Κανόνες Συσχέτισης Ι

Οι διαφάνειες στηρίζονται στο **P.-N. Tan, M.Steinbach, V. Kumar, «Introduction to Data Mining»**, Addison Wesley, 2006

Εισαγωγή

Market-Basket transactions
(Το καλάθι της νοικοκυράς!)

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

δοσοληψία

- Προώθηση προϊόντων
- Τοποθέτηση προϊόντων στα ράφια
- Διαχείριση αποθεμάτων

Το πρόβλημα: Δεδομένου ενός συνόλου δοσοληψιών (transactions), βρες κανόνες που προβλέπουν την εμφάνιση ενός στοιχείου (item) με βάση την εμφάνιση άλλων στοιχείων στις συναλλαγές

Παραδείγματα κανόνων συσχέτισης

{Diaper} → {Beer},
{Milk, Bread} → {Eggs, Coke},
{Beer, Bread} → {Milk}

Σημαίνει ότι εμφανίζονται μαζί, όχι ότι η εμφάνιση του ενός είναι η αιτία της εμφάνισης του άλλου (co-occurrence, not causality όχι έννοια χρόνου ή διάταξης)

Εισαγωγή

Δυαδική αναπαράσταση

Γραμμές: δοσοληψίες

Στήλες: Στοιχεία

1 αν το στοιχείο εμφανίζεται στη σχετική δοσοληψία

Μη συμμετρική δυαδική μεταβλητή (1 πιο σημαντικό από το 0)

- *Ένας περιορισμός είναι ότι χάνουμε πληροφορία για τις ποσότητες*

Παράδειγμα

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ορισμοί

- $I = \{i_1, i_2, .., i_k\}$ ένα σύνολο από διακριτά **στοιχεία (items)**
Παράδειγμα: {Bread, Milk, Diapers, Beer, Eggs, Coke}

- **Στοιχειοσύνολο (Itemset):** Ένα υποσύνολο του I
Παράδειγμα: {Milk, Bread, Diaper}

  - **k-στοιχειοσύνολο (k-itemset):** ένα στοιχειοσύνολο με *k* στοιχεία

- $T = \{t_1, t_2, .., t_N\}$ ένα σύνολο από **δοσοληψίες**, όπου κάθε t_i είναι ένα στοιχειοσύνολο

Πλάτος (width) δοσοληψίας: αριθμός στοιχείων

t_i **περιέχει** ένα στοιχειοσύνολο X, αν το X είναι υποσύνολο της t_i

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Ορισμοί

- **support count (σ) ενός στοιχειοσυνόλου**
Η συχνότητα εμφάνισης του στοιχειοσυνόλου

Παράδειγμα: σ({Milk, Bread, Diaper}) = 2
$\sigma(X) = \{ t_i \mid X \subseteq t_i, t_i \in T\}$

- **Υποστήριξη (Support (s)) ενός στοιχειοσυνόλου**
Το ποσοστό των δοσοληψιών που περιέχουν ένα στοιχειοσύνολο

Παράδειγμα: s({Milk, Bread, Diaper}) = 2/5

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Frequent Itemset – Συχνό Στοιχειοσύνολο**
Ένα στοιχειοσύνολο του οποίου η υποστήριξη είναι μεγαλύτερη ή ίση από κάποια τιμή κατωφλίου *minsup*

Ορισμοί

Κανόνας Συσχέτισης (Association Rule)
Είναι μια έκφραση της μορφής $X \rightarrow Y$,
όπου X και Y είναι στοιχειοσύνολα
$X \subseteq I, Y \subseteq I, X \cap Y = \varnothing$

Παράδειγμα: {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Υποστήριξη Κανόνα Support (s)
Το ποσοστό των δοσοληψιών που περιέχουν και το X και το Y ($X \cup Y$)
$\sigma(X \cup Y)/|T|$ ($|T|$ ο αριθμός των δοσοληψιών)

Εμπιστοσύνη - Confidence (c)
Πόσες από τις δοσοληψίες (ποσοστό) που περιέχουν το X περιέχουν και το Y
$\sigma(X \cup Y)/\sigma(X)$

$$s = \frac{\sigma\{\text{Milk, Diaper, Beer}\}}{|T|} = \frac{2}{5} = 0.4 \quad \{\text{Milk, Diaper}\} \rightarrow \text{Beer}$$

$$c = \frac{\sigma\{\text{Milk, Diaper, Beer}\}}{\sigma\{\text{Milk, Diaper}\}} = \frac{2}{3} = 0.67$$

Εξόρυξη Κανόνων Συσχέτισης

Παρατηρήσεις

- $s(X \rightarrow Y) = s(X \cup Y) = \sigma(X \cup Y)/N$

Ένας κανόνας με μικρή υποστήριξη μπορεί να εμφανίζεται τυχαία

Λιγότερη σημασία, γιατί αφορά μικρό αριθμό από συναλλαγές

Εξαιρεί κανόνες που δεν έχουν ενδιαφέρον

- $c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$

$c(X \rightarrow Y) = P(Y|X)$ δεσμευμένη πιθανότητα να εμφανίζεται το Y όταν εμφανίζεται το X

Εμπιστοσύνη μετρά την αξιοπιστία

Όσο μεγαλύτερη εμπιστοσύνη τόσο μεγαλύτερη η πιθανότητα εμφάνισης του Y σε κανόνες που περιέχουν το X

Εξόρυξη Κανόνων Συσχέτισης

Εύρεση Κανόνων Συσχέτισης

Είσοδος: Ένα σύνολο από δοσοληψίες T
Έξοδος: Όλοι οι κανόνες με
support ≥ *minsup*
confidence ≥ *minconf*

Εξόρυξη Κανόνων Συσχέτισης

Brute-force προσέγγιση:

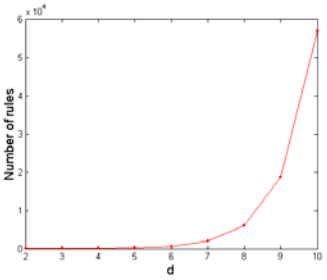
- Παρήγαγε όλους τους πιθανούς κανόνες συσχέτισης
- Υπολόγισε την υποστήριξη και την εμπιστοσύνη για τον καθένα
- Prune τους κανόνες που δεν ικανοποιούν το κατώφλι εμπιστοσύνης και υποστήριξης

⇒ Υπολογιστικά ακριβό!

