ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Αποθήκες Δεδομένων και Εξόρυξη Γνώσης από Δεδομένα

Κανόνες Συσχέτισης

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «**Ανοικτά Ακαδημαϊκά Μαθήματα στο Πανεπιστήμιο Αιγαίου**» έχει χρηματοδοτήσει μόνο τη αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Ενότητα 6: Κανόνες Συσχέτισης

Εισαγωγή

Καλάθι αγοράς

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)

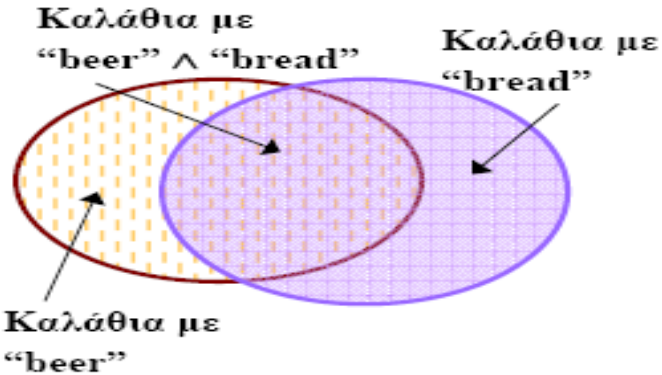
Παράδειγμα

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

- $T = \{t_1, i_2, .., t_N\}$ ένα σύνολο από **δοσοληψίες**, όπου κάθε t_i είναι ένα στοιχειοσύνολο
- **Πλάτος (width)** δοσοληψίας: αριθμός στοιχείων t_i που περιέχει ένα στοιχειοσύνολο X, αν το X είναι υποσύνολο της t_i



Ορισμοί

- **support count (σ) ενός στοιχειοσυνόλου:** Η συχνότητα εμφάνισης του στοιχειοσυνόλου
  - Παράδειγμα: σ({Milk, Bread, Diaper}) = 2
- **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

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→Y, όπου X και Y είναι στοιχειοσύνολα $X \subseteq I$, $Y \subseteq I$, $X \cap Y = \varnothing$
  - Παράδειγμα: {Milk, Diaper} → {Beer}
- Υποστήριξη Κανόνα Support (s) του X→Y : Το ποσοστό των δοσοληψιών που περιέχουν και το X και το Y ($X \cup Y$)
  - ή αλλιώς η πιθανότητα **P(X∪Y)**
- Εμπιστοσύνη - Confidence (c ή α) του X→Y: Πόσες από τις δοσοληψίες (ποσοστό) που περιέχουν το X περιέχουν και το Y
  - ή αλλιώς η πιθανότητα **P(X∪Y|X) = P(X∪Y)/P(X)**

Παράδειγμα

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

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
  - 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
  - Η εμπιστοσύνη μετρά την αξιοπιστία
  - Όσο μεγαλύτερη εμπιστοσύνη τόσο μεγαλύτερη η πιθανότητα εμφάνισης του Y σε κανόνες που περιέχουν το X

Εξόρυξη Κανόνων Συσχέτισης

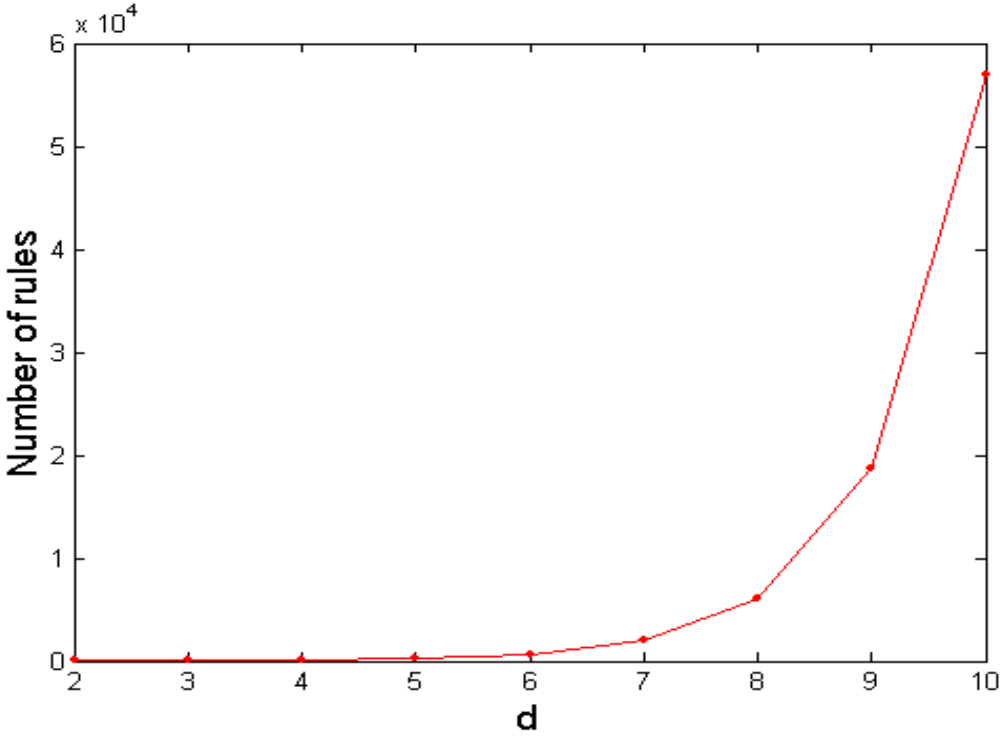
- Εύρεση Κανόνων Συσχέτισης
  - Είσοδος: Ένα σύνολο από δοσοληψίες T
  - Έξοδος: Όλοι οι κανόνες με
    - support ≥ *minsup*
    - confidence≥ *minconf*

- Προσέγγιση brute-force:
  - Παράθεση όλων των πιθανών κανόνων συσχέτισης
  - Για κάθε κανόνα
    - Υπολογισμός support και confidence
  - Αφαίρεση κανόνων που αποτυγχάνουν να καλύψουν τα κατώφλια *minsup* και *minconf*

⇒ ΥΠΟΛΟΓΙΣΤΙΚΑ ΑΠΑΓΟΡΕΥΤΙΚΗ

⇒ Για d στοιχεία, $3^d - 2^{d+1} + 1$

Εύρεση Συχνών Στοιχειοσυνόλων- Πολυπλοκότητα



- Για d μοναδικά στοιχεία
  - 2^d στοιχειοσύνολα
  - Αριθμός πιθανών κανόνων συσχέτισης:

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

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

Αν d = 6, R = 602 κανόνες

Εξόρυξη Κανόνων Συσχέτισης

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,

{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)

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 πιθανόν διαφορετική εμπιστοσύνη)

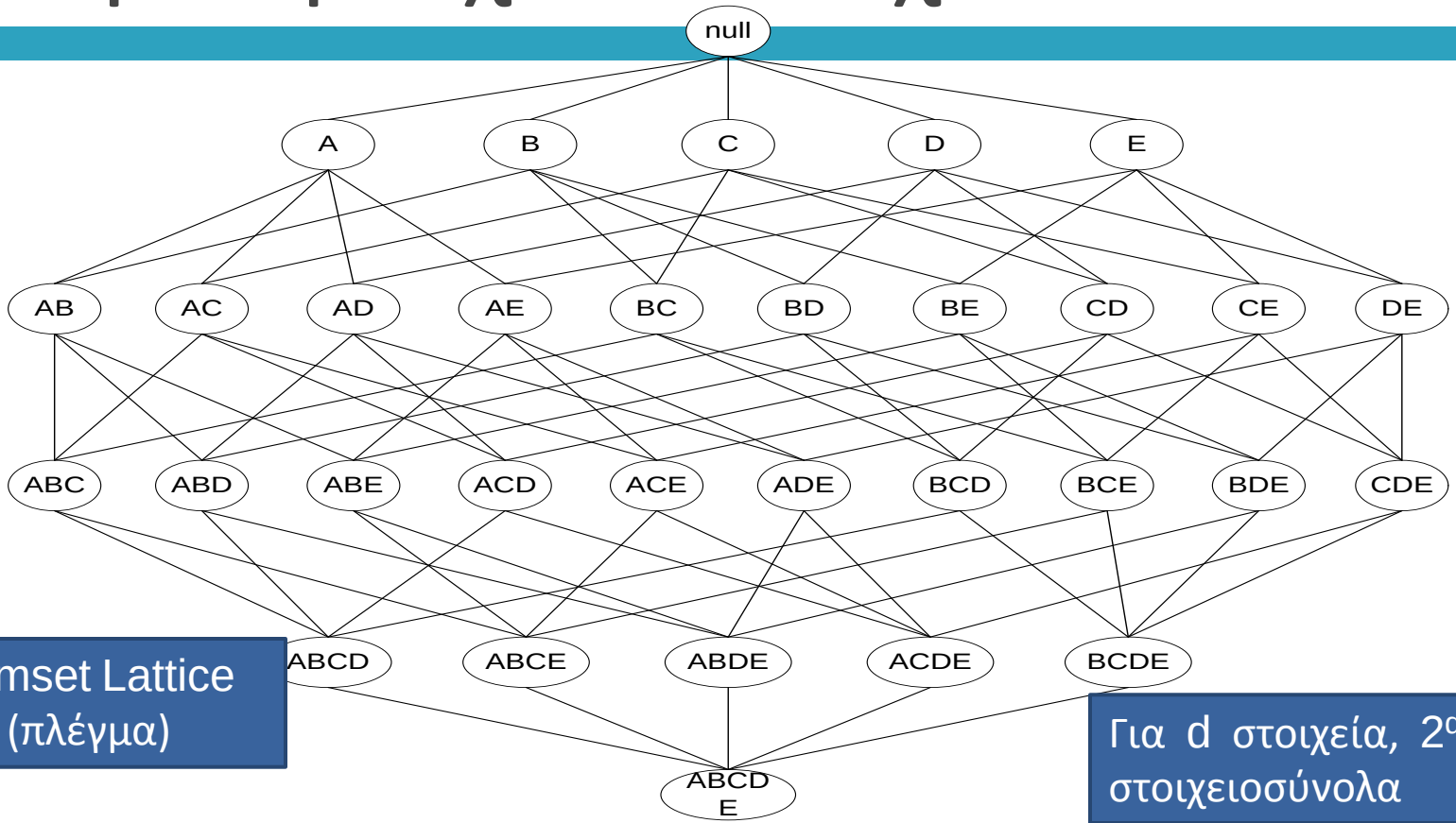
Αν είχαμε minsup = 0.5, θα αποκλείαμε και τους έξι κανόνες

Άρα μπορούμε να θεωρήσουμε 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 $\geq$ minsup
- Δημιουργία Κανόνων (Rule Generation)
  - Για κάθε στοιχειοσύνολο, δημιούργησε κανόνες με μεγάλη υποστήριξη, όπου κάθε κανόνας είναι μια δυαδική διαμέριση του συχνού στοιχειοσυνόλου
  - Η δημιουργία των συχνών στοιχειοσυνόλων είναι επίσης υπολογιστικά ακριβή

Εύρεση Συχνών Στοιχειοσυνόλων

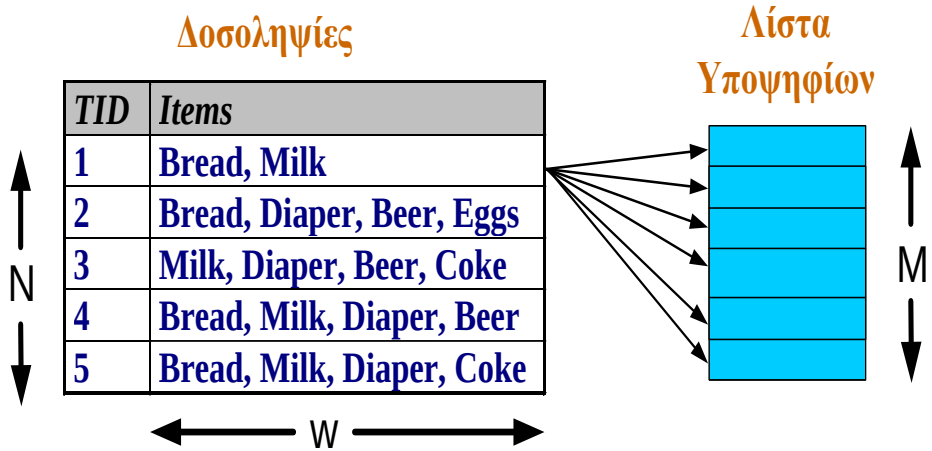


Itemset Lattice
(πλέγμα)

Για d στοιχεία, 2^d πιθανά στοιχειοσύνολα

Εύρεση Συχνών Στοιχειοσυνόλων

- Brute-force μέθοδος:
  - Κάθε στοιχειοσύνολο στο lattice είναι ένα υποψήφιο συχνό στοιχειοσύνολο
  - Υπολόγισε την υποστήριξη κάθε υποψήφιου στοιχειοσυνόλου διατρέχοντας (scanning) της βάση δεδομένων
- Ταίριαξε κάθε δοσοληψία με κάθε υποψήφιο
- Πολυπλοκότητα
  - ~ O(NMw) => Μεγάλη γιατί M = 2^d



Εύρεση Συχνών Στοιχειοσυνόλων: Στρατηγικές

- Μείωση του αριθμού των υποψηφίων στοιχειοσυνόλων (M)
  - Πλήρης αναζήτηση: $M=2^d$
  - Χρησιμοποίηση κάποιας τεχνικής pruning (ψαλιδίσματος - ελάττωσης) για να ελαττωθεί το M (πχ apriori)
- Μείωση του αριθμού των δοσοληψιών (N)
  - Ελάττωση του μεγέθους του N καθώς το μέγεθος του στοιχειοσυνόλου αυξάνεται
- Μείωση 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

Μείωση Υποψηφίων: Αρχή a-priori

- 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 στοιχειοσυνόλου είναι μικρότερη ή ίση της υποστήριξης οποιουδήποτε υποσυνόλου του :
  - **Anti-monotone** ιδιότητα της υποστήριξης

Αρχή a-priori

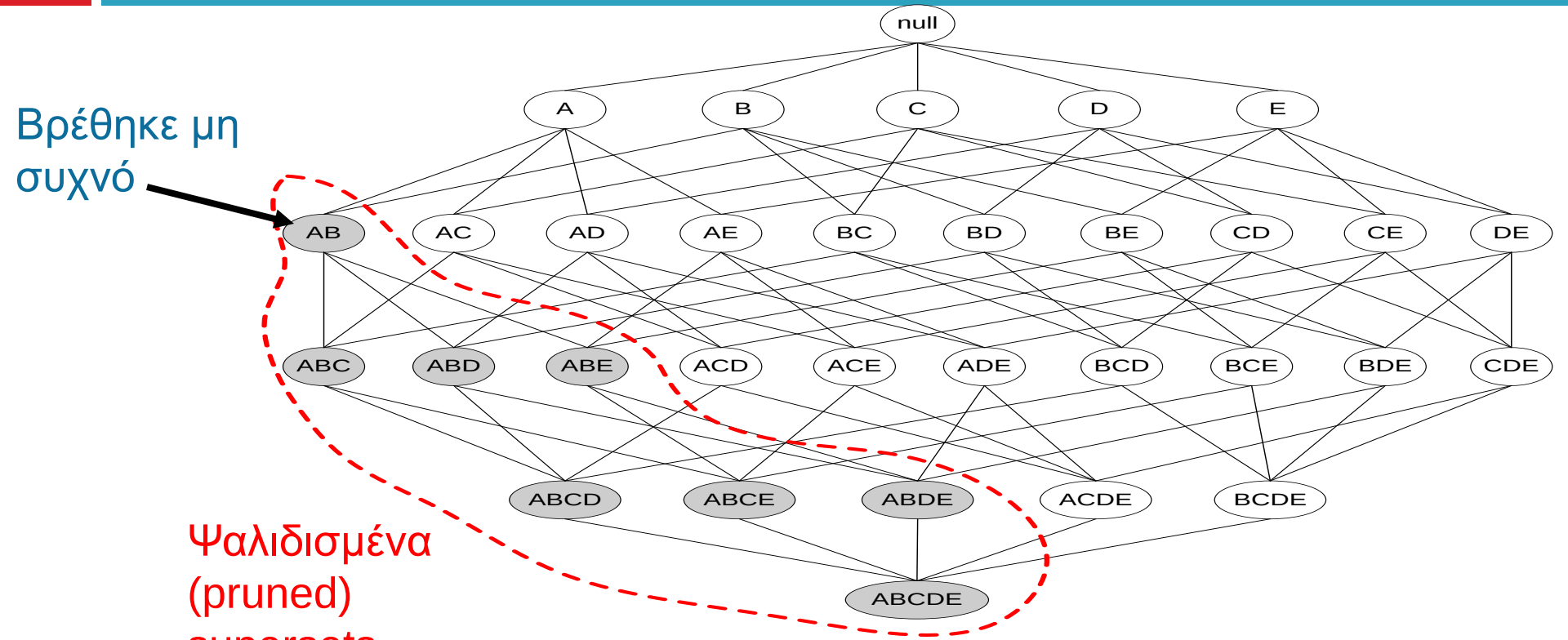
- Συχνό ονομάζεται το itemset που έχει υποστήριξη πάνω από ένα κατώφλι
- Παράδειγμα
  - (κατώφλι = 40%):
    - {Beer}
    - {Bread}
    - {PeanutButter}
    - {Bread, PeanutButter}

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Η ιδιότητα των συχνών itemsets:
  - Κάθε υποσύνολο ενός συχνού itemset είναι συχνό.
  - Αντιθέτως, αν ένα itemset δεν είναι συχνό, κανένα από τα υπερσύνολά του δεν μπορεί να είναι συχνό.

Παράδειγμα a-priori

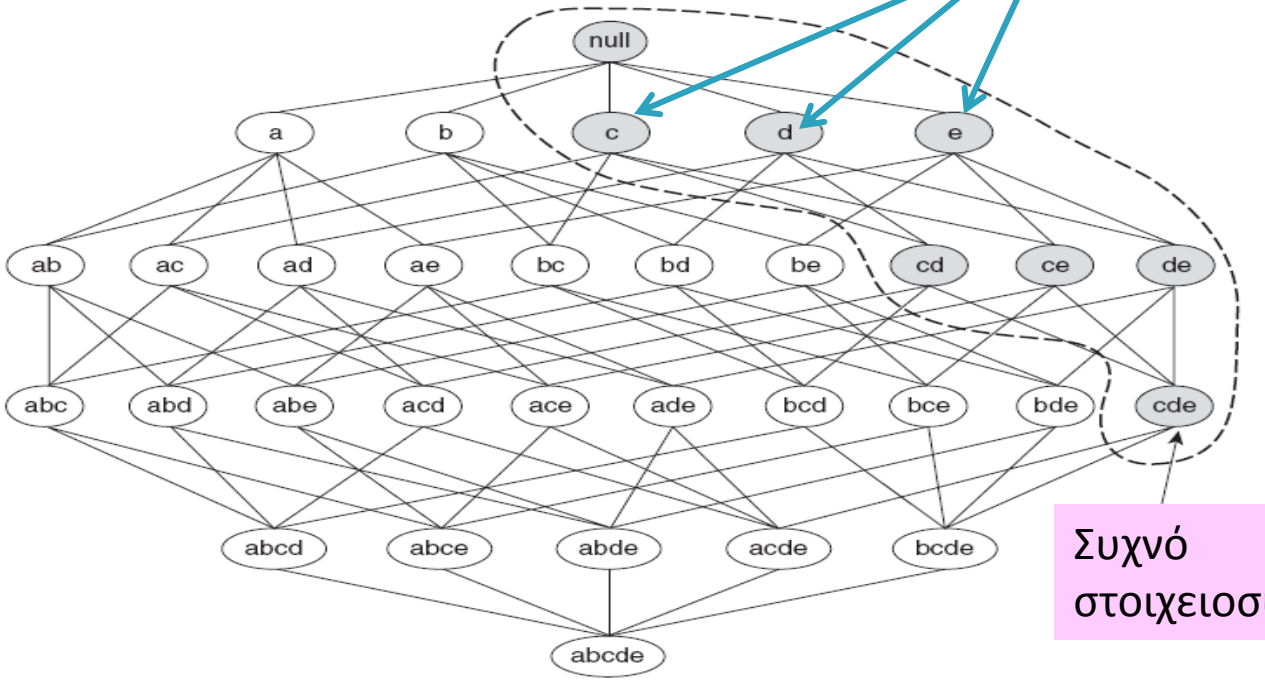
Βρέθηκε μη συχνό



Ψαλιδισμένα (pruned) supersets

Παράδειγμα a-priori

Όλα τα υποσύνολα του συχνά (κλειστό από πάνω)



Συχνό στοιχειοσύνολο

Figure 6.3. An illustration of the *Apriori* principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.

Στρατηγική a-priori

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

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

Τεμάχια (1-itemsets)

Ελάχιστη Υποστήριξη = 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:

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

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

Ζεύγη (2-itemsets)
Δεν δημιουργούνται υποψήφιοι με Coke ή Eggs

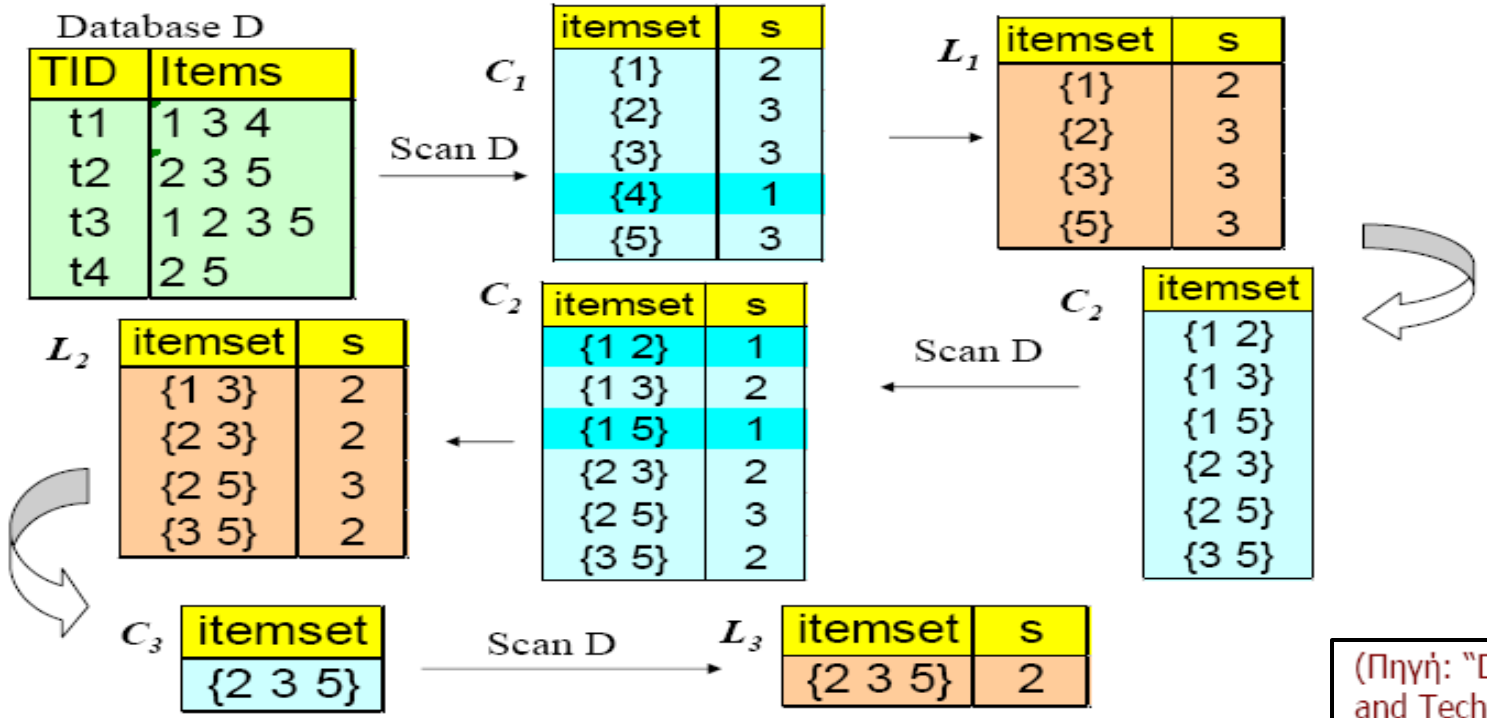
Τριπλέτες (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

Αλγόριθμος a-priori

- Έστω k=1
- Δημιουργία συχνών στοιχειοσυνόλων με μήκος 1
- Επανέλαβε έως ότου δεν προκύπτουν νέα συχνά στοιχειοσύνολα
  - Δημιουργία υποψηφίων στοιχειοσυνόλων μήκους (k+1) από μήκους k στοιχειοσύνολα
  - Ψαλίδισμα υποψηφίων στοιχειοσυνόλων που περιέχουν υποσύνολα μήκους k, που δεν είναι συχνά
  - Υπολογισμός της υποστήριξης κάθε υποψηφίου σαρώνοντας τη βάση
  - Αφαίρεση υποψηφίων που δεν είναι συχνοί, παραμένουν μόνον οι συχνοί

Απλό παράδειγμα με a-priori



(Πηγή: "Data Mining: Concepts and Techniques", Han & Kamber)

Στρατηγική a-priori: Δημιουργία Στοιχειοσυνόλων

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

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

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

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

{Beer, Diaper, Milk}

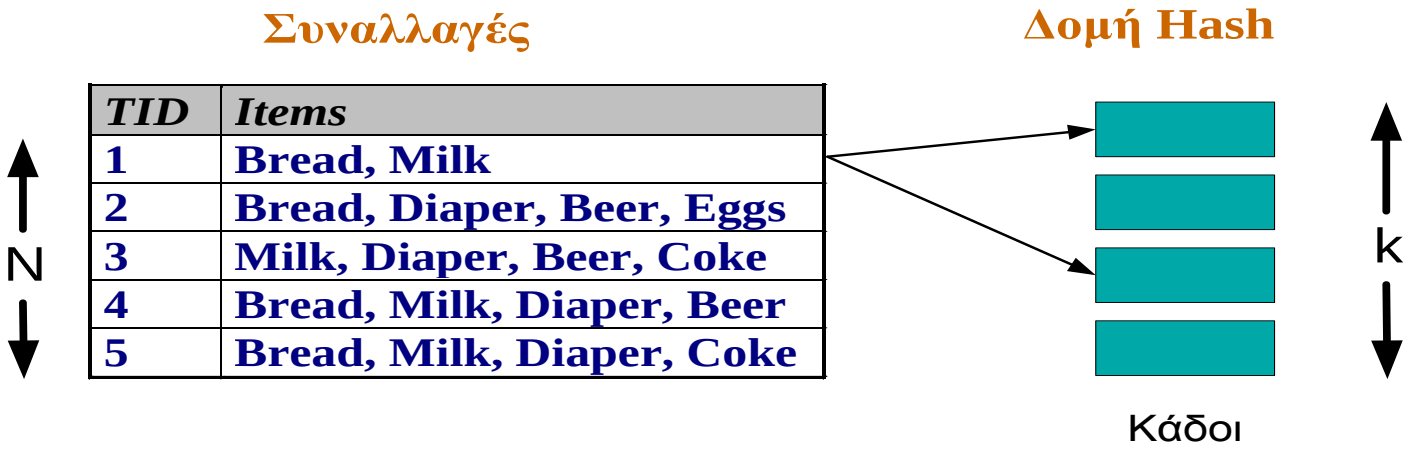
Δημιουργεί και κάποια περιττά, πχ το παραπάνω δεν είναι συχνό, γιατί το {Beer, Milk} δεν είναι συχνό

Στρατηγική a-priori: Δημιουργία Στοιχειοσυνόλων

- 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
- Π.χ. έστω το $\{i_1, i_2, i_3, i_4\}$ για να είναι συχνό, θα πρέπει να υπάρχουν τουλάχιστον 3 τρι-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 τα συχνά (k-1)-στοιχειοσύνολα

Ελάττωση του αριθμού των συγκρίσεων

- Μέτρηση υποψηφίων:
  - Σάρωση της βάσης και προσδιορισμός της υποστήριξης κάθε στοιχειοσυνόλου
  - Αποθήκευση κάθε υποψηφίου σε μια δομή κατακερματισμού (hash structure)
    - Αντί να ταιριάζουμε κάθε δοσοληψία με κάθε υποψήφιο στοιχειοσύνολο, **ταιριάζουμε κάθε δοσοληψία με τα υποψήφια στοιχειοσύνολα που περιέχονται σε κάδους κατακερματισμού**

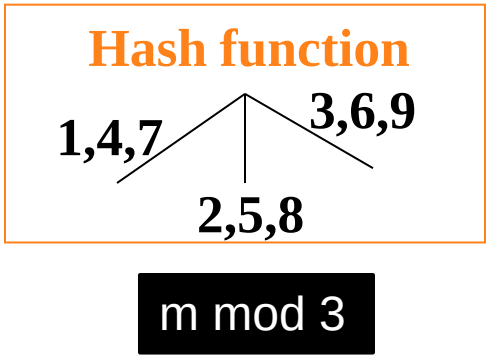


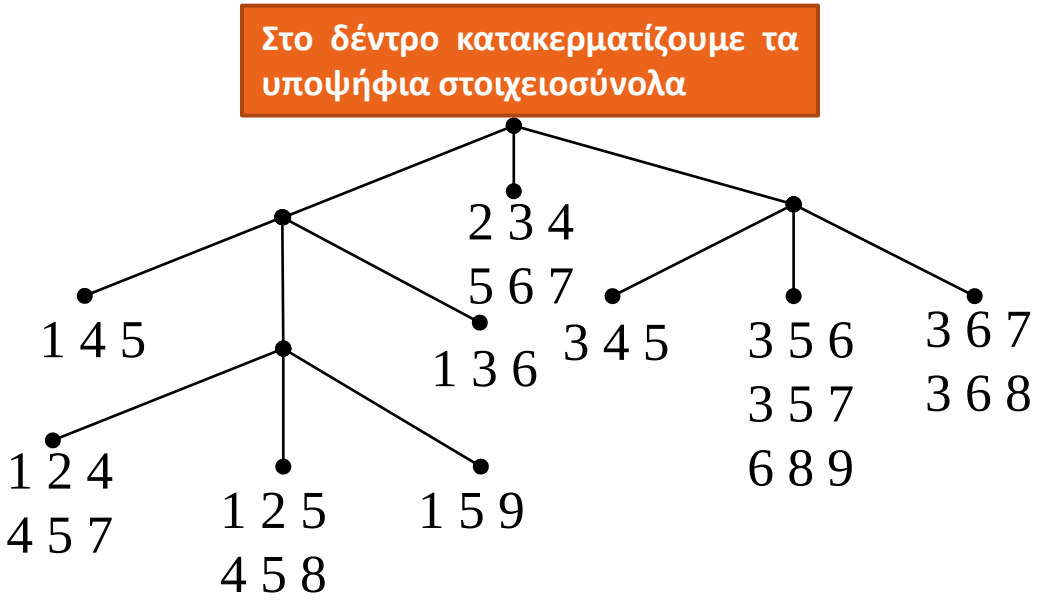
Δημιουργία δέντρου κατακερματισμού (hash tree)

- 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}
- Πρέπει να προσδιορίσουμε:
  - Συνάρτηση κατακερματισμού
  - 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)

