POGOSTI NABORI PREDMETOV

frequent itemsets

IN

POVEZOVALNA PRAVILA

association rules

doc. dr. Matej Guid

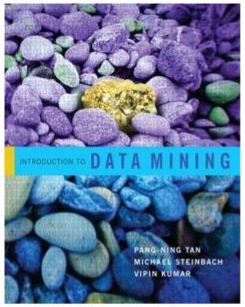
Fakulteta za računalništvo in informatiko
Univerza v Ljubljani

januar 2024

„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.

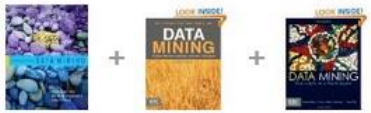
Want it Monday, Sept. 30? Order within **24 hrs 25 mins** and choose **One-Day Shipping** at checkout. Details

27 new from $70.85 **27 used** from $71.80

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS
Learn more amazonstudent

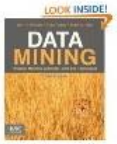
Formats	Amazon Price	New from	Used from
Hardcover	$81.43	$70.85	$71.80
Paperback	--	--	$85.00

Frequently Bought Together



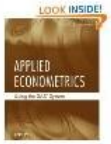
Price for all three: $166.19
Add all three to Cart | Add all three to Wish List
Show availability and shipping details

Customers Who Bought This Item Also Bought



Data Mining: Practical Machine Learning ...
› Ian H. Witten
(32)
Paperback
$37.80

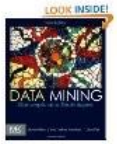
Regression Analysis by Example (Wiley Series ...
› Samprit Chatterjee
(4)
Hardcover
$76.61



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



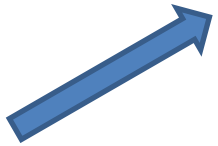
Data Mining: Concepts and Techniques, Third ...
› Jiawei Han
(19)
Hardcover
$46.96

*„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)

NAKUPOVALNE KOŠARICE

seznam transakcij

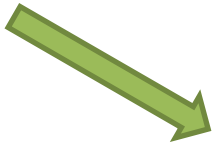
tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union



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 $T = \{ t_1, t_2, \ldots, t_N \}$ *transactions*

stvari, predmeti $I = \{ i_1, i_2, \ldots, i_d \}$ *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 $s\{\text{kruh, mleko}\} = 3/10 = 0.3$ *support*