Εύρεση Συχνών Στοιχειοσυνόλων

Υπολογιστική Πολυπλοκότητα

- Έστω d διαφορετικά στοιχεία:
  - Συνολικός αριθμός στοιχειοσυνόλων = 2^d (δυναμοσύνολο)
  - Συνολικός αριθμός πιθανών κανόνων συσχέτισης:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

If d = 6, R = 602 rules

Εξόρυξη Κανόνων Συσχέτισης

Μια σημαντική παρατήρηση

Η υποστήριξη ενός κανόνα $X \rightarrow Y$ εξαρτάται μόνο από την υποστήριξη του $X \cup Y$
Άρα κανόνες που ξεκινούν από το ίδιο στοιχειοσύνολο έχουν την ίδια υποστήριξη (αλλά πιθανών διαφορετική εμπιστοσύνη)

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Πιθανοί κανόνες με τα στοιχεία Milk, Diaper και Beer (στοιχειοσύνολο {Milk, Diaper, Beer})

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Αν είχαμε minsup = 0.5, θα αποκλείαμε και τους έξη κανόνες

Άρα μπορούμε να εξετάσουμε τους περιορισμούς για την υποστήριξη και την εμπιστοσύνη ξεχωριστά

Εξόρυξη Κανόνων Συσχέτισης

Χωρισμός του προβλήματος σε δύο υπο-προβλήματα:

- Εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation)
  Εύρεση όλων των στοιχειοσυνόλων με υποστήριξη ≥ minsup
- Δημιουργία Κανόνων (Rule Generation)
  Για κάθε στοιχειοσύνολο, δημιούργησε κανόνες με μεγάλη υποστήριξη, όπου κάθε κανόνες είναι μια δυαδική διαμέριση του συχνού στοιχειοσυνόλου

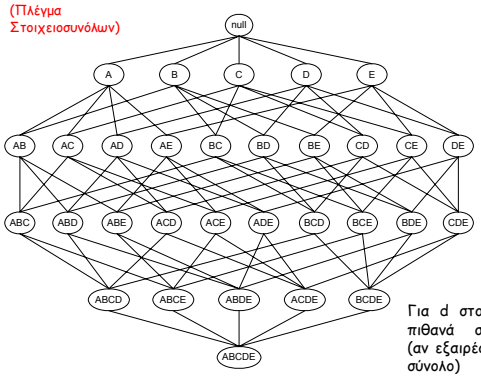
Η δημιουργία των συχνών στοιχειοσυνόλων είναι επίσης υπολογιστικά ακριβή

Εύρεση Συχνών Στοιχειοσυνόλων

Εύρεση Συχνών Στοιχειοσυνόλων

Itemset (Πλέγμα Στοιχειοσυνόλων)

Lattice



Για d στοιχεία, 2^d - 1 πιθανά στοιχειοσύνολα (αν εξαιρέσουμε το κενό σύνολο)

Εύρεση Συχνών Στοιχειοσυνόλων

Brute-force approach:

- Κάθε στοιχειοσύνολο στο πλέγμα είναι ένα υποψήφιο συχνό στοιχειοσύνολο
- Υπολόγισε την υποστήριξη κάθε υποψήφιου στοιχειοσυνόλου διατρέχοντας (scanning) τη βάση δεδομένων

Δοσοληψίες

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Κατάλογος Υποψηφίων

N

M

w

Ταίριαξε κάθε δοσοληψία με κάθε υποψήφιο

Πολυπλοκότητα ~ $O(NMw)$ => Μεγάλη γιατί $M = 2^d$!!!

N: αριθμός δοσοληψιών

w: μέγιστο πλάτος δοσοληψίας

Εύρεση Συχνών Στοιχειοσυνόλων

Διαφορετικές Στρατηγικές

Ελάττωση του αριθμού των υποψηφίων στοιχειοσυνόλων (M)

Πλήρης αναζήτηση: $M=2^d$

Χρησιμοποίησε κάποια τεχνική pruning (κλαδέματος - ελάττωσης) για να ελαττωθεί το M (πχ *apriori*)

Ελάττωση του αριθμού των δοσοληψιών (N)

Ελάττωση του μεγέθους του N καθώς το μέγεθος του στοιχειοσυνόλου αυξάνεται

(κάποιοι αλγόριθμοι βασισμένοι σε κατακερματισμό)

Ελάττωση του αριθμού των συγκρίσεων (NM)

Στόχος να αποφύγουμε να ταιριάξουμε κάθε υποψήφιο στοιχειοσύνολο με κάθε δοσοληψία

Χρήση *αποδοτικών δομών δεδομένων* για την αποθήκευση των υποψηφίων στοιχειοσυνόλων ή των δοσοληψιών

Αρχή apriori

Ελάττωση συχνών στοιχεισυνόλων

Αρχή Apriori

Αν ένα στοιχειοσύνολο είναι συχνό, τότε όλα τα υποσύνολα του είναι συχνά

Αντιθετοαντιστροφή: Αν ένα στοιχειοσύνολο δεν είναι συχνό, όλα τα υπερσύνολα του δεν είναι συχνά

Η αρχή Apriori ισχύει λόγω της παρακάτω ιδιότητας της υποστήριξης:

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

Η υποστήριξη ενός στοιχεισύνολου είναι *μικρότερη* ή *ίση* της υποστήριξης οποιουδήποτε υποσυνόλου του

