Gyakori elemhalmazok, asszociációs szabályok

Gyakori elemhalmazok

Alapprobléma

Sok üzleti vállalkozás nagy mennyiségű adatot halmoz fel a mindennapi működése során. Az élelmiszerboltok pénztárainál például minden egyes nap óriási mennyiségű vásárlói adat gyűlik össze.
A kiskereskedők számára érdekes lehet ezen adatok vizsgálata, ha többet akarnak megtudni a vevőik vásárlói szokásairól.
Az ilyen értékes információnak számos üzleti felhasználása lehet (pl. reklámok készítése, leltárvezetés, a vásárlói kapcsolattartás menedzselésének segítése).

Gyakori elemhalmazok

A különféle adatbázisokban gyakran együttesen előforduló jellemzők ismerete hasznos lehet

Marketing, adatbányászat

TranzakcióID	Termékek
1	{tej, kenyér, Hertz szalámi}
2	{sör, pelenka}
3	{sör, virsli}
4	{sör, bébiétel, pelenka}
5	{pelenka, kóla, kenyér}

Vásárlói kosarak megjelenési formái – Horizontális tárolás

TranzakcióID	Termékek
1	{tej, kenyér, Hertz szalámi}
2	{sör, pelenka}
3	{sör, virsli}
4	{sör, bébiétel, pelenka}
5	{pelenka, kóla, kenyér}

Vásárlói kosarak megjelenési formái – Vertikális/invertált tárolás

Termék	Kosarak
tej	{1}
kenyér	{1,5}
Hertz szalámi	{1}
sör	{2,3,4}
pelenka	{2,4,5}
virsli	{3}
bébiétel	{4}
kóla	{5}

Vásárlói kosarak megjelenési formái – Relációs tárolás

Kosár	Termék
1	tej
1	kenyér
1	Hertz szalámi
2	sör
2	pelenka
3	sör
3	virsli
4	sör
4	bébiétel
4	pelenka
5	pelenka
5	kóla
5	kenyér

Gyakori elemhalmazok

Az asszociációs elemzés más területekre is alkalmazható:
pl. bioinformatika, orvosi diagnosztizálás, web-bányászat, illetve tudományos adatok vizsgálata.
A földtudományi adatok elemzése során az asszociációs mintázatok érdekes összefüggéseket fedhetnek fel az óceáni, a szárazföldi és a légköri folyamatok között. Ezen információk segítségével a tudósok jobban megérthetik a Föld természeti erői közti kölcsönhatásokat.

Gyakori elemhalmazok - Plágiumellenőrzés

Intuíciónkkal ellentétben legyenek a kosaraink mondatok, a termékek pedig a dokumentumok

Konklúzió

A"termékek" alkotta "kosarak" és a "tartalmazás" fogalmát rugalmasan kell kezelni

Termékek közötti függést vizsgáljuk, nem a kosarak közötti hasonlóságot

Milyen jelentést társíthatunk a gyakori elemhalmazoknak/ asszociációs szabályoknak, ha a kosarak a dokumentumok, és a termékek a mondatok?

Gyakori elemhalmazok - Adatbányászat

Asszociációs szabályok segítségével kezdetleges osztályozót építhetünk

Hiányzó adatokat becsülhetünk meg a változók gyakori együttállásából

Változókat vonhatunk össze, ha azok túlzott (teljes) egyezést mutatnak

Gyakori elemhalmazok

Az asszociációs elemzés esetén két kulcsfontosságú szempontot kell figyelembe venni. Először is egy nagyméretű tranzakciós adathalmazban a mintázatok megtalálása jelentős számítási költséggel járhat. Másodsorban a felfedezett mintázatok egy része potenciálisan félrevezető, ugyanis elképzelhető, hogy csak véletlenül fordulnak elő.

az asszociációs elemzés alapfogalmai és hatékony mintázatkereső algoritmus bemutatása
a felfedezett mintázatok értékelése, melynek célja a hibás eredmények előállításának a megelőzése

Asszociációs szabályok legyűjtése

Cél: T tranzakciós adatbázis jellemzőinek (termékeinek) azon diszjunkt részhalmazainak meghatározása, melyek együttes *támogatottsága* (support) meghalad egy gyakorisági küszöbértéket, továbbá a kettőjük között (bármely irányban) fennálló implikációt jellemző *bizonyosság* (confidence) magas

