Metódy asociačných pravidiel a ich aplikácie

Vasyl Khorev

Vedúci bakalárskej práce: doc. RNDr. Ľubomír Antoni, PhD.
Konzultant: prof. RNDr. Stanislav Krajči, PhD.
Oponent: prof. RNDr. Gabriel Semanišin, PhD.

Prírodovedecká fakulta
Univerzita Pavla Jozefa Šafárika v Košiciach

Obhajoba bakalárskej práce
Košice, 19. jún 2024

Motivácia

- Hľadanie vzťahov (súvislosti) medzi atribútmi v dátach.
- **Analýza nákupného koša** môže zlepšiť marketingové stratégie a zvýšiť predaj.
- Aplikácia v **zdravotníctve**, napríklad nájdenie vzťahov medzi vykonanými procedúrami a stanovenými diagnózami.
- Zlepšenie **bezpečnosti na cestách** pomocou identifikácie vzťahov medzi podmienkami a nehodami.

Príklad

Asociačné pravidlo $\{\text{zemiakové lupienky}\} \Rightarrow \{\text{nápoj}\}$.

Ciele

- Popísať známe algoritmy pre generovanie asociačných pravidiel.
- Implementovať známe algoritmy pre generovanie asociačných pravidiel a porovnať ich výhody a nevýhody.
- Aplikovať implementované algoritmy na vybranú údajovú sadu.

Základné definície

Definícia

Uvažujme množinu $\mathcal{I} = \{i_1, i_2, \ldots, i_m\}$, pričom m je prirodzené číslo a prvky tejto množiny sú literály, ktoré budeme nazývať aj položky. Ľubovoľnú podmnožinu T množiny $\mathcal{I}$ budeme nazývať *množina položiek* alebo *transakcia*.

Príklad

Uvažujme množinu $\mathcal{I} = \{\text{mlieko}, \text{maslo}, \text{vajcia}, \text{chlieb}\}$. Potom transakciou môže byt množina $\{\text{vajcia}, \text{maslo}, \text{chlieb}\}$.

Základné definície

Definícia

Databáza $\mathcal{D}$ je postupnosť transakcií (položkových množín).

Príklad

Uvažujme postupnosť
$\mathcal{D} = (\{\text{chlieb}, \text{maslo}\}, \{\text{mlieko}\}, \{\text{vajcia}, \text{maslo}, \text{chlieb}\})$.

Tabuľka: Ekvivalentná reprezentácia databázy $\mathcal{D}$.

Transakcia	mlieko	maslo	vajcia	chlieb
1	0	1	0	1
2	1	0	0	0
3	0	1	1	1

Podpora pravidla (support)

Definícia

Definujme funkciu $\mathrm{support}(X)$ pre množinu položiek X takto:

$$\mathrm{support}(X) = \frac{\text{počet transakcií v } \mathcal{D} \text{ obsahujúcich } X}{\text{celkový počet transakcií}}$$

Príklad

Uvažujme databázu z predošlej snímky. Potom

$$\mathrm{support}(\{\mathrm{maslo}\}) = \frac{2}{3}.$$

Definícia

Množina X je *frekventovaná*, ak platí $\mathrm{support}(X) \geq \mathrm{min\_sup}$.

Spoľahlivosť pravidla (confidence)

Definícia

Definujme funkciu $\mathrm{confidence}(X \Rightarrow Y)$ pre pravidlo $X \Rightarrow Y$ takto:

$$\mathrm{confidence}(X \Rightarrow Y) = \frac{\mathrm{support}(X \cup Y)}{\mathrm{support}(X)}$$

Príklad

Uvažujme databázu z predošlej snímky. Potom

$$\mathrm{confidence}(\{\mathrm{maslo}\} \Rightarrow \{\mathrm{chlieb}\}) = \frac{\mathrm{support}(\{\mathrm{maslo}, \mathrm{chlieb}\})}{\mathrm{support}(\{\mathrm{maslo}\})} = 1.$$

Asociačné pravidlo

Definícia

Asociačné pravidlo $X \Rightarrow Y$ je dvojica disjunktných množín $X, Y \subseteq \mathcal{I}$, takých že:

1. množiny X a Y sú frekventované:

$$\text{support}(X \cup Y) \geq \text{min\_sup}$$

2. ak je X frekventovaná, tak aj Y je frekventovaná:

$$\text{confidence}(X \Rightarrow Y) \geq \text{min\_conf}$$

Hľadanie asociačných pravidiel

Problém hľadania asociačných pravidiel je možné rozdeliť do dvoch etáp:

1. Nájsť všetky frekventované množiny položiek.
2. Použitím nájdených frekventovaných položiek Y vygenerovať pravidlá:

$$X \Rightarrow Y \setminus X \qquad \text{pre } X \subset Y,\ X \neq \emptyset.$$

Prvé vlastné výsledky v práci:

- Príprava vlastných príkladov, tabuliek, diagramov a obrázkov pre metódu Apriori a FP-growth.
- Programátori väčšinou používajú tieto metódy priamo z knižníc, naším cieľom bolo tieto metódy matematicky popísať a simulovať ich výpočet do posledného detailu.
- Niektoré časti týchto algoritmov naprogramované aj od úplného základu.

Poznámka

Ak je množina Y nefrekventovaná, potom každá nadmnožina $X \supseteq Y$ je nefrekventovaná.

Postup algoritmu:

1. začíname jednoprvkovými frekventovanými množinami.
2. - generujeme $(k+1)$-prvkové množiny, tak že zjednotíme k-prvkové množiny, ak máju spoločných prvých $k-1$ prvkov.
   - odstránime tie, ktoré nespĺňajú minimálnu prahovú hodnotu podpory.
3. takto pokračujeme, kým nedostaneme prázdnu množinu.

Množina	Podpora
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Tabuľka: L_2

Množina	Podpora
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2
$\{I_1, I_3, I_5\}$	1
$\{I_2, I_3, I_4\}$	0
$\{I_2, I_4, I_5\}$	0
$\{I_2, I_3, I_4\}$	0

Tabuľka: C_3

Transakcia
I_1, I_2, I_5
I_2, I_4
I_2, I_3
I_1, I_2, I_4
I_1, I_3
I_2, I_3
I_1, I_3
I_1, I_2, I_3, I_5
I_1, I_2, I_3

Tabuľka: DB

FP-growth

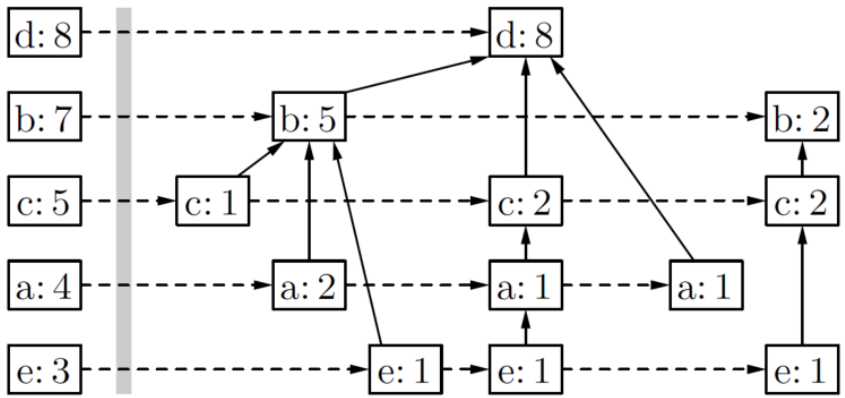
Umožňuje hľadanie frekventovaných množín bez generovania kandidátnych množín.

- **Krok 1:** dvoma prechodmi databázou zostrojiť kompaktný **FP-strom**.
  1. zostupne usporiadať položky v transakciách podľa hodnoty podpory; nefrekventovane položky vymazať.
  2. zostrojiť strom.
- **Krok 2:** Vyextrahovať frekventované množiny z **FP-stromu**.
  - od listov smerom ku koreňu (zdola nahor): najprv nájsť frekventované množiny obsahujúce $\{e\}$, $\{e, a\}$ . . . , potom $\{e,c\}$,$\{e,c,b\}$ atď..

Konštrukcia FP-stromu

Zostupne usporiadame frekventované položky $i \in \mathcal{I}$ podlá hodnoty $\mathrm{support}(i)$.

