Εξόρυξη Δεδομένων

9: Κανόνες συσχέτισης (ΙΙ)

Επανάληψη

Market-Basket transactions (καλάθι αγορών

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 (transaction)

στοιχείο (item)

- Το πρόβλημα: Δεδομένου ενός συνόλου δοσοληψιών (transactions), βρες κανόνες που προβλέπουν την εμφάνιση στοιχείων (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)

Ορισμοί

- στοιχειοσύνολο (itemset): Ένα υποσύνολο του συνόλου των στοιχείων
- k-στοιχειοσύνολο (k-itemset): ένα στοιχειοσύνολο με k στοιχεία
- support count (σ) ενός στοιχειοσυνόλου: ο αριθμός εμφανίσεων του στοιχείου
- Υποστήριξη (Support (s)) ενός στοιχειοσυνόλου Το ποσοστό των δοσοληψιών που περιέχουν ένα στοιχειοσύνολο
- Συχνό Στοιχειοσύνολο (Frequent Itemset) Ένα στοιχειοσύνολο του οποίου η υποστήριξη είναι μεγαλύτερη ή ίση από κάποια τιμή κατωφλίου minsup

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

<u>Κανόνας Συσχέτισης (Association Rule)</u>

Είναι μια έκφραση της μορφής $X \rightarrow Y$, όπου X και Y είναι στοιχειοσύνολα

$X \subseteq I$, $Y \subseteq I$, $X \cap Y = \emptyset$

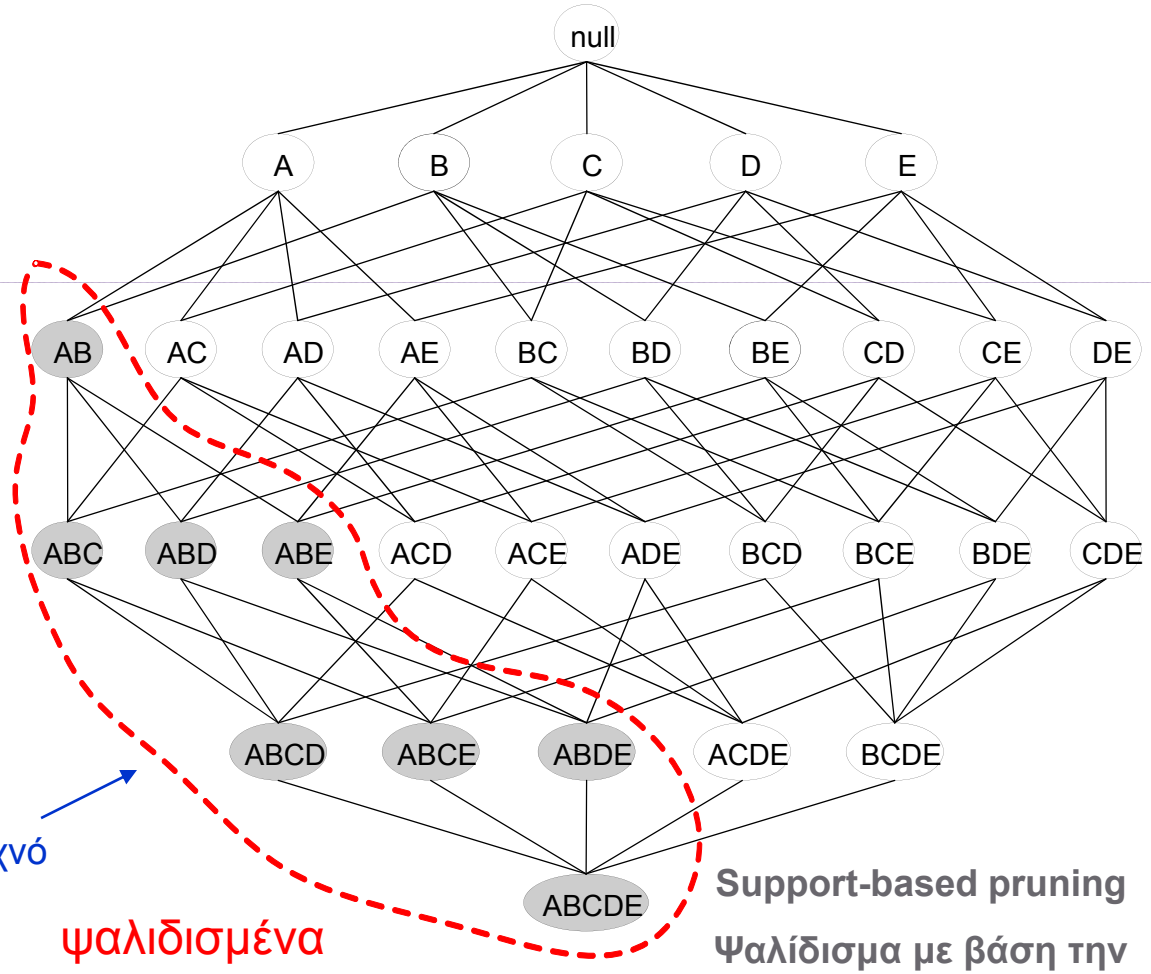
- Υποστήριξη Κανόνα Support (s)
  - Το ποσοστό των δοσοληψιών που περιέχουν και το X και το Y (X →Y)
- Εμπιστοσύνη - Confidence (c)
  - Πόσες από τις δοσοληψίες (ποσοστό) που περιέχουν το X περιέχουν και το Y
- Χωρισμός του προβλήματος σε δύο υπο-προβλήματα:
  - Εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation) (υποστήριξη >= minsup)
  - Δημιουργία Κανόνων (Rule Generation)

Εύρεση Συχνών Στοιχειοσυνόλων: Στρατηγική apriori

Αρχή Apriori

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

Ή ισοδύναμα αν ένα στοιχειοσύνολο είναι μη συχνό, όλα τα υπερσύνολα του είναι μη συχνά



βρέθηκε μη συχνό

ψαλιδισμένα υπερσύνολα

Support-based pruning

Ψαλίδισμα με βάση την υποστήριξη

Γενικός Αλγόριθμος για την Εύρεση Συχνών Στοιχειοσυνόλων

Έστω **k** = 1 *#k: μήκος στοιχειοσυνόλου*

Παρήγαγε τα συχνά **1-στοιχειοσύνολα**

Repeat until να μην παράγονται νέα συχνά στοιχειοσύνολα

1. Παρήγαγε υποψήφια **(k+1)**-στοιχειοσύνολα
2. Ψαλίδισε τα υποψήφια στοιχειοσύνολα που περιέχουν μη συχνά στοιχειοσύνολα μεγέθους k
3. Υπολόγισε την υποστήριξη κάθε υποψήφιου (k+1)-στοιχειοσυνόλου διασχίζοντας τη βάση των δοσοληψιών
4. Σβήσε τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά
5. k = k + 1

Παραγωγή Κανόνων

Δοθέντος ενός συχνού στοιχειοσυνόλου L,
βρες όλα τα μη κενά υποσύνολα $f \subset L$ τέτοια ώστε:
ο κανόνας $f \rightarrow L - f$ να ικανοποιεί τον περιορισμό της ελάχιστης εμπιστοσύνης

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

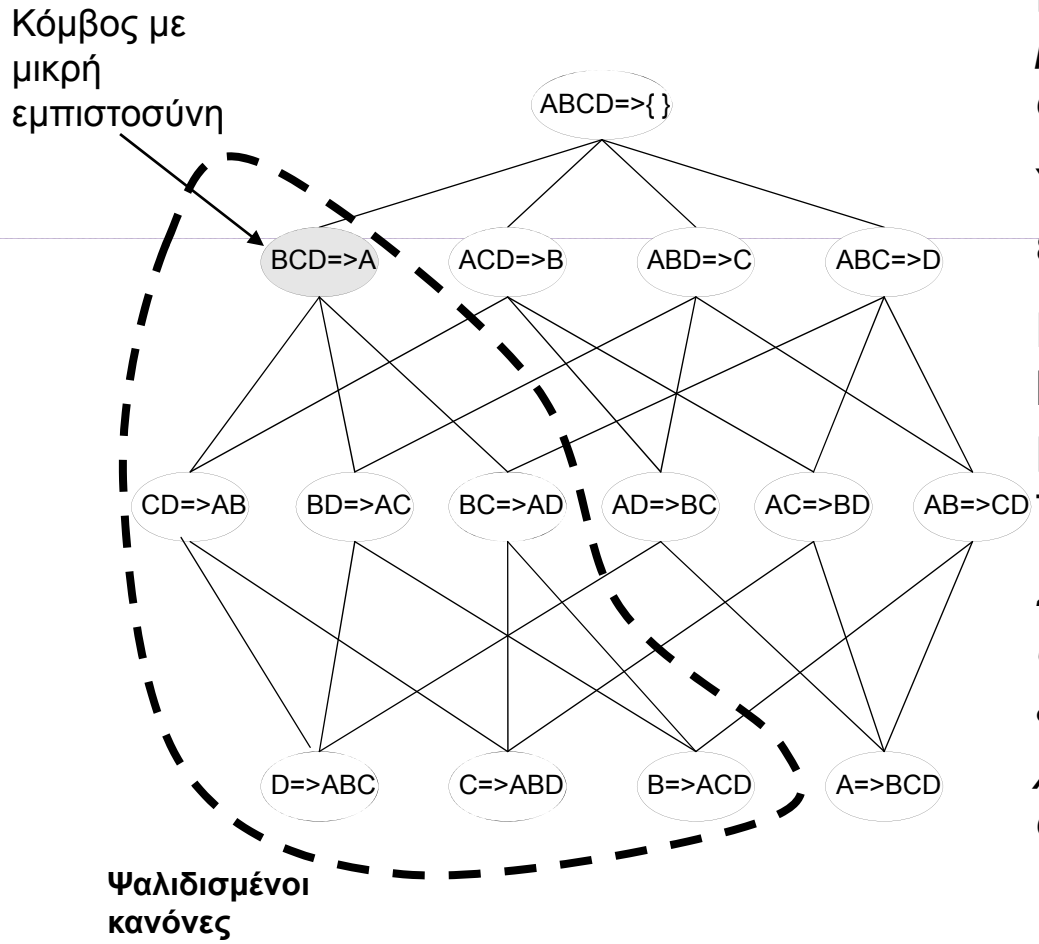
Για παράδειγμα L = {A,B,C,D}:

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

Η εμπιστοσύνη είναι αντι-μονότονη σε σχέση με τον αριθμό των στοιχείων στο RHS του κανόνα

Πλέγμα Κανόνων

Ψαλίδισμα με βάση την εμπιστοσύνη



Για κάθε συχνό στοιχειοσύνολο, ξεκινάμε με έναν κανόνα που έχει *μόνο k = **1** στοιχείο στο δεξί μέρος του*

Υπολογίζουμε την εμπιστοσύνη

Παράγουμε κανόνες με **k+1** στοιχεία στο δεξί μέρος και υπολογίζουμε την εμπιστοσύνη τους

Σημείωση: Για τον υπολογισμό της εμπιστοσύνης δεν χρειάζεται να διαπεράσουμε τη βάση

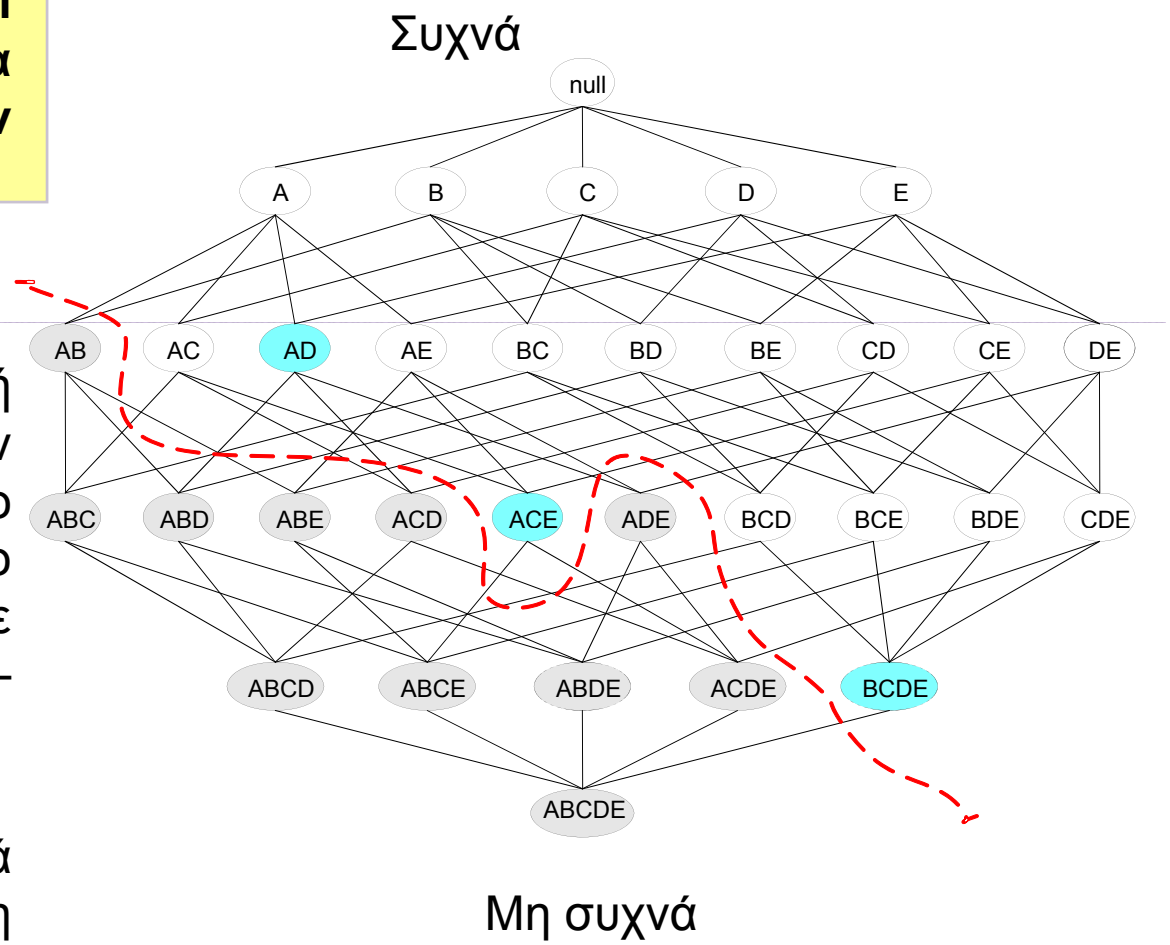
Maximal frequent itemsets

Ένα itemset είναι maximal συχνό αν κανένα από τα άμεσα υπερσύνολά του δεν είναι συχνό

δηλαδή είναι <u>όλα μη συχνά</u>

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

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



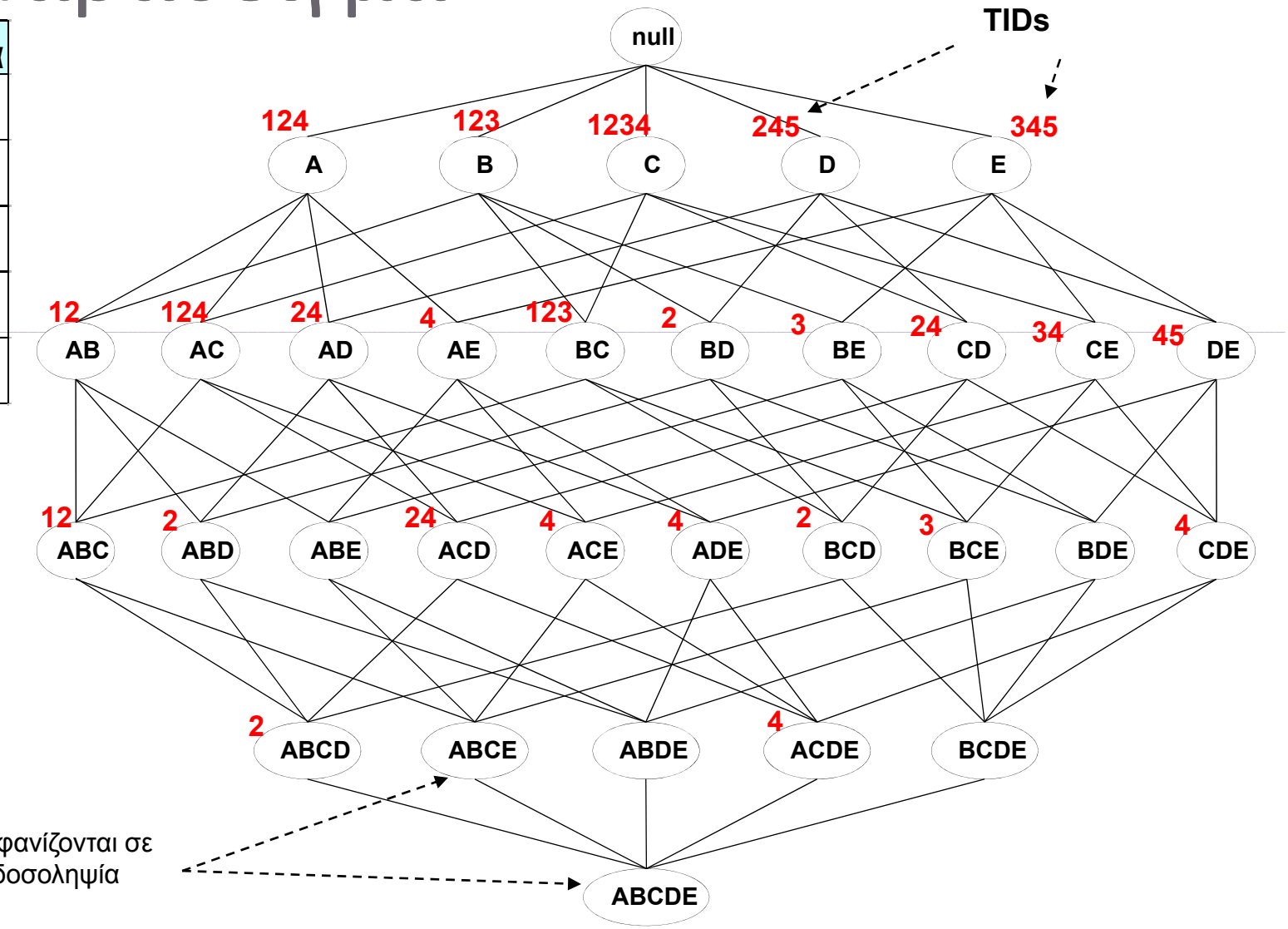
Closed frequent itemsets

- Ένα στοιχειοσύνολο είναι **κλειστό (closed)** αν κανένα από τα άμεσα υπερσύνολα του δεν έχει την ίδια υποστήριξη με αυτό (δηλαδή, έχει μικρότερη υποστήριξη)
- Ένα στοιχειοσύνολο είναι **κλειστό *συχνό* στοιχειοσύνολο** αν είναι κλειστό και συχνό (δηλαδή, η υποστήριξη του είναι μεγαλύτερη ή ίση με minsup)
- Πάλι τα υποσύνολά τους μας δίνουν όλα τα συχνά υποσύνολα, τώρα όμως μπορούμε να υπολογίσουμε την υποστήριξη των υποσυνόλων τους
- *Πως: Η υποστήριξη ενός μη κλειστού στοιχειοσυνόλου πρέπει να είναι ίση με την μεγαλύτερη υποστήριξη ανάμεσα στα υπερσύνολά του*

