Frekventni skupovi objekata

Matej Mihelčić

Prirodoslovno-matematički fakultet, Sveučilište u Zagrebu

matmih@math.hr

12. travnja, 2026.

Traženje frekventnih skupova objekata

- Traženje frekventnih skupova objekata spada u zadatke **deskriptivnog** modeliranja.
- Podaci iz kojih se traže **ne sadrže** ciljnu varijablu.
- Originalna motivacija za traženje frekventnih skupova objekata dolazi iz potrebe za analizom **transakcijskih podataka** u supermarketima.
- Cilj je bio analizirati ponašanje kupaca na temelju kupljenih proizvoda.
- Frekventni skupovi objekata opisuju **koliko često** su neki proizvodi **kupljeni zajedno**.
- Objekte međutim možemo smatrati i puno općenitije, kao npr. svojstvo koje ima određena transakcija (entitet) u skupu podataka.
- Označimo s $\mathcal{I}$ skup svih objekata (proizvoda).
- **Transakcija** nad $\mathcal{I}$ je par $T = (tid, I)$, gdje je tid identifikator transakcije, a I je skup objekata iz $\mathcal{I}$.
- Baza podataka D nad $\mathcal{I}$ je skup transakcija nad $\mathcal{I}$ takvih da svaka transakcija ima jedinstveni identifikator.

Traženje frekventnih skupova objekata

- Transakcija $T = (tid, I)$ **podupire** skup X ako $X \subseteq I$.
- **Pokrivanje** (eng. cover) skupa X iz D se sastoji od skupa identifikatora transakcija iz D koje podupiru X.
  $Pok(X, D) = \{tid_s \in T_s,\ T_s \in D |\ X \subseteq I_s\}$.
- **Potpora** (eng. support) skupa X u D je broj transakcija sadržanih u pokrivanju skupa X iz D. $Pot(X, D) = |Pok(X)|$.
- **Frekvencija** skupa X u D je vjerojatnost da se X javlja u transakciji.
  $Frek(X, D) = \dfrac{Pot(X)}{|D|}$.
- Skup objekata zovemo **frekventnim** ako njegova potpora nije manja od predefiniranog praga σ, gdje $0 \leq \sigma \leq |D|$.
- Ukoliko koristimo frekvencije, definiramo prag minimalne frekvencije $0 \leq \sigma_f \leq 1$. Tada, $\sigma = \lceil \sigma_f \cdot |D| \rceil$.

Traženje frekventnih skupova objekata

Definicija

Neka je D zadani skup transakcija (baza podataka) nad skupom objekata $\mathcal{I}$, a σ prag minimalne potpore. Kolekcija frekventnih skupova objekata iz D s obzirom na σ se definira kao: $\mathcal{F}(D, \sigma) = \{X \subseteq \mathcal{I} | \ Pot(X, D) \geq \sigma\}$.

Definicija

Neka je zadan skup objekata $\mathcal{I}$, baza transakcija D nad $\mathcal{I}$ i minimalni prag potpore σ, **zadatak traženja frekventnih skupova objekata** je pronaći $\mathcal{F}(D, \sigma)$.

Traženje frekventnih skupova objekata - primjer

Pretpostavimo da imamo skup objekata $\mathcal{I} = \{\text{pivo, čips, pizza, vino}\}$.

tid	skup objekata
1	{pivo, čips}
2	{pizza, vino}
3	{čips, pizza}
4	{pivo, vino, čips}

Tražimo skupove frekventnih objekata uz minimalni prag potpore $\sigma = 1$.

Traženje frekventnih skupova objekata - primjer

Skup	Pokrivanje	Potpora	Frekvencija
{}	{1, 2, 3, 4}	4	1
{pivo}	{1, 4}	2	0.5
{čips}	{1, 3, 4}	3	0.75
{pizza}	{2, 3}	2	0.5
{vino}	{2, 4}	2	0.5
{pivo, čips}	{1, 4}	2	0.5
{pivo, vino}	{4}	1	0.25
{pizza, vino}	{2}	1	0.25
{čips, pizza}	{3}	1	0.25
{vino, čips}	{4}	1	0.25
{pivo, vino, čips}	{4}	1	0.25

Traženje frekventnih skupova objekata

- Prostor pretraživanja je **eksponencijalan** u broju objekata iz $\mathcal{I}$.
- Prostor objekata sadrži $2^{|\mathcal{I}|}$ različitih skupova objekata.
- Generiranje svih podskupova **nije računalno realistična** opcija.
- Čak i nakon uvođenja uvjeta na minimalnu potporu skupa objekata na jednu transakciju ($\sigma = 1$), prostor je i dalje suviše velik za učinkovito računanje na većim bazama transakcija.
- Računanje uz uvođenje minimalnog praga potpore većeg od 1 dovodi do **problema s memorijom** (nemogućnost pamćenja generiranih skupova objekata).
- **Parcijalno rješenje** je u korištenju sljedećeg svojstva:

Definicija

Neka je zadana baza transakcija D nad $\mathcal{I}$ i dva skupa $X, Y \subseteq \mathcal{I}$.
Monotonost potpore se definira kao, $X \subseteq Y \Rightarrow Pot(Y, D) \leq Pot(X, D)$.

Traženje frekventnih skupova objekata

- Iz prethodne definicije slijedi:
  - Ukoliko skup objekata **nije frekventan**, tada niti njegovi nadskupovi **ne mogu** biti frekventni.
  - Ukoliko **je skup frekventan**, tada **svi njegovi podskupovi** također **moraju biti frekventni**.
- Ovo svojstvo monotonosti frekventnih skupova objekata se zove svojstvo zatvorenosti prema dolje. Skup frekventnih objekata je zatvoren prema dolje pošto je svaki podskup frekventnog skupa objekata također frekventan.
- Skupovi objekata koji nisu frekventni imaju **svojstvo zatvorenosti prema gore**.

Traženje frekventnih skupova objekata

- Računanje potpore skupova objekata zahtjeva pristup bazi. Baze zbog njihove veličine najčešće ne možemo spremiti u radnu memoriju.
- Važnu ulogu prilikom izračunavanja ima reprezentacija baze (skupa podataka).
- Bazu možemo reprezentirati kao dvodimenzionalnu binarnu matricu, gdje svaki redak predstavlja jednu transakciju, a stupci predstavljaju objekte iz $\mathcal{I}$.
- Takvu matricu možemo implementirati na nekoliko načina:
    - Najčešće korištena je **horizontalna reprezentacija**. Svaka transakcija ima poseban identifikator i listu objekata koji se javljaju u transakciji.
    - **Vertikalna reprezentacija** u redcima reprezentira skupove objekata, a svaki stupac retka sadrži informaciju o tome sadrži li transakcija taj skup objekata ili ne.
- Ukoliko koristimo horizontalnu reprezentaciju, pri računanju potpore skupa objekata X, moramo iterirati kroz cijelu bazu transakcija i testirati svaku transakciju T (računamo vrijedi li $X \subseteq I_T$).

Traženje frekventnih skupova objekata

- Postupak računanja potpore skupova objekata se može odjednom provesti za cijele kolekcije skupova objekata.
- Vertikalna reprezentacija omogućava brzo računanje potpore skupa objekata X jednostavnim računanjem presjeka između dva podskupa $Y, Z \subseteq X$ takvih da $Y \cup Z = X$.
- Međutim, vertikalna reprezentacija zahtijeva smještanje velikog broja skupova objekata u radnu memoriju, što često nije moguće napraviti.
- Npr. pokrivanja skupova objekata veličine 1 već sadrže cijeli skup (bazu) podataka.

