

POGOSTI NABORI PREDMETOV

frequent itemsets

IN

POVEZOVALNA PRAVILA

association rules

doc. dr. Matej Guid

Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

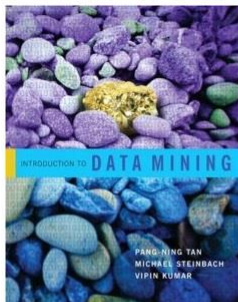
januar 2022

„67% kupcev, ki kupijo vino in sir, kupi tudi grozdje.“

42,7% vseh statistik je izmišljenih ☺



PRIMER ANALIZE NAKUPOVALNIH KOŠARIC



Click to open expanded view



See all 4 customer images
Share your own customer images

Publisher: learn how customers can search inside this book.

Introduction to Data Mining [Hardcover]
Pang-Ning Tan (Author), Michael Steinbach (Author), Vipin Kumar (Author)
★★★★☆ (24 customer reviews)

Buy New
\$81.43 & FREE Shipping. [Details](#)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Want it Monday, Sept. 30? Order within **24 hrs 25 mins** and choose **One-Day Shipping** at checkout. [Details](#)

27 new from \$70.85 27 used from \$71.80

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS
[Learn more](#)

Formats	Amazon Price	New from	Used from
<input checked="" type="checkbox"/> Hardcover	\$81.43	\$70.85	\$71.80
<input type="checkbox"/> Paperback	--	--	\$85.00

Frequently Bought Together

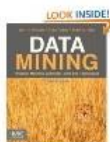


Price for all three: \$166.19

[Add all three to Cart](#) [Add all three to Wish List](#)

Show availability and shipping details

Customers Who Bought This Item Also Bought



Data Mining: Practical Machine Learning ...
> Ian H. Witten
★★★★☆ (32)
Paperback
\$37.80



Regression Analysis by Example (Wiley Series ...
> Samprit Chatterjee
★★★★★ (4)
Hardcover
\$76.61



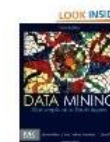
Applied Econometrics Using the SAS System
Vivek Ajmani
★★★★★ (4)
Paperback
\$81.21



Statistical and Machine-Learning ...
> Bruce Ratner
★★★★★ (9)
Hardcover
\$52.36



Logistic Regression Using SAS: Theory ...
> Paul D. Allison
★★★★★ (4)
Paperback
\$48.22



Data Mining: Concepts and Techniques, Third ...
> Jiawei Han
★★★★★ (19)
Hardcover
\$46.96

„A widely used example of **cross selling** on the web with market basket analysis is Amazon.com's use of "customers who bought book A also bought book B" (Wikipedia)

seznam transakcij

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

pogosti nabori stvari (primeri)

{mleko, plenice}

{kruh, mleko, union}

povezovalna pravila (primeri)

{kruh} → {mleko}

{kruh} → {mleko, plenice}

{kruh, union} → {mleko}

vzročnost 

sopojavitve 

POGOSTI NABORI STVARI: DEFINICIJE

primeri, transakcije $T = \{ t_1, t_2, \dots, t_N \}$ *transactions*

stvari, predmeti $I = \{ i_1, i_2, \dots, i_d \}$ *items*

nabor predmetov npr. {kruh, mleko}

nabor enega ali več predmetov *itemset*

k-nabor: vsebuje k predmetov *k-itemset*

širina transakcije število predmetov v transakciji

število podpornih transakcij $\sigma\{\text{kruh, mleko}\} = 3$

$\sigma(X) = | \{ t_i \mid X \subseteq t_i, t_i \in T \} |$ *support count*

podpora nabora predmetov $s\{\text{kruh, mleko}\} = 3/10 = 0.3$

$$s(X) = \frac{\sigma(X)}{N}$$

nabori predmetov s slabo podporo so lahko naključni!

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

⊂

{kruh, mleko} je vsebovano v {kruh, mleko, plenice}

{union, mleko} ni vsebovano v {union, kruh}

⊄

$$s(X) = \frac{\sigma(X)}{N}$$

→ št. podpornih transakcij
→ št. vseh transakcij

$s\{\text{mleko, union}\} = s\{\text{union, mleko}\}$

- vrstni red pojavitev predmetov ni pomemben
- količine predmetov nas ne zanimajo