Παράδειγμα

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

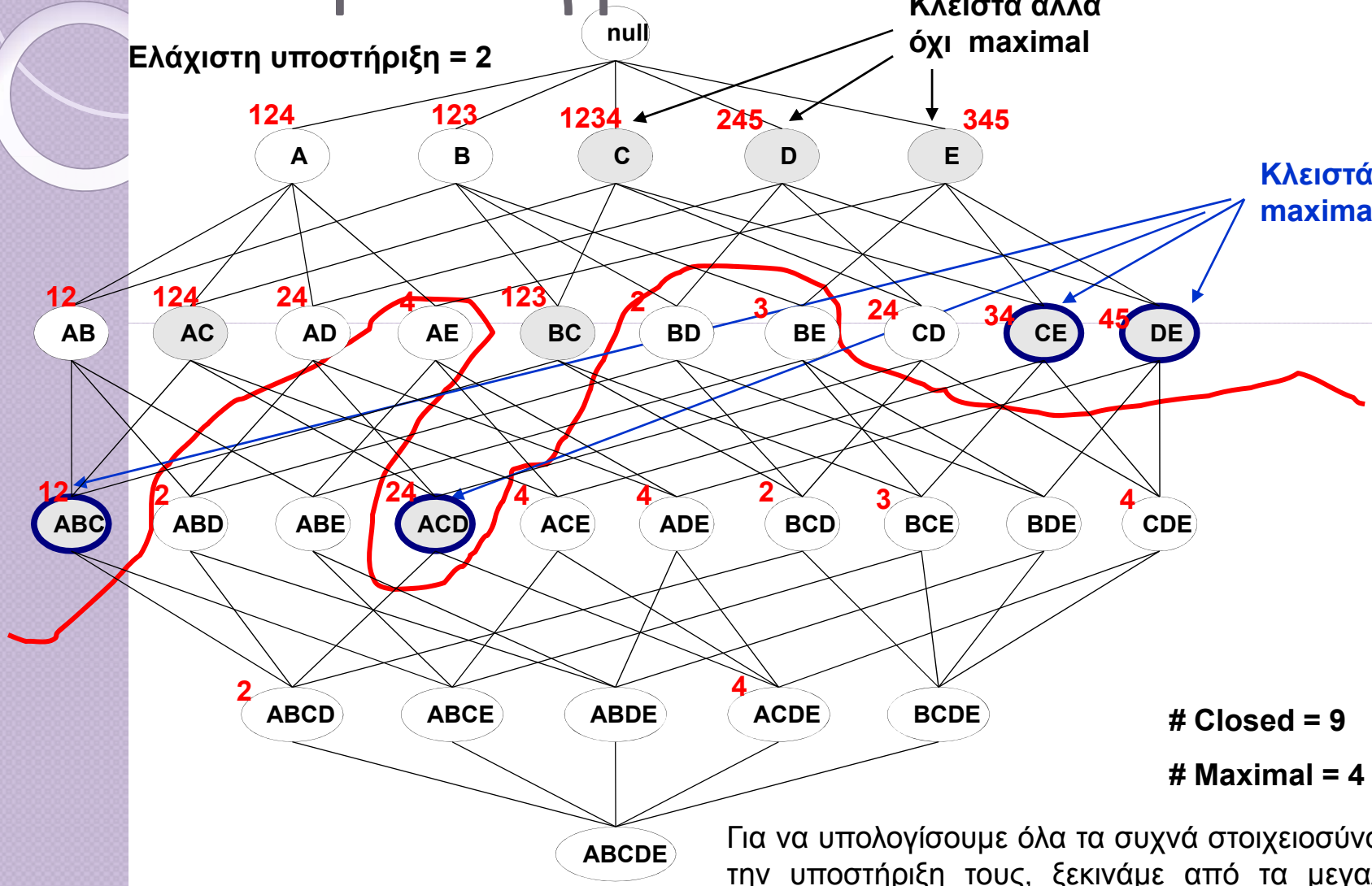
Maximal vs Closed Itemsets



Δεν εμφανίζονται σε καμιά δοσοληψία

Παράδειγμα

Ελάχιστη υποστήριξη = 2



Κλειστά αλλά όχι maximal

Κλειστά και maximal

 # Closed = 9

 # Maximal = 4

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

Άλλες μέθοδοι για frequent itemsets

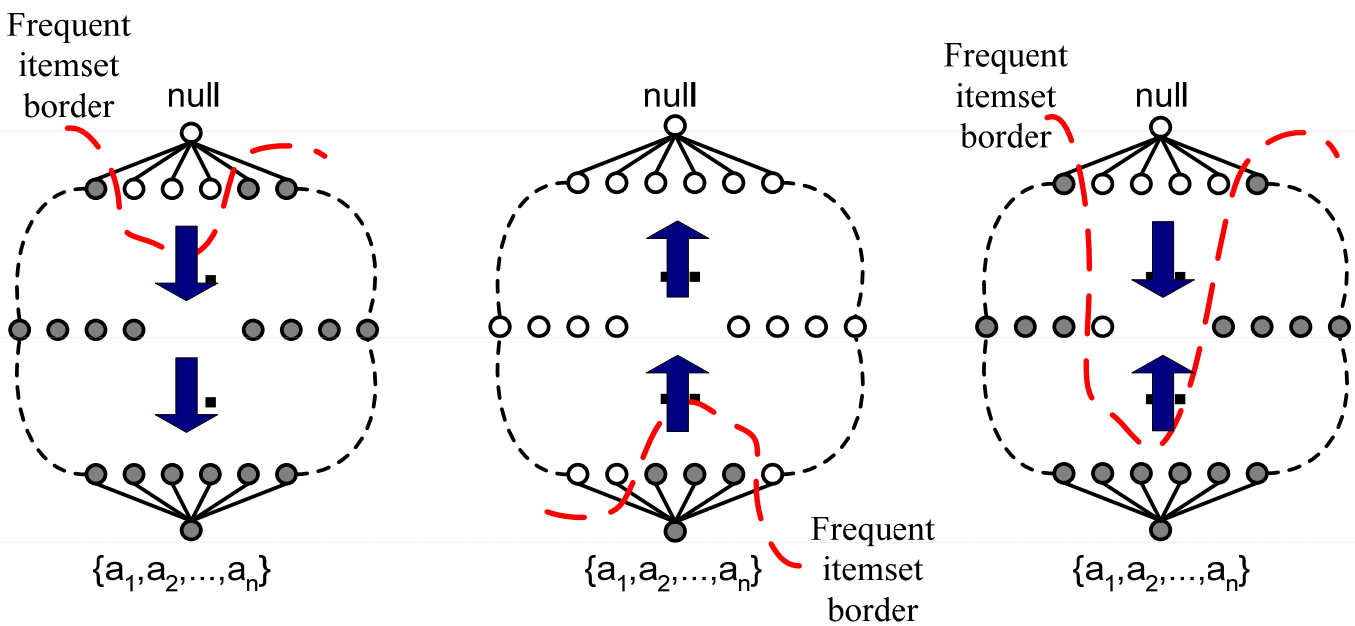
- Ο Apriori
  - Συχνά μεγάλο I/O επειδή κάνει πολλαπλά περάσματα στη βάση των δοσοληψιών
  - Κακή απόδοση όταν οι δοσοληψίες έχουν μεγάλο πλάτος
- Άλλες μέθοδοι
  - Διαφορετικές διασχίσεις του πλέγματος των στοιχειοσυνόλων
  - Αναπαράσταση Συνόλου Δοσοληψιών

Διάσχιση του Πλέγματος των Στοιχειοσυνόλων

k -> k – 1 (συγκεκριμένο-προς-γενικό)

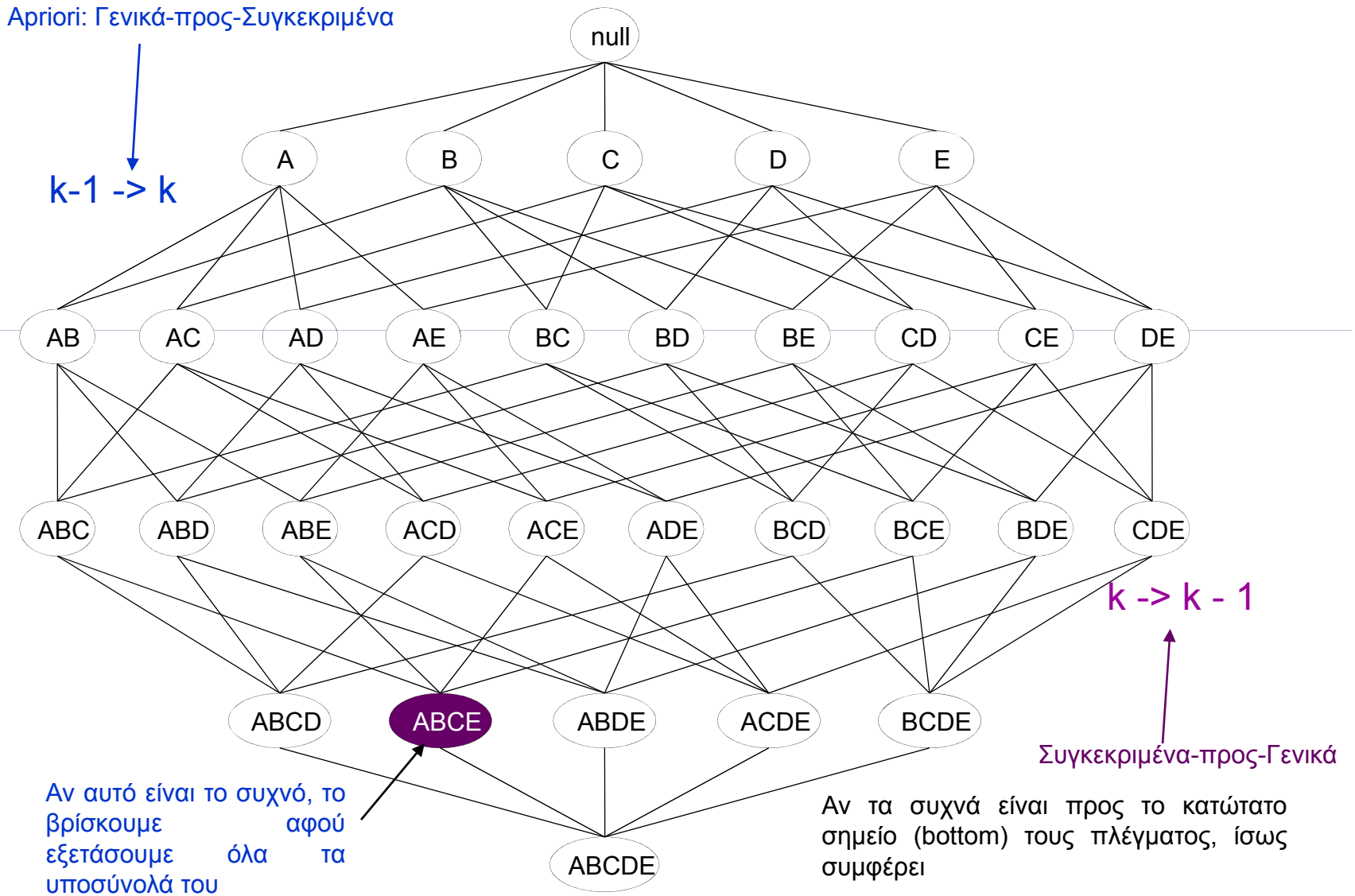
Πιο χρήσιμο για τον εντοπισμό maximal συχνών στοιχειοσυνόλων σε πυκνές (δηλ, με μεγάλο πλάτος) δοσοληψίες όπου το συχνό στοιχειοσύνολο βρίσκεται κοντά στο κατώτατο σημείο του πλέγματος

Αν συχνό, δε χρειάζεται να ελέγξουμε κανένα από τα υποσύνολά του



(a) General-to-specific (b) Specific-to-general (c) Bidirectional

Παράδειγμα

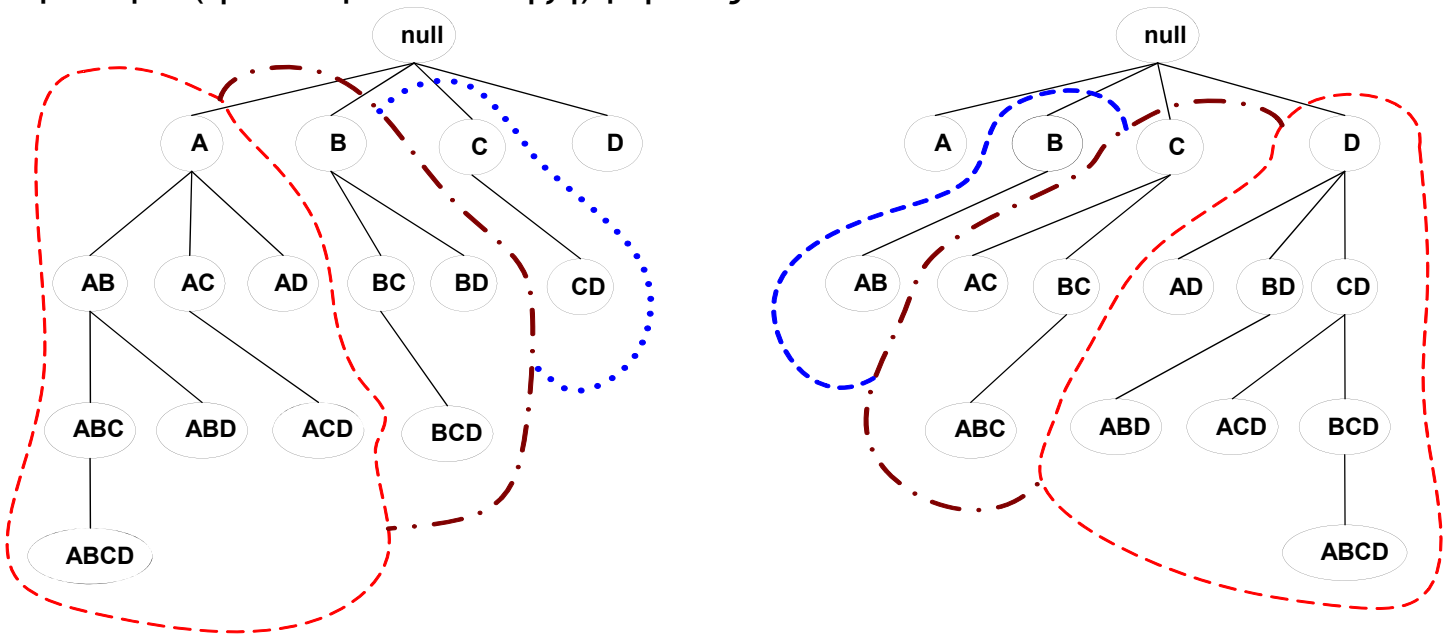

Apriori: Γενικά-προς-Συγκεκριμένα
k-1 -> k
null
A
B
C
D
E
AB
AC
AD
AE
BC
BD
BE
CD
CE
DE
ABC
ABD
ABE
ACD
ACE
ADE
BCD
BCE
BDE
CDE
ABCD
ABCE
ABDE
ACDE
BCDE
ABCDE
k -> k - 1
Συγκεκριμένα-προς-Γενικά
Αν αυτό είναι το συχνό, το βρίσκουμε αφού εξετάσουμε όλα τα υποσύνολά του
Αν τα συχνά είναι προς το κατώτατο σημείο (bottom) τους πλέγματος, ίσως συμφέρει

Κλάσεις Ισοδυναμίας

Χωρισμός των στοιχειοσυνόλων του πλέγματος σε ξένες μεταξύ τους ομάδες (κλάσεις ισοδυναμίας) και εξέταση των στοιχειοσυνόλων ανά κλάσεις

Apriori: ορίζει τις κλάσεις με βάση το μήκος k των στοιχειοσυνόλων, πρώτα αυτά μήκους 1, μετά μήκους 2 κοκ

Prefix (Suffix): Δύο στοιχειοσύνολα ανήκουν στην ίδια κλάση αν έχουν κοινό πρόθεμα (ή επίθεμα-κατάληξη) μήκους k

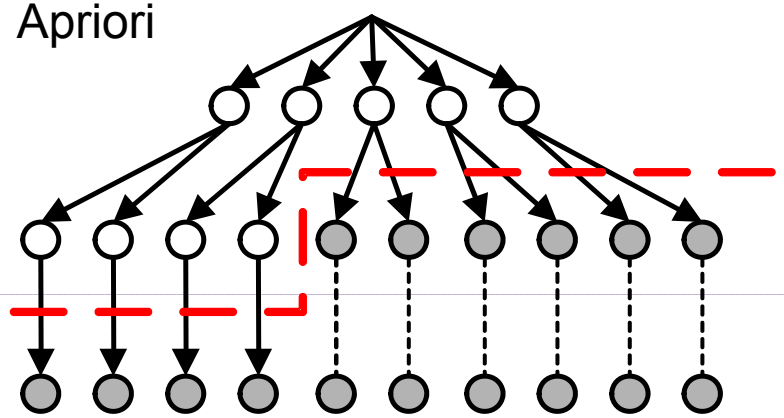


(a) Prefix tree

(b) Suffix tree

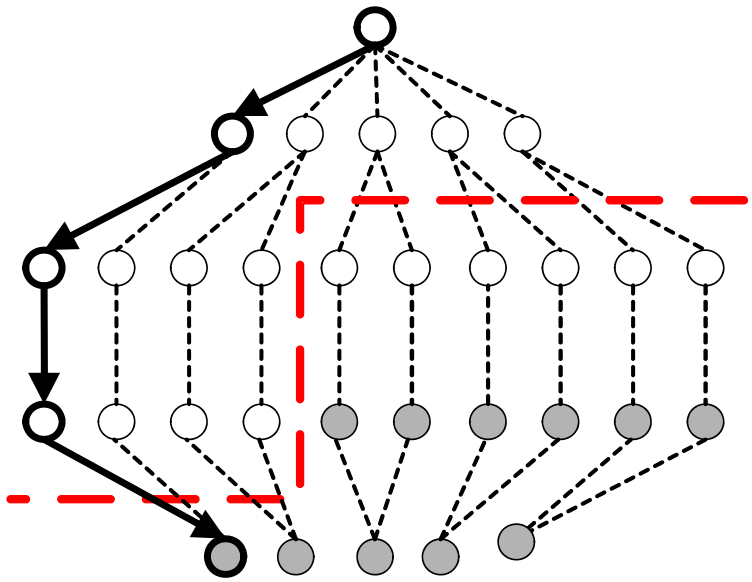
BFS vs DFS

Apriori



BFS: Breadth-First-Search
Διάσχιση κατά Πλάτος

DFS: Depth-First-Search
Διάσχιση κατά Βάθος



Χρήσιμο για την εύρεση maximal συχνών στοιχειοσυνόλων γιατί τα εντοπίζει πιο γρήγορα από το BFS

Μόλις εντοπιστεί το maximal, είναι δυνατόν να κλαδευτούν πολλά υποσύνολα του

Αναπαράσταση της Βάσης Δεδομένων των Δοσοληψιών: Οριζόντια vs Κάθετη

Αυτό χρησιμοποιεί ο apriori

Εναλλακτικά:

Για κάθε στοιχείο σε ποιες δοσοληψίες εμφανίζεται

Οριζόντια Διάρθρωση 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 Διάρθρωση 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				

Η υποστήριξη υπολογίζεται παίρνοντας τις τομές των TID-λιστών

Η υποστήριξη υπολογίζεται παίρνοντας τις τομές των TID-λιστών

A
1
4
5
6
7
8
9

$\wedge$

B
1
2
5
7
8
10

$\rightarrow$

AB
1
5
7
8

- Η υποστήριξη ενός k-στοιχειοσυνόλου υπολογίζεται παίρνοντας τις τομές των TID-λιστών δύο από τα (k-1)-ύπο-στοιχειοσύνολα του.

- Πλεονέκτημα: πολύ γρήγορος υπολογισμός της υποστήριξης
- Πρόβλημα, αν οι TID-λίστες είναι μεγάλες και δε χωρούν στη μνήμη

Ο Αλγόριθμος FP-Growth

Συνοπτικά

- Ο αλγόριθμος χρησιμοποιεί μια συμπιεσμένη αναπαράσταση της βάσης των δοσοληψιών με τη μορφή ενός FP-δέντρου
  - Το δέντρο μοιάζει με προθεματικό δέντρο - prefix tree (trie)
  - Ο αλγόριθμος κατασκευής διαβάζει μια δοσοληψία τη φορά, απεικονίζει τη δοσοληψία σε ένα μονοπάτι του FP-δέντρου
  - Μερικά μονοπάτια μπορεί να επικαλύπτονται: όσο περισσότερα μονοπάτια επικαλύπτονται, τόσο καλύτερη συμπίεση
- Μόλις κατασκευαστεί το FP-δέντρο, ο αλγόριθμος χρησιμοποιεί μια αναδρομική διαίρει-και-βασίλευε (divide-and-conquer) προσέγγιση για την εξόρυξη των συχνών στοιχειοσυνόλων

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 FP-δέντρο είναι ένα προθεματικό δέντρο
- Επειδή έχουμε σύνολα, κάπως πρέπει να τα διατάξουμε ώστε να βρίσκουμε προθέματα
- Δηλαδή δε μπορεί το ένα σύνολο να είναι {A, B} και το άλλο {B, C, A} γιατί χάνουμε το κοινό πρόθεμα AB (ή BA)
- Άρα τα στοιχεία σε κάθε σύνολο πρέπει να ακολουθούν κάποια **διάταξη**, έστω τη *λεξικογραφική*

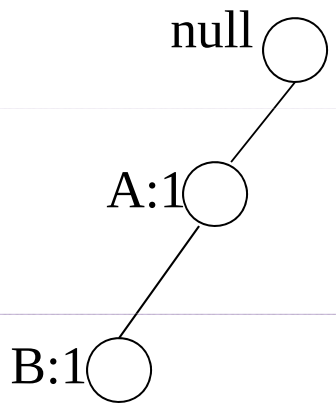
Αρχικά, το δέντρο κενό

null

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}

Διάβασμα TID=1:



Κάθε κόμβος έχει μια ετικέτα: ποιο στοιχείο και ποια η συχνότητα εμφάνισης (υποστήριξη) – πόσες δοσοληψίες φτάνουν σε αυτόν

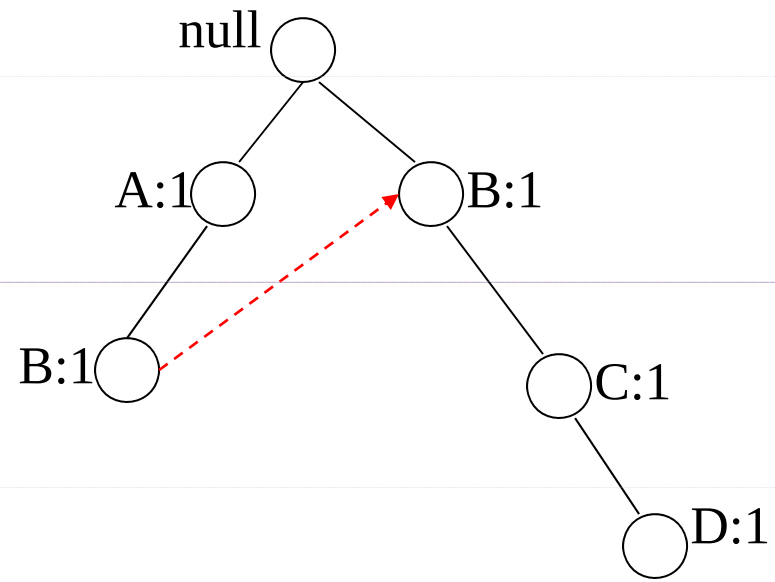
Ετικέτα κόμβου <**ΣΤΟΙΧΕΙΟ: ΥΠΟΣΤΗΡΙΞΗ**>

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}

Διάβασμα TID=1:

Διάβασμα TID=2:



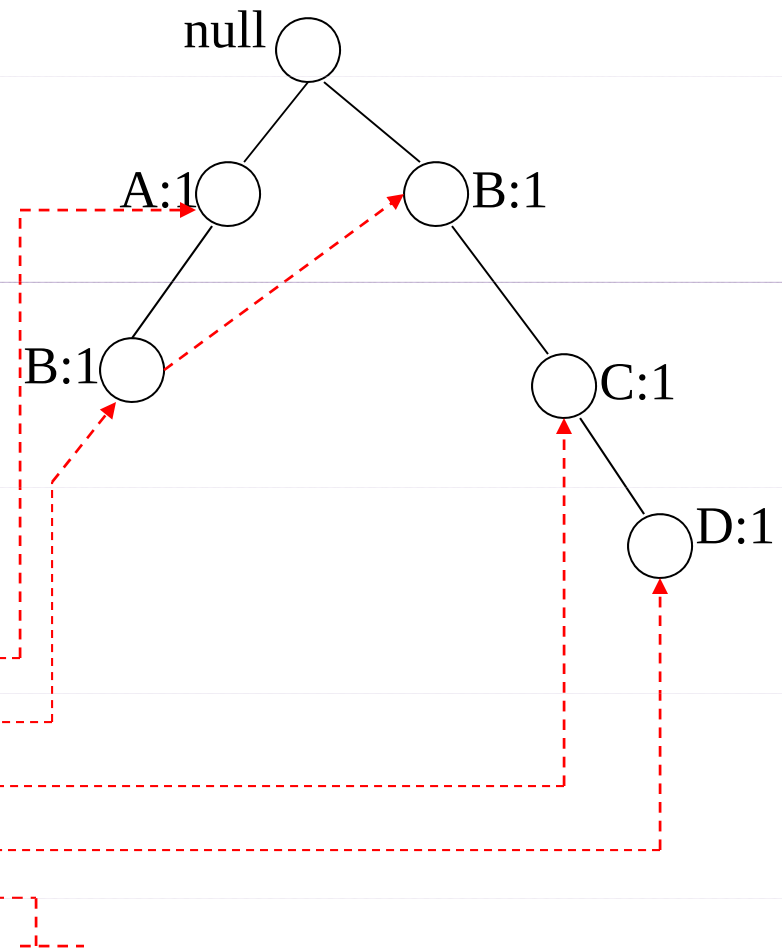
Κάθε κόμβος ετικέτα, ποιο στοιχείο και τη συχνότητα εμφάνισης (υποστήριξη) – πόσες δοσοληψίες φτάνουν σε αυτόν

Επίσης, δείκτες μεταξύ των κόμβων που αναφέρονται στο ίδιο στοιχείο

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}

Διάβασμα TID=1, 2:



Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

Επίσης, κρατάμε πίνακα δεικτών για να βοηθήσουν στον υπολογισμό των συχνών στοιχειοσυνόλων

Κατασκευή FP-δέντρου

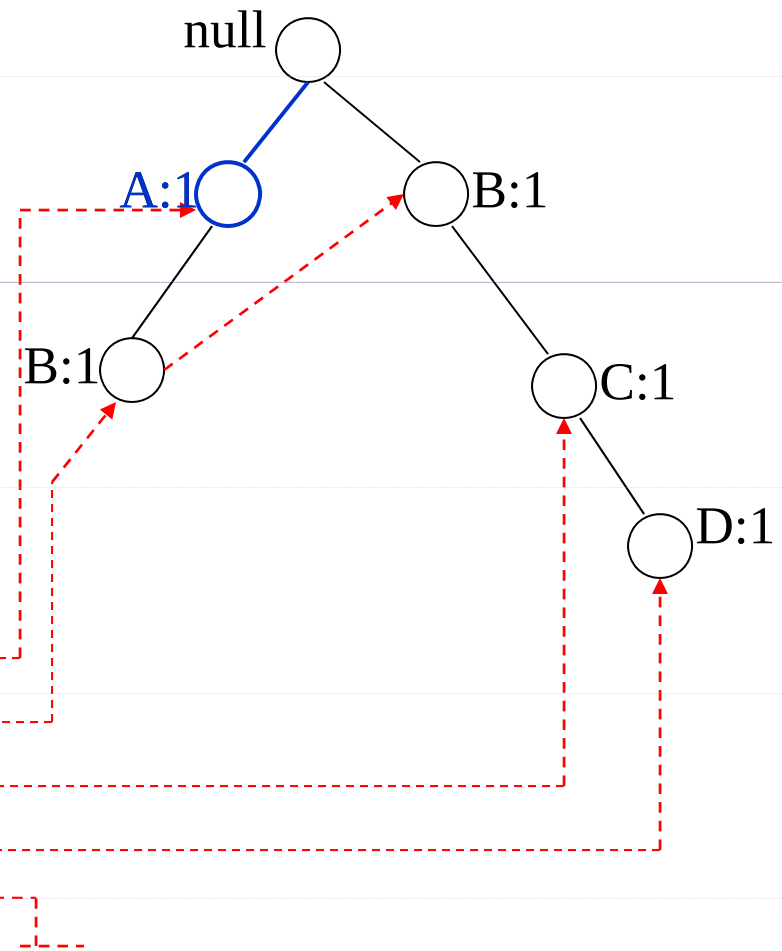
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:

Διάβασμα TID=3

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	



Κατασκευή FP-δέντρου

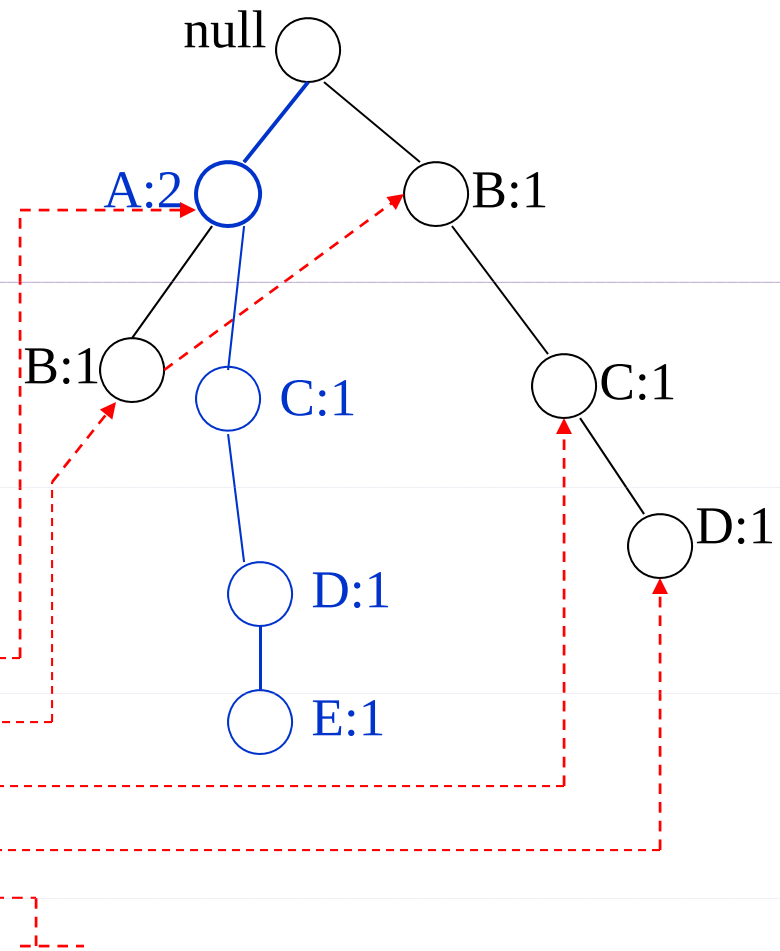
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:

Διάβασμα TID=3

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	



Κατασκευή FP-δέντρου

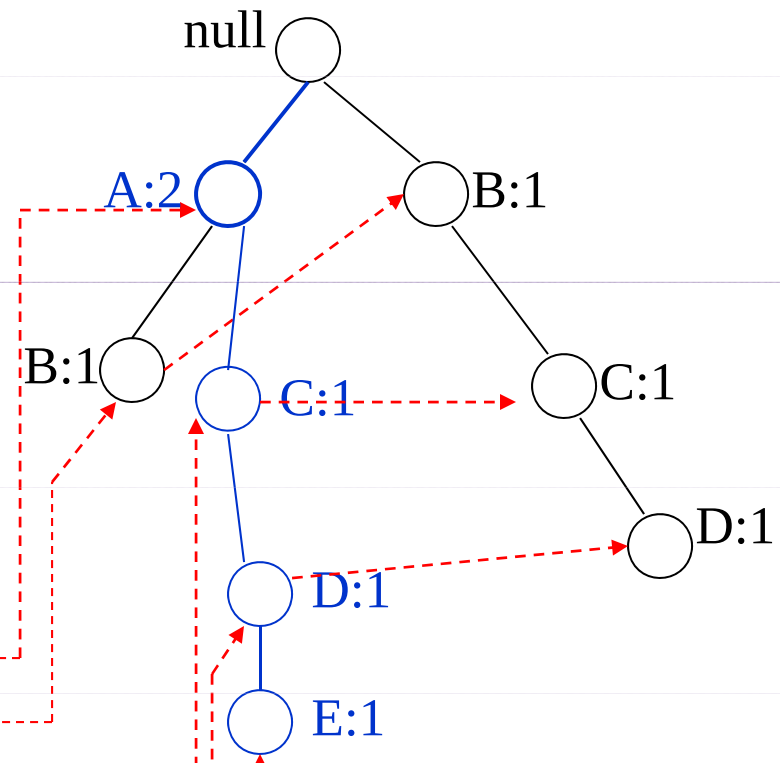
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Διάβασμα TID=1, 2:

Διάβασμα TID=3

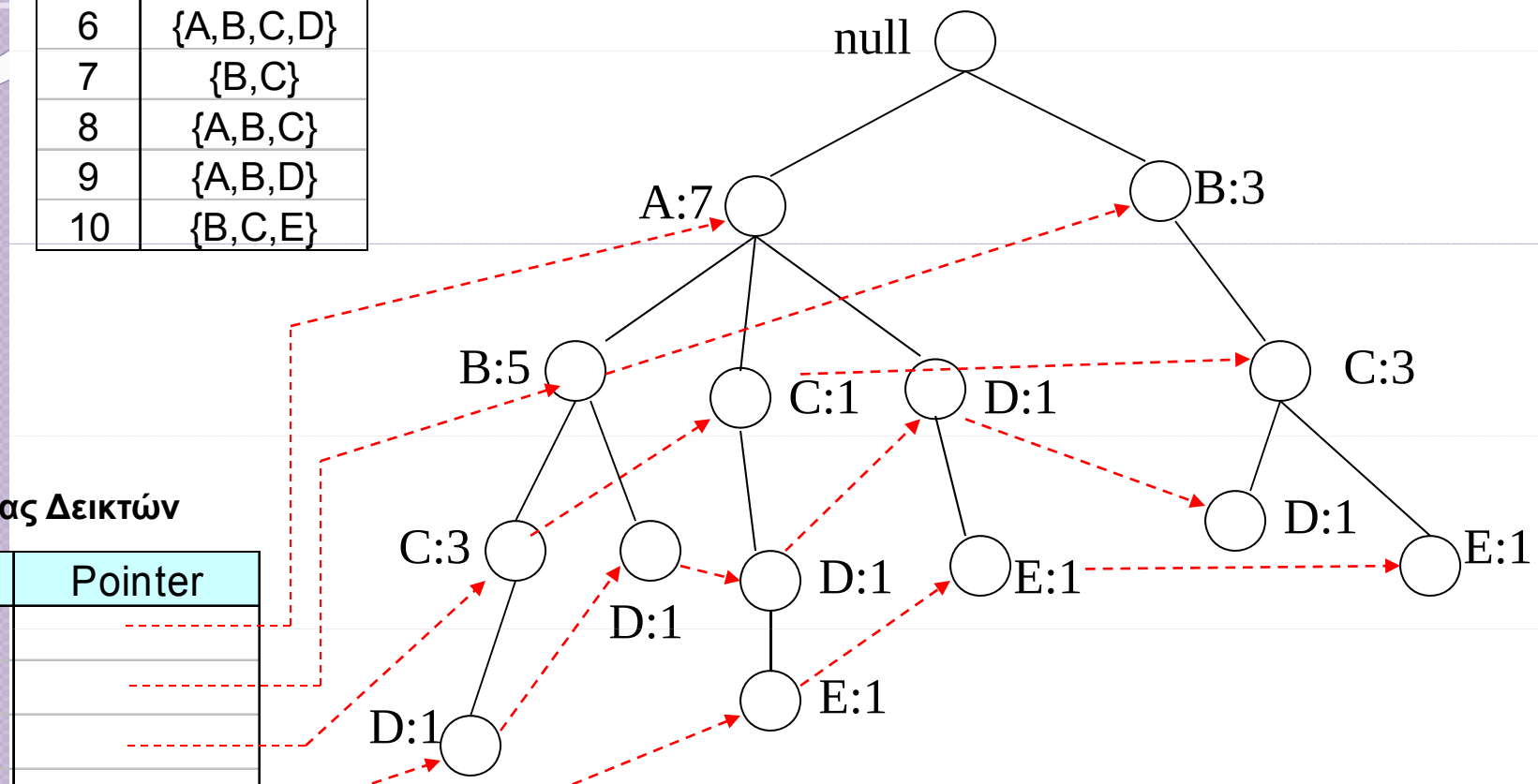
Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	



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}



Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

Μέγεθος FP-δέντρου

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

Τελικό δέντρο

- Το τελικό δέντρο, εξαρτάται από τη **διάταξη**:
  - άλλη διάταξη -> άλλα προθέματα
- (Συνήθως) μικρότερο δέντρο, αν διατάξουμε με βάση τη συχνότητα εμφάνισης και όχι λεξικογραφικά
  - Αρχικά, διαβάζουμε όλα τα δεδομένα μια φορά ώστε να υπολογιστεί ο μετρητής υποστήριξης κάθε στοιχείου, και διατάσσουμε τα στοιχεία με βάση αυτό

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}

- **Επίσης, αγνοούμε όσα στοιχεία είναι μη συχνά**

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

Αλγόριθμος εύρεσης συχνών στοιχειοσυνόλων

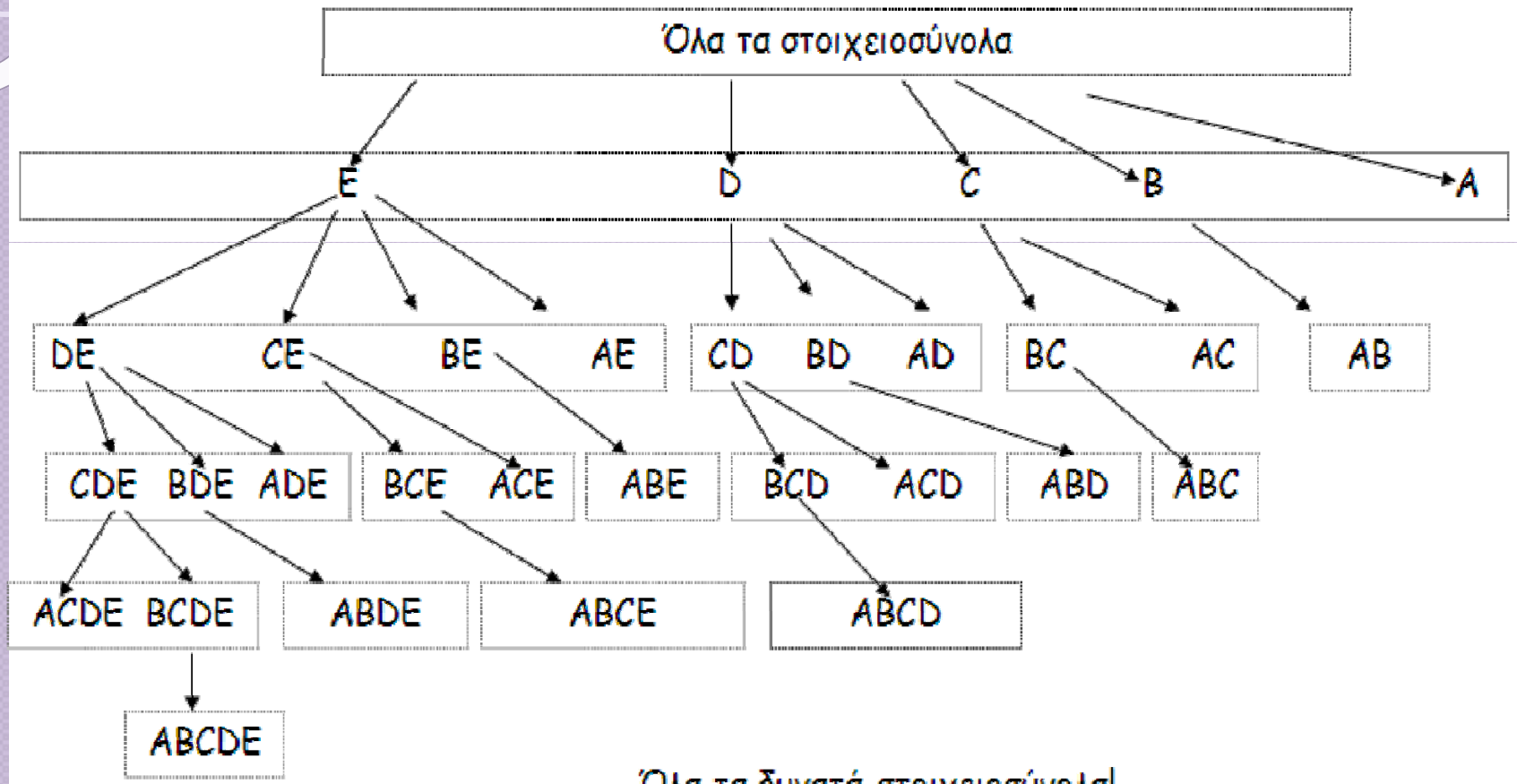
Είσοδος: FP-δέντρο

Έξοδος: Συχνά στοιχειοσύνολα και η υποστήριξη τους

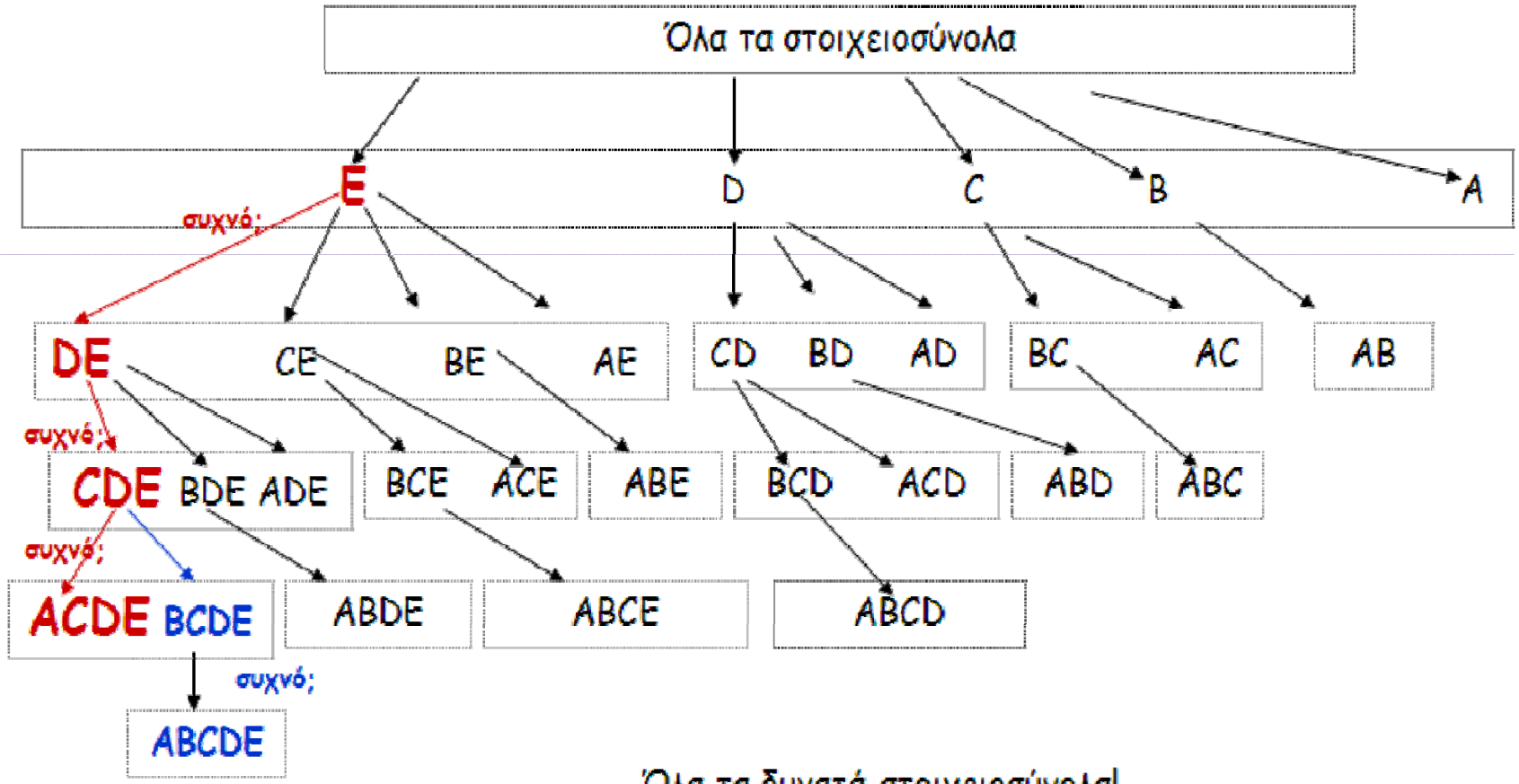
Μέθοδος:

- Διαίρει-και-Βασίλευε
  - Χωρίζουμε τα στοιχειοσύνολα σε αυτά που τελειώνουν σε E, D, C, B, A
    - Μετά αυτά που τελειώνουν σε E σε αυτά που τελειώνουν σε DE, CE, BE, AE κοκ