Algoritam Apriori algoritam za traženje frekventnih skupova objekata

Ulaz: D, σ
Izlaz: $\mathcal{F}(D, \sigma)$
$C_1 := \{\{i\} \mid i \in \mathcal{I}\}$
$k := 1$
while $C_k \neq \emptyset$ **do**
    for sve transakcije $(tid, I) \in D$ **do**
        for kandidate $X \in C_k$ **do**
            if $X \subseteq I$ **then**
                Povećaj $X.potpora$ za 1
            end if
        end for
    end for
    $\mathcal{F}_k := \{X \in C_k \mid X.potpora \geq \sigma\}$
    $C_{k+1} := \emptyset$
    for all $X, Y \in \mathcal{F}_k$, takvi da $X[i] = Y[i]$ za $1 \leq i \leq k-1$ i $X[k] < Y[k]$ **do**
        $I := X \cup \{Y[k]\}$
        if $\forall J \subset I, |J| = k : J \in \mathcal{F}_k$ **then**
            Dodaj I u C_{k+1}
        end if
    end for
    Povećaj k za 1
end while

Traženje frekventnih skupova objekata - Apriori

- Koristi svojstvo monotonosti potpore za računanje frekventnih skupova objekata.
- Pretpostavljamo da su objekti u transakcijama i skupovima sortirani u leksikografskom poretku.
- Algoritam izvodi pretraživanje u širinu (BFS) prostora svih skupova objekata.
- Iterativno generira i računa potporu skupova kandidata.
- Skup objekata je kandidat ako smo pobrojali sve njegove podskupove i svi su frekventni.
- U svakoj iteraciji generiramo kolekciju C_{k+1}, kandidata skupova objekata veličine $k+1$.
- C_1 sadrži sve objekte iz $\mathcal{I}$.
- U svakoj iteraciji stvaramo nove kandidate računanjem $X \cup Y$, gdje su $X, Y \in \mathcal{F}_k$, ako i samo ako X i Y imaju identičan $k-1$ prefiks.

Traženje frekventnih skupova objekata - Apriori

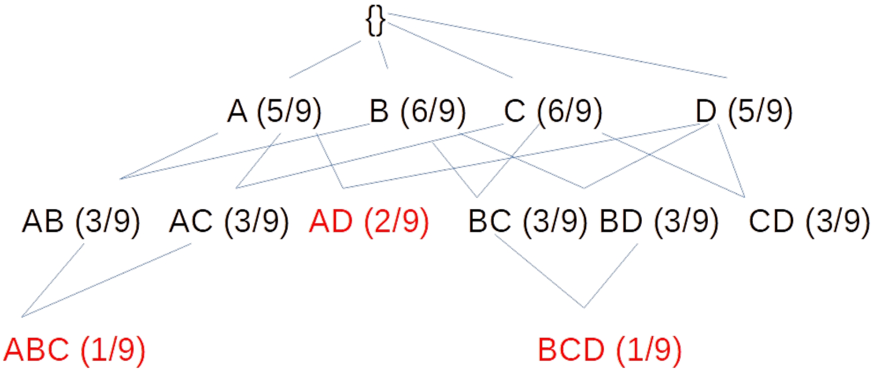
- Potpore novoizračunatih skupova objekata se računaju prolazom po bazi transakcija.
- Frekventne skupove objekata ubacujemo u F_k.
- U slučaju velikih skupova podataka, algoritam se može modificirati da iterativno računa sve kandidate jednog sloja (po djelovima). Djelovi su toliko veliki da mogu stati u memoriju.
- Nakon što se generiraju svi kandidati k-te razine, pokreće se računanje sljedeće ($k + 1$-ve) razine.
- Najveća mana algoritma je neučinkovito računanje potpora skupova objekata. Npr. pri računanju potpore svih kandidata veličine 5 prema transakcijama koje imaju maksimalno 20 objekata, trebamo računati preko 15000 provjera jednakosti skupova.
- Provjere jesu li skupovi objekata kandidata sadržani u transakcijama također može biti računalno skup korak ukoliko imamo puno kandidata skupova objekata.

Traženje frekventnih skupova objekata - Apriori

- Obično koristimo stabla hashiranja ili prefiksna stabla za učinkovito spremanje skupova objekata kandidata.

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

Pretpostavimo da je $\sigma = 3$ (30% transakcija).


{}
A (5/9)
B (6/9)
C (6/9)
D (5/9)
AB (3/9)
AC (3/9)
AD (2/9)
BC (3/9)
BD (3/9)
CD (3/9)
ABC (1/9)
BCD (1/9)

Traženje frekventnih skupova objekata - Eclat

- Ukoliko prikažemo bazu transakcija vertikalnom reprezentacijom, potporu skupova objekata možemo računati kao presjeke dva podskupa koji u uniji čine taj skup.
- Glavni problem koji susrećemo kod tog pristupa je što ne možemo uvijek držati sve generirane skupove objekata u memoriji, što je potrebno za računanje daljnjih kandidata.
- Možemo znatno reducirati količinu kandidata koje moramo držati u memoriji generiranjem kolekcija skupova kandidata koristeći pretraživanje u dubinu (eng. DFS).
- Dodatno ubrzanje možemo postići pronalaskom frekventnih objekata i 2-skupova, te primjenom pretraživanja u dubinu od tih frekventnih kandidata.

Traženje frekventnih skupova objekata - Eclat

- Za danu bazu transakcija D i prag minimalne potpore σ, označimo skup svih frekventnih skupova objekata sa identičnim prefiksom $I \subseteq \mathcal{I}$ s $F[I](D,\sigma)$. $\mathcal{F}[\{\}](D,\sigma) = \mathcal{F}(D,\sigma)$.
- Pretraga koju proučavamo će koristiti svojstva i-uvjetnih baza (D_i), gdje je $i \in \mathcal{I}$ zadani objekt. Uvjetne baze se sastoje od transakcija iz D koje sadrže objekt i, međutim svi stupci koji reprezentiraju i te objekte leksikografski manje od i brišemo iz D_i.
- Uvjetne baze možemo generalizirati i na skup objekata I. Tada se D_I sastoji od svih transakcija koje sadrže I, međutim svi stupci koji reprezentiraju objekte prije zadnjeg objekta u I (uključujući taj objekt) su obrisani iz D_I.
- Tada svakom frekventnom skupu objekata pronađenom na D_I dodajemo I. Time dobijemo frekventne skupove objekata koji sadrže I, ali ne sadrže niti jedan objekt koji se nalazi prije objekata sadržanih u I.

Traženje frekventnih skupova objekata - `Eclat`

Algoritam Eclat algoritam za traženje frekventnih skupova objekata

Ulaz: $D, \sigma, I \subseteq \mathcal{I}$ (inicijalno pozivamo s $I = \emptyset$)
Izlaz: $\mathcal{F}[I](D, \sigma)$
$\mathcal{F}[I] := \emptyset$
for sve $i \in \mathcal{I}$ koji se javljaju u D **do**
  $\mathcal{F}[I] := \mathcal{F}[I] \cup \{I \cup \{i\}\}$
  $D_i := \emptyset$
  for sve $j \in \mathcal{I}$ koji se javljaju u D takve da $j > i$ **do**
    $C :=$ Pokrivanje$(\{i\}) \cap$ Pokrivanje$(\{j\})$
    if $|C| \geq \sigma$ **then**
      $D_i := D_i \cup \{(j, C)\}$
    end if
  end for
  Računaj $\mathcal{F}[I \cup \{i\}](D_i, \sigma)$ rekurzivno   *// DFS rekurzija*
  $\mathcal{F}[I] := \mathcal{F}[I] \cup \mathcal{F}[I \cup \{i\}]$
end for

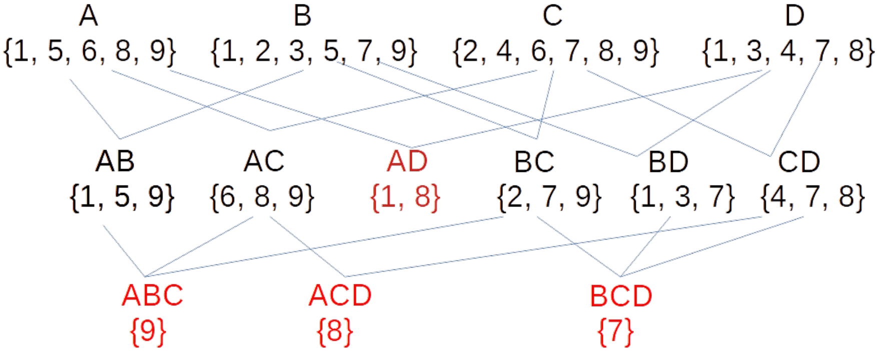
Traženje frekventnih skupova objekata - Eclat

- Pretpostavka: svi objekti u bazi transakcija su frekventni. Inače, možemo izračunati i selektirati podskup frekventnih objekata prolaskom kroz bazu podataka i brojanjem transakcija koje sadrže objekt.
- Kandidati generirani `Eclat`-om su oblika $I \cup \{i, j\}$.
- Algoritam ne koristi svojstvo monotonosti potpore nego generira skup kandidata na bazi frekvencije dva njegova podskupa objekata.
- Broj generiranih skupova objekata je znatno veći nego kod `Apriori` BFS pretrage.
- `Eclat` u principu koristi samo korak spajanja algoritma `Apriori`. Pošto radimo DFS, nemamo izračunate skupove koje `Apriori` koristi za podrezivanje.

