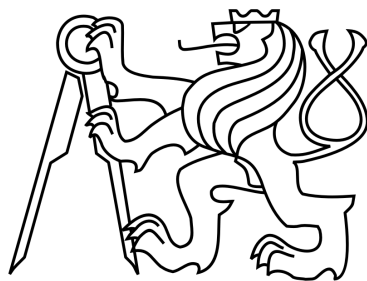




Tvorba asociačních pravidel a hledání častých skupin položek





- ❖ **Asociace a transakce**
- ❖ Časté skupiny položek
- ❖ Apriori vlastnost podmnožin
- ❖ Asociační pravidla
- ❖ Aplikace

† Příklad transakcí



TID	Products
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Asociace



Nechť I je množina položek. Množinu $Z \subseteq I$ označujeme jako **asociaci**, pokud frekvence, s jakou se Z vyskytuje, je výrazně odlišná od toho, co bychom očekávali na základě frekvencí výskytů jednotlivých prvků $X \in Z$.

Příklad:

- ❖ Necht' párky P se vyskytují v nákupním košíku u 10% zákazníků a hořčice H u 4%.
- ❖ Očekávali bychom tedy, že $\{P, H\}$ se bude vyskytovat u 0,4 % zákazníků.
- ❖ Ovšem frekvence $\{P, H\}$ je ve skutečnosti 1 %.
- ❖ Tedy $\{P, H\}$ je asociace.

Databáze transakcí příkladu



TID	Položky transakce	TID	Products
1	A, B, E	1	MILK, BREAD, EGGS
2	B, D	2	BREAD, SUGAR
3	B, C	3	BREAD, CEREAL
4	A, B, D	4	MILK, BREAD, SUGAR
5	A, C	5	MILK, CEREAL
6	B, C	6	BREAD, CEREAL
7	A, C	7	MILK, CEREAL
8	A, B, C, E	8	MILK, BREAD, CEREAL, EGGS
9	A, B, C	9	MILK, BREAD, CEREAL

POLOŽKY:

A = milk

B= bread

C= cereal

D= sugar

E= eggs

Jednotlivé instance = transakce položek

Databáze transakcí příkladů



Uniformní reprezentace: položky převedeny na binární údaje

TID	Položky transakce
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Definice



- ❖ **Položka** (*item*): pár tvaru *atribut = hodnota* nebo jen hodnota
 - ◆ Obvykle se z každého atributu vytvoří několik binárních *údajů* odpovídajících možným hodnotám, např. **Produkt = "A"** se píše jako **"A"**

- ❖ **Množina položek** *I* (*itemset*) : podmnožina všech uvažovaných položek
 - ◆ Příklad: $I = \{A, B, E\}$ (pořadí není podstatné)

- ❖ **Transakce**: (TID, množina položek)
 - ◆ TID je identifikátor transakce (její ID)



- ❖ Asociace a transakce
- ❖ **Časté skupiny položek**
- ❖ Apriori vlastnost podmnožin
- ❖ Asociační pravidla
- ❖ Aplikace

Podpora a časté množiny položek



❖ Podpora (support) množiny položek I

◆ $\text{sup}(I)$ = počet transakcí t ,
které svědčí pro (tj. obsahují) I

❖ Pro náš příklad:

◆ $\text{sup}(\{A,B,E\}) = 2$, $\text{sup}(\{B,C\}) = 4$

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

❖ Množina častých položek I je ta množina,

jejíž podpora je vyšší než předem stanovené minimum **minsup** :

$$\text{sup}(I) \geq \text{minsup}$$

❖ *Je-li všech položek n , pak možných podmnožin je 2^n . Je to problém?*

❖ *Můžeme najít všechny množiny častých položek tak, že postupně prohlédneme všechny podmnožiny položek?*

Jak rostou zajímavé funkce?