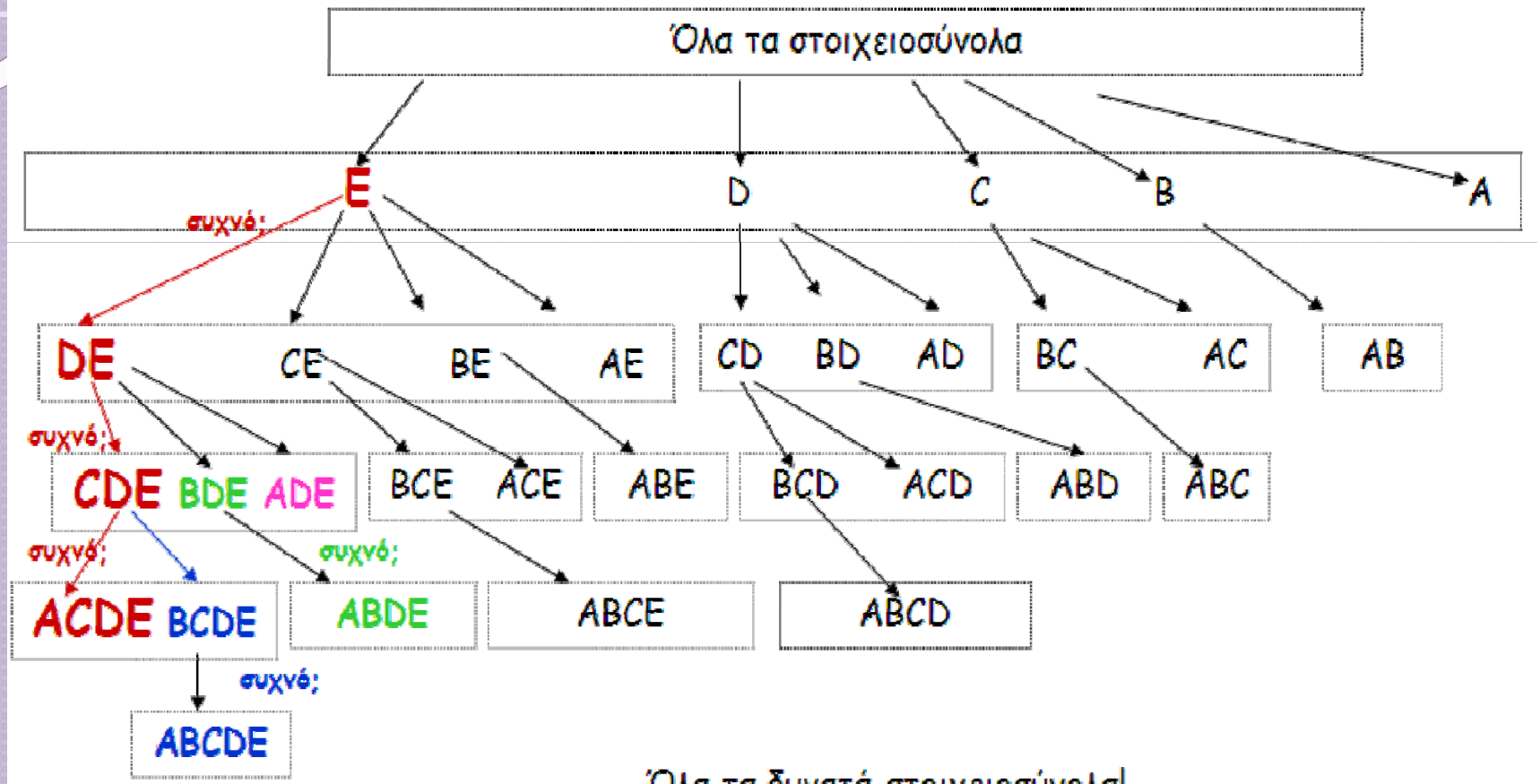
Αλγόριθμος FP-Growth



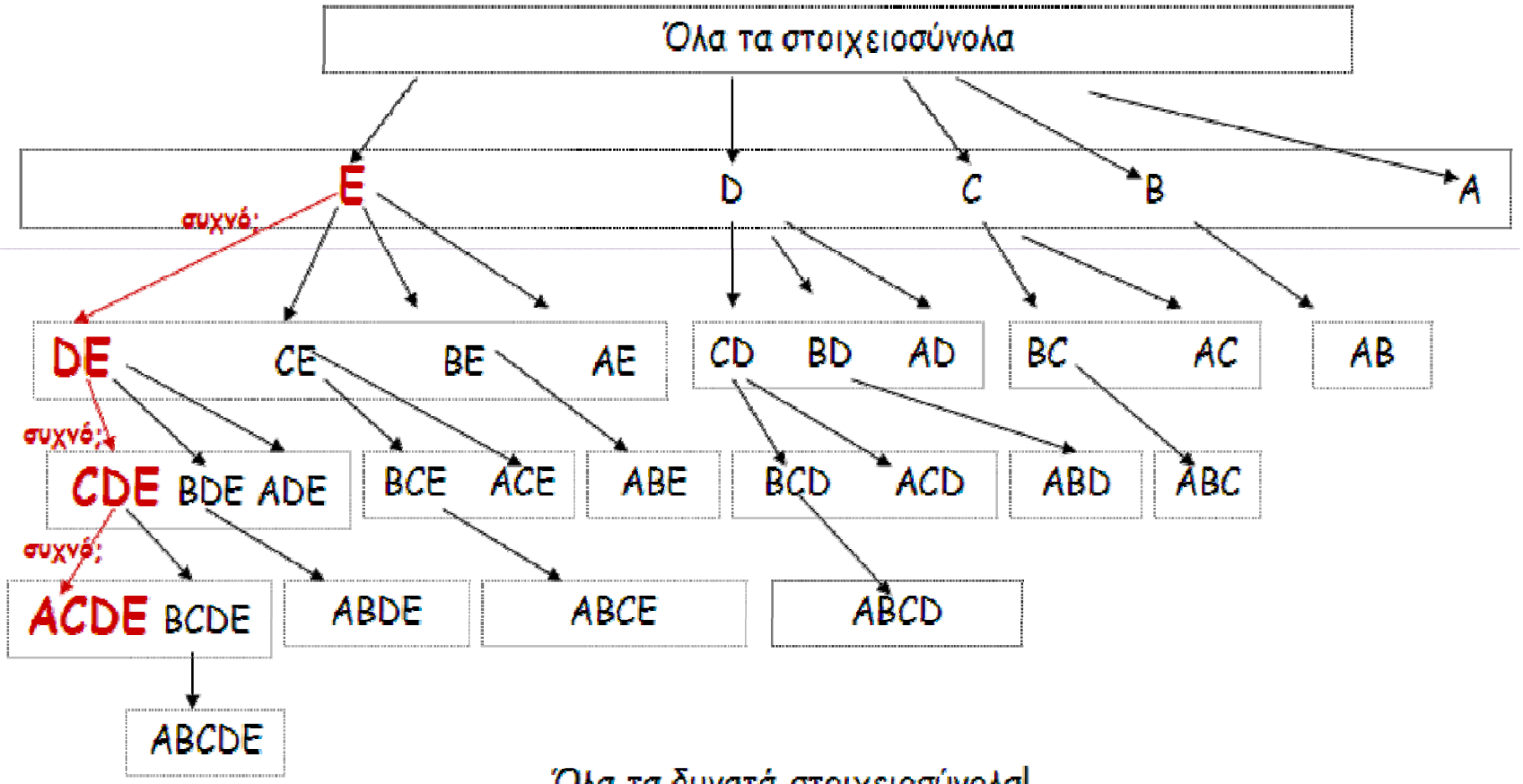
Όλα τα δυνατά στοιχειοσύνολα!



Όλα τα δυνατά στοιχειοσύνολα!



Όλα τα δυνατά στοιχειοσύνολα!

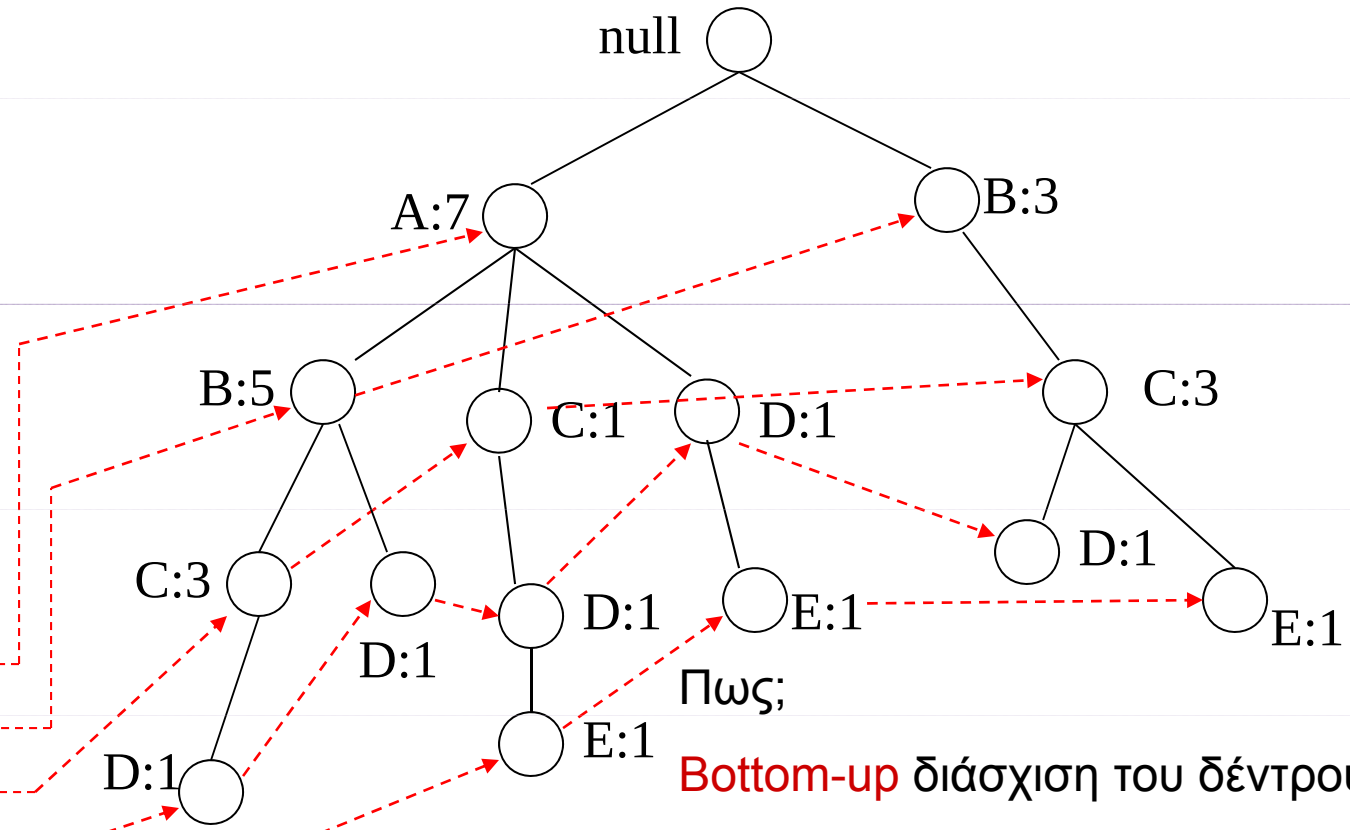


Όλα τα δυνατά στοιχειοσύνολα!

Στο δέντρο μπορεί να υπάρχουν λιγότερα!

Παράδειγμα

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

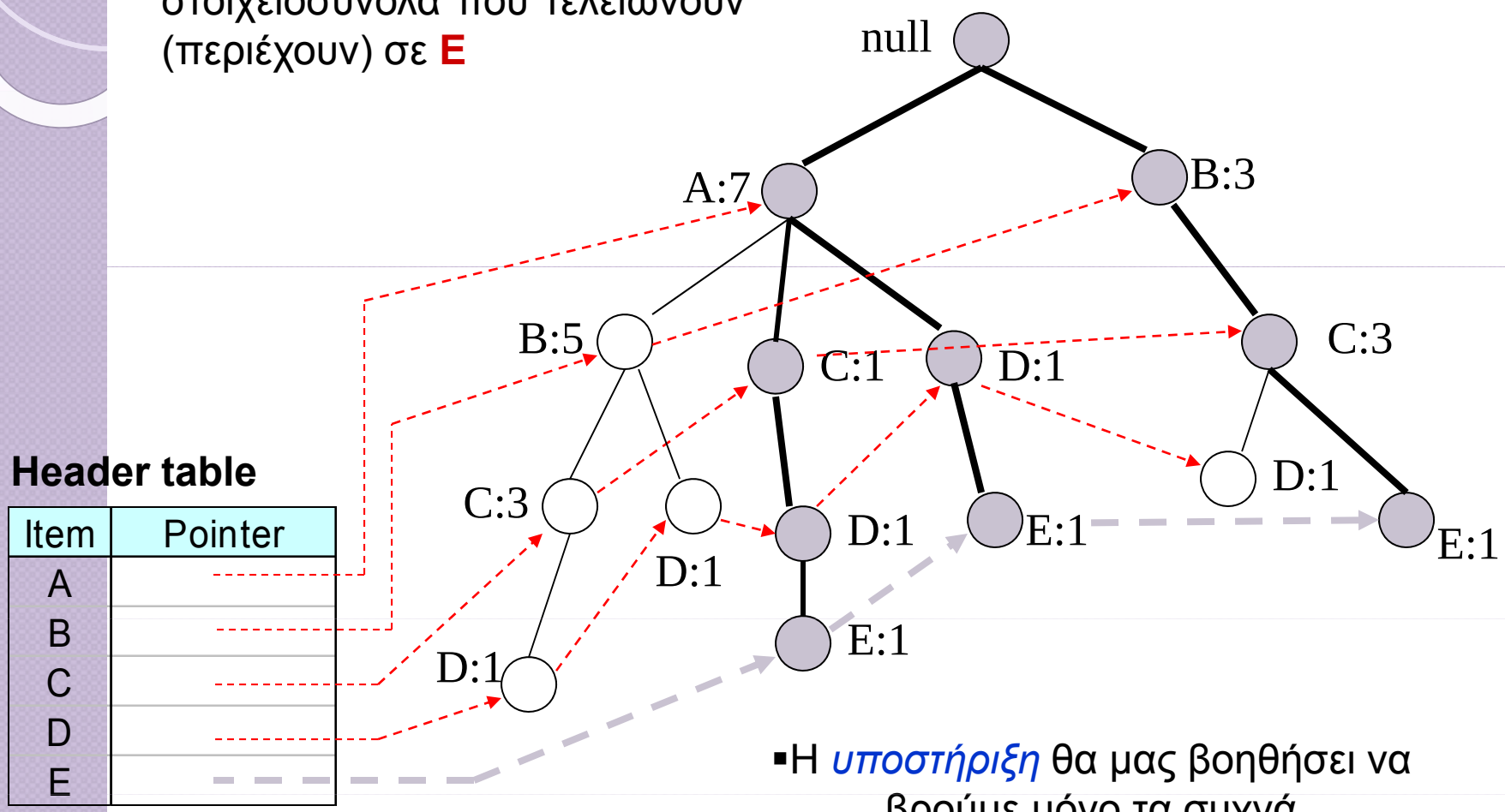
Item	Pointer
A	
B	
C	
D	
E	

Πως;

Bottom-up διάσχιση του δέντρου

Από το Header Table βρίσκω αυτά που τελειώνουν σε E, μετά αυτά που τελειώνουν σε D, C, B και τέλος A – suffix-based classes (επίθεμα – κατάληξη)

Υποπρόβλημα: Βρες συχνά στοιχειοσύνολα που τελειώνουν (περιέχουν) σε **E**

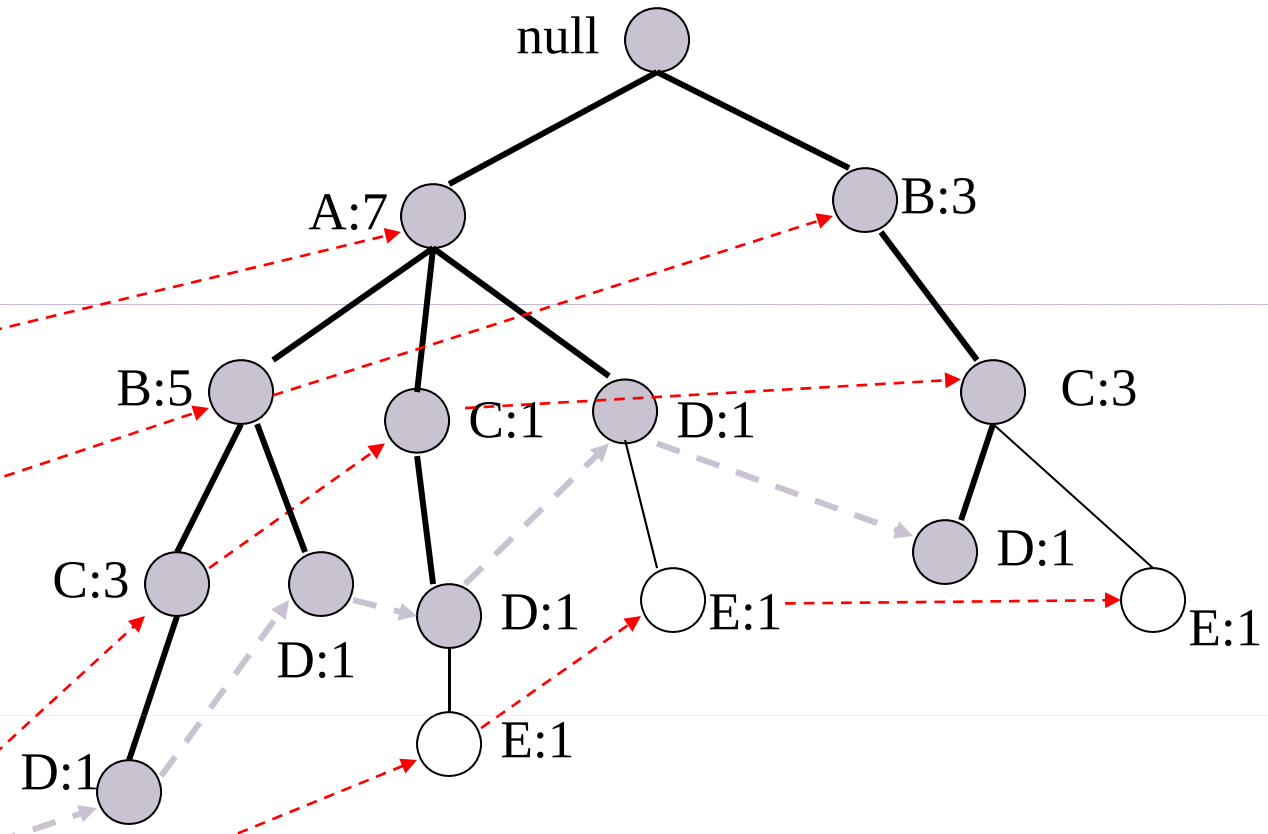


- Η *υποστήριξη* θα μας βοηθήσει να βρούμε μόνο τα συχνά στοιχειοσύνολα

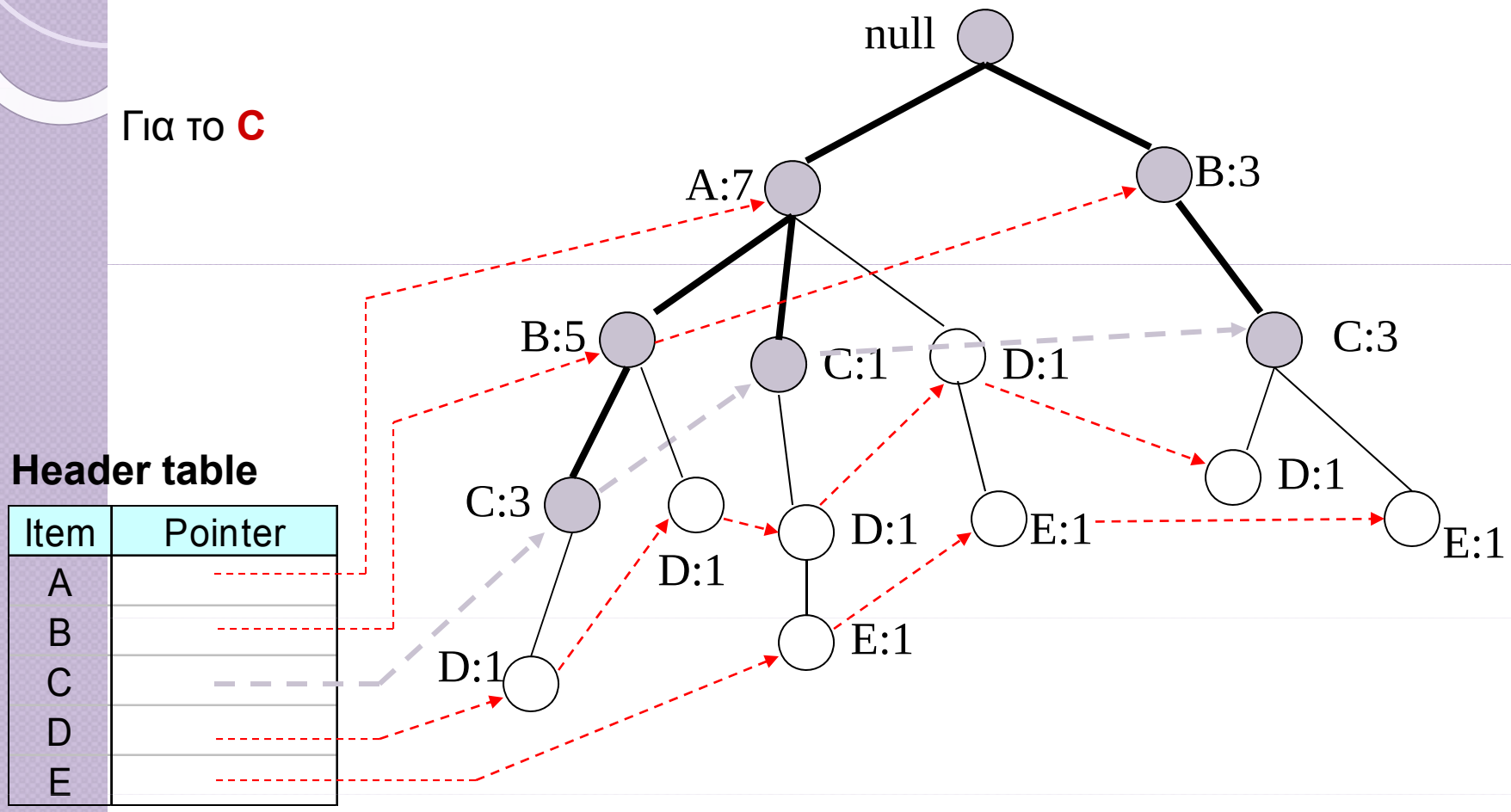
Για το **D**

Header table

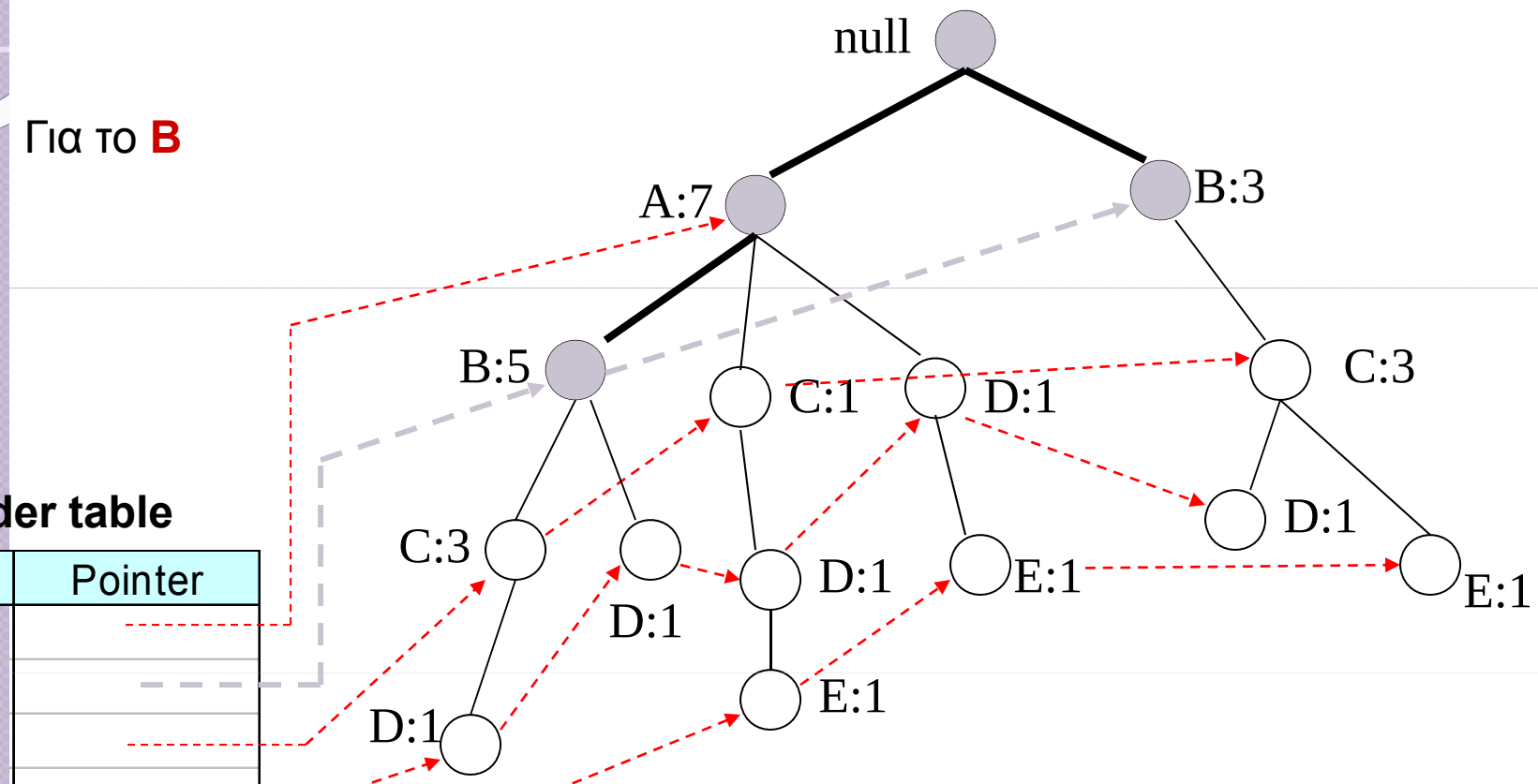
Item	Pointer
A	
B	
C	
D	
E	



Για το **C**


null
A:7
B:3
B:5
C:1
D:1
C:3
C:3
D:1
D:1
D:1
E:1
E:1
D:1
E:1
Header table
Item
Pointer
A
B
C
D
E