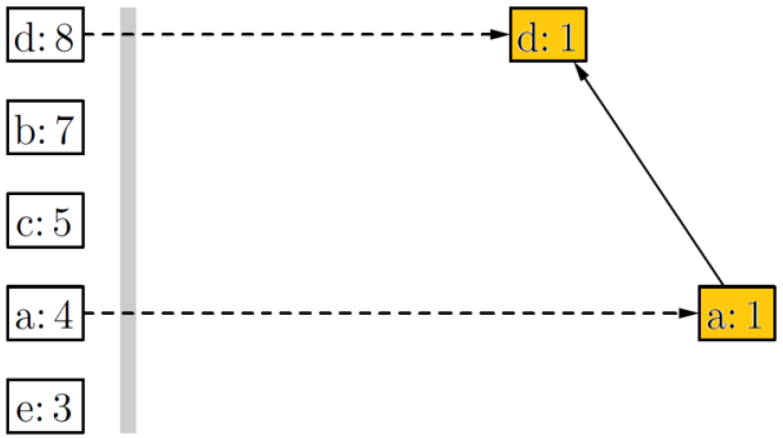
Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



f a g nie sú frekventované pri $\mathrm{min\_sup} = 3$

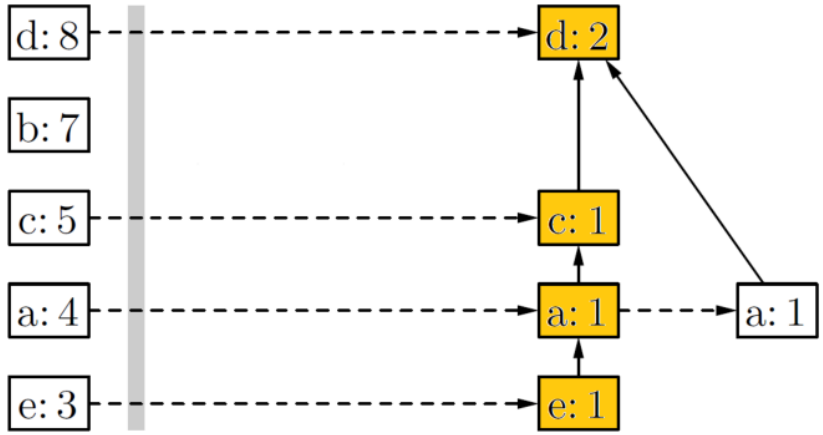
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



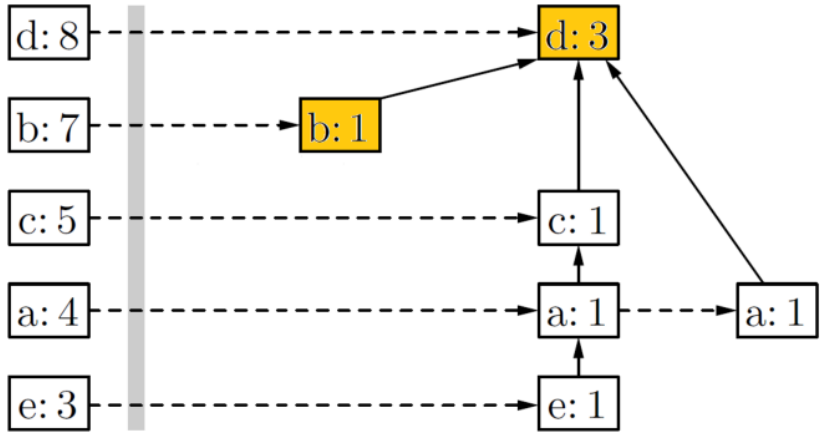
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



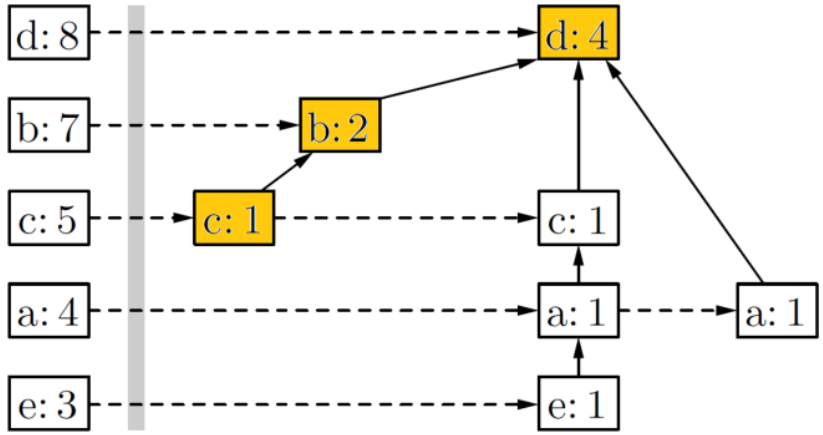
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