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

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

Nabori in podnabori

$\subset$

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

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

$\not\subset$

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

$\sigma(X)$ → število podpornih transakcij

N → število vseh 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

s{mleko, union} = s{union, mleko}

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

s{mleko, plenice} = ? 0.6

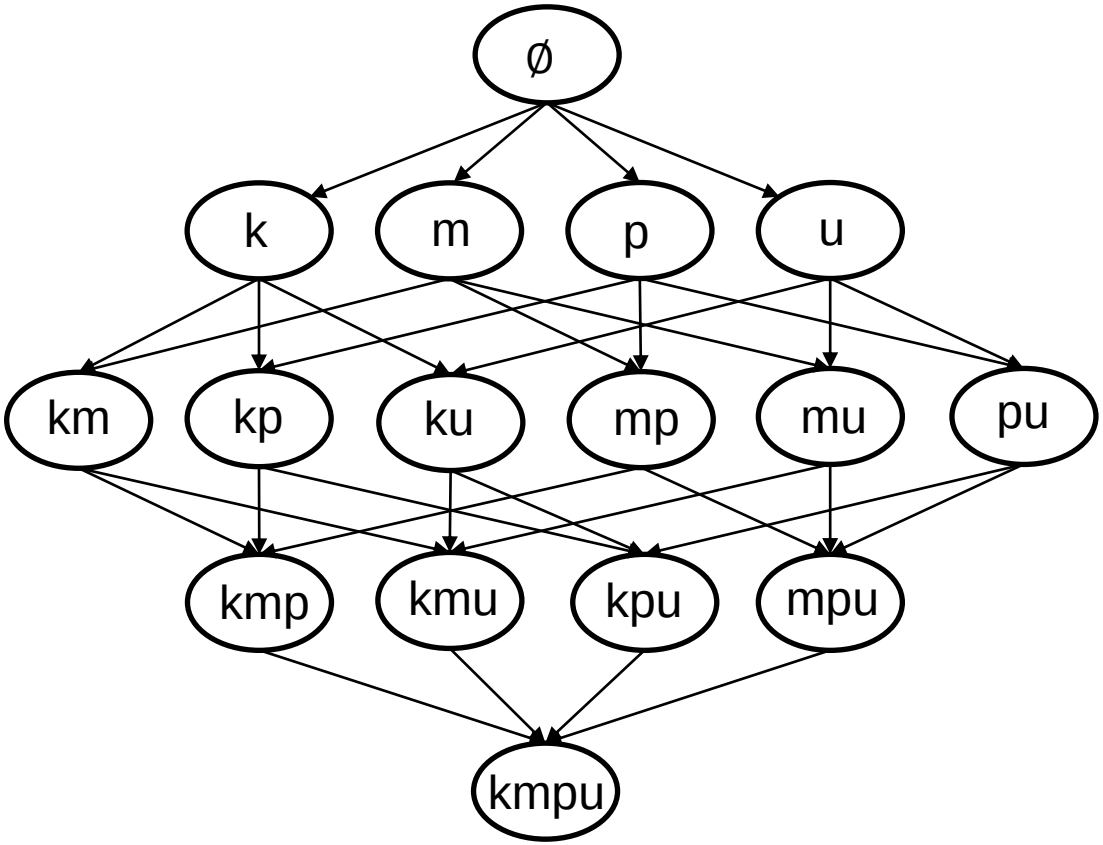
s{kruh, plenice} = ? 0.3

s{mleko, union, plenice} = ? 0.4

Mreža naborov

število možnih naborov je 2^d

d = 5 → 32
d = 6 → 64
d = 20 → 1048576



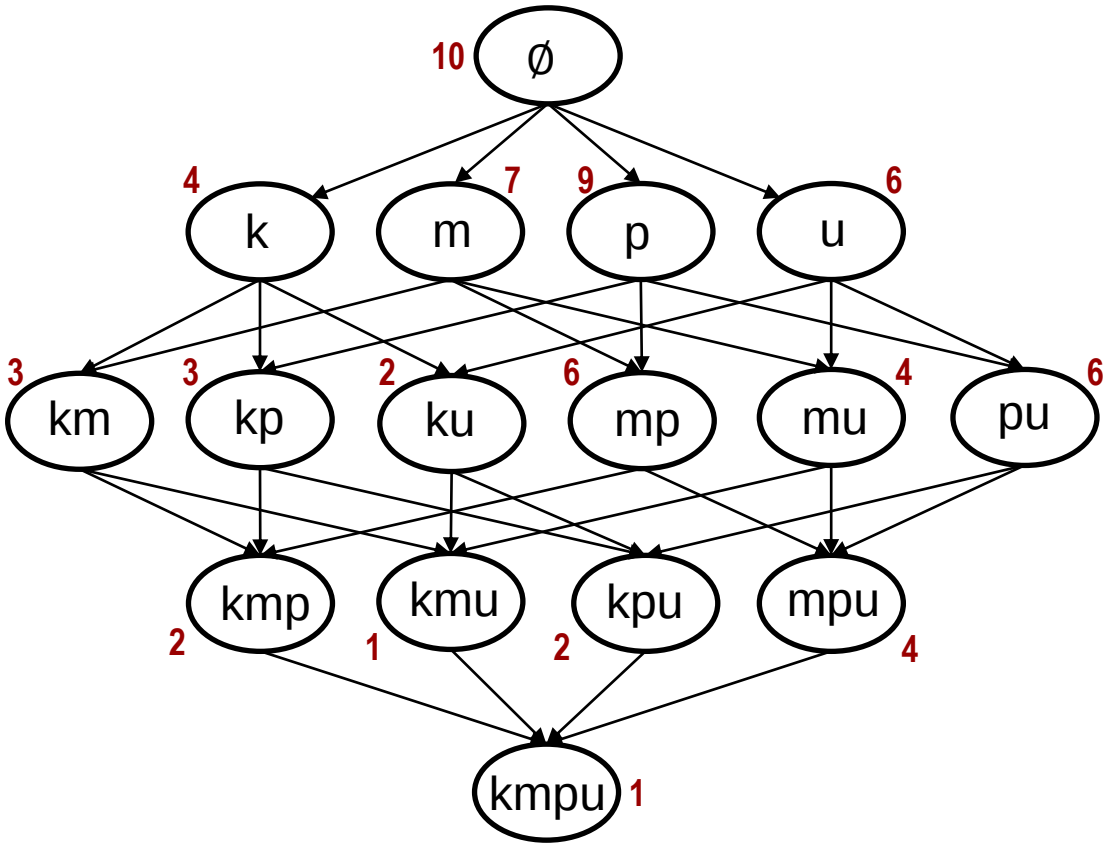
tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

CILJ: kateri so nabori X, pri katerih velja s(X) ≥ min_supp ?