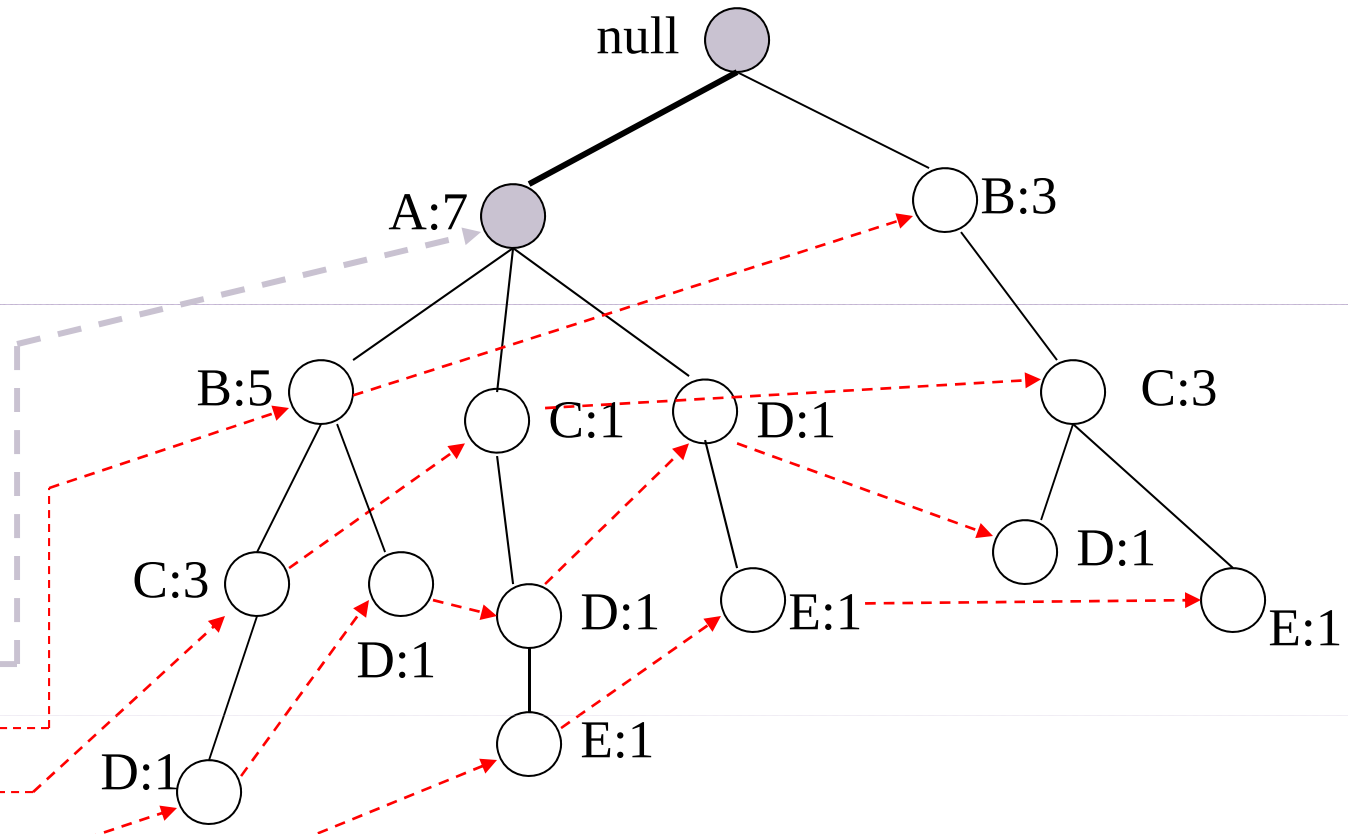
Για το **B**



Header table

Item	Pointer
A	
B	
C	
D	
E	

Για το **A**



Header table

Item	Pointer
A	
B	
C	
D	
E	

Συνοπτικά

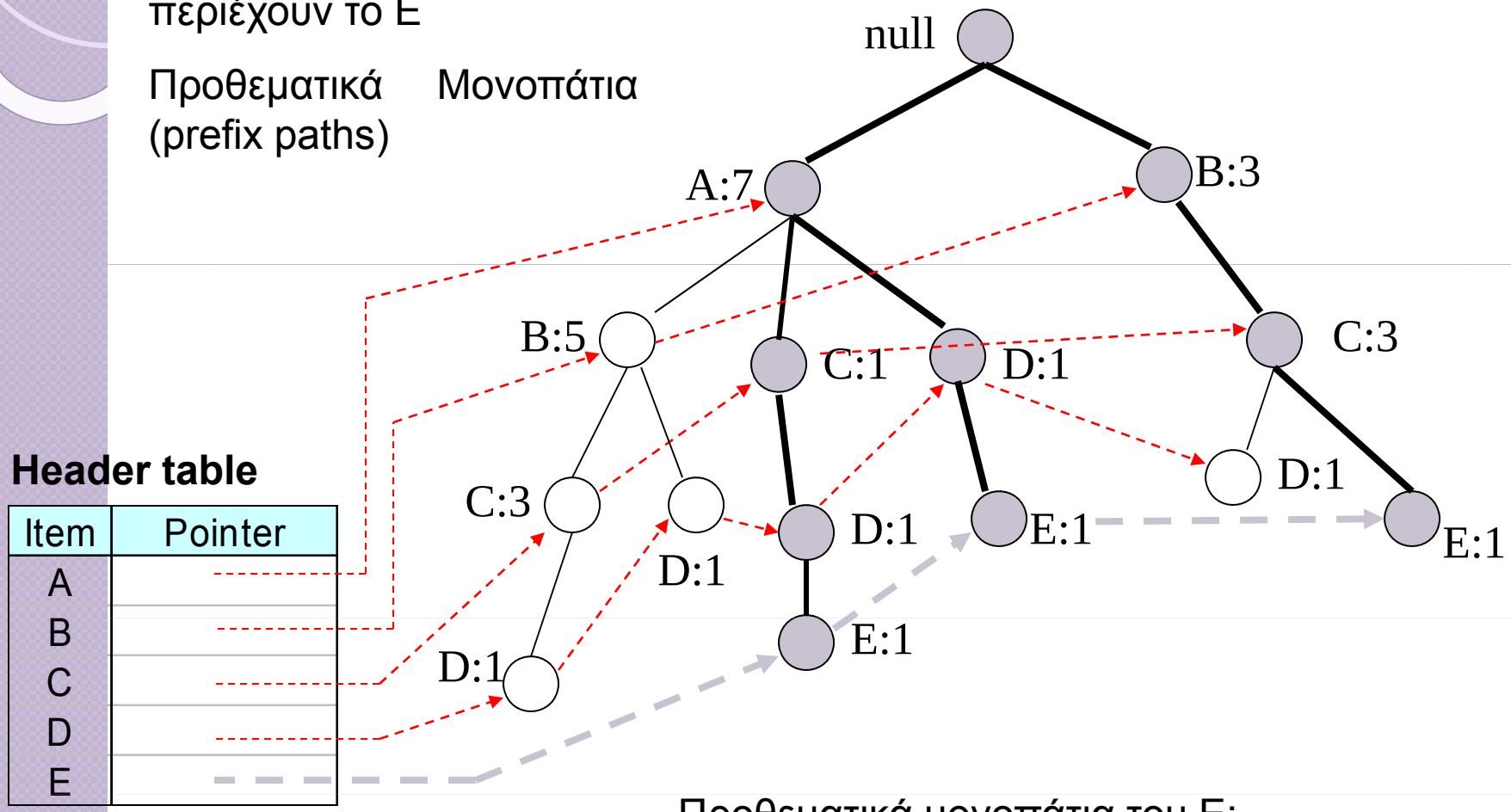
Σε κάθε βήμα, για το suffix (επίθεμα) X

- Φάση 1
  - Κατασκευάζουμε το προθεματικό δέντρο για το X και υπολογίζουμε την υποστήριξη χρησιμοποιώντας τον πίνακα
- Φάση 2
  - Αν είναι συχνό, κατασκευάζουμε το υπο-συνθήκη δέντρο για το X, σε βήματα
    - επανα-υπολογισμός υποστήριξης
    - περικοπή κόμβων με μικρή υποστήριξη
    - περικοπή φύλλων

Φάση 1 – κατασκευή προθεματικού δέντρου

Όλα τα μονοπάτια που περιέχουν το E

Προθεματικά Μονοπάτια (prefix paths)



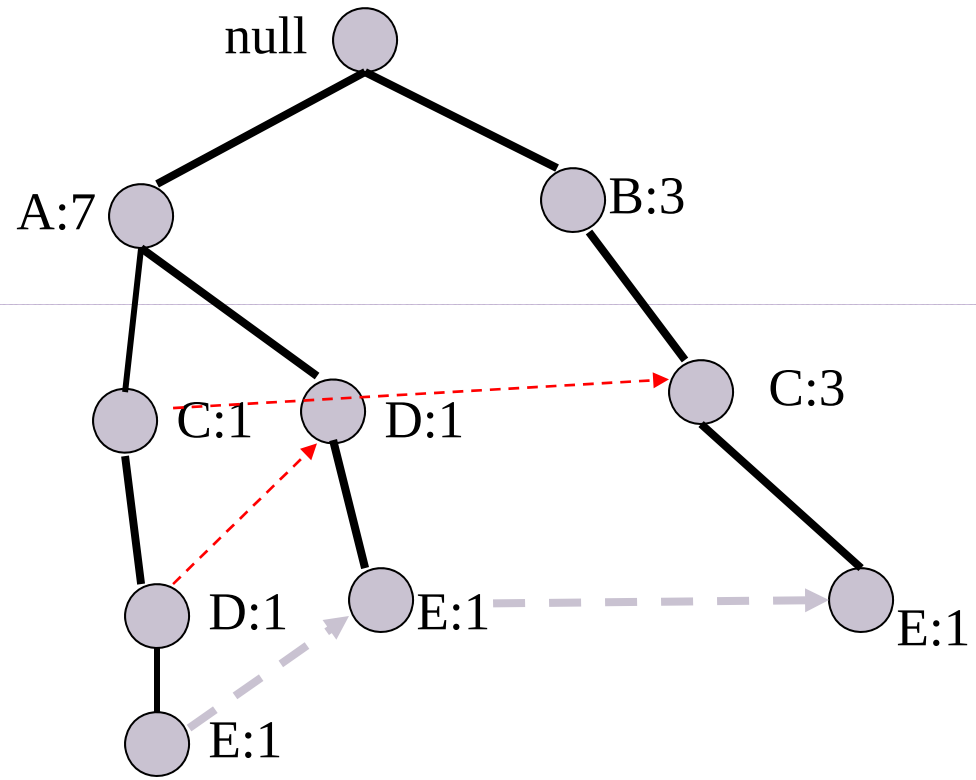
Προθεματικά μονοπάτια του E:

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

Φάση 1 – κατασκευή προθεματικού δέντρου

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}

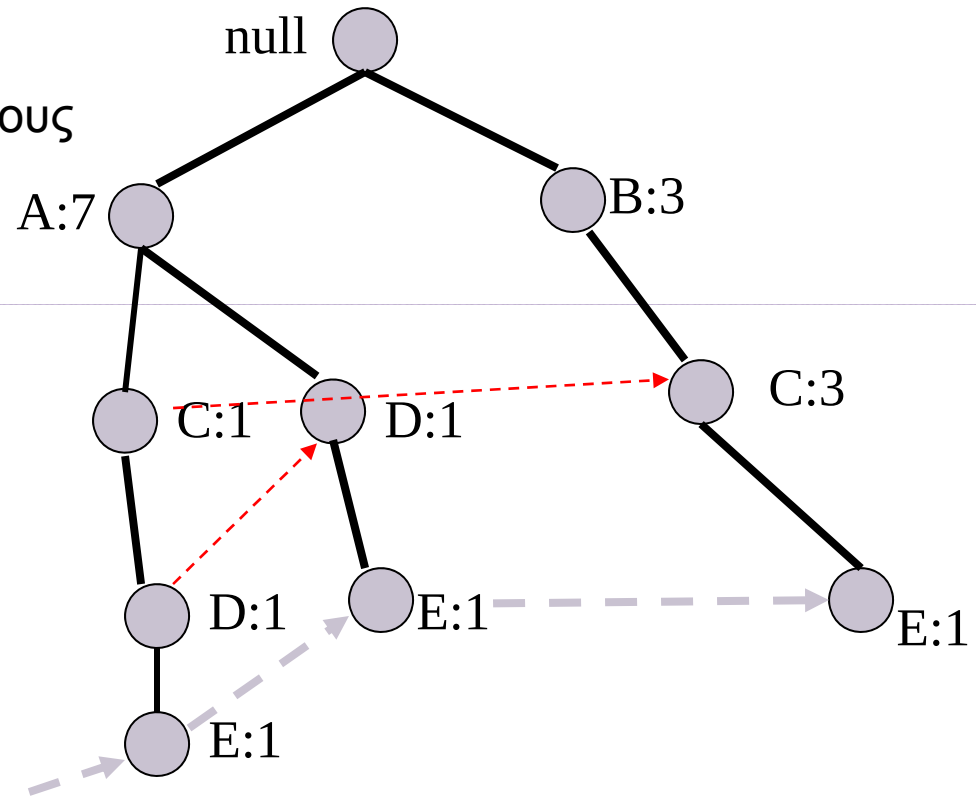
Έστω minsup = 2

Βρες την υποστήριξη του {E}

Πως;

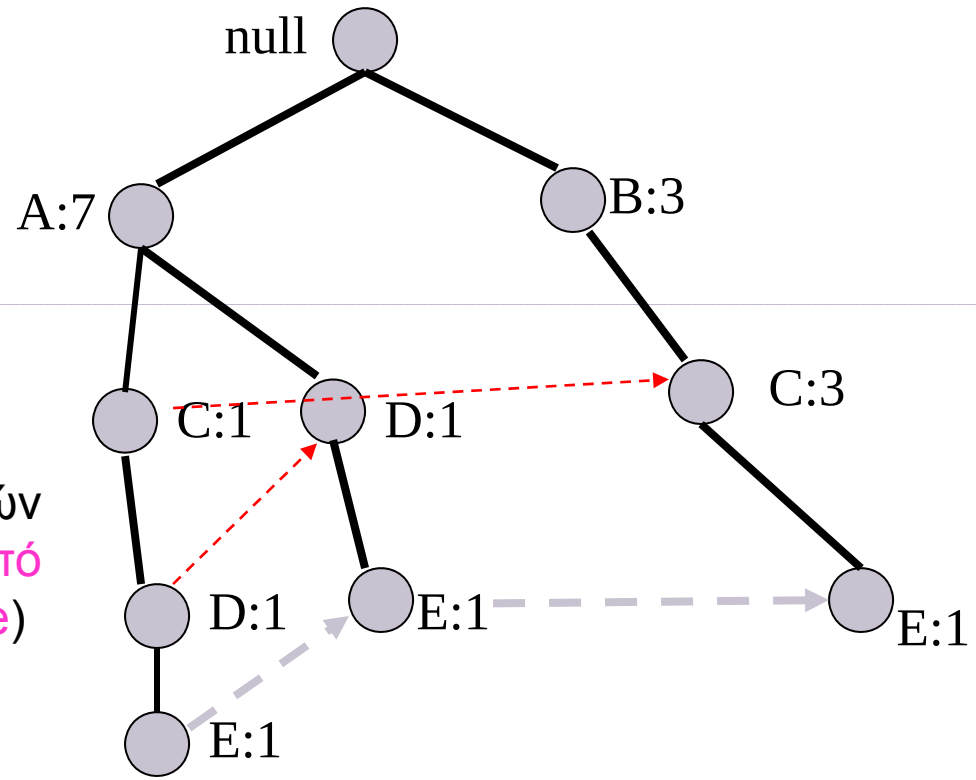
Ακολούθησε τους συνδέσμους αθροίζοντας 1+1+1=3>2

Οπότε {E} συχνό



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

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



Φάση 2

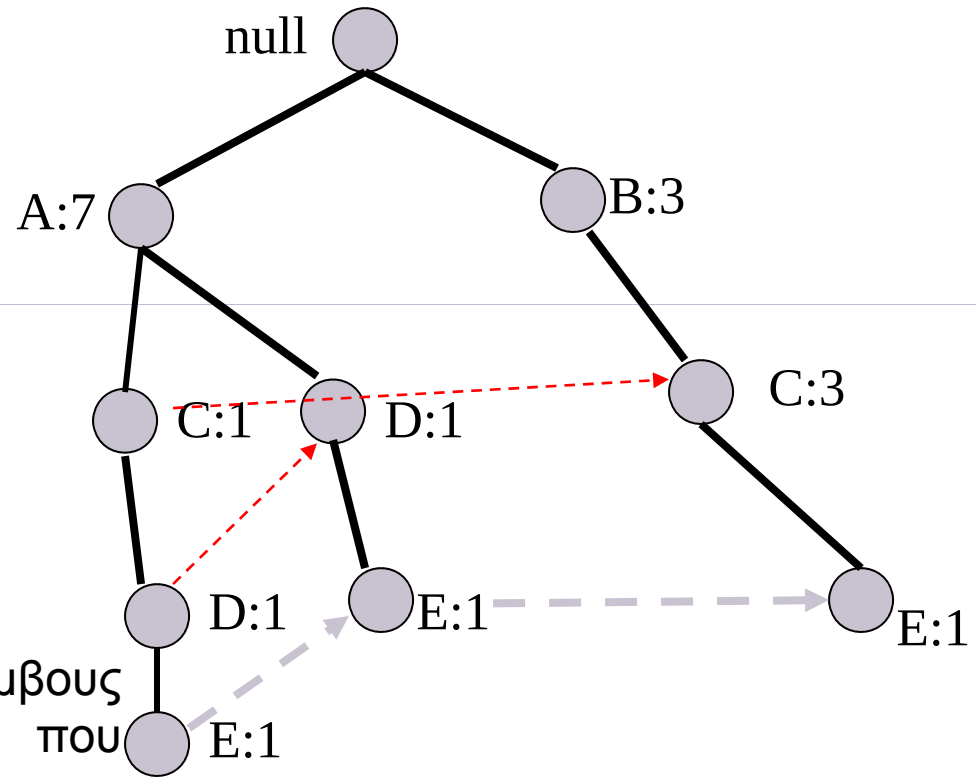
Μετατροπή των προθεματικών δέντρων σε FP-δέντρο υπό συνθήκες (conditional FP-tree)