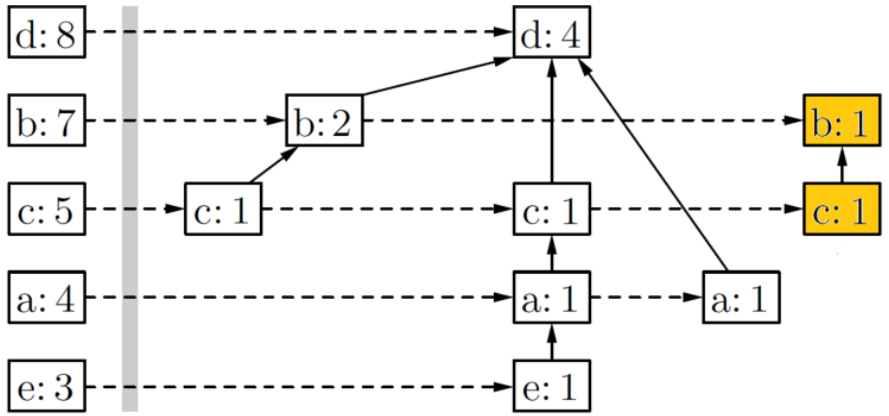
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



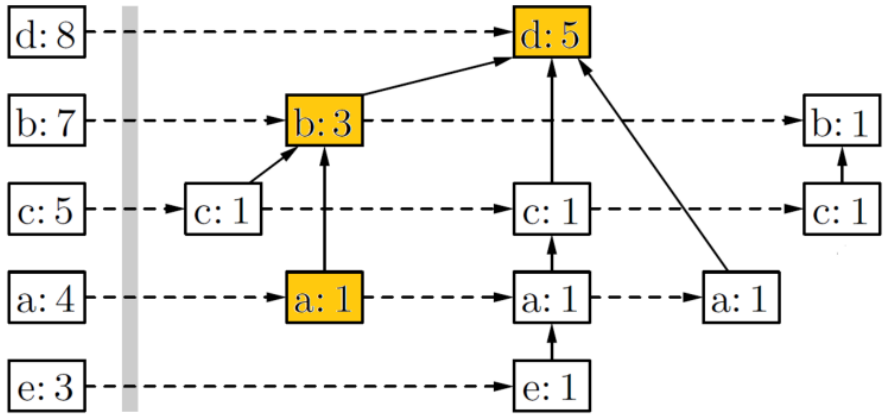
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



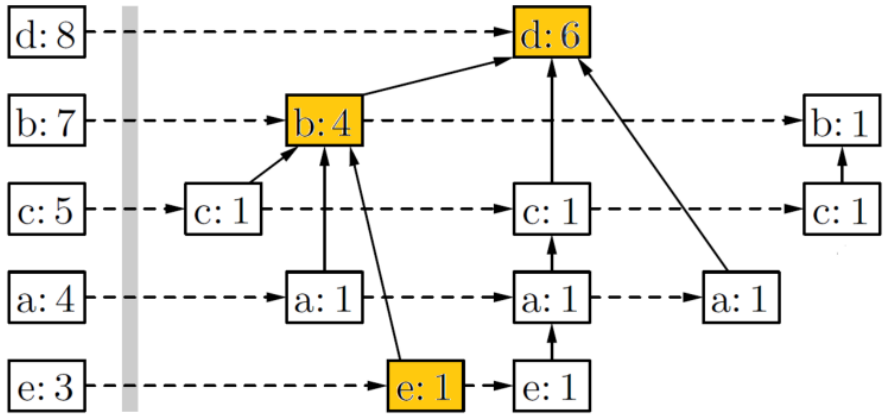
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



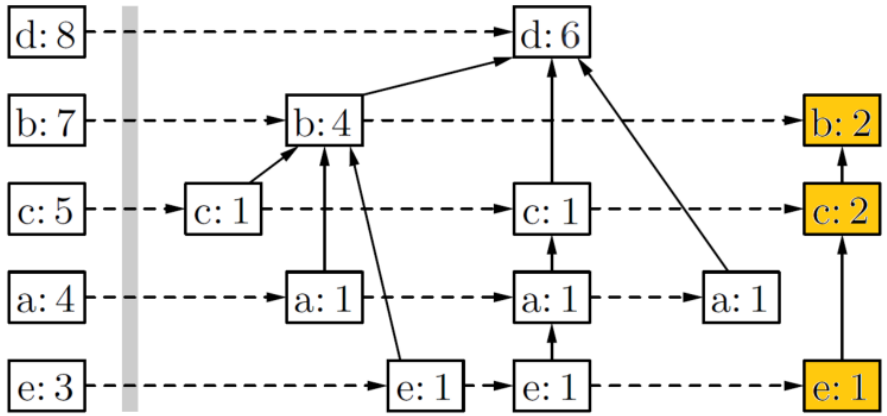
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



