- Kandydujące reguły budowane są z każdego niepustego zbioru częstego.
- Niech  $Z$  będzie danym niepustym zbiorem częstym. W iteracji  $i$  tworzone są reguły kandydujące postaci:

$$Z \setminus Y \rightarrow Y,$$

gdzie  $Y \subset Z$  i  $|Y| = i$ .

25

**Odkrywanie reguł asocjacyjnych z użyciem algorytmu AprioriRuleGen**

- Własność.** Niech  $r_1: Z \setminus Y \rightarrow Y$  i  $r_2: Z \setminus Y' \rightarrow Y'$ , gdzie  $Y \subset Y'$ , będą regułami asocjacyjnymi.
- $conf(r_1) \geq conf(r_2)$ ,
- Jeśli  $conf(r_1) \leq minConf$ , to  $conf(r_2) \leq minConf$ .
- Aby zredukować liczbę reguł kandydujących, następni o długości  $i + 1$  są budowane wyłącznie z następników o długości  $i$  i silnych reguł asocjacyjnych.

26

**Przykład: Odkrywanie ARs...**

Częste zbiory pozycji ( $minSup = 1$ ):  $\emptyset_8$

$a_6 b_5 c_4 e_4 f_4 h_3$   
 $ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 ce_2 cf_2 ch_2 ef_2$   
 $abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$   
 $abce_2$

Niech  $minConf = 60\%$ ,  $Z = abce$ .

**Iteracja 1:**

- Następniki reguł kandydujących:  $Y_1 = \{a, b, c, e\}$ .
- Kandydujące reguły:
  - $bce \rightarrow a$  [2, 2/2];
  - $ace \rightarrow b$  [2, 2/2];
  - $abe \rightarrow c$  [2, 2/3];
  - $abc \rightarrow e$  [2, 2/3].
- Silne reguły asocjacyjne:
  - $bce \rightarrow a$  [2, 2/2];
  - $ace \rightarrow b$  [2, 2/2];
  - $abe \rightarrow c$  [2, 2/3];
  - $abc \rightarrow e$  [2, 2/3].

27

**Przykład: Odkrywanie ARs...**

Częste zbiory pozycji ( $minSup = 1$ ):  $\emptyset_8$

$a_6 b_5 c_4 e_4 f_4 h_3$   
 $ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 ce_2 cf_2 ch_2 ef_2$   
 $abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$   
 $abce_2$

**Iteracja 2 ( $minConf = 60\%$ ,  $Z = abce$ ):**

- Następniki dla ARs, znalezione w iteracji 1:  $Y_1 = \{a, b, c, e\}$ .
- Następniki reguł kandydujących:  $Y_2 = \{ab, ac, ae, bc, be, ce\}$ .
- Kandydujące reguły:
  - $ce \rightarrow ab$  [2, 2/2];  $ae \rightarrow bc$  [2, 2/3];
  - $be \rightarrow ac$  [2, 2/4];  $ac \rightarrow be$  [2, 2/4];
  - $bc \rightarrow ae$  [2, 2/3];  $ab \rightarrow ce$  [2, 2/4];
- Silne reguły asocjacyjne:
  - $ce \rightarrow ab$  [2, 2/2];
  - $bc \rightarrow ae$  [2, 2/3];
  - $ae \rightarrow bc$  [2, 2/3].

28

**Przykład: Odkrywanie ARs**

Częste zbiory pozycji ( $minSup = 1$ ):  $\emptyset_8$

$a_6 b_5 c_4 e_4 f_4 h_3$   
 $ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 ce_2 cf_2 ch_2 ef_2$   
 $abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$   
 $abce_2$

**Iteracja 3 ( $minConf = 60\%$ ,  $Z = abce$ ):**

- Następniki dla ARs, znalezione w iteracji 2:  $Y_2 = \{ab, ae, bc\}$ .
- Następniki reguł kandydujących:  $Y_3 = \{abe\}$ .
- Kandydujące reguły:
  - $c \rightarrow abe$  [2, 2/4]
- Silne reguły asocjacyjne: Brak ARs w iteracji 3.