POGOSTI NABORI

pogosti nabori s(X) ≥ 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

Iskanje pogostih naborov

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

{∅}
{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}

M

- primerjava vsake transakcije z vsakim kandidatom
- ~ O(NMw)

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

Načelo Apriori

Teorem 1

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

$$s(X) \geq min\_supp \Rightarrow s(Y) \geq min\_supp, Y \subset X$$

Kako nam to lahko pomaga?

Teorem 2

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

$$s(X) \not\geq min\_supp \Rightarrow s(Y) \not\geq min\_supp, X \subset Y$$

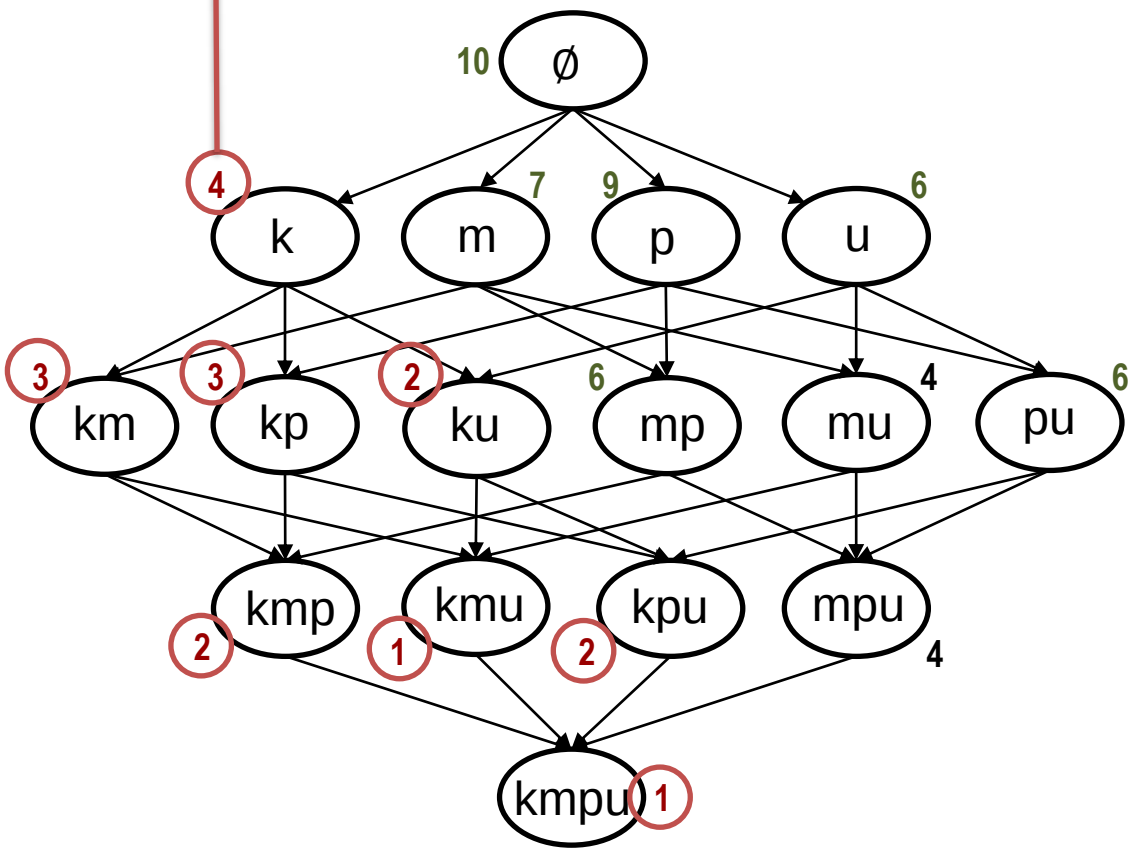
ANTI-MONOTONOST: podpora nabora nikoli ne presega nabora njegovega podnabora.