$s\{\text{mleko, plenice}\} = ?$ 0.6

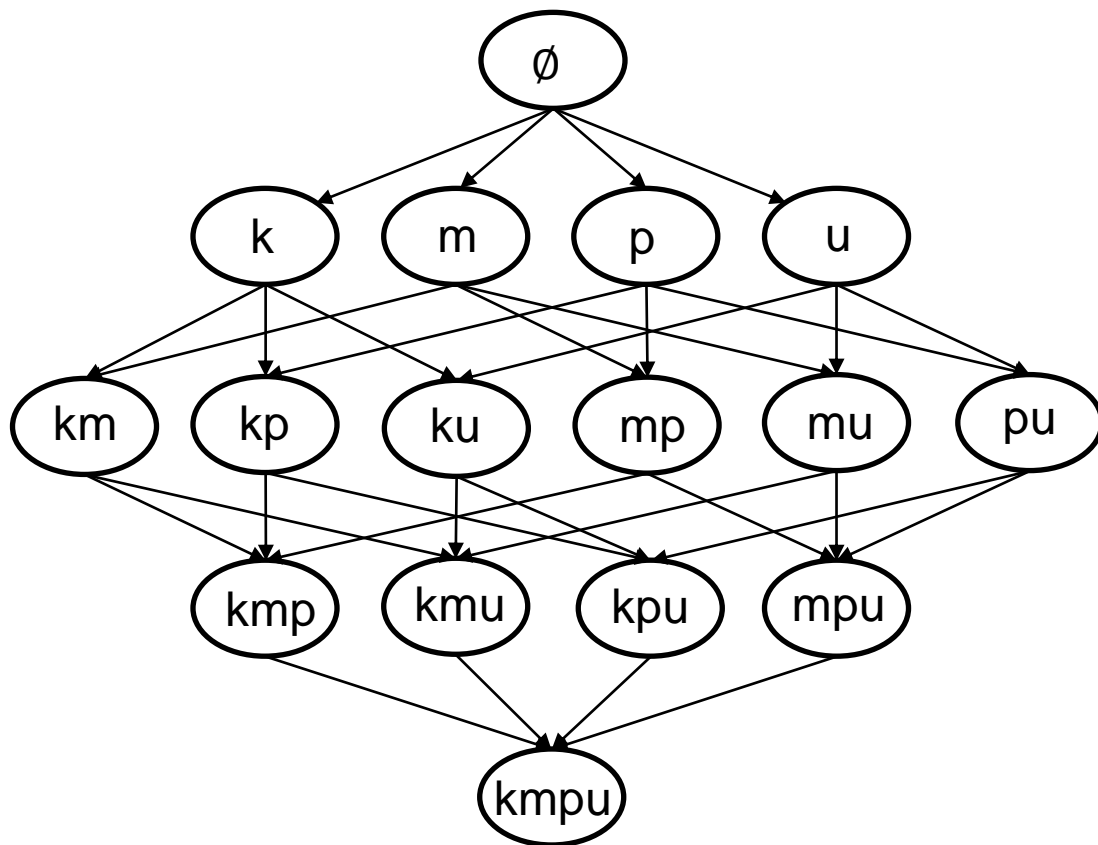
$s\{\text{kruh, plenice}\} = ?$ 0.3

$s\{\text{mleko, union, plenice}\} = ?$ 0.4

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

število možnih naborov je 2^d

$d = 5 \rightarrow 32$
 $d = 6 \rightarrow 64$
 $d = 20 \rightarrow 1048576$



tid	vsebina košarice
1	kruh, mleko
2	kruh, pšenice, unio
3	mleko, pšenice, unio
4	kruh, mleko, pšenice, unio
5	kruh, mleko, pšenice
6	mleko, pšenice
7	pšenice
8	mleko, unio, pšenice
9	pšenice, unio
10	mleko, pšenice, unio

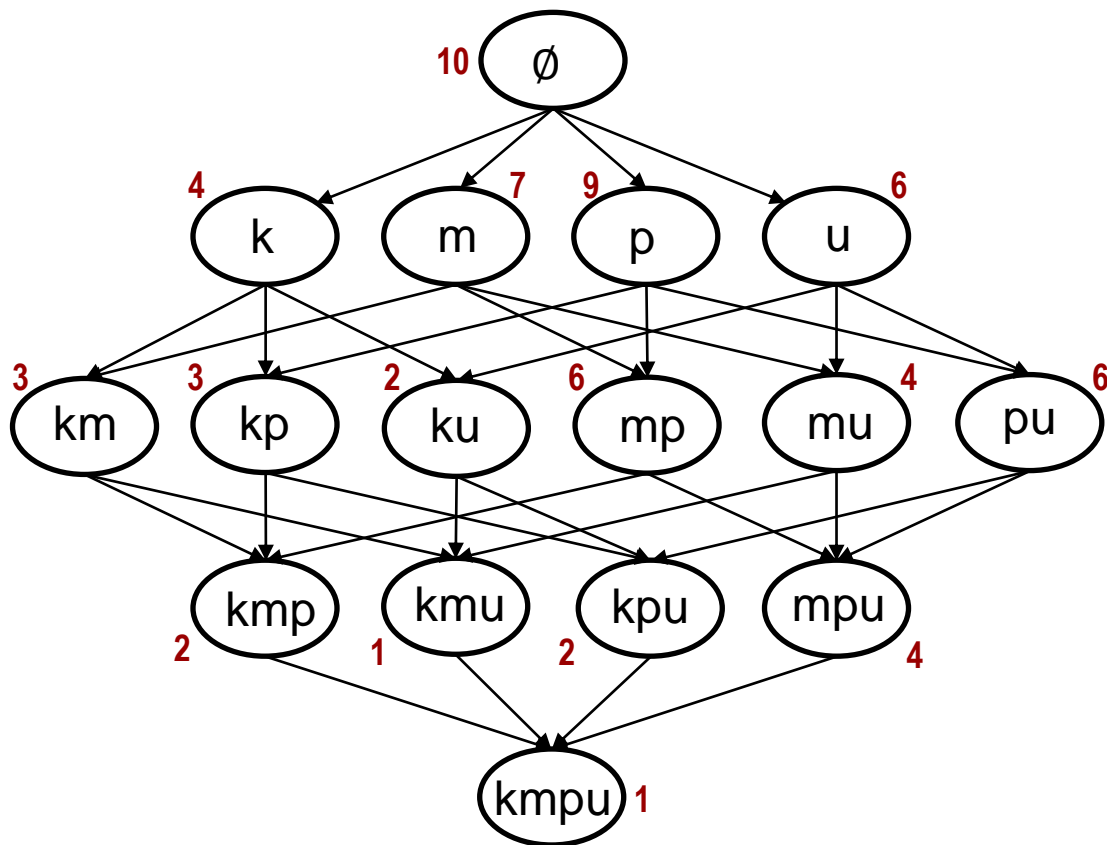
CILJ: kateri so nabori X, pri katerih velja $s(X) \geq \text{min_supp}$?