Traženje frekventnih skupova objekata - `Eclat`

- Optimizacija koja donosi značajne dobitke u brzini izvođenja algoritma računa i sprema razlike između pokrivanja od I i pokrivanja $k-1$ prefiksa od I umjesto spremanja i računanja pokrivanja k-skupova objekata. Razlike između pokrivanja se zovu skupovi razlike (eng. *diffset*).
- Veličinu potpore skupa objekata I tada možemo izračunati kao razliku veličine potpore nekog njegovog $k-1$ prefiksa i veličine odgovarajućeg skupa razlike.
- Skup razlike skupa objekata $I \cup \{i, j\}$ možemo računati iz skupa razlike dva njegova podskupa $I \cup \{i\}$ i $I \cup \{j\}$, uz $i < j$ kao: $\text{skupRazlike}(I \cup \{i, j\}) = \text{skupRazlike}(I \cup \{j\}) \setminus \text{skupRazlike}(I \cup \{i\})$.
- Optimizirana verzija `Eclat` algoritma koja računa skupove razlike se zove `dEclat`.
- Može se pokazati da je veličina skupova razlike, skupova objekata koji se generiraju rekurzijom prilikom DFS poziva, **znatno manja** od veličine pokrivanja istih tih skupova objekata.

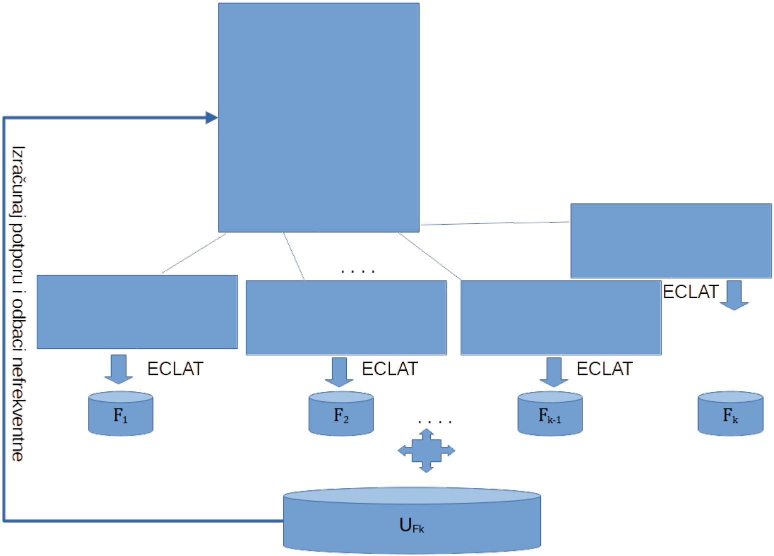
- Prikazujemo rad algoritma `Eclat` na istom primjeru kao i `Apriori` uz $\sigma = 3$.



Traženje frekventnih skupova objekata - `Partition`

- Particionira bazu transakcija na disjunktne podskupove. Svaka podbaza sadrži sve objekte, ali podskup transakcija.
- Na svakom podskupu traži frekventne skupove objekata (nekim algoritmom, npr. `Eclat`).
- Gornjim postupkom algoritam dobije nužno nadskup skupa svih frekventnih skupova objekata. Razlog tome je što je svaki frekventni skup objekata (na cijeloj bazi) ujedno **relativno frekventan** i na barem jednoj particiji.
- Skup objekata je relativno frekventan na particiji D_p, ako pokrivanje skupa objekata, pronađenih na particiji D_p, sadrži minimalno $\frac{\sigma}{|D|} \cdot |D_p|$ transakcija.
- Frekventni skupovi objekata dobiveni na svakoj particiji se spajaju u jedan skup.
- Potpora svih kandidata se evaluira na cijeloj bazi transakcija i zadržavaju se samo frekventni skupovi objekata.

Traženje frekventnih skupova objekata - Partition



Traženje frekventnih skupova objekata - `Uzorkovanje`

- Izaberi uzorak transakcija iz baze podataka.
- Na uzorku pronađi sve relativno frekventne skupove objekata.
- Izračunaj potporu relativno frekventnih skupova objekata na cijeloj bazi podataka.
- Ukoliko nisu generirani svi frekventni skupovi objekata, moguće ih je dopuniti generiranjem preostalih skupova objekata drugim prolaskom kroz bazu.
- Vjerojatnost neuspjeha detekcije frekventnih skupova na uzorku se može proizvoljno smanjiti smanjivanjem praga minimalne potpore. Međutim, takvo smanjivanje znatno povećava broj generiranih skupova objekata (kombinatorna eksplozija).
- Sve relativno frekventne skupove objekata pronađemo na uzorku, npr. korištenjem `Eclat` algoritma, izračunamo potporu tih relativno frekventnih skupova objekata na cijeloj bazi podataka, pretragu nastavimo nekim algoritmom koji pretražuje po razinama (npr. `Apriori`) i njime dovršavamo pretragu preostalih frekventnih skupova objekata.

Traženje frekventnih skupova objekata - `FP-stabla`

- Algoritam konstrukcije FP stabala je uz `Apriori` i `Eclat` jedan od najpoznatijih i najkorištenijih algoritama za traženje frekventnih objekata.
- Slično algoritmu `Eclat` izvodi pretraživanje u dubinu (eng. DFS) i stvara i-uvjetne baze D_i.
- Umjesto računanja potpore kandidata presjecima potpore podskupova skupova objekata, koristi posebnu strukturu koja se zove `FP`-stablo.
- Transakcije se spremaju u strukturu prefiksnog stabla. Umjesto da spremamo potpore svakog frekventnog skupa objekata, u prefiksno stablo spremamo razne prefikse skupova objekata i njihove potpore.
- Količina potrebne memorije za izvođenje algoritma je uglavnom znatno veća nego za izvođenje `Eclat` algoritma.
- Veću prednost pristup postiže u slučaju kada sve transakcije i-uvjetne baze sadrže identičan prefiks. Tada možemo pronaći sve frekventne skupove objekata uvjetne baze (bez tog prefiksa) i dodati sve podskupove prefiksa pronađenim frekventnim skupovima objekata uvjetne baze.

Traženje frekventnih skupova objekata - `FP-stabla`

- Podjednako efikasan pristup se može koristiti i u `Eclat` algoritmu baziran na zatvorenju skupa.

Definicija

FP-stablo je struktura podataka stabla koja:

1. Sastoji se od korjena označenog s *null*, skupa podstabala prefiksa objekata koja su djeca korjena i tablice frekventnih objekata i pokazivača na prvo pojavljivanje frekventnog objekta u *FP*-stablu.
2. Svaki čvor u podstablu prefiksa objekata se sastoji od tri polja: imena objekta, broja pojavljivanja i veze čvorova. Ime objekta reprezentira objekt, broj pojavljivanja označava broj transakcija koje reprezentira put u stablu koji dolazi do čvora, dok veza čvorova povezuje čvor sa sljedećim čvorom u *FP*-stablu koje ima isto ime objekta ili *null* ako takav čvor ne postoji.
3. Svaki element u tablici frekeventnih objekata i pokazivača se sastoji od dva polja: a) imena objekta i b) pokazivača koji pokazuje na prvi čvor *FP*-stabla koji sadrži ime objekta.