Funkce \ N	10	50	100	300	1000
$N * \log N$	33	282	665	2469	9966
N^2	100	2500	10 000	90 000	10^6
N^3	1000	125 000	10^6	$27 * 10^6$	10^9
2^N	1024	10^{14}	10^{29}	10^{89}	10^{300}
$N!$	$3,6 * 10^6$	10^{63}	10^{159}	10^{621}	
N^N	10^{10}	10^{83}	10^{200}	10^{742}	

- Počet protonů ve známém vesmíru $\approx 10^{77}$
- Vzdálenost „Big bang“ - dnešek? $\approx 10^{22} \mu s = 10^{16} s$

Úlohy exp. složitosti nelze řešit hrubou silou!



- ❖ Asociace a transakce
- ❖ Časté skupiny položek
- ❖ **Apriori vlastnost podmnožin**
- ❖ Asociační pravidla
- ❖ Aplikace

Apriori vlastnost podmnožin



Každá podmnožina množiny častých položek je také množinou častých položek!

❖ **Otázka: Proč?**

❖ **Příklad:** Předpokládejme, že $\{A,B\}$ je častý pár položek. Protože každý výskyt svědčící pro $\{A,B\}$ musí obsahovat i položku A, musí být i položka A častá (obdobně pro B)

❖ Podobný argument platí i pro větší množiny položek

❖ *Takřka všechny algoritmy pro konstrukci asocičních pravidel využívají této vlastnosti množin častých položek !*

Hledání častých položek



❖ Postup zdola nahoru: vyhledáme nejdřív všechny *jednoprvkové množiny častých položek* (to je lehké)

❖ **JAK?**

❖ Stačí spočítat frekvence jednotlivých položek

❖ *Princip pro další kroky:*

- ◆ 1-prvkové množiny častých položek lze použít k vytvoření kandidátů na 2-prvkové množiny častých položek. **Jen kandidáti se prověřují** a z nich jsou vybrány SKUTEČNÉ 2-prvkové množiny častých položek !
- ◆ 2-prvkové k vytvoření 3-prvkových množin, ...

Hledání množin častých položek



- ◆ Pokud $\{A, B\}$ je častá množina položek, pak $\{A\}$ i $\{B\}$ musí být rovněž častými množinami položek!
- ◆ Obecně: pokud X je množina častých položek s k -prvky, pak všechny její podmnožiny četnosti $(k-1)$, tj. ty obsahující právě $(k-1)$ položek, musí být rovněž množinami častých položek.

Apriori algoritmus (Agrawal & Srikant)

Najdi množiny častých položek o 1 prvku a pokračuj na četnosti vždy o 1 vyšší pomocí cyklu, který tvoří následující dva kroky:

- ❖ **Krok „spoj“**: Kandidáty na množiny častých položek o k -prvcích lze zkonstruovat jako „rozšíření“ množin častých prvků o $(k-1)$ -prvcích.
- ❖ **Krok „prořez“**: Z množiny kandidátů se pak vyberou jen ty podmnožiny, které mají v databázi transakcí dostatečnou podporu.

† Příklad



❖ *Předpokládejme, že máme 5 množin častých položek o 3 prvcích*

(A B C), (A B D), (A C D), (A C E), (B C D)

❖ *Které množiny o 4 prvcích jsou vhodnými kandidáty ?*

(A B C D) **Otázka: Je OK?**

Odpověď: *Ano*, protože všechny její 3-prvkové podmnožiny jsou časté !

(A C D E) **Otázka: Je OK?**

Odpověď: *Ne*, protože (C D E) není množinou častých položek

❖ *Hledání odpovědí usnadňuje, když položky v množinách uvádíme v lexicografickém uspořádání !*



- ❖ Asociace a transakce
- ❖ Časté skupiny položek
- ❖ Apriori vlastnost podmnožin
- ❖ **Asociační pravidla a jejich tvorba**
- ❖ Aplikace

Asociační pravidla



❖ Asociační pravidlo $R : Množina_pol1 \Rightarrow Množina_pol2$

Význam: "Pokud transakce obsahuje prvky z $Množina_pol1$, pak také obsahuje prvky z $Množina_pol2$ "