POGOSTI NABORI

pogosti nabori

$$s(X) \geq \text{min_supp}$$

frequent itemsets



tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

Kako bi izgledal pristop z uporabo grobe sile?

transakcije

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

število transakcij

N

w

širina transakcije

seznam kandidatov

$\{\emptyset\}$
$\{\text{kruh}\}$
$\{\text{mleko}\}$
$\{\text{plenice}\}$
$\{\text{union}\}$
$\{\text{kruh, mleko}\}$
$\{\text{kruh, plenice}\}$
$\{\text{kruh, union}\}$
$\{\text{mleko, plenice}\}$
$\{\text{mleko, union}\}$
$\{\text{plenice, union}\}$
$\{\text{kruh, mleko, plenice}\}$
$\{\text{kruh, mleko, union}\}$
$\{\text{kruh, plenice, union}\}$
$\{\text{mleko, plenice, union}\}$
$\{\text{kruh, mleko, plenice, union}\}$

M

- ❑ primerjava vsake transakcije z vsakim kandidatom
- ❑ $\sim O(NMw)$

problem: $M = 2^d$!!!

Teorem 1

Če je nabor pogost, so pogosti tudi vsi njegovi podnabori.

$$s(X) \geq \text{min_supp} \Rightarrow s(Y) \geq \text{min_supp}, Y \subset X$$

Kako nam to lahko pomaga?

Teorem 2

Če nabor ni pogost, so nepogosti tudi vsi nabori, ki ga vsebujejo.

$$s(X) \not\geq \text{min_supp} \Rightarrow s(Y) \not\geq \text{min_supp}, X \subset Y$$

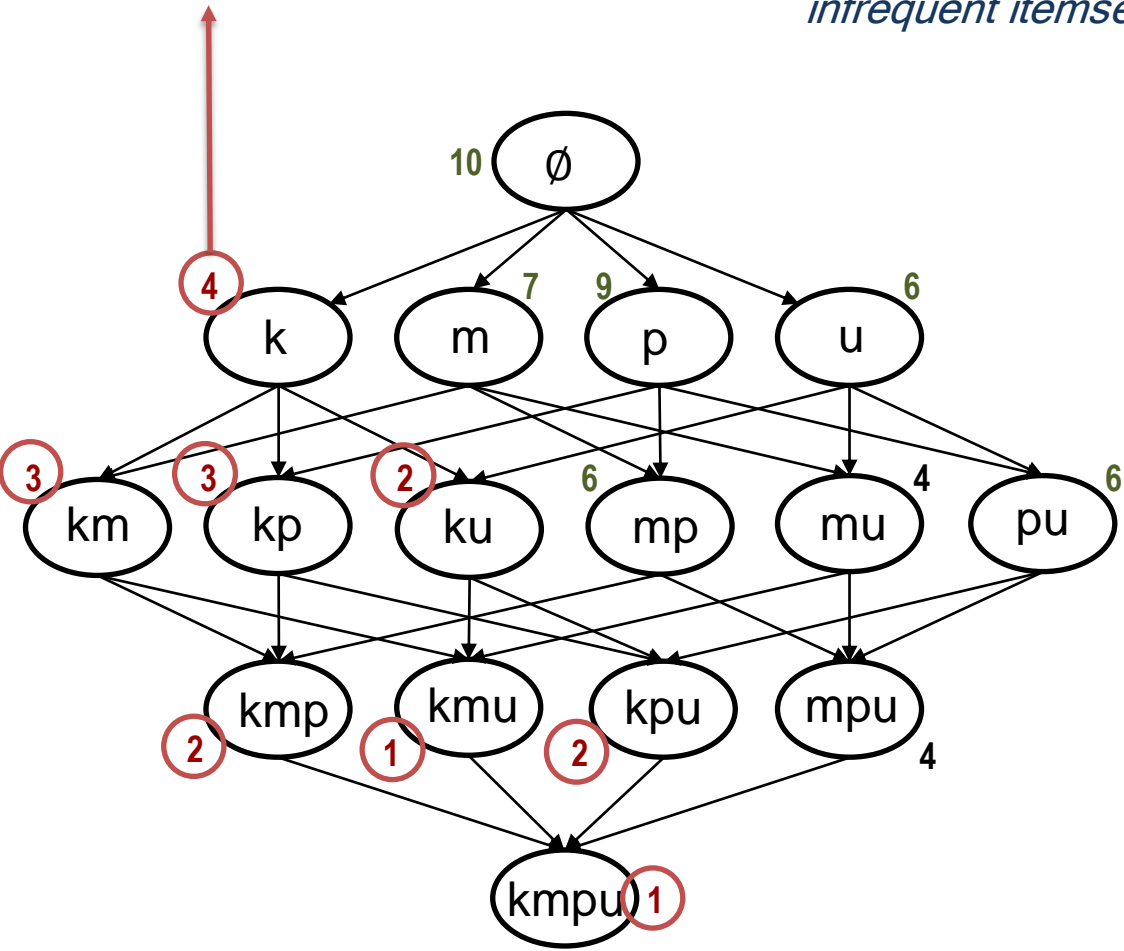
ANTI-MONOTONOST: podpora nabora nikoli ne presega nabora njegovega podnabora.

