ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

Εξόρυξη Δεδομένων στον Παγκόσμιο Ιστό

Κανόνες Συσχέτισης

Μανώλης Μαραγκουδάκης

Τμήμα Μηχανικών Πληροφοριακών και Επικοινωνιακών Συστημάτων

Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

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

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)

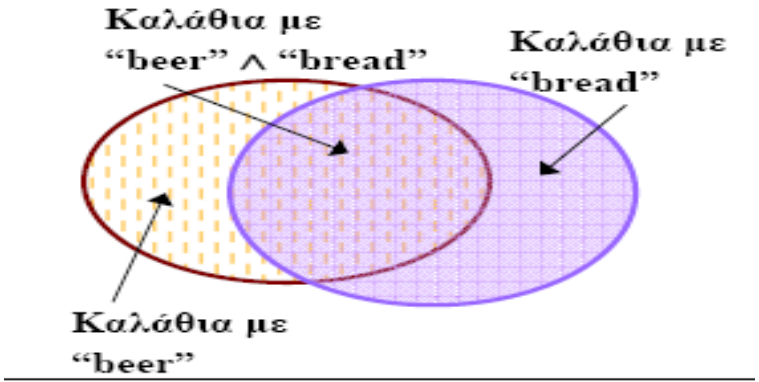
Παράδειγμα

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

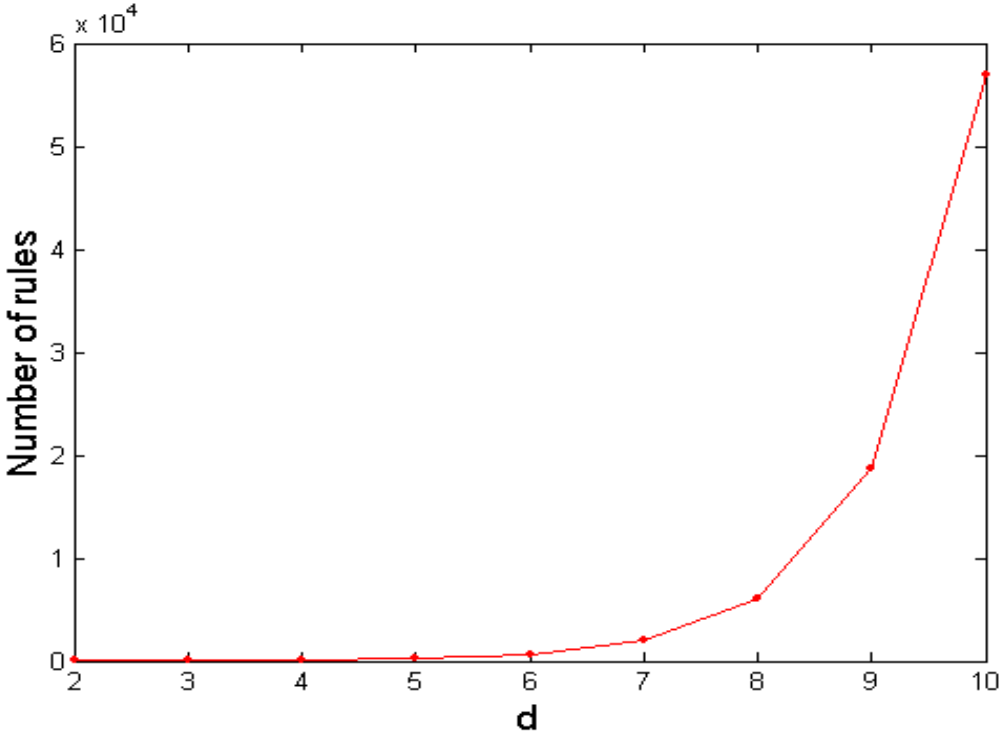
- Εύρεση Κανόνων Συσχέτισης
  - Είσοδος: Ένα σύνολο από δοσοληψίες T
  - Έξοδος: Όλοι οι κανόνες με
    - support ≥ *minsup*
    - confidence≥ *minconf*

- 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)

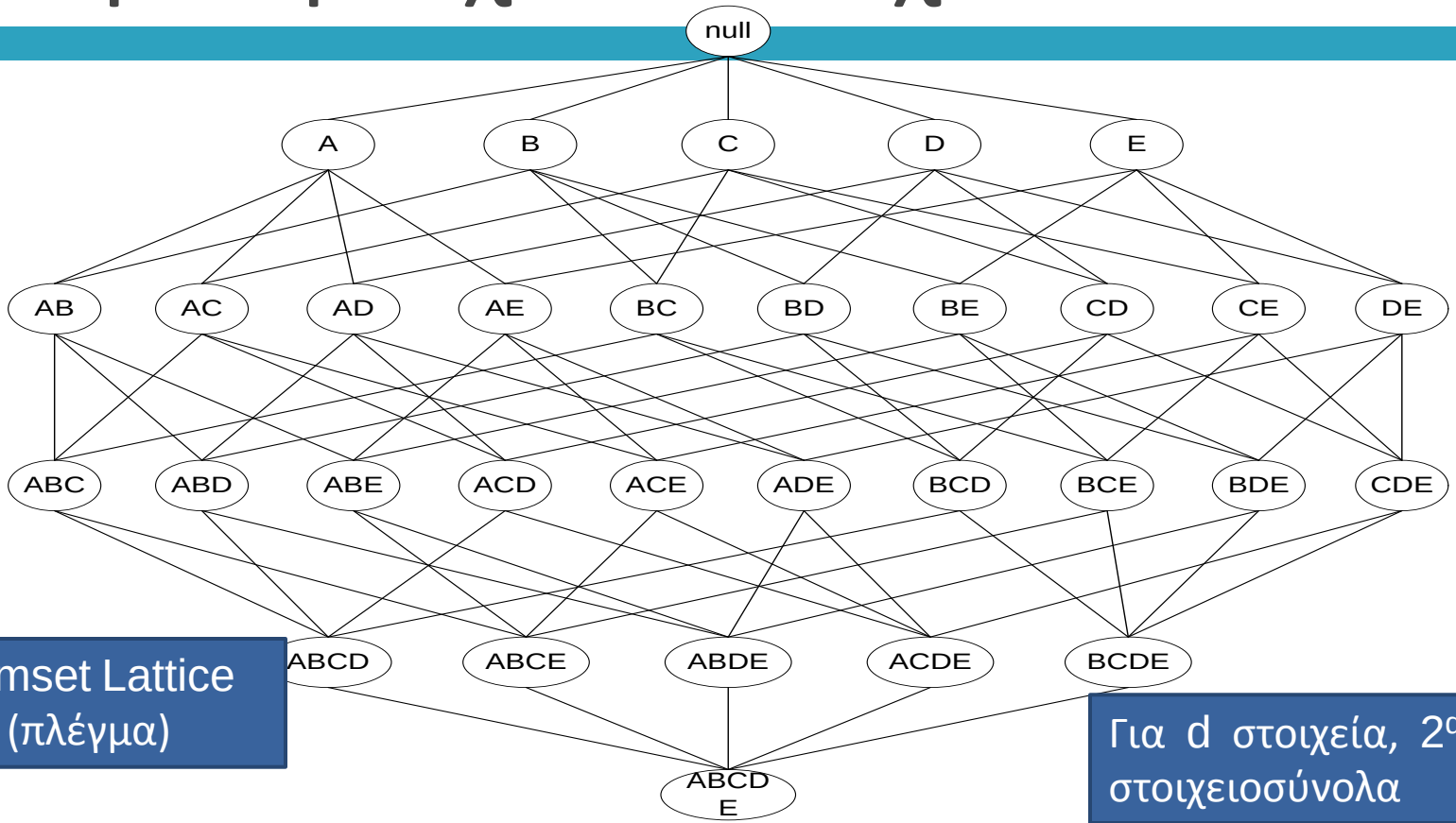
Αν είχαμε minsup = 0.5, θα αποκλείαμε και τους έξι κανόνες

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

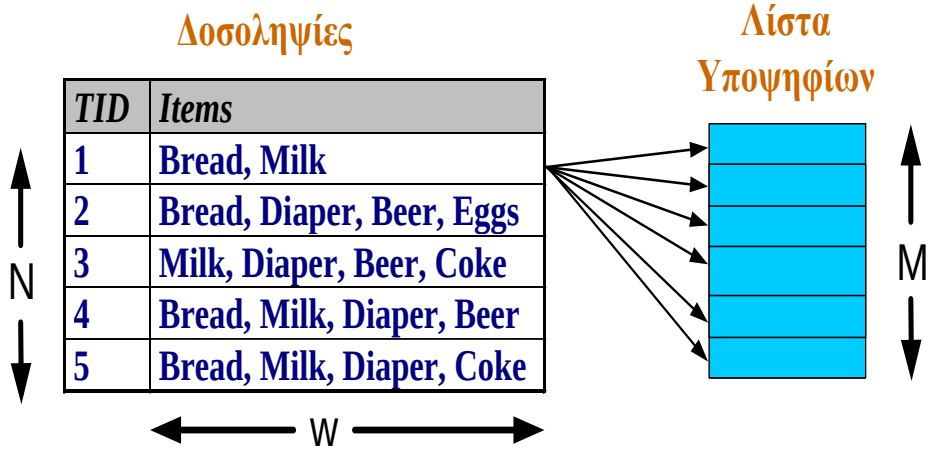


Itemset Lattice
(πλέγμα)

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

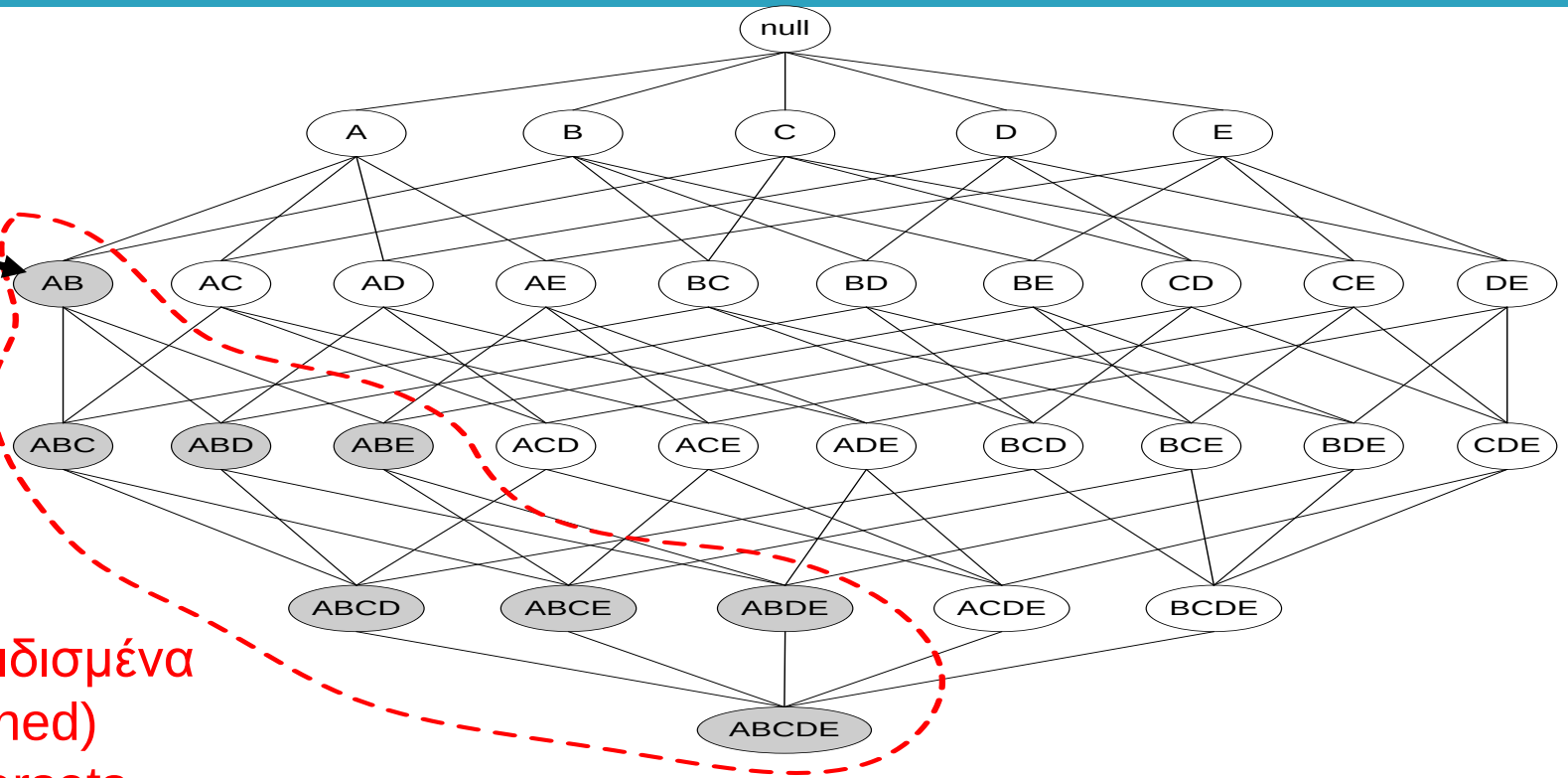
- Συχνό ονομάζεται το itemset που έχει υποστήριξη πάνω από ένα κατώφλι
- Παράδειγμα
  - (κατώφλι = 40%):
    - {Beer}
    - {Bread}
    - {PeanutButter}
    - {Bread, PeanutButter}