Δύο αλλαγές

(1) Αλλαγή των μετρητών

(2) Περικοπή

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

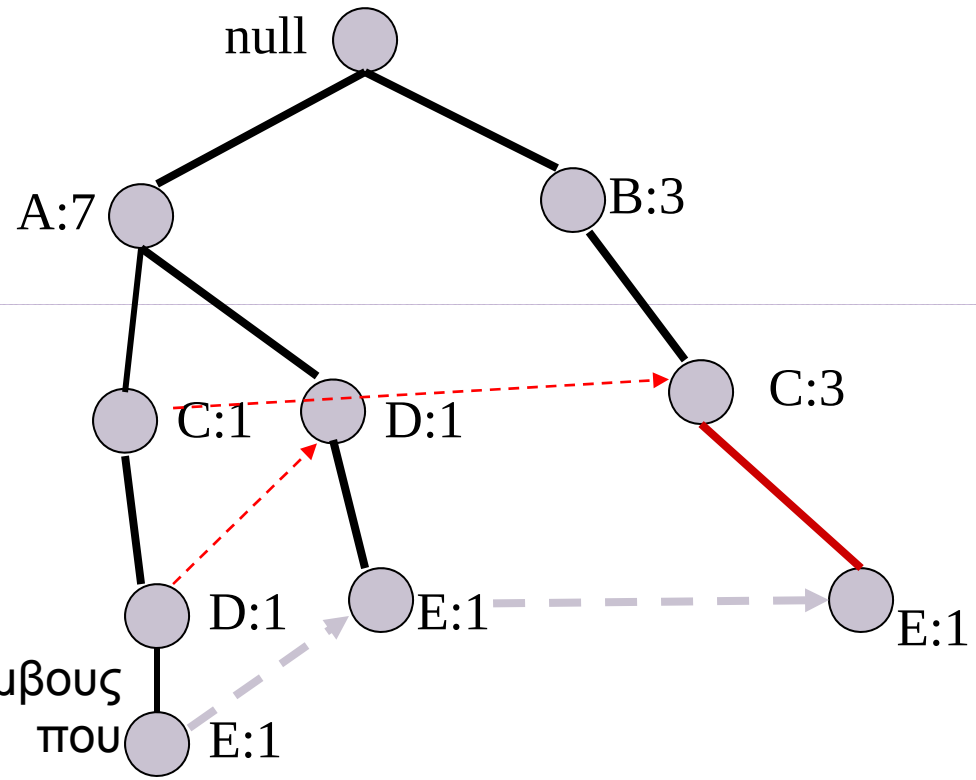


1) Αλλαγή μετρητών

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

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

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

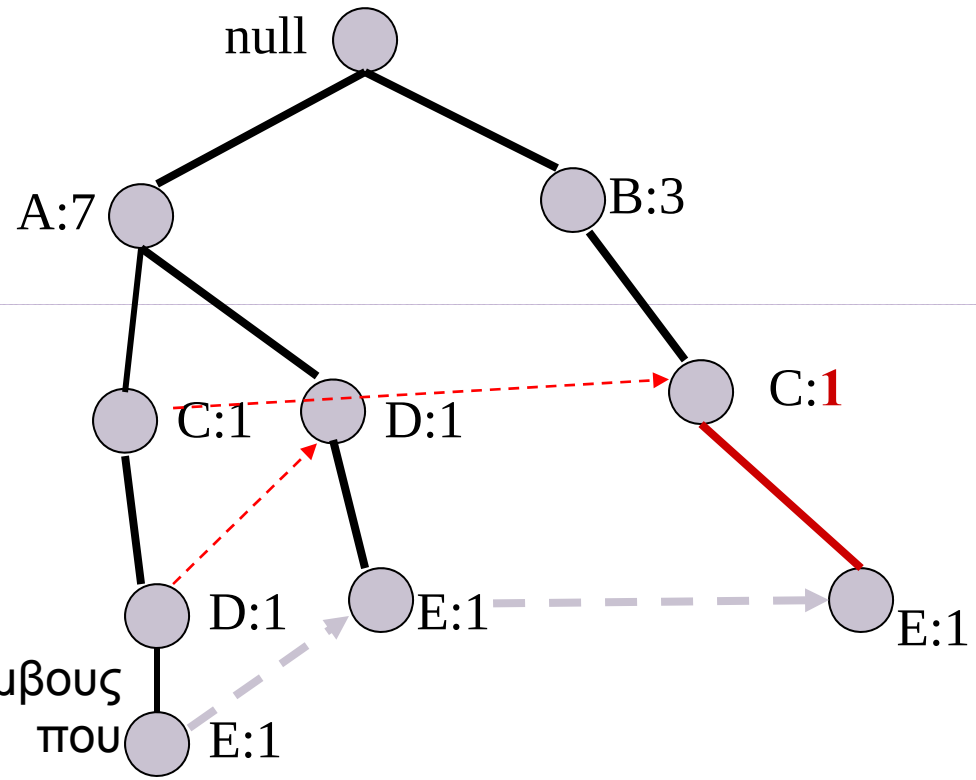


1) Αλλαγή μετρητών

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

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

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

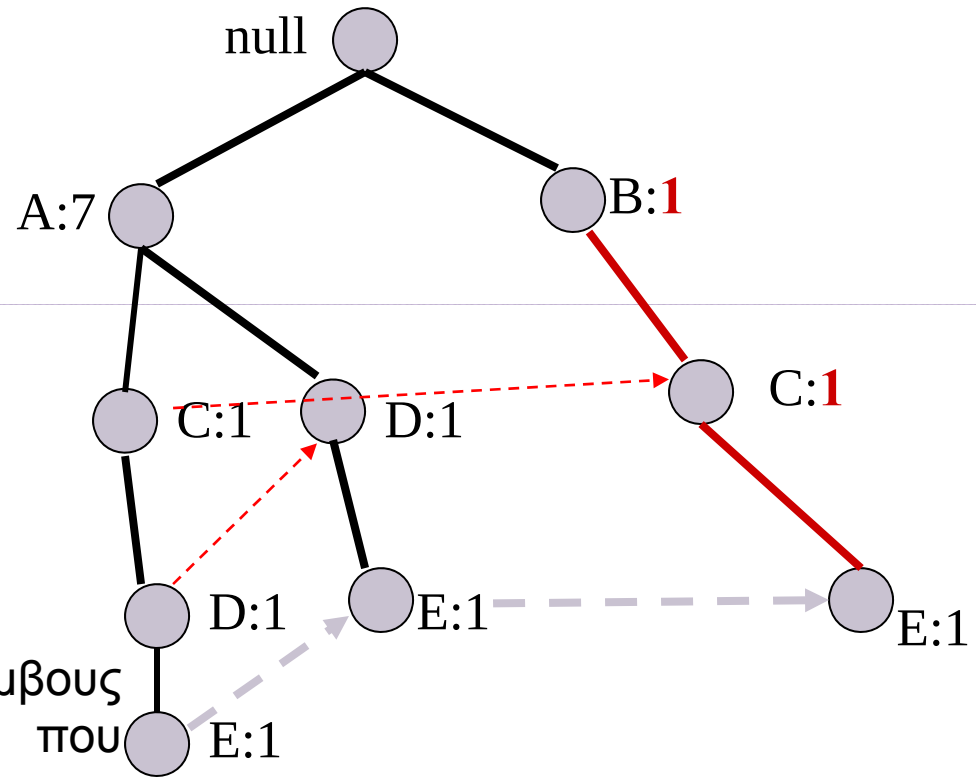


1) Αλλαγή μετρητών

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

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

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

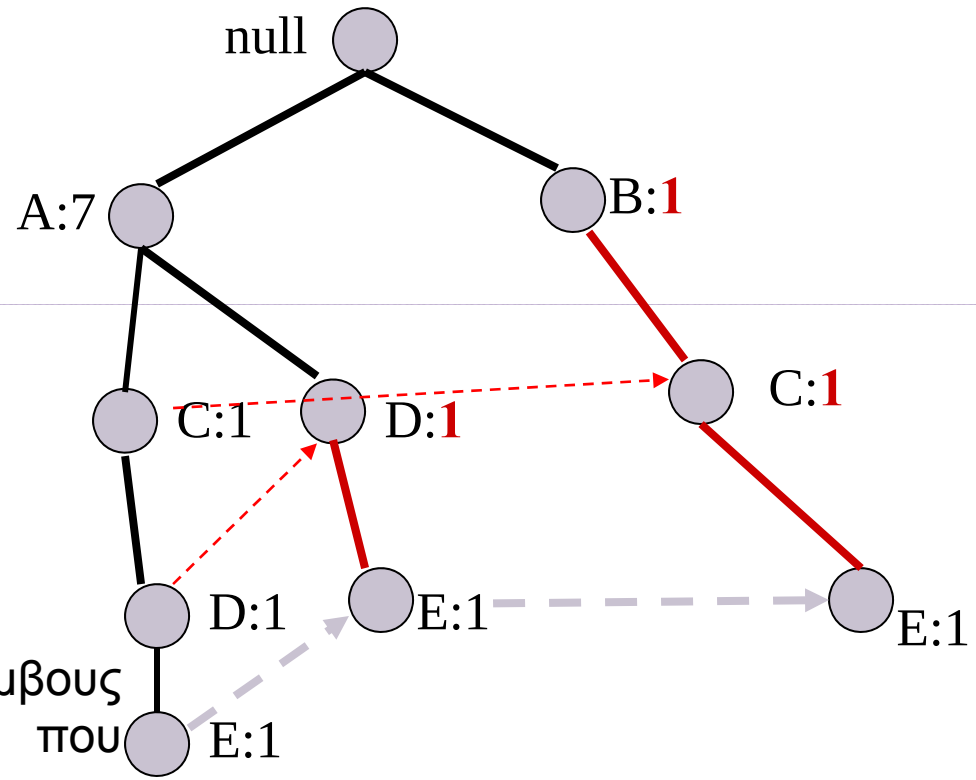


1) Αλλαγή μετρητών

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

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

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

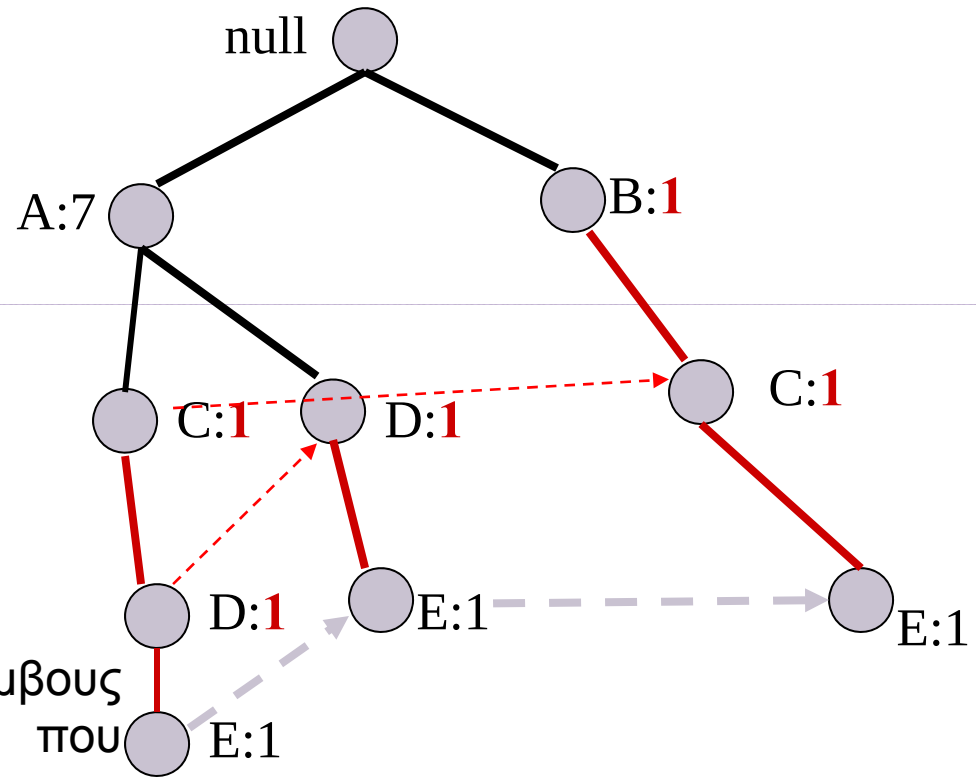


1) Αλλαγή μετρητών

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

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

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

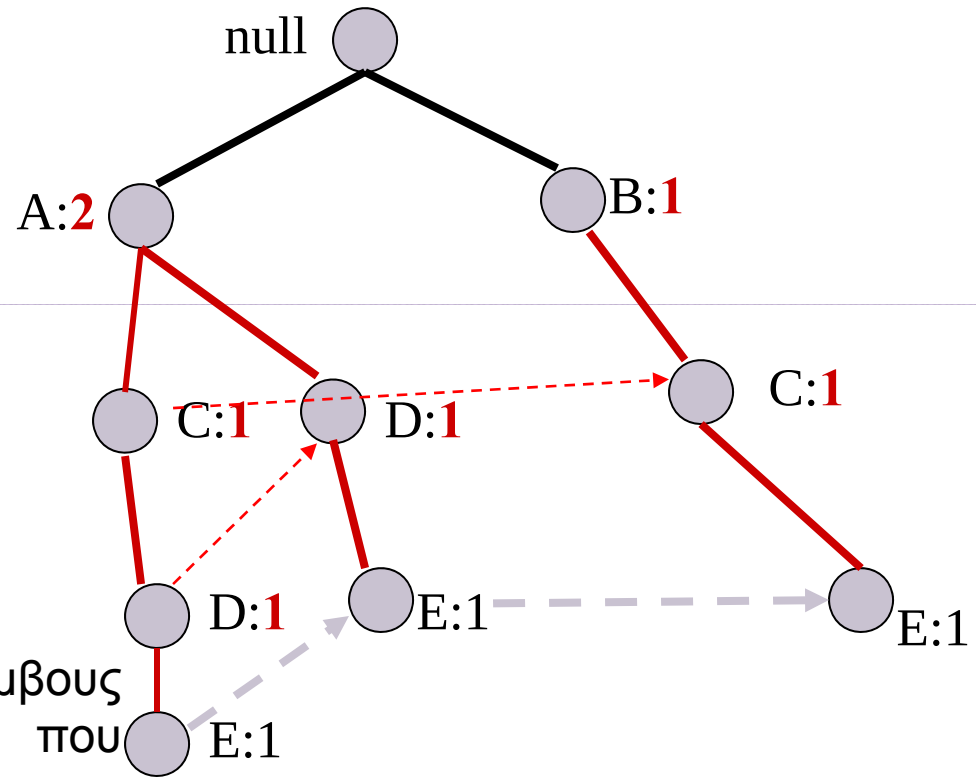


1) Αλλαγή μετρητών

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

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

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

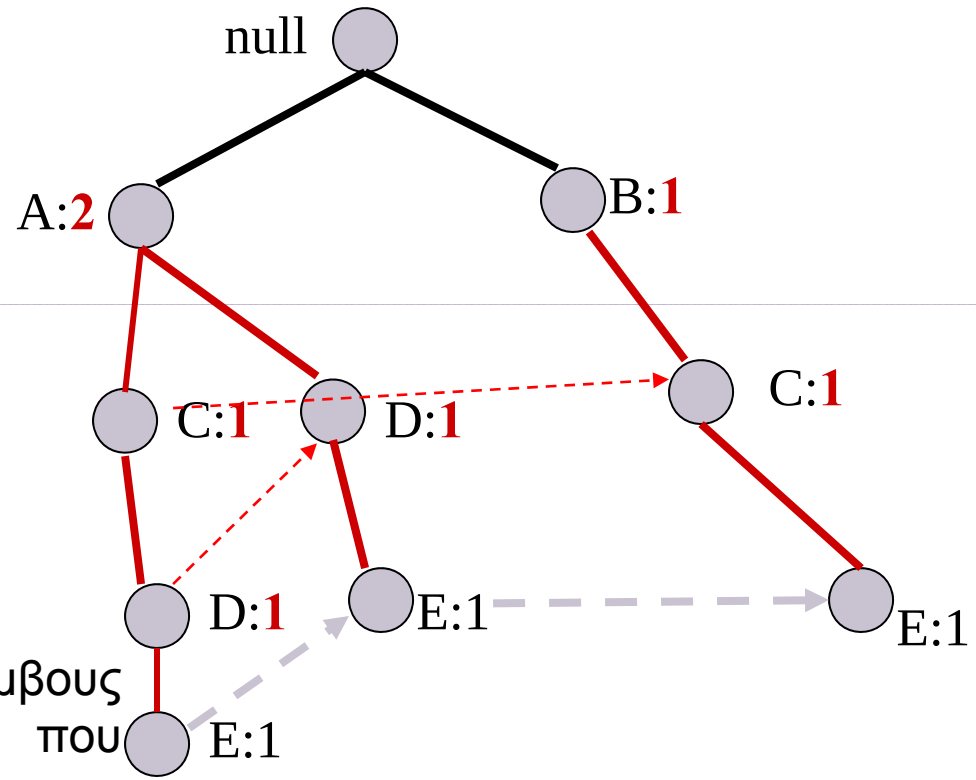


1) Αλλαγή μετρητών

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

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

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

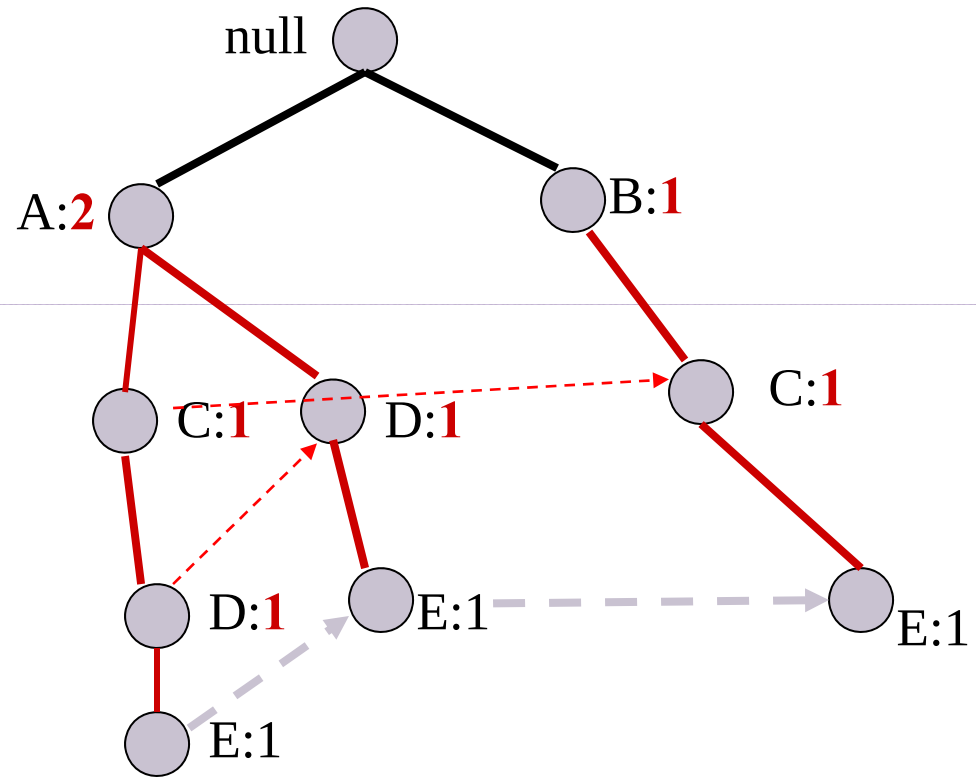


1) Αλλαγή μετρητών

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

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

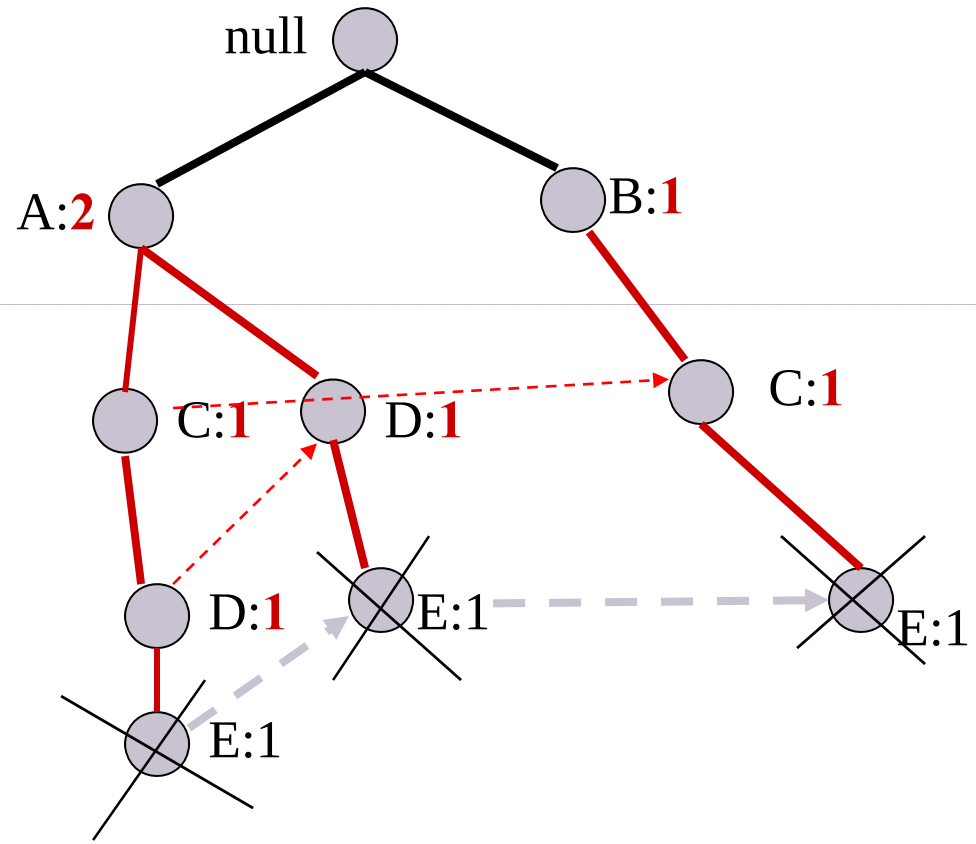
{E} συχνό άρα προχωράμε για DE, CE, BE, AE



2) Περικοπή (truncate)

Σβήσε τους κόμβους του E

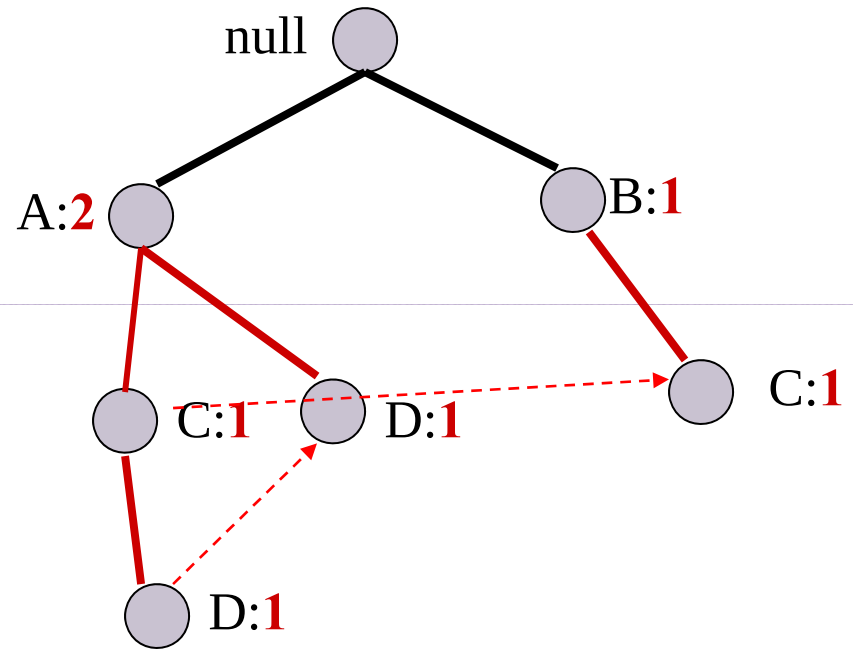
{E} συχνό άρα προχωράμε για DE, CE, BE, AE



2) Περικοπή (truncate)

Σβήσε τους κόμβους του E

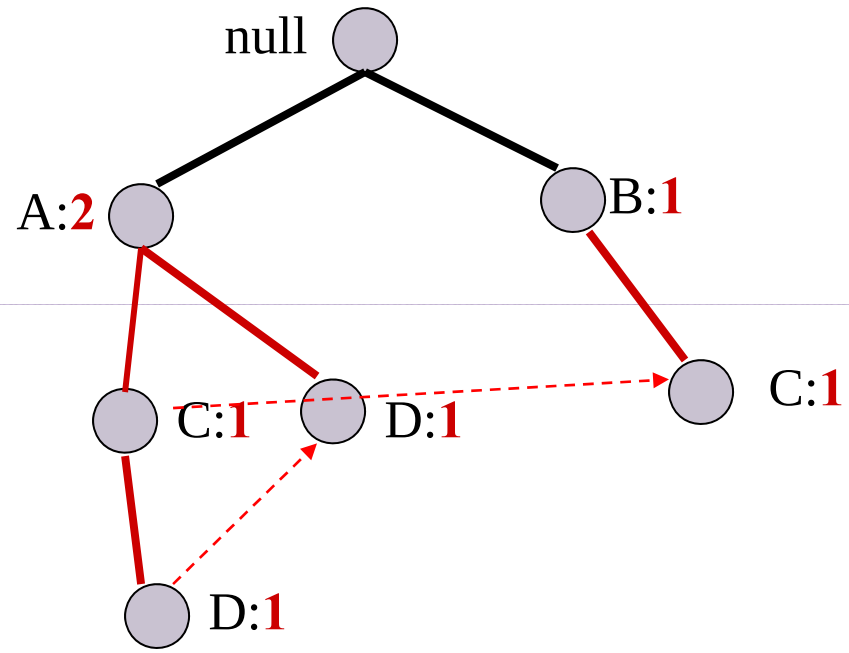
{E} συχνό άρα προχωράμε για DE, CE, BE, AE



2) Περικοπή (truncate)