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

NEPOGOSTI NABORI IN „REZANJE“ MREŽE

nepogost nabor

infrequent itemset

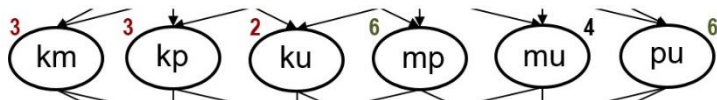


tid	vsebina košarice
1	kruh, mleko
2	kruh, pšenice, unio
3	mleko, pšenice, unio
4	kruh, mleko, pšenice, unio
5	kruh, mleko, pšenice
6	mleko, pšenice
7	pšenice
8	mleko, unio, pšenice
9	pšenice, unio
10	mleko, pšenice, unio

Naj bo $\text{min_supp} = 0.6 \longrightarrow \sigma(X) \geq 6$

□ Zmanjšati število kandidatov (M)

- s pomočjo tehnik rezanja (načelo Apriori!)



tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice

{}
{kruh}
{mleko}
{plenice}
{union}
{kruh, mleko}
{kruh, plenice}
{kruh, union}
{mleko, plenice}
{mleko, union}
{plenice, union}
{kruh, mleko, plenice}
{kruh, mleko, union}
{kruh, plenice, union}
{mleko, plenice, union}
{kruh, mleko, plenice, union}

□ Zmanjšati število transakcij (N)

- transakcije, ki ne vsebujejo pogostih k-naborov, ne morejo vsebovati pogostih (k+1)-naborov in jih zato lahko ignoriramo

□ Zmanjšati število primerjav (NM)

- uporaba učinkovitih podatkovnih struktur za shranjevanje kandidatov ali transakcij
- ni potrebe po primerjavi vsakega kandidata z vsako transakcijo

NAČELO APRIORI: ILUSTRACIJA

predmeti (1-nabori)

PREDMET	ŠT. TRANSAKCIJ
kruh	4
mleko	7
plenice	9
union	6



pari (2-nabori)

NABOR PREDMETOV	ŠT. TRANSAKCIJ
{mleko, plenice}	6
{mleko, union}	4
{plenice, union}	6

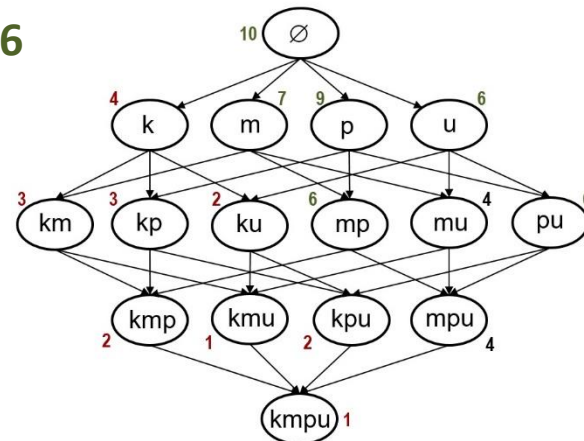
ni potrebno generirati kandidatov, ki vsebujejo predmet **kruh**



trojke (3-nabori)

NABOR PREDMETOV	ŠT. TRANSAKCIJ
{mleko, plenice, union}	4

$$\sigma(X) \geq 6$$



S postopnim generiranjem k-naborov kandidatov bistveno zmanjšamo čas iskanja pogostih naborov.

Dokler je število predmetov v novem naboru večje od 0:

1. Ustvari seznam kandidatov dolžine k (k-nabori)
2. S pregledom baze transakcij za vsak k-nabor preveri njegovo podporo
3. Obdrži pogoste k-nabore in s preostalimi predmeti ustvari (k+1)-nabore

predmeti (1-nabori)

PREDMET	ŠT. TRANSAKCIJ
kruh	4
mleko	7
plenice	9
union	6



pari (2-nabori)

NABOR PREDMETOV	ŠT. TRANSAKCIJ
{mleko, plenice}	6
{mleko, union}	4
{plenice, union}	6



trojke (3-nabori)

NABOR PREDMETOV	ŠT. TRANSAKCIJ

APRIORI: ALGORITEM (I)

$k = 1$

$F_k = \{i \mid i \in I \wedge \sigma(\{i\}) \geq N * \text{min_supp}\}$ # poišči pogoste 1-nabore

ponovi:

$k = k + 1$

$C_k = \text{kandidati}(F_{k-1})$ # ustvari seznam kandidatov

izračunaj podporo naborom v C_k # preveri podporo kandidatov

$F_k = \{c \mid c \in C_k \wedge \sigma(c) \geq N * \text{min_supp}\}$ # obdrži pogoste kandidate

dokler $F_k = 0$

vrni $\bigcup F_k$

APRIORI: ALGORITEM (II)

$k = 1$

$F_k = \{i \mid i \in I \wedge \sigma(\{i\}) \geq N * \text{min_supp}\}$ # poišči pogoste 1-nabore

ponovi:

$k = k + 1$

$C_k = \text{kandidati}(F_{k-1})$ # ustvari seznam kandidatov

izračunaj podporo naborom v C_k # preveri podporo kandidatov

$F_k = \{c \mid c \in C_k \wedge \sigma(c) \geq N * \text{min_supp}\}$ # obdrži pogoste kandidate

dokler $F_k = 0$

vrni $\bigcup F_k$

$\sigma(c) = 0$ za vsak $c \in C_k$

za vsako $t \in T$: # transakcije

$C_t = \{c \mid c \in C_k \wedge c \subset T\}$

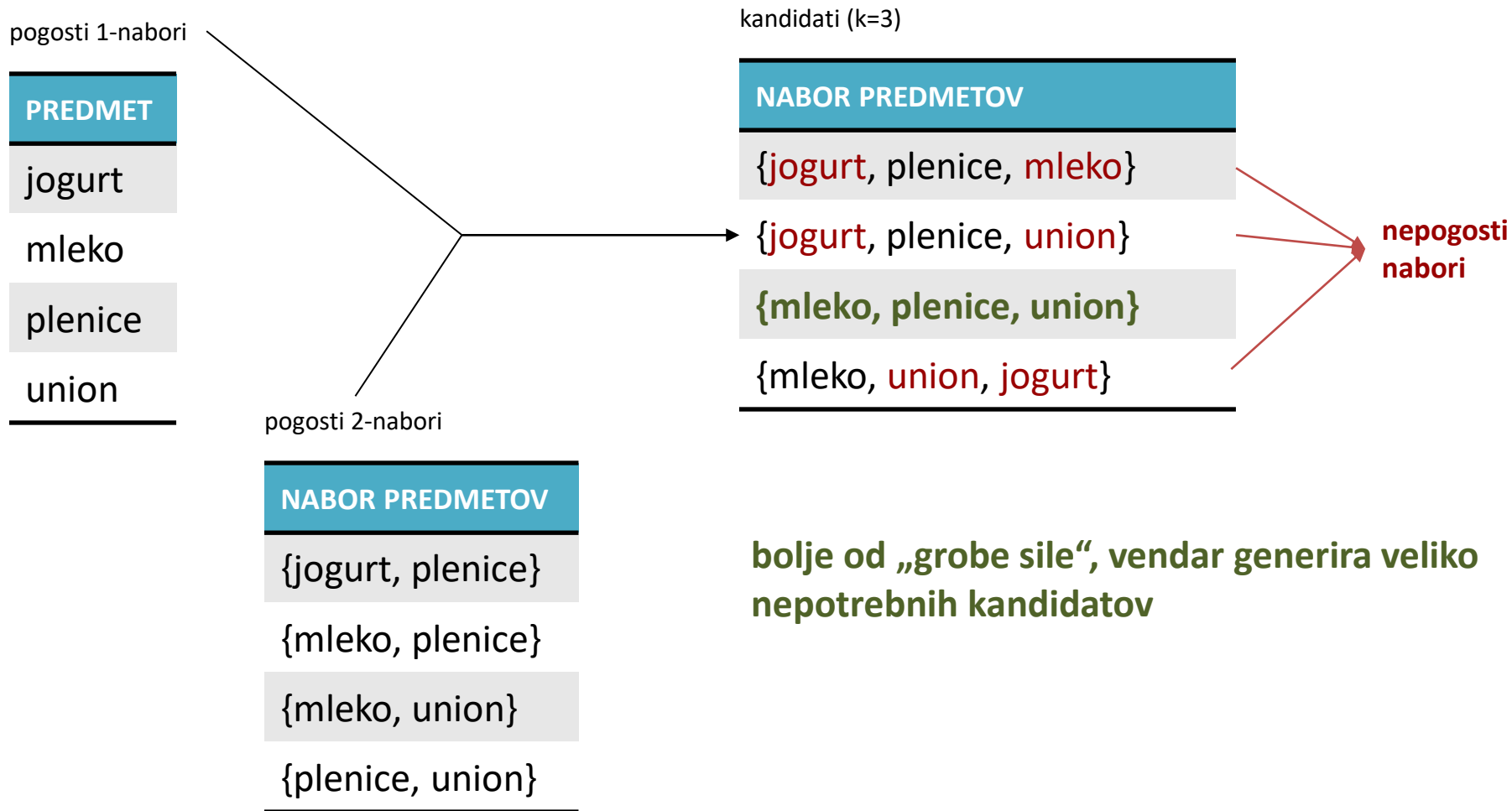
za vsak $c \in C_t$:

$\sigma(c) = \sigma(c) + 1$ # povečaj števec podpore

GENERIRANJE KANDIDATOV (I)

uporaba „grobe sile“: enostavno, a izčrpno: vsaka terka je kandidat za pogosti nabor $O(d \cdot 2^{d-1})$

Metoda $F_{k-1} \times F_1$ razširitev (k-1)-nabora s pogostim predmetom

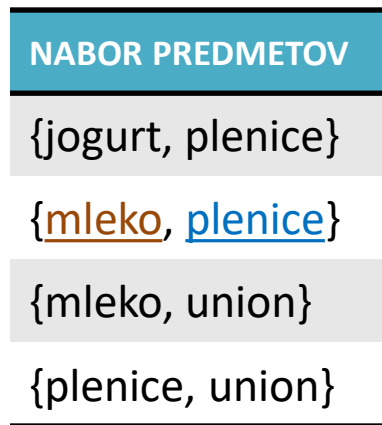


bolje od „grobe sile“, vendar generira veliko nepotrebnih kandidatov

GENERIRANJE KANDIDATOV (II)

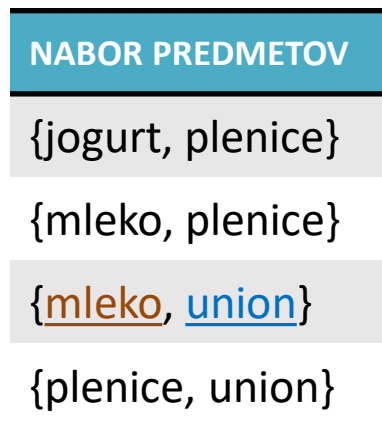
Metoda $F_{k-1} \times F_{k-1}$ združitev para pogostih (k-1)-naborov, če je prvih k-2 predmetov enakih

pogosti 2-nabori

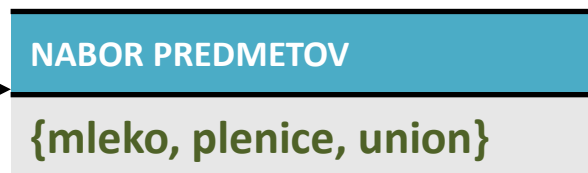


$$a_i = b_i \text{ {for } i = 1, 2 \dots k-2} \text{ and } a_{k-1} \neq b_{k-1}$$

pogosti 2-nabori



kandidati (k=3)



Problem možnega podvajanja rešimo z leksikografsko ureditvijo:

$$\{mleko, plenice\} \cap \{mleko, union\} \rightarrow \{mleko, plenice, union\}$$

$$\{mleko, plenice\} \cap \{union, plenice\} \rightarrow /$$

- ❑ Izbira **minimalne podpore** (*min_supp*)
 - nižji *min_supp* vodi do večjega števila pogostih naborov
 - lahko poveča število kandidatov in največjo dolžino pogostih naborov

- ❑ Dimenzionalnost baze transakcij: **število različnih predmetov**
 - več prostora za shranjevanje podatkov o podpori
 - s povečevanjem števila pogostih naborov se povečuje čas računanja

- ❑ Velikost baze: **število transakcij**
 - algoritem Apriori večkrat „skenira“ bazo, zato se s številom transakcij tipično povečuje čas izvajanja

- ❑ Povprečna **širina transakcije**
 - možno povečanje dolžine pogostih naborov

ZMANJŠEVANJE ŠTEVILA TRANSAKCIJ

Transakcije, ki ne vsebujejo pogostih k-naborov, ne morejo vsebovati pogostih (k+1)-naborov in jih zato lahko ignoriramo!

$$\sigma(X) \geq 4$$

pogosti 1-nabori

PREDMET
češnje
ananas
limone
<u>hruške</u>

vse transakcije vsebujejo vsaj en pogost predmet

pogosti 2-nabori

NABOR PREDMETOV
{ananas, češnje}
<u>{ananas, limone}</u>

tid	vsebina košarice
1	ananas, limone, hruške, grozdje
2	ananas, limone
3	hruške, slive, jabolka
4	češnje, ananas, kivi, mango
5	češnje, ananas
6	hruške, grozdje, avokado
7	češnje, ananas, limone
8	ananas, kivi, mango, hruške, jabolka
9	mango, hruške, slive
10	češnje, ananas, limone

nekatero transakcije ne vsebujejo nobenega pogostega para!

POVEZOVALNA PRAVILA: DEFINICIJE

Povezovalna pravila

association rules

$$X \rightarrow Y \quad \text{npr. } \{\text{union, plenice}\} \rightarrow \{\text{mleko}\}$$