Algoritam FP-stablo

Ulaz: baza transakcija D, minimalni prag potpore σ.
Izlaz: Stablo frekventnih skupova objekata (FP-stablo) baze transakcija D.
for all objekt $o \in I$ **do**
  $F \leftarrow F \cup$ frekventan(o,D)
end for
$FList \leftarrow F.sort()$ silazno po potpori.
fptree $T \leftarrow null$
for all transakcija $tr \in D$ **do**
  Izaberi frekventne objekte iz tr i sortiraj ih u poretku određenom $FList$.
  Neka je sortirana lista $[p \mid P]$.
  Pozovi $insert\_tree([p \mid P], T)$.
end for

Algoritam insert_tree

Procedure $insert\_tree([p \mid P], T)$
if T ima dijete N takvo da $N.oznaka = p.oznaka$ **then**
  $N.brojac \leftarrow N.brojac + 1$
else
  Stvori novi čvor N s brojačem 1.
  Poveži $N.roditelj \leftarrow T$ i osvježi strukturu veza čvorova.
end if
if P nije prazan **then**
  Call $insert\_tree(P, N)$ rekurzivno.
end if

Traženje frekventnih skupova objekata - primjer izgradnje FP-stabla

- Konstrukciju stabla demonstriramo na dolje definiranoj bazi transakcija.

tid	Kupljeni predmeti
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

Traženje frekventnih skupova objekata - primjer izgradnje FP-stabla

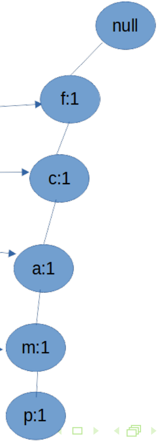
- Jednim prolaskom kroz cijelu bazu transakcija izračunamo frekventne objekte (uz $\sigma = 3$).
- U primjeru $\{(f:4),(c:4),(a:3),(b:3),(m:3),(p:3)\}$.
- Sada možemo napraviti tablicu frekventnih objekata i pokazivače na čvorove *FP*-stabla.

Objekt	Pokazivač na čvor u FP-stablu
f	$\rightarrow$ null
c	$\rightarrow$ null
a	$\rightarrow$ null
b	$\rightarrow$ null
m	$\rightarrow$ null
p	$\rightarrow$ null

Traženje frekventnih skupova objekata - primjer izgradnje FP-stabla

- Ponovo iteriramo po transakcijama baze i stvaramo *FP*-stablo.
- Čitanjem prve transakcije stvaramo prvu granu *FP*-stabla: $\{(f:1),(c:1),(a:1),(m:1),(p:1)\}$.

Objekt	Pokazivač na čvor u FP-stablu
f	
c	
a	
b	null
m	
p	

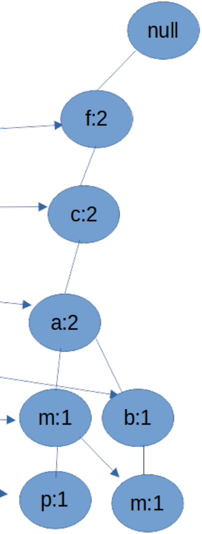


Traženje frekventnih skupova objekata - primjer izgradnje FP-stabla

- Frekventna lista objekata za drugu transakciju je $< f, c, a, b, m >$. Ona dijeli prefix $< f, c, a >$ s postojećim putem $< f, c, a, m, p >$.
- Broj pojavljivanja u svakom čvoru prefiksa inkrementiramo za 1.
- Novi čvor $(b : 1)$ se stvara i povezuje kao djete od $(a : 2)$.
- Novi čvor $(m : 1)$ se stvara i povezuje kao djete od $(b : 1)$.

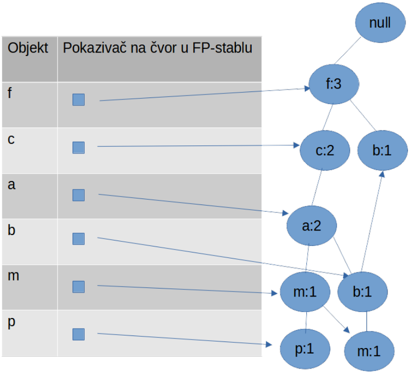
Traženje frekventnih skupova objekata - primjer izgradnje `FP-stabla`

Objekt	Pokazivač na čvor u FP-stablu
f	
c	
a	
b	
m	
p	



Traženje frekventnih skupova objekata - primjer izgradnje FP-stabla

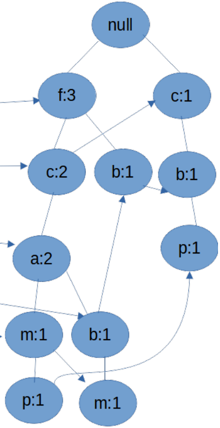
- Frekventna lista objekata treće transakcije $< f, b >$ dijeli samo čvor f s postojećim putem $< f, c, a, m, p >$.
- Inkrementiramo broj pojavljivanja f za 1.
- Stvaramo čvor $(b : 1)$ i povezujemo kao djete čvora $(f : 3)$.



Traženje frekventnih skupova objekata - primjer izgradnje `FP-stabla`

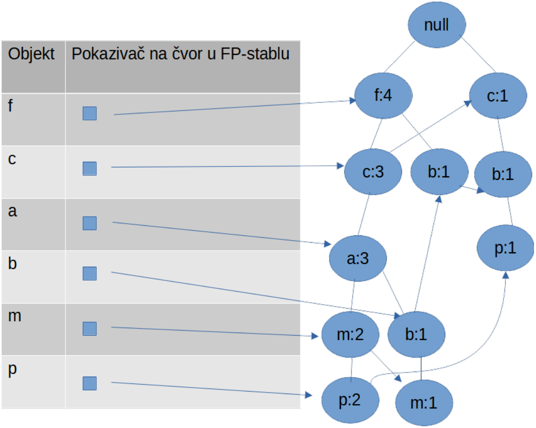
- Frekventna lista objekata četvrte transakcije je $< c, b, p >$, stoga stvaramo novu granu $\{(c : 1), (b : 1), (p : 1)\}$.

Objekt	Pokazivač na čvor u FP-stablu
f	
c	
a	
b	
m	
p	



Traženje frekventnih skupova objekata - primjer izgradnje `FP-stabla`

- Frekventna lista objekata pete transakcije je $< f, c, a, m, p >$, stoga inkrementiramo broj pojavljivanja čvorova prve grane *FP*-stabla.



Traženje frekventnih skupova objekata - svojstva `FP-stabla`

Lema

Za zadanu bazu transakcija D i prag potpore σ. Iz *FP*-stabla možemo dohvatiti potpun skup frekventnih projekcija transakcija iz baze D.

- Frekventna projekcija objekata je mapirana jednom putu *FP*-stabla svake transakcije iz baze D.
- Za put $a_1 a_2 \ldots a_k$ od korjena do čvora *FP*-stabla. Neka je c_{a_k} broj pojavljivanja čvora a_k i c'_{a_k} suma pojavljivanja djece od a_k.
- Prema konstrukciji *FP*-stabla, put stabla registrira projekciju frekventnih objekata od $c_{a_k} - c'_{a_k}$ transakcija.
- *FP*-stablo registrira potpun skup projekcija frekventnih objekata bez dupliciranja.

Traženje frekventnih skupova objekata - svojstva `FP-stabla`

Lema

Za zadanu bazu transakcija D i prag potpore σ. Bez uzimanja u obzir čvora (*null* - korjena), veličina *FP*-stabla je ograničena s $\sum_{T \in D} |frek(T)|$, a dubina stabla je ograničena s $\max_{T \in D} |frek(T)|$, gdje frek(T) označava projekciju frekventnih objekata transakcije T.

- Važna posljedica leme je da dubina i veličina *FP*-stabla ovise o broju transakcija u bazi transakcija D. Time izbjegavamo eksponencijalan rast koji susrećemo kod `Apriori` algoritma.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

Za svaki frekventni objekt a_i, svi mogući skupovi objekata koji sadrže samo frekventne objekte i frekventni objekt a_i mogu se konstruirati slijedeći poveznice čvorova a_i, počevši od pokazivača na a_i iz tablice koja sadrži pokazivače objekta na čvorove *FP*-stabla.