X halmaz **támogatottság**a: $s(X) = |\{t_i | t_i \in T \wedge X \subseteq t_i\}|$

Relatív támogatottság: $s(X)$ normalizálása a tranzakciós adatbázis sorainak számával

$A \rightarrow B$ szabály **bizonyosság** a $c(A \rightarrow B) = \dfrac{s(A \cup B)}{s(A)} \approx P(B|A)$

$A \rightarrow B$ szabály jelentése: amennyiben egy kosárban benne vannak A elemei, úgy az valószínű B-t is tartalmazni fogja

Asszociációs szabályok érdekessége

Minden gyakorisági és bizonyossági küszöböt meghaladó szabály érdekel bennünket?

Egy szabály bizonyossága egyszerűen attól is lehet magas, hogy a szabály jobb oldalán lévő elem a bal oldalon állóktól függetlenül gyakori.

$A \rightarrow B$ szabály érdekessége = $c(A \rightarrow B) - P(B)$

Érdekesek a magas abszolútértékkel jellemzett szabályok.

A küszöbszámok olyanok legyenek, hogy a megkapott érdekes szabályok feldogozhatók legyenek

Az aszociációs szabályok legyártásának sémája

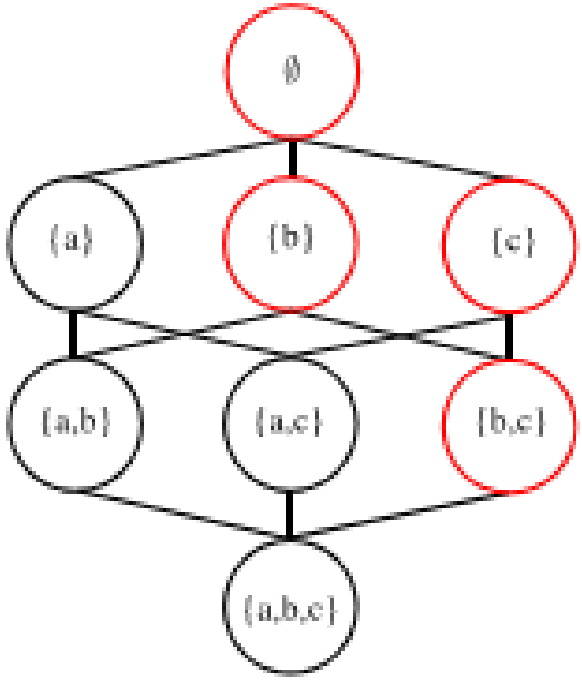
t (relatív) gyakorisági küszöböt meghaladó gyakori elemhalmazok *F* halmazának legyűjtése

a gyakori elemhalmazok particionálása (diszjunkt részhalmazokra bontása), bizonyosságuk mértékének vizsgálata

Apriori elv

Az I elemhalmaz gyakori $\Rightarrow \forall J \subseteq I$ elemhalmaz gyakori

Anti-monoton tulajdonság miatt: f függvény anti-monoton, ha $\forall X, Y \in P(U)$: $X \subseteq Y \Rightarrow f(X) \geq f(Y)$



Ha egy mérték rendelkezik az antimonoton tulajdonsággal, akkor közvetlenül beépíthető az adat-bányászati algoritmusokba az elemhalmazjelöltek exponenciális keresési terének hatékony nyesése érdekében.

Az Apriori elv érvényesülése

Legyen 3 a gyakorisági küszöbérték

Termék	Gyakoriság
sör	3
kenyér	4
kóla	2
pelenka	4
tej	4
virsli	1

Az Apriori elv érvényesülése

Legyen 3 a gyakorisági küszöbérték

Termék	Gyakoriság
sör	3
kenyér	4
kóla	2
pelenka	4
tej	4
virsli	1

Termékek	Gyakoriság
{kenyér,sör}	2
{pelenka,sör}	3
{sör,tej}	2
{kenyér,pelenka}	3
{kenyér,tej}	3
{pelenka,tej}	3