❖ $Množina_pol1$ i $Množina_pol2$ jsou disjunktní (mají prázdný průnik) a $Množina_pol2$ je **neprázdná!**

Příklady

$A, B \Rightarrow C$

$A, B \Rightarrow C, E$

$A \Rightarrow B, C$

$A, B \Rightarrow D$

TID	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Od množin častých položek k asociačním pravidlům



❖ *Otázka: Jaká asociační pravidla lze sestavit z množiny častých položek {A,B,E} ?*

- ◆ **A => B, E**
- ◆ **A, B => E**
- ◆ **A, E => B**
- ◆ **B => A, E**
- ◆ **B, E => A**
- ◆ **E => A, B**
- ◆ **__ => A,B,E** (prázdné pravidlo), zapisované také jako **true => A,B,E**

Pravidla klasifikační X asociační



Klasifikační pravidla (~ stromy)

- ❖ Úlohou je určit cílovou třídu (hodnota specifického atributu)
- ❖ Třída musí být předem známa pro všechny zpracovávané příklady
- ❖ **Používaná míra : přesnost klasifikace** na testovací množině dat

Asociační pravidla

- ❖ Neomezují se na jediný cílový atribut
- ❖ Má smysl použít i v případě neklasifikovaných příkladů
- ❖ **Používané míry : podpora, spolehlivost, zdvih**



Asociační pravidlo $R : I \Rightarrow J$: jeho podpora a spolehlivost

Definujeme základní míry významu pravidla R , kterými jsou

❖ **podpora** (*support*): $\text{sup}(R) = \text{sup}(I \cup J)$

odpovídající podpoře množiny všech položek, které se v R vyskytují, tj. $I \cup J$

❖ **spolehlivost** (*confidence*): $\text{conf}(R) = \text{sup}(I \cup J) / \text{sup}(I)$

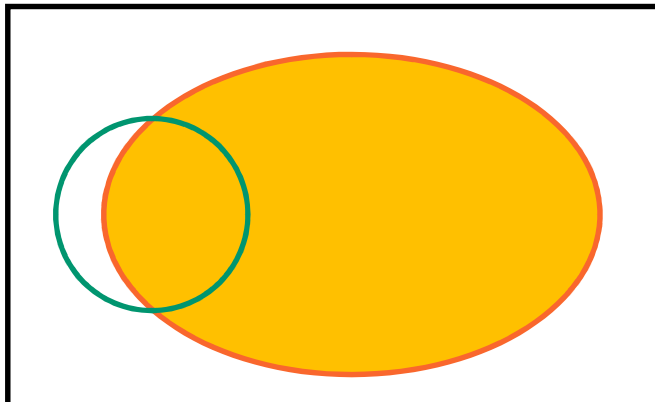
vyjadřující informaci o tom, jaká část z transakcí, ve kterých se vyskytují všechny položky tvořící množinu předpokladů I pravidla R , obsahuje také položky J tvořící závěr tohoto pravidla



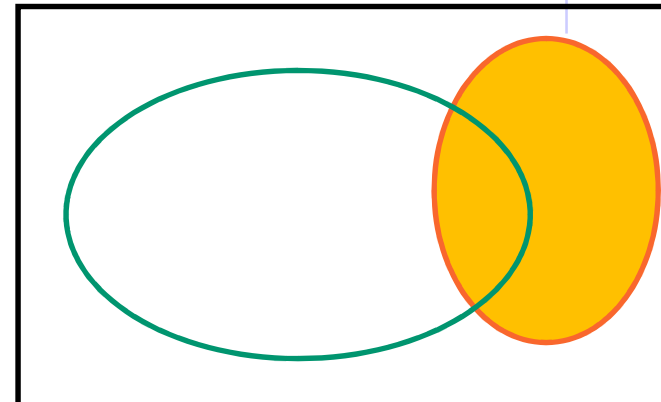
Jaká je spolehlivost pravidla $A \Rightarrow B$?