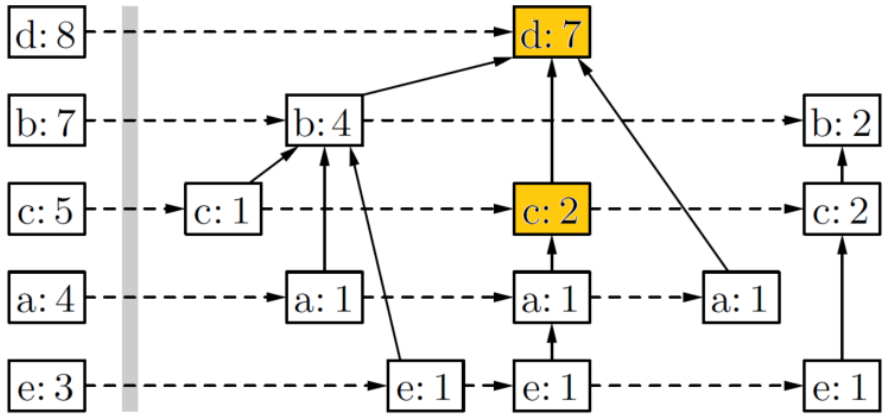
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



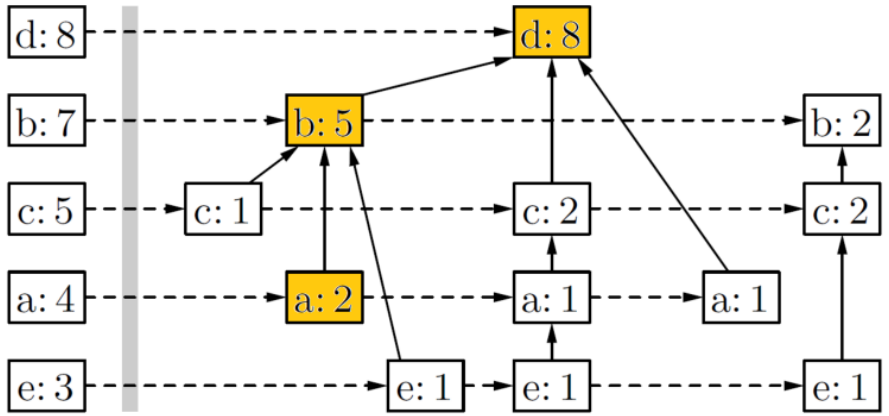
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