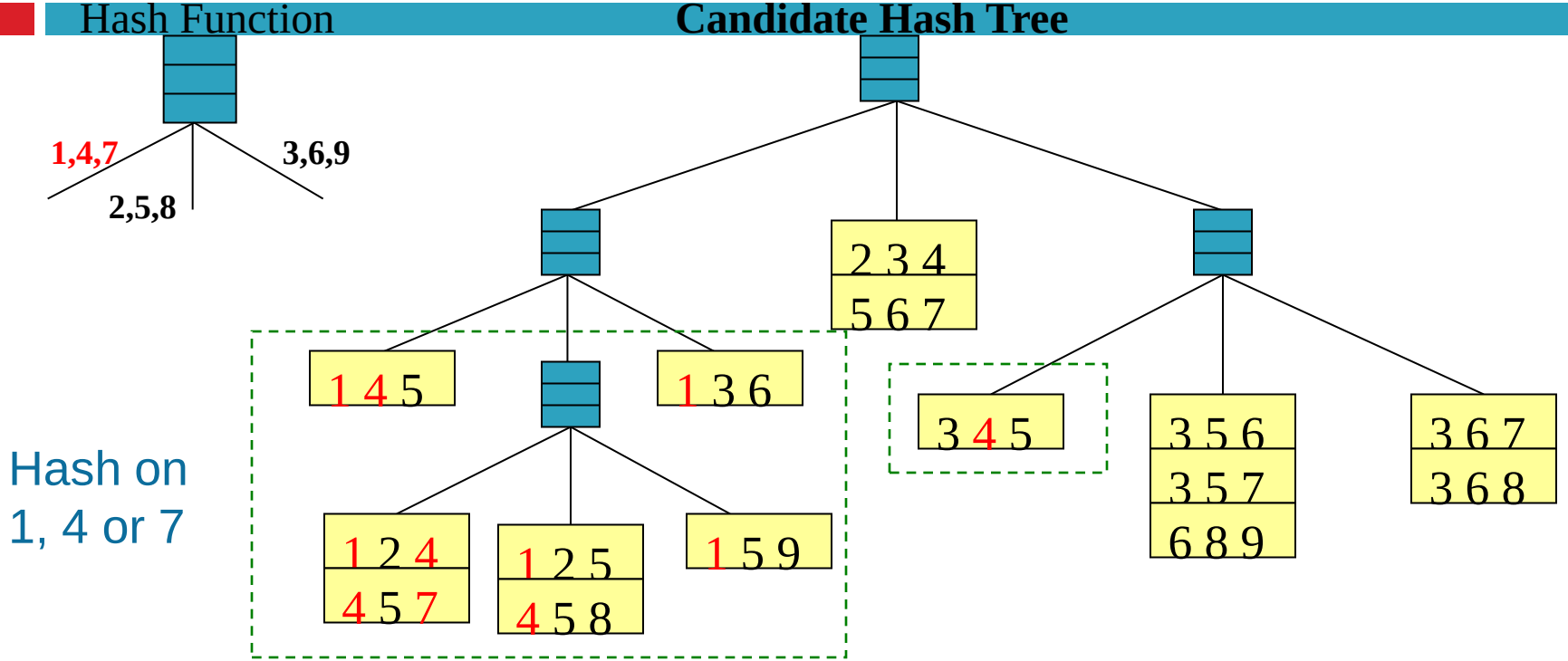
Δημιουργία δέντρου κατακερματισμού (hash tree)

- {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}

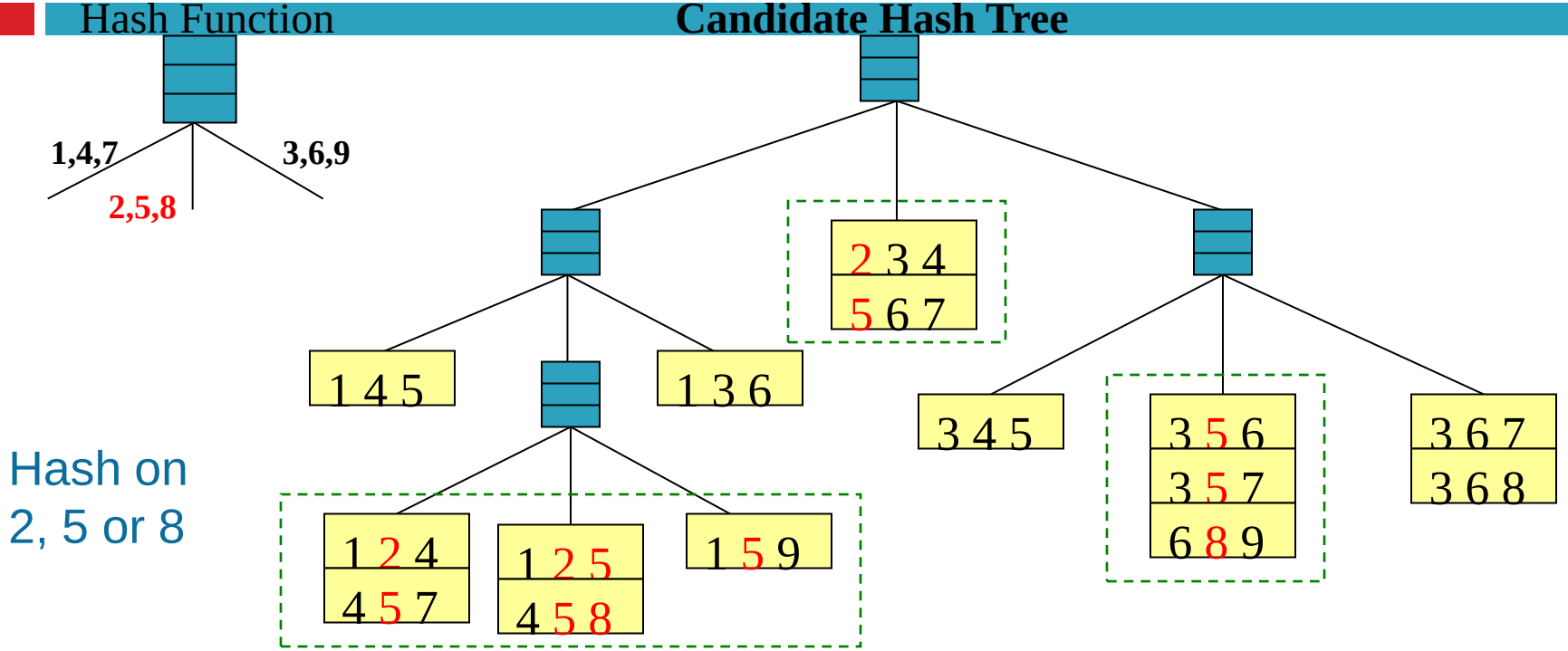




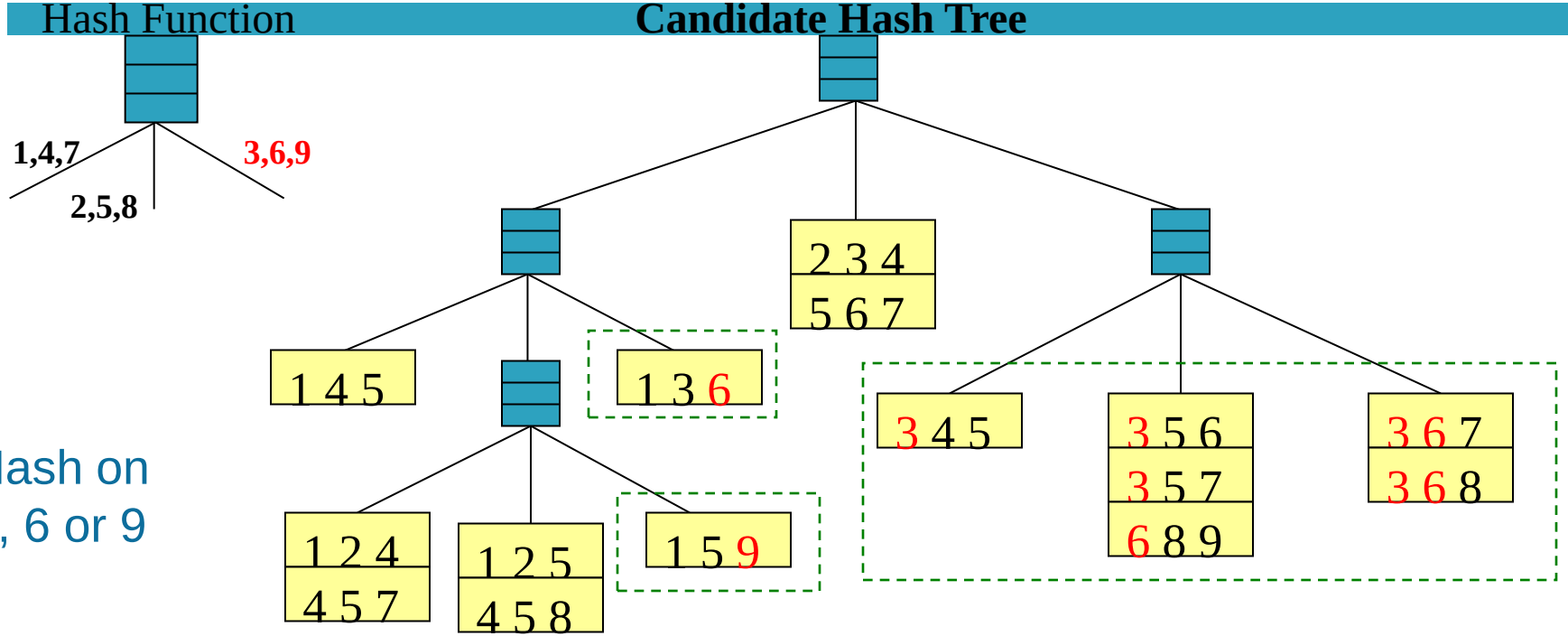
Δημιουργία δέντρου κατακερματισμού (hash tree)



Δημιουργία δέντρου κατακερματισμού (hash tree)

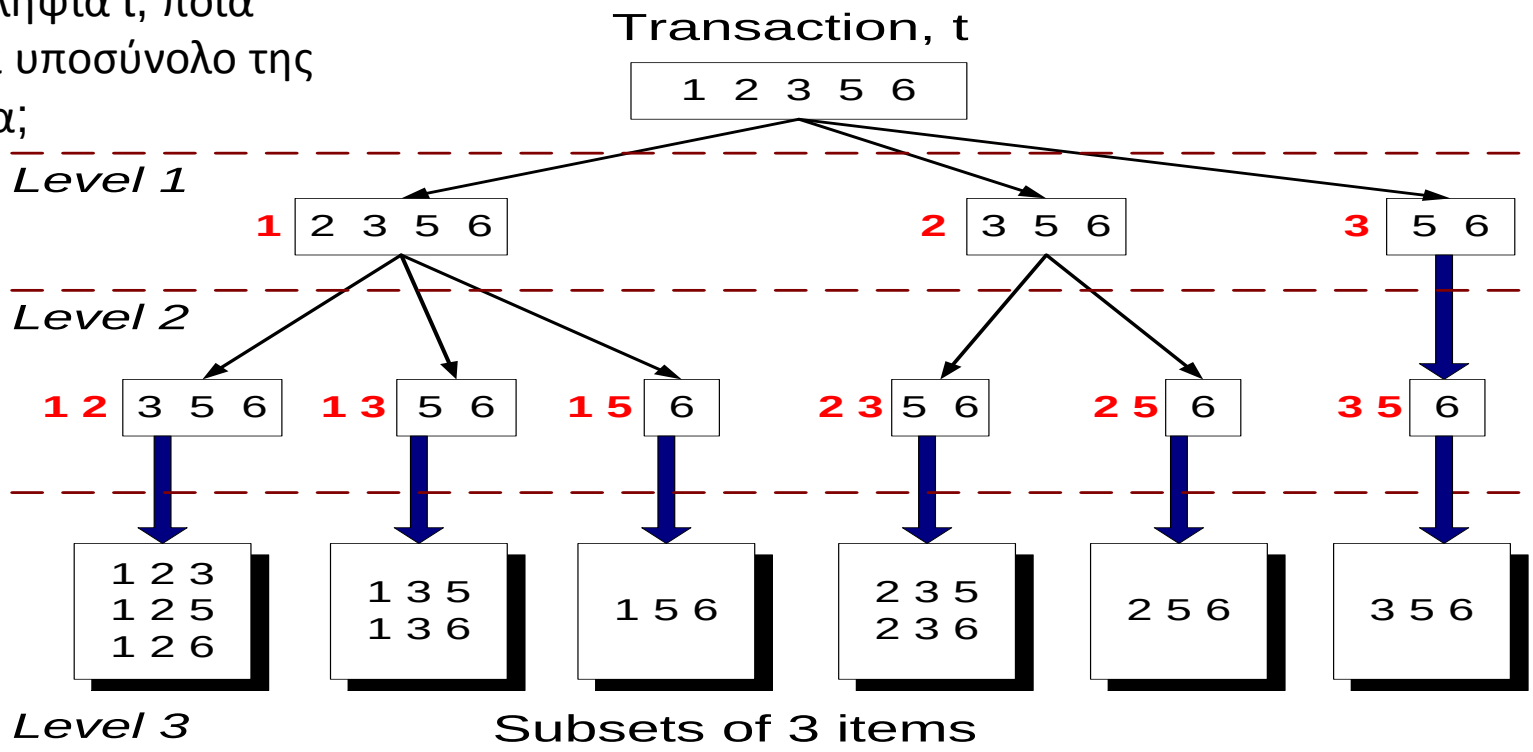


Δημιουργία δέντρου κατακερματισμού (hash tree)

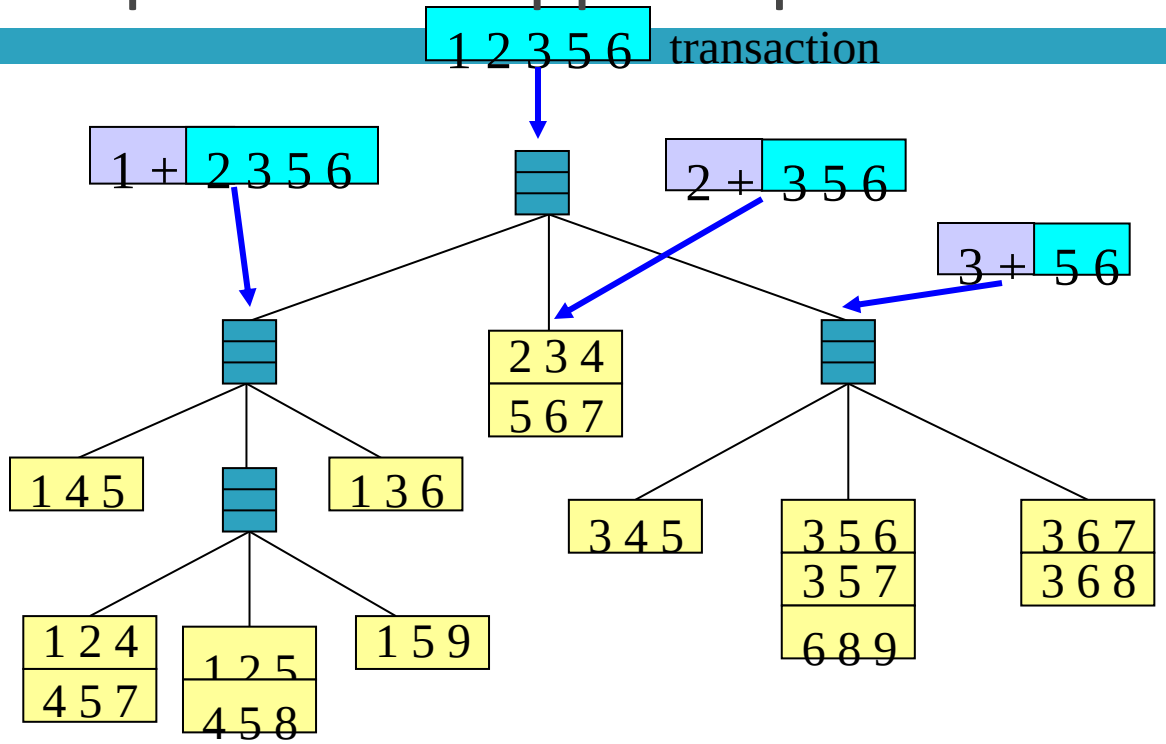


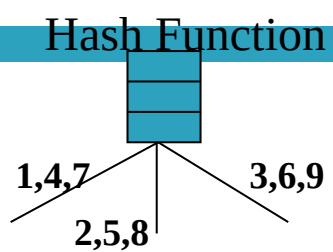
Απαρίθμηση Υποσυνόλων

Έστω μια δοσοληψία t, ποια είναι τα πιθανά υποσύνολο της με τρία στοιχεία;



Απαρίθμηση Υποσυνόλων με χρήση του Δέντρου Κατακερματισμού



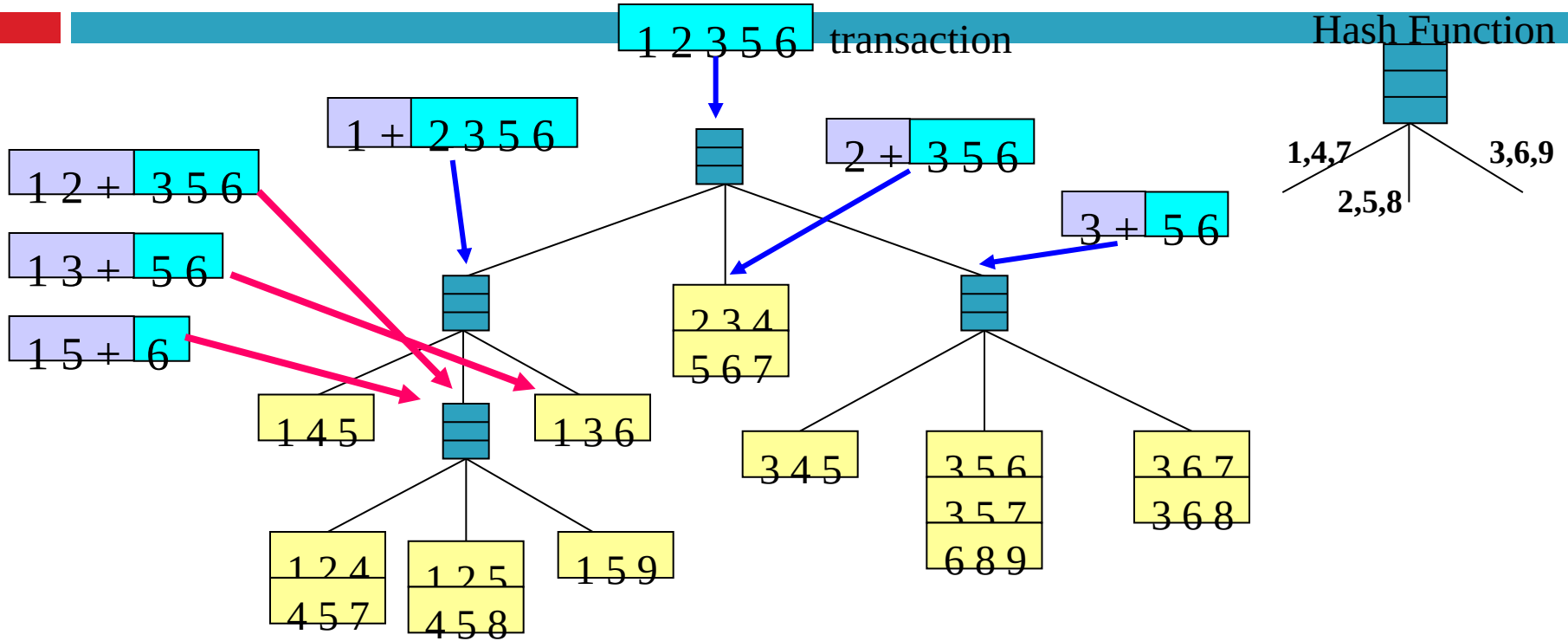


Έχοντας κατασκευάσει το δέντρο κατακερματισμού για τα 3-στοιχειοσύνολα,

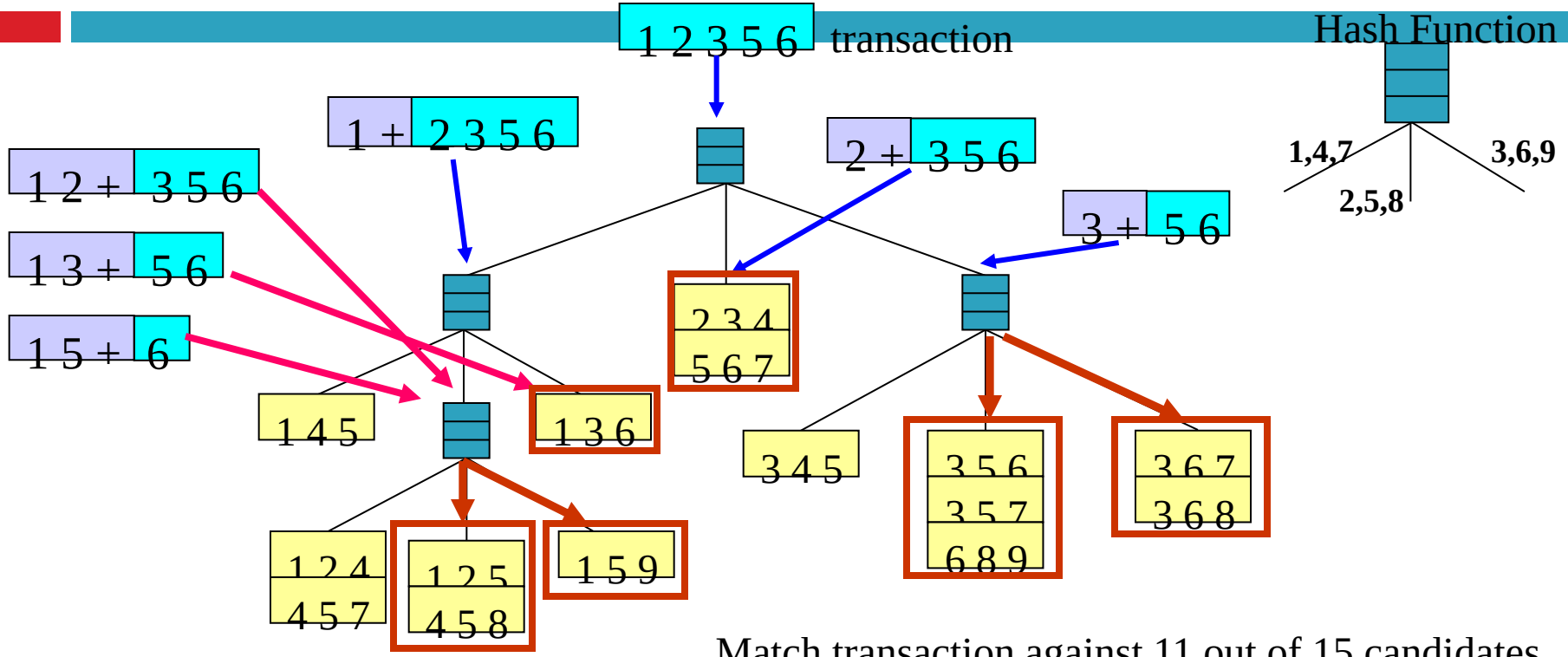
κατακερματίζουμε όλα τα 3-στοιχειοσύνολα της δοσοληψίας στο δέντρο

και αυξάνουμε τον αντίστοιχο μετρητή

Απαρίθμηση Υποσυνόλων με χρήση του Δέντρου Κατακερματισμού



Απαρίθμηση Υποσυνόλων με χρήση του Δέντρου Κατακερματισμού



Match transaction against 11 out of 15 candidates

A-priori: Πολυπλοκότητα

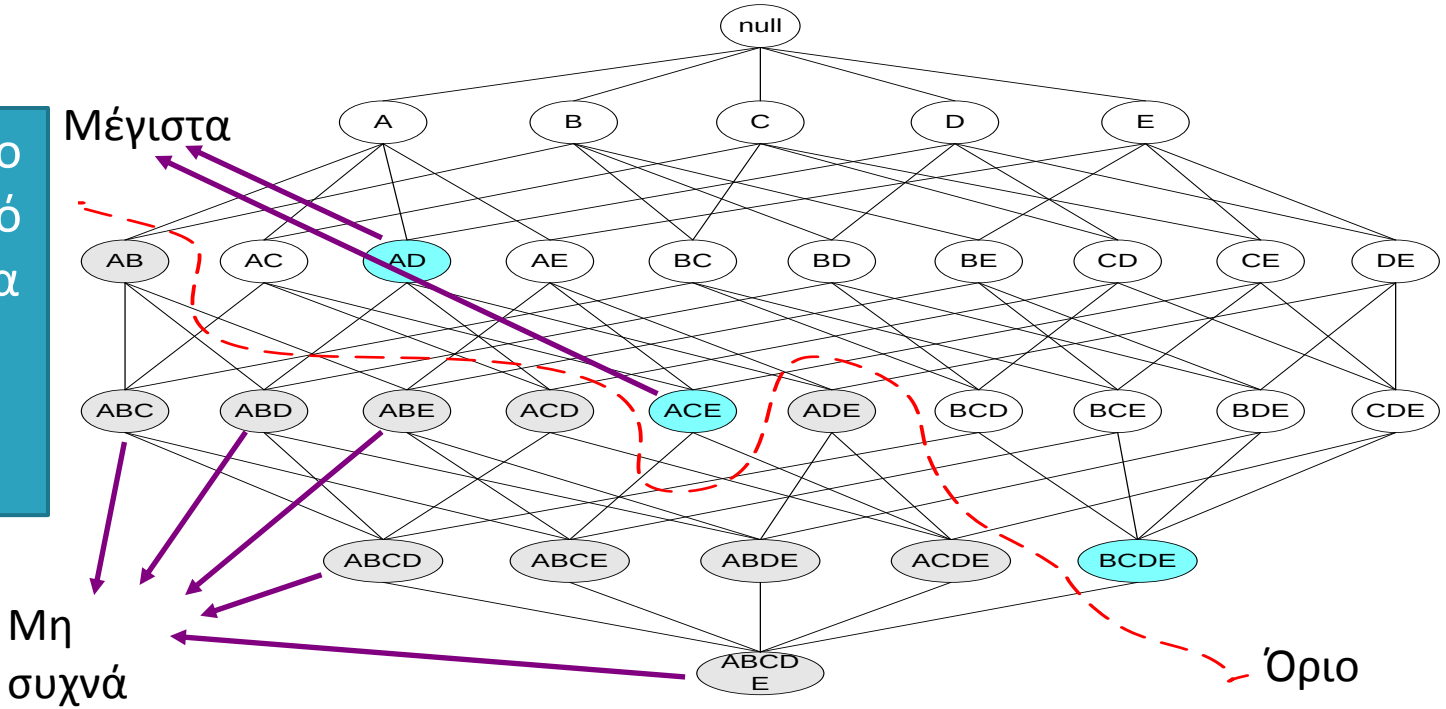
- Επιλογή κατωφλίου υποστήριξης
  - Η μείωση του κατωφλίου επιφέρει περισσότερα συχνά στοιχειοσύνολα
  - Μπορεί να αυξηθεί ο αριθμός των υποψηφίων και το μέγιστο μήκος των συχνών στοιχειοσυνόλων
- Διαστατικότητα (αριθμών στοιχείων)
  - Περισσότερος χώρος για να αποθηκευθεί η υποστήριξη ενός στοιχείου
- Μέγεθος της βάσης
  - Αφού ο apriori κάνει πολλαπλά περάσματα, ο χρόνος εκτέλεσης του αλγόριθμου αυξάνει με τον αριθμό των δοσοληψιών
- Μέσο πλάτος δοσοληψιών
  - Το πλάτος αυξάνει σε πιο πυκνά σύνολα δεδομένων

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
    - $\{b\} \rightarrow \{d, e\}$ περιττός
    - $\{b, c\} \rightarrow \{d, e\}$

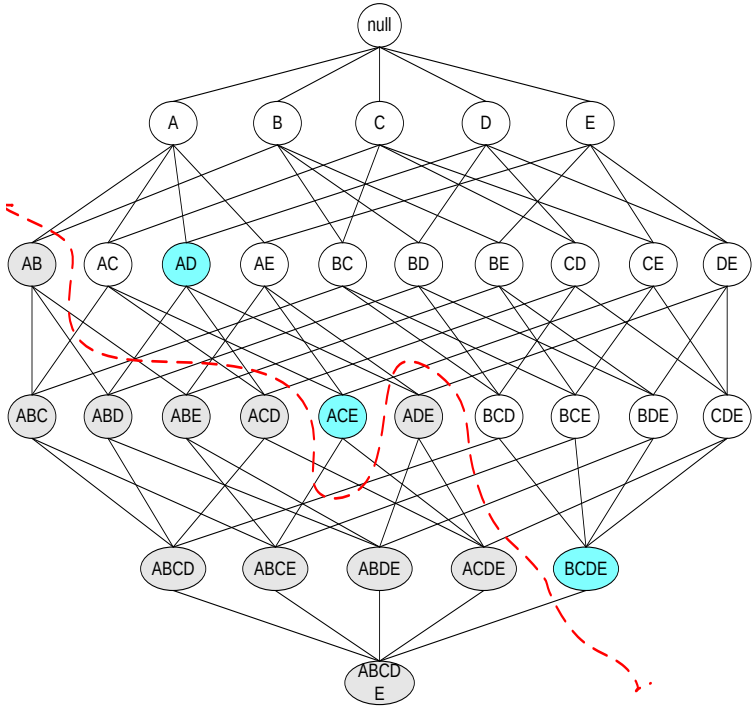
Στοιχειοσύνολο μέγιστης συχνότητας

Ένα στοιχειοσύνολο είναι μέγιστα συχνό όταν κανένα από τα αμέσως υπερσύνολα του δεν είναι συχνό



Αναπαράσταση στοιχειοσυνόλων

- Προσφέρουν μια συνοπτική αναπαράσταση των συχνών στοιχειοσυνόλων
- Το μικρότερο σύνολο στοιχειοσυνόλων από το οποίο μπορούμε να πάρουμε όλα τα συχνά στοιχειοσύνολα – είναι τα υποσύνολά τους
- Βέβαια χρειαζόμαστε 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



Κλειστό στοιχειοσύνολο

- Ένα στοιχειοσύνολο είναι κλειστό αν κανένα από τα αμέσως υπερσύνολα του δεν 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

Κλειστό συχνό στοιχειοσύνολο

- 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,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