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

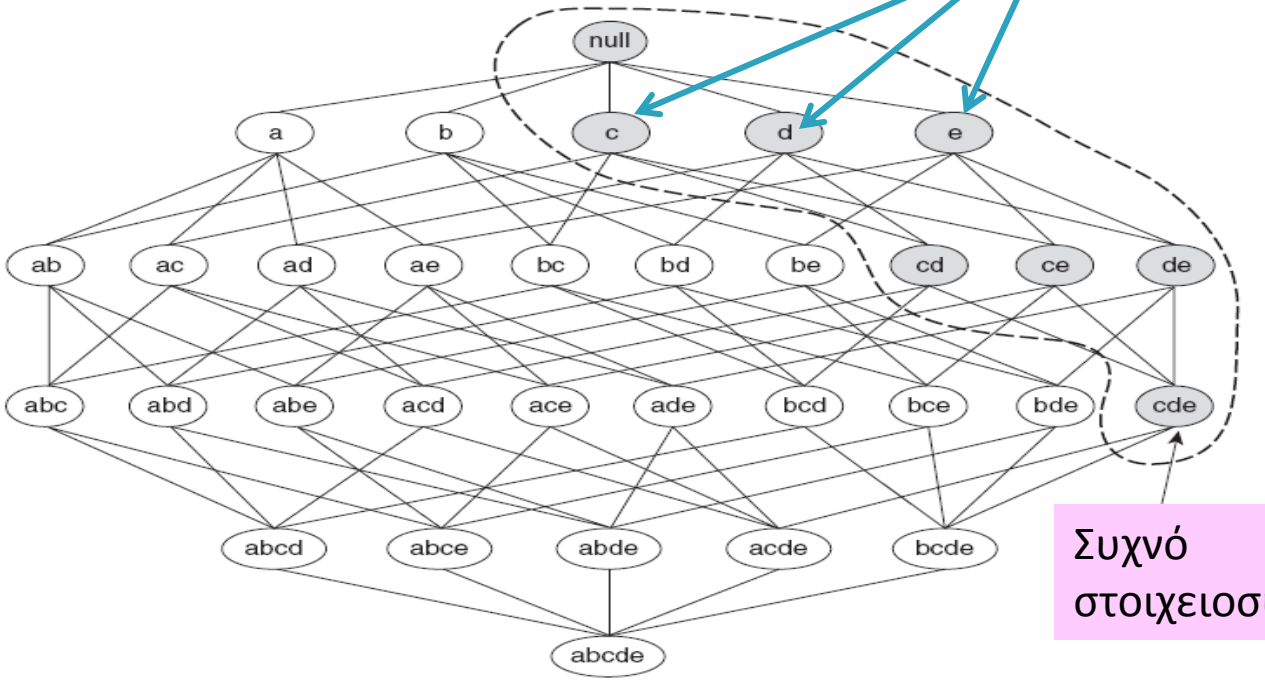
Βρέθηκε μη συχνό



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

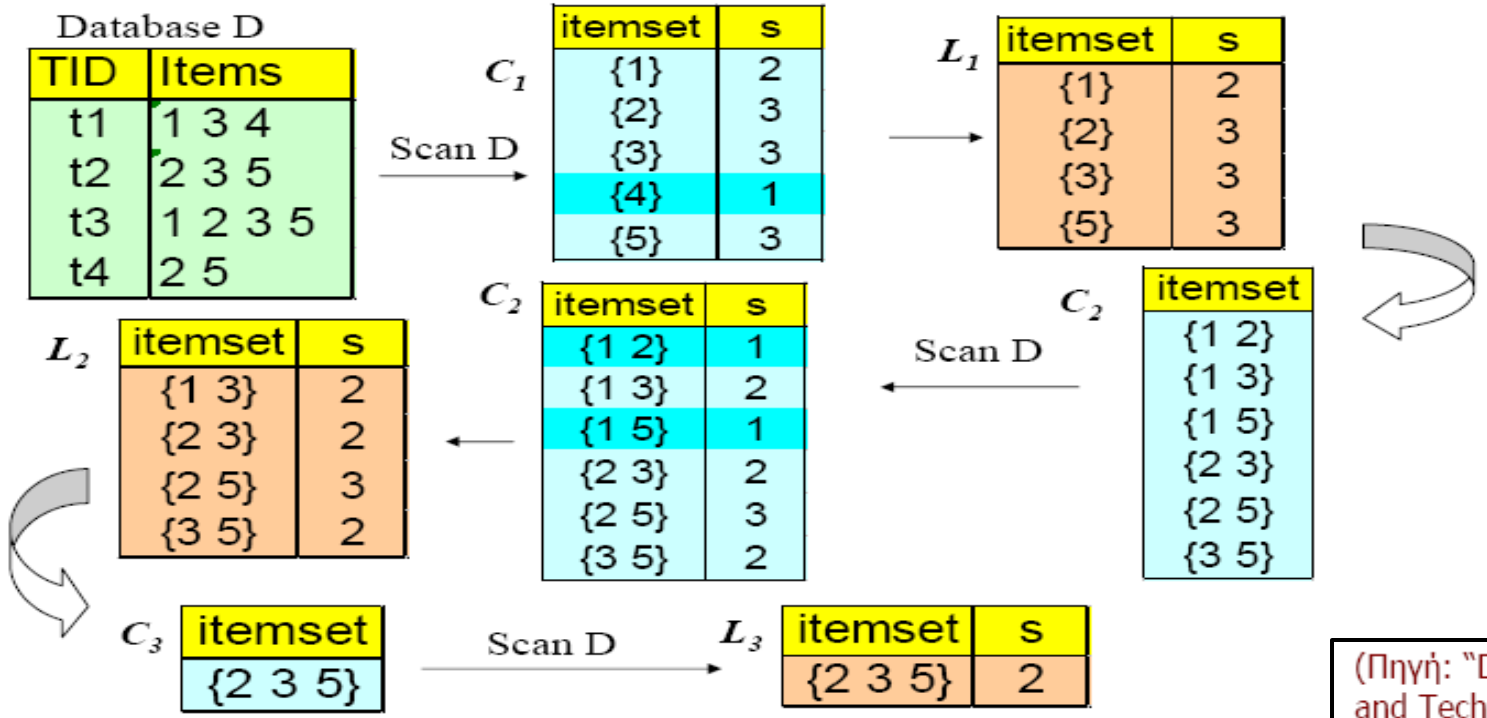
Τριπλέτες (3-itemsets)

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

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

Ο αλγόριθμος FP-growth

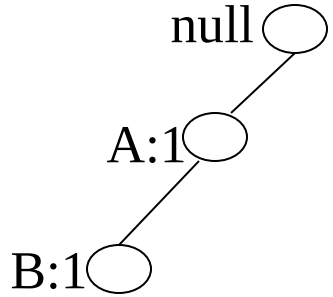
- Χρησιμοποιεί μια συμπιεσμένη αναπαράσταση της βάσης δεδομένων με βάση ένα FP-tree

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

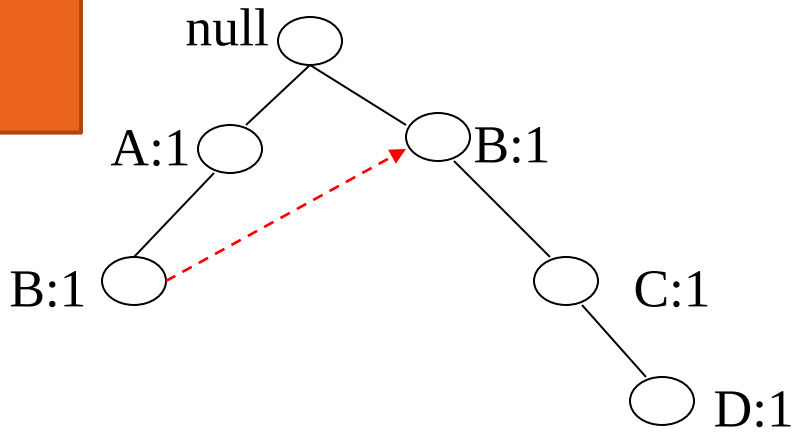
FP-tree

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:



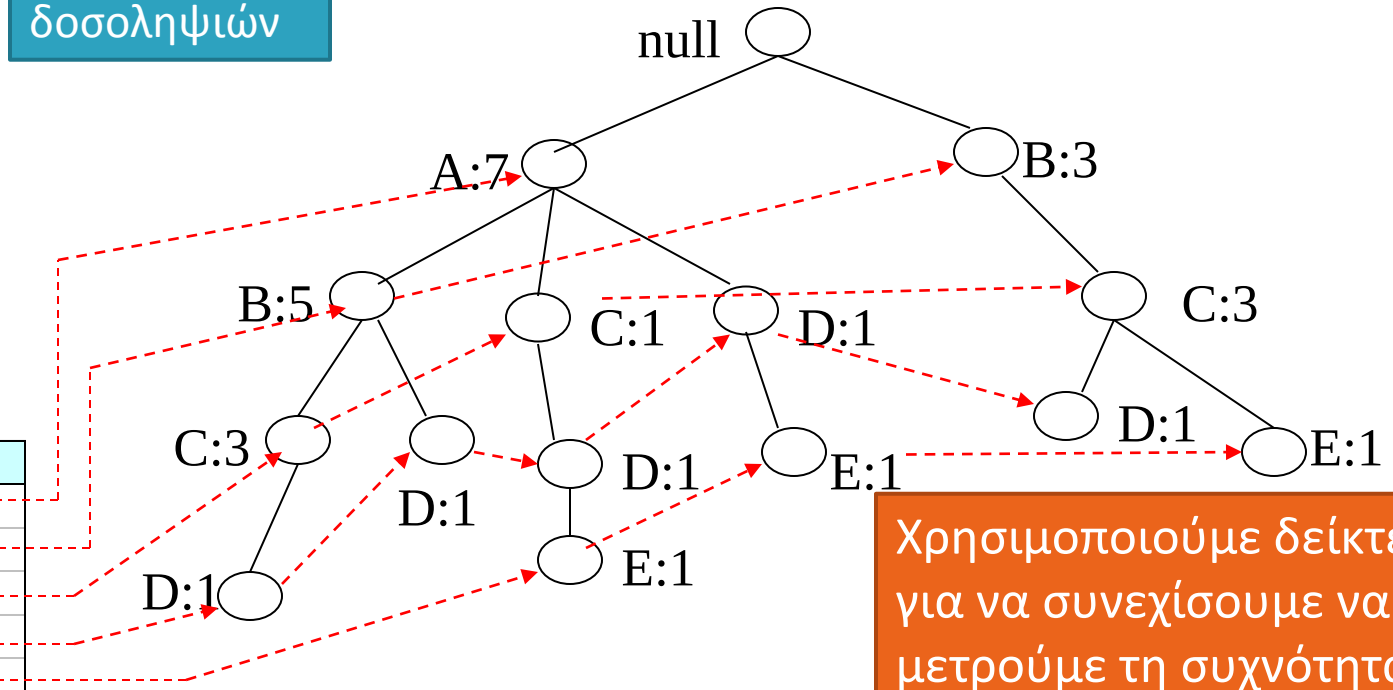
FP-Tree

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

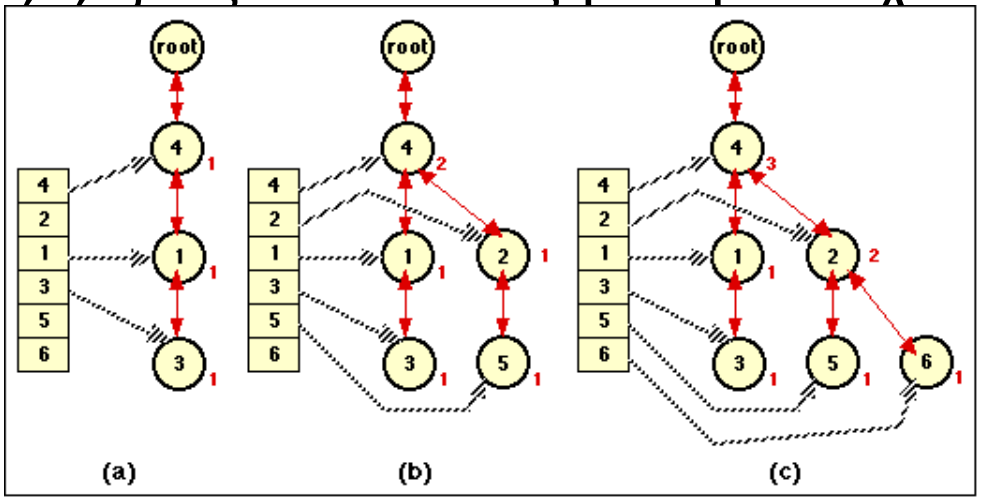
Item	Pointer
A	
B	
C	
D	
E	



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}

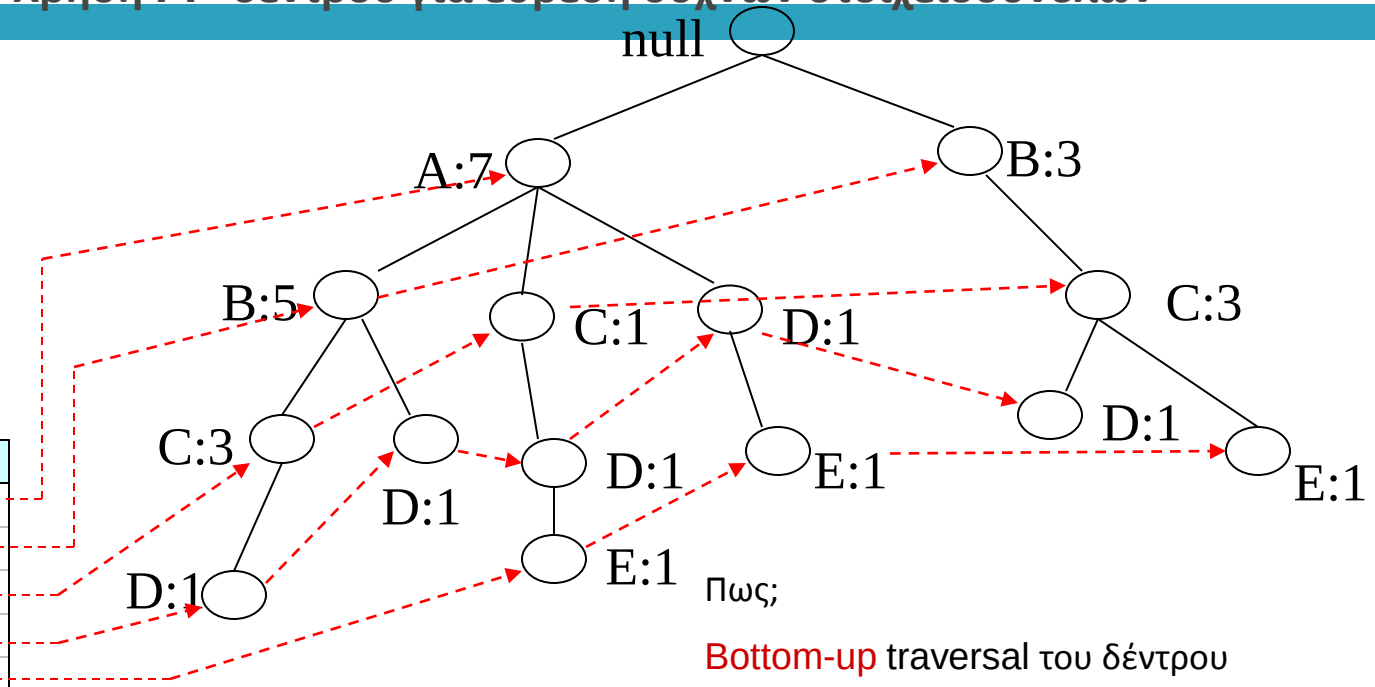
Αλγόριθμος FP-Growth

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	



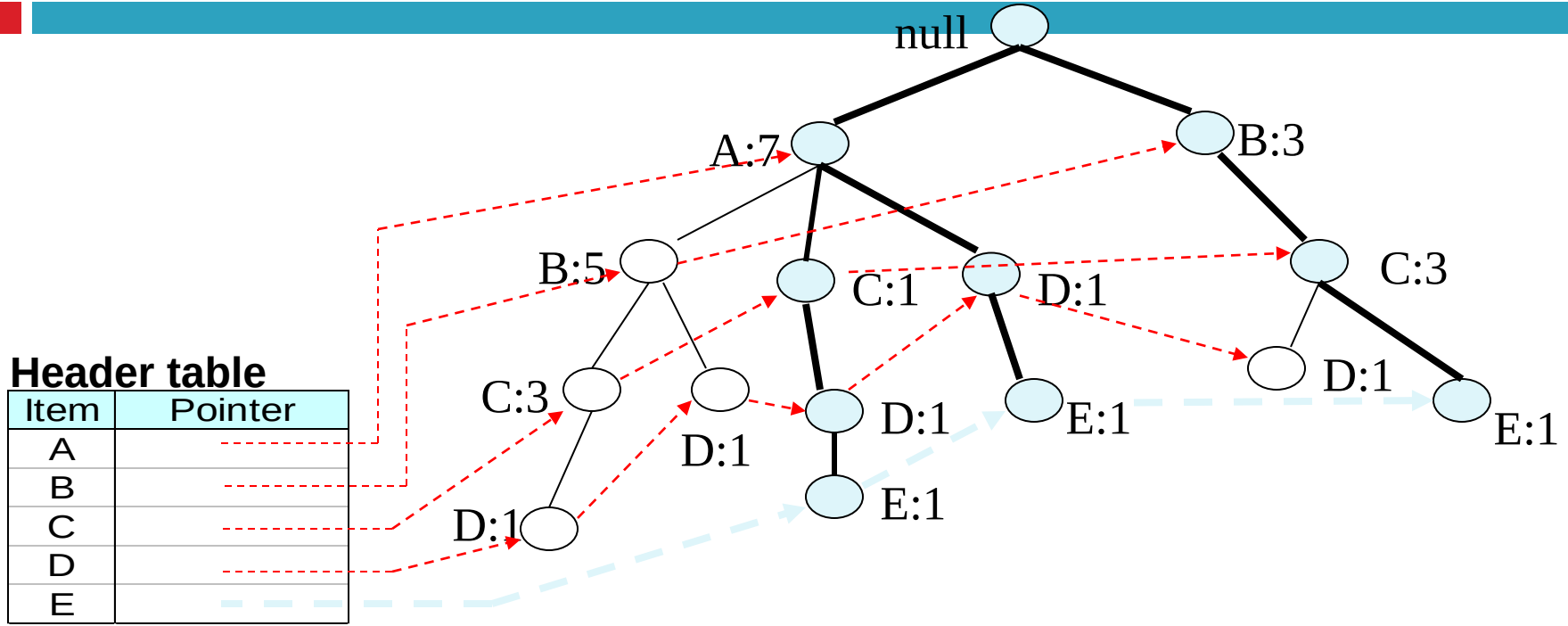
Πως;