Svojstvo **prefiksa puta** kaže da možemo izračunati frekventne skupove objekata sa sufiksom a_i tako da akumuliramo prefikse podputeva čvorova označenih s a_i u *FP*-stablu. Broj pojavljivanja svakog čvora prefiksa puta treba biti identičan broju pojavljivanja čvora a_i u tom putu.

- Označimo čvorove na putu p s $a_1, \ldots, a_n$ u takvom poretku da je a_1 korjen prefiksnog podstabla, a_n je list tog podstabla. Neka je a_i $(1 \leq i \leq n)$ promatrani čvor.
- Za svaki prefiks čvora a_k $(1 \leq k < i)$, prefiks podputa čvora a_i u P se javlja zajedno s a_k točno onoliko puta koliko se javlja čvor a_i. Stoga svaki takav prefiks mora imati identičan broj pojavljivanja kao čvor a_i.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

- Postfiks čvor a_m ($i < m \leq n$) se također javlja zajedno s čvorom a_i. Međutim skupove objekata koji sadrže *am* generiramo u skoraku kada promatramo sufiks *am*, inače bi generirali redundantne skupove objekata. Stoga promatramo samo prefiksni podput od a_i u P.

Lema

Neka je α skup objekata u D, B je uvjetna baza skupova objekata od α (skup frekventnih skupova objekata koji kao sufiks imaju α). Neka je β skup objekata iz B. Potpora skupa objekata $\alpha \cup \beta$ u D je ekvivalentan potpori od β u B.

Tvrdnja slijedi iz definicije uvjetne baze skupova objekata.

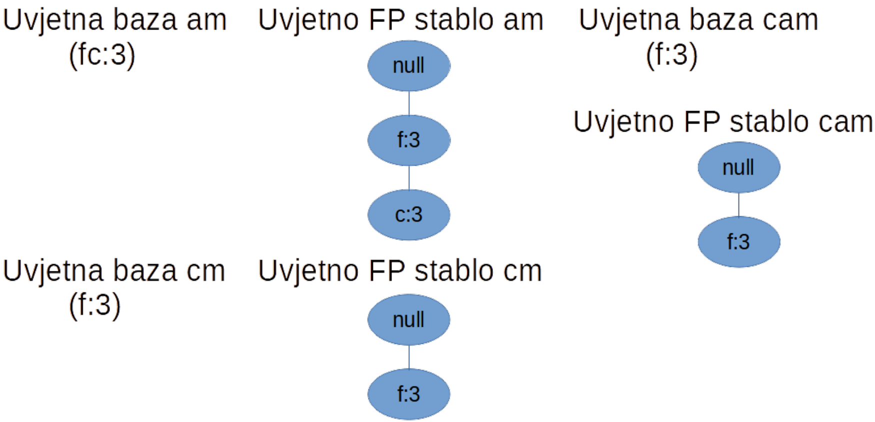
Korolar

Neka je α frekventan skup objekata u D, a B je uvjetna baza uzoraka od α. Neka je β skup objekata iz B. Tada $\alpha \cup \beta$ je frekventan u D ako i samo ako je β frekventan u B.

Traženje frekventnih skupova objekata korištenjem FP-stabla

- Čvor p odgovara frekventnom skupu objekata $p : 3$, te se javlja u dva puta u FP-stablu, $\{(fcam : 2), (cb : 1)\}$ koja čine uvjetnu bazu od p.
- Stvaranjem FP-stabla na toj uvjetnoj bazi dobijemo uvjetno FP-stablo od p.
- Uvjetno FP-stablo od p se sastoji od jednog čvora $c : 3$.
- Čvor m odgovara frekventnom skupu $(m : 3)$ i javlja se u dva puta FP-stabla, $\{(fca : 2), (fcab : 1)\}$ koja čine njegovu uvjetnu bazu.
- Uvjetno FP-stablo čvora m sadrži jedan put $< f : 3, c : 3, a : 3 >$.
- Iz tog uvjetnog FP-stabla dalje rekurzivno računamo $< f : 3, c : 3, a : 3 > |m >$. To uključuje ispitivanje tri čvora $(a), (c), (f)$.
- Računanjem $< f : 3 > |m$ dobijemo uvjetno stablo s jednim čvorom $(f : 3)$, stoga je jedini frekventni skup objekata $(fm : 3)$.

Traženje frekventnih skupova objekata korištenjem FP-stabla



Traženje frekventnih skupova objekata korištenjem `FP-stabla`

- Uvjetna baza dobivena računanjem $< a : 3|m >$ je $\{fc : 3\}$, nakon čega računamo $< f : 3, c : 3|am >$. Iz čega dobijemo $(cam : 3)$ uz uvjetnu bazu $\{(f : 3)\}$, te $(fam : 3)$. Računanjem $< f : 3 > |cam$ dobijemo $(fcam : 3)$.
- Računanjem $< c : 3|m >$ dobijemo $(cm : 3)$ i uvjetnu bazu $\{(f : 3)\}$. Računanjem $< f : 3 > |cm$ dobijemo $(fcm : 3)$.
- Familija frekventnih skupova objekata koja sadrži m je $\{(m : 3), (am : 3), (cm : 3), (fm : 3), (cam : 3), (fam : 3), (fcam : 3), (fcm : 3)\}$.
- Možemo uočiti da kod *FP*-stabala koje sadrže jedan put, sve kombinacije čvorova na putu čine frekventne skupove objekata.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

- Postupak traženja frekventnih skupova objekata se može optimizirati ukoliko *FP*-stablo ima posebnu strukturu koja se zove *FP*-stablo s jednim prefiksnim putem.
- *FP*-stablo s jednim prefiksnim pute je *FP*-stablo koje se sastoji od samo jednog puta ili jednog prefiksnog puta od korjena do prvo čvora stabla na kojem se stablo grana (sadrži više od jednog djeteta).

Lema

Pretpostavimio da se *FP*-stablo T sastoji od jednog puta P. Potpuni skup frekventnih skupova objekata od T možemo generirati enumeracijom svih kombinacija podputeva od P, gdje je potpra jednaka minimalnoj potpori čvorova sadržanih u podputu.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

- Neka je jedini put P FP-stabla $(a_1 : s_1, a_2 : s_2, \ldots, a_k : s_k)$.
- Pošto FP-stablo sadrži samo put P, potpora s_i svakog a_i $(1 \leq i \leq k)$ je potpora od a_i u ko-pojavljivanju sa prefiksom.
- Svaka kombinacija objekata na putu, npr. $(a_i, \ldots, a_j)$ $(1 \leq i, j \leq k)$ je frekventan skup objekata uz potporu ko-pojavljivanja jednakoj minimalnoj potpori sadržanih objekata.
- Pošto je svaki objekt u putu P jedinstven, ne možemo dobiti redundantne skupove objekata takvim postupkom generiranja. Također, ne možemo generirati frekventne skupove objekata izvan FP-stabla.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

Lema

Pretpostavimo da se *FP*-stablo T sastoji od jednog prefiksnog puta P i dijela s višestrukim putevima Q, koji možemo promatrati kao nezavisno *FP*-stablo, s pseudo-korjenom R. Potpuni skup frekventnih skupova objekata od T se sastoji od sljedećih djelova:

1. Frekventne skupove objekata generiramo iz P enumeracijom svih kombinacija objekata na putu P, gdje je potpora svakog skupa objekata jednaka minimalnoj potpori objekta sadržanog u skupu.
2. Generiramo frekventne skupove objekata iz Q generalnim algoritmom (uvjetne baze i *FP*-stabla) uz izmjenu $R \leftrightarrow$ null.