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

NEPOGOSTI NABORI IN „REZANJE“ MREŽE

nepogost nabor

infrequent itemset



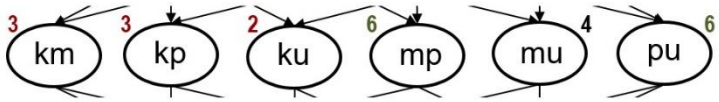
tid	vsebina košarice
1	kruh, mleko
2	kruh, plenice, union
3	mleko, plenice, union
4	kruh, mleko, plenice, union
5	kruh, mleko, plenice
6	mleko, plenice
7	plenice
8	mleko, union, plenice
9	plenice, union
10	mleko, plenice, union

Naj bo min_supp = 0.6 ⟶ **σ(X) ≥ 6**

Strategije za iskanje pogostih naborov

- **Zmanjšati število kandidatov** (M)
  - s pomočjo tehnik rezanja (načelo Apriori!)

~~{∅}~~
~~{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}~~



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

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

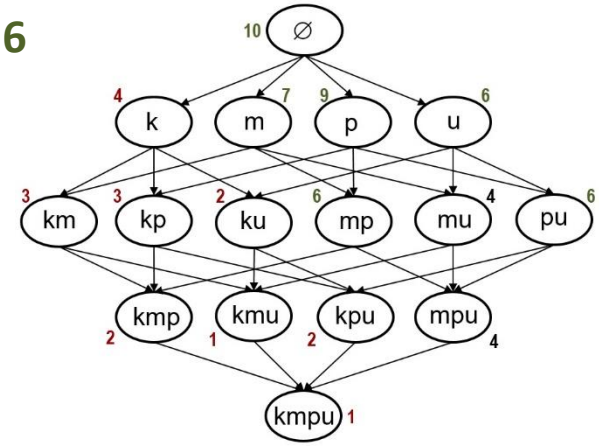
- **Zmanjšati število primerjav** (NM)
  - uporaba učinkovitih podatkovnih struktur za shranjevanje kandidatov ali transakcij
  - ni potrebe po primerjavi vsakega kandidata z vsako transakcijo

Načelo Apriori: ilustracija

predmeti (1-nabori)

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

$\sigma(X) \geq 6$



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

pari (2-nabori)

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

trojke (3-nabori)

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

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

APRIORI: ALGORITEM

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 (1)

```
k = 1
F_k = {i | i ∈ I ∧ σ({i}) ≥ N * min_supp}   # poišči pogoste 1-nabore
ponovi:
        k = k + 1
        C_k = kandidati (F_{k-1})   # ustvari seznam kandidatov
        izračunaj podporo naborom v C_k   # preveri podporo kandidatov
        F_k = {c | c ∈ C_k ∧ σ(c) ≥ N * min_supp}   # obdrži pogoste kandidate
dokler F_k = 0
vrni ∪ F_k
```

APRIORI: ALGORITEM (II)

```
k = 1
F_k = {i | i ∈ I ∧ σ({i}) ≥ N * min_supp}   # poišči pogoste 1-nabore
ponovi:
        k = k + 1
        C_k = kandidati (F_{k-1})   # ustvari seznam kandidatov
        izračunaj podporo naborom v C_k   # preveri podporo kandidatov
        F_k = {c | c ∈ C_k ∧ σ(c) ≥ N * min_supp}   # obdrži pogoste kandidate
dokler F_k = 0
vrni ∪ F_k
```

```
σ(c) = 0 za vsak c ∈ C_k
za vsako t ∈ T:   # transakcije
        C_t = {c | c ∈ C_k ∧ c ⊂ T}
        za vsak c ∈ C_t:
                σ(c) = σ(c) + 1   # povečaj števec podpore
```

Generiranje kandidatov (I)

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

Metoda F_{k-1} x F_1 razširitev (k-1)-nabora s pogostim predmetom

pogosti 1-nabori

PREDMET
jogurt
mleko
plenice
union

pogosti 2-nabori

NABOR PREDMETOV
{jogurt, plenice}
{mleko, plenice}
{mleko, union}
{plenice, union}

kandidati (k=3)