Gyakori elemhalmazok meghatározása

Algoritmus1 Gyakori elemhalmazok számításának pszeudokódja

Input: U termékuniverzum, T tranzakciós adatbázis, t gyakorisági küszöbérték

Output: gyakori elemhalmazok

1: $C_1 := U$

2: Számítsuk ki C_1 elemeinek támogatottságait

3: $F_1 := \{x | x \in C_1 \wedge s(x) \geq t\}$

4: **for** (k=2; k<$|U|$ && $F_{k-1} \neq \emptyset$; k++) **do**

5: C_k meghatározása F_{k-1} elemeiből

6: C_k elemeinek támogatottságának kiszámítása

7: $F_k := \{X | X \in C_k \wedge s(x) \geq t\}$

8: **endfor**

9: **return** $\cup_{i=1}^{k} F_i$

Gyakori elemhalmazok meghatározása

Az algoritmus először végigmegy az adathalmazon és megállapítja az elemek támogatottságát. Ezen lépés végére ismert lesz az összes gyakori 1-elemhalmaz. Az algoritmus ezután új k-elemhalmaz jelölteket állít elő iteratív módon az előző iterációban talált gyakori (k−1)-elemhalmazokból. A jelöltek támogatottsági értékének megállapításához az algoritmusnak ismét végig kell mennie az adathalmazon. A részhalmaz függvény mindazon elemhalmaz-jelölteket adja vissza, melyek szerepelnek az egyes t tranzakciók-ban. A támogatottsági értékek kiszámítása után az algoritmus kizárja az összes olyan elemhalmazjelöltet, amelyeknek a támogatottsága kisebb, mint a küszöbérték.
Az algoritmus akkor áll le, amikor már nem tud több gyakori elemhalmazt előállítani, azaz $F_k = \emptyset$.

Az Apriori algoritmus

Az Apriori algoritmus gyakori elemhalmazokat előállító részének két fontos jellemzője van:

szintenkénti algoritmus, ugyanis az elemhalmazhálót szintenként járja be, vagyis a gyakori 1-elemhalmazoktól folyamatosan halad a leghosszabb gyakori elemhalmazokig.

generál-és-tesztel stratégiát használ a gyakori elemhalmazok megtalálásához. Az új elemhalmaz jelölteket minden egyes iterációban az előző lépésben talált gyakori elemhalmazokból állítja elő. Az algoritmus ezután kiszámolja a jelöltek támogatottságát és ezt összeveti a küszöbértékkel. Az algoritmusnak összesen $k_{max}+1$ iterációra van szüksége, ahol k_{max} a legnagyobb gyakori elemhalmaz méretét jelöli.

C_k halmazok hatékony számítási módja

F_1 elemeire támaszkodva nem hatékony

F_{k-1} és F_1 elemeinek kombinálásából.

F_{k-1} és F_{k-1} elemeinek kombinálásából: előnye, hogy nem generál fölöslegesen többször elemhalmazokat, csak akkor von össze, ha

$$\exists\, A=\{a_1,a_2,\ldots,a_{k-1}\} \in F_{k-1},\ B=\{b_1,b_2,\ldots,b_{k-1}\} \in F_{k-1}:$$

$$\forall\, a_i = b_i,\ 1 \le i \le k-2 \wedge\ a_{k-1} \neq b_{k-1}$$

F elemein megéri értelmezni egy rendezést (pl. egészekké leképezni őket), és így tárolni őket