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

Συμπιεσμένη αναπαράσταση συχνών στοιχειοσυνόλων

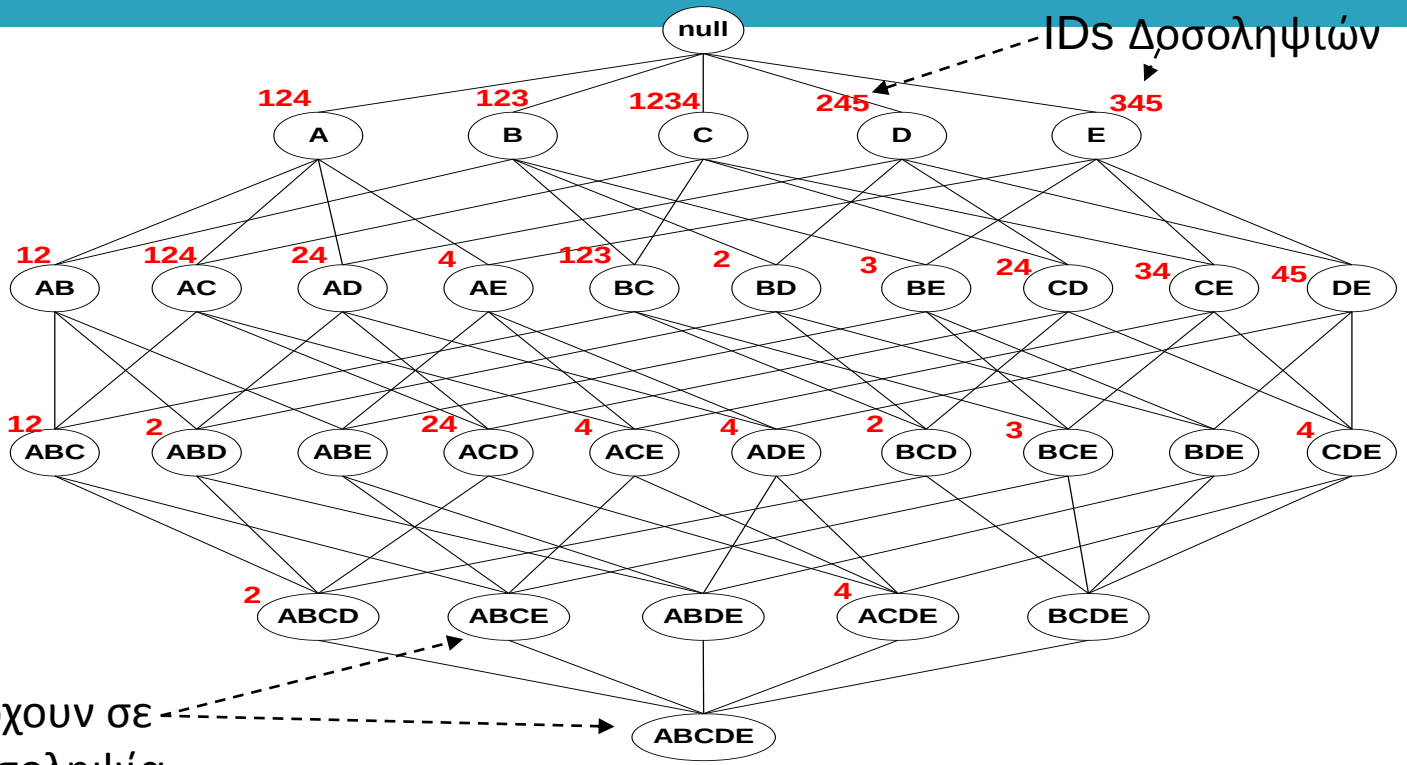
- Ορισμένα στοιχειοσύνολα είναι περιττά επειδή έχουν ίδια υποστήριξη με τα υπερσύνολα τους

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

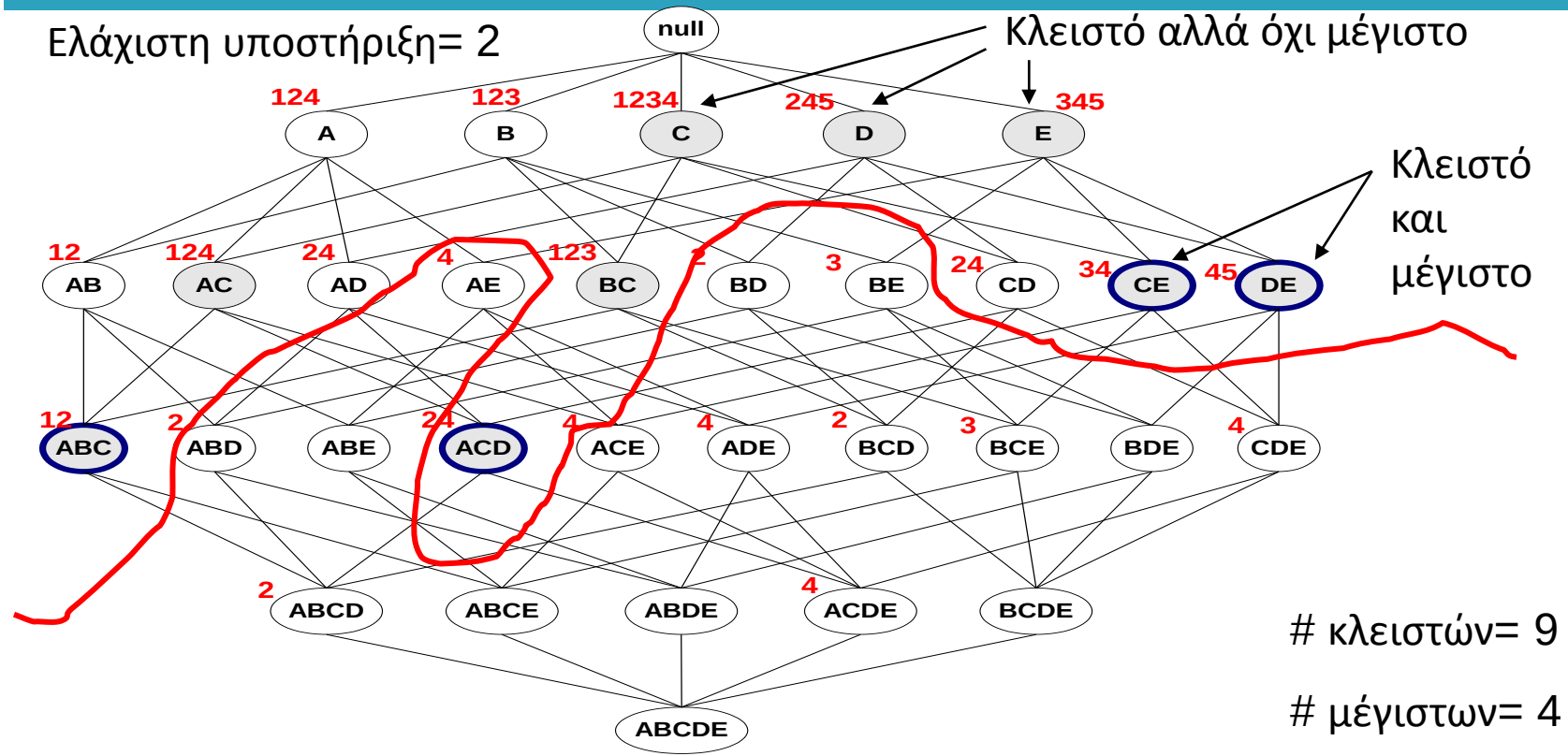
- Αριθμός στοιχειοσυνόλων $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$
- Απαιτείται πιο συμπιεσμένη αναπαράσταση

Μέγιστο vs Κλειστό Στοιχειοσύνολο

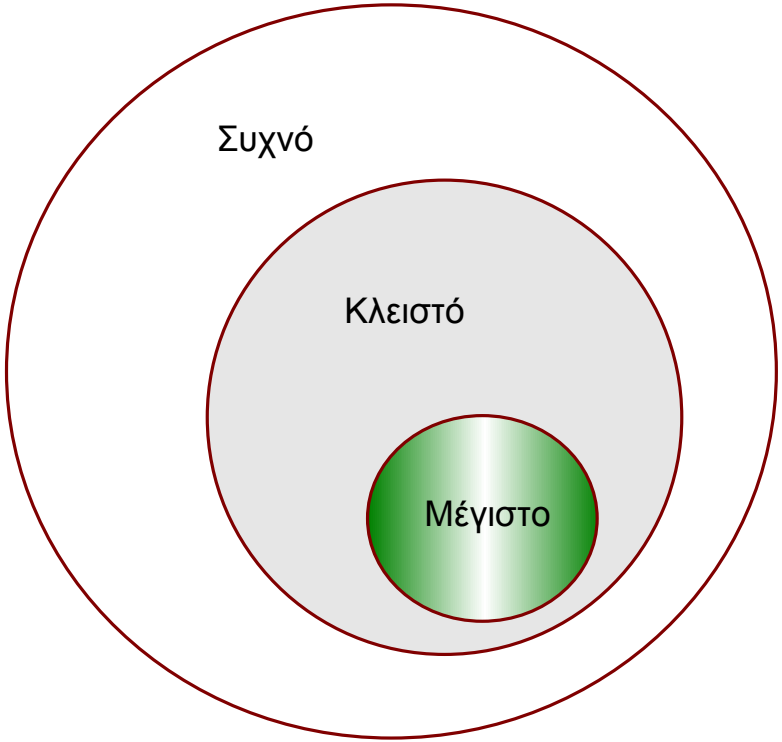
TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Μέγιστο vs Κλειστό Στοιχειοσύνολο

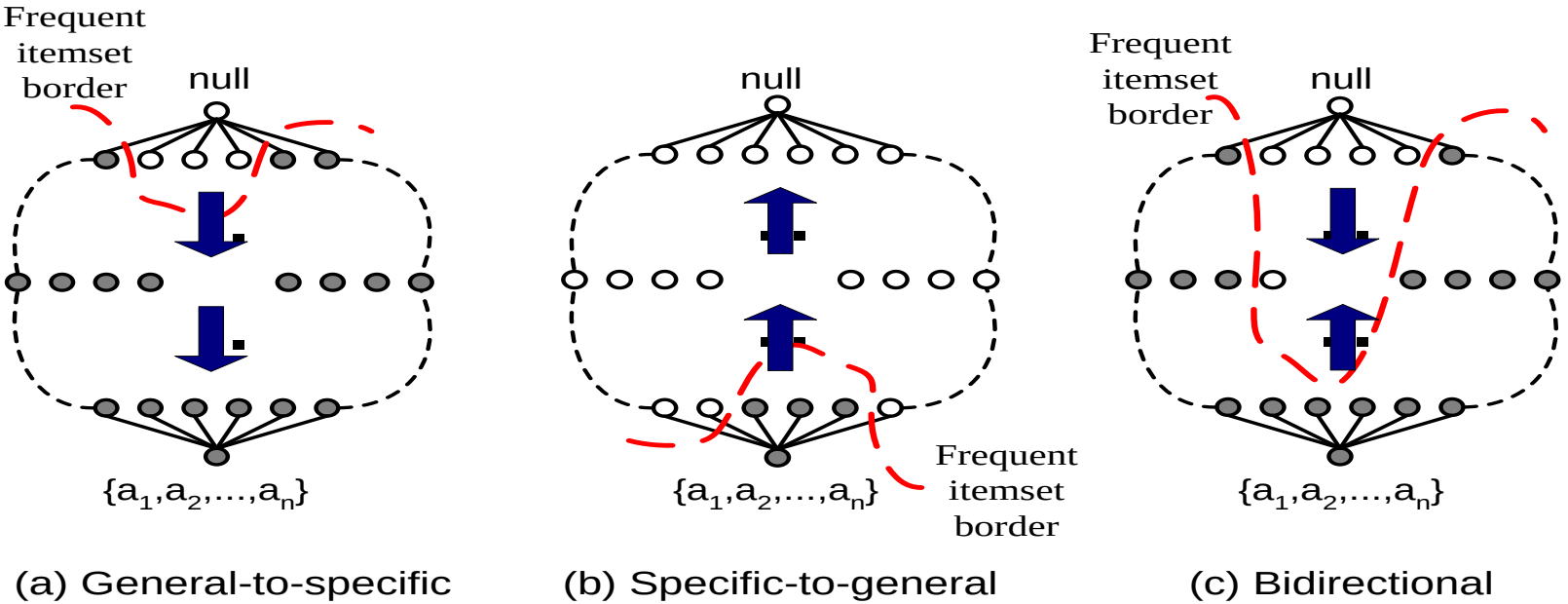


Μέγιστο vs Κλειστό Στοιχειοσύνολο



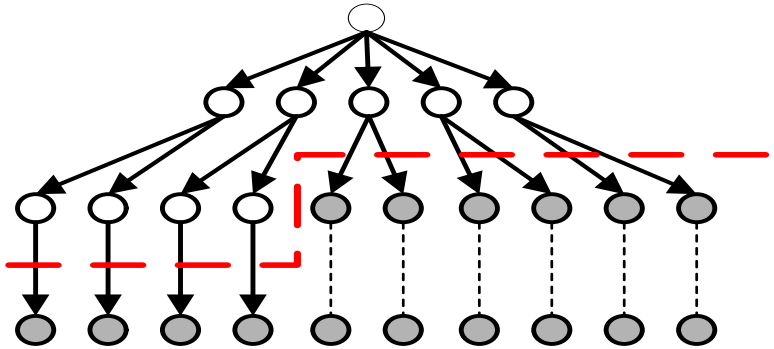
Εναλλακτικές μέθοδοι για δημιουργία συχνών στοιχειοσυνόλων

- Να διασχίσουμε το Itemset Lattice
  - Γενικό-προς-ειδικό και αντίστροφα

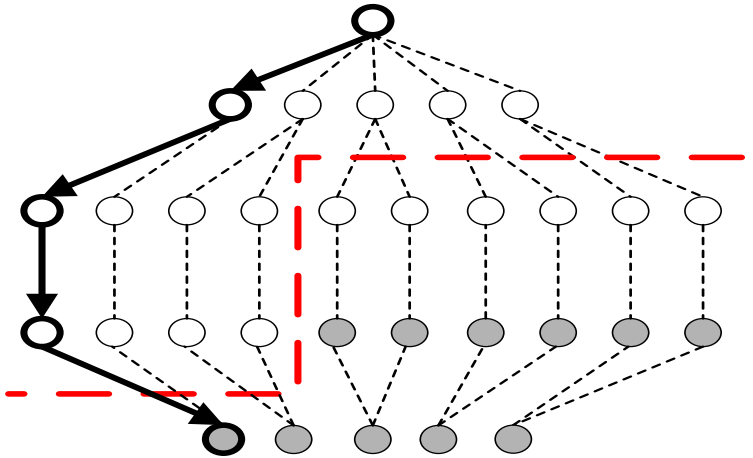


Εναλλακτικές μέθοδοι για δημιουργία συχνών στοιχειοσυνόλων

- Να διασχίσουμε το Itemset Lattice
  - Κατά πλάτος ή κατά βάθος



(a) Breadth first



(b) Depth first

Εναλλακτικές μέθοδοι για δημιουργία συχνών στοιχειοσυνόλων

- Αναπαράσταση βάσης
  - Οριζόντια ή κάθετη

Οριζόντια παράθεση δεδομένων

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Κάθετη παράθεση δεδομένων

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

Ο αλγόριθμος FP-growth

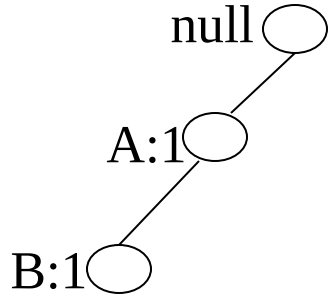
- Χρησιμοποιεί μια συμπιεσμένη αναπαράσταση της βάσης δεδομένων με βάση ένα FP-tree

- Όταν κατασκευασθεί ένα, χρησιμοποιεί μια μέθοδο ***διαίρει και βασίλευε*** για να εξορύξει τα συχνά στοιχειοσύνολα

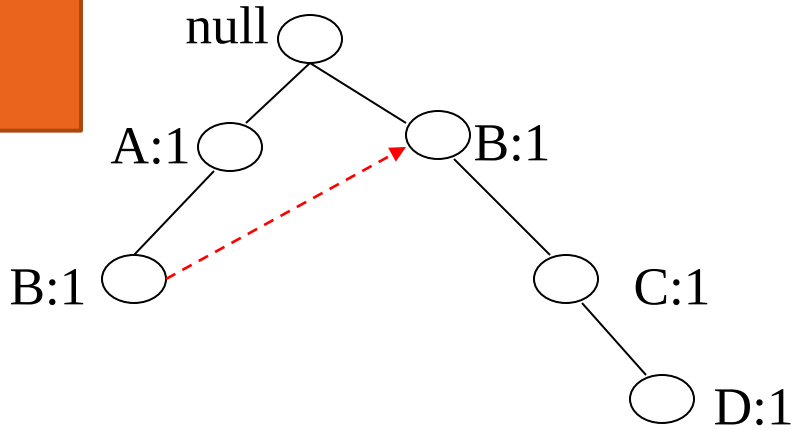
FP-tree

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Μόλις περάσει την εγγραφή TID=1:



Μόλις περάσει την εγγραφή TID=2:



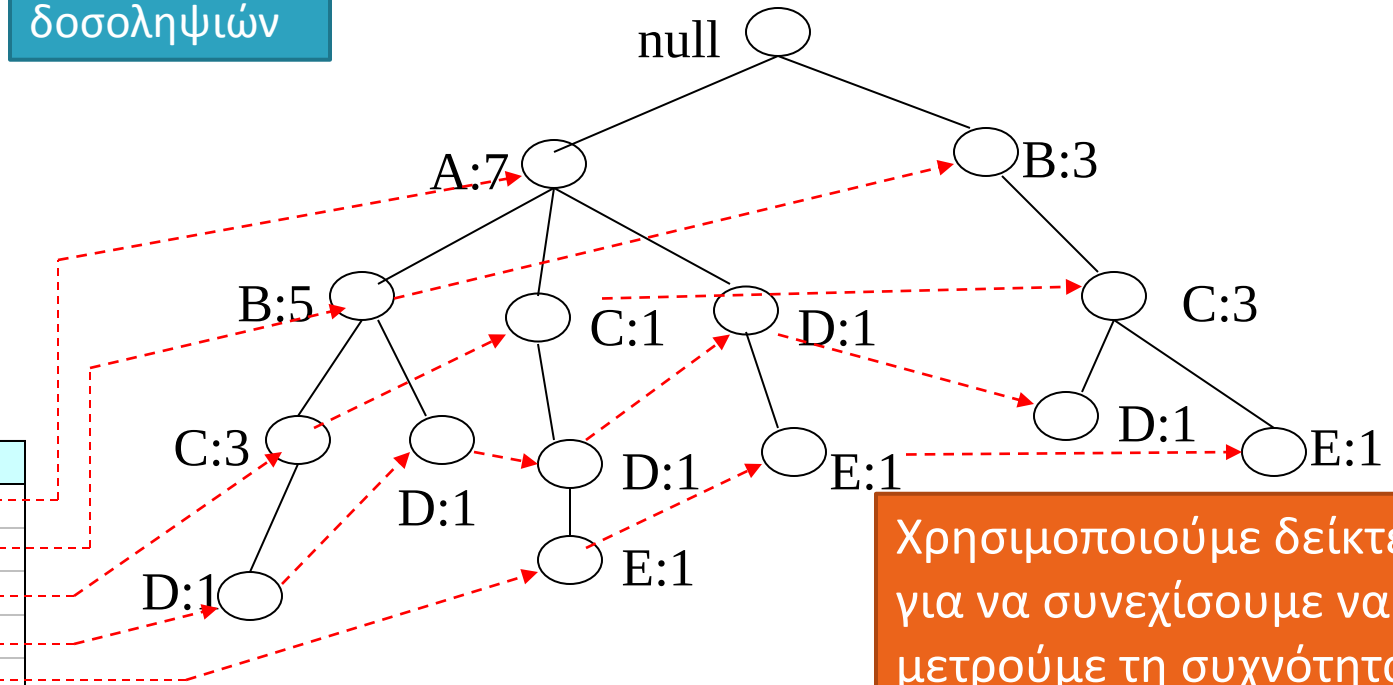
FP-Tree

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Βάση δοσοληψιών

Header table

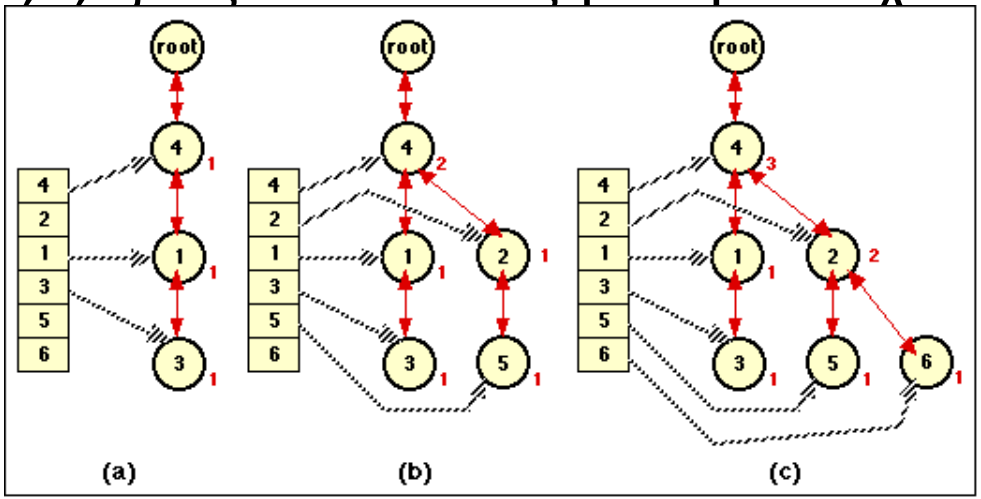
Item	Pointer
A	
B	
C	
D	
E	



Χρησιμοποιούμε δείκτες για να συνεχίσουμε να μετρούμε τη συχνότητα

FP-Growth (Παράδειγμα)

- Έστω το σύνολο: {1, 3, 4} {2, 4, 5} {2, 4, 6}
- Υπολογίζουμε τα support κάθε item και έχουμε:
  - {4,2,1,3,5,6} ως νέα διάταξη. Στη συνέχεια:



FP-Tree (Μέγεθος)

- Κάθε δοσοληψία αντιστοιχεί σε ένα μονοπάτι από τη ρίζα
- Το μέγεθος του δέντρου συνήθως μικρότερο των δεδομένων, αν υπάρχουν κοινά προθέματα
- Αν όλες οι δοσοληψίες τα ίδια δεδομένα, μόνο ένα κλαδί
- Αν όλες διαφορετικές, ο χώρος μεγαλύτερος (γιατί αποθηκεύεται περισσότερη πληροφορία, όπως δείκτες μεταξύ των κόμβων αλλά και συχνότητες εμφάνισης)

Κατασκευή FP-δέντρου

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Το τελικό δέντρο, εξαρτάται από τη διάταξη: άλλη διάταξη -> άλλα προθέματα

(Συνήθως) Μικρότερο δέντρο, αν όχι λεξικογραφικά, αλλά με βάση τη συχνότητα εμφάνισης -> Αρχικά, διαβάζουμε όλα τα δεδομένα μια φορά ώστε να υπολογιστεί ο μετρητής υποστήριξης κάθε στοιχείου, και διατάσουμε τα στοιχεία με βάση αυτό

Για το παράδειγμα, σ(A)=7, σ(B)=8, σ(C)=7, σ(D)=5, σ(E)=3

Άρα, διάταξη B,A,C,D,E

TID	Items
1	{B,A}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{B,A,C}
6	{B,A,C,D}
7	{B,C}
8	{B,A,C}
9	{B,A,D}
10	{B,C,E}

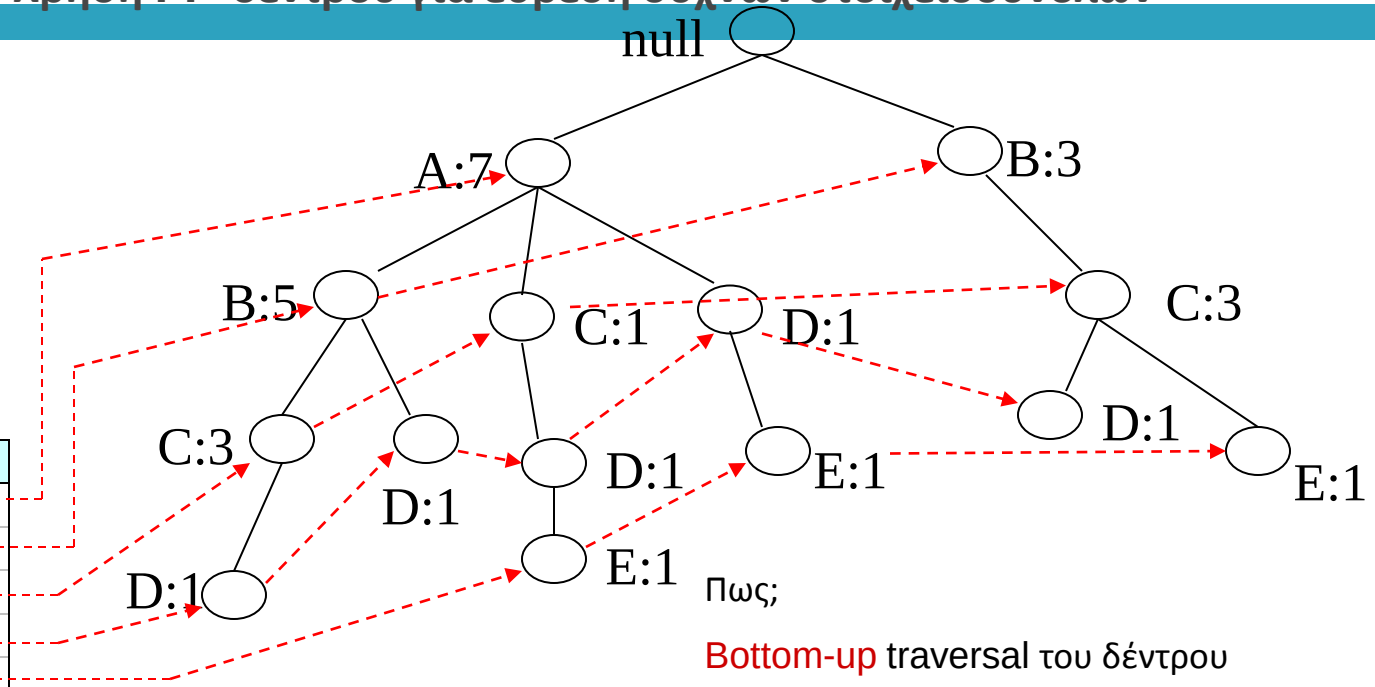
Αλγόριθμος FP-Growth

Χρήση FP-δέντρου για εύρεση συχνών στοιχειοσυνόλων

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Header table

Item	Pointer
A	
B	
C	
D	
E	



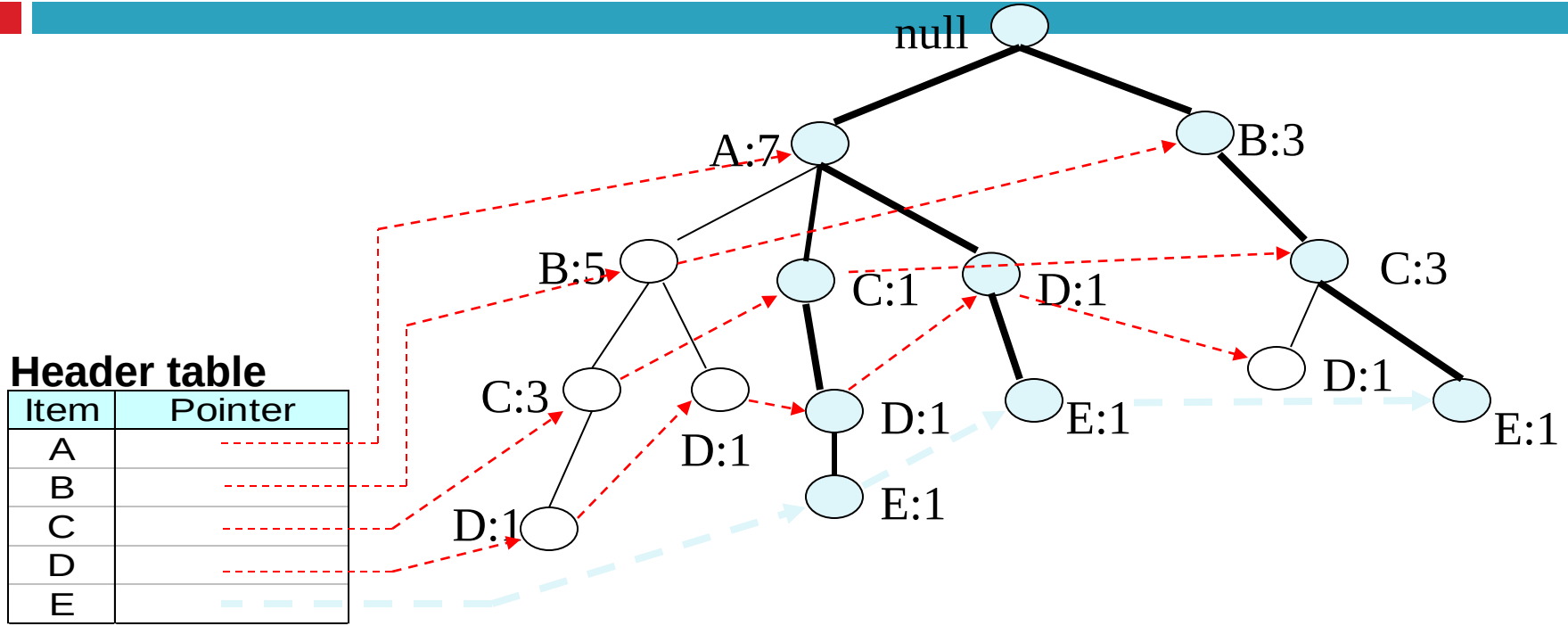
Πως;

Bottom-up traversal του δέντρου

Αυτά που τελειώνουν σε E, μετά αυτά που τελειώνουν σε D, C, B και τέλος A -suffix-based classes