Σβήσε τους κόμβους του E

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



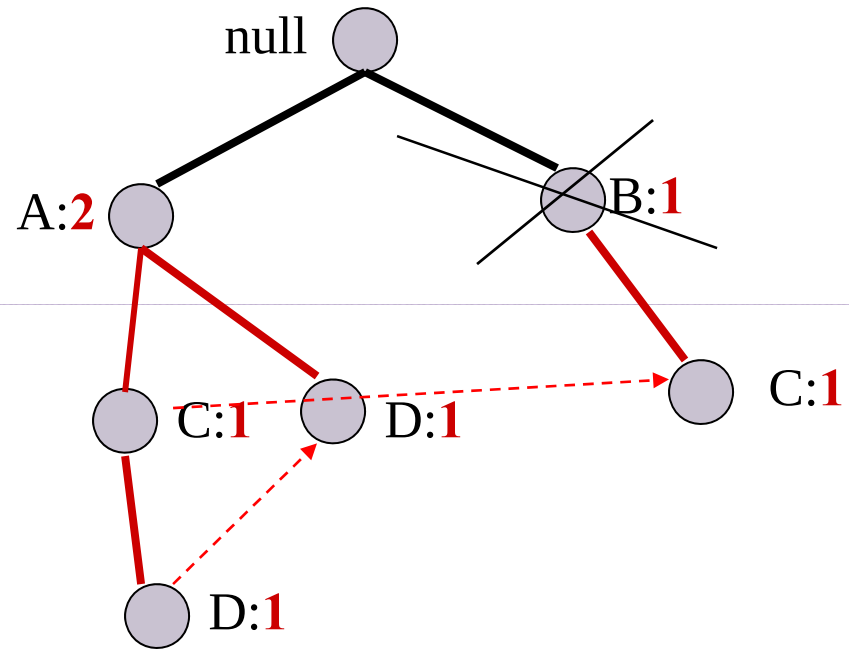
Πιθανή περαιτέρω περικοπή

Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης

Πχ το B -> περικοπή

Αυτό σημαίνει ότι το B εμφανίζεται μαζί με το E λιγότερο από minsup φορές

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



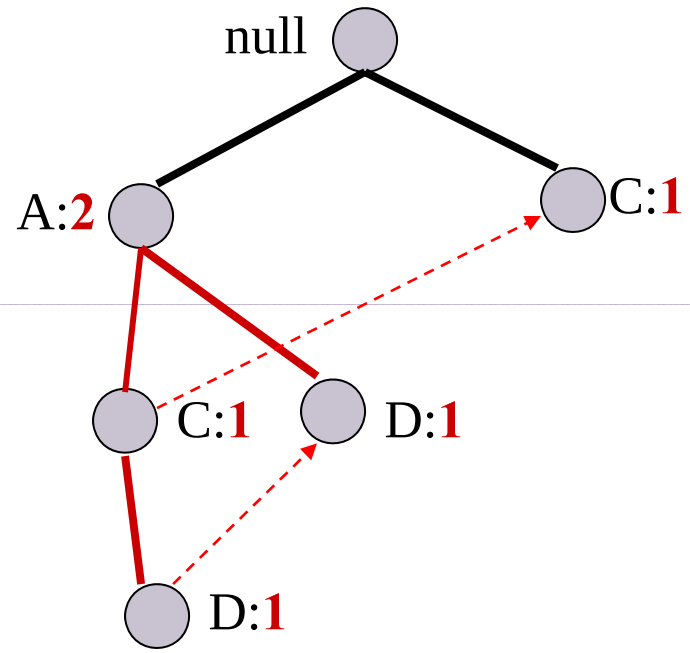
Πιθανή περαιτέρω περικοπή

Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης

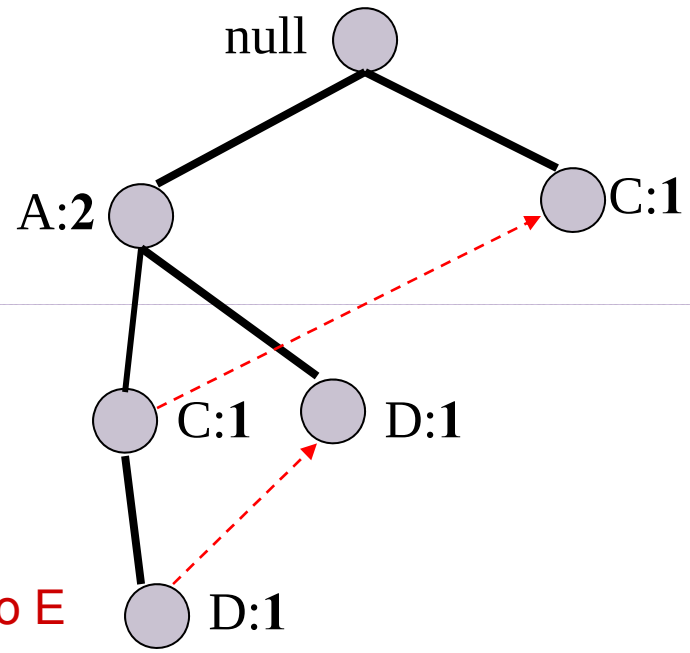
Πχ το B -> περικοπή

Αυτό σημαίνει ότι το B εμφανίζεται μαζί με το E λιγότερο από minsup φορές

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

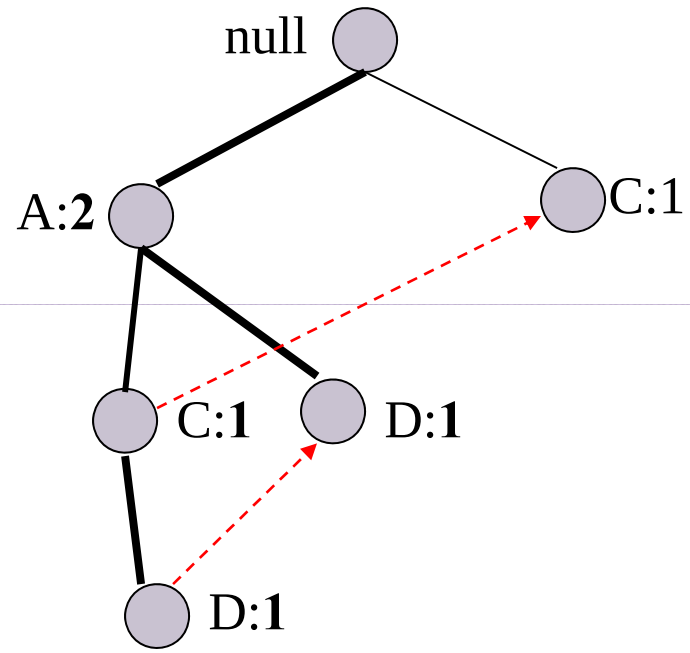


{E} συχνό άρα προχωράμε για DE, CE, BE, AE



Υπο-συνθήκη FP-δέντρο για το E

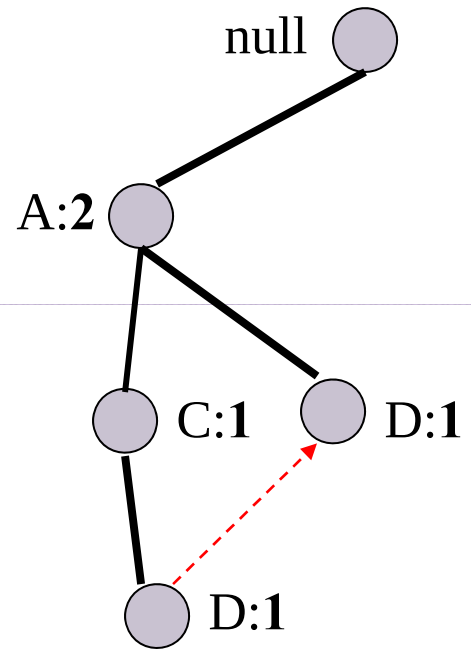
Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, {C, E}, {A, E}



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

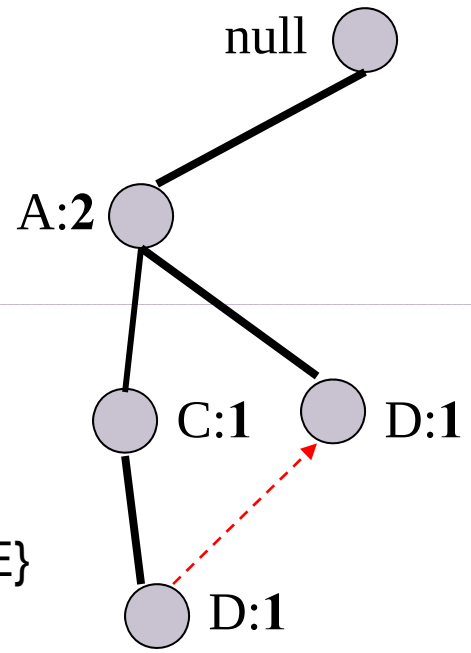
Προθεματικά Μονοπάτια (prefix paths)



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)

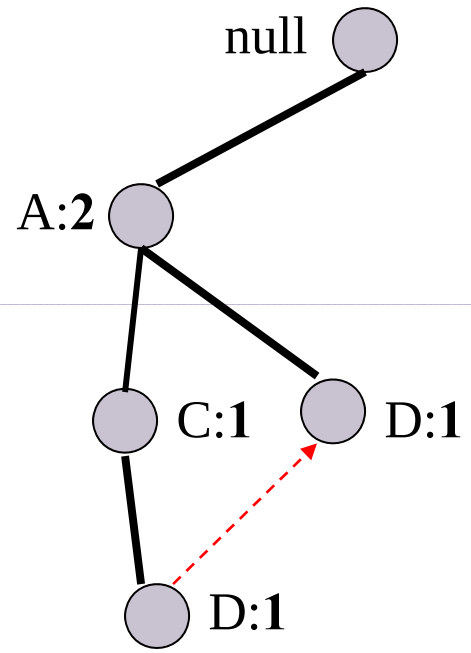


Βρες την υποστήριξη του {D, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

Οπότε {D, E} συχνό

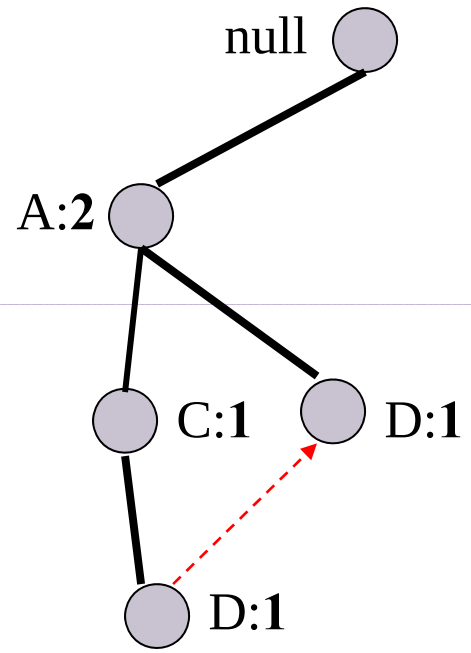


Φάση 2

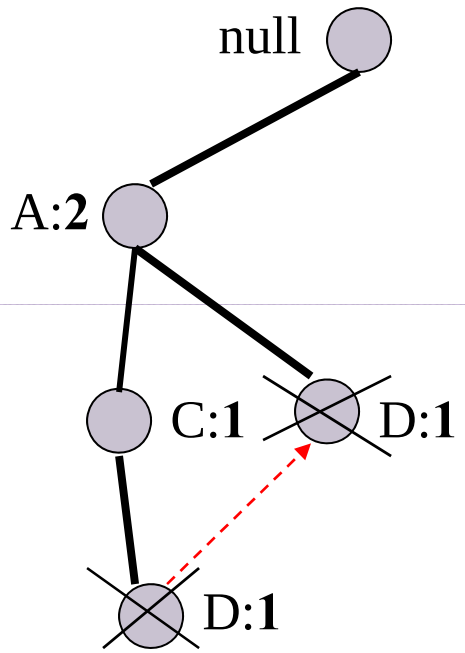
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {D, E} (για τα A,B,C)

1. Αλλαγή υποστήριξης

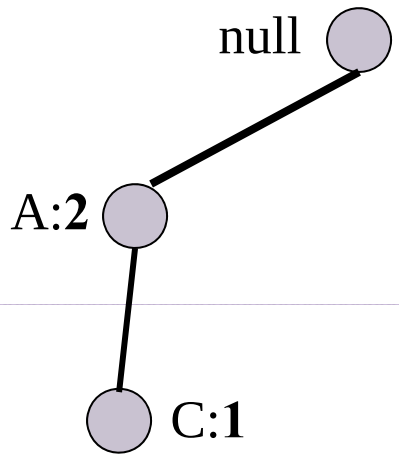
2. Περικοπές κόμβων



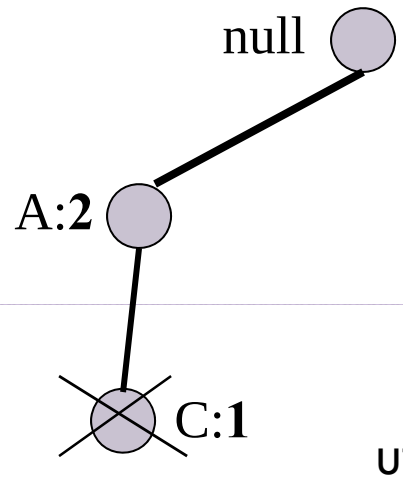
1. Αλλαγή υποστήριξης



2. Περικοπές κόμβων



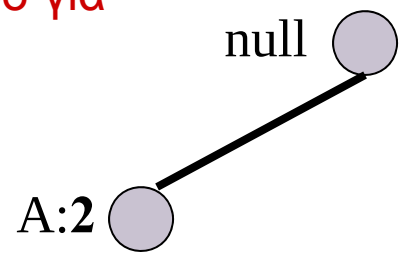
2. Περικοπές κόμβων



Μικρή υποστήριξη

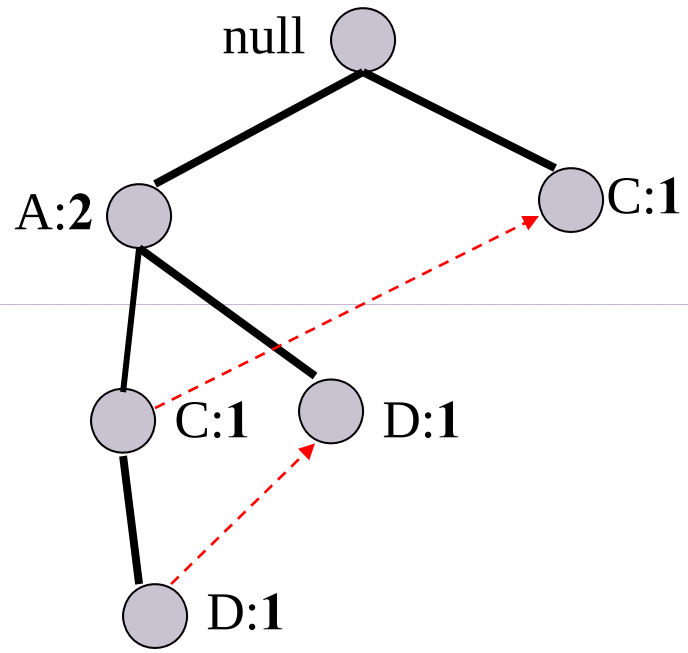
2. Περικοπές κόμβων

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



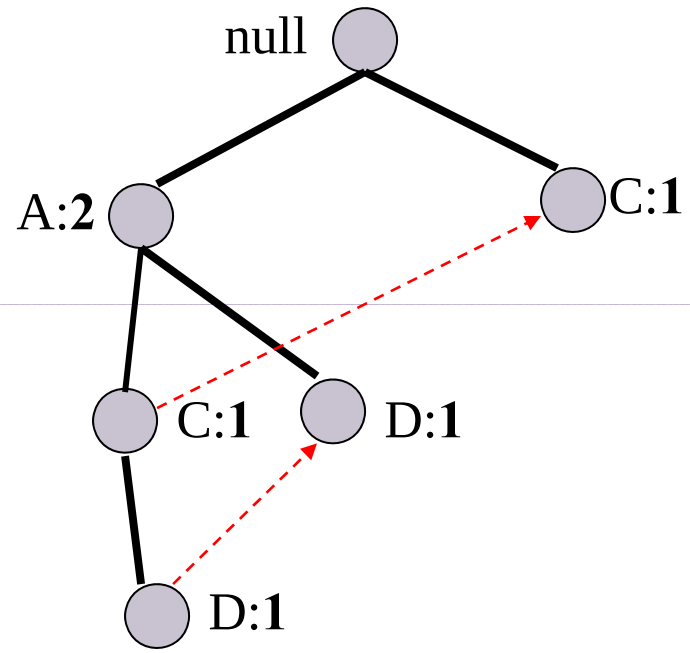
Υποστήριξη του A είναι ≥ minsup -> {A, D, E} συχνό

Αφού έχουμε μόνο έναν κόμβο, επιστροφή στο επόμενο υποπρόβλημα



Υπο-συνθήκη FP-δέντρο για το E

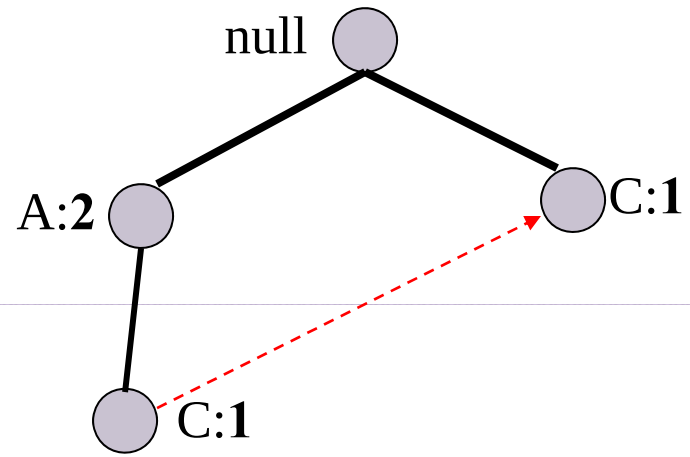
Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, **{C, E},** {A, E}



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

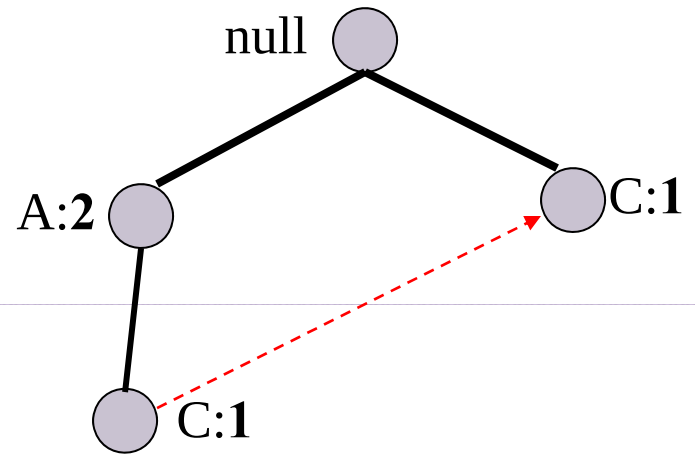
Προθεματικά Μονοπάτια (prefix paths)



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)

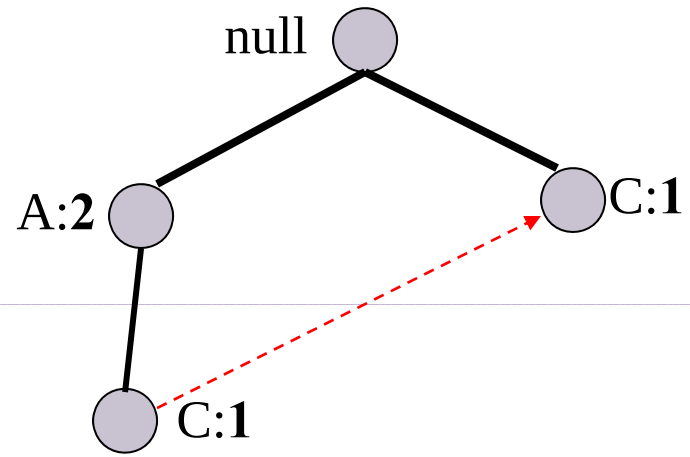


Βρες την υποστήριξη του {C, E}

Πως;

Ακολούθησε τους συνδέσμους αθροίζοντας 1+1=2≥2

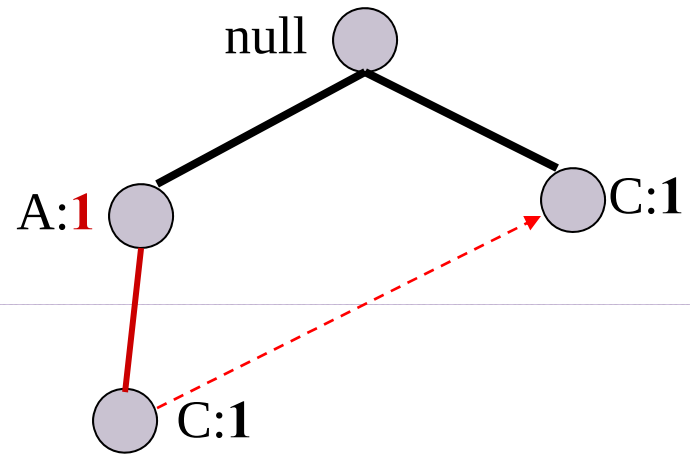
Οπότε {C, E} συχνό



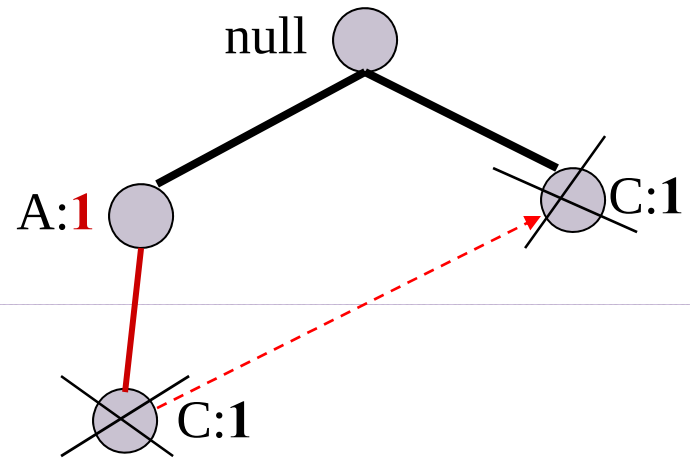
Κατασκεύασε το υπο-συνθήκη FP-δέντρο για το {C, E} (για τα A,B)