NABOR PREDMETOV
{jogurt, plenice, mleko}
{jogurt, plenice, union}
{mleko, plenice, union}
{mleko, union, jogurt}

nepogosti nabori

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

GENERIRANJE KANDIDATOV (II)

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

pogosti 2-nabori

NABOR PREDMETOV
{jogurt, plenice}
{mleko, plenice}
{mleko, union}
{plenice, union}

pogosti 2-nabori

NABOR PREDMETOV
{jogurt, plenice}
{mleko, plenice}
{mleko, union}
{plenice, union}

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

kandidati (k=3)

NABOR PREDMETOV
{mleko, plenice, union}

Problem možnega podvajanja rešimo z leksikografsko ureditvijo:

{mleko, plenice} ∩ {mleko, union} → {mleko, plenice, union}

{mleko, plenice} ∩ {union, plenice} → /

Faktorji vpliva na računsko zahtevnost

- **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
hruške

vse transakcije vsebujejo vsaj en pogost predmet

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

pogosti 2-nabori

NABOR PREDMETOV
{ananas, češnje}
{ananas, limone}

nekatere transakcije ne vsebujejo nobenega pogostega para!

POVEZOVALNA PRAVILA: DEFINICIJE

Povezovalna pravila *association rules*

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

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

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

podpora *support*

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

kako pogosto je pravilo prisotno v množici transakcij

zaupanje *confidence*

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

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

Povezovalna pravila: primeri

X→Y ≡ {kruh} → {plenice, mleko}

s(X→Y) = ? 0.2

c(X→Y) = ? 0.5

X→Y ≡ {union} → {plenice}

s(X→Y) = ? 0.6

c(X→Y) = ? **1.0**

X→Y ≡ {plenice} → {union}

s(X→Y) = ? 0.6

c(X→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 ≥ *min_supp*
- ❑ zaupanje ≥ *min_conf*

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

računsko zahtevno!

korak 1: iskanje pogostih naborov

poišči vse nabore, za katere velja podpora ≥ *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 $X \rightarrow Y$)

Povezovalna pravila: možna pravila

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

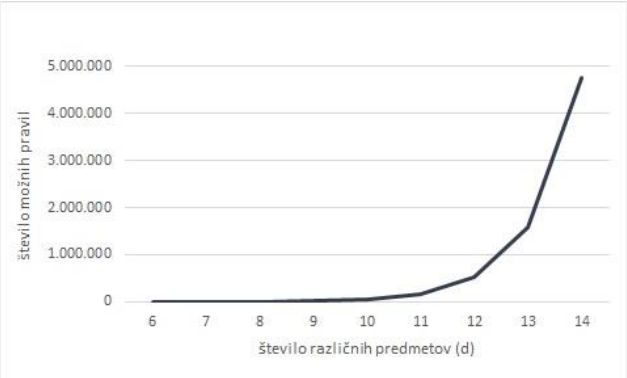
kmp →d	kmd →p	kpd →m	mpd →k
k →mpd	m →kpd	p →kmd	d →kmp
km →pd	kp →md	kd →mp	mp →kd
md →kp	pd →km	∅→kmpu	kmpu→∅