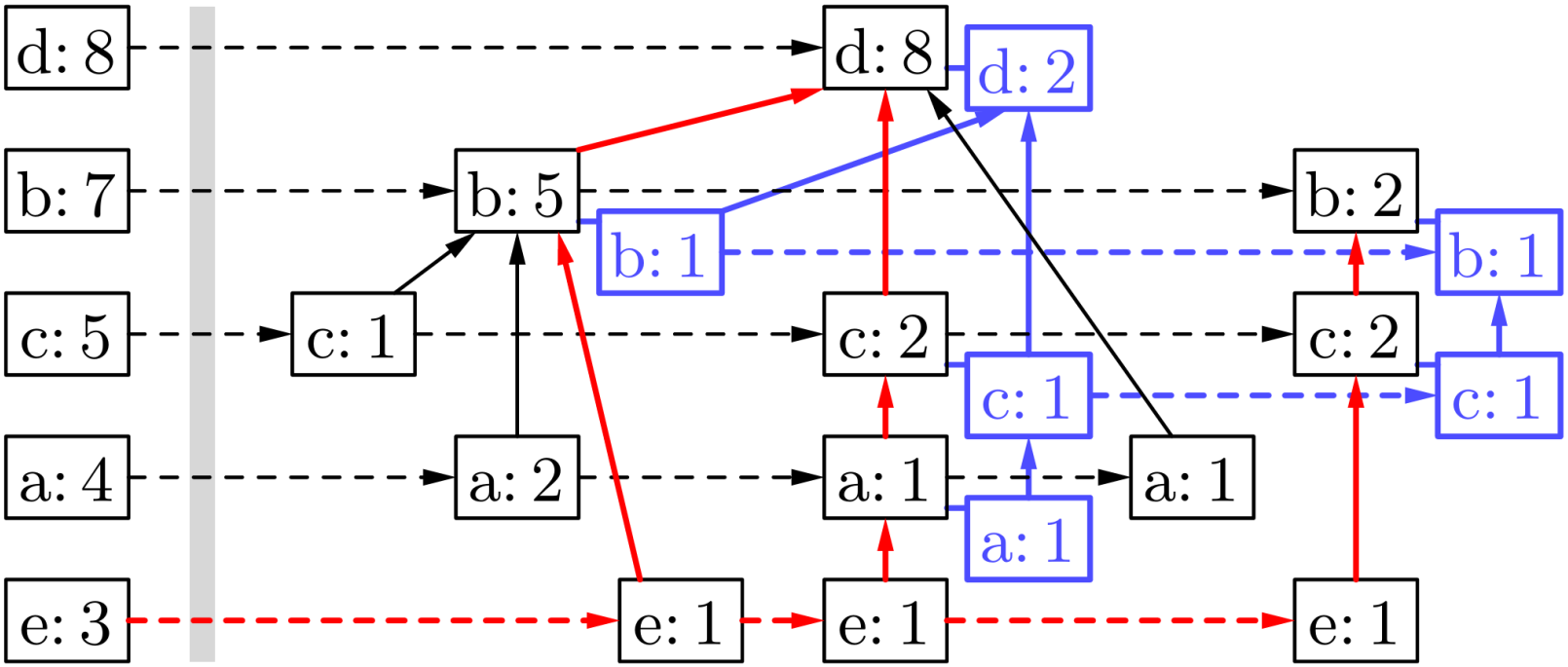
Konštrukcia FP-stromu

Transakcie	Usporiadané
a,d,f	d,a
a,c,d,e	d,c,a,e
b,d	d,b
b,c,d	d,b,c
b,c	b,c
a,b,d	d,b,a
b,d,e	d,b,e
b,c,e,g	b,c,e
c,d,f	d,c
a,b,d	d,b,a



Podmienený FP-strom

Strom transakcií, ktoré obsahujú položku e.



Obr.: Podmienený FP-strom $T|e$

Dataset

Údaje o predaji z pekárne „The Bread Basket“ v Edinburghu:

- 20507 záznamov
- viac ako 9000 riadkov
- 5 atribútov

No.	**Items**	**DateTime**	**Daypart**	**DayType**
1	Bread	30. 10. 2016 9:58	Morning	Weekend
2	Scandinavian	30. 10. 2016 10:05	Morning	Weekend
2	Scandinavian	30. 10. 2016 10:05	Morning	Weekend
⋮	⋮	⋮	⋮	⋮
9683	Coffee	04. 09. 2017 14:57	Afternoon	Weekend
9683	Pastry	04. 09. 2017 14:57	Afternoon	Weekend
9684	Smoothies	04. 09. 2017 15:04	Afternoon	Weekend

Tabuľka: Ukážka použitého dátového súboru

Parametre a výsledný dátový rámec

Prahové hodnoty parametrov:

- podpora: $\min\_sup = 0{,}03$.
- spoľahlivosť: $\min\_conf = 0{,}5$.

No.	Položka 1	...	Položka 94	Afternoon	Evening	Morning	Night
1	0	...	0	0	0	1	0
2	0	...	0	0	0	1	0
⋮	⋮	...	⋮	⋮	⋮	⋮	⋮
9683	0	...	0	0	1	0	0
9684	0	...	0	0	1	0	0