- implikacija, če X potem Y
- presečna množica prazna, $X \cap Y = \emptyset$

podpora

support

$$s(X \rightarrow Y) = \frac{\sigma(XUY)}{N} \quad s = \frac{\sigma(\text{union, plenice, mleko})}{N} = \frac{4}{10} = 0.4$$

kako pogosto je pravilo prisotno v množici transakcij

zaupanje

confidence

$$c(X \rightarrow Y) = \frac{\sigma(XUY)}{\sigma(X)} \quad c = \frac{\sigma(\text{union, plenice, mleko})}{\sigma(\text{union, plenice})} = \frac{4}{6} = 0.67$$

kako pogosto se predmeti v Y pojavljajo v transakcijah, ki vsebujejo X

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

$X \rightarrow Y \equiv \{\text{kruh}\} \rightarrow \{\text{plenice, mleko}\}$

$s(X \rightarrow Y) = ?$ 0.2

$c(X \rightarrow Y) = ?$ 0.5

$X \rightarrow Y \equiv \{\text{union}\} \rightarrow \{\text{plenice}\}$

$s(X \rightarrow Y) = ?$ 0.6

$c(X \rightarrow Y) = ?$ **1.0**

$X \rightarrow Y \equiv \{\text{plenice}\} \rightarrow \{\text{union}\}$

$s(X \rightarrow Y) = ?$ 0.6

$c(X \rightarrow Y) = ?$ **0.67**

- isti nabor
- enaka podpora
- **različno zaupanje!**

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

POVEZOVALNA PRAVILA: ISKANJE PRAVIL

Podana je množica transakcij. Poiščite pravila, za katera velja:

- podpora $\geq \text{min_supp}$
- zaupanje $\geq \text{min_conf}$

korak 1: iskanje pogostih naborov  **računsko zahtevno!**

poišči vse nabore, za katere velja podpora $\geq \text{min_supp}$

korak 2: generiranje pravil iz pogostih naborov

- iz pogostih naborov generiramo pravila z visokim zaupanjem
- vsako pravilo razdeli pogost nabor na dva dela (nabor $\{X \cup Y\}$ v pravilo $X \rightarrow Y$)

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

POVEZOVALNA PRAVILA: MOŽNA PRAVILA

Pravila, kjer nastopajo vsi elementi {k,m,p,u}: $R = 2^d - 2$

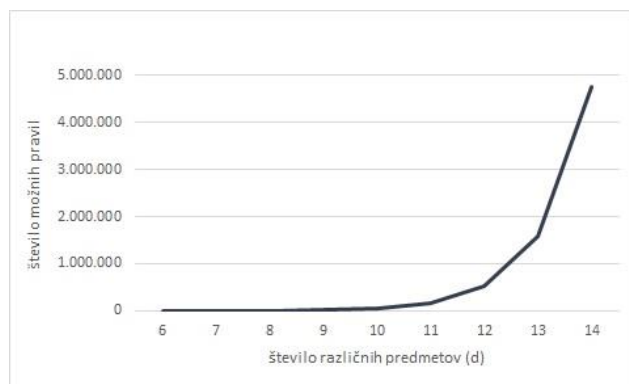
kmp \rightarrow d	kmd \rightarrow p	kpd \rightarrow m	mpd \rightarrow k
k \rightarrow mpd	m \rightarrow kpd	p \rightarrow kmd	d \rightarrow kmp
km \rightarrow pd	kp \rightarrow md	kd \rightarrow mp	mp \rightarrow kd
md \rightarrow kp	pd \rightarrow km	$\emptyset \rightarrow$ kmpu	kmpu \rightarrow \emptyset