Bottom-up traversal του δέντρου

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

Αλγόριθμος FP-Growth

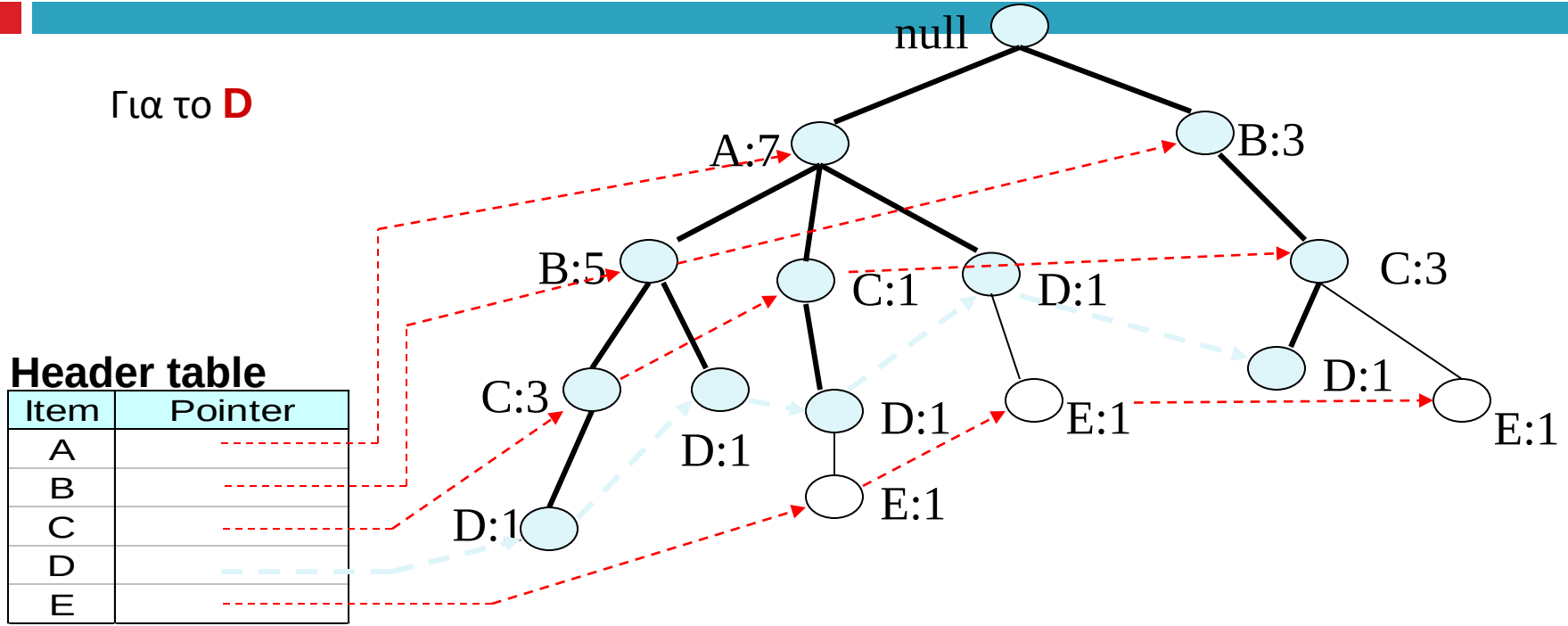
Υποπρόβλημα: Βρες συχνά στοιχειοσύνολα που τελειώνουν σε **E**



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

Αλγόριθμος FP-Growth

Για το **D**

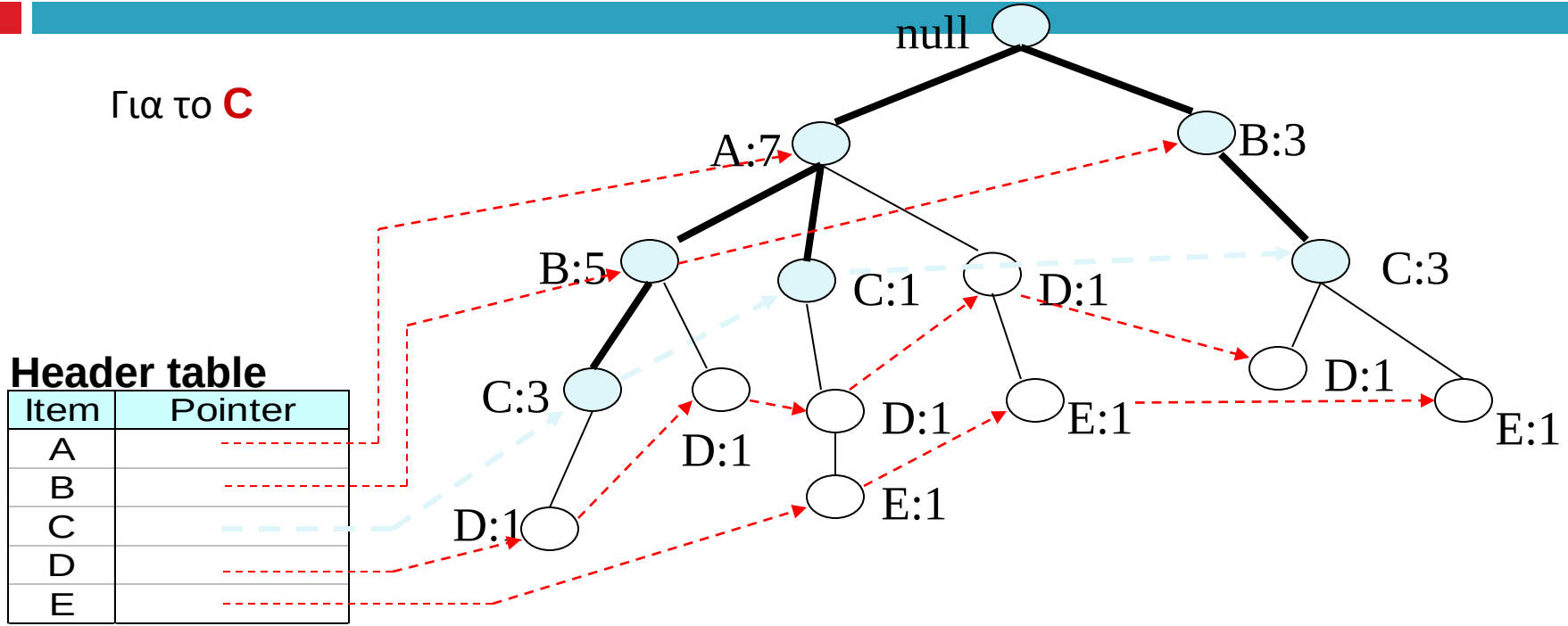


Header table

Item	Pointer
A	
B	
C	
D	
E	

Αλγόριθμος FP-Growth

Για το **C**

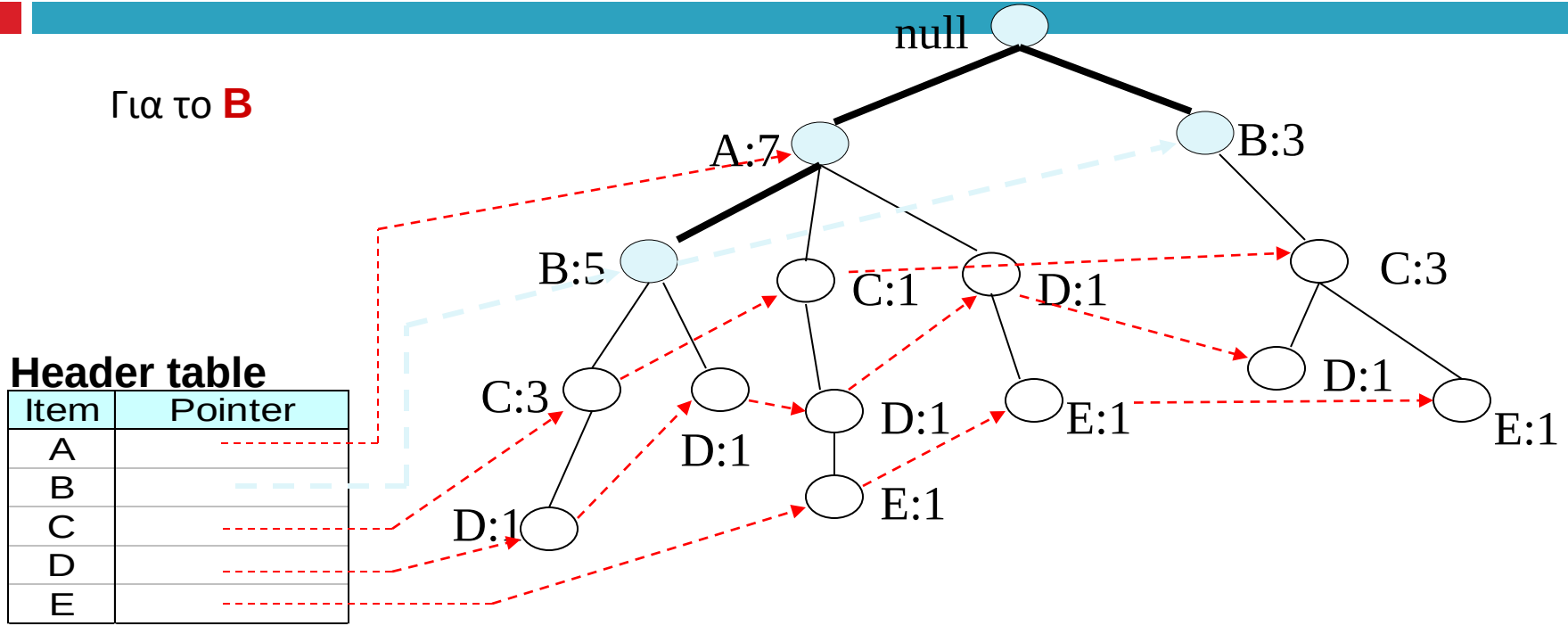


Header table

Item	Pointer
A	
B	
C	
D	
E	

Αλγόριθμος FP-Growth

Για το **B**

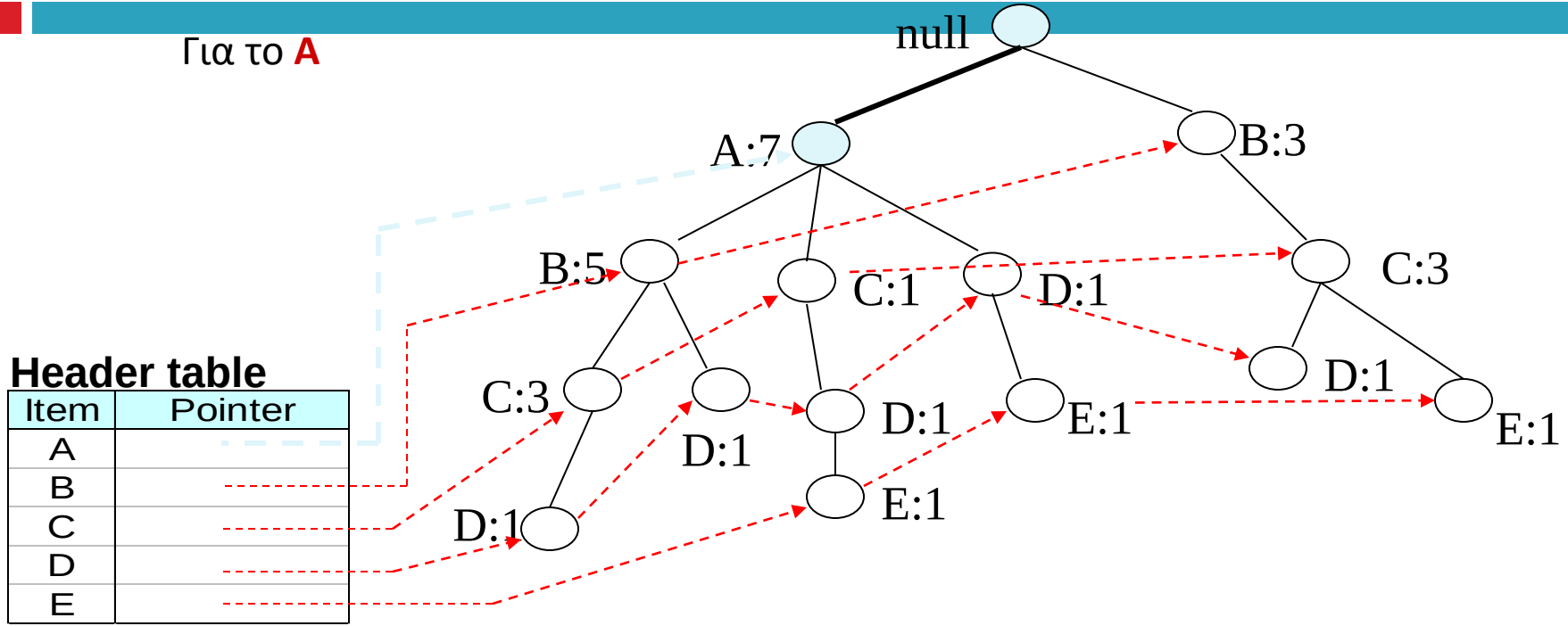


Header table

Item	Pointer
A	
B	
C	
D	
E	

Αλγόριθμος FP-Growth

Για το **A**

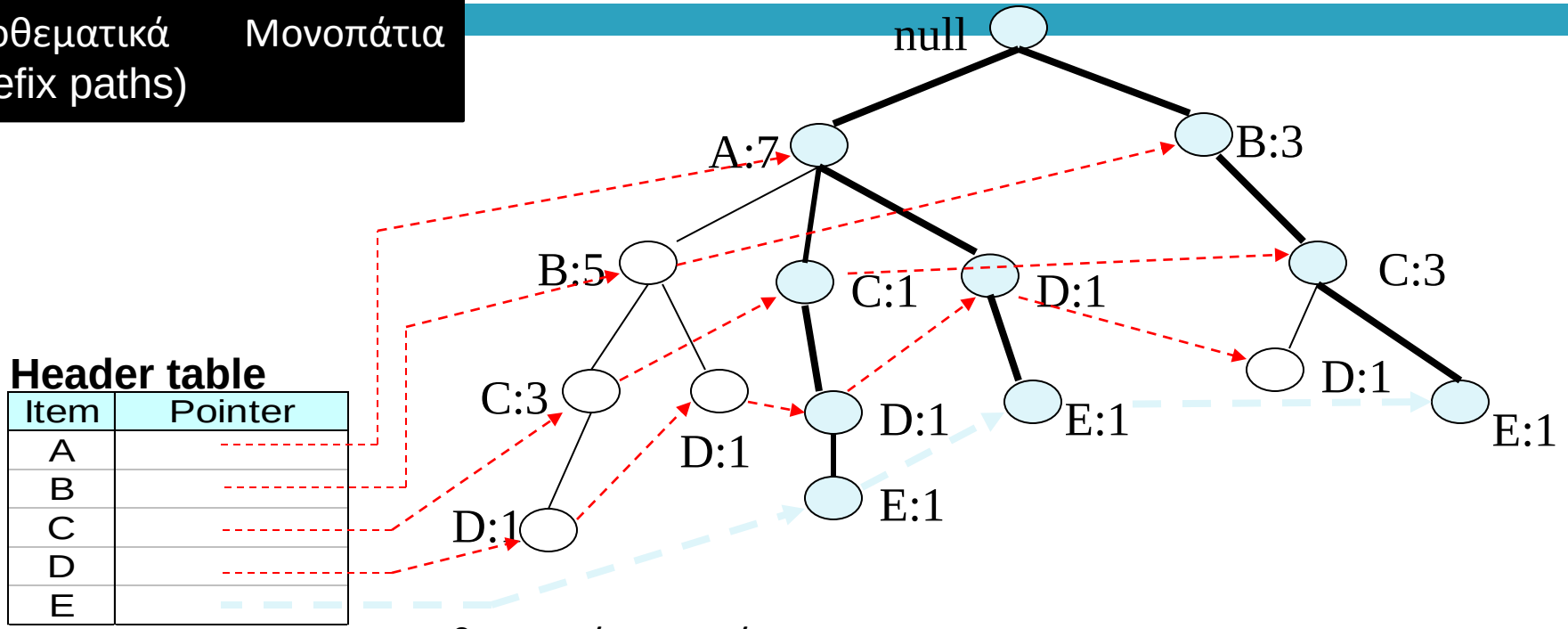


Αλγόριθμος FP-Growth

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:

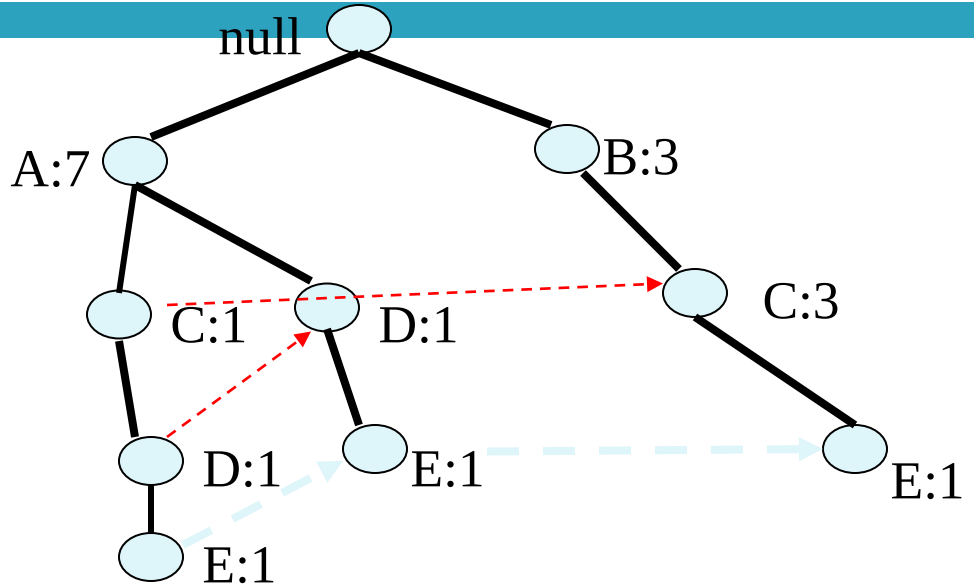
{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}

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

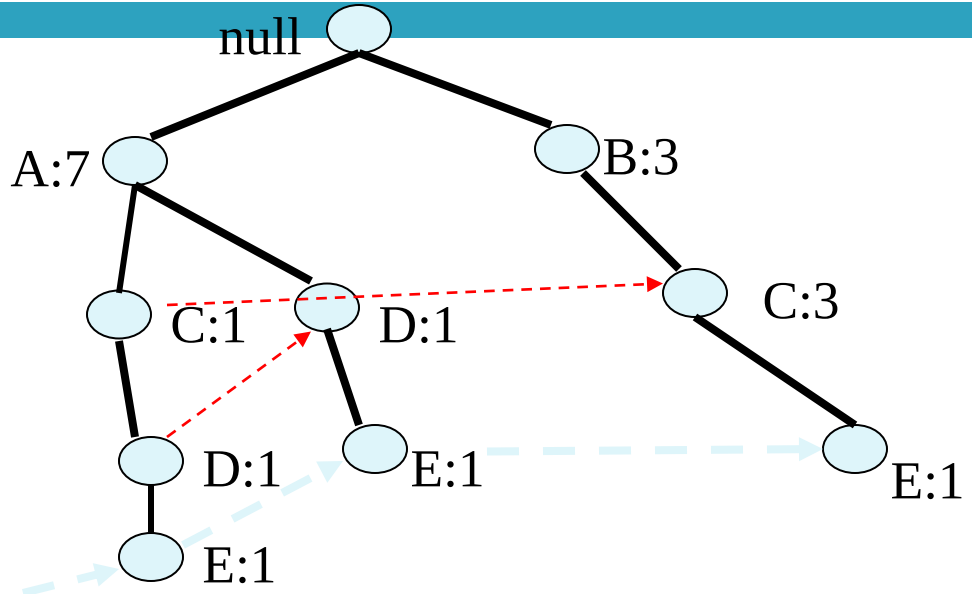
Έστω minsup = 2

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

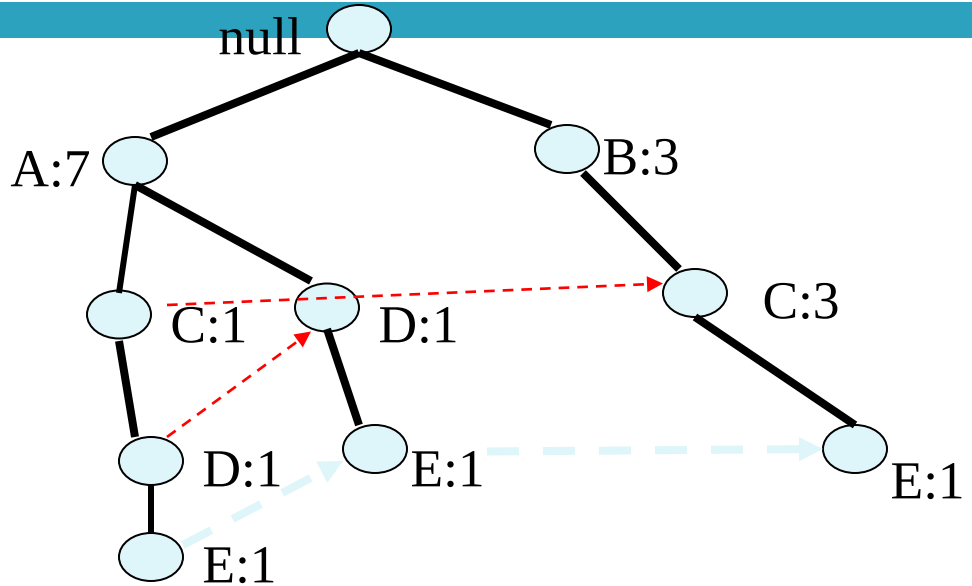
Αλγόριθμος FP-Growth

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

Δύο αλλαγές

(1) Αλλαγή των μετρητών

(2) Περικοπή



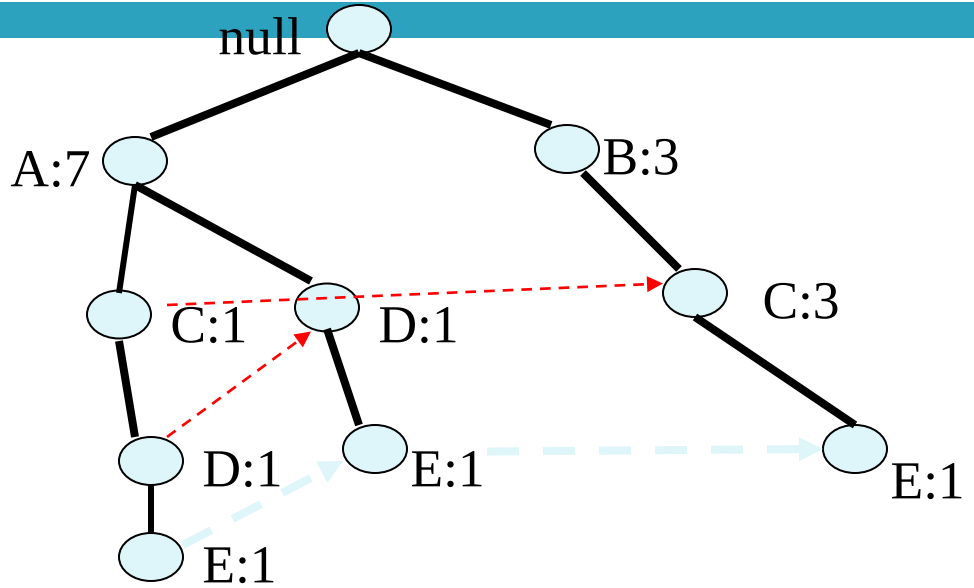
{E} συχνό άρα προχωράμε για DE, CE, BE, AE

Αλγόριθμος FP-Growth

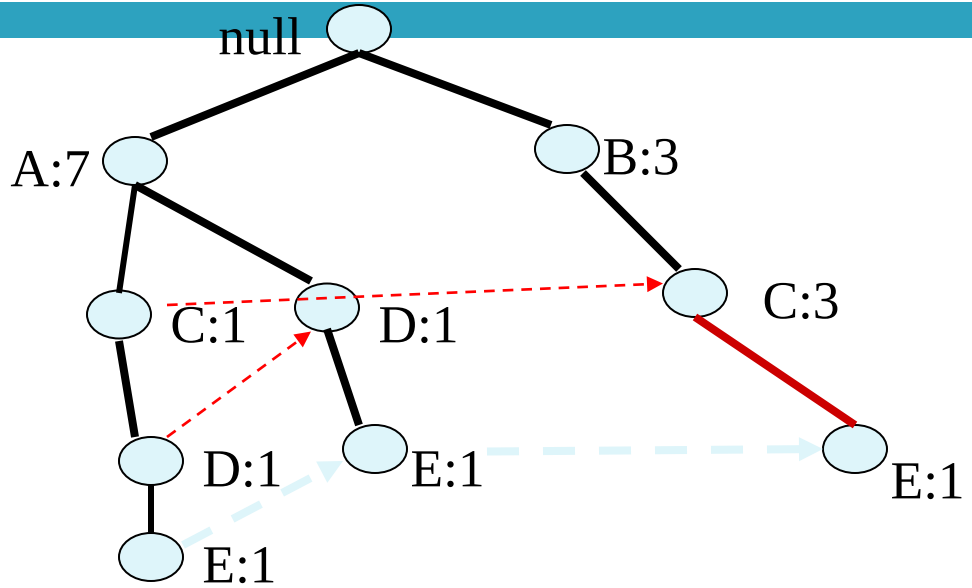
Αλλαγή μετρητών

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

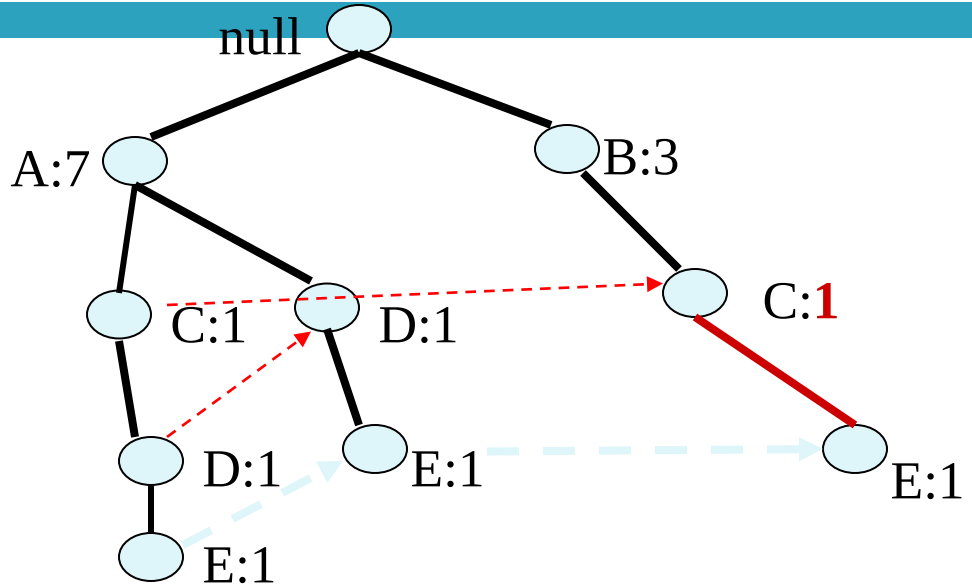
Πχ στο null->B->C->E μετράμε και την {B, C}