Αρχή apriori

Αντι-μονότονη (anti-monotone) ιδιότητα της υποστήριξης

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

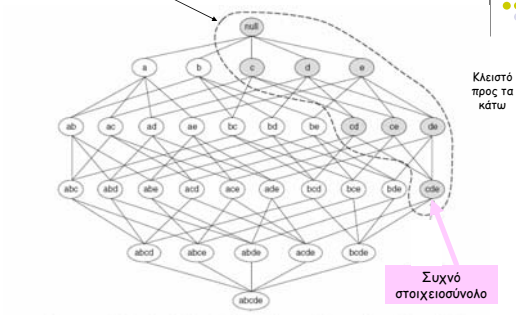
s: downwards closed

Μονότονη ιδιότητα ή upwards closed

$$\forall X, Y : (X \subseteq Y) \Rightarrow f(X) \leq f(Y)$$

Αρχή apriori

Όλα τα υποσύνολα του συχνά



Κλειστό προς τα κάτω

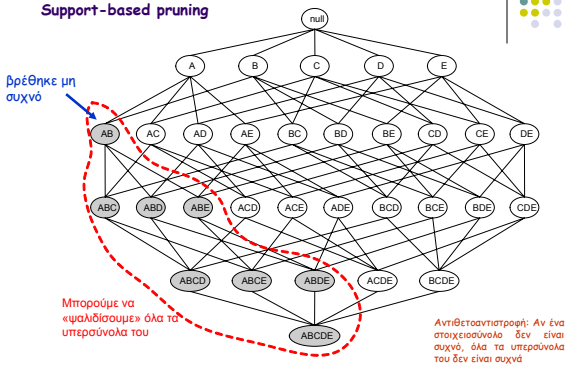
Συχνό στοιχειοσύνολο

Αν το {c, d, e} είναι συχνό, όλα τα υποσύνολα του είναι συχνά

Στρατηγική apriori

Support-based pruning

βρέθηκε μη συχνό



Μπορούμε να «ψαλιδίσουμε» όλα τα υπερσύνολα του

Αντιθετοαντιστροφή: Αν ένα στοιχειοσύνολο δεν είναι συχνό, όλα τα υπερσύνολα του δεν είναι συχνά

Στρατηγική apriori

Παράδειγμα

Minimum Support = 3

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Στοιχεία (1-στοιχειοσύνολα)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Ζεύγη (2-στοιχειοσύνολα)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

(Δε χρειάζεται να παραχθούν υποψήφιοι με Coke ή Eggs)

Τριάδες (3-στοιχειοσύνολα)

Itemset	Count
{Bread,Milk,Diaper}	3

Αν όλα τα δυνατά στοιχειοσύνολα:

$$\binom{6}{1}+\binom{6}{2}+\binom{6}{3}=6+15+20=41$$

Μετά την ελάττωση με βάση την υποστήριξη:

$$\binom{6}{1}+\binom{4}{2}+1=6+6+1=13$$

Στρατηγική apriori

Γενικός Αλγόριθμος

k = 1

Δημιούργησε όλα τα συχνά στοιχειοσύνολα μήκους 1

Repeat until δεν δημιουργούνται νέα στοιχειοσύνολα

- Δημιούργησε υποψήφια στοιχειοσύνολα μήκους (k+1) από τα συχνά στοιχειοσύνολα μήκους k
- Prune τα υποψήφια στοιχειοσύνολα που περιέχουν υποσύνολα μήκους k που δεν είναι συχνά
- Υπολόγισε την υποστήριξη (support) κάθε υποψηφίου στοιχειοσύνολου διαβάζοντας από τη βάση δεδομένων
- Σβήσε τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά, αφήνοντας μόνο τα συχνά

Στρατηγική apriori

Γενικός Αλγόριθμος

- Διατρέχει το πλέγμα ανά επίπεδο
- Generate-and-Test στρατηγική

Σε κάθε βήμα k:

Δημιουργία υποψήφιων k-στοιχειοσυνόλων με βάση τα συχνά k-1 στοιχειοσύνολα

Υπολογισμός της υποστήριξής τους και pruning όσων έχουν μικρή υποστήριξη

- k_{max} περάσματα, όπου k_{max} μέγεθος (αριθμός στοιχείων) του μεγαλύτερου στοιχειοσυνόλου

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

Σε κάθε βήμα k:

Δημιουργία υποψήφιων k-στοιχειοσυνόλων με βάση τα συχνά k-1 στοιχειοσύνολα

- Όλα τα υποσύνολα του πρέπει να είναι συχνά
- Δεν πρέπει να δημιουργούμε ένα στοιχειοσύνολο πολλές φορές
- complete – δεν πρέπει να χάνουμε κάποιο συχνό

Πως;

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

Θα δούμε δύο τρόπους:

- Μέθοδος $F_{k-1} \times F_1$
- Μέθοδος $F_{k-1} \times F_{k-1}$

Για να αποφύγουμε τη δημιουργία του ίδιου στοιχειοσυνόλου, κρατάμε κάθε στοιχειοσύνολο (λεξικογραφικά) ταξινομημένο

Και στις δύο περιπτώσεις, έλεγχος αν τα παραγόμενα στοιχειοσύνολα είναι συχνά με βάση τα υποσύνολά τους.

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

Μέθοδος $F_{k-1} \times F_1$

Επέκταση κάθε συχνού (k-1) στοιχειοσυνόλου με άλλα συχνά στοιχεία

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

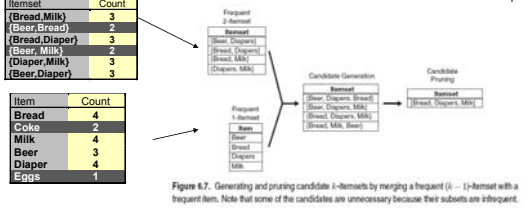


Figure 6.7. Generating and pruning candidate k-itemsets by merging a frequent (k − 1)-itemset with a frequent item. Note that some of the candidates are unnecessary because their subsets are infrequent.