Υποπρόβλημα: Βρες συχνά στοιχειοσύνολα που τελειώνουν σε **E**

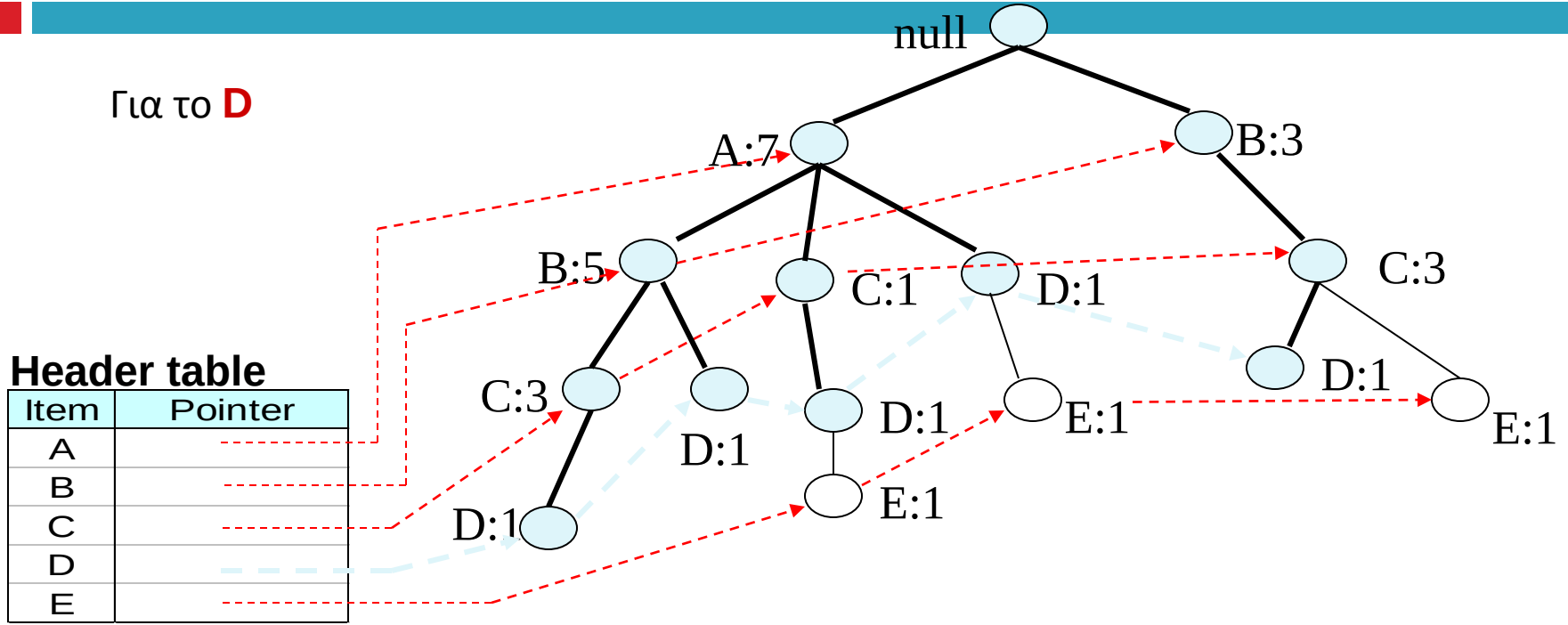
Αλγόριθμος FP-Growth



Θα δούμε στη συνέχεια πως υπολογίζεται η υποστήριξη για τα πιθανά στοιχειοσύνολα

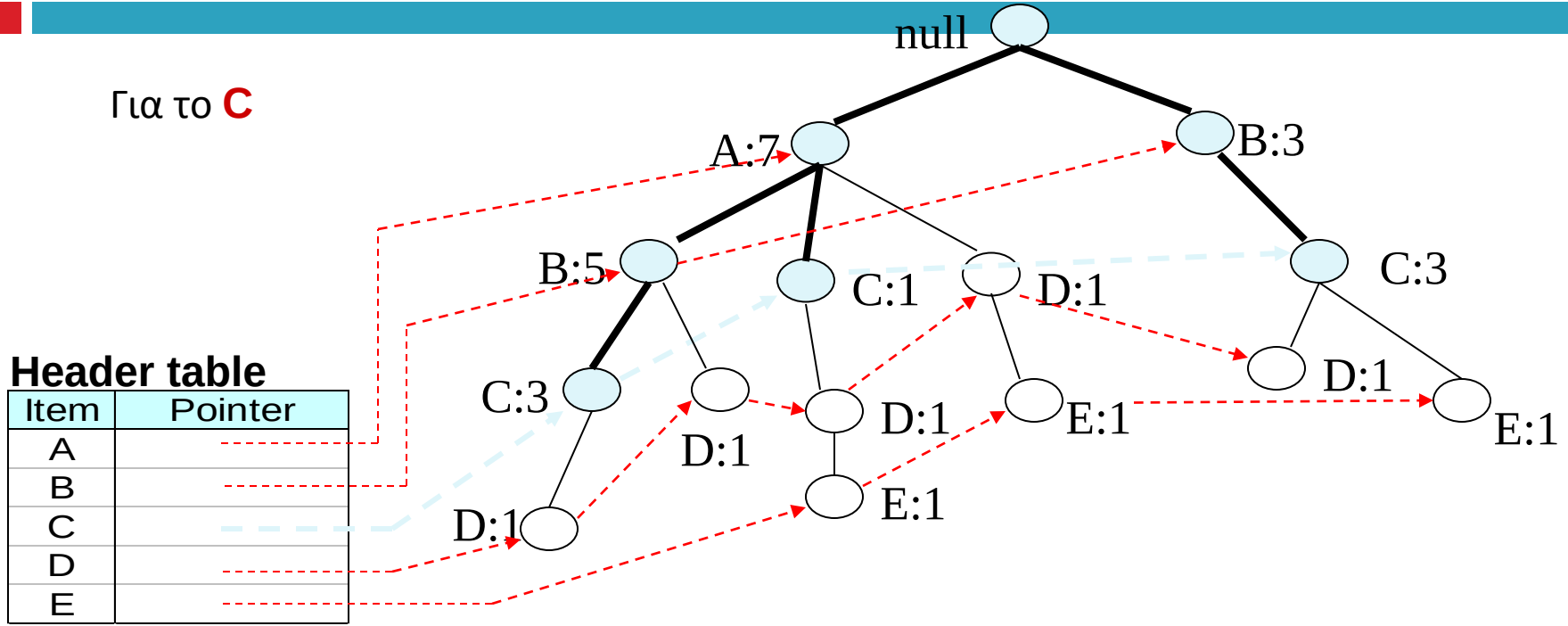
Αλγόριθμος FP-Growth

Για το **D**



Αλγόριθμος FP-Growth

Για το **C**

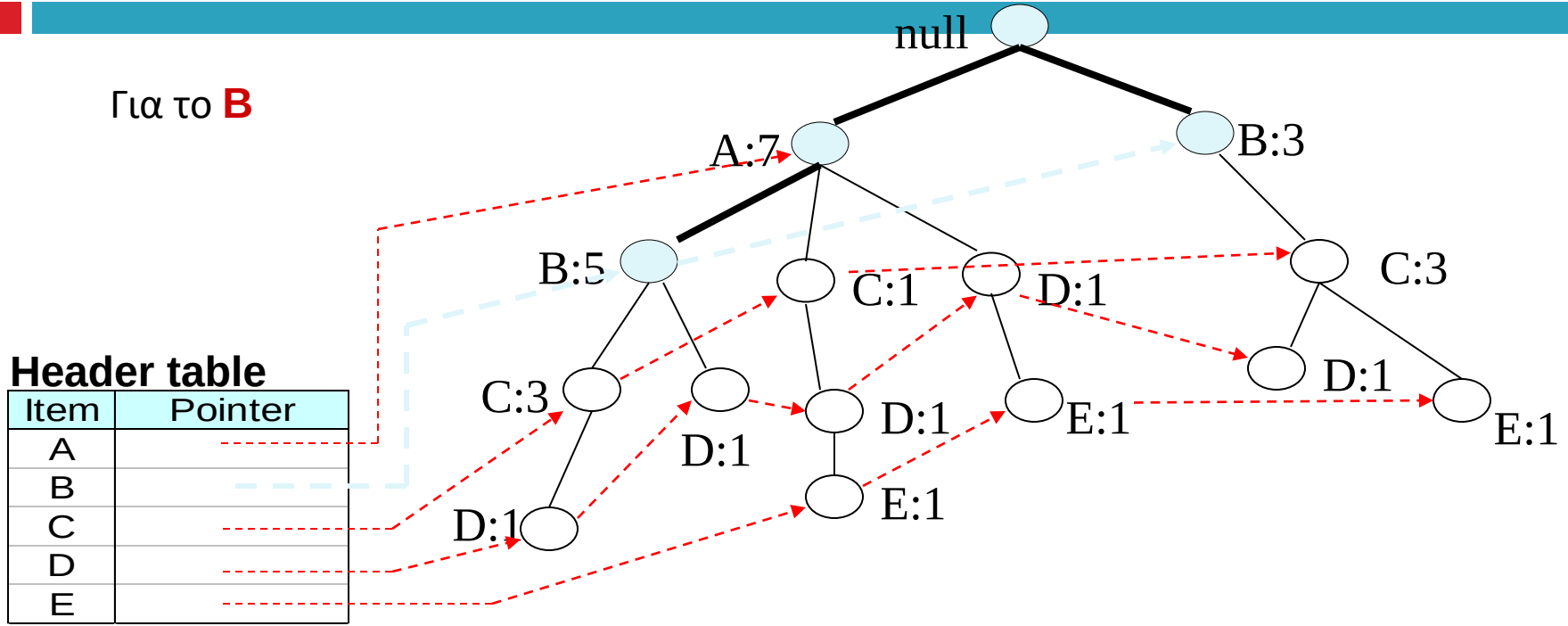


Header table

Item	Pointer
A	
B	
C	
D	
E	

Αλγόριθμος FP-Growth

Για το **B**

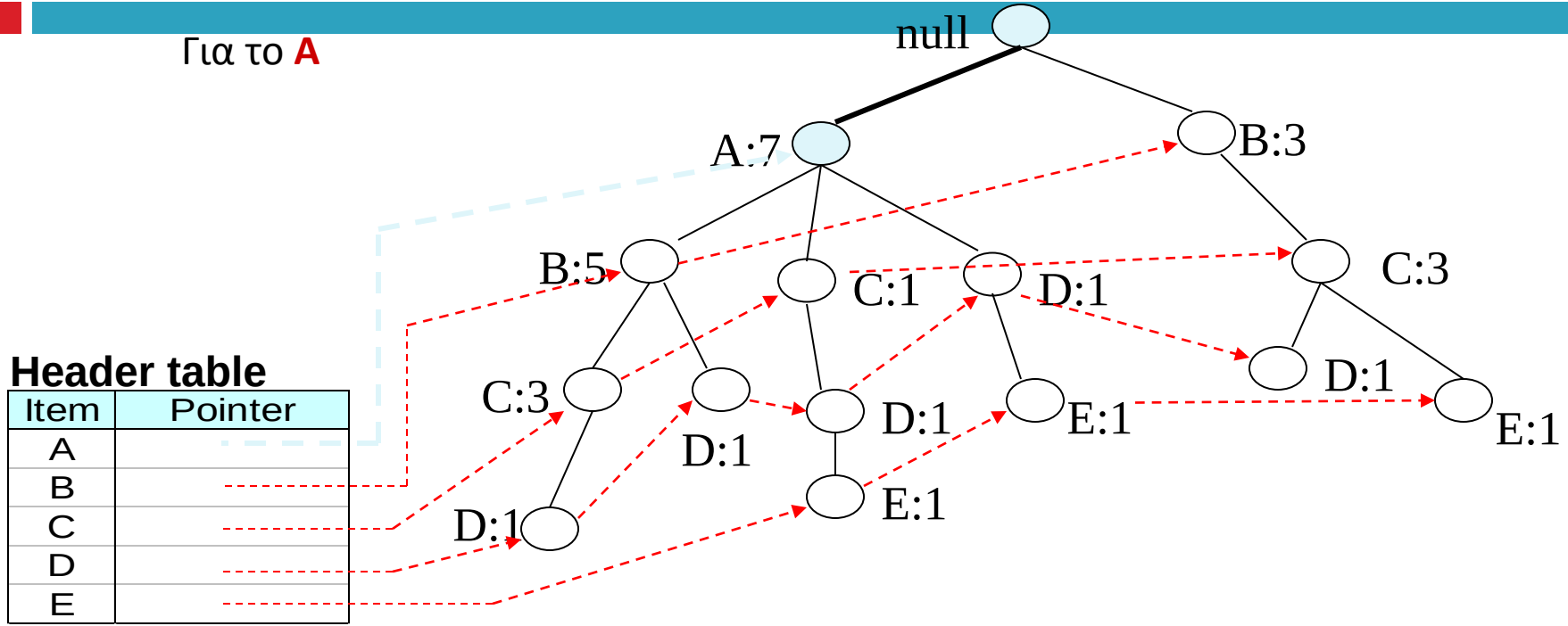


Header table

Item	Pointer
A	
B	
C	
D	
E	

Αλγόριθμος FP-Growth

Για το **A**

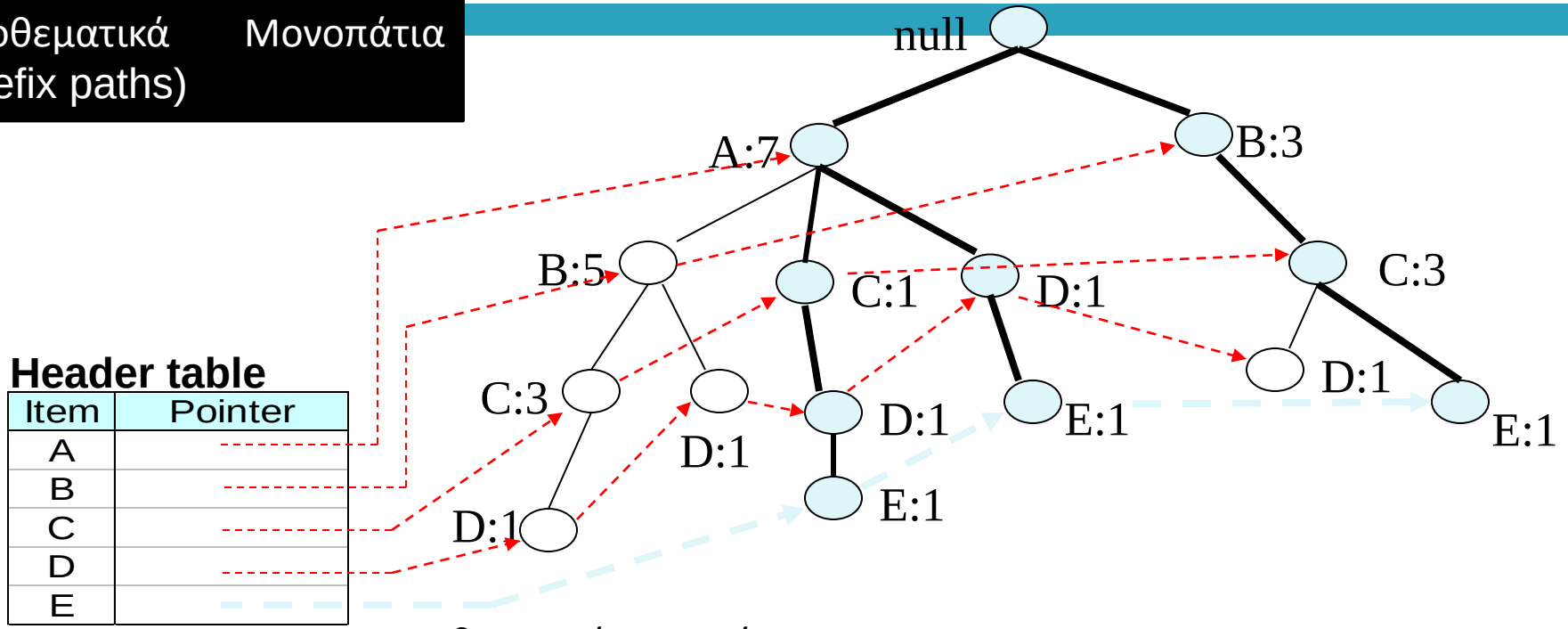


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το E

Προθεματικά Μονοπάτια (prefix paths)



Header table

Item	Pointer
A	
B	
C	
D	
E	

Προθεματικά μονοπάτια του E:

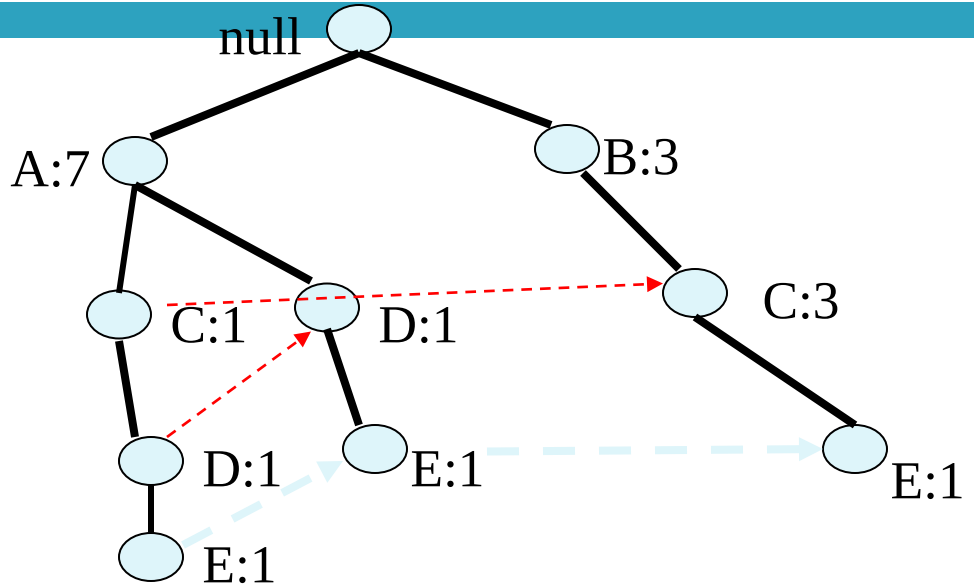
{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}

Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το E

Προθεματικά Μονοπάτια (prefix paths)



Προθεματικά μονοπάτια του E:

{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}

Αλγόριθμος FP-Growth

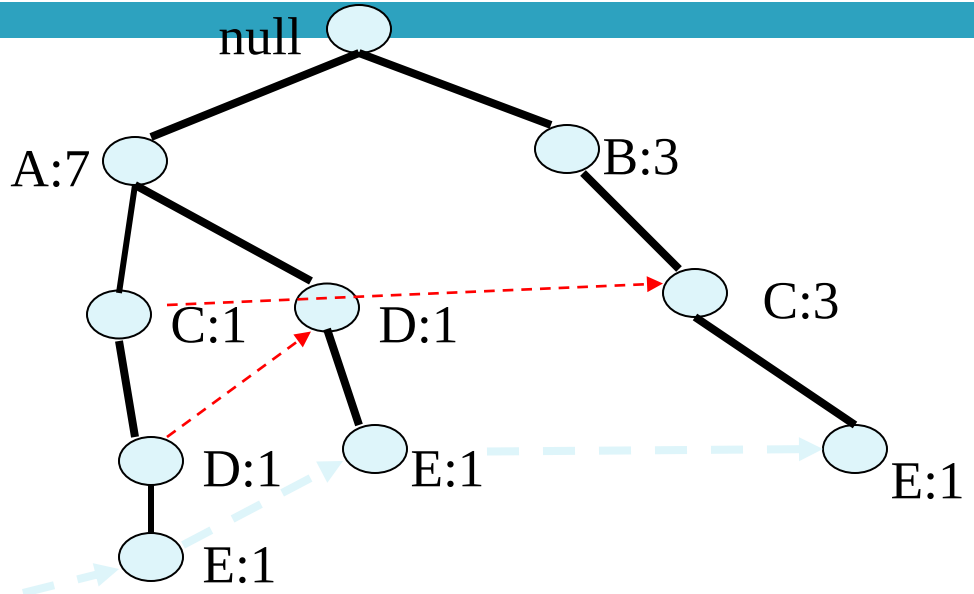
Έστω minsup = 2

Βρες την υποστήριξη του {E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1+1=3>2

Οπότε {E} συχνό



{E} συχνό άρα προχωράμε για DE, CE, BE, AE

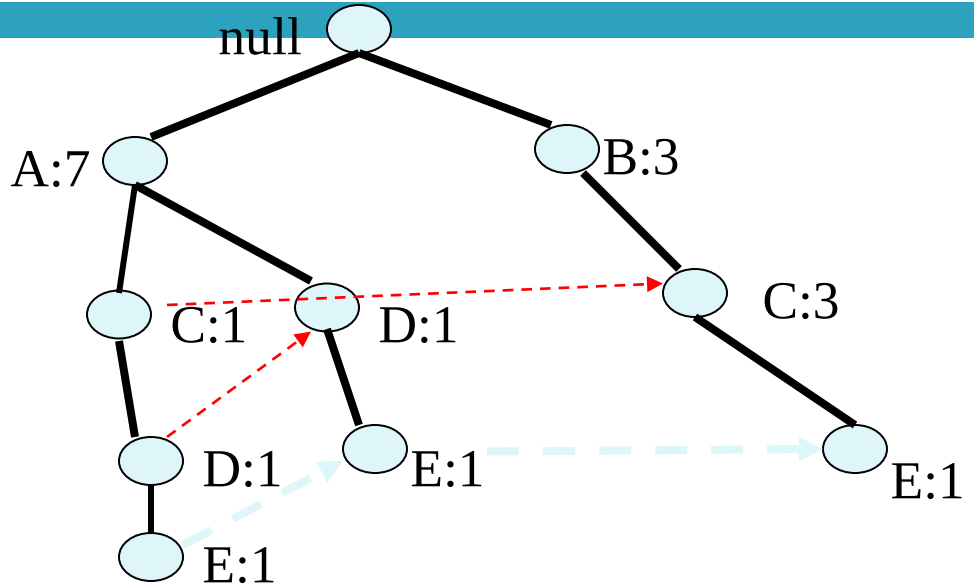
Αλγόριθμος FP-Growth

Μετατροπή των προθεματικών δέντρων σε FP-δέντρο υπό συνθήκες (conditional FP-tree)

Δύο αλλαγές

(1) Αλλαγή των μετρητών

(2) Περικοπή



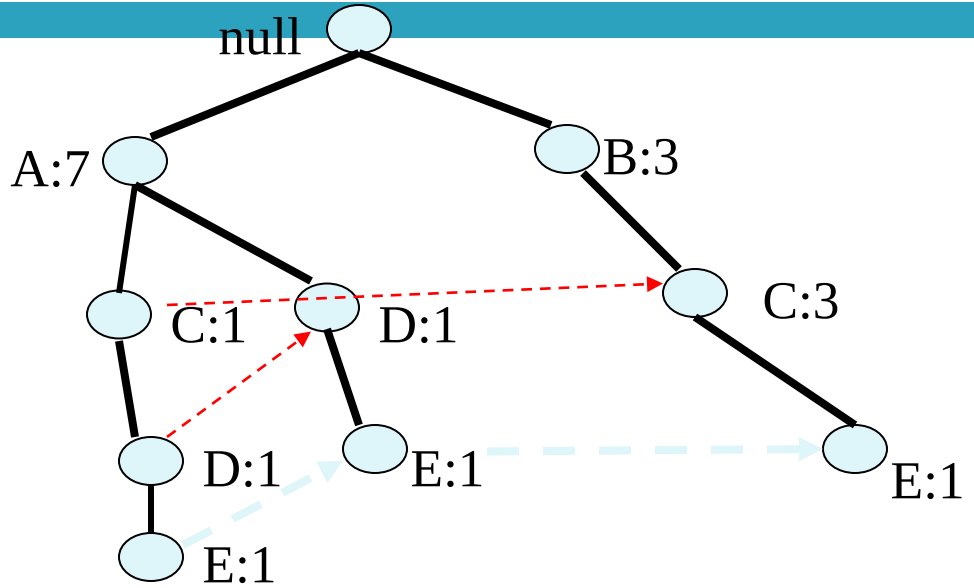
{E} συχνό άρα προχωράμε για DE, CE, BE, AE

Αλγόριθμος FP-Growth

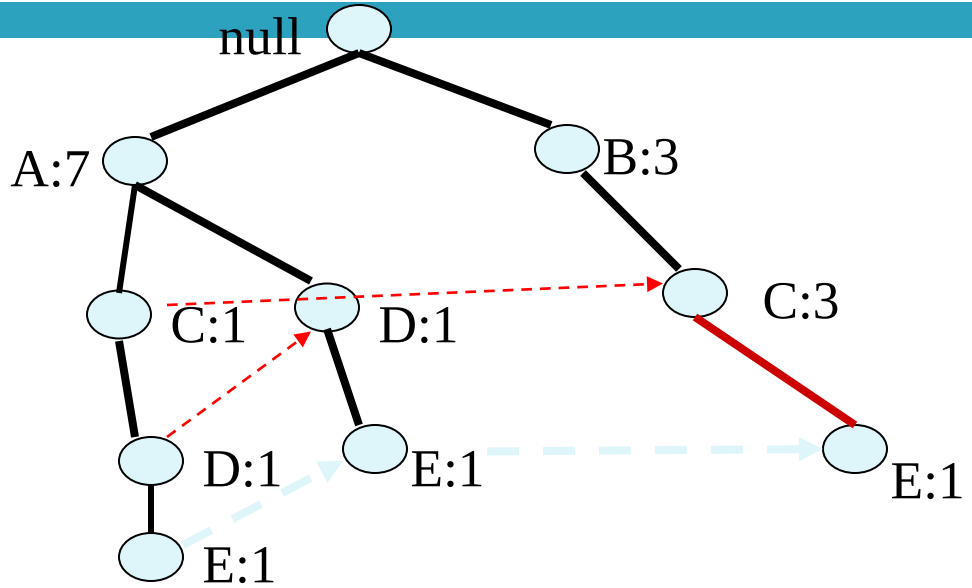
Αλλαγή μετρητών

Οι μετρητές σε κάποιους κόμβους περιλαμβάνουν δοσοληψίες που δεν έχουν το E

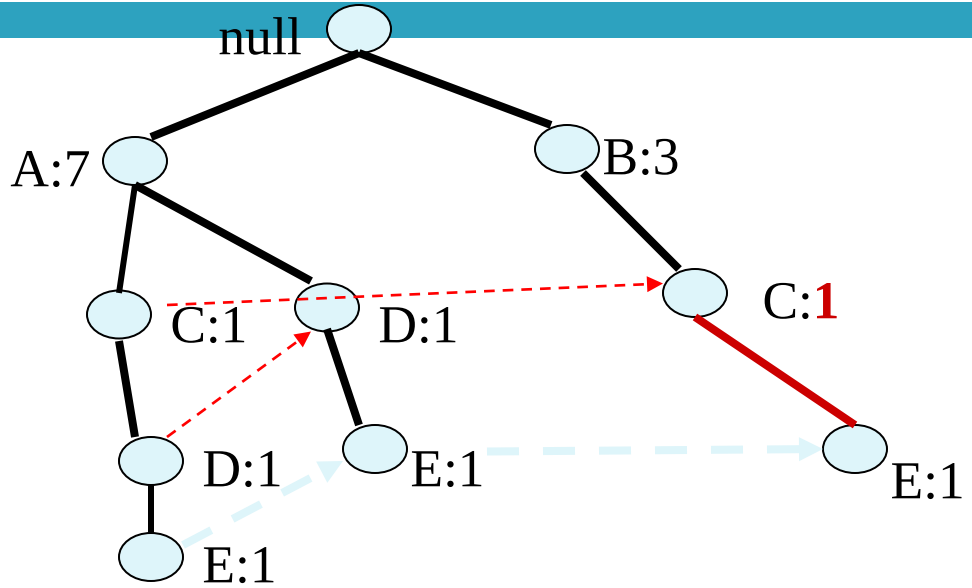
Πχ στο null->B->C->E μετράμε και την {B, C}



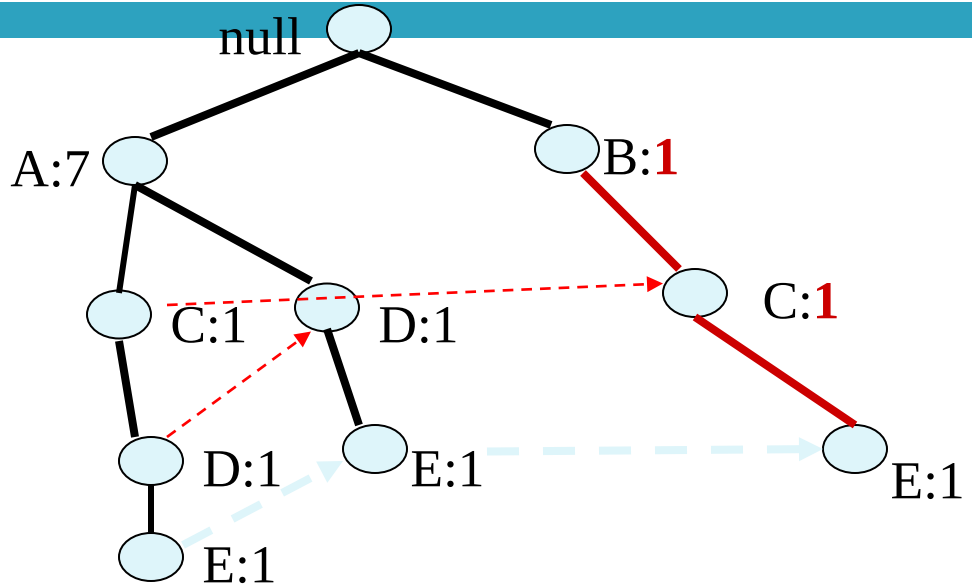
Αλγόριθμος FP-Growth



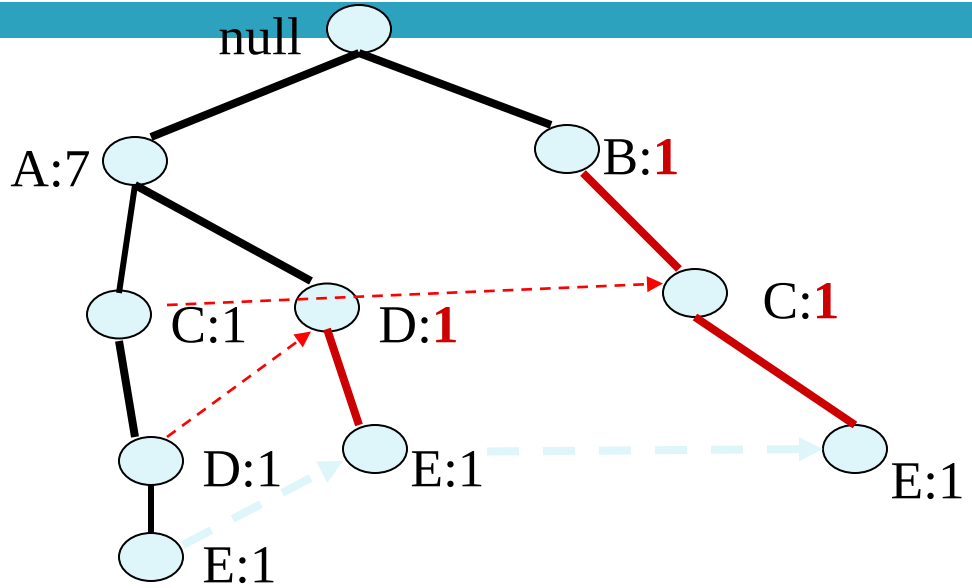
Αλγόριθμος FP-Growth



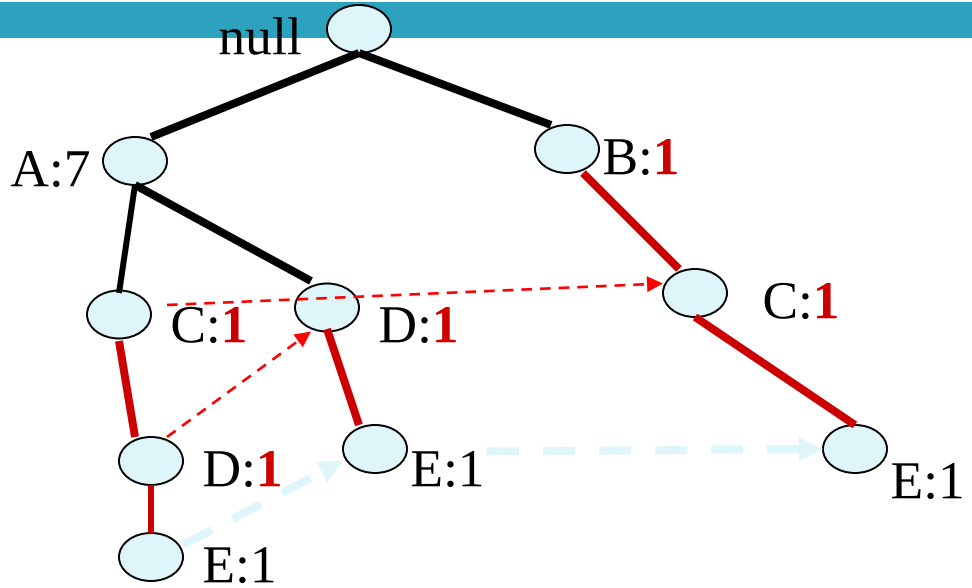
Αλγόριθμος FP-Growth



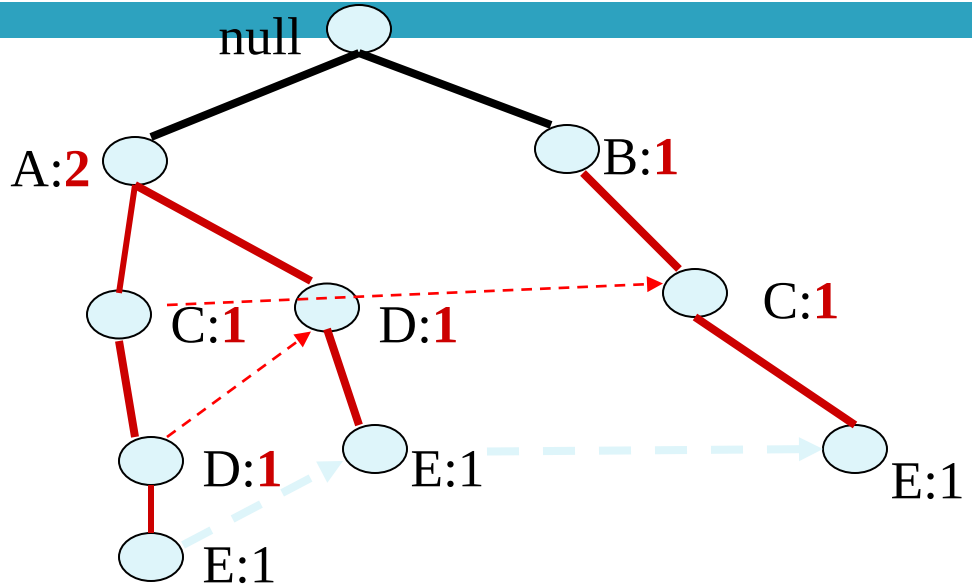
Αλγόριθμος FP-Growth



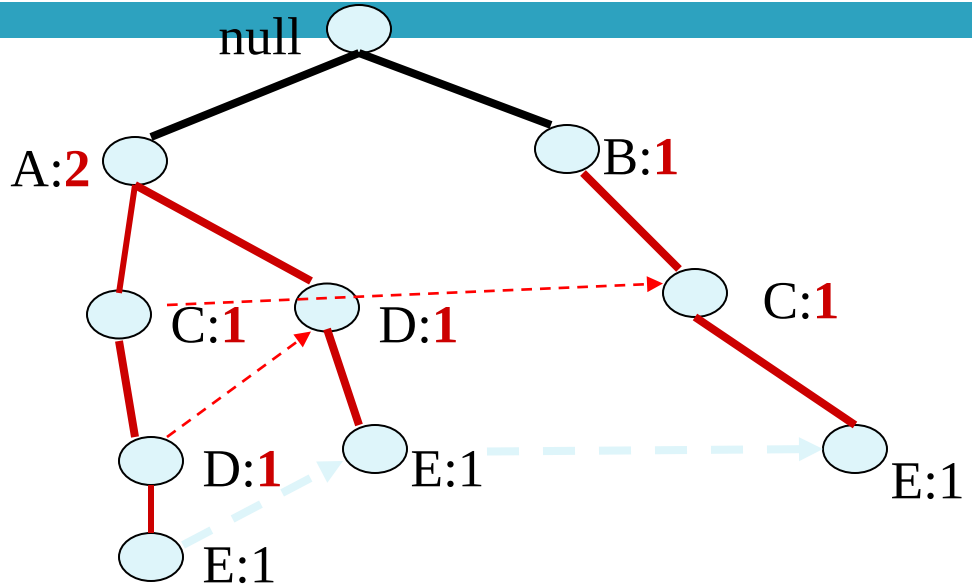
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth