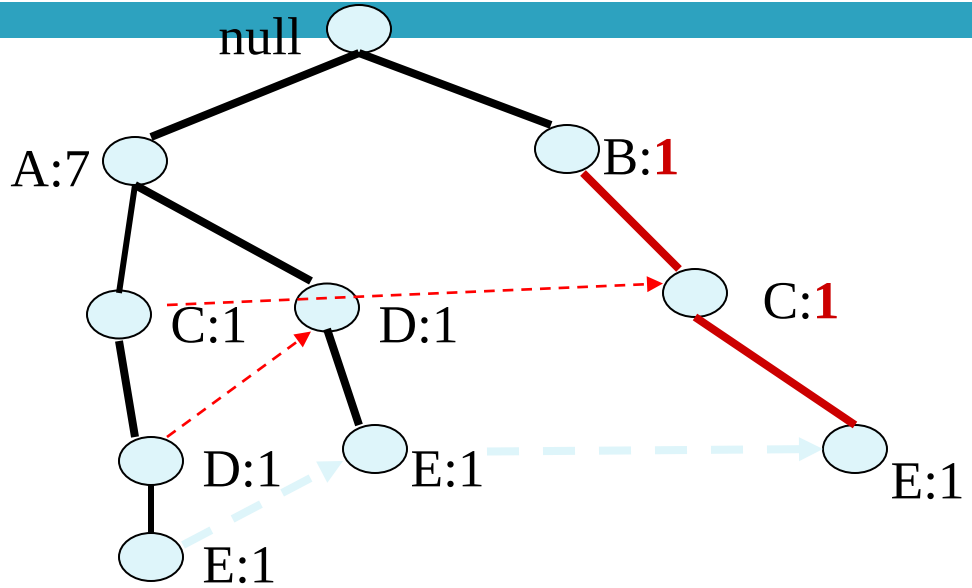
Αλγόριθμος FP-Growth



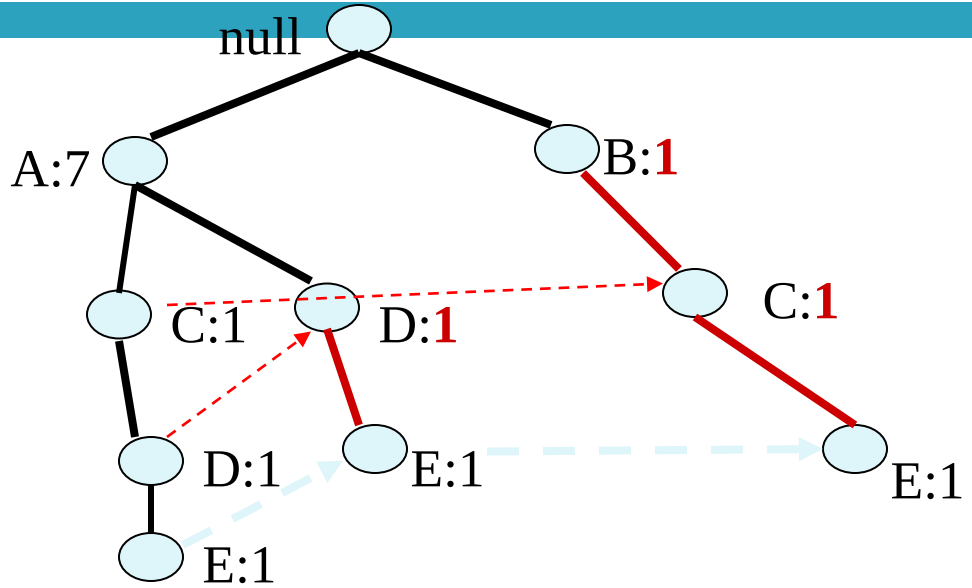
Αλγόριθμος FP-Growth



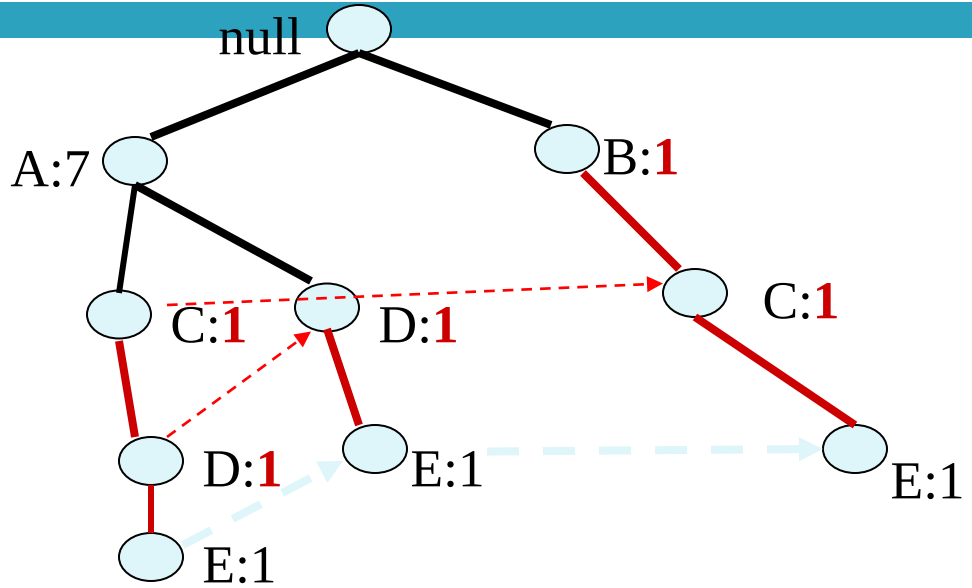
Αλγόριθμος FP-Growth



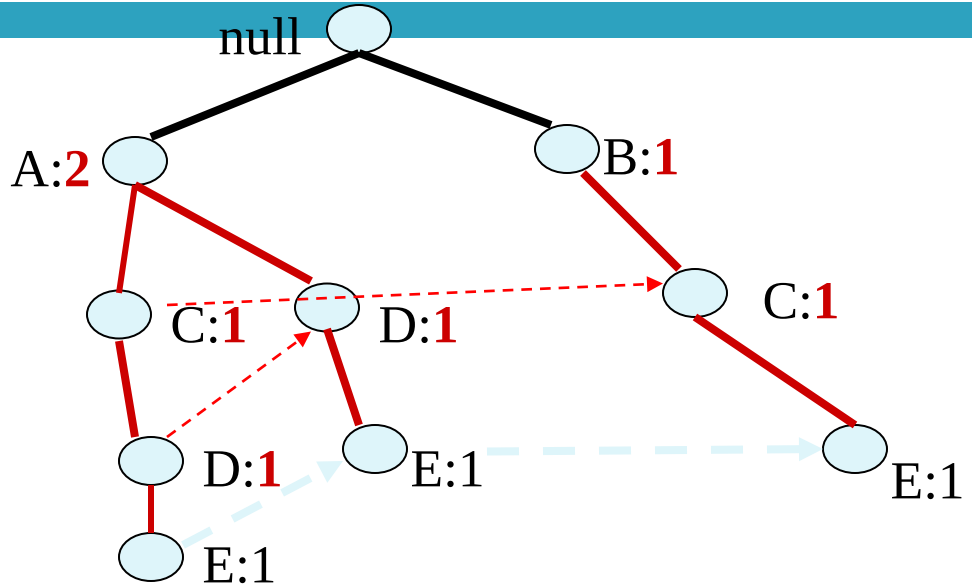
Αλγόριθμος FP-Growth



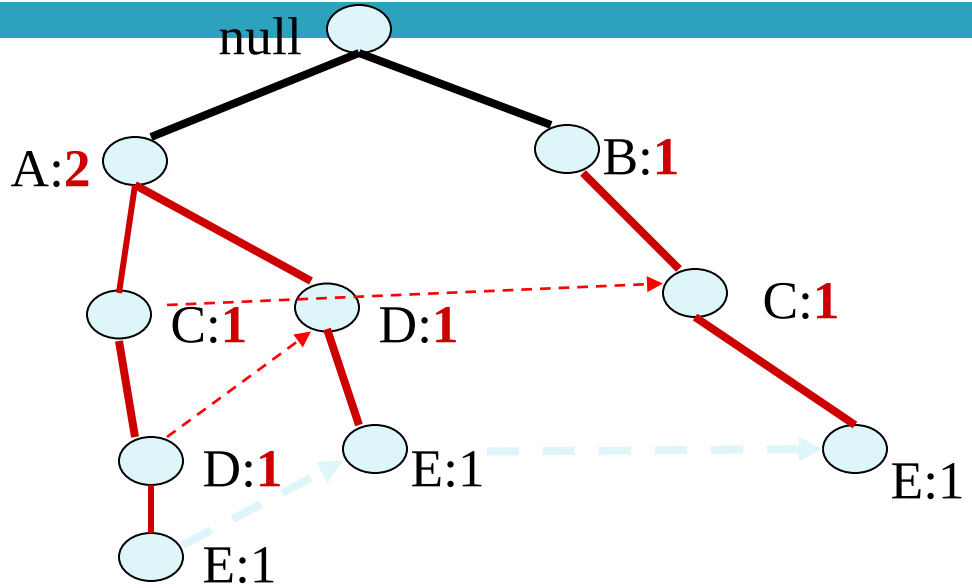
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth



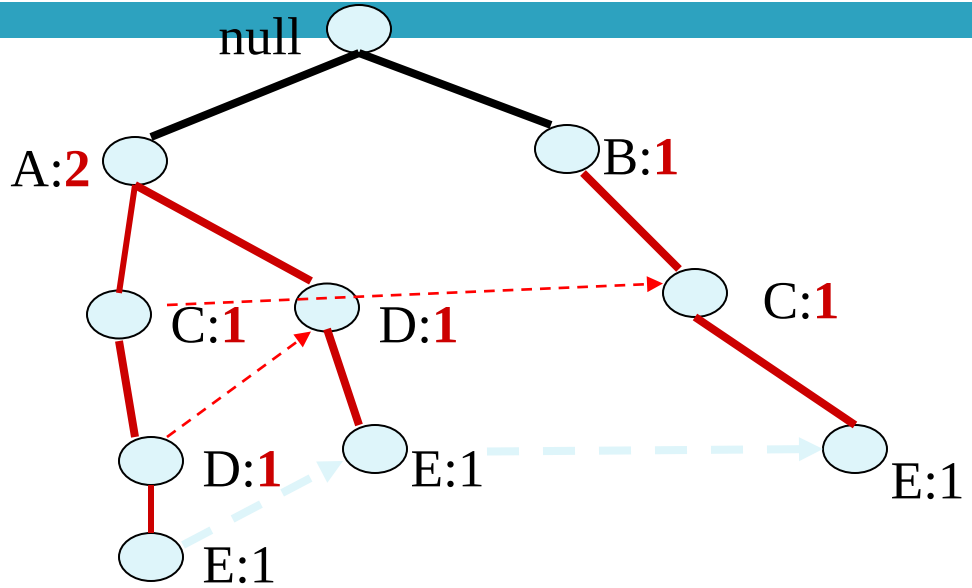
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Περικοπή (truncate)

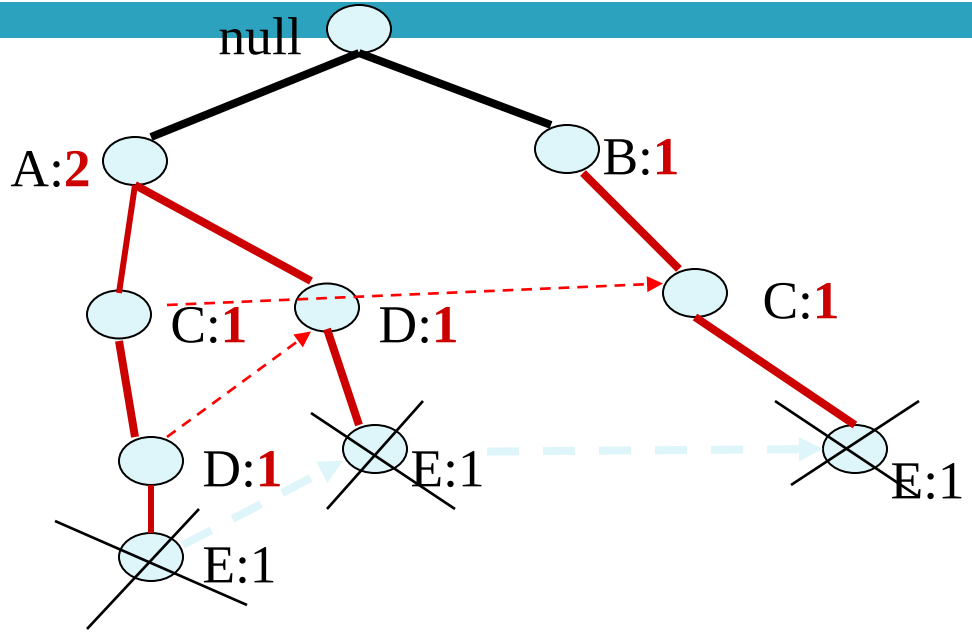
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth

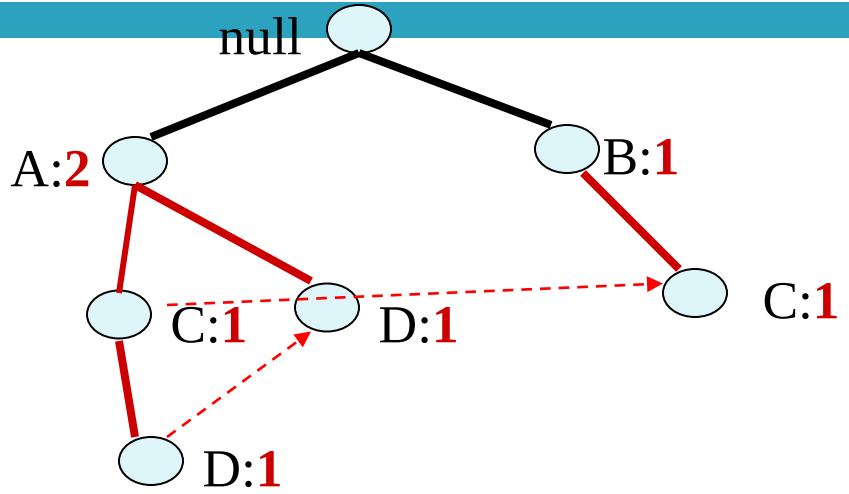
Περικοπή (truncate)

Σβήσε τους κόμβους του E



Περικοπή (truncate)

Σβήσε τους κόμβους του E

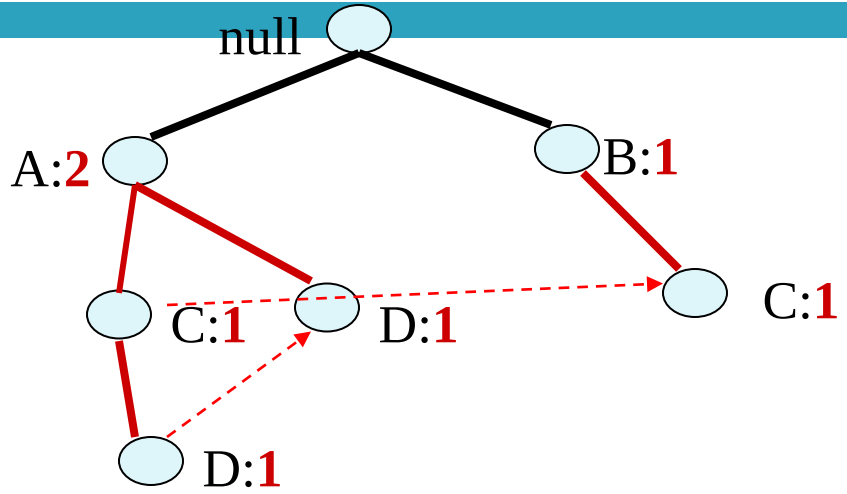


Αλγόριθμος FP-Growth

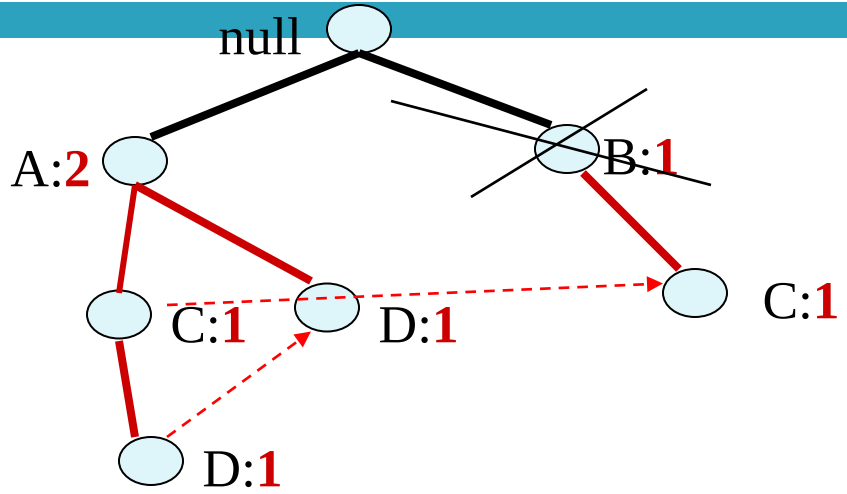
Πιθανή περαιτέρω περικοπή

Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης

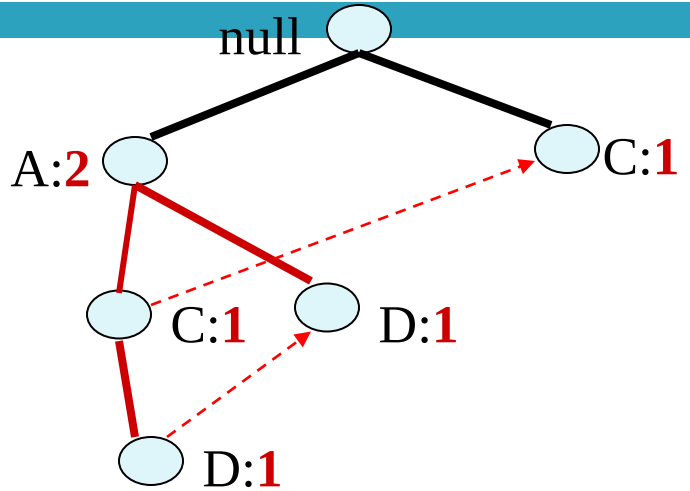
Πχ το B -> περικοπή



Αλγόριθμος FP-Growth



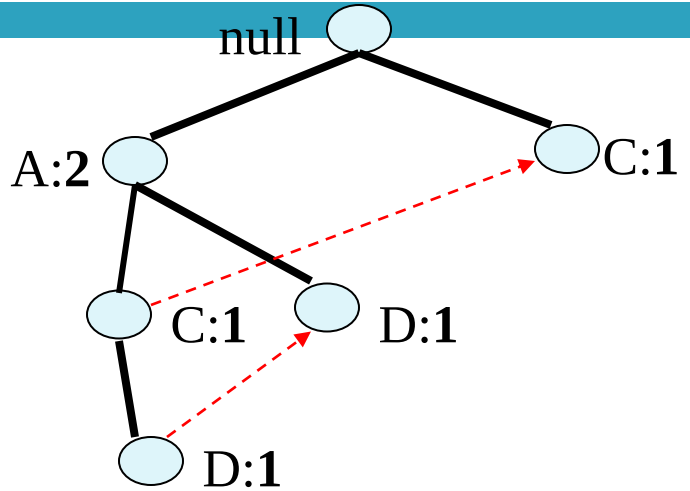
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