29

**Przykład: Odkryte ARs**

Częste zbiory pozycji ( $minSup = 1$ ):  $\emptyset_8$

$a_6 b_5 c_4 e_4 f_4 h_3$   
 $ab_4 ac_4 ae_3 af_3 ah_2 bc_3 be_4 bf_2 ce_2 cf_2 ch_2 ef_2$   
 $abc_3 abe_3 ace_2 acf_2 ach_2 bce_2 bef_2$   
 $abce_2$

Strong association rules ( $minConf = 60\%$ ,  $Z = abce$ ):

- $bce \rightarrow a$  [2, 2/2];
- $ace \rightarrow b$  [2, 2/2];
- $abe \rightarrow c$  [2, 2/3];
- $abc \rightarrow e$  [2, 2/3];
- $ce \rightarrow ab$  [2, 2/2];
- $bc \rightarrow ae$  [2, 2/3];
- $ae \rightarrow bc$  [2, 2/3].

30

## Kwestie implementacyjne



### Ważne operacje w algorytmach Apriori i AprioriRuleGen

- Ważnymi, czasochłonnymi operacjami w *Apriori* są:
  - wyszukiwanie podzbiorów danej transakcji wśród zbiorów kandydujących o ustalonej długości,
  - sprawdzanie czy wszystkie podzbiory kandydata, krótsze od niego o jedną pozycję, są częste.
- Ważną, czasochłonną operacją w *AprioriRuleGen* jest wyszukiwanie podzbiorów częstych o zadanej długości (poprzedników reguł kandydujących) danego zbioru częstego (bazy reguły kandydującej) w celu poznania ich wsparć.
- W powyższych przypadkach wyszukiwane są ustalonej długości podzbiory danego zbioru (odpowiednio: transakcji, kandydata lub poprzednika kandydującej reguły).

32



### Użycie drzewa mieszającego

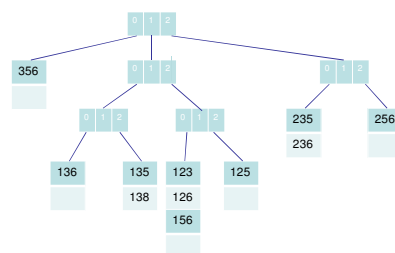
- Drzewo mieszające (ang. *hash tree*) jest używane w celu wydajnego wyszukiwania i pozycyjnych podzbiorów danego zbioru.
- W szczególności, wszystkie *i* pozycyjne zbiory kandydujące są przechowywane w drzewie mieszającym.

33



### Przykład: Tworzenie drzewa mieszającego z kandydującymi 3-zbiorami

zbiór	reprezentacja
abc	123
abe	125
abf	126
ace	135
acf	136
ach	138
aef	156
bce	235
bcf	236
bef	256
cef	356

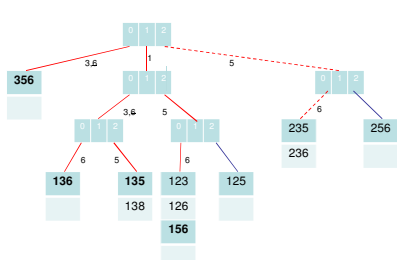


34



### Przykład: Wyszukiwanie podzbiorów w drzewie mieszającym z kandydującymi 3-zbiorami

zbiór	reprezentacja
abc	123
abe	125
abf	126
...	...
transakcja	reprezentacja
...	...
acfh	1356
...	...



- Znalaziono 4 podzbiory transakcji *acfh* (reprezentowanej przez: 1356).

35

## Odkrywanie ARs z danych nie (tylko) transakcyjnych

### Dane transakcyjne + taksonomie

Tid	Items
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

**J**  
Jadalne

**O**  
Owoc

a

b

ananas   banan

**S**  
Słodycze

c

czekolada

**V**  
Witamina

e

wit. e

**M**  
Muzyczne instr.

f

h

fujarka   harfa

37

### Dane transakcyjne z uwzględnieniem taksonomii

Tid	Items
1	abceOSVJ
2	abcefOSVMJ
3	abchOSMJ
4	abeOVJ
5	acfhOSMJ
6	befOVMJ
7	hM
8	afOMJ

**J**  
Jadalne

**O**  
Owoc

a

b

ananas   banan

**S**  
Słodycze

c

czekolada

**V**  
Witamina

e

wit. e

**M**  
Muzyczne instr.

f

h

fujarka   harfa

38

### Dane transakcyjne + pozycje zanegowane...

Tid	Items
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

6 poz. pozycji {abcefh}

6 neg. pozycji {abcefh}

Tid	Items
1	abcefh
2	abcefh
3	abchef
4	abcefh
5	acfhbe
6	befach
7	habcef
8	afbceh