število vseh možnih pravil

$$R = 3^d - 2^{d+1} + 1$$

d = 6 \longrightarrow 602 pravil

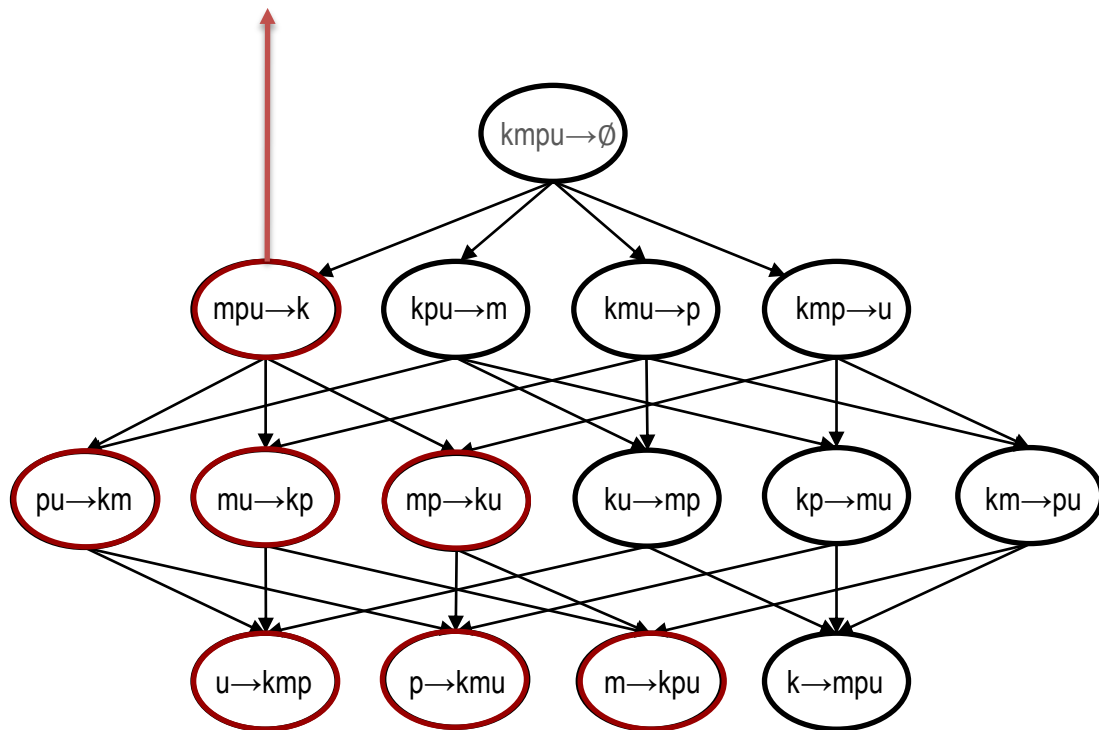
d = 14 \longrightarrow 4.750.202 pravil



tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

GENERIRANJE PRAVIL IN APRIORI ALGORITEM

pravilo z nizkim zaupanjem



ANTI-MONOTONOST velja za pravila iz istega nabora

$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

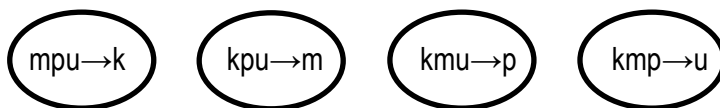
GENERIRANJE PRAVIL: POSTOPEK

$R = \{\}$ seznam pravil

Za vsak pogost nabor f_k , ki ima vsaj dva predmeta:

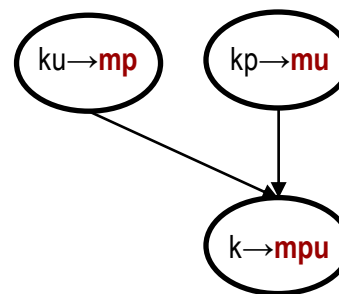
$k > 1$ npr. $f_k = \{kmpu\}$

- ustvari seznam kandidatov (pravil) $H_1 = \{i \mid i \in f_k\}$



pravila dolžine 1 na desni strani

- preveri zaupanje kandidatov: **izbriši** kandidate z nizkim zaupanjem
- ustvari nov seznam kandidatov H_2
- preveri zaupanje kandidatov: **izbriši** kandidate z nizkim zaupanjem
- ustvari nov seznam kandidatov H_3
- ...



novе kandidate se tvori s pomočjo združevanja pravil: unija elementov na desni strani pravila

Subjektivno

- s pomočjo predznanja: cene, hierarhije konceptov, poznavanje strukture pravil...
- vizualizacija

Tipično je veliko povezovalnih pravil nezanimivih ali redundantnih.

Objektivno

- različni tipi kontingenčnih tabel
- **interes (ang. INTEREST), dvig (ang. LIFT)**

Primeri subjektivnega vrednotenja pravil:

- visoka podpora, visoko zaupanje („preveč očitna“)
- zmerna podpora, nizko zaupanje („nezanesljiva“)
- nizka podpora, nizko zaupanje („očitno nezanimiva“)
- **nizka podpora, visoko zaupanje („potencialno zelo zanimivo“)**
- npr. mleko → kruh
- npr. mleko → tuna
- npr. bučno olje → detergent
- npr. vodka → kaviar

VREDNOTENJE Z METODO DVIG (ANGL. LIFT)

	kava	kava	
čaj	150	50	200
čaj	650	150	800
	800	200	1000

čaj → kava „kdor pije čaj, pije tudi kavo“

$$s(\text{čaj} \rightarrow \text{kava}) = 0.15$$

$$c(\text{čaj} \rightarrow \text{kava}) = 150/200 = 0.75$$

Ampak: kavo pije 80% anketirancev!

$$\text{Interest} = |c(X \rightarrow Y) - s(Y)| \quad \text{Interest} = |0.75 - 0.8| = 0.05 \rightarrow \text{nezanimivo!}$$

zanesljivost glede na izhodiščno verjetnost Y

$$\{\text{union}\} \rightarrow \{\text{kruh}\} \quad \text{Interest} = 0.35$$

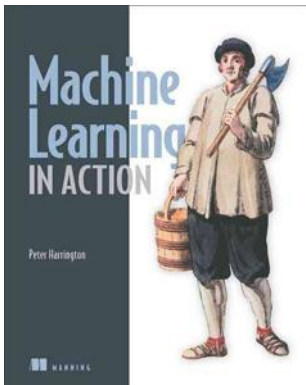
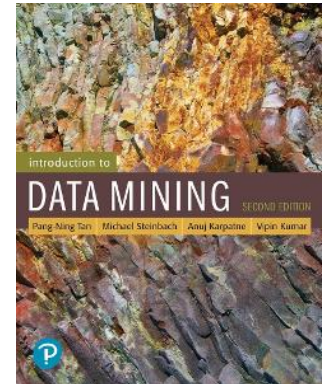
$$\{\text{union}\} \rightarrow \{\text{plenice}\} \quad \text{Interest} = 0.07$$

tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

- Tan P.-N., Steinbach M. in Kumar V. **Introduction to Data Mining**, Pearson Addison Wesley, 2006.

<http://www-users.cs.umn.edu/~kumar/dmbook/>

Association Analysis: Basic Concepts and Algorithms (šesto poglavje)



Implementacija v programskem jeziku python

- Harrington, P. *Machine Learning in Action*. Manning Publications Co., 2012.

Association analysis with the Apriori algorithm (enajsto poglavje)