Traženje frekventnih skupova objekata korištenjem `FP-stabla`

Lema

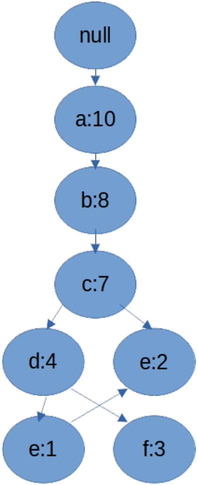
3. Familiju frekventnih skupova objekata dobijemo kombiniranjem frekventnih skupova objekata dobivenih iz P s frekventnih skupovima objekata dobivenih iz Q računanjem kartezijevog umnoška skupova $FSO(P) \times FSO(Q)$, gdje $FSO(P)$ označava familiju frekventnih skupova objekata od P. Rezultantni skupovi objekata su unija jednog frekventnog skupa objekata iz P i jednog frekventnog skupa objekta iz Q. Potpora je jednaka minimumu potpora ta dva skupa objekata.

- Prema pravilima konstrukcije FP-stabla, svaki čvor a_i na jedinom prefiksnom putu FP-stabla se javlja samo jednom u stablu.
- Jedno-prefiksni put FP-stabla formira novo FP-stablo P, a dio FP-stabla koje sadrži više putova formira drugo FP-stablo Q.
- Ova dva stabla ne dijele čvorove koji reprezentiraju isti objekt, stoga ta dva FP-stabla možemo konstruirati nezavisno.

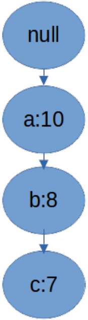
Traženje frekventnih skupova objekata korištenjem FP-stabla

- Prvo pokazujemo da je svaki skup objekata generiran iz jednog od tri moguća dijela, slijedeći pravila generiranja, različit i frekventan.
  - Prema Lemi, svaki skup objekata generiran iz P (FP-stabla koji se sastoji od jednog prefiksnog puta) je različit i frekventan.
  - Skup frekventnih skupova objekata generiranih iz Q na način da preimenujemo R u `null` je također različit i frekventan pošto tako dobivene skupove ne kombiniramo s objektima iz P.
  - Skup frekventnih objekata generiranih kombiniranjem P i Q, računanjem kartezijevog umnoška je također različit i frekventan pošto su takvi skupovi kombinacije skupova dobivenih iz P i skupova dobivenih iz Q, a svaki frekventni skup objekata generiran na P je frekventan u uvjetnoj bazi frekventnih skupova iz Q.
- Pokazujemo da ne postoje frekventni skupovi objekata koji se ne nalaze u jednom od ranije navedena tri dijela. Potpunost slijedi iz činjenice da cijelo (globalno) FP-stablo sadrži potpunu familiju frekventnih skupova objekata, a sve frekventne skupove objekata generiramo ili iz P ili iz Q ili iz njihove kombinacije.

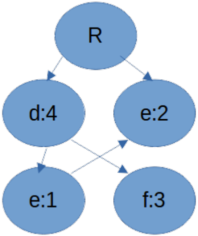
Traženje frekventnih skupova objekata korištenjem FP-stabla



Stablo koje sadrži
jedno-prefiksni put



Dio stabla koji
sadrži jedan put P



Dio stabla koji sadrži
Grananje Q

Traženje frekventnih skupova objekata korištenjem FP-stabla

- Put P generira familiju frekventnih skupova objekata $\{(a:10),(b:8),(c:7),(ab:8),(ac:7),(bc:7),(abc:7)\}$.
- Iz *FP*-stabla Q dobijemo familiju frekventnih skupova objekata $\{(d:4),(e:3),(f:3),(df:3)\}$.
- Za svaki frekventni skup objekata dobiven iz Q, možemo skup frekventnih objekata dobivenih iz P promatrati kao uvjetnu bazu.
- Računanjem kartezijevog umnoška $(d:4)$ s frekventnim skupovima objekata dobivenih iz P dobijemo $\{(ad:4),(bd:4),(cd:4),(abd:4),(acd:4),(bcd:4),(abcd:4)\}$.
- Unija frekventnih skupova objekata dobivenih iz P iz Q i iz kartezijevog umnoška P i Q čini familiju svih frekventnih skupova objekata.

Algoritam FP-growth

Ulaz: Baza D, reprezentirana FP-stablom; prag minimalne potpore σ.
Izlaz: Potpuna familija frekventnih skupova objekata.

Method: pozovi FP-growth(FP-tree, null)
Procedure FP-growth($Tree, \alpha$)
if *Tree* sadrži jedno-prefiksni put **then**
  Neka je P jedno-prefiksni put od *Tree*
  Neka je Q dio stabla s više putova, gdje najviši čvor grananja zamijenimo s korjenom `null`
  for all kombinacije β čvorova puta P **do**
    generiraj skup objekata $\beta \cup \alpha$ s $potpora =$ min potpora čvorova iz β
  end for
  Neka je $frek\_skup\_obj(P)$ familija tako generiranih skupova objekata
else
  $Q \leftarrow Tree$
end if
for all objekt $a_i \in Q$ **do**
  generiraj skup objekata $\beta = a_i \cup \alpha$ s $potpora = a_i.potpora$
  konstruiraj uvjetnu bazu od β i uvjetno FP-stablo od β, nazovimo ga $Tree_\beta$
  if $Tree_\beta \neq \emptyset$ **then**
    pozovi FP-growth($Tree_\beta, \beta$)
  end if
  Neka je $frek\_skup\_obj(Q)$ familija tako generiranih skupova objekata
end for
return $(frek\_skup\_obj(P) \cup frek\_skup\_obj(Q) \cup (frek\_skup\_obj(P) \times frek\_skup\_obj(Q)))$

Maksimalni frekventni skupovi objekata

- Pošto je kolekcija frekventnih skupova objekata zatvorena prema dolje, možemo ju reprezentirati njenim maksimalnim elementima. Takve elemente zovemo **maksimalni frekventni skupovi objekata**.
- Glavni dodaci su korištenje postupaka evaluacije budućih koraka i efikasna provjera podskupova.
- Nakon generiranja $k+1$-skupova, particioniramo ih tako da svi skupovi koji dijele isti k-prefiks pripadaju istoj particiji. Ukoliko označimo s X objekte koji čine prefiks, svaki skup objekata sadržan u particiji dodaje jedan objekt X-u.
- Označimo skup dodanih objekata s I. Kada se utvrdi da je nadskup od $X \cup I$ frekventan, skupovi objekata sadržani u particiji se mogu brisati, pošto niti jedan više ne može biti maksimalno frekventan. Ne moramo više niti brojati njihove potpore.

Maksimalni frekventni skupovi objekata

- Drugi postupak se zove **određivanje donje ograde potpore**. Nakon računanja potpore svakog kandidata (skupa objekata) $X \cup \{i\}$, moguće je izračunati donju ogradu potpore njegovih nadskupova koristeći nejednakost: $Pot(X \cup J) \geq Pot(X) - \sum_{i \in J} (Pot(X) - Pot(X \cup \{i\}))$.
- Za svaki dio sa prefiksnim skupom X, ova ograda se računa počevši od skupa J koji sadrži najfrekventniji objekt, nakon kojeg se ostali objekti dodaju u silaznom poretku po potpori sve dok objekti zadovoljavaju uvjet praga minimalne potpore.
- Konačno, $X \cup J$ se dodaje maksimalnom skupu objekata, a svi njegovi podskupovi se brišu.

Zatvoreni frekventni skupovi objekata

- Skup objekata se zove **zatvorenim** ukoliko je njegova potpora različita od potpore njegovih nadskupova.
- Iako postoji mogućnost da su svi frekventni skupovi objekata zatvoreni, u praksi se pokazuje da puno frekventnih skupova nisu zatvoreni skupovi.
- Glavni postupak podrezivanja podrazumijeva provjeru je li potpora generiranog skupa objekata jednaka potpori nekog njegovog podskupa. Ukoliko je to slučaj, novi objekt iz generiranog skupa možemo dodati svim frekventnim nadskupovima tog podskupa, a generirani skup više ne moramo razmatrati pošto on ne može biti zatvoren.
- Efikasni postupci provjere potpore su nužni za provjeru postoji li neki zatvoreni nadskup s istom potporom kao skup objekata koji smo trenutno generirali. Ukoliko postoji, ažuriramo objekte i generirani podskup ne moramo razmatrati.