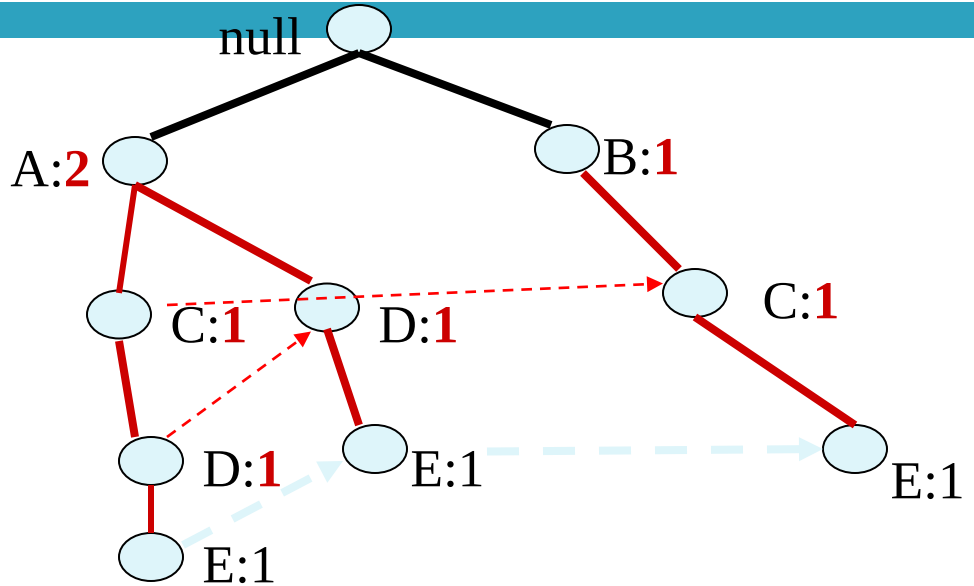
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Περικοπή (truncate)

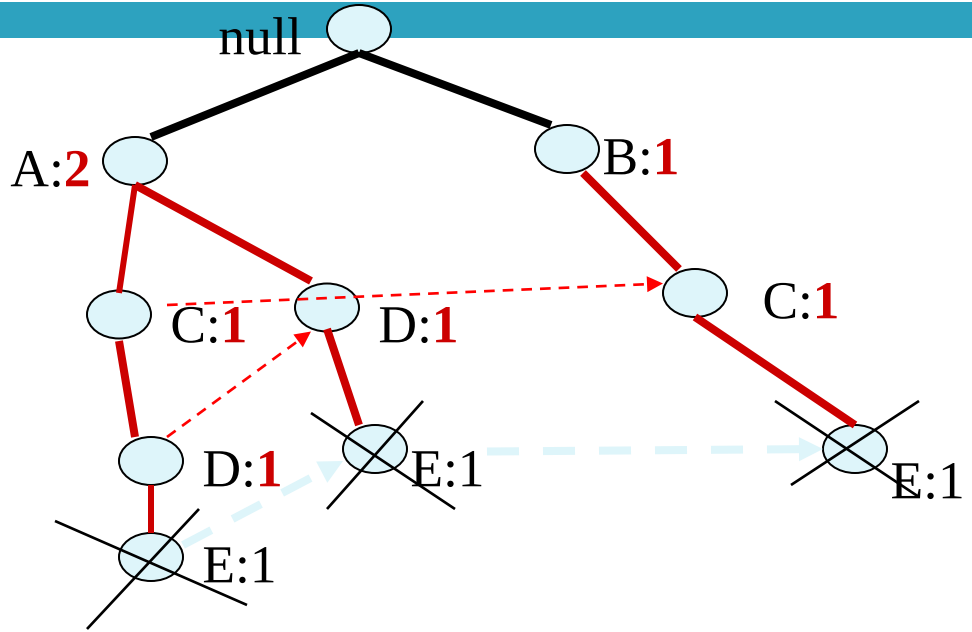
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth

Περικοπή (truncate)

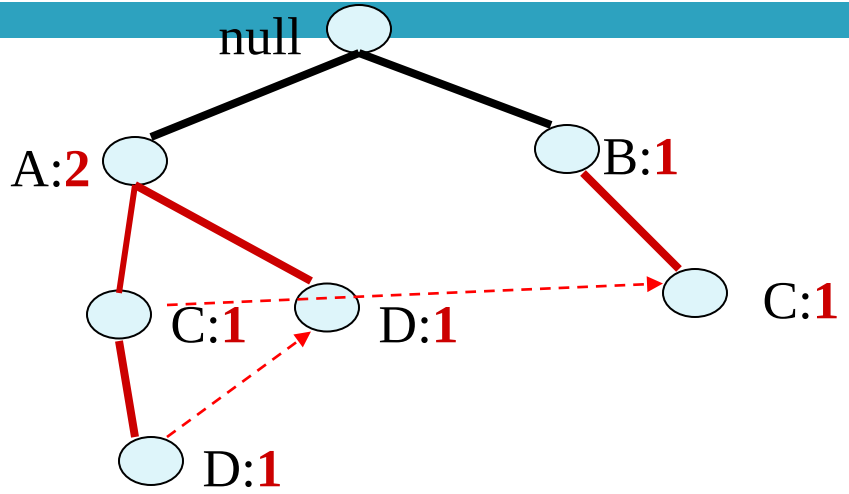
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth

Περικοπή (truncate)

Σβήσε τους κόμβους του E

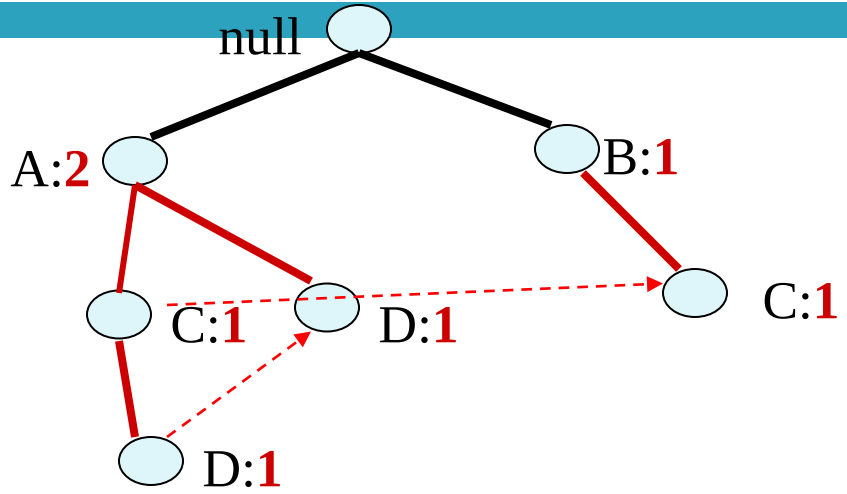


Αλγόριθμος FP-Growth

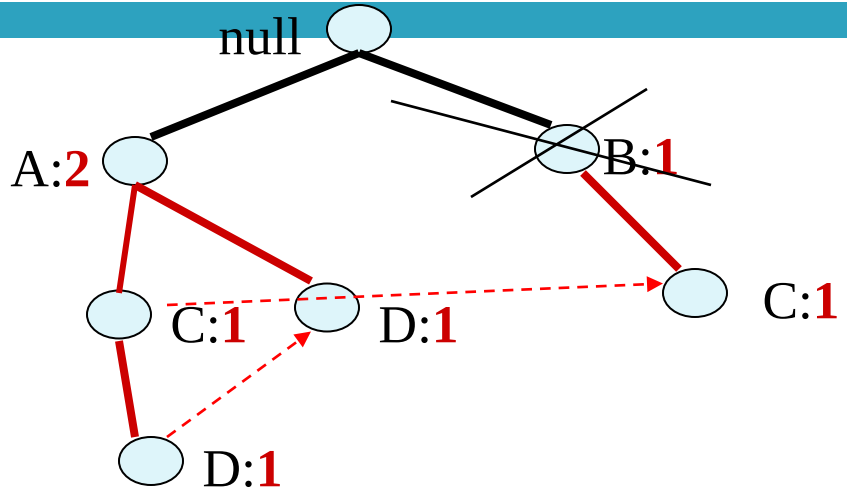
Πιθανή περαιτέρω περικοπή

Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης

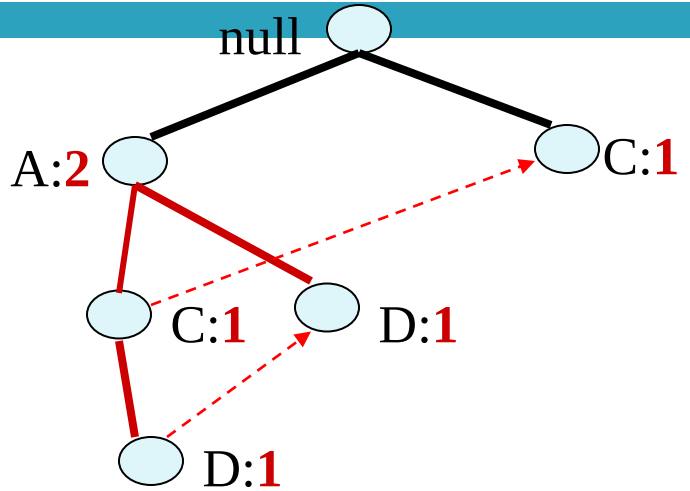
Πχ το B -> περικοπή



Αλγόριθμος FP-Growth



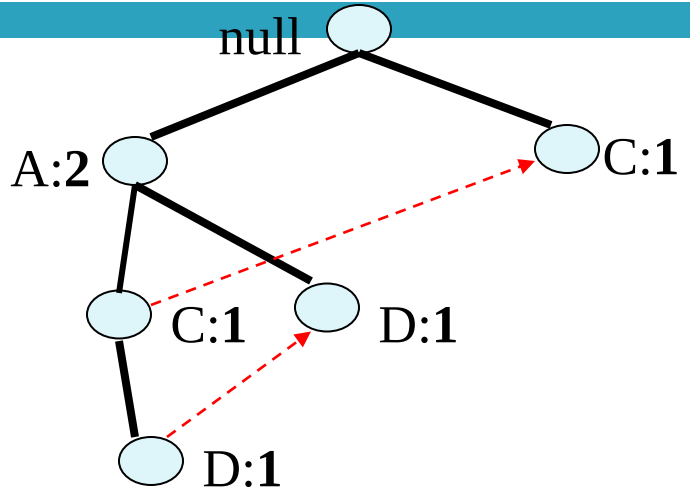
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, {C, E}, {A, E}

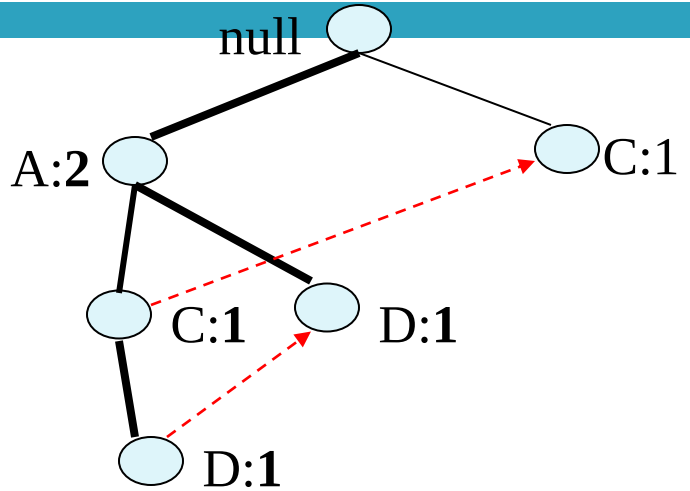


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)

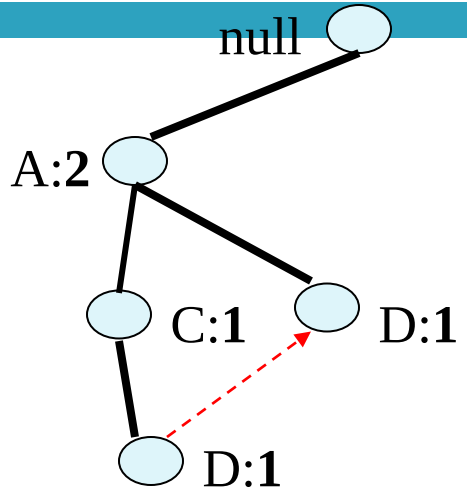


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



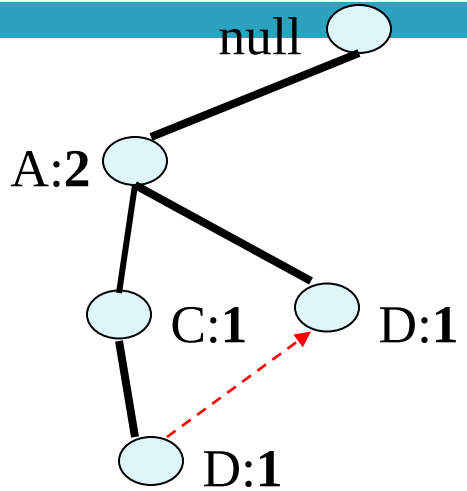
Αλγόριθμος FP-Growth

Βρες την υποστήριξη του {D, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

Οπότε {D, E} συχνό

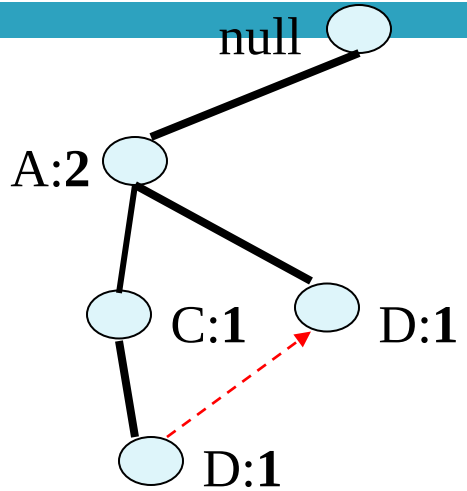


Αλγόριθμος FP-Growth

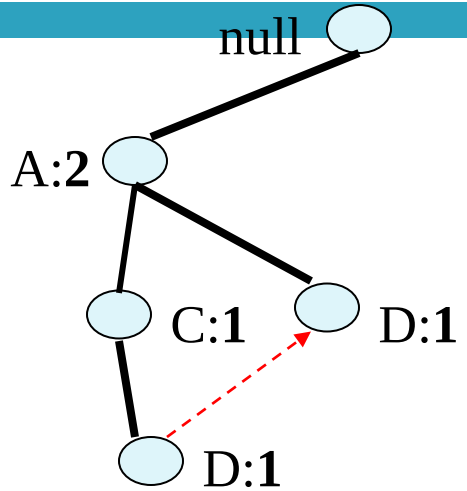
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {D, E}

1. Αλλαγή υποστήριξης

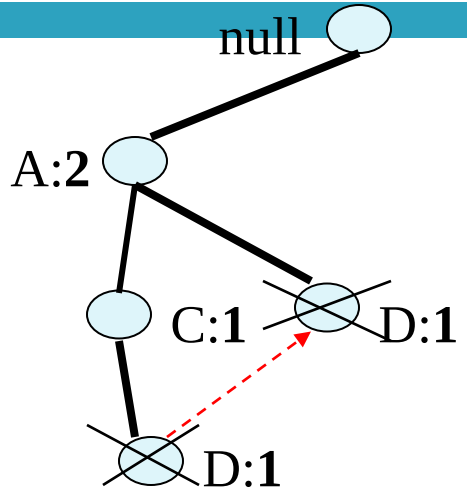
2. Περικοπές κόμβων



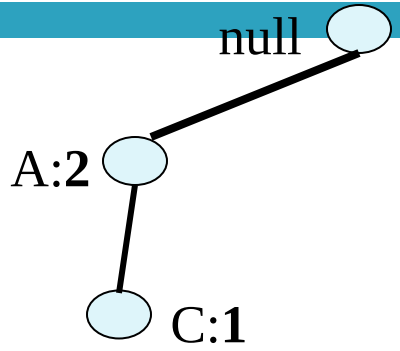
1. Αλλαγή υποστήριξης



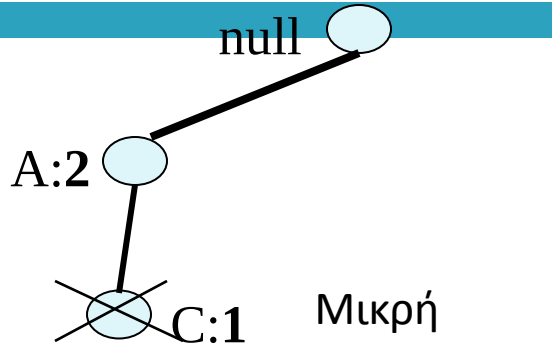
2. Περικοπές κόμβων



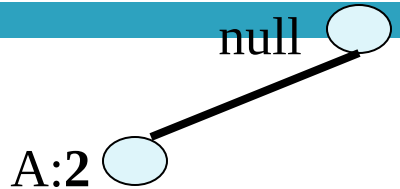
2. Περικοπές κόμβων



2. Περικοπές κόμβων



Τελικό υπο-συνθήκη FP-δέντρο για το {D, E}



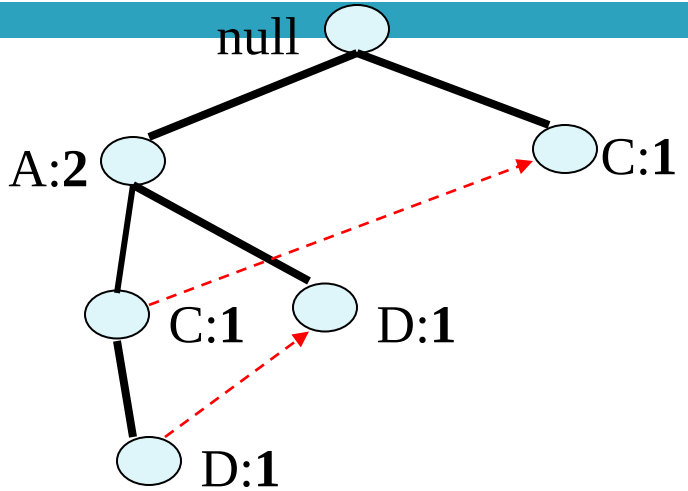
Υποστήριξη του A είναι ≥ minsup -> {A, D, E} συχνό

Αφού μόνο έναν κόμβο, επιστροφή στο επόμενο υποπρόβλημα

Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, **{C, E},** {A, E}

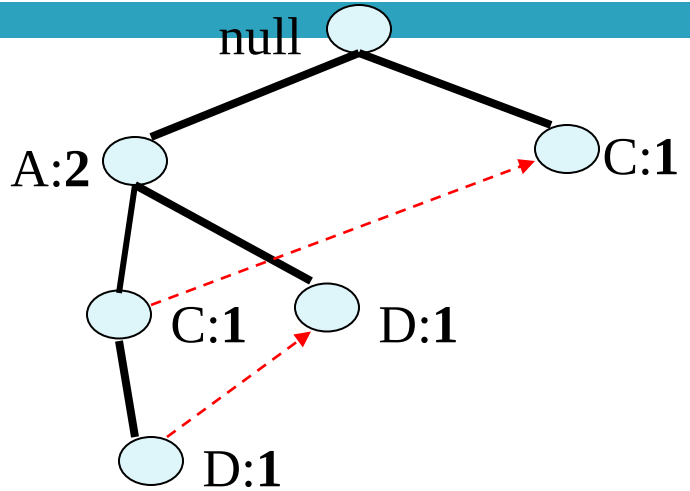


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)

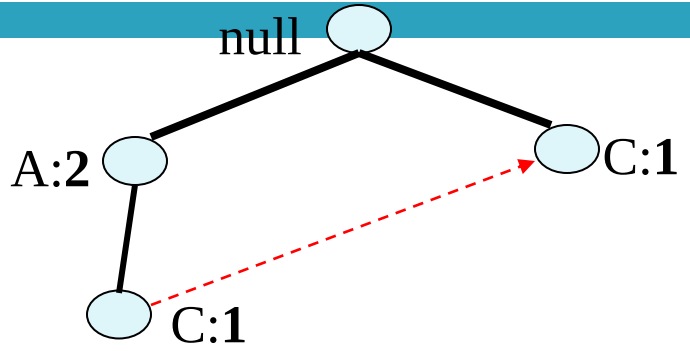


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



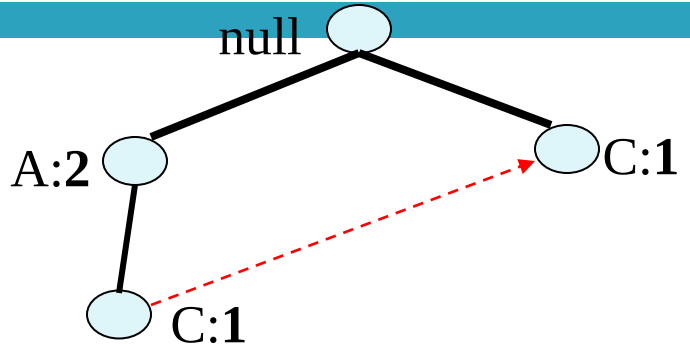
Αλγόριθμος FP-Growth

Βρες την υποστήριξη του {C, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

Οπότε {C, E} συχνό

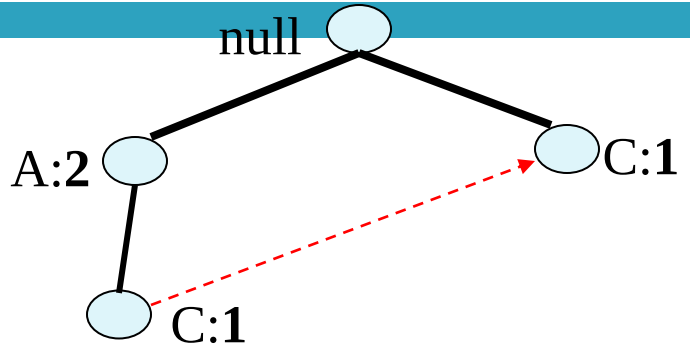


Αλγόριθμος FP-Growth

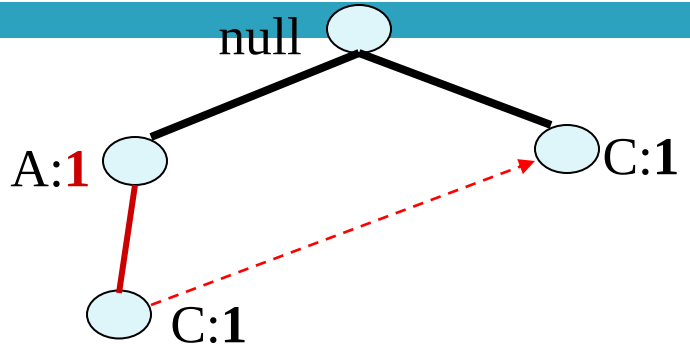
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {C, E}

1. Αλλαγή υποστήριξης

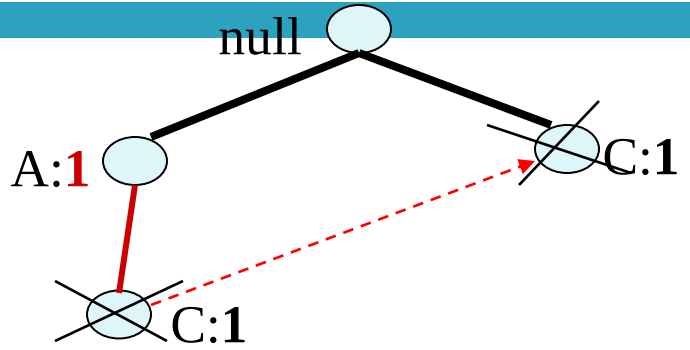
2. Περικοπές κόμβων



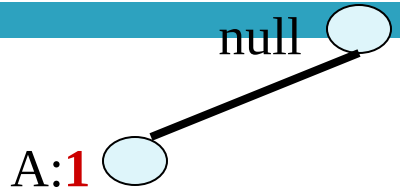
1. Αλλαγή υποστήριξης



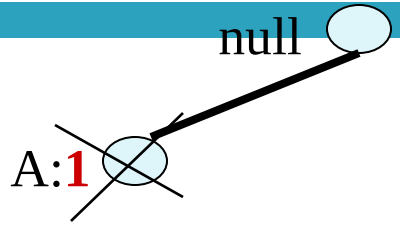
2. Περικοπή Κόμβων



2. Περικοπή Κόμβων



2. Περικοπή Κόμβων



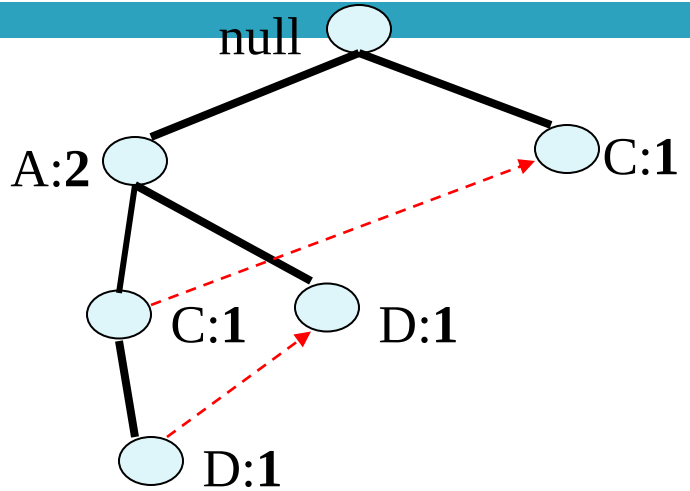
2. Περικοπή Κόμβων

Άρα, επιστροφή στο επόμενο υποπρόβλημα

Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

Ο αλγόριθμος επαναλαμβάνεται για το ~~{D, E}~~, ~~{C,~~ E}, **{A, E}**

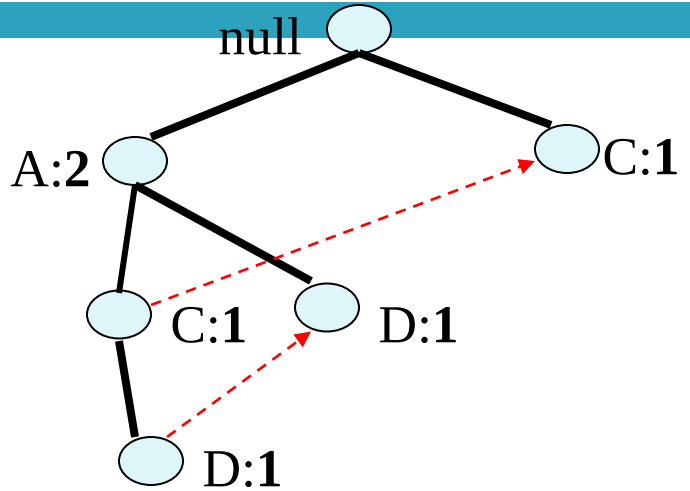


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)

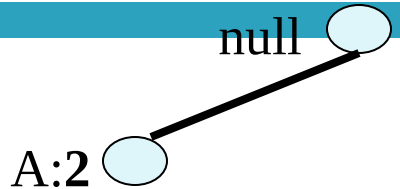


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)