Κάθε στοιχειοσύνολο (λεξικογραφικά) ταξινομημένο – κάθε (k-1) συχνό στοιχεισύνολο επεκτείνεται με συχνά στοιχεία που είναι λεξικογραφικά μεγαλύτερα του

$O(|F_{k-1}| \times |F_1|)$

{Beer, Diaper, Milk}

Δημιουργεί και κάποια *περιττά*, πχ το παραπάνω δεν είναι συχνό, γιατί το {Beer, Milk} δεν είναι συχνό

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

$F_{k-1} \times F_1$

Επέκταση κάθε συχνού (k-1) στοιχειοσυνόλου με άλλα συχνά στοιχεία

Διάφοροι ευριστικοί για να μειωθεί ο αριθμός των στοιχειοσυνόλων που δημιουργούνται και δεν είναι συχνά

Πχ έστω το $\{i_1, i_2, i_3, i_4\}$ για να είναι συχνό πρέπει όλα τα 3-στοιχειοσύνολα που είναι υποσύνολα του να είναι συχνά,

Πχ θα πρέπει να υπάρχουν τουλάχιστον 3 3-στοιχειοσύνολα που περιέχουν πχ το i_4 ($\{i_1, i_2, i_4\}$, $\{i_1, i_3, i_4\}$ και $\{i_2, i_3, i_4\}$)

Γενικά, κάθε στοιχείο ενός k-στοιχειοσυνόλου θα πρέπει να περιέχεται σε τουλάχιστον k-1 από το συχνά (k-1)-στοιχειοσύνολα

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

$F_{k-1} \times F_{k-1}$

Συγχώνευση δύο συχνών (k-1) στοιχειοσυνόλου αν τα πρώτα k-2 στοιχεία τους είναι τα ίδια

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Συγχώνευση δύο συχνών (k-1)-στοιχειοσυνόλων αλλά πρέπει *επιπρόσθετα να ελέγξουμε* ότι και τα υπόλοιπα k-2 υποσύνολα είναι συχνά

Στρατηγική apriori

Γενικός Αλγόριθμος

k = 1

Δημιούργησε όλα τα συχνά στοιχειοσύνολα μήκους 1

Repeat until δεν δημιουργούνται νέα στοιχειοσύνολα

- Δημιούργησε υποψήφια στοιχειοσύνολα μήκους (k+1) από τα συχνά στοιχειοσύνολα μήκους k
- Prune τα υποψήφια στοιχειοσύνολα που περιέχουν υποσύνολα μήκους k που δεν είναι συχνά
- Υπολόγισε την υποστήριξη (support) κάθε υποψηφίου στοιχειοσύνολου διαβάζοντας από τη βάση δεδομένων
- Σβήσε τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά, αφήνοντας μόνο τα συχνά

Στρατηγική apriori: Υπολογισμός Υποστήριξης

Υπολογισμός υποστήριξης: για κάθε νέο υποψήφιο συχνό στοιχειοσύνολο, πρέπει να υπολογίσουμε την υποστήριξή του

Brute Force:

- Διαπέρασε τη βάση των δοσοληψιών για τον υπολογισμό της υποστήριξης κάθε υποψήφιου στοιχειοσυνόλου

Αν σε ένα βήμα έχουμε m συχνά στοιχειοσύνολα, τότε διαπέραση της βδ m φορές

Δοσοληψίες

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

For each item

for i = 1 to N

if t_i περιέχει το item

c(item)++

N

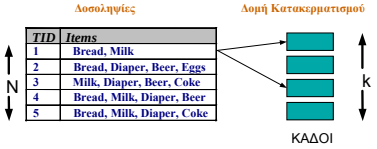
Πχ έστω item = {Beer, Bread}

Στρατηγική apriori: Υπολογισμός Υποστήριξης

Ελάττωση του αριθμού των συγκρίσεων

Για να *μειώσουμε* τον αριθμό των συγκρίσεων, αποθήκευση των υποψηφίων στοιχειοσυνόλων σε μια δομή κατακερματισμού

- Αντί να ταιριάζουμε κάθε δοσοληψία με κάθε υποψήφιο στοιχειοσύνολο, **ταίριαξε κάθε δοσοληψία με τα υποψήφια στοιχειοσύνολα που περιέχονται σε κάδους κατακερματισμού**



Στρατηγική apriori: Υπολογισμός Υποστήριξης

Απαρίθμηση Στοιχειο-συνόλων

Βασική ιδέα του κατακερματισμού

1. Κατά τη διάρκεια του apriori τα συχνά στοιχειοσύνολα που παράγονται κατακερματίζονται σε κάδους και αποθηκεύονται σε ένα δέντρο κατακερματισμού

2. Στη συνέχεια, κάθε δοσοληψία (για την ακρίβεια, κάθε στοιχειοσύνολο που περιέχει) κατακερματίζεται με την ίδια συνάρτηση και τη συγκρίνουμε όχι με όλα τα πιθανά στοιχειοσύνολα, *αλλά μόνο με τα στοιχειοσύνολα στους αντίστοιχους κάδους*

Ας δούμε πως

Στρατηγική apriori: Υπολογισμός Υποστήριξης

1. Δημιουργία του δέντρου κατακερματισμού υποψηφίων στοιχειοσυνόλων

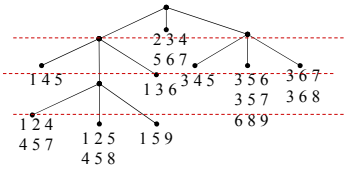
Έστω ότι έχουμε 15 υποψήφια 3-στοιχειοσύνολα:

{1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}

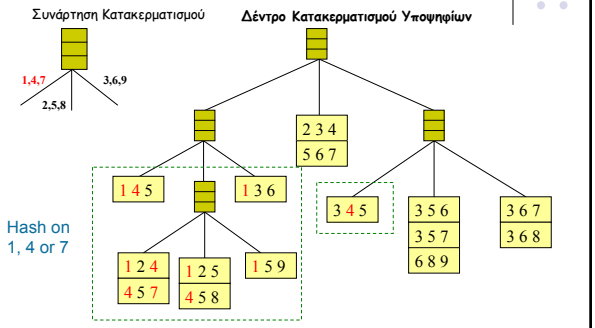
Τα αποθηκεύουμε στα φύλλα (κάδους) του δέντρου