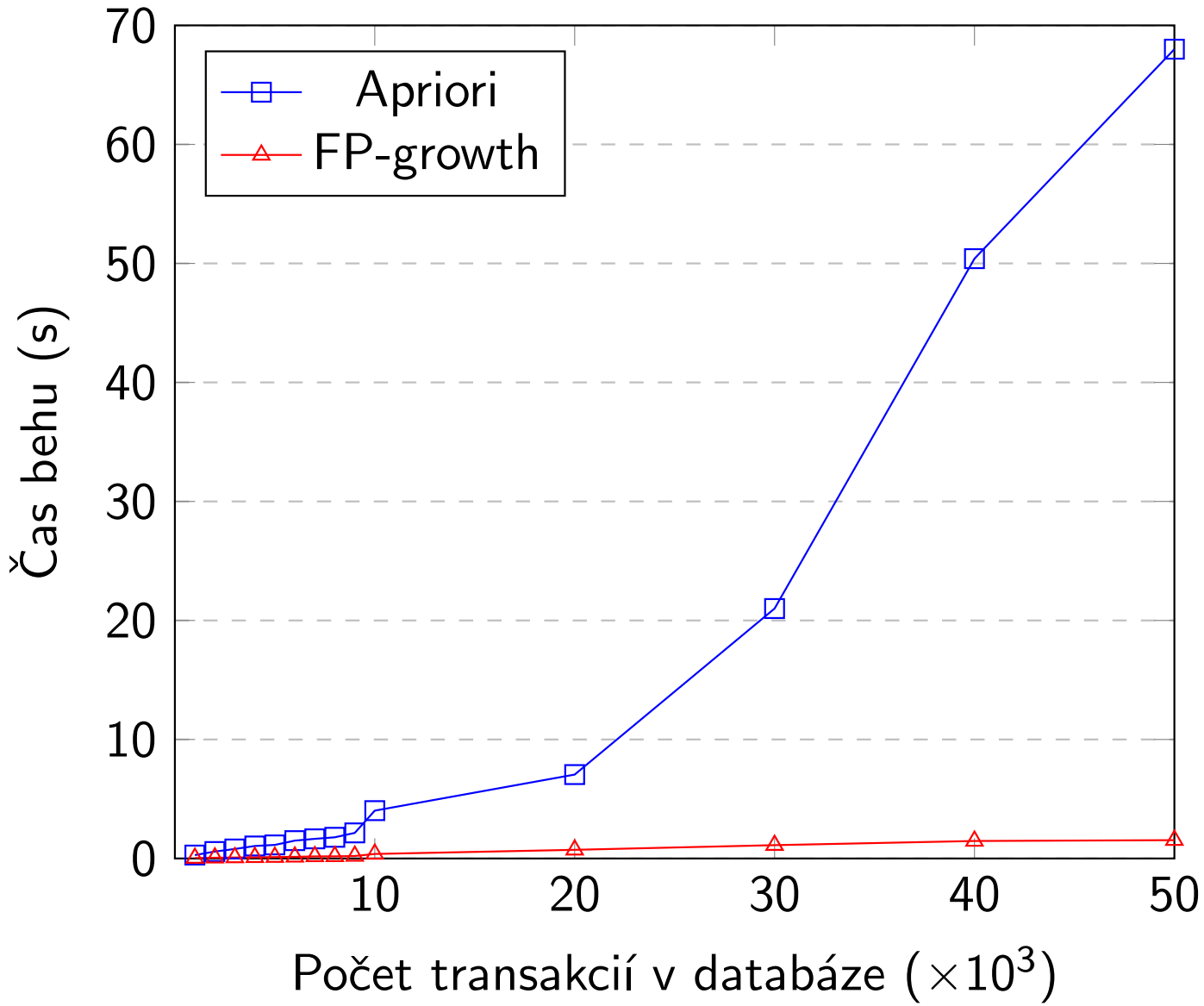
Tabuľka: Výsledná tabuľka po transformácii.