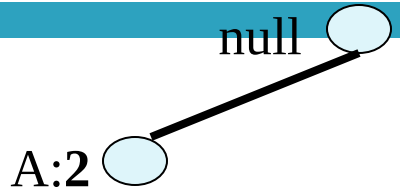


Αλγόριθμος FP-Growth

Βρες την υποστήριξη του {A, E}

Οπότε {A, E} συχνό

Δε χρειάζεται να φτιάξουμε υπο-συνθήκη FP-δέντρο για το {A, E}



Αλγόριθμος FP-Growth

Άρα για το E

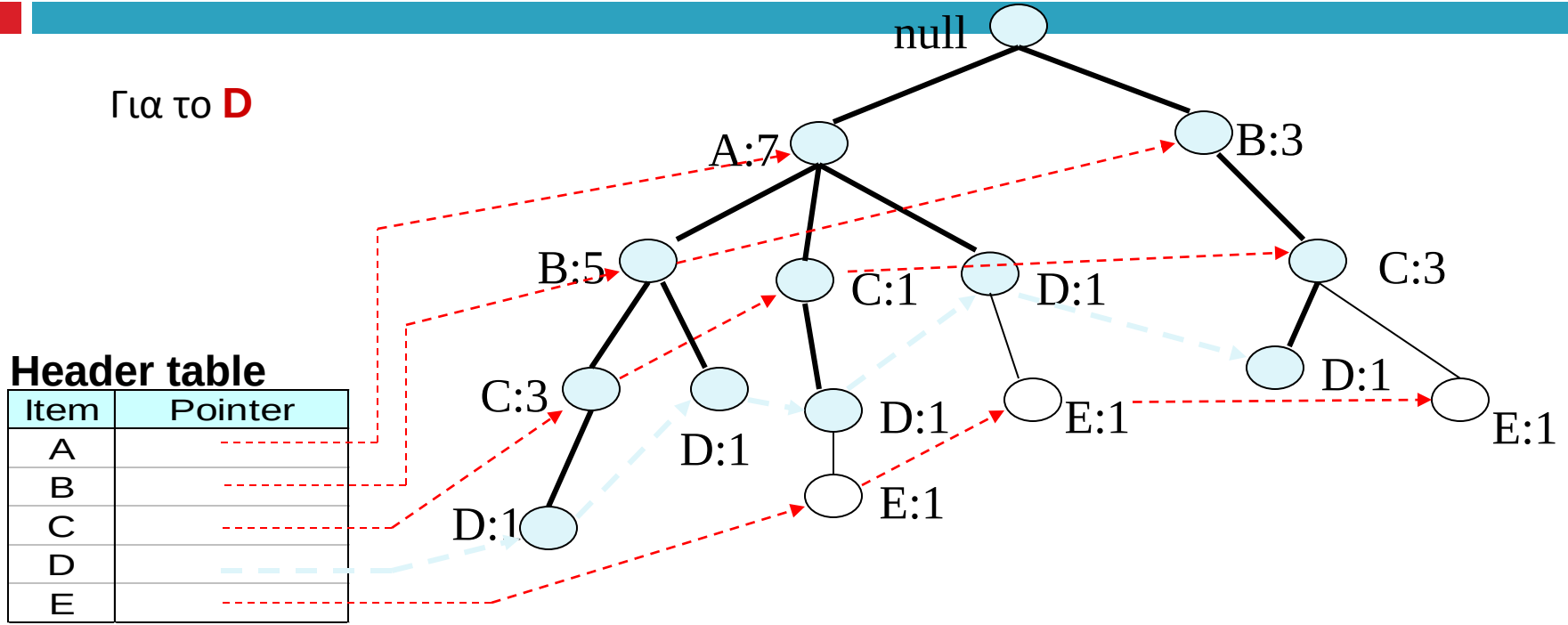
Έχουμε τα εξής συχνά στοιχειοσύνολα

{E} {D, E} {A, D, E} {C, E} {A, E}

Συνεχίζουμε για το D

Αλγόριθμος FP-Growth

Για το **D**

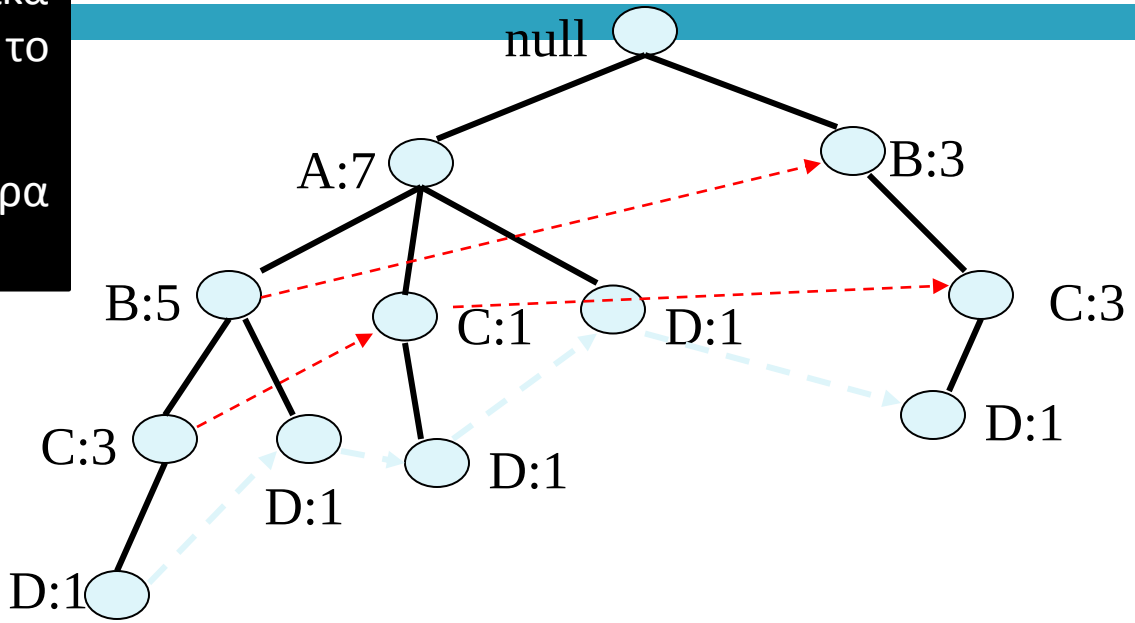


Αλγόριθμος FP-Growth

Φάση 1

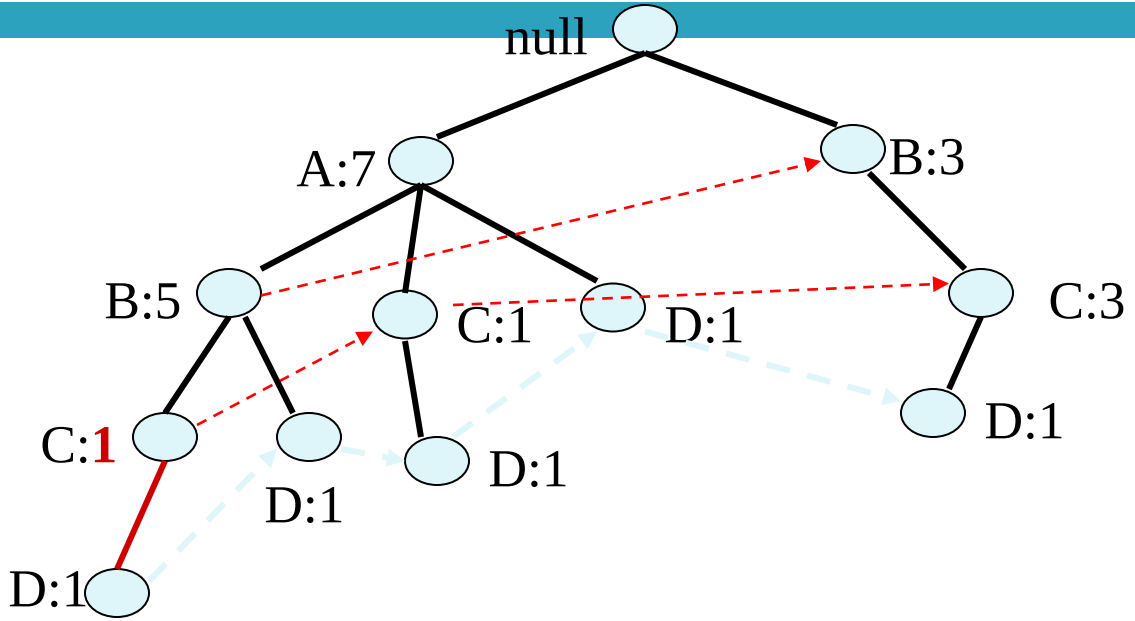
Όλα τα προθεματικά μονοπάτια που περιέχουν το D

Υποστήριξη 5>2 -> άρα συχνό

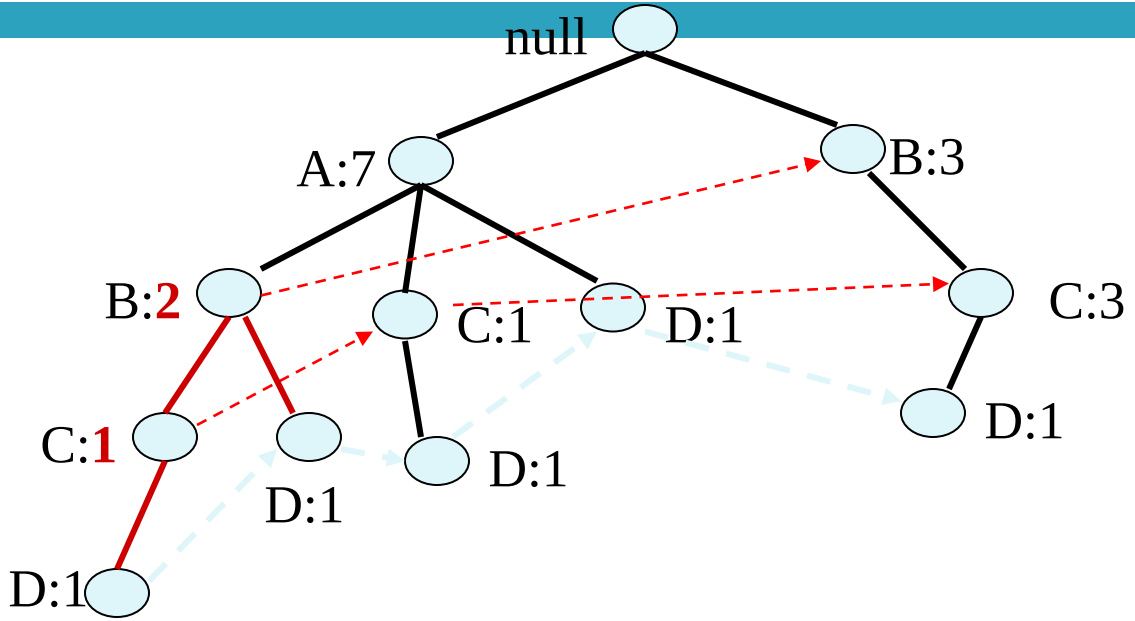


Μετατροπή του προθεματικού δέντρου σε FP-δέντρο υπό συνθήκη

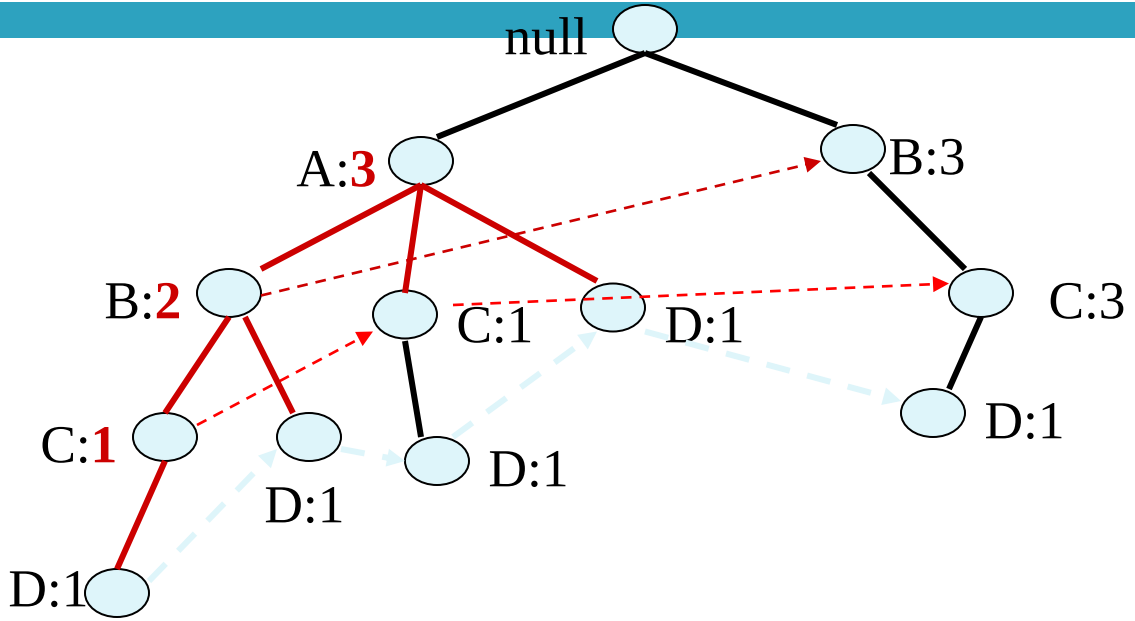
1. Αλλαγή υποστήριξης



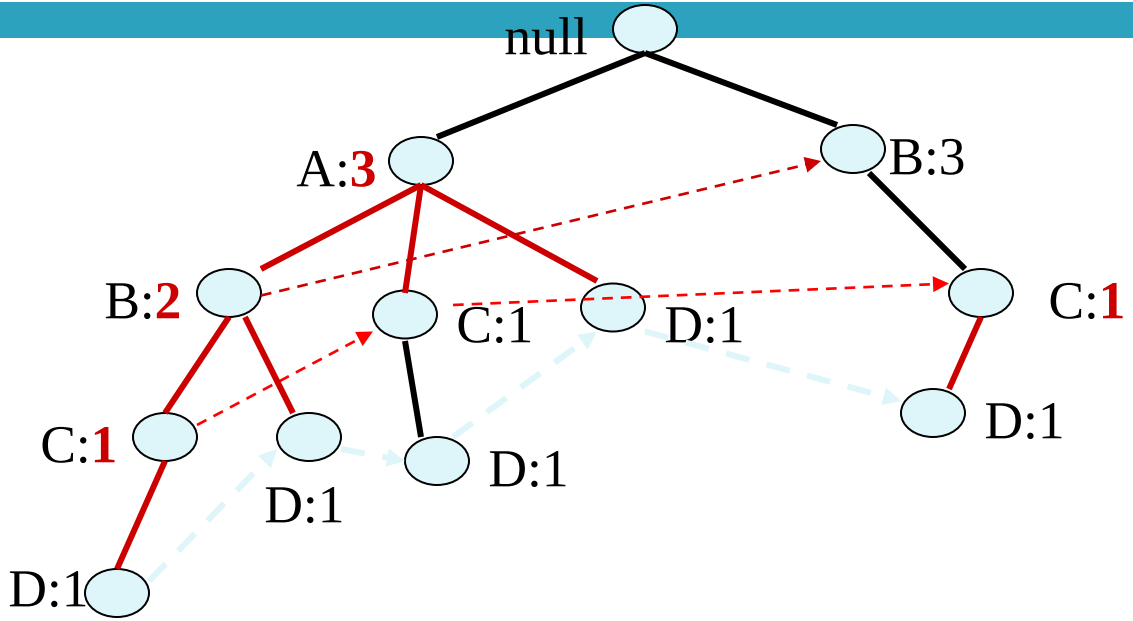
1. Αλλαγή υποστήριξης



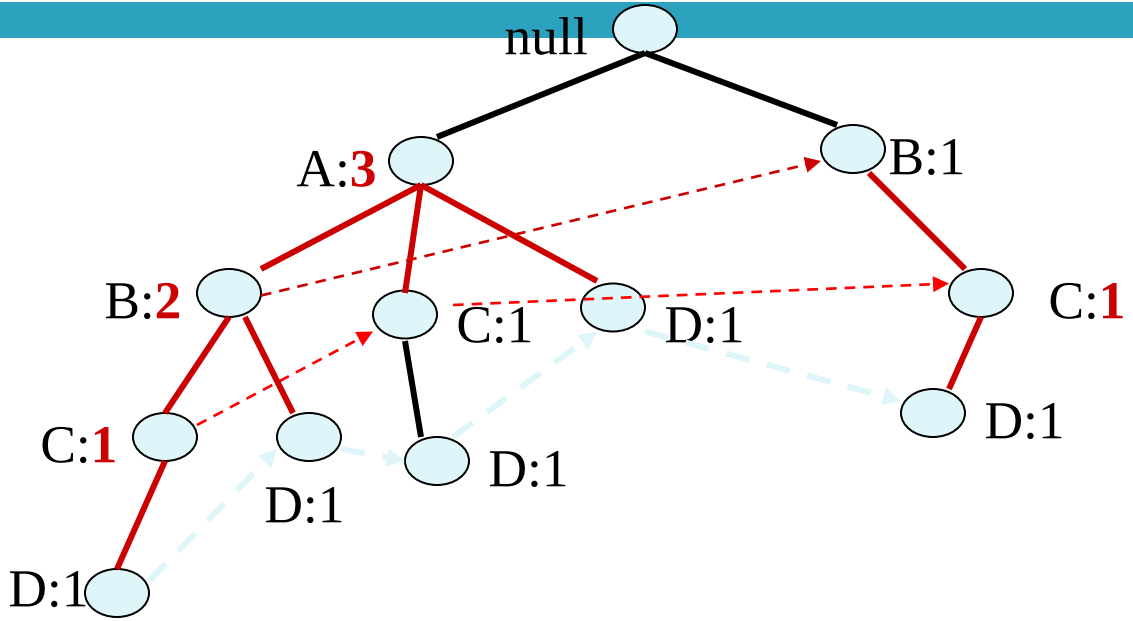
1. Αλλαγή υποστήριξης



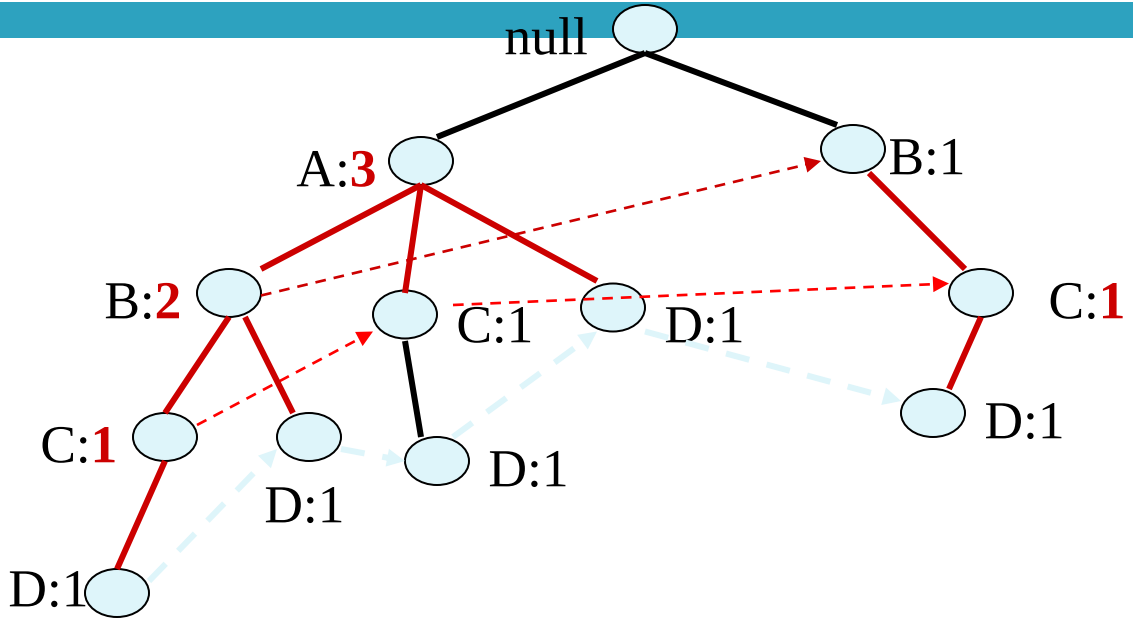
1. Αλλαγή υποστήριξης



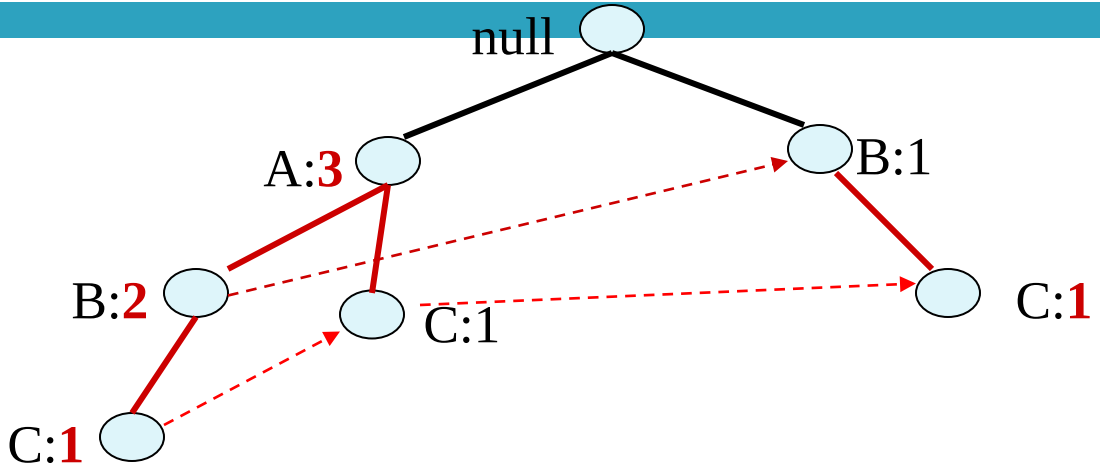
1. Αλλαγή υποστήριξης



2. Περικοπή Κόμβων



2. Περικοπή Κόμβων

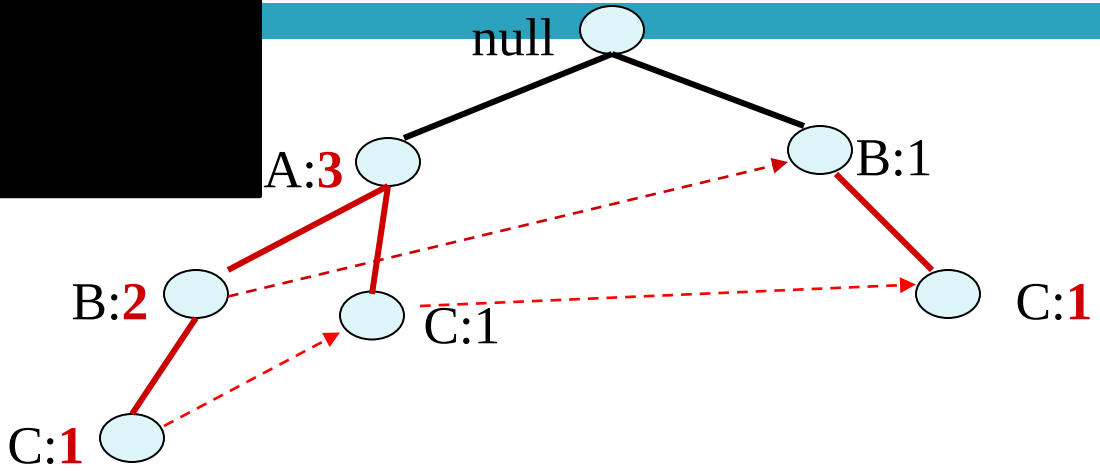


Αλγόριθμος FP-Growth

Προθεματικά δέντρα και υποσυνθήκη δέντρα

Για τα AD, BD και CD

κοκ



ECLAT

- Για κάθε στοιχείο, αποθήκευσε μια λίστα δοσοληψιών (tids)

Horizontal
Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

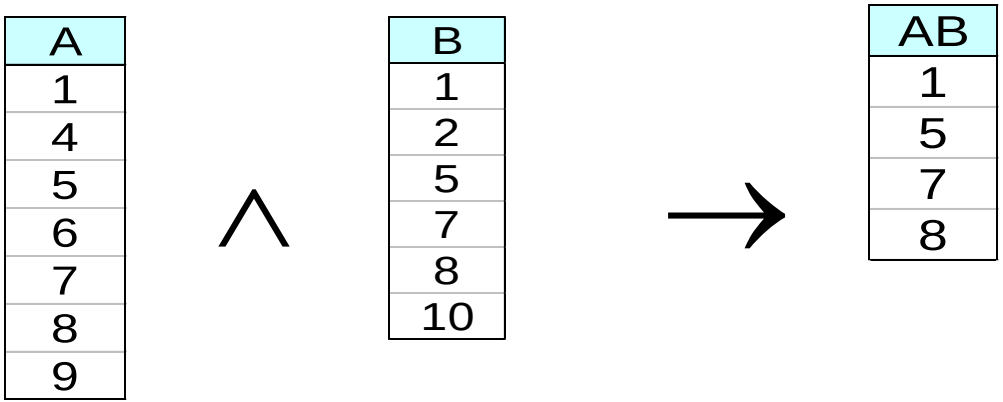
Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

Λίστα TID

ECLAT

- Καθορισμός της υποστήριξης κάθε Κ-οστού στοιχειοσυνόλου διασταυρώνοντας τις tid-λίστες των δυο (κ-1) υποσυνόλων. Π.χ.:



- 3 επιλογές:
  - top-down, bottom-up και υβριδική
- Πλεονέκτημα: πολύ γρήγορη μέτρηση υποστήριξης
- Μειονέκτημα: οι ενδιάμεσες tid-λίστες μπορεί να γίνουν μεγάλες για τη μνήμη

Παραγωγή Κανόνων

- 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 →D, ABD →C, ACD →B, BCD →A,
    A →BCD, B →ACD, C →ABD, D →ABC
    AB →CD, AC → BD, AD → BC, BC →AD,
    BD →AC, CD →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 υπολογίσει την υποστήριξή του

Παραγωγή Κανόνων

Πως να παράγουμε αποδοτικά τους κανόνες από τα συχνά στοιχειοσύνολα;

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

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

Γενικά έστω {p} → {q} με εμπιστοσύνη c1

- Και {p, r} → {q} με εμπιστοσύνη c2

  Μπορεί c2 > c1, c2 < c1 ή c2 = c1

- Έστω {p} → {q, r} με εμπιστοσύνη c3

  $c3 \leq c1$

- Επίσης, $c3 \leq c2$

Παραγωγή Κανόνων

- Η εμπιστοσύνη για τους κανόνες που παράγονται από τα ίδια στοιχειοσύνολα έχει μια αντι-μονότονη ιδιότητα

  Για παράδειγμα L = {A,B,C,D}:

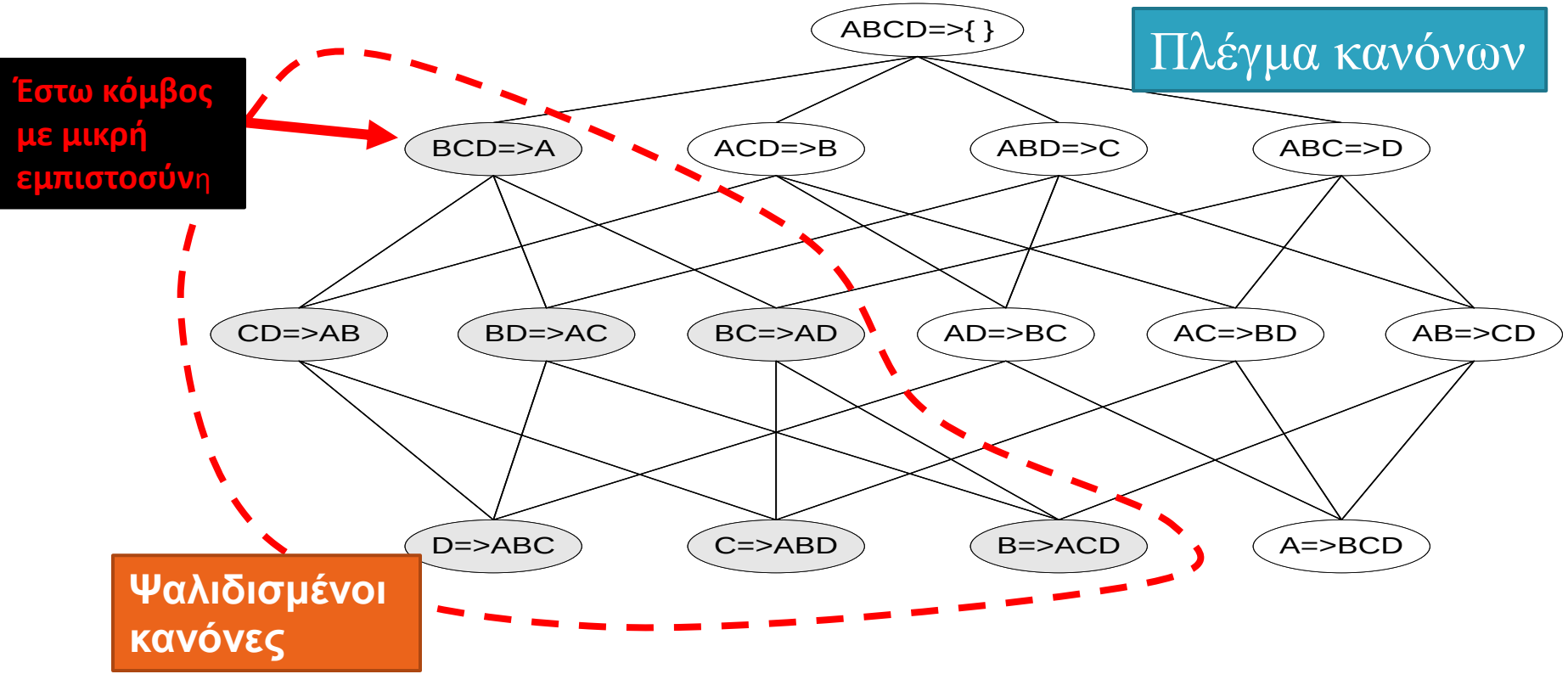
$$c(ABC \rightarrow \mathbf{D}) \geq c(AB \rightarrow C\mathbf{D}) \geq c(A \rightarrow BC\mathbf{D})$$

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

  Τυπικά:

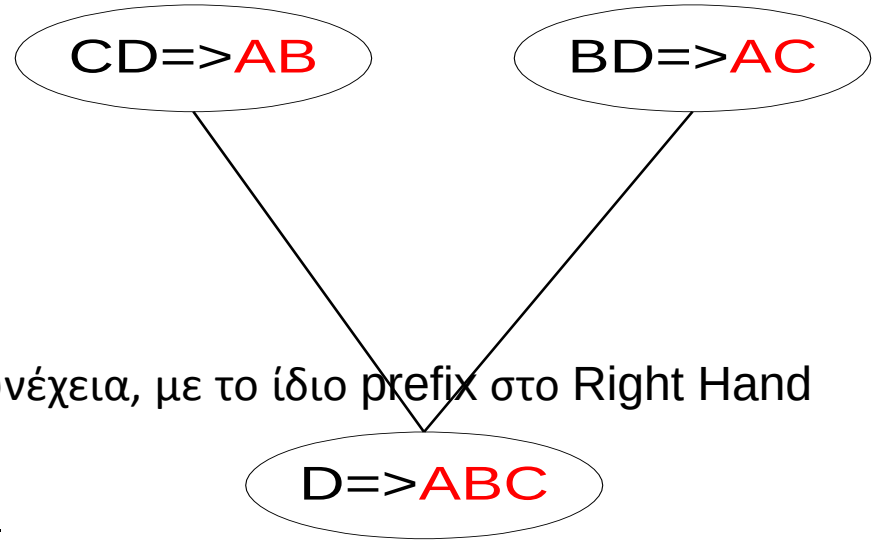
  Αν ο κανόνας $X \rightarrow X - Y$ δεν ικανοποιεί το κατώφλι εμπιστοσύνης, τότε και ο κανόνας $X' \rightarrow X' - Y'$ ($X' \subseteq X$) δεν τον ικανοποιεί