Στο δέντρο κατακερματίζουμε τα υποψήφια στοιχειοσύνολα

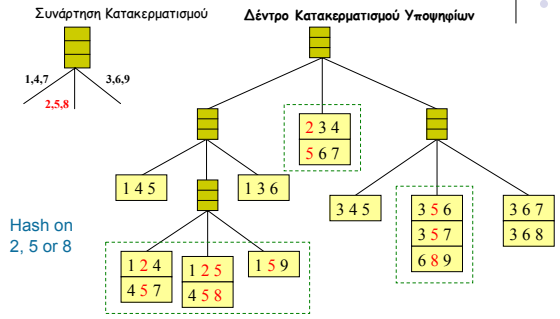
- Συνάρτηση κατακερματισμού (ποιο κλαδί θα ακολουθήσουμε σε κάθε επίπεδο)
- Μέγιστο Μήκος Φύλλου: μέγιστο αριθμό στοιχειοσυνόλων που θα αποθηκευτούν σε κάθε φύλλο (αν ο αριθμός των στοιχειοσυνόλων υπερβεί το μέγιστο μέγεθος του φύλλου, διαχώρισε τον κόμβο – χρήση κατακερματισμού στο επόμενο στοιχείο)



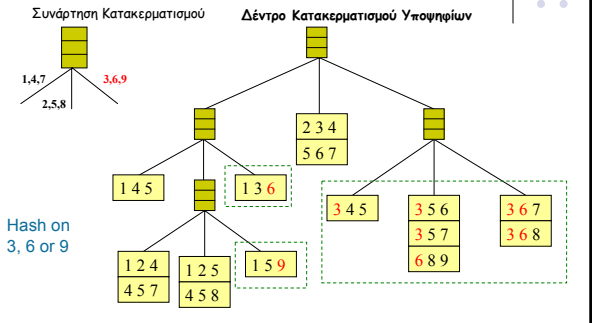
Στρατηγική apriori: Υπολογισμός Υποστήριξης



Στρατηγική apriori: Υπολογισμός Υποστήριξης



Στρατηγική apriori: Υπολογισμός Υποστήριξης



Στρατηγική apriori: Υπολογισμός Υποστήριξης

2. Απαρίθμηση Υποσυνόλων με χρήση του Δέντρου Κατακερματισμού

Έχοντας κατασκευάσει το δέντρο κατακερματισμού (για τα 3-στοιχειοσύνολα),

Για κάθε δοσοληψία,

κατακερματίζουμε όλα τα 3-στοιχειοσύνολα της δοσοληψίας στο δέντρο

και αυξάνουμε τον αντίστοιχο μετρητή

Στρατηγική apriori: Υπολογισμός Υποστήριξης

Απαρίθμηση Στοιχειο-συνόλων

Πχ έστω ότι είμαστε στο 3 βήμα και έχουμε δημιουργήσει όλα τα πιθανά 3-στοιχειο-σύνολα

Έστω μια δοσοληψία t με 5 στοιχεία {1, 2, 3, 5, 6}

Θα πρέπει να ελέγξουμε για καθένα από αυτά αν το περιέχει η t

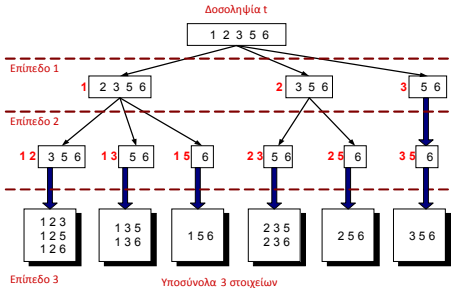
Αν το περιέχει η t θα πρέπει να αυξήσουμε την υποστήριξη του κατά 1

Ας δούμε πρώτα ένα συστηματικό τρόπο για την απαρίθμηση όλων των 3-στοιχειοσυνόλων της t

Στρατηγική apriori: Υπολογισμός Υποστήριξης

Απαρίθμηση Στοιχειο-συνόλων

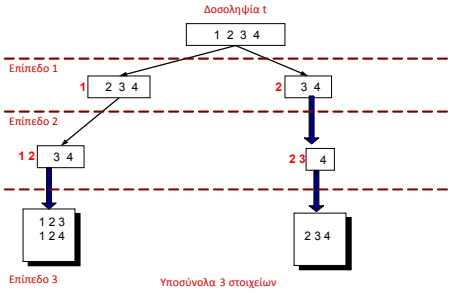
Έστω μια δοσοληψία t με 5 στοιχεία {1, 2, 3, 5, 6} - Απαρίθμηση όλων των πιθανών υποσυνόλων της με τρία στοιχεία (3-στοιχειοσύνολα) με λεξικογραφική διάταξη



Στρατηγική apriori: Υπολογισμός Υποστήριξης

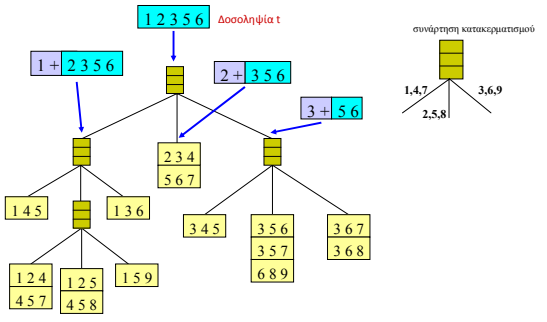
Απαρίθμηση Στοιχειο-συνόλων

Έστω μια δοσοληψία t με 4 στοιχεία {1, 2, 3, 4} - Απαρίθμηση όλων των πιθανών υποσυνόλων της με τρία στοιχεία (3-στοιχειοσύνολα) με λεξικογραφική διάταξη



Στρατηγική apriori: Υπολογισμός Υποστήριξης

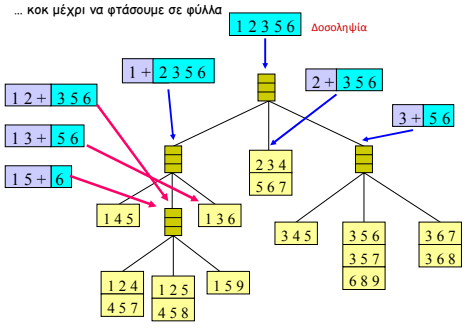
Με βάση το δέντρο απαρίθμησης για την t = {1, 2, 3, 5, 6} όλα τα δυνατά στοχειοσύνολα αρχίζουν από 1, 2ή 3 => στη ρίζα κατακερματίζουμε χωριστά τα 1, 2 και 3 – δηλαδή με βάση τα στοιχεία του πρώτου επιπέδου