Različiti objekti baze

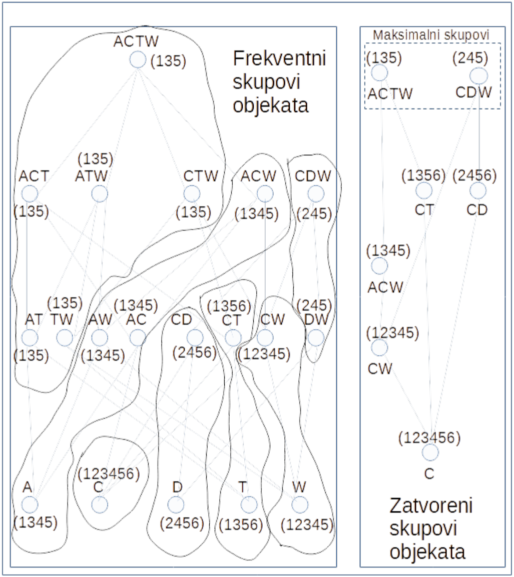
Jane Austin	Agatha Christie	Sir Arthur Conan Doyle	Mark Twain	P.G. Wodehouse
A	C	D	T	W

Baza transakcija

tid	Objekti
1	ACTW
2	CDW
3	ACTW
4	ACDW
5	ACDTW
6	CDT

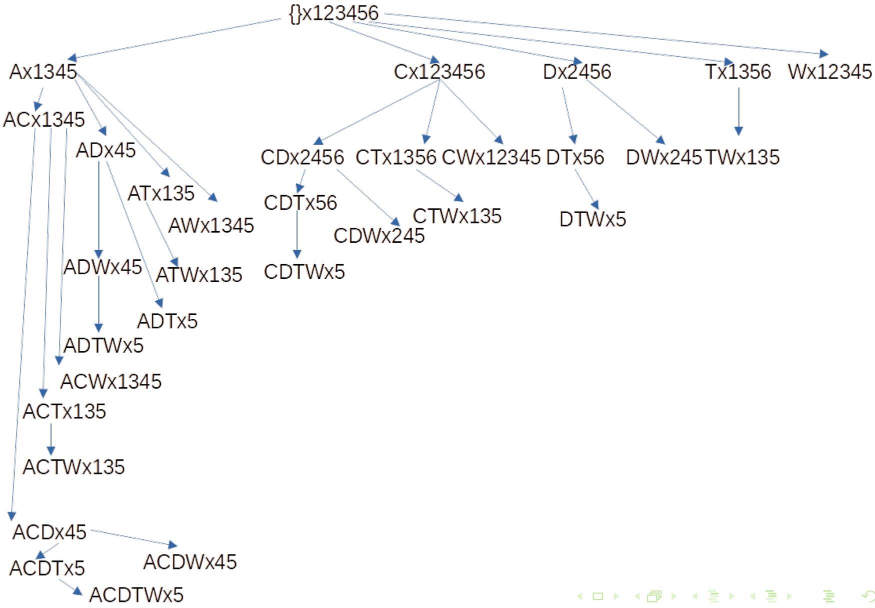
Svi frekventni skupovi objekata min. frek. = 0.5

Frek. (Pot.)	Skup objekata
1 (6)	C
0.83 (5)	W, CW
0.67 (4)	A, D, T, AC, AW, CD, CT, ACW
0.5 (3)	AT, DW, TW, ACT, ATW, CDW, CTW, ACTW



- Označimo skup objekata s $\mathcal{I}$. $p(X, k) = X[1:k]$ označava prefiks duljine k od X.
- Θ_k jer relacija ekvivalencije bazirana na gore definiranom prefiksu. $\forall X, Y \subseteq \mathcal{I},\ X \equiv_{\Theta_k} X \Leftrightarrow p(X, k) = p(Y, k)$. Dva skupa objekata su u istoj klasi ekvivalencije ukoliko dijele prefiks duljine k.
- Algoritam `CHARM` pretražuje zatvorene frekventne skupove objekata korištenjem posebne vrste stabla koja kao čvorove sadrži parove (skup objekata, skup transakcija). Ti parovi se često zapisuju u obliku $X \times t(X)$ te čine klasu ekvivalencije s obzirom na svojstvo dijeljenja prefiksa duljine k.
- Sva djeca nekog čvora X pripadaju njegovoj klasi ekvivalencije.
- Prednost korištenja klasa ekvivalencije je da originalni problem dijelimo na manje neovisne pod-probleme.
- Za svako dijete nekog čvora X, možemo pod-stablo s korijenom u djetetu promatrati kao poseban problem. Nakon pronalaska skupova objekata, svima dodamo prefiks X.

Zatvoreni frekventni skupovi objekata - Charm



Zatvoreni frekventni skupovi objekata - Charm

- Traženje frekventnih skupova objekta je jednostavno korištenjem stabla koje sadrži čvorove s parovima skupova objekata i skupova transakcija.
- Za zadani čvor ili klasu prefiksa, računamo presjeke skupova transakcija svih parova skupova objekata koji pripadaju istoj klasi, te provjeravamo prelazi li njihova potpora prag minimalne potpore.
- Računanje potpore radimo simultano s generiranjem skupova objekata.
- Koristi se operacija zatvorenja za provjeru je li neki skup objekata zatvoren ili nije.
- Zatvorenje skupa objekata X, u oznaci $z(X)$, je najmanji zatvoreni skup koji sadrži X.
- Sa $o(Y)$ označavamo skup objekata koji je zajednički svim transakcijama iz Y, a s $t(X)$ označavamo skup identifikatora transakcija koje sadrže sve objekte iz X.
- Zatvorenje skupa objekata X računamo tako da prvo izračunamo $t(X)$, zatim $o(t(x))$. Rezultatni skup objekata **mora biti zatvoren**.
- X je zatvoren ako i samo ako $X = z(X)$.

Zatvoreni frekventni skupovi objekata - Charm

- Skup objekata ACW je zatvoren pošto $z(ACW) = o(t(ACW)) = o(1345) = ACW$.
- Definiramo $f : P(\mathcal{I}) \mapsto N$ kao mapiranje koje pridijeljuje skupu objekata neki cijeli broj.
- Za svaka dva čvora u stablu skupova objekata i skupova transakcija, npr. $X_i \times t(X_i)$, $X_j \times t(X_j)$, ako $X_i \subseteq X_j$ tada $t(X_j) \subseteq t(X_i)$.
- Možemo definirati uređaj, $X_i \leq_f X_j$ ako i samo ako $f(X_i) \leq f(X_j)$.
- f može biti npr. leksikografski uređaj ili može odgovarati potpori skupova objekata.
- `CHARM` koristi četiri glavna svojstva čvorova stabla za brzo pronalaženje zatvorenih skupova.

Zatvoreni frekventni skupovi objekata - Charm

Teorem

Neka su $X_i \times t(X_i)$ i $X_j \times t(X_j)$ bilo koja dva člana klase $[P]$, s $X_i \leq_f X_j$, gdje je f potpuni uređaj. Tada vrijede sljedeća četiri svojstva:

1. Ako vrijedi $t(X_i) = t(X_j)$, tada $z(X_i) = z(X_j) = z(X_i \cup X_j)$.
2. Ako $t(X_i) \subset t(X_j)$ tada $z(X_i) \neq z(X_j)$, ali $z(X_i) = z(X_i \cup X_j)$.
3. Ako $t(X_i) \supset t(X_j)$ tada $z(X_i) \neq z(X_j)$, ali $z(X_j) = z(X_i \cup X_j)$.
4. Ako $t(X_i) \neq t(X_j)$ tada $z(X_i) \neq z(X_j) \neq z(X_i \cup X_j)$.