Παραγωγή Κανόνων για τον αλγόριθμο apriori



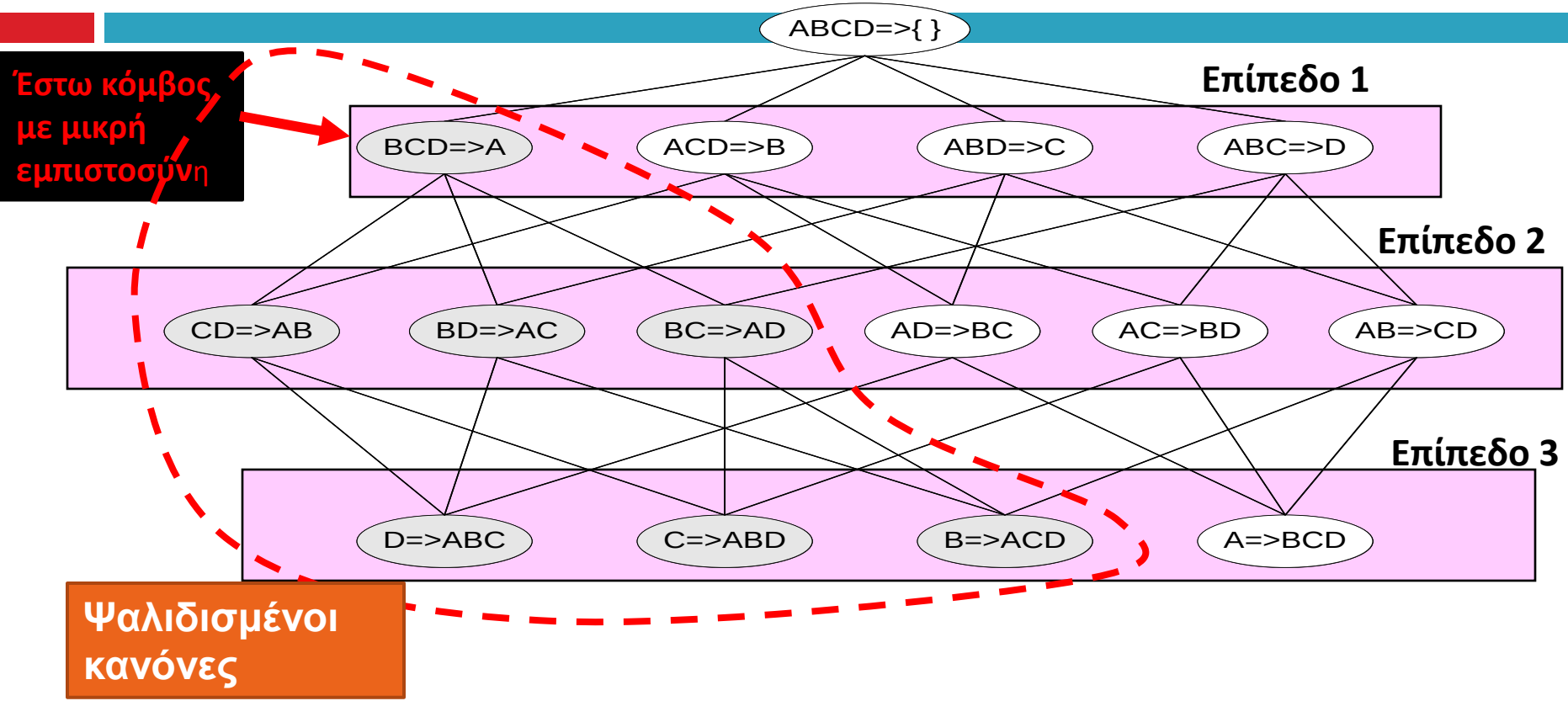
Παραγωγή Κανόνων για τον αλγόριθμο apriori

- Ο υποψήφιος κανόνας δημιουργείται με τη συγχώνευση δυο κανόνων που μοιράζονται το ίδιο πρόθεμα στον κανόνα που προκύπτει
- Η ένωση (CD=>AB,BD=>AC) θα παρήγαγε τον κανόνα
  - D => ABC
- Η ένωση (ACD=>B, ABD=>C) μας δίνει
  - AD=>BC
- Όπως και στα συχνά στοιχειοσύνολα, στη συνέχεια, με το ίδιο prefix στο Right Hand Side
  - join(CD=>AB,BD=>AC) μας δίνει D => ABC
- Ο κανόνας D=>ABC ψαλιδίζεται εάν το υποσύνολο του
  - AD=>BC δεν έχει μεγάλη εμπιστοσύνη



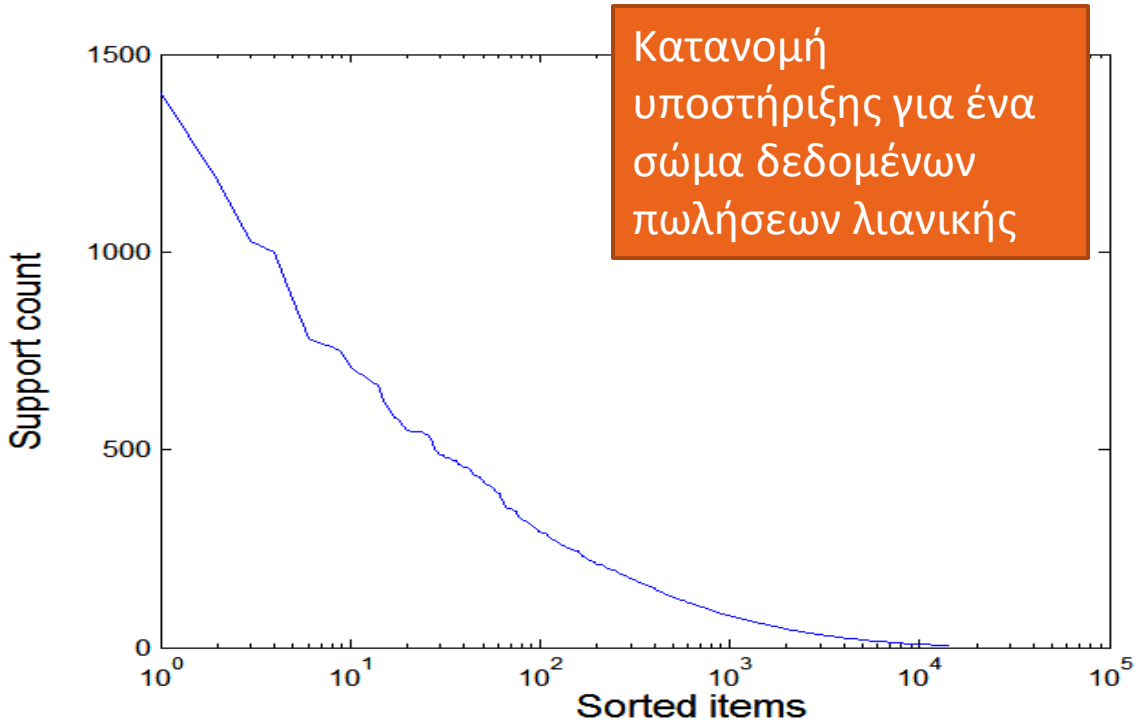
Παραγωγή Κανόνων για τον αλγόριθμο apriori

Πλέγμα κανόνων



Επίδραση της κατανομής υποστήριξης

- Πολλά πραγματικά σώματα δεδομένων έχουν διαστρεβλωμένη κατανομή υποστήριξης



Επίδραση της κατανομής υποστήριξης

- Πως ορίζουμε το κατάλληλο κατώφλι *minsup*?
  - Εάν τεθεί πολύ ψηλά
    - Χάνουμε στοιχειοσύνολα με ενδιαφέροντα σπάνια στοιχεία (π.χ. ακριβά προϊόντα)
  - Εάν τεθεί πολύ χαμηλά
    - Υπολογιστικά ακριβό γιατί θα προκύψει μεγάλος αριθμός στοιχειοσυνόλων
- Η χρήση ενός μόνο κατωφλίου για την υποστήριξη ίσως να μην είναι η ενδεδειγμένη λύση

Πολλαπλές ελάχιστες υποστηρίξεις

- Πως εφαρμόζονται οι πολλαπλές υποστηρίξεις;
  - MS(i): minimum support (ελάχιστη υποστήριξης) του i
    - MS(Milk)=5%, MS(Coke) = 3%, MS(Broccoli)=0.1%, MS(Salmon)=0.5%
  - MS({Milk, Broccoli}) = min (MS(Milk), MS(Broccoli))= 0.1%
  - Πρόκληση: Η υποστήριξη δεν είναι πλέον αντί-μονότονη
    - Έστω: Support(Milk, Coke) = 1.5% και Support(Milk, Coke, Br

Λόγω του Broccoli που κατεβάζει το minsup

    - το {Milk,Coke} είναι μη συχνό ενώ το {Milk,Coke,Broccoli} είναι

Πολλαπλές ελάχιστες υποστηρίξεις (Liu 1999)

- Ταξινόμησε τα στοιχεία με βάση την ελάχιστη υποστήριξη σε αύξουσα σειρά
  - Π.χ: MS(Milk)=5%, MS(Coke) = 3%, MS(Broccoli)=0.1%, MS(Salmon)=0.5%
  - Ταξινόμηση: Broccoli, Salmon, Coke, Milk

- Αλλάζει και ο **apriori** στα εξής:
  - L_1 : συχνά στοιχεία
  - F_1 : στοιχεία με υποστήριξη≥ MS(1) όπου MS(1) το min_i(MS(i))
  - C_2 : τα υποψήφια στοιχειοσύνολα μεγέθους 2 δημιουργούνται από το F_1 αντί του L_1

Πολλαπλές Τιμές Υποστήριξης

- Τροποποιήσεις στον Apriori (Βήμα Ψαλιδίσματος):
  - Στον παραδοσιακό Apriori,
    - Ένα υποψήφιο (k+1)-στοιχειοσύνολο δημιουργείται συγχωνεύοντας δυο συχνά k-στοιχειοσύνολα
    - Το υποψήφιο ψαλιδίζεται αν περιέχει ένα (οποιοδήποτε) μη συχνό k-στοιχειοσύνολο
  - Τροποποίηση βήματος ψαλιδίσματος:
    - Ψαλίδισε μόνο αν το υποσύνολο περιέχει το πρώτο στοιχείο

    πχ Candidate={Broccoli, Coke, Milk} (διατεταγμένα με βάση την μικρότερη ελάχιστη υποστήριξη)

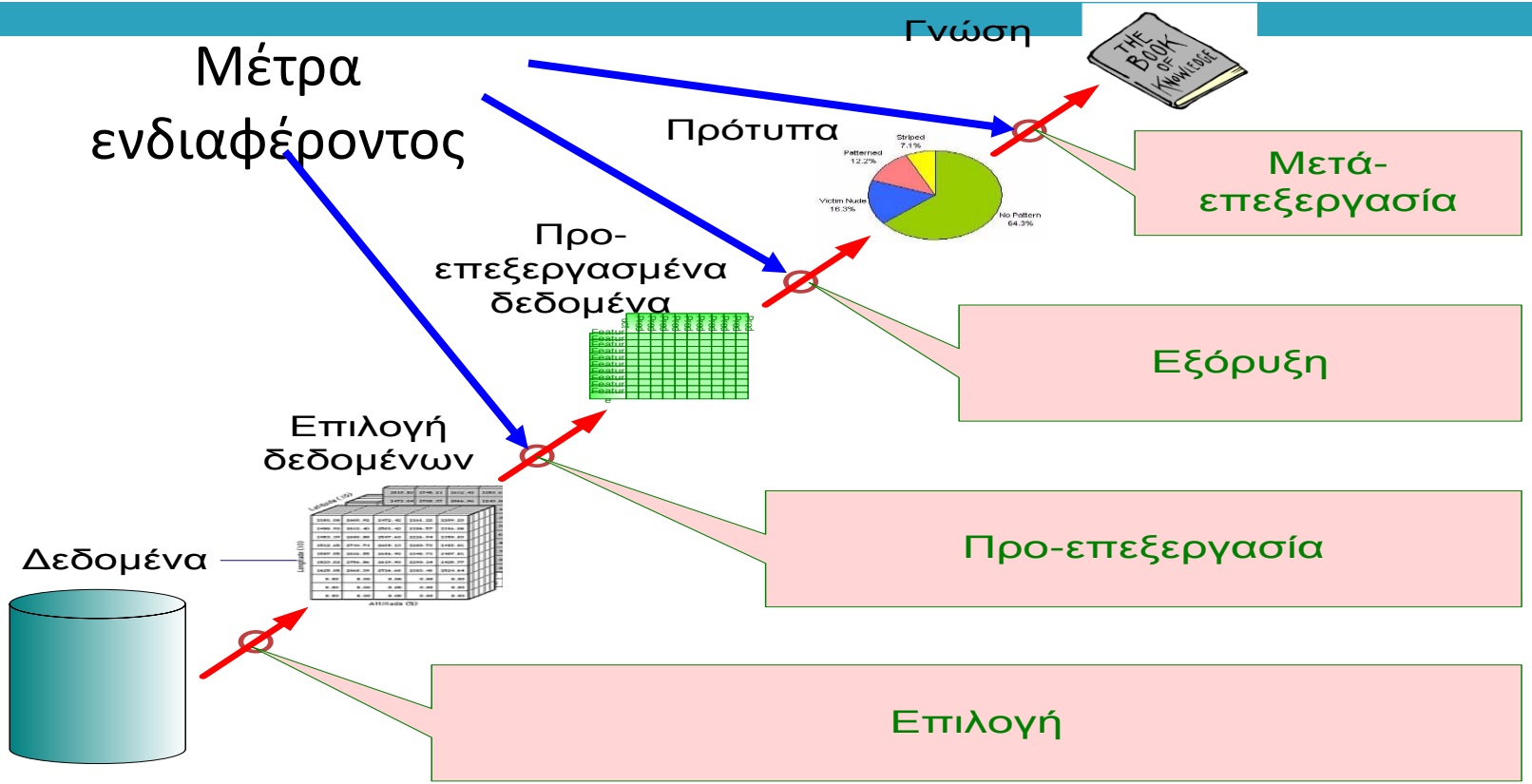
    {Broccoli, Coke} και {Broccoli, Milk} είναι συχνά αλλά {Coke, Milk} είναι μη συχνό

      - Το Candidate δε σβήνεται γιατί το {Coke,Milk} δεν περιέχει το πρώτο

Αξιολόγηση προτύπων

- Οι αλγόριθμοι κανόνων συσχέτισης τείνουν να παράγουν πολλούς κανόνες
  - Πολλοί από αυτούς δεν έχουν ενδιαφέρον οι είναι πλεονάζοντες
  - Πλεονάζων αν
    - $\{A,B,C\} \rightarrow \{D\}$ και $\{A,B\} \rightarrow \{D\}$ έχουν ίδια εμπιστοσύνη και υποστήριξη
- Μπορούν να χρησιμοποιηθούν μέτρα ενδιαφέροντος (Interestingness measures) για να ελαττώσουν (prune) ή να ιεραρχήσουν (rank) τα παραγόμενα πρότυπα
- Στην αρχική διατύπωση του προβλήματος της εξόρυξης κανόνων συσχέτισης χρησιμοποιήθηκαν ως μέτρα μόνο η *υποστήριξη* και η *εμπιστοσύνη*

Εφαρμογές μέτρησης ενδιαφέροντος



Υπολογισμός του μέτρου ενδιαφέροντος

- Δοσμένου ενός κανόνα $X \rightarrow Y$, η πληροφορία που χρειάζεται για να υπολογισθεί το μέτρο ενδιαφέροντος πηγάζει από τον ακόλουθο πίνακα

Πίνακας Συνάφειας $\overline{X} \rightarrow Y$

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	$\|T\|$

Μέτρηση συχνότητας εμφάνισης

f_{11}: υποστήριξη του X και Y
f_{10}: υποστήριξη του X και $\overline{Y}$
f_{01}: υποστήριξη του $\overline{X}$ και Y
f_{00}: υποστήριξη του $\overline{X}$ και $\overline{Y}$

Χρησιμοποιείται για διάφορα μέτρα

- υποστήριξη, εμπιστοσύνη, lift, Gini,J-measure, κτλ.

Μειονέκτημα της εμπιστοσύνης

Ενδιαφερόμαστε για τη σχέση μεταξύ αυτών που πίνουν καφέ και αυτών που πίνουν τσάι

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Κανόνας : Tea $\rightarrow$ Coffee

Εμπιστοσύνη=P(Coffee|Tea)=15/20=0.75

αλλά P(Coffee) = 90/100=0.9

$\Rightarrow$Ενώ έχει μεγάλη εμπιστοσύνη, ο κανόνας είναι παραπλανητικός

$\Rightarrow$P(Coffee|$\overline{\text{Tea}}$) = 75/80=0.9375

Αγνοεί την υποστήριξη του Right Hand

Στατιστική ανεξαρτησία

- Έστω πληθυσμός 1000 φοιτητών
  - 600 ξέρουν κολύμπι (S)
  - 700 ξέρουν ποδήλατο (B)
  - 420 ξέρουν κολύμπι και ποδήλατο (S,B)

  - $P(S \wedge B) = 420/1000 = 0.42$
  - $P(S) \times P(B) = 0.6 \times 0.7 = 0.42$

- $P(S \wedge B) = P(S) \times P(B)$ => στατιστική ανεξαρτησία independence
- $P(S \wedge B) > P(S) \times P(B)$ => Positively correlated (θετική συσχέτιση)
- $P(S \wedge B) < P(S) \times P(B)$ => Negatively correlated (αρνητική συσχέτιση)

Μέτρα στατιστικής εξάρτησης $X \rightarrow Y$

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

$$PS = P(X,Y) - P(X)P(Y)$$

$$\varphi\text{-coefficient} = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

Παράδειγμα: Lift/Interest

	Coffee	$\overline{\text{Coffee}}$	
Tea	15	5	20
$\overline{\text{Tea}}$	75	5	80
	90	10	100

Κανόνας: Tea $\rightarrow$ Coffee

Εμπιστοσύνη= P(Coffee|Tea) = =15/20=0.75

ενώ P(Coffee) =90/100= 0.9

$\Rightarrow$Lift = 0.75/0.9= 0.8333 (< 1, επομένως είναι αρνητικά συσχετιζόμενα)

Μειονέκτημα των Lift & Interest

	Y	$\overline{Y}$	
X	10	0	10
$\overline{X}$	0	90	90
	10	90	100

$$Interest = \frac{0.1}{(0.1)(0.1)} = 10$$

Μεγαλύτερο αν και σπάνια εμφανίζονται μαζί

	Y	$\overline{Y}$	
X	90	0	90
$\overline{X}$	0	10	10
	90	10	100

$$Interest = \frac{0.9}{(0.9)(0.9)} = 1.11$$

Στατιστική ανεξαρτησία:

αν P(X,Y)=P(X)P(Y) => Interest = 1

Παράδειγμα: ϕ-coefficient

ϕ-coefficient: είναι ανάλογος του συντελεστή συσχέτισης για συνεχείς μεταβλητές

	Y	$\overline{Y}$	
$\underline{X}$	60	10	70
X	10	20	30
	70	30	100

$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}}$$

$$= 0.5238$$

	Y	$\overline{Y}$	
$\underline{X}$	20	10	30
X	10	60	70
	30	70	100

$$\phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}}$$

$$= 0.5238$$

Ο ϕ-coefficient είναι ίδιος

Έχουν προταθεί πολλές μέτρα ανάλογα με την εφαρμογή

Με ποια κριτήρια θα επιλέξουμε ένα καλό μέτρο;

Πως έναν Apriori-style support based pruning επηρεάζει αυτά τα

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A), P(A\|B))$
12	Laplace (L)	$\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$
13	Conviction (V)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\left(\frac{P(B\|A)-P(B)}{1-P(B)}, \frac{P(A\|B)-P(A)}{1-P(A)}\right)$
18	Added Value (AV)	$\max(P(B\|A)-P(B), P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})} \times \frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

Ιδιότητες ενός καλού μέτρου

- Piatetsky-Shapiro:

  Ένα καλό μέτρο M πρέπει να έχει 3 ιδιότητες:

  - M(A,B) = 0 αν A και B στατιστικά ανεξάρτητα
  - M(A,B) αυξάνει μονότονα με το P(A,B) όταν τα P(A) και P(B) παραμένουν αμετάβλητα
  - M(A,B) μειώνεται μονότονα με το P(A) [ή το P(B)] όταν τα P(A,B) και P(B) [ή P(A)] παραμένουν αμετάβλητα

Σύγκριση διαφορετικών μέτρων

10 παραδείγματα πινάκων συνάφειας

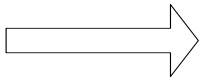
Example	f_{11}	f_{10}	f_{01}	f_{00}
E1	8123	83	424	1370
E2	8330	2	622	1046
E3	9481	94	127	298
E4	3954	3080	5	2961
E5	2886	1363	1320	4431
E6	1500	2000	500	6000
E7	4000	2000	1000	3000
E8	4000	2000	2000	2000
E9	1720	7121	5	1154
E10	61	2483	4	7452

Ιεράρχηση των πινάκων με βάση τα διάφορα μέτρα (1 το πιο ενδιαφέρον, 10 το λιγότερο ενδιαφέρον):

#	ϕ	λ	α	Q	Y	κ	M	J	G	s	c	L	V	I	IS	PS	F	AV	S	ζ	K
E1	1	1	3	3	3	1	2	2	1	3	5	5	4	6	2	2	4	6	1	2	5
E2	2	2	1	1	1	2	1	3	2	2	1	1	1	8	3	5	1	8	2	3	6
E3	3	3	4	4	4	3	3	8	7	1	4	4	6	10	1	8	6	10	3	1	10
E4	4	7	2	2	2	5	4	1	3	6	2	2	2	4	4	1	2	3	4	5	1
E5	5	4	8	8	8	4	7	5	4	7	9	9	9	3	6	3	9	4	5	6	3
E6	6	6	7	7	7	7	6	4	6	9	8	8	7	2	8	6	7	2	7	8	2
E7	7	5	9	9	9	6	8	6	5	4	7	7	8	5	5	4	8	5	6	4	4
E8	8	9	10	10	10	8	10	10	8	4	10	10	10	9	7	7	10	9	8	7	9
E9	9	9	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	8	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

Αλλαγή διάταξης μεταβλητών

	B	$\overline{B}$
A	p	q
$\overline{A}$	r	s



	A	$\overline{A}$
B	p	r
$\overline{B}$	q	s

Ισχύει ότι M(A,B) = M(B,A)?

Συμμετρικά μέτρα:

- support (υποστήριξη) lift, collective strength, cosine, Jaccard, κλπ.

Μη-συμμετρικά μέτρα:

- confidence (εμπιστοσύνη), conviction, Laplace, J-measure,

Κλιμάκωση γραμμής/στήλης
Row/Column Scaling

Παράδειγμα Βαθμός-Φύλλο (Mosteller, 1968):

	Male	Female	
High	2	3	5
Low	1	4	5
	3	7	10

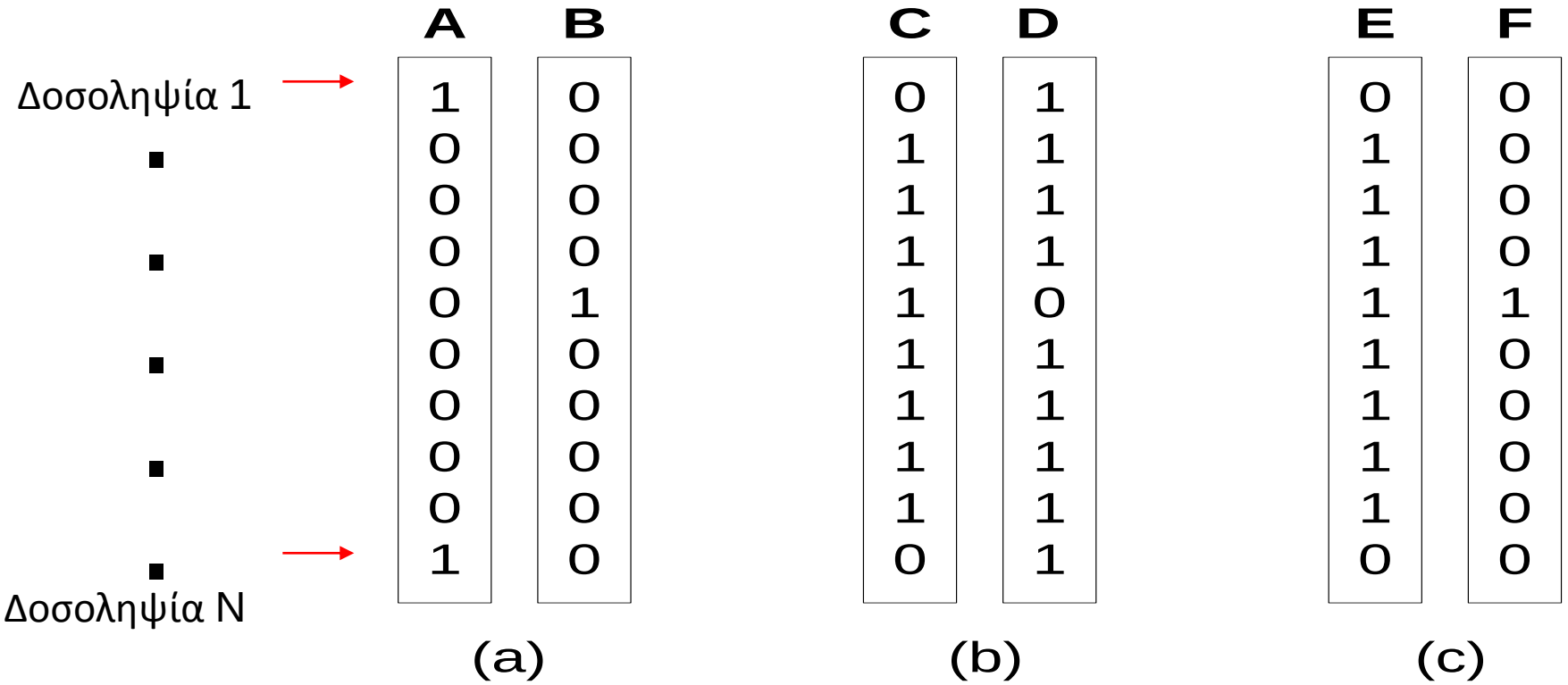
	Male	Female	
High	4	30	34
Low	2	40	42
	6	70	76

2x 10x

Mosteller:

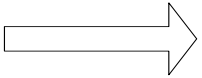
Η συσχέτιση πρέπει να είναι ανεξάρτητη από το σχετικό αριθμό αγοριών-κοριτσιών στο δείγμα

Αντιστροφή - Inversion Operation



Προσθήκη μη-συσχετιζόμενων στοιχείων

	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s



	B	$\bar{B}$
A	p	q
$\bar{A}$	r	s + k

Δεν επηρεάζονται από την αύξηση του f_{00} όταν οι άλλες τιμές παραμένουν αμετάβλητες

Αναλλοίωτα μέτρα

- support, cosine, Jaccard, κτλ

Μη αναλλοίωτα μέτρα:

- correlation, Gini, mutual information, odds ratio, κτλ

Διαφορετικά μέτρα έχουν διαφορετικές ιδιότητες

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 … 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 … 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 … 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 … 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 … 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 … 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 … 1 … ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 … 1 … ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}-1}\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0 \ldots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Το παράδοξο του Simpson

- Εμφανίζεται όταν μια Τρίτη (πιθανώς κρυμμένη) μεταβλητή προκαλεί τις παρατηρηθείσες σχέσεις μεταξύ δυο μεταβλητών να αλλάξουν κατεύθυνση
- Παράδειγμα: εγώ και ο φίλος μου σουτάρουμε 20 βολές. Ποιος είναι ο καλύτερος;

	me
make	10
miss	10
total	20

	my friend
make	8
miss	12
total	20

Το παράδοξο του Simpson

- Ποιος είναι όμως ο καλύτερος όταν λάβουμε υπόψη και την απόσταση;

	me		
	far	**close**	**total**
make	1	9	10
miss	3	7	10
total	4	16	20

	my friend		
	far	**close**	**total**
make	5	3	8
miss	10	2	12
total	15	5	20

Το παράδοξο του Simpson

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99	81	180
No	54	66	120
	153	147	300

Students

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	1	9	10
No	4	30	34

Working adults

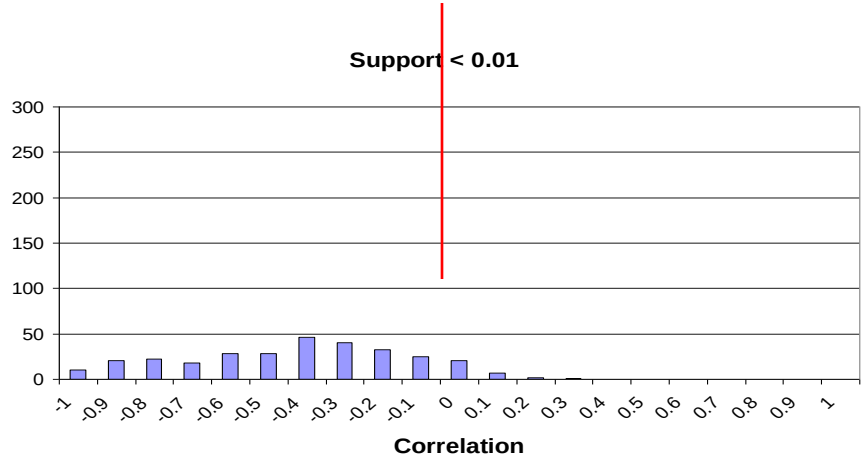
Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	98	72	170
No	50	36	86

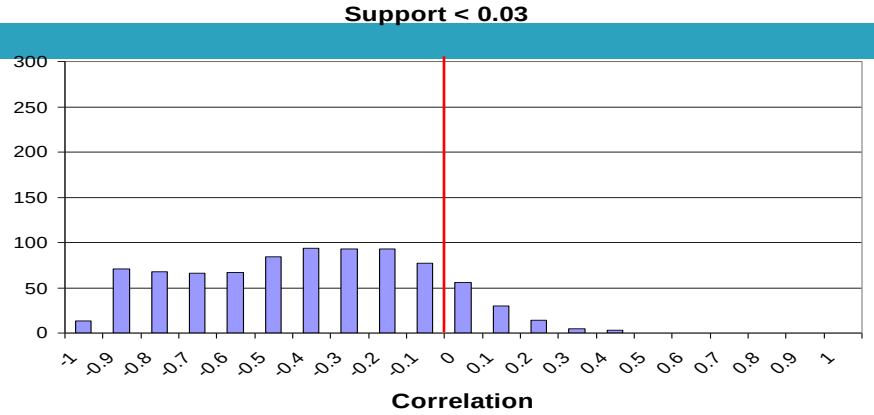
Ελάττωση με βάση την υποστήριξη support-based pruning

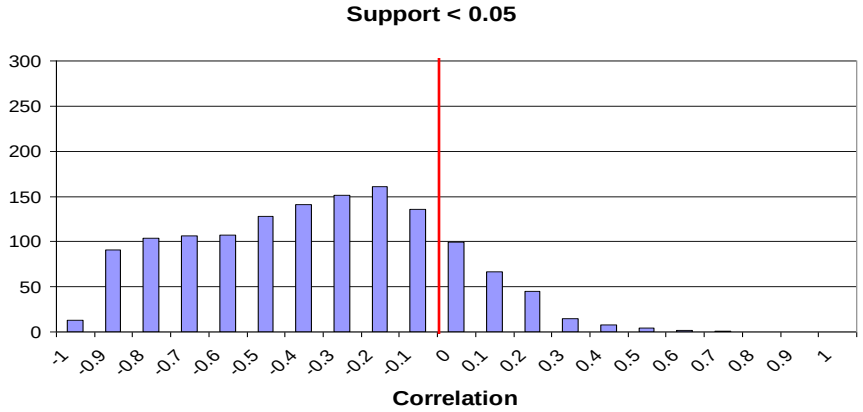
- Οι περισσότεροι αλγόριθμοι για την εξόρυξη κανόνων συσχέτισης χρησιμοποιούν την υποστήριξη για να μειώσουν (prune) κανόνες και στοιχειοσύνολα
- Μελέτη του αποτελέσματος της μείωσης στη συσχέτιση των στοιχειοσυνόλων
  - Δημιουργία 10.000 τυχαίων contingency tables
  - Υπολογισμός της υποστήριξης και της ανά δύο συσχέτισης των πινάκων
  - Εφαρμογή της ελάττωσης με βάση την υποστήριξη και μελέτη των πινάκων που αφαιρέθηκαν