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

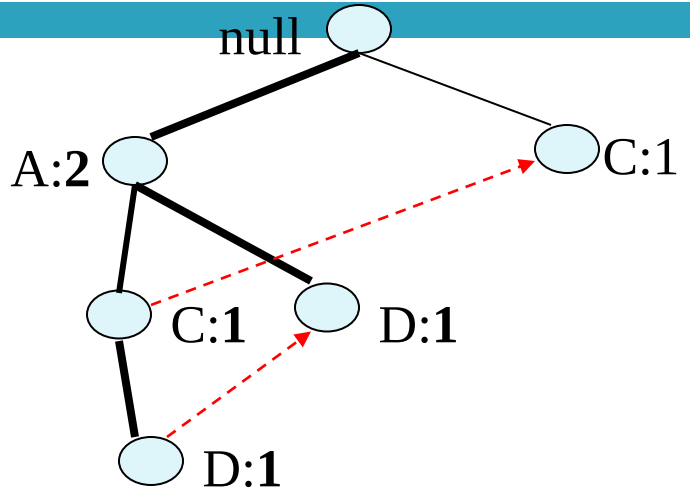


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)

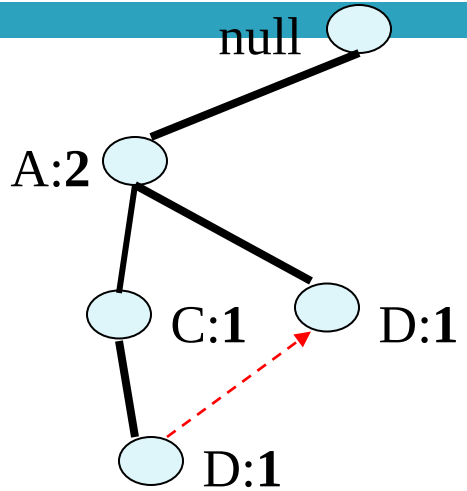


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



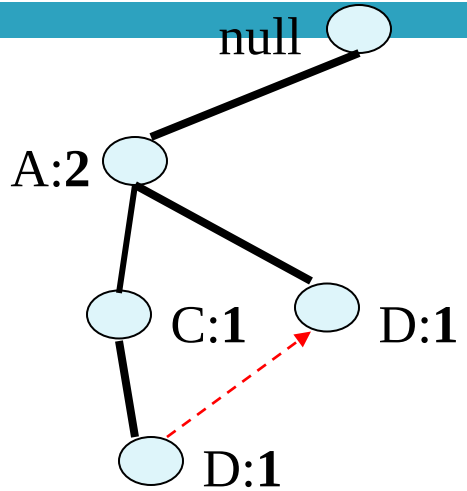
Αλγόριθμος FP-Growth

Βρες την υποστήριξη του {D, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

Οπότε {D, E} συχνό

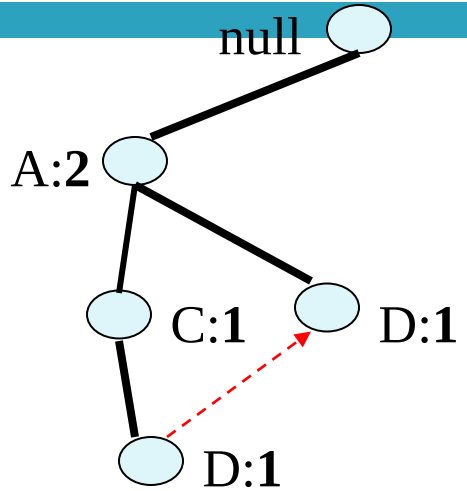


Αλγόριθμος FP-Growth

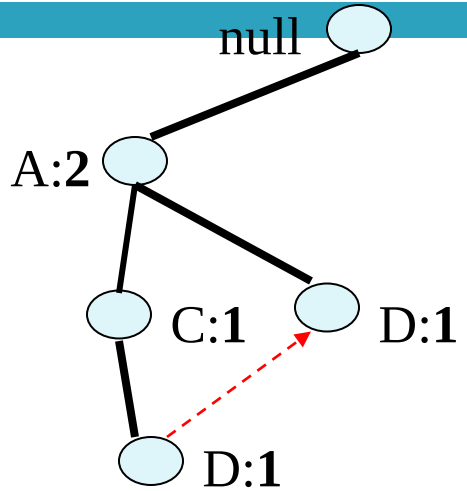
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {D, E}

1. Αλλαγή υποστήριξης

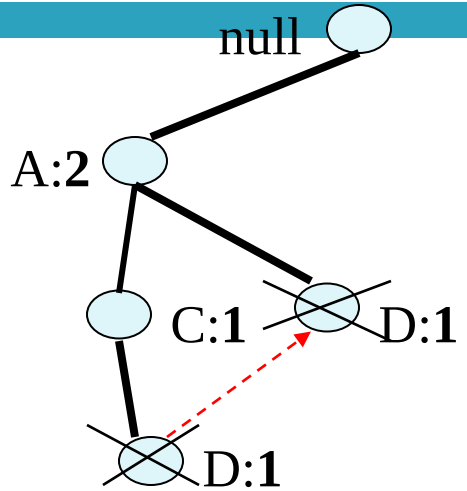
2. Περικοπές κόμβων



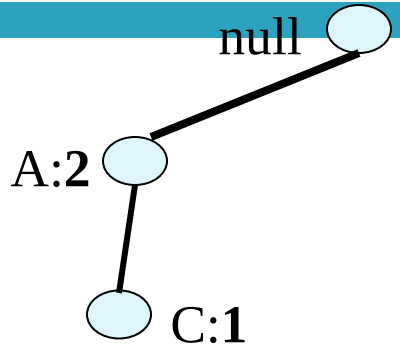
1. Αλλαγή υποστήριξης



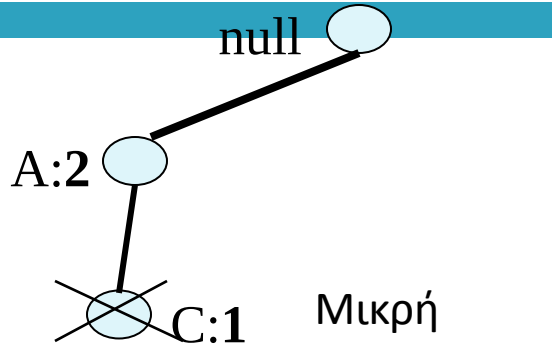
2. Περικοπές κόμβων



2. Περικοπές κόμβων

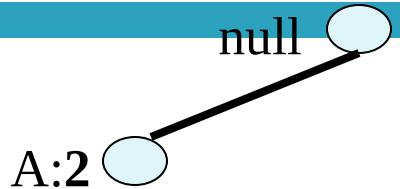


2. Περικοπές κόμβων



Αλγόριθμος FP-Growth

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



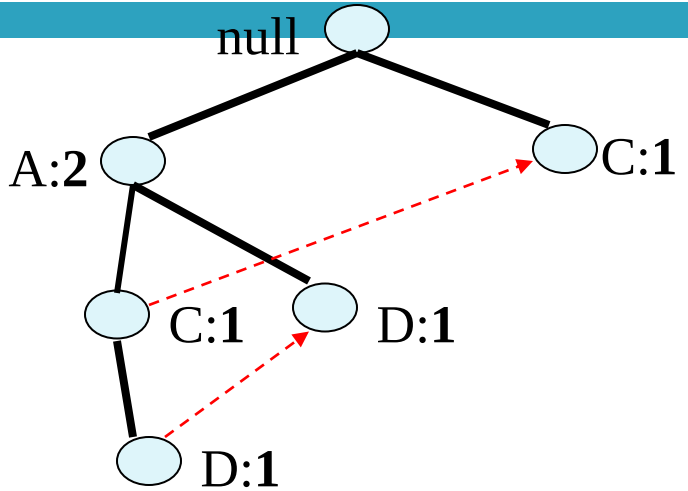
Υποστήριξη του A είναι ≥ minsup -> {A, D, E} συχνό

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}

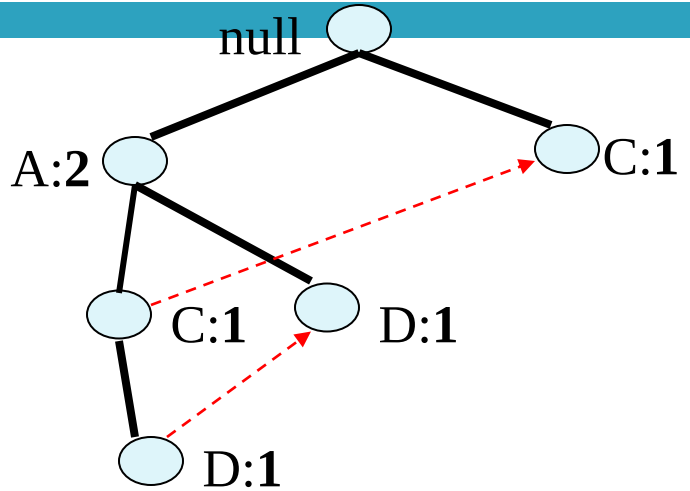


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)

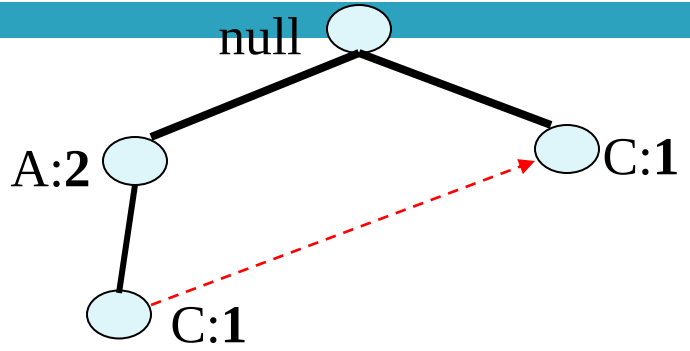


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



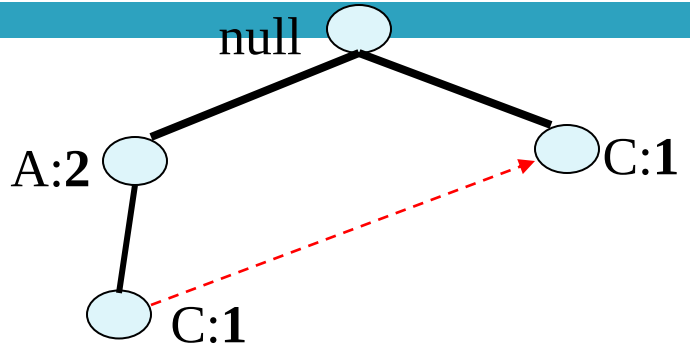
Αλγόριθμος FP-Growth

Βρες την υποστήριξη του {C, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

Οπότε {C, E} συχνό

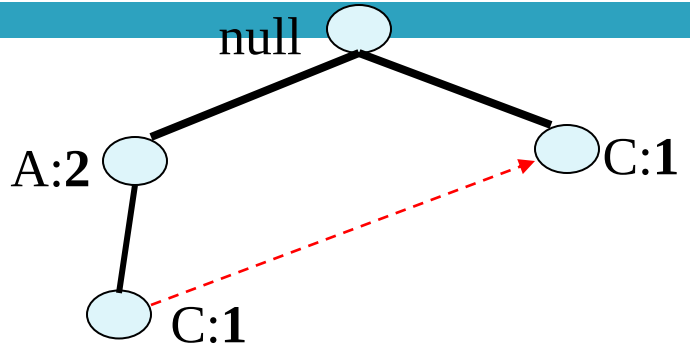


Αλγόριθμος FP-Growth

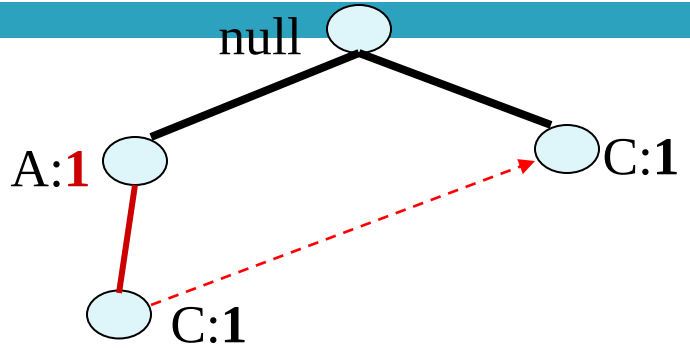
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {C, E}

1. Αλλαγή υποστήριξης

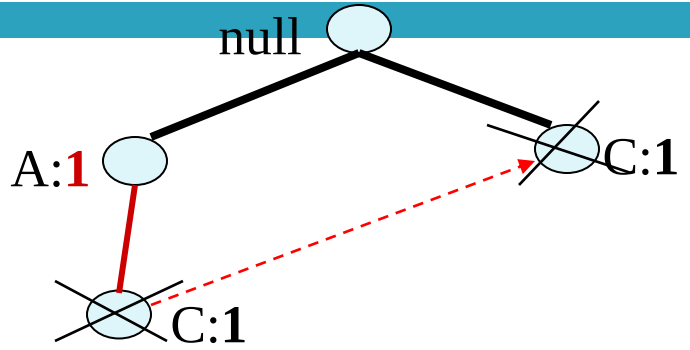
2. Περικοπές κόμβων



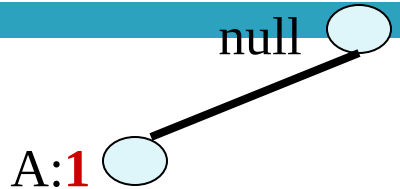
1. Αλλαγή υποστήριξης



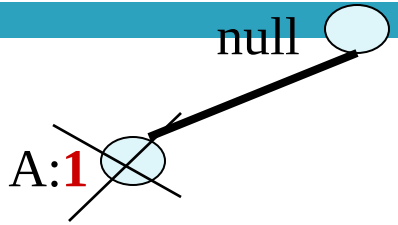
2. Περικοπή Κόμβων



2. Περικοπή Κόμβων



2. Περικοπή Κόμβων



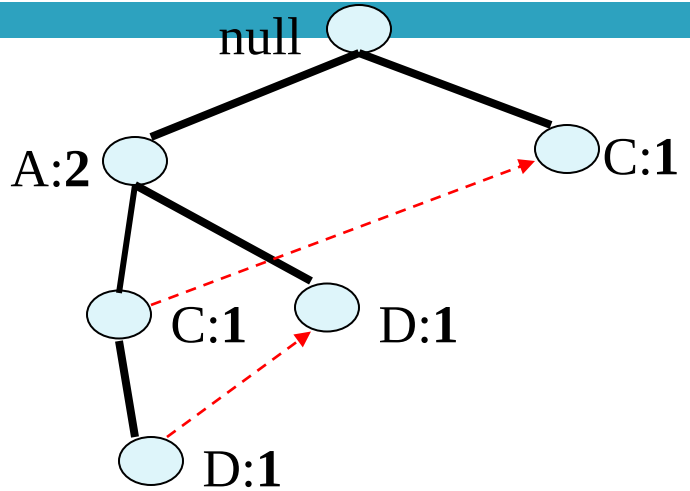
2. Περικοπή Κόμβων

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}**

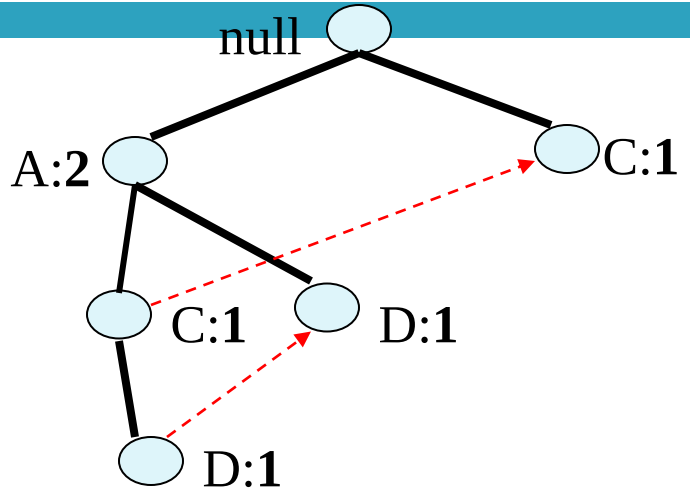


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)