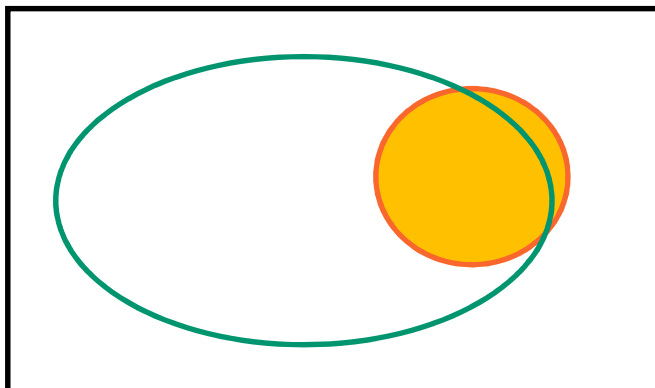
Nechť $A \cup B$ mají dostatečnou podporu: výčet všech možných situací



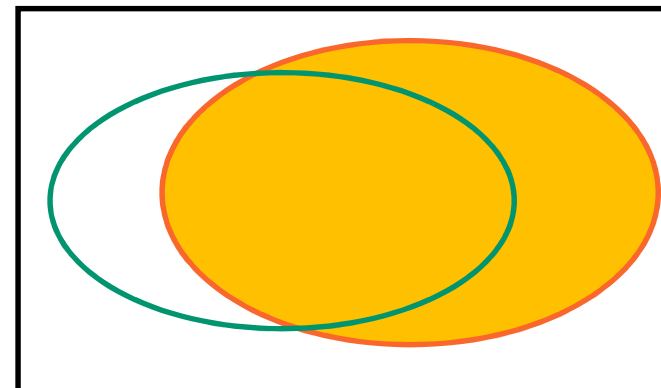
*Conf($A \Rightarrow B$) je vysoká
Conf($B \Rightarrow A$) je nízká*



*Conf($A \Rightarrow B$) je nízká
Conf($B \Rightarrow A$) je nízká*



*Conf($A \Rightarrow B$) je nízká
Conf($B \Rightarrow A$) je vysoká*



*Conf($A \Rightarrow B$) je vysoká
Conf($B \Rightarrow A$) je vysoká*

Nejčastěji používané míry pro $Ant \Rightarrow Suc$



a je mohutnost množiny všech transakcí s položkami $Ant \cup Suc$

	<i>Suc</i>	Non (<i>Suc</i>)	Σ
<i>Ant</i>	a	b	$r = a+b$
Non (<i>Ant</i>)	c	d	$s = c+d$
Σ	$a+c = k$	$l = b+d$	$n = r+s$

spolehlivost(*confidence*) = a/r

podpora(*support*) = a/n

pokrytí(*cover*) = a/k

Zdvih (*lift*, „above average“) = $an/rk = p$

Poměr spolehlivosti uvažovaného pravidla a/r (tedy frekvence *Suc* mezi transakcemi, které splňují *Ant*) a spolehlivosti pravidla $\emptyset \Rightarrow Suc$ (frekv. *Suc* mezi všemi transakcemi, tj. k/n) určuje hodnotu p , která říká kolikrát se díky předpokladu *Ant* spolehlivost zvýší, tj. $a/r = p * (k/n)$

† Příklad:



Úloha: Najděte všechna pravidla s $minsup = 2$ a $minconf = 50%$ pro množinu častých položek $\{A, B, E\}$.

$$A, B \Rightarrow E : conf = 2/4 = 50\%$$

$$A, E \Rightarrow B : conf = 2/2 = 100\%$$

$$B, E \Rightarrow A : conf = 2/2 = 100\%$$

$$E \Rightarrow A, B : conf = 2/2 = 100\%$$

Naopak následující pravidla požadovanou podmínku nesplňují

$$A \Rightarrow B, E : conf = 2/6 = 33\% < 50\%$$

$$B \Rightarrow A, E : conf = 2/7 = 28\% < 50\%$$

$$_ \Rightarrow A, B, E : conf = 2/9 = 22\% < 50\%$$

$$conf(I \Rightarrow J) = \frac{sup(I \cup J)}{sup(I)}$$

TID	List of items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Generování asociačních pravidel