Επίδραση του Support-based Pruning

Επίδραση του Support-based Pruning





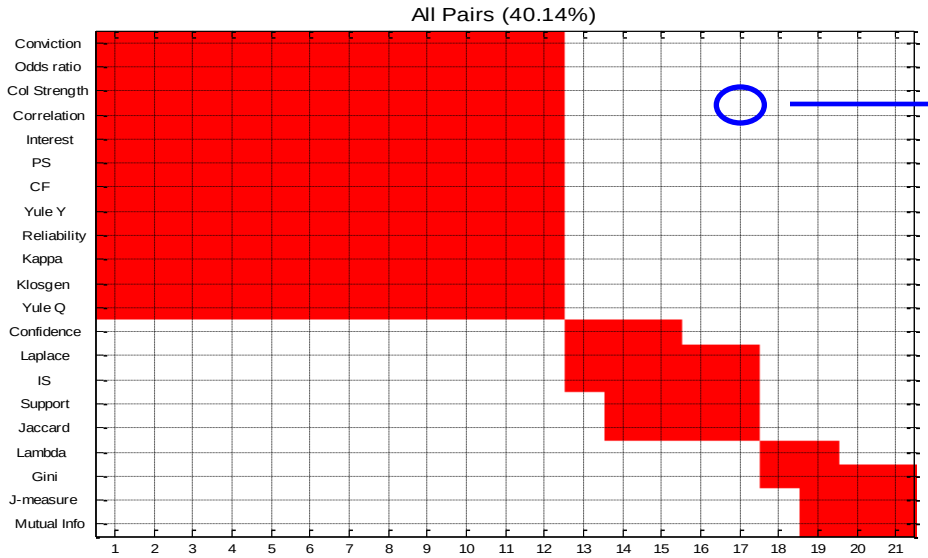


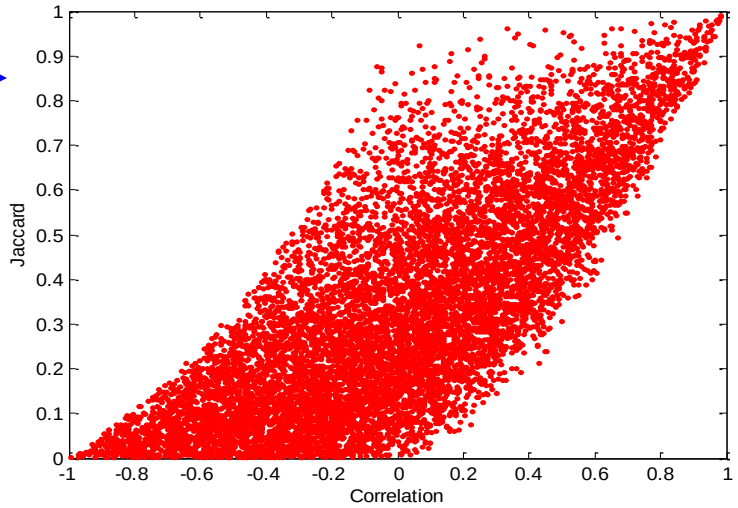
Κυρίως αφαιρεί στοιχειοσύνολα που σχετίζονται αρνητικά

Επίδραση του Support-based Pruning

- Μελέτη του πως επηρεάζει τα άλλα μέτρα
- Βήματα
  - Δημιουργία 10.000 contingency tables
  - Ιεράρχηση κάθε πίνακα με βάση τα διαφορετικά μέτρα
  - Υπολογισμός της ανά-δύο συσχέτισης μεταξύ των μέτρων

Χωρίς ελάττωση (όλα τα ζεύγη)

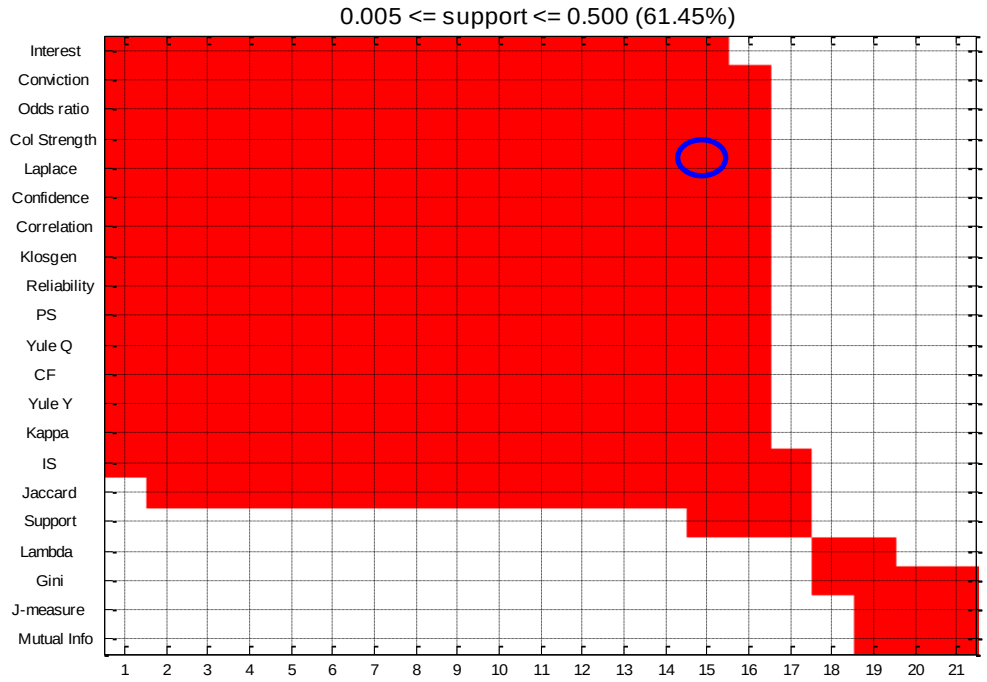


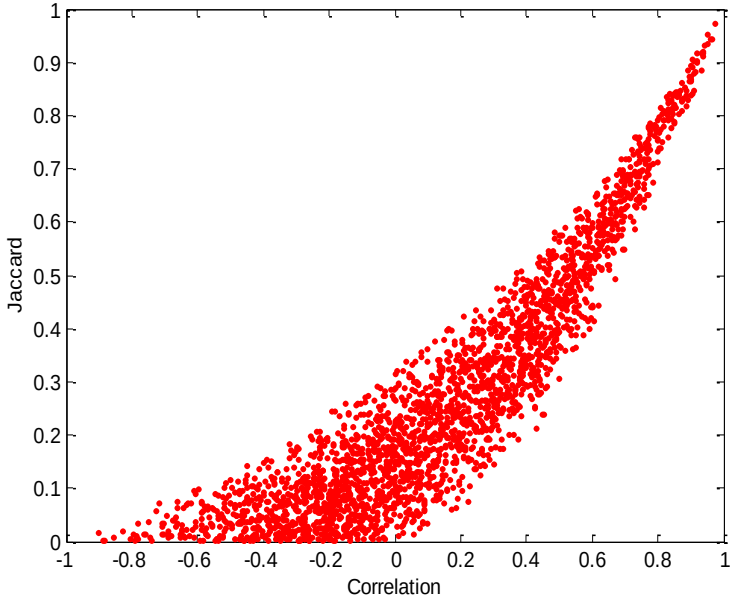


- Τα κόκκινα κελιά δείχνουν correlation μεταξύ των ζευγών των μέτρων > 0.85
- 40.14% των ζευγών έχουν correlation >

Διάγραμμα διασποράς μεταξύ Correlation & Jaccard Measure

0.5% ≤ support ≤ 50%

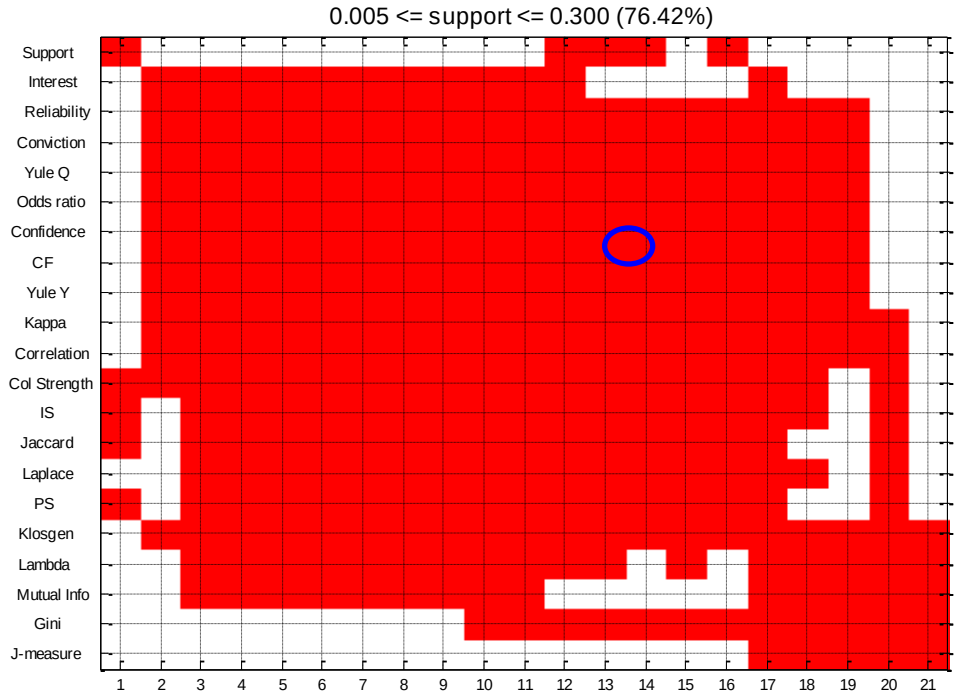


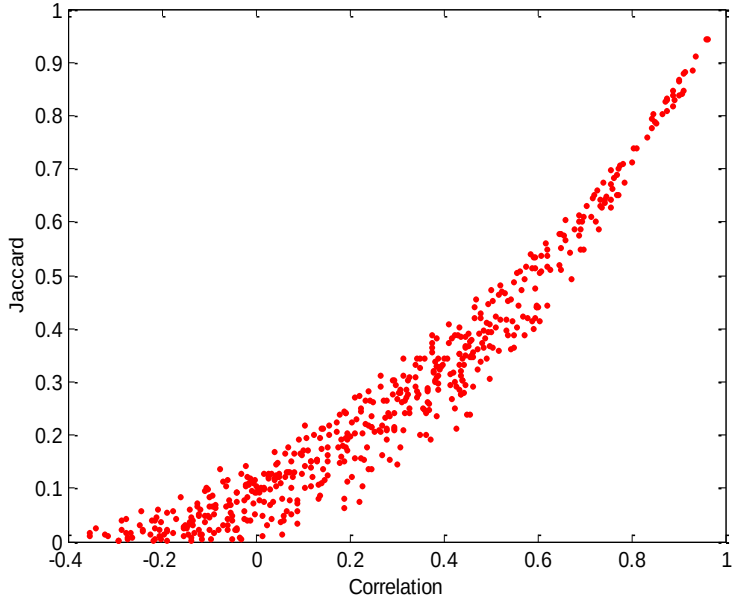


Διάγραμμα διασποράς μεταξύ Correlation & Jaccard Measure

- 61.45% των ζευγών έχουν correlation >

0.5% ≤ support ≤ 30%





Διάγραμμα διασποράς μεταξύ Correlation & Jaccard Measure

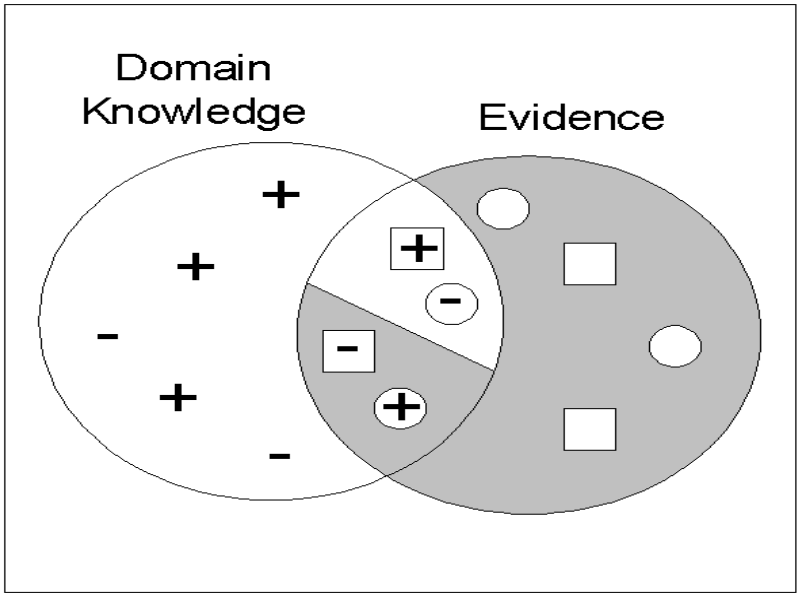
- 76.42% των ζευγών έχουν correlation >

Υποκειμενικά μέτρα ενδιαφέροντος

- Αντικειμενικό μέτρο:
  - Ιεράρχηση προτύπων με βάση τα στατιστικά που υπολογίσθηκαν από τα δεδομένα
    - 21 μέτρα συσχέτισης (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).
- Υποκειμενικό μέτρο:
  - Ιεράρχηση προτύπων με βάση την ερμηνεία του χρήστη
    - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is actionable (Silberschatz & Tuzhilin)

Υποκειμενικά μέτρα ενδιαφέροντος

- Χρειάζεται μοντελοποίηση της γνώσης των χρηστών(domain knowledge)



+ Πρότυπο που αναμένεται να είναι συχνό

- Πρότυπο που αναμένεται να είναι μη συχνό

Πρότυπο που βρέθηκε να είναι συχνό

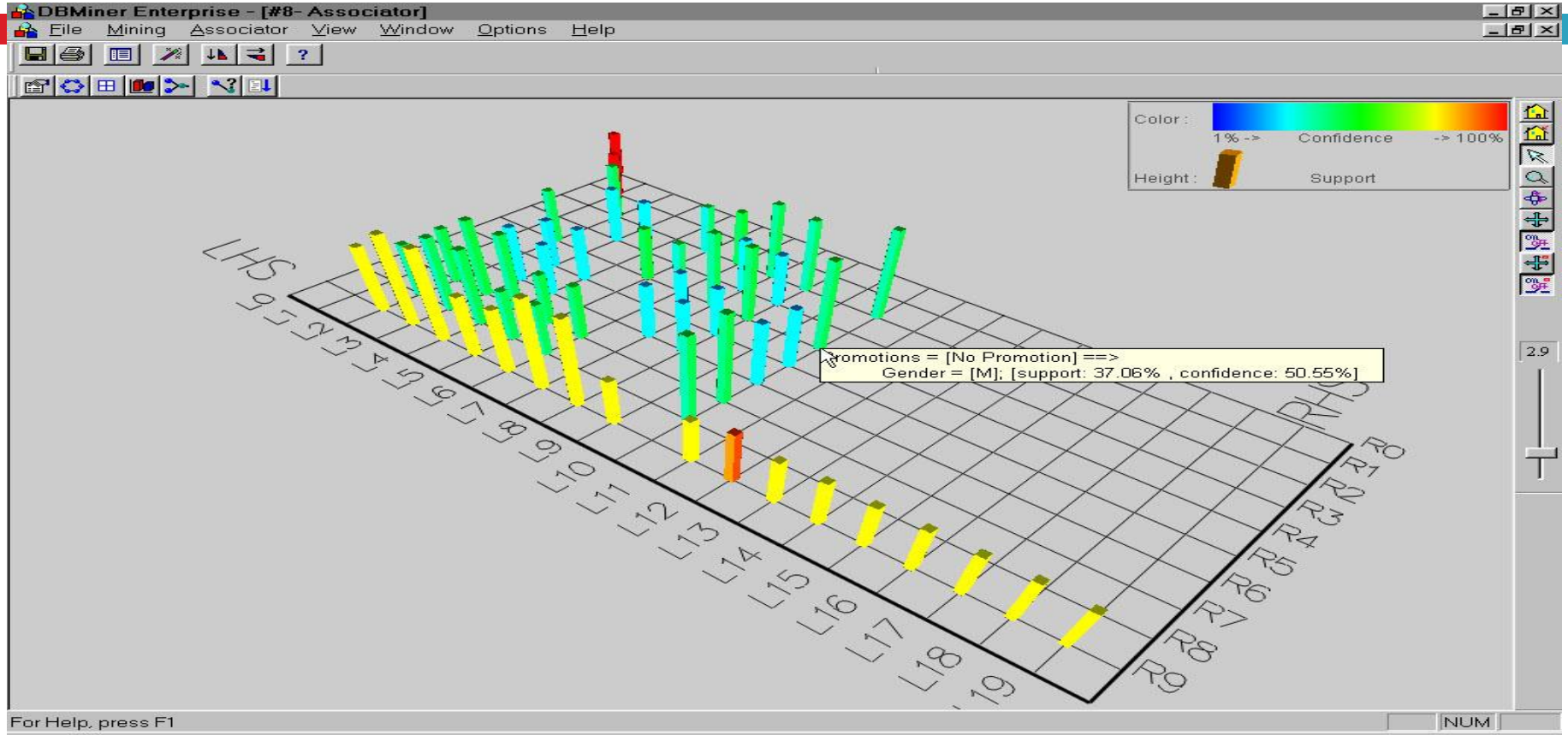
Πρότυπο που βρέθηκε να είναι μη συχνό

Προσδοκώμενα πρότυπα

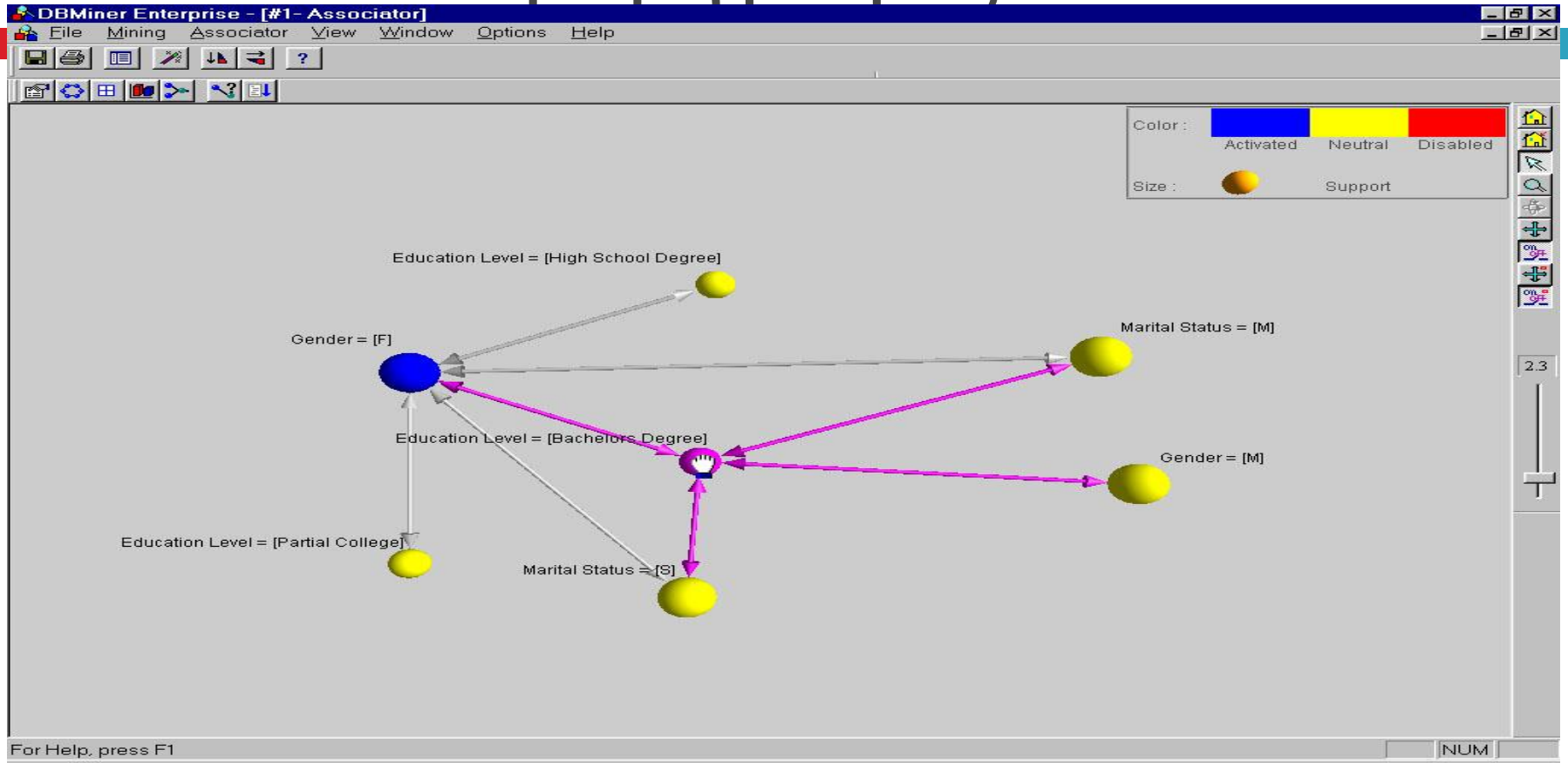
Μη- προσδοκώμενα πρότυπα

- Χρειάζεται να συνδυάσουμε τη γνώση των χρηστών με δεδομένα

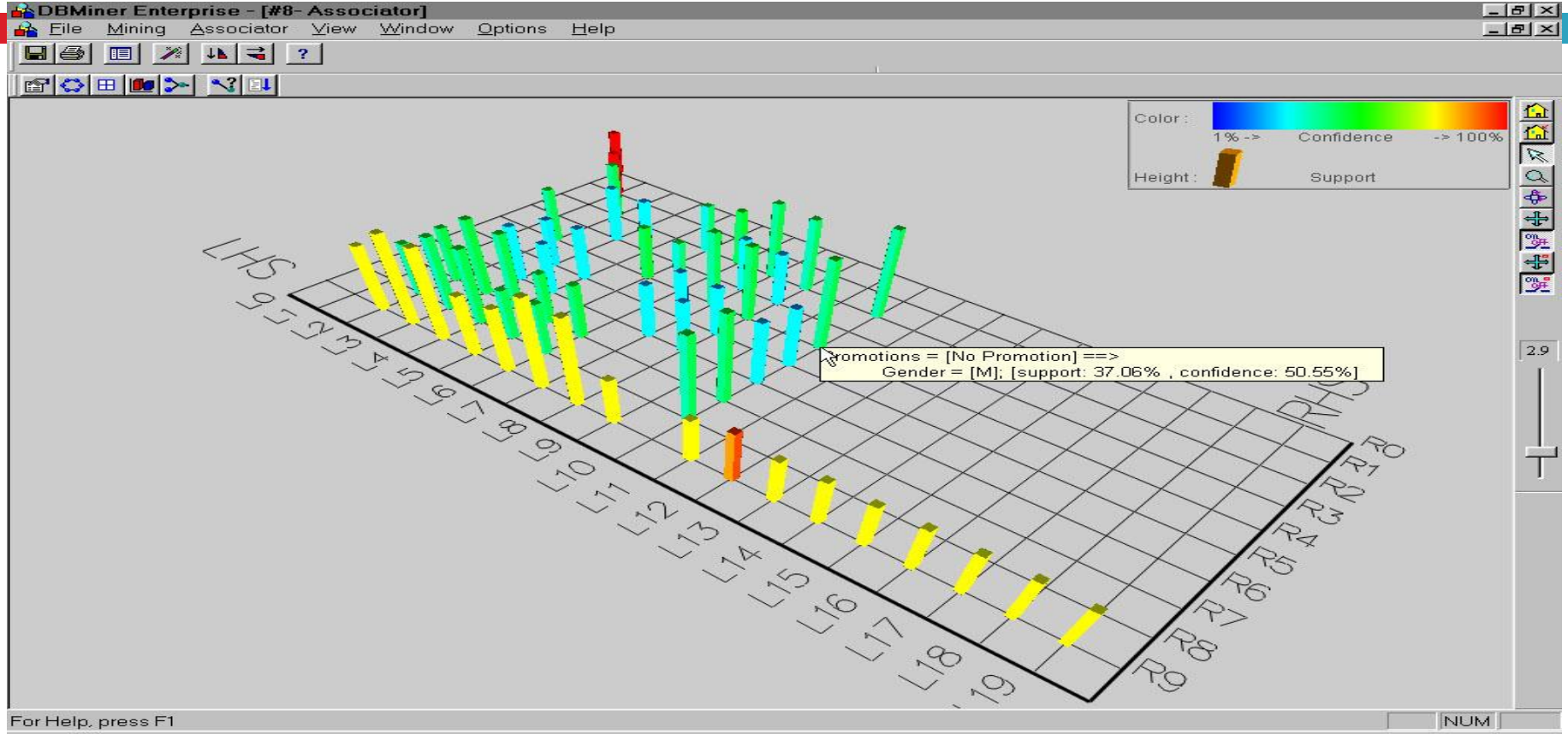
Οπτικοποίηση: απλός γράφος



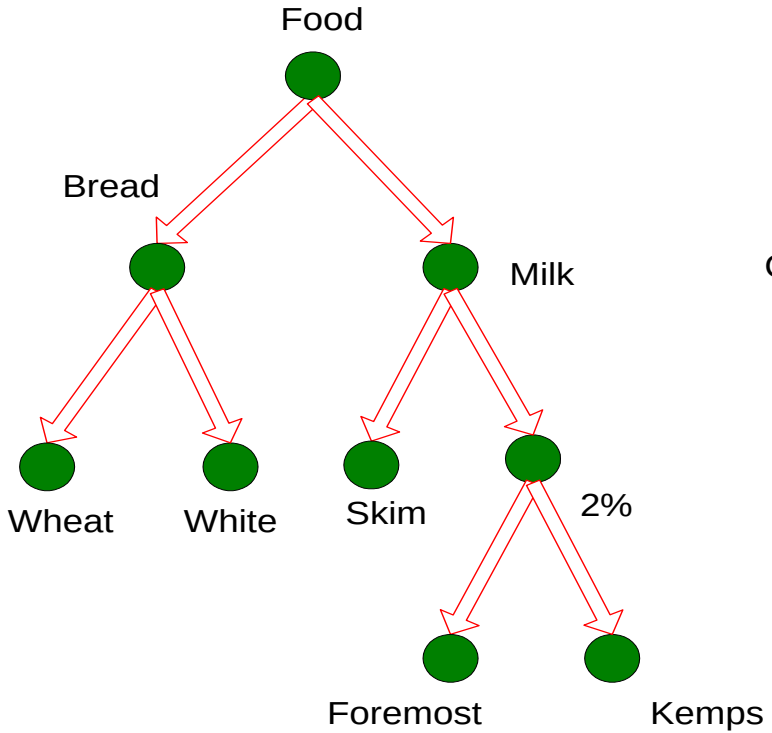
Οπτικοποίηση: γράφος κανόνων

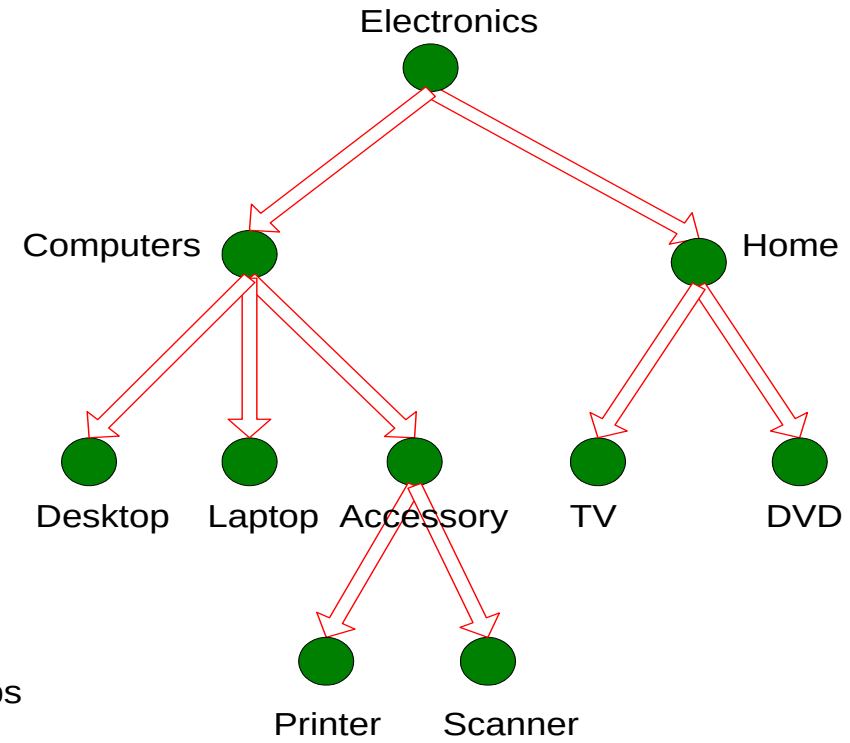


Οπτικοποίηση: (SGI/MineSet 3.0)



Κανόνες συσχέτισης πολλών επιπέδων





Κανόνες συσχέτισης πολλών επιπέδων

- Γιατί είναι χρήσιμοι;
  - Οι κανόνες στα χαμηλότερα επίπεδα δεν έχουν αρκετή υποστήριξη σε κανένα στοιχειοσύνολο
- Οι κανόνες στα χαμηλότερα επίπεδα είναι πάρα πολύ συγκεκριμένοι
  - π.χ. skim milk →white bread, 2% milk → wheat bread, skim milk → wheat bread, etc.
    - είναι ενδεικτικοί της συσχέτισης μεταξύ γάλατος και ψωμιού

Κανόνες συσχέτισης πολλών επιπέδων

- Προσέγγιση 1:
  - Επέκταση κάθε δοσοληψίας με στοιχεία από τα υψηλότερα επίπεδα της ιεραρχίας
    - Αρχική Δοσοληψία: {skim milk, wheat bread}
    - Επαυξημένη Δοσοληψία: {skim milk, wheat bread, milk,bread, food}
- Θέματα:
  - Τα στοιχεία στα υψηλότερα επίπεδα θα εμφανίζονται πολύ συχνά, μεγάλους μετρητές υποστήριξης
    - εάν το κατώφλι της υποστήριξης είναι πολύ χαμηλά, προκύπτουν πολλά συχνά πρότυπα από τα ανώτερα επίπεδα
  - Αυξημένη διαστατικότητα των δεδομένων

Κανόνες συσχέτισης πολλών επιπέδων

- Πως μεταβάλλονται εμπιστοσύνη (conf) και υποστήριξη (σ) με τη διάσχιση της ιεραρχίας των εννοιών
  - Αν X το πατρικό στοιχείο των X1 and X2, then $\sigma(X) \leq \sigma(X1) + \sigma(X2)$
  - If $\sigma(X1 \cup Y1) \geq$ minsup,
    and X πατρικός του X1, Y πατρικός του Y1
    then $\sigma(X \cup Y1) \geq$ minsup, $\sigma(X1 \cup Y) \geq$ minsup
    $\sigma(X \cup Y) \geq$ minsup
  - If conf$(X1 \Rightarrow Y1) \geq$ minconf,
    then conf$(X1 \Rightarrow Y) \geq$ minconf

Κανόνες συσχέτισης πολλών επιπέδων

- Προσέγγιση 2:
  - Αρχικά→Δημιουργία συχνών προτύπων στο υψηλότερο επίπεδο
  - Έπειτα→Δημιουργία συχνών προτύπων στα επόμενα υψηλότερα επίπεδα και ούτω καθεξής
- Ζητήματα:
  - Οι απαιτήσεις I/O θα αυξηθούν δραματικά επειδή χρειάζεται να γίνονται περισσότερα περάσματα των δεδομένων
  - Ίσως χαθούν ορισμένοι ενδιαφέροντες κανόνες συσχέτισης