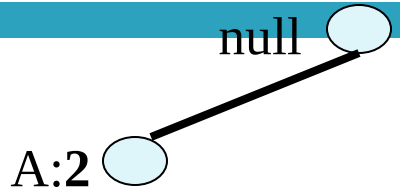


Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)

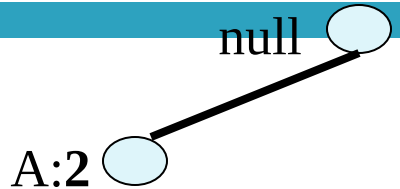


Αλγόριθμος FP-Growth

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

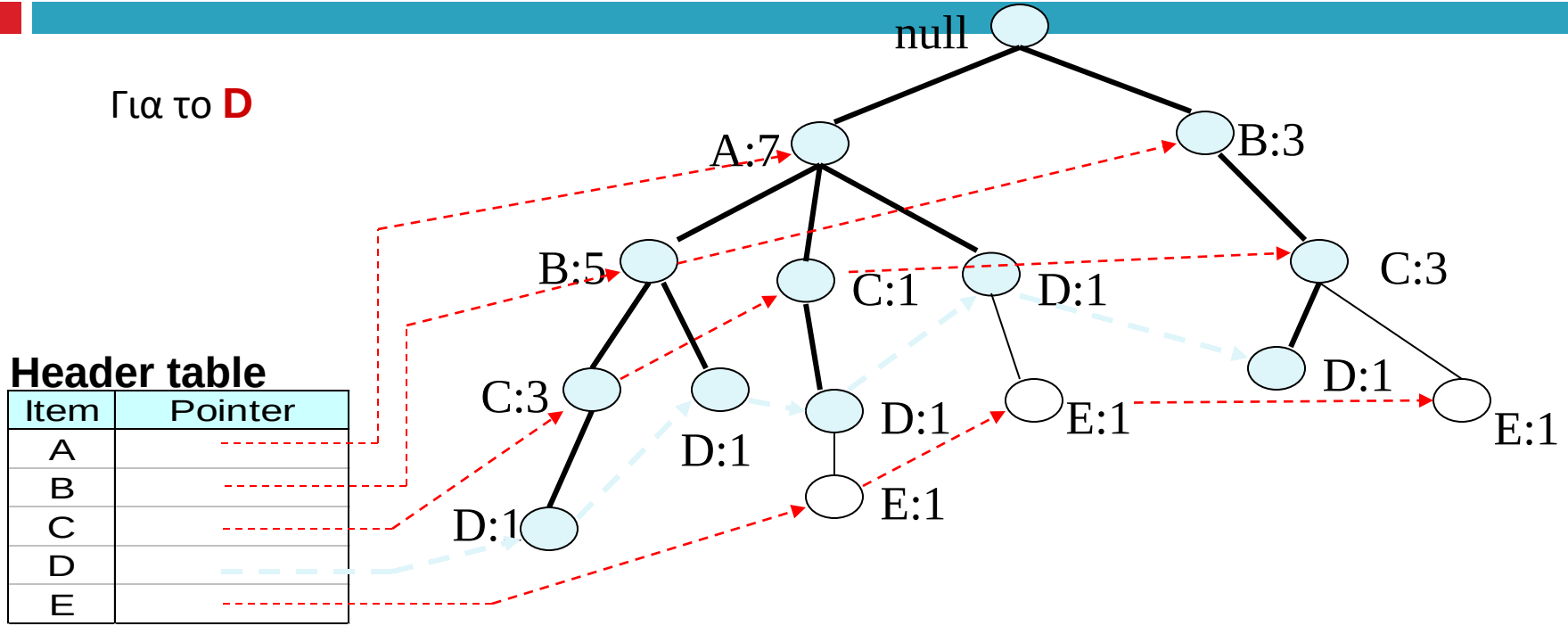
Έχουμε τα εξής συχνά στοιχειοσύνολα

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

Συνεχίζουμε για το D

Αλγόριθμος FP-Growth

Για το **D**



Header table

Item	Pointer
A	
B	
C	
D	
E	

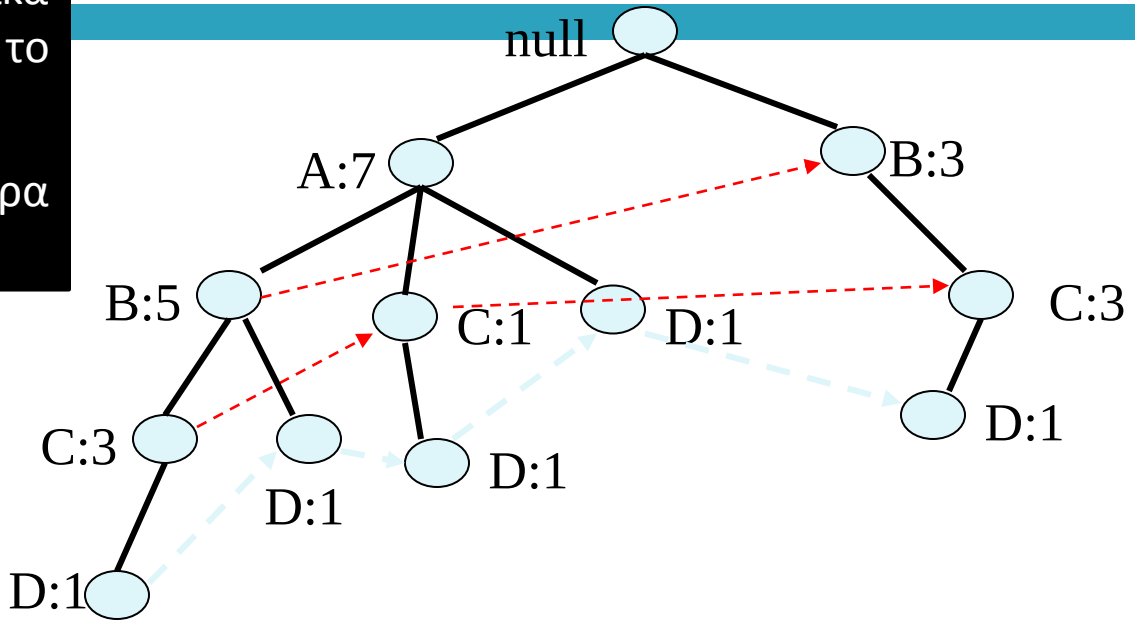
Αλγόριθμος FP-Growth

Φάση 1

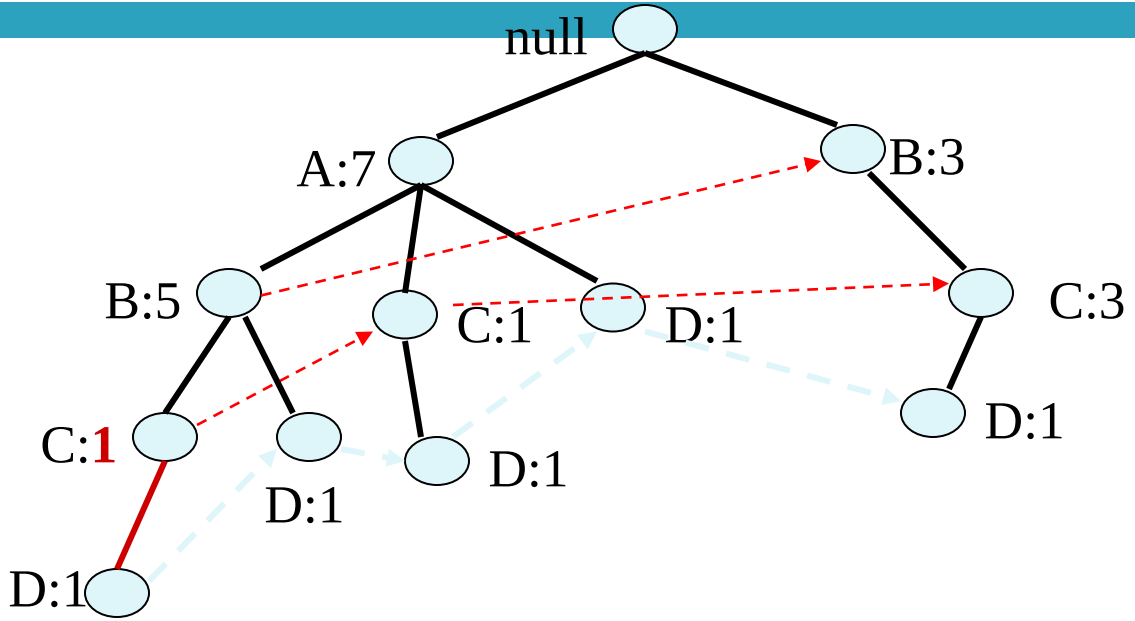
Όλα τα προθεματικά μονοπάτια που περιέχουν το D

Υποστήριξη 5>2 -> άρα συχνό

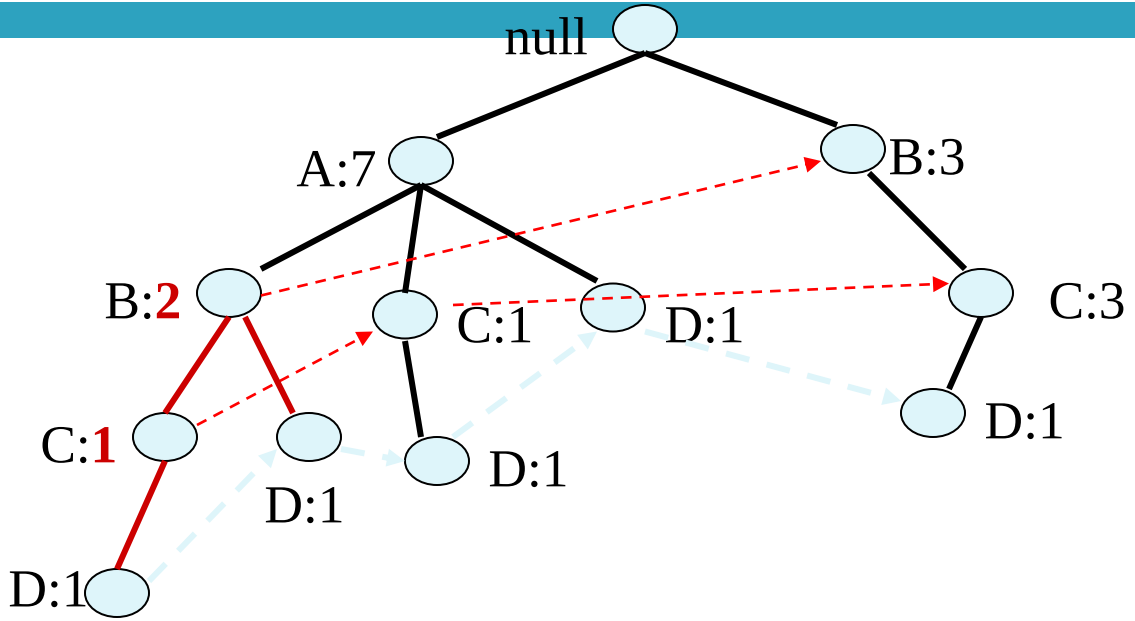
Μετατροπή του προθεματικού δέντρου σε FP-δέντρο υπό συνθήκη