Výsledky

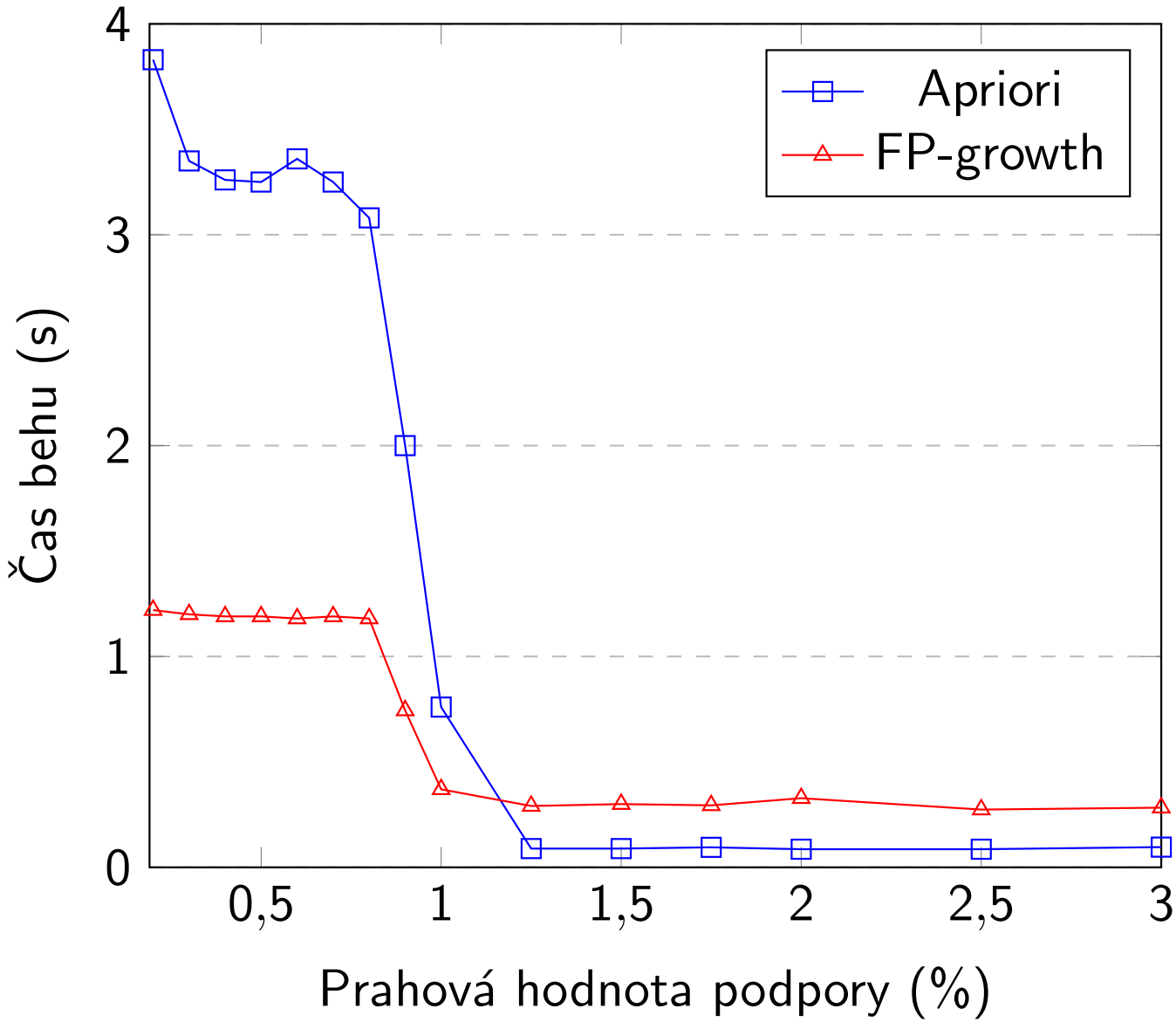
No.	Antecedents	Consequents	Support	Confidence
1	{Soup}	{Afternoon}	0, 032647	0, 947853
2	{Sandwich, Coffee}	{Afternoon}	0, 033492	0, 875691
⋮	⋮	⋮	⋮	⋮
12	{Medialuna}	{Coffee}	0, 035182	0, 569231
13	{Hot chocolate}	{Afternoon}	0, 033069	0, 567029
14	{Morning, Pastry}	{Coffee}	0, 033492	0, 554196
⋮	⋮	⋮	⋮	⋮
18	{Sandwich}	{Coffee}	0, 038246	0, 532352
19	{Cake}	{Coffee}	0, 054727	0, 526958
⋮	⋮	⋮	⋮	⋮
22	{Morning}	{Coffee}	0, 223244	0, 514989
23	{Bread}	{Afternoon}	0, 164395	0, 502422

Tabuľka: Získané asociačné pravidla so spoľahlivosťou viac ako 50 %.

Porovnanie času behu v závislosti od rozmeru DB



Vplyv prahovej hodnoty podpory na čas behu



Literatúra

- Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. SIGMOD Rec. 29, 2 (June 2000), 1–12. https://doi.org/10.1145/335191.335372
- Christian Borgelt. 2005. An implementation of the FP-growth algorithm. In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (OSDM '05). Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/1133905.1133907
- Voroncov Konstantin. 2023. Mašynnoe obučenie (Strojové učenie, v ruštine) [online][cit. 02.04.2024]. Dostupne na: www.machinelearning.ru

Ďakujem za pozornosť

Otázky

- **Vo výsledkoch uvádzate asociačné pravidlá, ktoré majú na pravej strane pravidla len jednu položku. Je možné generovať aj viacero položiek súčasne na pravej strane pravidla?**
  - Áno, každý z implementovaných algoritmov dokáže generovať aj takéto pravidlá. Dôvod, prečo takéto pravidlá neboli vygenerované, spočíva v charakteristikách použitého datasetu.
- **Zoznam použitej literatúry je pomerne dlhý. Používali ste ju naozaj v takom rozsahu?**
- **Poznáte metódu FCA - Formal Concept Analysis? Líši sa nejako principiálne od vami skúmaných metód?**