1. Iz $t(X_i) = t(X_j)$ slijedi $o(t(X_i)) = o(t(X_j))$, stoga $z(X_i) = z(X_j)$. Također slijedi $t(X_i \cup X_j) = t(X_i)$ pa i $o(t(X_i \cup X_j)) = o(t(X_i))$, što daje $z(X_i \cup X_j) = z(X_i)$.
2. Ako $t(X_i) \subset t(X_j)$, tada $t(X_i \cup X_j) = t(X_i) \cap t(X_j) = t(X_i) \neq t(X_j)$. Iz čega slijedi $z(X_i \cup X_j) = z(X_i) \neq z(X_j)$.
3. Analogno 2.
4. Ako $t(X_i) \neq t(X_j)$ tada $t(X_i \cup X_j) = t(X_i) \cap t(X_j) \neq t(X_i) \neq t(X_j)$, što daje $z(X_i \cup X_j) \neq z(X_i) \neq z(X_j)$.

Zatvoreni frekventni skupovi objekata - Charm

CHARM $(\mathcal{D}, \sigma)$:
$[P] = \{X_i \times t(X_i) : X_i \in \mathcal{I} \wedge Pot(X_i) \geq \sigma\}$
CHARM-EXTEND $([P], \mathcal{C} = \emptyset)$
return $\mathcal{C}$

CHARM-EXTEND $([P], \mathcal{C})$:
for svaki $X_i \times t(X_i)$ iz $[P]$ **do**
    $[P_i] = \emptyset$ i $\mathrm{X} = X_i$
    for svaki $X_j \times t(X_j)$ u $[P]$ s $X_j \geq_f X_i$ **do**
        $\mathrm{X} = \mathrm{X} \cup X_j$, $\mathrm{Y} = t(X_i) \cap t(X_j)$
        CHARM-PROPERTY$([P], [P_i])$
    end for
    if $[P_i] \neq \emptyset$ **then**
        CHARM-EXTEND $([P_i], \mathcal{C})$
    end if
    delete $[P_i]$
    $\mathcal{C} = \mathcal{C} \cup \mathrm{X}$
end for

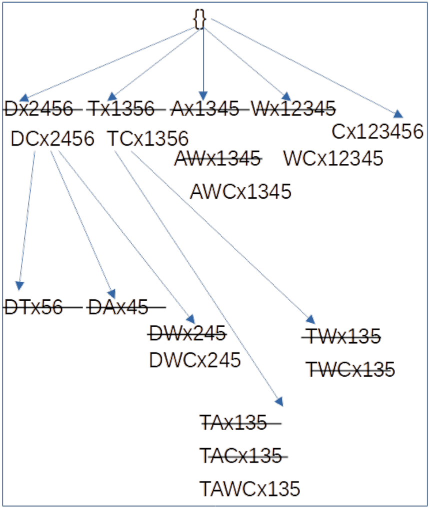
CHARM-PROPERTY $([P], [P_i])$:
if $(Pot(\mathrm{X}) \geq \sigma)$ **then**
    if $t(X_i) = t(X_j)$ **then**
        Briši X_j iz $[P]$
        Zamijeni sve X_i s X
    else if $t(X_i) \subset t(X_j)$ **then**
        Zamijeni sve X_i s X
    else if $t(X_i) \supset t(X_j)$ **then**
        Briši X_j iz $[P]$
        Dodaj $\mathrm{X} \times \mathrm{Y}$ u $[P_i]$
    else if $t(X_i) \neq t(X_j)$ **then**
        Dodaj $\mathrm{X} \times \mathrm{Y}$ u $[P_i]$
    end if
end if

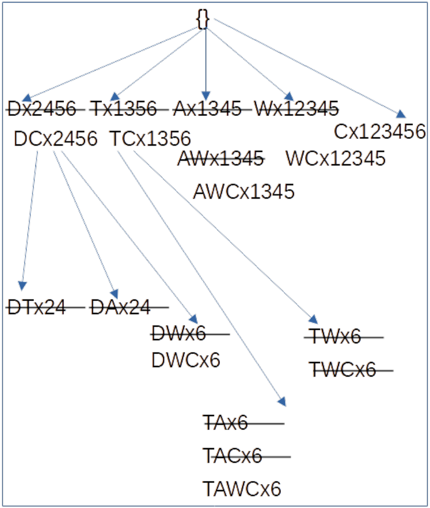
- Za dva skupa objekata X_i i X_j kažemo da X_i **obuhvaća** X_j ako i samo ako $X_j \subset X_i$ i $Pot(X_i) = Pot(X_j)$.
- Ukoliko za neki skup objekata X_j postoji skup objekata koji ga obuhvaća, tada X_j ne može biti zatvoren.
- Ovo svojstvo u kombinaciji s hashiranjem možemo koristiti za brzu eliminaciju kandidata prilikom pretrage.
- Hash funkcija koju koristi `CHARM` se definira kao: $h(X) = \sum\limits_{T \in t(X)} T$.
- Provjera generiranog skupa objekata se sastoji od računanja hash funkcije, provjere potpore skupa objekta sa svim zatvorenim skupovima objekata koji imaju istu vrijednost hash funkcije, te računanje provjere $X \subset C$, za zatvorene skupove s istom vrijednosti hash funkcije. Ukoliko postoji C t.d $X \subset C$, X odbacujemo jer je obuhvaćen.

Zatvoreni frekventni skupovi objekata - Charm

- `CHARM` može koristiti i skupove razlike pokrivanja skupova objekata za traženje zatvorenih skupova objekata.
- Označimo s $m(X_i), m(X_j)$ brojeve neslaganja u skupovima razlike $d(X_i)$ i $d(X_j)$. Svojstva iz teorema detektiramo na sljedeći način:
  - $m(X_i) = 0$ i $m(X_j) = 0$, tada $d(X_i) = d(X_j)$ ili $t(X_i) = t(X_j)$ — **Svojstvo 1**
  - $m(X_i) > 0$ i $m(X_j) = 0$, tada $d(X_i) \supset d(X_j)$ ili $t(X_i) \subset t(X_j)$ — **Svojstvo 2**
  - $m(X_i) = 0$ i $m(X_j) > 0$, tada $d(X_i) \subset d(X_j)$ ili $t(X_i) \supset t(X_j)$ — **Svojstvo 3**
  - $m(X_i) > 0$ i $m(X_j) > 0$, tada $d(X_i) \neq d(X_j)$ ili $t(X_i) \neq t(X_j)$ — **Svojstvo 4**

Zatvoreni frekventni skupovi objekata - Charm





Ne izvodljivi frekventni skupovi objekata

- Moguće je konstruirati puno bolje ograde na potpore skupova kandidata I korištenjem principa uključivanja isključivanja uz potpore svih podskupova od I.
- Za svaki podskup $J \subseteq I$ računamo donju ili gornju ogradu podtpore od I korištenjem formula:

$$Pot(I) \leq \sum_{J \subseteq X} (-1)^{|I \setminus X|+1} Pot(X)$$

, za neparni $|I \setminus J|$.

$$Pot(I) \geq \sum_{J \subseteq X} (-1)^{|I \setminus X|+1} Pot(X)$$

, za paran $|I \setminus J|$.

Ne izvodivi frekventni skupovi objekata

- Kada je najmanja gornja ograda manja od paraga minimalne potpore, više ne moramo računati potporu tog skupa. Ukoliko je najveća donja ograda jednaka najmanjoj gornjoj ogradi potpore skupa, tada također više ne moramo računati potporu tog skupa pošto su te ograde jednake stvarnoj potpori.
- Takav skup se zove **izvodiv** pošto se njegova potpora može izvesti iz potpore njegovih podskupova. Inače se skup zove **ne izvodiv**.
- Ne izvodiva kolekcija frekventnih skupova objekata je zatvorena prema dolje. Svaki podskup ne izvodivog skupa je također ne izvodiv.
- Veličina najvećeg ne izvodivog skupa je najviše $1 + \log |D|$, gdje $|D|$ označava ukupan broj transakcija u bazi.
- Ima smisla generirati samo ne izvodive frekventne skupove pošto izvodivi skupovi ne daju nove informacije o bazi.

Traženje skupova objekata - važniji alati

Softver za kreiranje frekventnih skupova objekata:

- Java - **SPMF biblioteka,** `https://www.philippe-fournier-viger.com/spmf/index.php?link=algorithms.php` - kolekcija algoritama (`Apriori`, `Eclat`, `FP-growth`, `Charm`).