39

### Dane transakcyjne + pozycje zanegowane...

Tid	Items
1	abce
2	abcef
3	abch
4	abe
5	acfh
6	bef
7	h
8	af

6 poz. pozycji {abcefh}

6 neg. pozycji {abcefh}

$sup(\{befh\}) = 2$

$conf(\{bf\} \rightarrow \{eh\}) = 2/3$

$conf(\{bh\} \rightarrow \{ef\}) = 2/4$

Tid	Items
1	abcefh
2	abcefh
3	abchef
4	abcefh
5	acfhbe
6	befach
7	habcef
8	afbceh

40

### Dane transakcyjne + pozycje zanegowane

Tid	Items
1	abcefh
2	abcefh
3	abchef
4	abcefh
5	acfhbe
6	befach
7	habcef
8	afbceh

Tid	Items
1	1 2 3 4 11 12
2	1 2 3 4 5 12
3	1 2 3 6 11 12
4	1 2 4 10 11 12
5	1 3 5 6 8 10
6	2 4 5 7 9 12
7	6 7 8 9 10 11
8	1 5 8 9 10 12

Item	a	b	c	e	f	h	a	b	c	e	f	h
item id	1	2	3	4	5	6	7	8	9	10	11	12

41

### Dane relacyjne → Dane transakcyjne...

Wzrost	Kolor	Ocena
wysoki	zielony	5
niski	czarny	4
niski	zielony	4

→

Tid	Items
1	1 3 6
2	2 4 5
3	2 3 5

Item	(W=wys.)	(W=nis.)	(K=ziel.)	(K=czarny)	(O=4)	(O=5)
item id	1	2	3	4	5	6
atrybut	1	1	2	2	3	3

42

**Dane relacyjne → Dane transakcyjne**

Wzrost	Kolor	Ocena
wysoki	zielony	5
niski	czarny	4
niski	zielony	4

→

Tid	Items
1	1 3 6
2	2 4 5
3	2 3 5

Item	(W=w)	(W=n)	(K=z)	(K=c)	(O=4)	(O=5)
item id	1	2	3	4	5	6
atrybut	1	1	2	2	3	3

$\{2\} \rightarrow \{5\} [2, 2/2]$ . Stąd:  $(W=n) \rightarrow (O=4) [2, 100\%]$ .  
 $\{3\} \rightarrow \{5\} [1, 1/2]$ . Stąd:  $(K=z) \rightarrow (O=4) [1, 50\%]$ .  
 $\{2,3\} \rightarrow \{5\} [1, 1/1]$ . Stąd:  $(W=n) \wedge (K=z) \rightarrow (O=4) [1, 100\%]$ .

**Literatura**

- Rakesh Agrawal, Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases. [VLDB 1994](#): 487-499
- Jiawei Han, Micheline Kamber, Jian Pei: Data Mining: Concept and Techniques, The Morgan Kaufmann Series in Data Management Systems, 2011
- Tadeusz Morzy, Eksploracja danych: Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013

44

**Literatura dodatkowa - odkrywanie zbiorów częstych**

- Ferenc Bodon: A fast APRIORI implementation. FIMI 2003
- Mohammed Javeed Zaki: Scalable Algorithms for Association Mining. IEEE Trans. Knowl. Data Eng. 12(3): 372-390 (2000)
- Frequent Itemset Mining Implementations Repository: <http://fimi.ua.ac.be/src/>

45

**Ćwiczenia...**

- Za pomocą algorytmu Apriori wyznacz zbiory częste na podstawie zbioru danych z tabeli na slajdzie 18. Przyjmij, że wartość progowa wsparcia  $minSup$  wynosi 2.
- Za pomocą algorytmu AprioriRuleGen wyznacz reguły asocjacyjne o bazie  $\{acf\}$  na podstawie informacji o zbiorach częstych przedstawionych na slajdzie 27. Przyjmij, że wartość progowa zaufania  $minConf$  wynosi 50%.
- Korzystając ze slajdu 37 (lub slajdu 38), wyznacz wsparcie i zaufanie hierarchicznych reguł asocjacyjnych:  $\{O\} \rightarrow \{M\}$ ,  $\{M\} \rightarrow \{O\}$ ,  $\{M\} \rightarrow \{cO\}$ .

**Ćwiczenia**

- Korzystając ze slajdu 39, wyznacz wsparcie i zaufanie reguły asocjacyjnej z negacją:  $\{a\} \rightarrow \{ch\}$ .
- Wyznacz zbiory częste i reguły asocjacyjne z tabeli na slajdzie 42. Przyjmij, że wartość progowa wsparcia  $minSup$  wynosi 1, a wartość progowa zaufania  $minConf$  wynosi 50%.