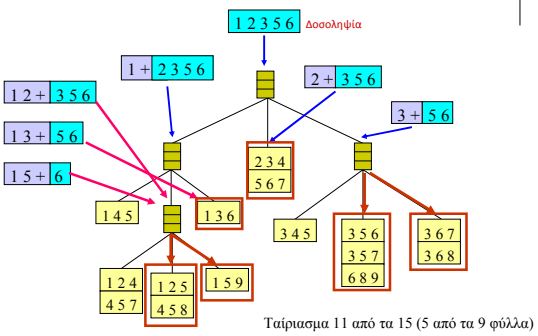
Στρατηγική apriori: Υπολογισμός Υποστήριξης

στη συνέχεια κατακερματίζουμε με βάση τα αντίστοιχα στοιχεία του δεύτερου επιπέδου: 2, 3, 5 (για το 1) 3, 5 (για το 2) 5 (για το 3)

... κοκ μέχρι να φτάσουμε σε φύλλα



Στρατηγική apriori: Υπολογισμός Υποστήριξης



Ταίριασμα 11 από τα 15 (5 από τα 9 φύλλα)

Στρατηγική apriori

Γενικός Αλγόριθμος (ξανά)

k = 1

Δημιούργησε όλα τα συχνά στοιχειοσύνολα μήκους 1

Repeat until δεν δημιουργούνται νέα στοιχειοσύνολα

- Δημιούργησε υποψήφια στοιχειοσύνολα μήκους (k+1) από τα συχνά στοιχειοσύνολα μήκους k (είτε $F_{k-1} \times F_1$ είτε $F_{k-1} \times F_{k-1}$)
- Prune τα υποψήφια στοιχειοσύνολα που περιέχουν υποσύνολα μήκους k που δεν είναι συχνά
- Υπολόγισε την υποστήριξη (support) κάθε υποψηφίου στοιχειοσύνολου διαβάζοντας από τη βάση δεδομένων (πχ χρησιμοποίησε το δέντρο κατακερματισμού)
- Σβήσε τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά, αφήνοντας μόνο τα συχνά

Στρατηγική apriori: Πολυπλοκότητα

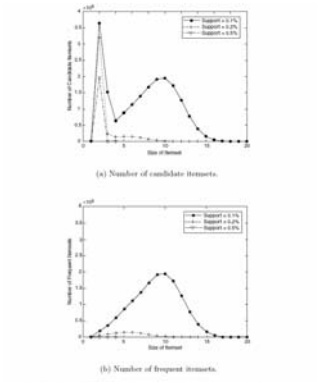


Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.

- Επιλογή της τιμής του κατωφλιού για την ελάχιστη υποστήριξη
  - Μικρή τιμή => πολλά συχνά στοιχειοσύνολα
  - Αύξηση υποψήφιων στοιχειοσυνόλων (πολυπλοκότητα) και το μέγιστο μήκος των συχνών στοιχειοσυνόλων (περισσότερα περάσματα στα δεδομένα)

Στρατηγική apriori: Πολυπλοκότητα

Αριθμός διαστάσεων - Dimensionality (αριθμός στοιχείων) του συνόλου δεδομένων

- Περισσότερος χώρος για την αποθήκευση της υποστήριξης κάθε στοιχείου
- Αύξηση του αριθμού των συχνών στοιχείων, αύξηση του υπολογιστικού κόστους και του κόστους I/O

Μέγεθος της βάσης

Επειδή ο Apriori κάνει πολλαπλά περάσματα, ο χρόνος εκτέλεσης μπορεί να αυξηθεί

Στρατηγική apriori: Πολυπλοκότητα

- Μέσο πλάτος δοσοληψίας
  - Το μέγιστο μήκος των συχνών στοιχειοσύνολων τείνει να αυξηθεί με την αύξηση του μέσου πλάτους των δοσοληψιών, άρα και ο αριθμός των υποψηφίων σε κάθε βήμα
  - Επίσης, αύξηση των περασμάτων του δέντρου

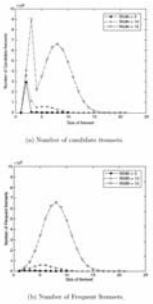


Figure 6.14. Effect of average transaction width on the number of candidate and frequent itemsets.

Στρατηγική apriori: Πολυπλοκότητα

1. Δημιουργία συχνών 1-στοχειοσυνόλων

$O(Nw)$

2. Δημιουργία υποψήφιων στοιχειοσυνόλων

Έστω $F_{k-1} \times F_{k-1}$

k-2 συγκρίσεις για κοινό prefix

Στη χειρότερη περίπτωση, ταιριάζουν όλα $\Sigma_{k=2,w}|F_{k-1}|^2$

Επίσης κατασκευάζουμε το δέντρο, μέγιστο ύψος k, άρα $\Sigma_{k=2,w}k|F_{k-1}|^2$

Έλεγχος, για τα k-2 υποσύνολα με χρήση του δέντρου

3. Υπολογισμός της Υποστήριξης

Κάθε δοσοληψία έχει k από |t| k-στοιχειοσύνολα

Δημιουργία Κανόνων

Παραγωγή Κανόνων

Παραγωγή Κανόνων (Rule Generation)

- Δοθέντος ενός συχνού στοιχειοσυνόλου L, βρες όλα τα μη κενά υποσύνολα $f \subset L$ τέτοια ώστε ο κανόνας $f \rightarrow L - f$ ικανοποιεί τον περιορισμό της ελάχιστης εμπιστοσύνης
- Παράδειγμα αν {A,B,C,D} υποψήφιοι κανόνες:
  - $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
    $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$
    $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
    $BD \rightarrow AC$, $CD \rightarrow AB$,

  Όλοι έχουν την ίδια υποστήριξη, πρέπει να ελέγξουμε την εμπιστοσύνη
- Αν $|L| = k$, τότε υπάρχουν $2^k - 2$ υποψήφιοι κανόνες συσχέτισης (εξαιρώντας τον $L \rightarrow \varnothing$ και τον $\varnothing \rightarrow L$)