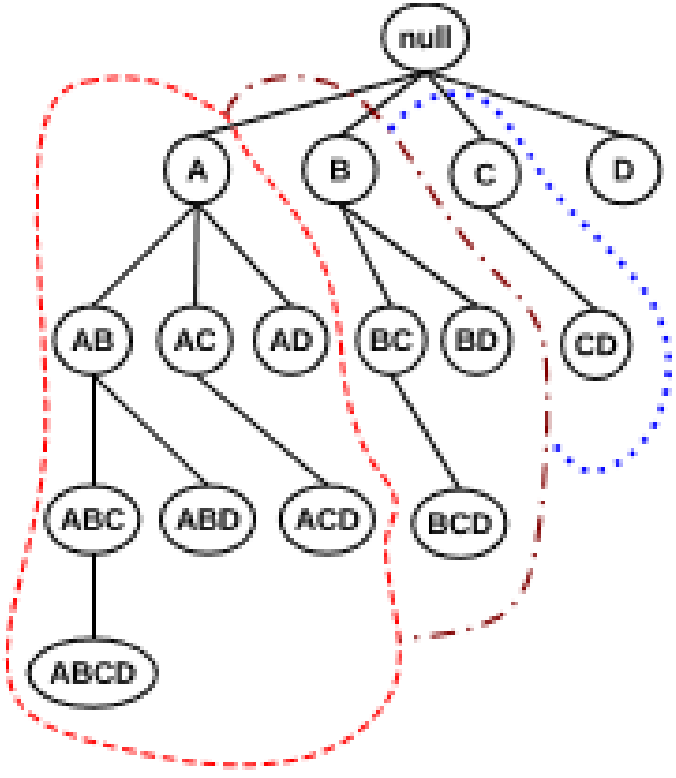
Apriori általánosítása – prefix és szuffix szemlélet

A háló csúcsait előbb diszjunkt csoportokba (ekvivalencia-osztályokba) választjuk szét, majd a gyakori elemhalmazokat előbb egy bizonyos ekvivalencia-osztályban keresi meg, majd innen továbblép egy másik ekvivalencia-osztályra.
Apriori algoritmus szintenkénti stratégiája: a hálót az elemhalmazok mérete alapján osztjuk fel (előbb a gyakori 1-elemhalmazok, majd a nagyobb méretű elemhalmazok).

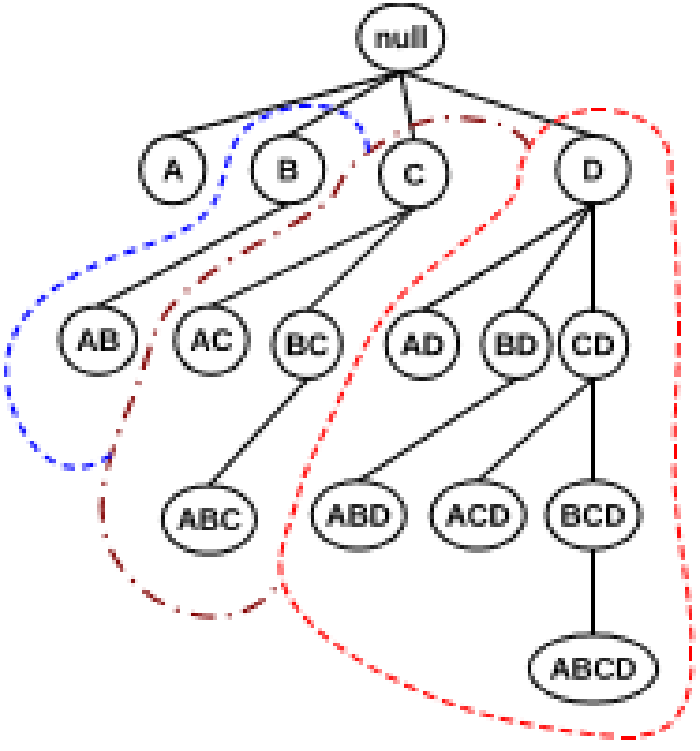
Apriori általánosítása – prefix és szuffix szemlélet

Az ekvivalencia-osztályokat az elemhalmazok elő- és utótagjai szerint is definiálhatjuk. Ebben az esetben két elemhalmaz egyazon ekvivalencia-osztályba tartozik, ha az elő- vagy utótagjuk *k* hosszúságú része megegyezik. Az előtag-alapú megközelítés esetén az algoritmus először az *a* előtaggal rendelkező gyakori elemhalmazokat keresi meg. Ezt követik a *b* előtagú gyakori elemhalmazok, majd a *c* előtagúak, stb. Mind az előtag-alapú (prefix), mind az utótag-alapú (suffix) ekvivalencia-osztályokat egy fa-struktúrával tudjuk szemléltetni.

Apriori általánosítása – prefix és szuffix szemlélet



(a) Prefix fa



(b) Szuffix fa

Adatbányászat - Gyakori elemhalmazok, asszociációs és döntési szabályok című tananyag alapján készült (részben)