Proces postupující **zdola nahoru** (od pravidel vytvořených z množin častých položek s nejmenší mohutností směrem k množinám s více prvky) rozdělený do 2 fází:

1. **Výstup** := \emptyset , $k := 1$
2. Najděte množiny častých položek dané mohutnosti k , např. algoritmem Apriori. Pokud žádná taková množina není, pak KONEC
3. Pro každou množinu I častých položek nalezněte asociační pravidla platná ve studovaných datech
 - ❖ Pro každou neprázdnou podmnožinu J množiny I
 - ❖ Prověřte platnost asociačního pravidla: $I - J \Rightarrow J$ a platné pravidlo přidejte do množiny **Výstup**
4. $k := k + 1$
5. Jdi na bod 2.

Apriori vlastnost přispívá k efektivitě celého procesu v bodě 2 i 3 ! (díky tomu, že některé mohutnosti byly na datech spočítány už v předchozích krocích výpočtu)

Příklad na generování pravidel



Množina častých položek
z dat „golf/tenis“:

Je toto množina častých
položek?

{Humidity = normal,
Windy = False,
Play = Yes }

Podpora je 4

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



Určete **spolehlivost** pro všechna pravidla ze zvolené množiny častých položek:
 {Humidity = Normal, Windy = False, Play = Yes}

7 potenciálních pravidel (dva typy zápisu) :

**If Humidity = Normal & Play = Yes then
 Windy = False**

**{Humidity = Normal, Play = Yes} =>
 {Windy = False}**

4 / 6

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

If Humidity = Normal and Play = Yes then Windy = False 4 / 6
If Windy = False and Play = Yes then Humidity = Normal 4 / 6
If Humidity = Normal then Windy = False and Play = Yes 4 / 7
If Windy = False then Humidity = Normal and Play = Yes 4 / 8
If Play = Yes then Humidity = Normal and Windy = False 4 / 9
If True then Humidity = Normal and Windy = False and Play = Yes 4 / 14

If Humidity = Normal & Windy = False then Play = Yes 4 / 4

Pravidla pro data „golf/tenis“

❖ **Jaká jsou pravidla mající sup > 1 a conf = 100% ?**

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

	Association rule		Sup.	Conf.
1	Humidity=Normal & Windy=False	⇒Play=Yes	4	100%
2	Temperature=Cool	⇒Humidity=Normal	4	100%
3	Outlook=Overcast	⇒Play=Yes	4	100%
4	Temperature=Cold & Play=Yes	⇒Humidity=Normal	3	100%
...
58	Outlook=Sunny & Temperature=Hot	⇒Humidity=High	2	100%

❖ **Celkem: 3 pravidla se sup=4, 5 s sup=3 a 51 se sup=2**

Filtrování asociačních pravidel



❖ **Problém:** Každý rozsáhlý datový soubor vede ke vzniku obrovského množství různých asociačních pravidel, a to i v případě, že požadujeme rozumnou minimální spolehlivost a podporu. Asociační pravidla, která mají vyšší hodnotu **podpory** a **spolehlivosti**, než je předem zvolená hodnota pro **min_sup** a **min_conf**, se nazývají "**silná (strong)**" pravidla

❖ *Spolehlivost sama o sobě nestačí k výběru opravdu zajímavých pravidel !*

- ◆ Proč? Necht' všechny transakce v souboru obsahují množinu položek **Z**,
Jakou spolehlivost má libovolné pravidlo tvaru

$$\mathbf{I} \Rightarrow \mathbf{Z} \quad ?$$

Filtrování asociačních pravidel



❖ **Problém:** Každý rozsáhlý datový soubor vede ke vzniku obrovského množství různých asociačních pravidel, a to i v případě, že požadujeme rozumnou minimální spolehlivost a podporu. Asociační pravidla, která mají vyšší hodnotu **podpory** a **spolehlivosti**, než je předem zvolená hodnota pro **min_sup** a **min_conf**, se nazývají "**silná (strong)**" pravidla

❖ *Spolehlivost sama o sobě nestačí k výběru opravdu zajímavých pravidel !*

- ◆ Např. pokud všechny transakce v souboru obsahují množinu položek **Z**, pak libovolné pravidlo tvaru

I => Z bude mít spolehlivost 100%.