število vseh možnih pravil

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

d = 6 ⟶ 602 pravil

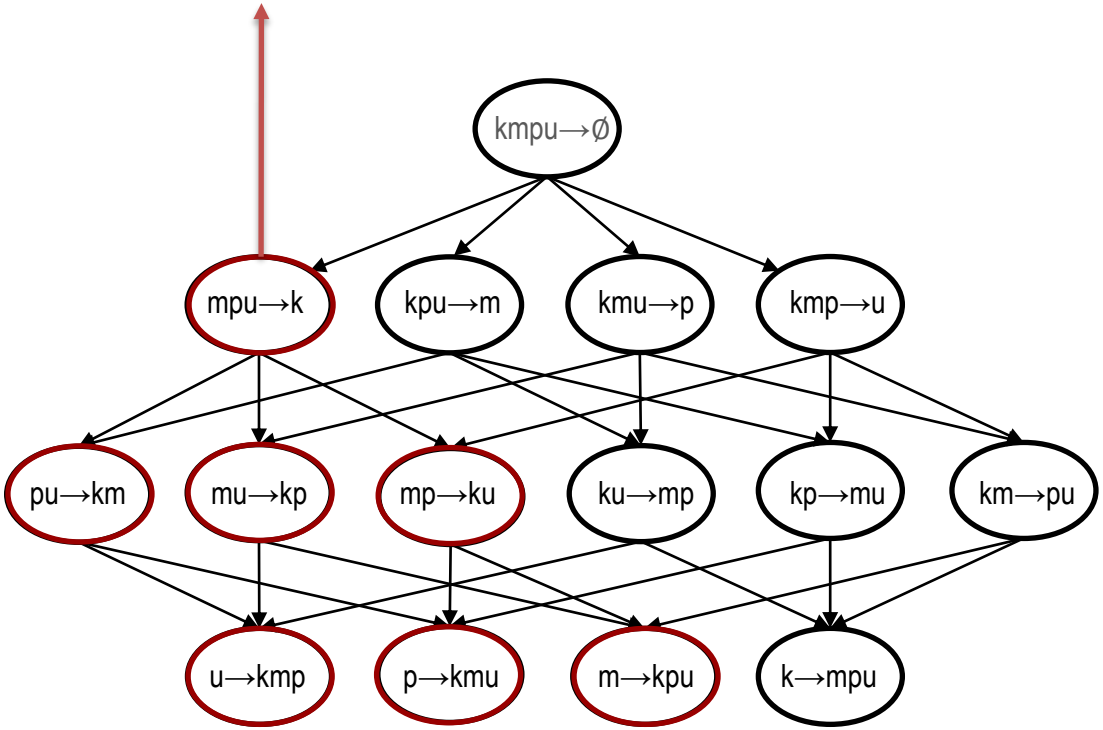
d = 14 ⟶ 4.750.202 pravil



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



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

ANTI-MONOTONOST velja za pravila iz istega nabora

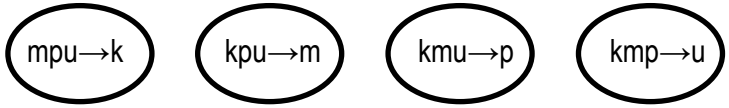
$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

GENERIRANJE PRAVIL: POSTOPEK

R = {} seznam pravil

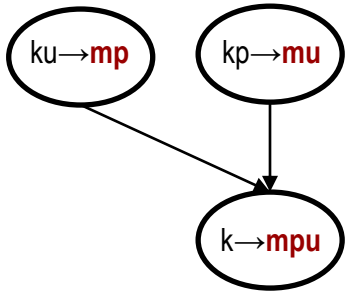
Za vsak pogost nabor f_k, ki ima vsaj dva predmeta: **k>1** npr. f_k = {kmpu}

- 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
- …



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

VREDNOTENJE

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“)
  - npr. mleko ⟶ kruh
- zmerna podpora, nizko zaupanje („nezanesljiva“)
  - npr. mleko ⟶ tuna
- nizka podpora, nizko zaupanje („očitno nezanimiva“)
  - npr. bučno olje ⟶ detergent
- **nizka podpora, visoko zaupanje** („potencialno zelo zanimivo“)
  - 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(čaj → kava) = 0.15

c(čaj → kava) = 150/200 = 0.75

Ampak: kavo pije 80% anketirancev!

Interest = |c(X → Y) – s(Y)| Interest = |0.75 – 0.8| = 0.05 → nezanimivo!

zanesljivost glede na izhodiščno verjetnost 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

{union} → {kruh} **Interest = 0.35**

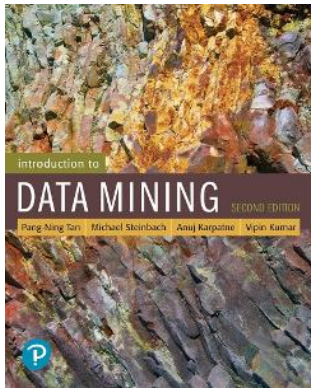
{union} → {plenice} **Interest = 0.07**

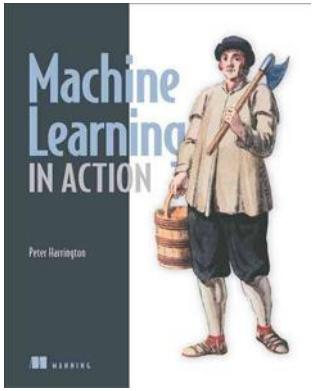
Viri in literatura

- 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)