Παραγωγή Κανόνων

Υπολογισμός Εμπιστοσύνης

- Παρατήρηση: Δε χρειάζεται να διαπεράσουμε πάλι τα δεδομένα για να υπολογίσουμε την εμπιστοσύνη ενός κανόνα που προκύπτει από ένα συχνό στοιχειοσύνολο:
  - $ABC \rightarrow D$, $ABD \rightarrow C$, $ACD \rightarrow B$, $BCD \rightarrow A$,
    $A \rightarrow BCD$, $B \rightarrow ACD$, $C \rightarrow ABD$, $D \rightarrow ABC$
    $AB \rightarrow CD$, $AC \rightarrow BD$, $AD \rightarrow BC$, $BC \rightarrow AD$,
    $BD \rightarrow AC$, $CD \rightarrow AB$

Γιατί; Πχ $c(CD \rightarrow AB) = \sigma\{A, B, C, D\}/\sigma\{C, D\}$

Από την αντι-μονότονη ιδιότητα της υποστήριξης, το {C, D} είναι συχνό στοιχειοσύνολο άρα έχουμε ήδη υπολογίζει την υποστήριξή του

Παραγωγή Κανόνων

Πως μπορούν να παραχθούν αποδοτικά οι κανόνες από τα συχνά στοιχειοσύνολα;

- Γενικά, η αντι-μονότονη ιδιότητα δεν ισχύει για την εμπιστοσύνη

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

Δηλαδή, η εμπιστοσύνη του $X \rightarrow Y$ μπορεί να είναι μεγαλύτερη, μικρότερη ή ίση της εμπιστοσύνης ενός κανόνα $X' \rightarrow Y'$ όπου $X' \subseteq X$ και $Y' \subseteq Y$

Γενικά έστω $\{p\} \rightarrow \{q\}$ με εμπιστοσύνη c_1

- Και $\{p, r\} \rightarrow \{q\}$ με εμπιστοσύνη c_2 (το αριστερό μέρος – LHS - υπερσύνολο)

  Μπορεί $c_2 > c_1$, $c_2 < c_1$ ή $c_2 = c_1$
- Έστω $\{p\} \rightarrow \{q, r\}$ με εμπιστοσύνη c_3 (το δεξί μέρος – RHS - υπερσύνολο)

  $c_3 \leq c_1$
- Επίσης, $c_3 \leq c_2$

Παραγωγή Κανόνων

Η εμπιστοσύνη για τους κανόνες που παράγονται από το ίδιο στοιχειοσύνολο έχει μια αντι-μονότονη ιδιότητα

Για παράδειγμα L = {A,B,C,D}:

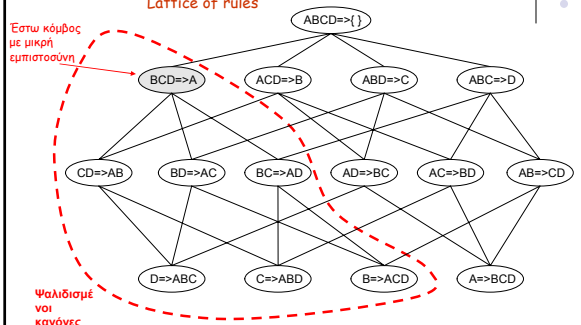
$$c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$$

- Η εμπιστοσύνη είναι αντι-μονότονη σε σχέση με των αριθμό των στοιχείων στο RHS του κανόνα (ή ισοδύναμα μονότονη στον αριθμό των στοιχείων στο LHS)

Pruning Rule:
Αν ο κανόνας $X \rightarrow Y - X$ δεν ικανοποιεί το κατώφλι εμπιστοσύνης, τότε και ο κανόνας $X' \rightarrow Y - X'$ ($X' \subseteq X$) δεν τον ικανοποιεί

Παραγωγή Κανόνων για τον Αλγόριθμο apriori

Πλέγμα Κανόνων
Lattice of rules



Παραγωγή Κανόνων για τον Αλγόριθμο apriori

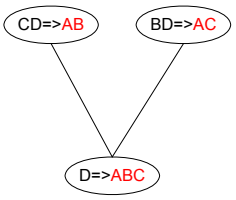
Οι κανόνες παράγονται σε επίπεδα με βάση τα στοιχεία στο RHS

Αρχικά, θεωρούμε όλους τους κανόνες **με ένα στοιχείο** στο RHS

Στη συνέχεια, οι υποψήφιοι κανόνες παράγονται συγχωνεύοντας το RHS δυο υποψηφίων κανόνων
Πχ
Join(AC=>B, AD=>C) μας δίνει AD=>BC

Όπως και στα συχνά στοιχειοσύνολα, στη συνέχεια, με το ίδιο prefix στο RHS
join(CD=>AB, BD=>AC) μας δίνει D => ABC

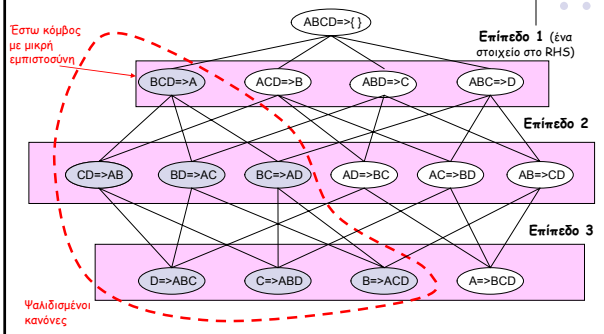
Prune τον κανόνα D=>ABC, αν το υποσύνολο AD=>BC δεν έχει επαρκή εμπιστοσύνη



- Σε αντίθεση με την περίπτωση των συχνών στοιχειοσυνόλων, δε χρειάζεται να διαβάσουμε τις δοσοληψίες για να υπολογίσουμε την εμπιστοσύνη

Παραγωγή Κανόνων για τον Αλγόριθμο apriori

Πλέγμα Κανόνων



Αναπαράσταση Κανόνων Συσχέτισης

Αναπαράσταση Στοιχειοσυνόλων

Τα στοιχειοσύνολα που παράγονται είναι πολλά, κάποια ίσως περιττά

Ποια να κρατήσουμε;

Αντιπροσωπευτικά συχνά στοιχειοσύνολα

Περιττός κανόνας

$X \rightarrow Y$, αν υπάρχει ένας κανόνας $X' \rightarrow Y'$, όπου $X \subseteq X'$ και $Y \subseteq Y'$ με την ίδια υποστήριξη και εμπιστοσύνη