❖ Pak je třeba využít i další míry pro filtrování pravidel

† Další míry pro pravidlo $Ant \Rightarrow Suc$

podpora (*support*) = a/n

spolehlivost (*confidence*) = a/r

pokrytí (*cover*) = a/k

lift (*zdvih*) = $(a/r)/(k/n) = a*n/(r*k)$

“Zdvih určuje stupeň, o nějž pravidlo zlepšuje přesnost defaultní predikce svého důsledku Suc na výchozích datech”

páka (*leverage*) = $(a - r*k/n)/n = a*(1 - 1/lift)/n$ “Hodnota udávající ‘*procento*’ transakcí, které pravidlo pokývá navíc oproti odhadu, který bychom udělali za předpokladu, že Ant a Suc jsou nezávislé”

přesvědčivost (*conviction*) = $r*I/(b*n) = (I/n)/(b/r)$ “Podobá se zdvihu, ale soustředí se na transakce, které nejsou pokryty Suc . Proto využívá převrácený poměr četností!”

	Suc	Non (Suc)	Σ
Ant	a	b	r = a+b
Non (Ant)	c	d	s = c+d
Σ	k = a+c	l = b+d	n = r+s

ZDVIH asoc. pravidla: interpretace



❖ **Zdvih** (lift) asociačního pravidla $I \Rightarrow J$ je definován takto:

- ◆ Zdvih: $\text{lift}(I \Rightarrow J) = P(J | I) / P(J)$
- ◆ Připomenutí: $P(J) = (\text{podpora } J) / (\text{počet všech transakcí})$
- ◆ Poměr mezi spolehlivostí pravidla $I \Rightarrow J$ a předpokládanou spolehlivostí tak, jak platí na výchozích datech

❖ **Interpretace:**

- ◆ Pokud $\text{lift}(I \Rightarrow J) > 1$, pak jsou I a J pozitivně korelované
pokud $\text{lift}(I \Rightarrow J) < 1$, pak jsou I a J negativně korelované
je-li $\text{lift}(I \Rightarrow J) = 1$, pak jsou I a J nezávislé



- ❖ Transakce
- ❖ Časté skupiny položek
- ❖ Apriori vlastnost podmnožin
- ❖ Asociační pravidla
- ❖ **Aplikace**

Aplikace



- ❖ **Analýza spotřebního koše** může být využita pro
 - ❖ Změnu způsobu prezentace zboží (např. uspořádání zboží v prostoru)
 - ❖ Tvorbu cílené nabídky dalšího zboží (např. v internet. Knihkupectví typu Amazon)
 - ❖ Identifikaci překvapivých jevů jako např. „*Mléko se obvykle kupuje současně s chlebem. Toto však neplatí o sojovém mléce.*“ Tyto jevy mohou být významné nebo mohou vypovídat třeba o organizaci výchozího experimentu, např. „*Pokud dlouhodobý kuřák přestává kouřit, jeho zdravotní stav se zhoršuje.*“
 - ❖ ...
- ❖ Při analýze **dat o pojistných událostech** mohou asoc.pravidla upozornit na jevy, které si zaslouží uvažovat o novém typu služby
 - ❖ if (Car=Porsche & Gender=Male & Age < 20) then
(Risk=high & Insurance = high)
- ❖ ..., např. **identifikace změny** (WSARE – What is Strange About Recent Events), ...

* Doporučená literatura



Jan Rauch, Milan Šimůnek: Dobývání znalostí z dat, LISp-
Miner a GUHA, OECONOMICA, Nakladatelství VŠE, Praha
2014