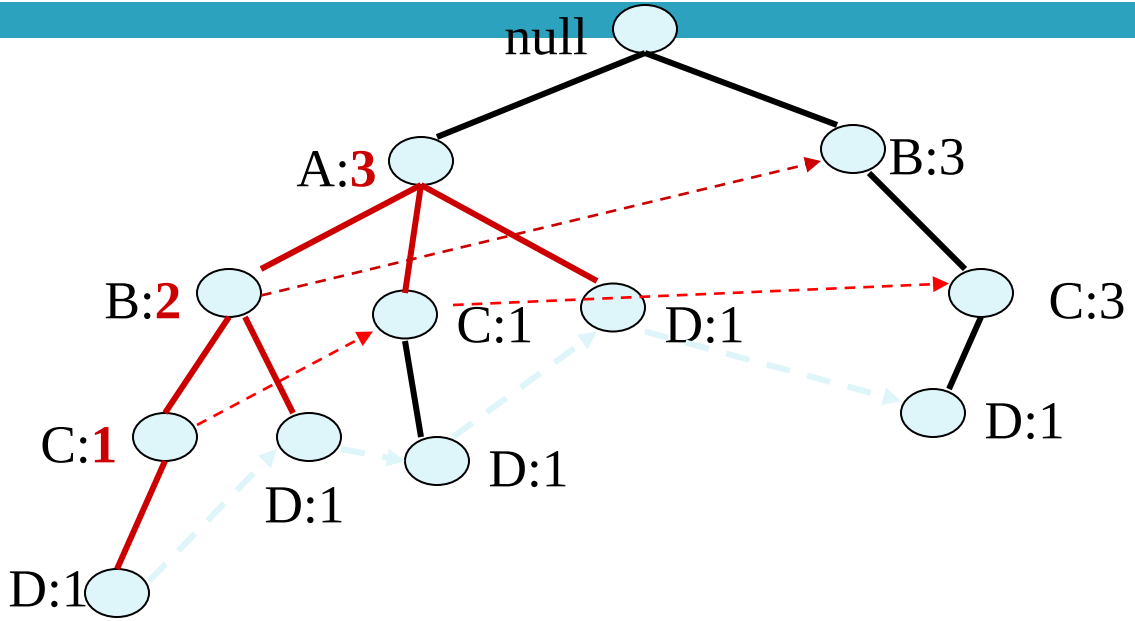
1. Αλλαγή υποστήριξης



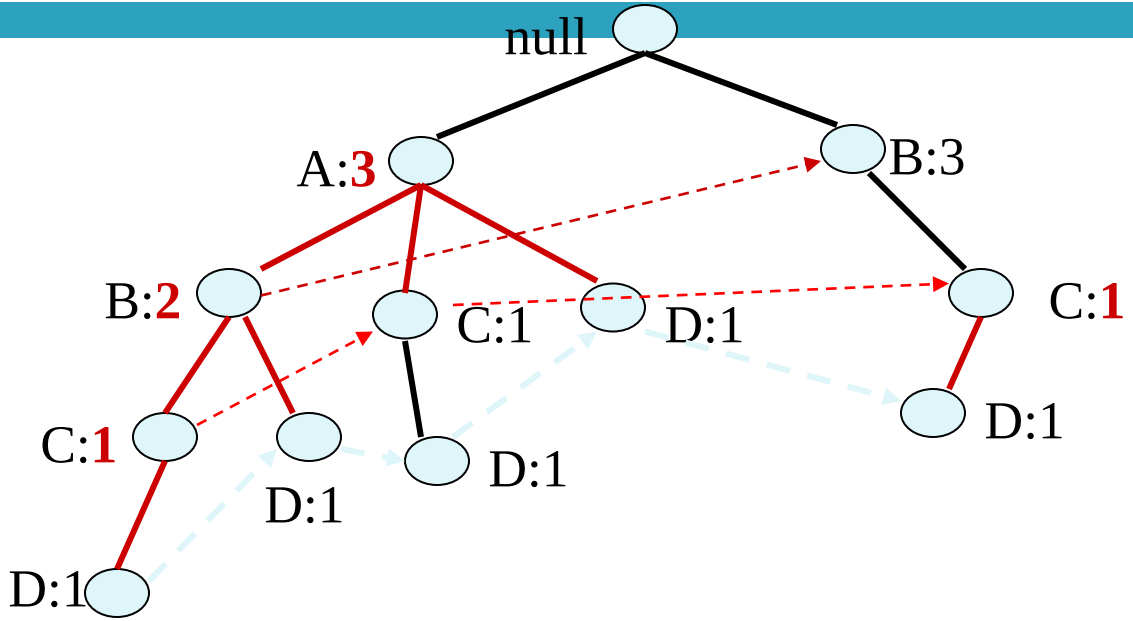
1. Αλλαγή υποστήριξης



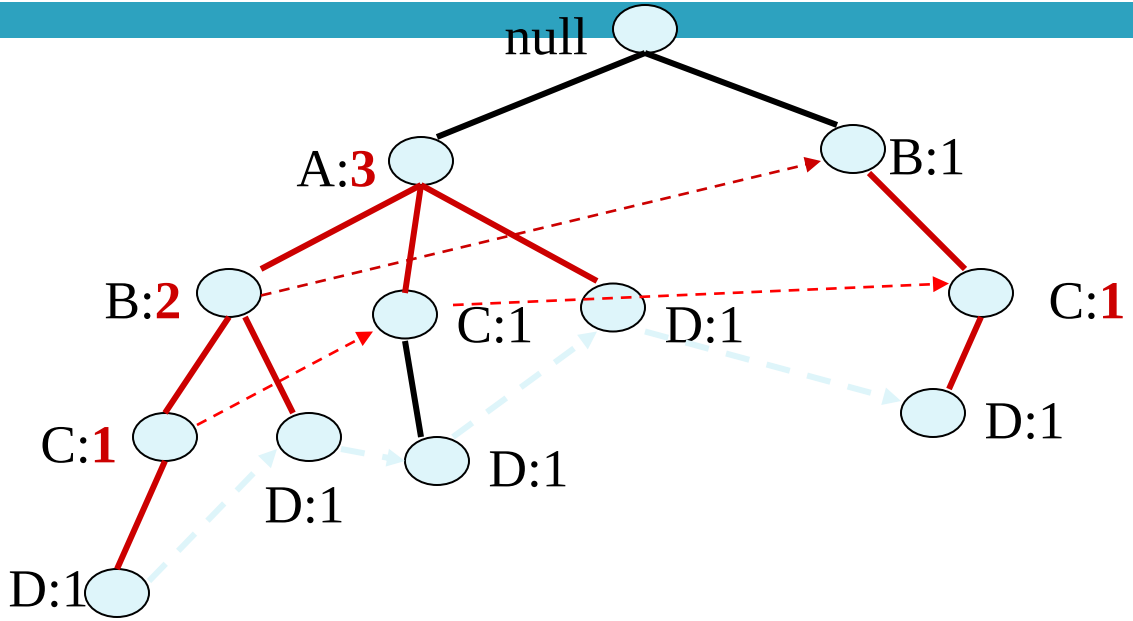
1. Αλλαγή υποστήριξης



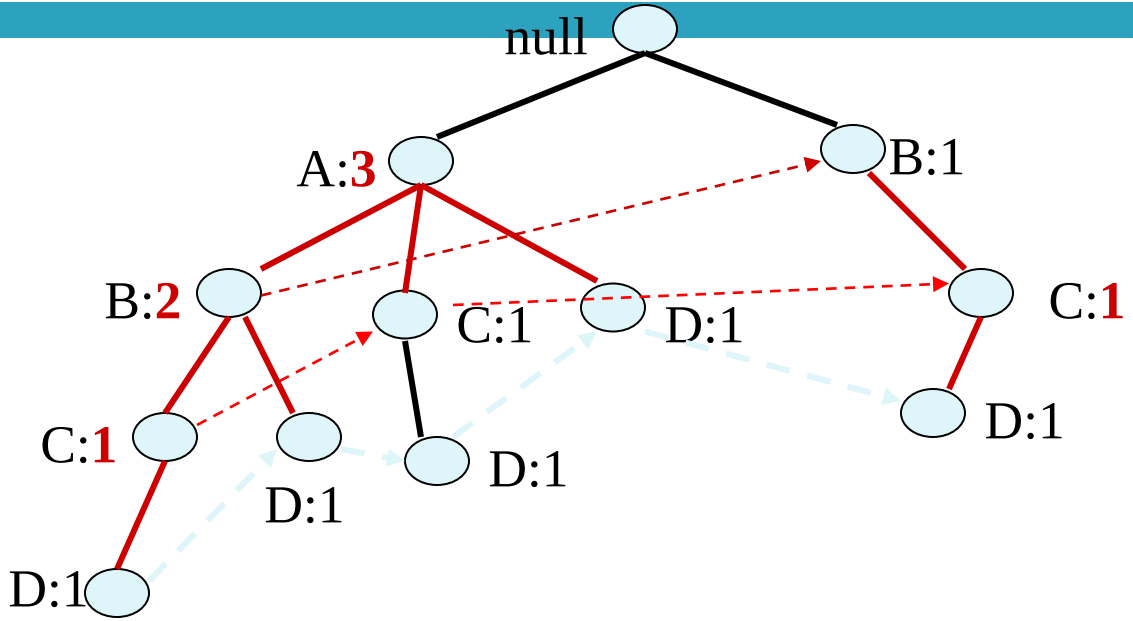
1. Αλλαγή υποστήριξης



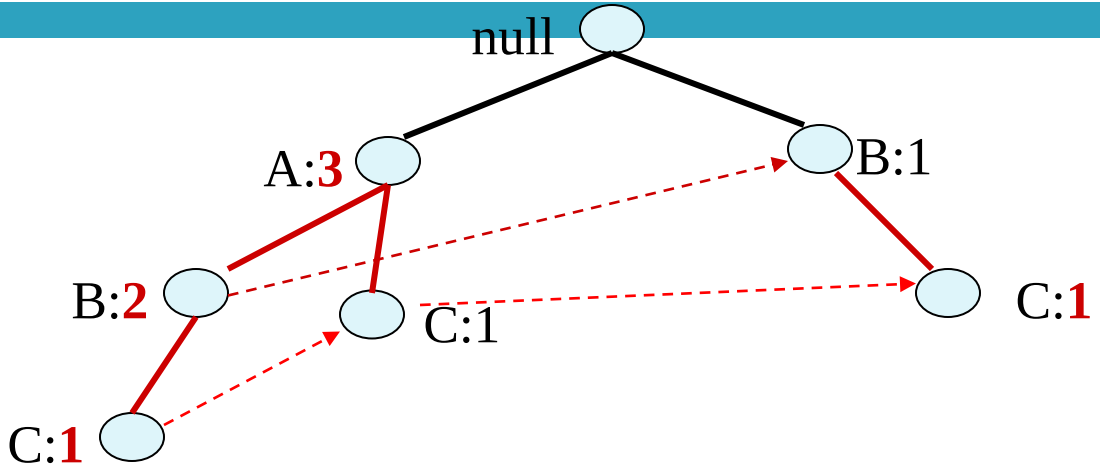
1. Αλλαγή υποστήριξης



2. Περικοπή Κόμβων



2. Περικοπή Κόμβων

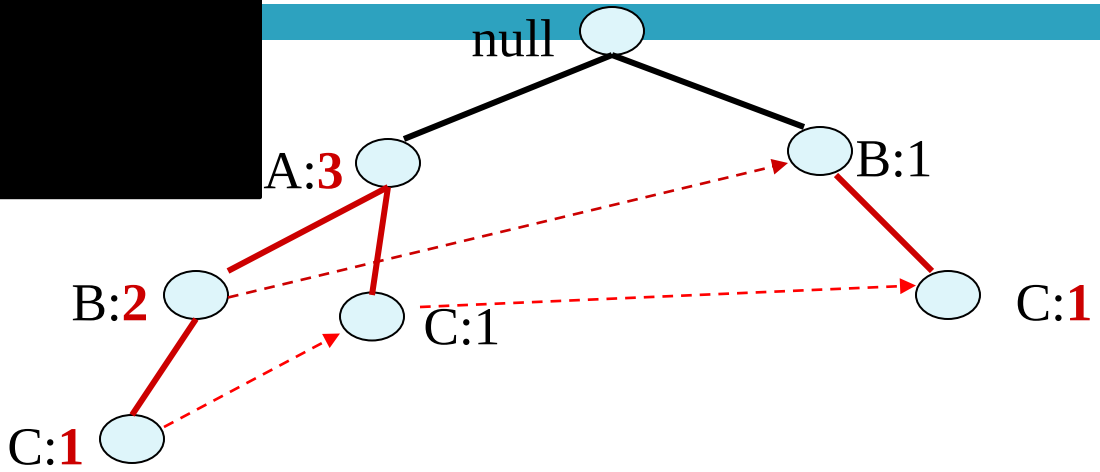


Αλγόριθμος FP-Growth

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

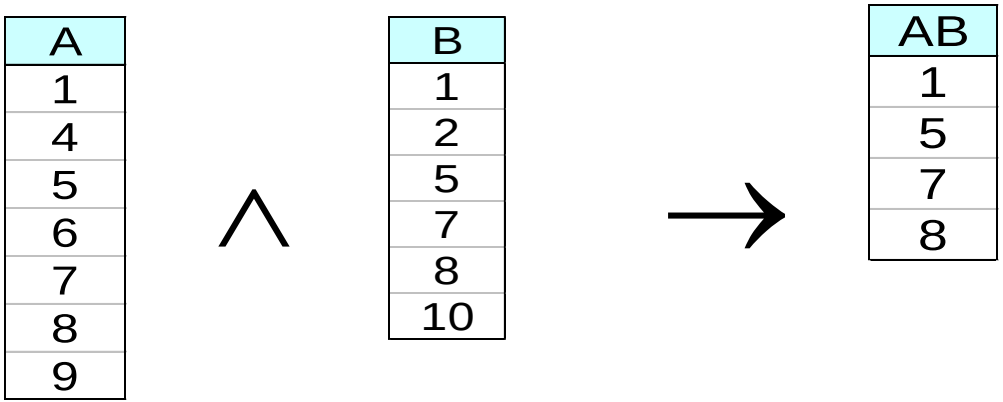
Vertical Data Layout

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 K-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}:

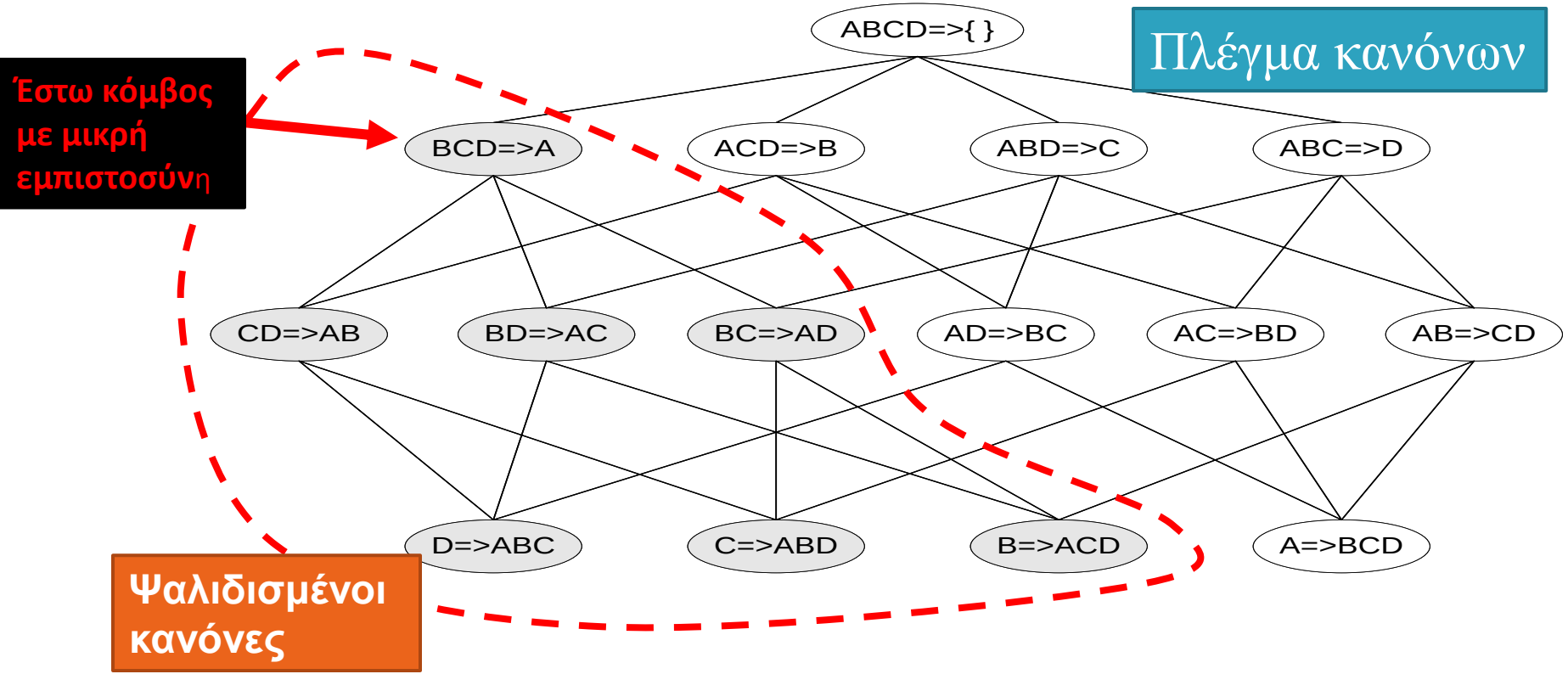
$$c(ABC \rightarrow \mathbf{D}) \geq c(AB \rightarrow C\mathbf{D}) \geq c(A \rightarrow BC\mathbf{D})$$

- 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:

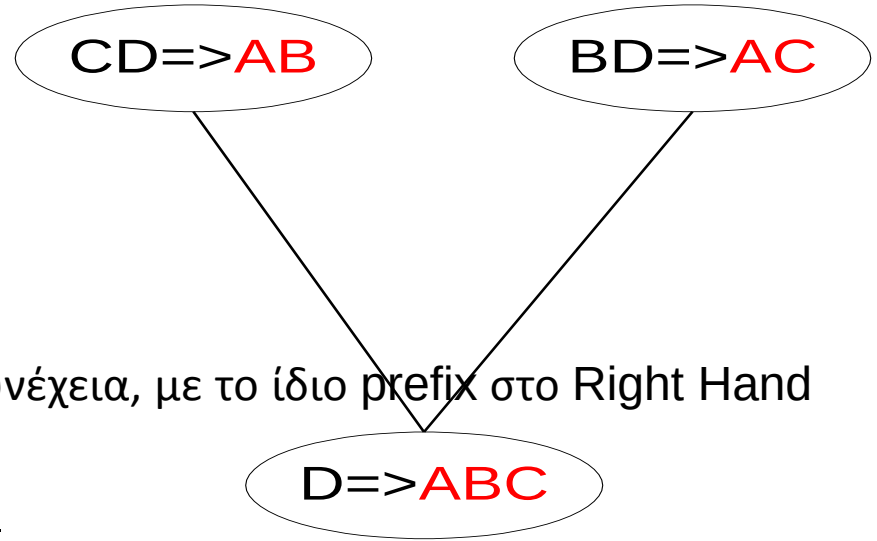
  Αν ο κανόνας $X \rightarrow X - Y$ δεν ικανοποιεί το κατώφλι εμπιστοσύνης, τότε και ο κανόνας $X' \rightarrow X' - Y'$ ($X' \subseteq X$) δεν τον ικανοποιεί

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