Πχ., $\{b\} \rightarrow \{d, e\}$ περιττός

Αν ο $\{b, c\} \rightarrow \{d, e\}$, έχει την ίδια υποστήριξη και εμπιστοσύνη

Αναπαράσταση Στοιχειοσυνόλων

Έστω οι παρακάτω 15 δοσοληψίες με 30 στοιχεία

Έστω, υποστήριξη 20%

TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

Αριθμός συχνών στοιχειοσυνόλων $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$

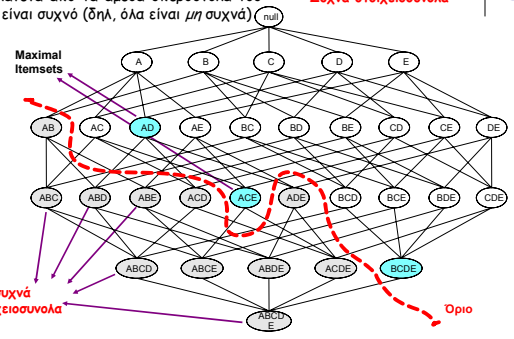
Μερικά στοιχειοσύνολα είναι πλεονάζοντα, έχουν την ίδια υποστήριξη με το τα υπερσύνολα τους

Πιθανή συνοπτική αναπαράσταση {A1, A2, A3, A4, A5, A6, A7, A8, A9, A10}, {B1, B2, B3, B4, B5, B6, B7, B8, B9, B10}, {C1, C2, C3, C4, C5, C6, C7, C8, C9, C10}

Αναπαράσταση Στοιχειοσυνόλων

Ένα στοιχειοσύνολο είναι maximal συχνό αν κανένα από τα άμεσα υπερσύνολά του δεν είναι συχνό (δηλ, όλα είναι *μη* συχνά)

Συχνά στοιχειοσύνολα

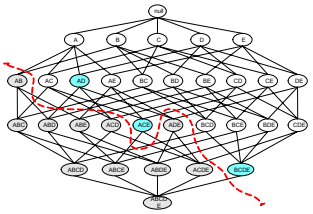


Αναπαράσταση Στοιχειοσυνόλων

Προσφέρουν μια συνοπτική αναπαράσταση των συχνών στοιχειοσυνόλων: το μικρότερο σύνολο στοιχειοσυνόλων από το οποίο μπορούμε να πάρουμε όλα τα συχνά στοιχειοσύνολα (είναι όλα τα υποσύνολά τους)

Βέβαια, αυτό έχει νόημα μόνο αν έχουμε έναν αποδοτικό αλγόριθμο για τον υπολογισμό τους που δεν παράγει όλα τα δυνατά υποσύνολα τους

ΜΕΙΟΝΕΧΤΗΜΑ: Δεν προσφέρουν καμιά πληροφορία για την υποστήριξη των υποσυνόλων τους



Αναπαράσταση Στοιχειοσυνόλων

Ένα στοιχειοσύνολο είναι **κλειστό (closed)** αν κανένα από τα άμεσα υπερσύνολα του δεν έχει την ίδια υποστήριξη με αυτό

Δεν είναι κλειστό αν κάποιο άμεσο υπερσύνολό του έχει την ίδια υποστήριξη

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Αναπαράσταση Στοιχειοσυνόλων

Ένα στοιχειοσύνολο είναι **κλειστό συχνό στοιχειοσύνολο** αν είναι κλειστό και η υποστήριξη του είναι μικρότερη ή ίση με minsup

Ο αλγόριθμος υπολογισμού της υποστήριξης βασίζεται στο ότι:

Η υποστήριξη ενός μη κλειστού στοιχειοσυνόλου πρέπει να είναι ίση με την μεγαλύτερη υποστήριξη ανάμεσα στα υπερσύνολά του

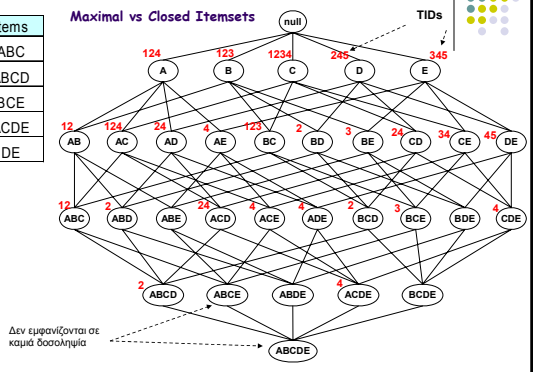
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

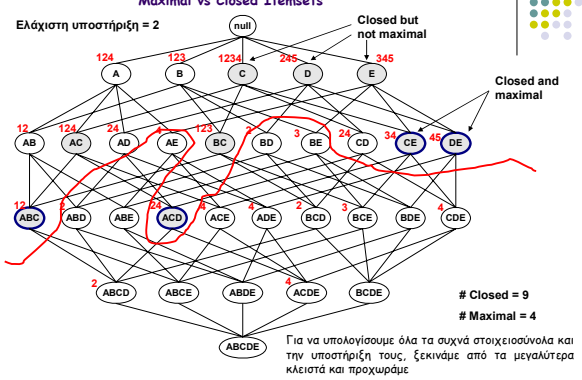
Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Αναπαράσταση Στοιχειοσυνόλων

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Αναπαράσταση Στοιχειοσυνόλων



Για να υπολογίσουμε όλα τα συχνά στοιχειοσύνολα και την υποστήριξη τους, ξεκινάμε από τα μεγαλύτερα κλειστά και προχωράμε

Αναπαράσταση Στοιχειοσυνόλων

Περιττός κανόνας

$X \rightarrow Y$, αν υπάρχει ένας κανόνας $X' \rightarrow Y'$, όπου $X \subseteq X'$ και $Y \subseteq Y'$ με την ίδια υποστήριξη και εμπιστοσύνη

{b} → {d, e} περιττός

{b, c} → {d, e}

Παρατήρηση: Θα κρατήσουμε μόνο το {b, c, d, e}

Αναπαράσταση Στοιχειοσυνόλων

Maximal vs Closed Itemsets