1. Αλλαγή υποστήριξης

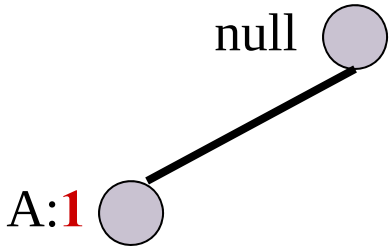
2. Περικοπές κόμβων



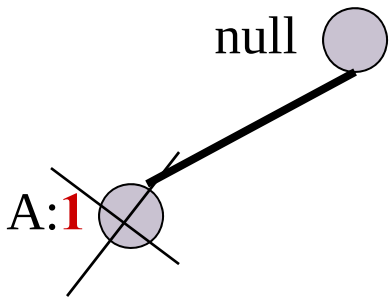
1. Αλλαγή υποστήριξης



2. Περικοπή Κόμβων



2. Περικοπή Κόμβων



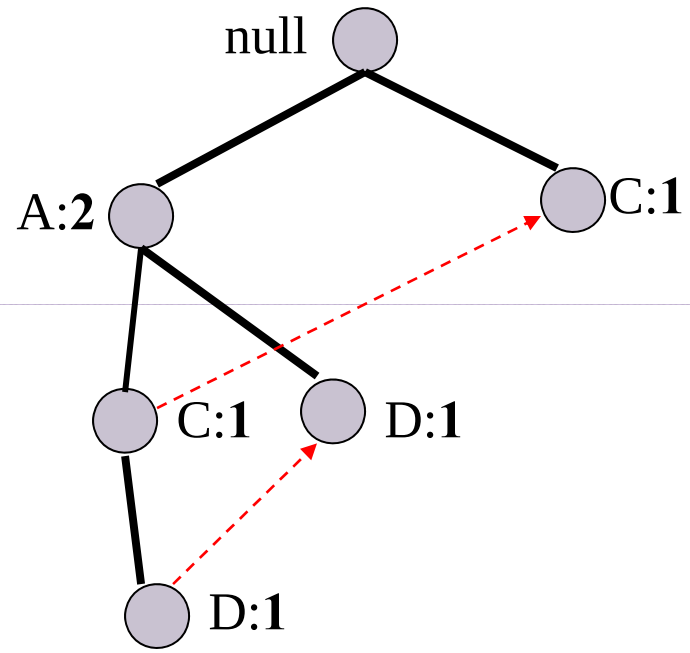
2. Περικοπή Κόμβων

Αλγόριθμος FP-Growth

null

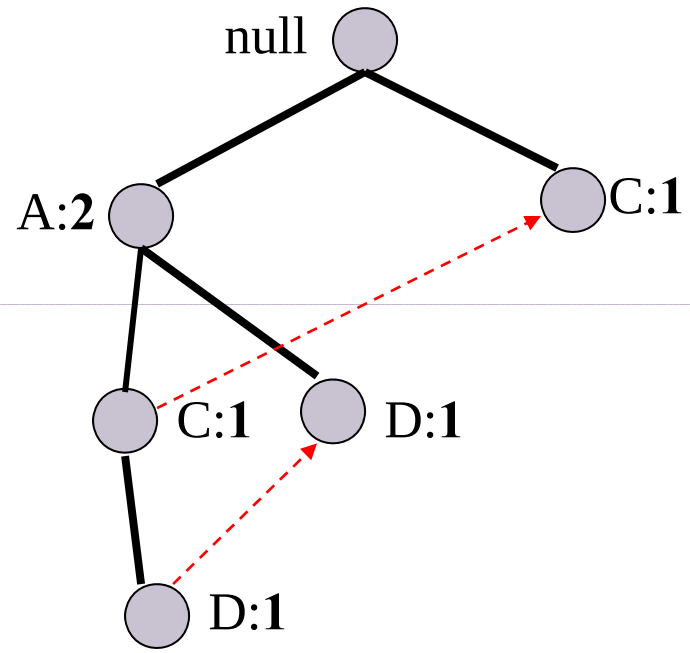
2. Περικοπή Κόμβων

Άρα, επιστροφή στο επόμενο υποπρόβλημα



Υπο-συνθήκη FP-δέντρο για το E

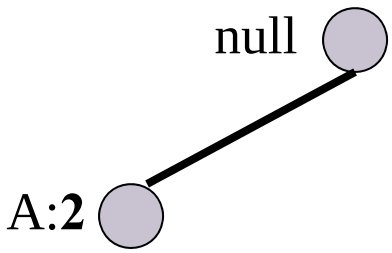
Ο αλγόριθμος επαναλαμβάνεται για το {D, E}, {C, E}, **{A, E}**



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

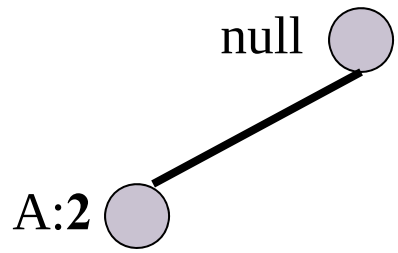
Προθεματικά Μονοπάτια (prefix paths)



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)



Βρες την υποστήριξη του {A, E}

Οπότε {A, E} συχνό

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

Άρα για το Ε

Έχουμε τα εξής συχνά στοιχειοσύνολα

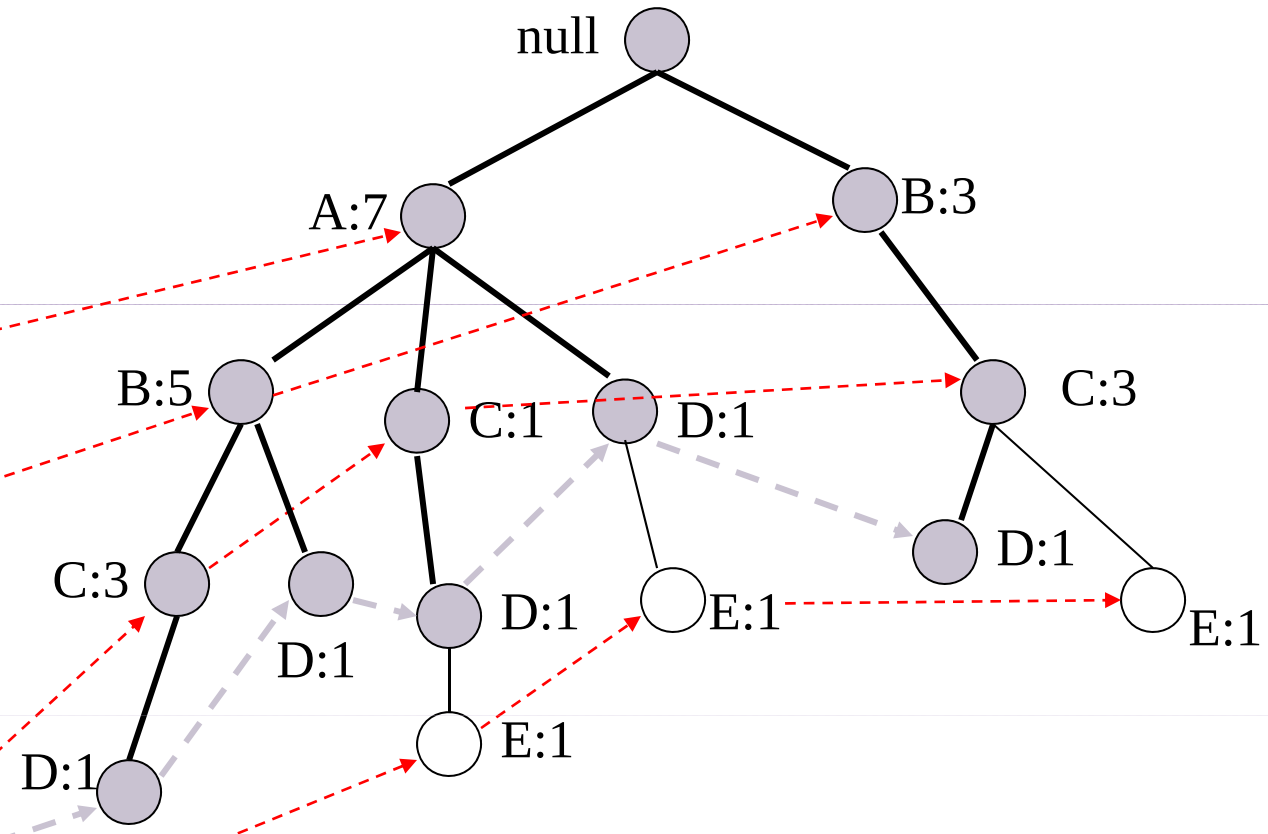
{E} {D, E} {A, D, E} {C, E} {A, E}

Συνεχίζουμε για το D

Για το **D**

Header table

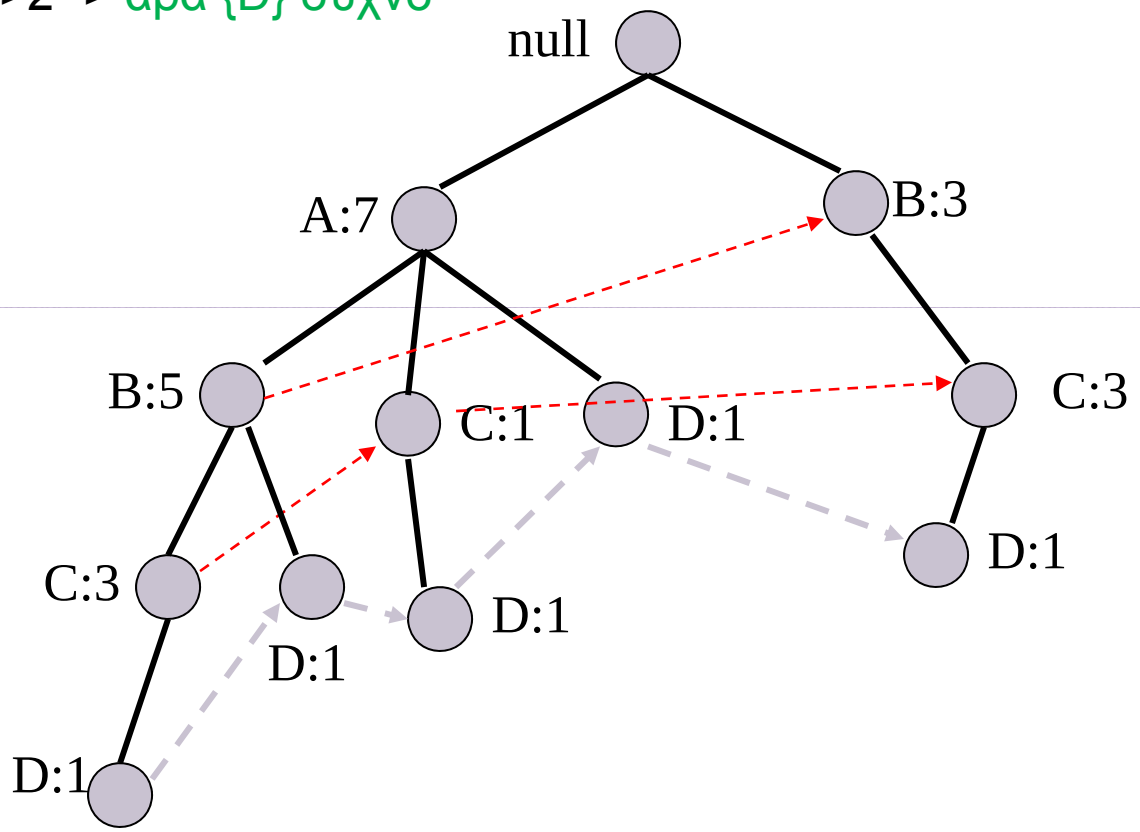
Item	Pointer
A	
B	
C	
D	
E	



Φάση 1

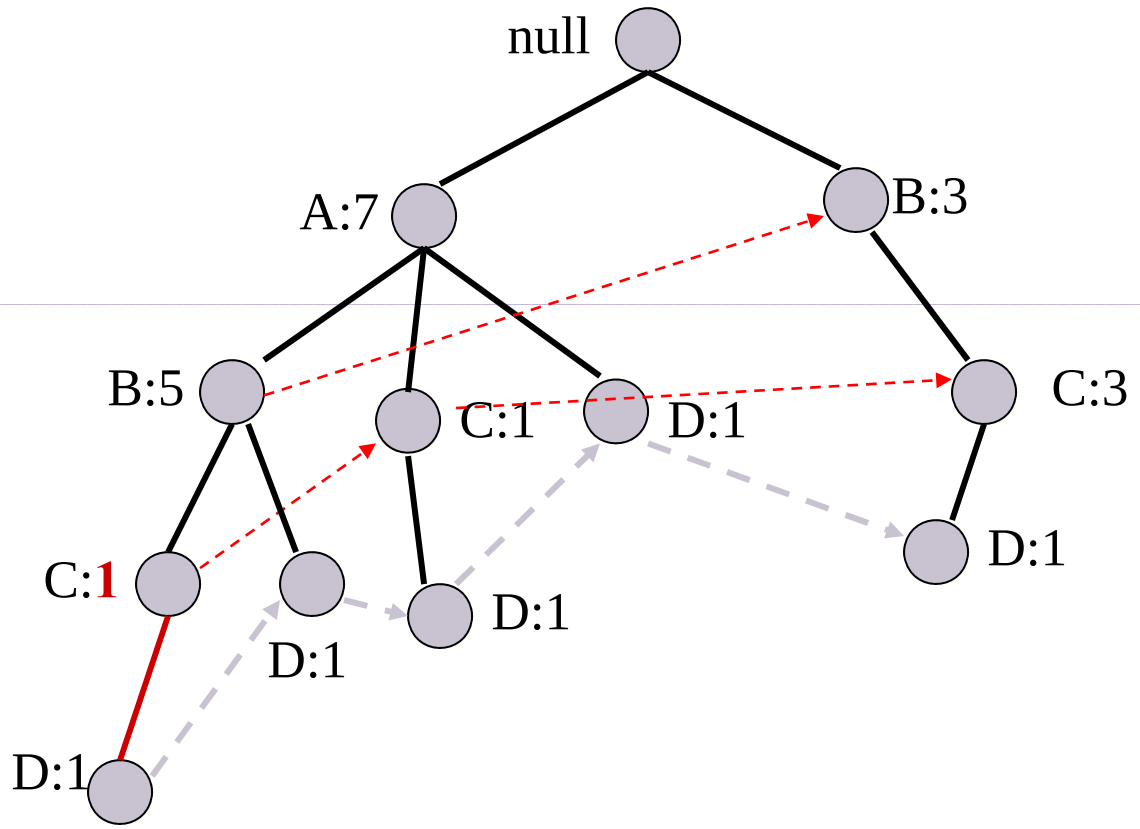
Όλα τα προθεματικά μονοπάτια που περιέχουν το D

Υποστήριξη 5>2 -> άρα {D} συχνό

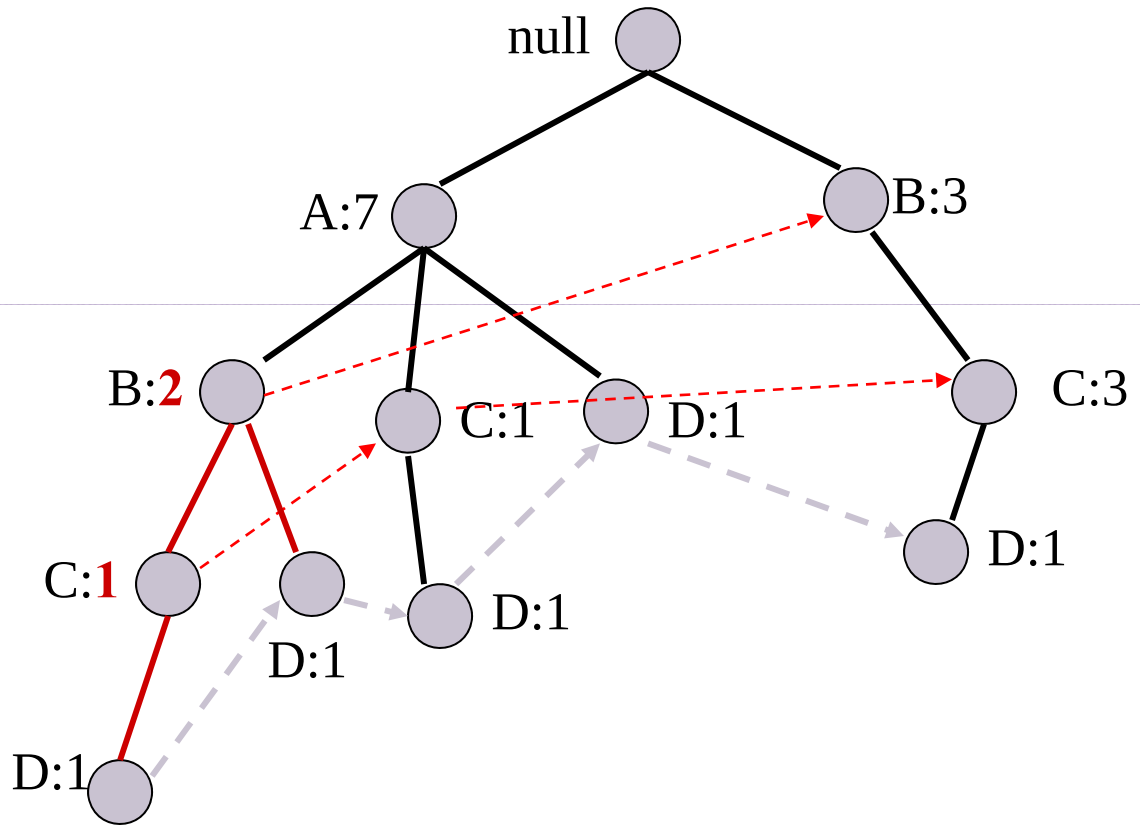


Μετατροπή του προθεματικού δέντρου σε FP-δέντρο υπό συνθήκη

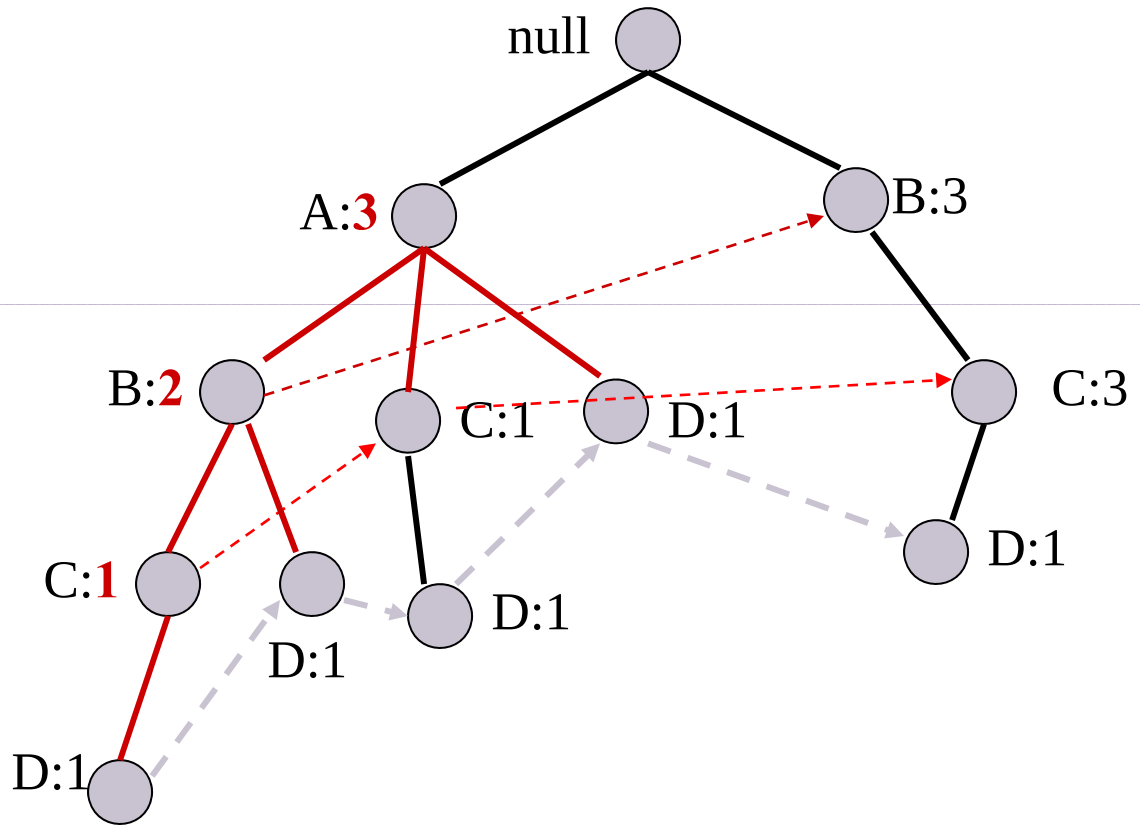
1. Αλλαγή υποστήριξης



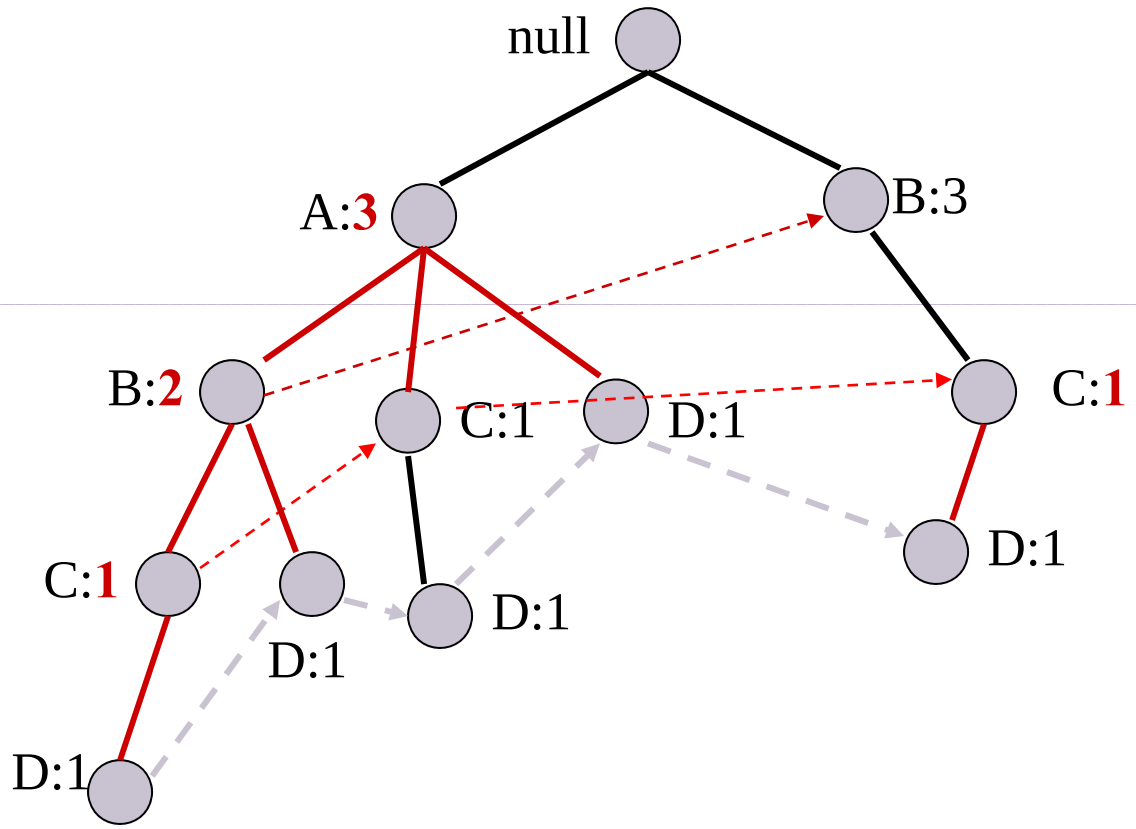
1. Αλλαγή υποστήριξης



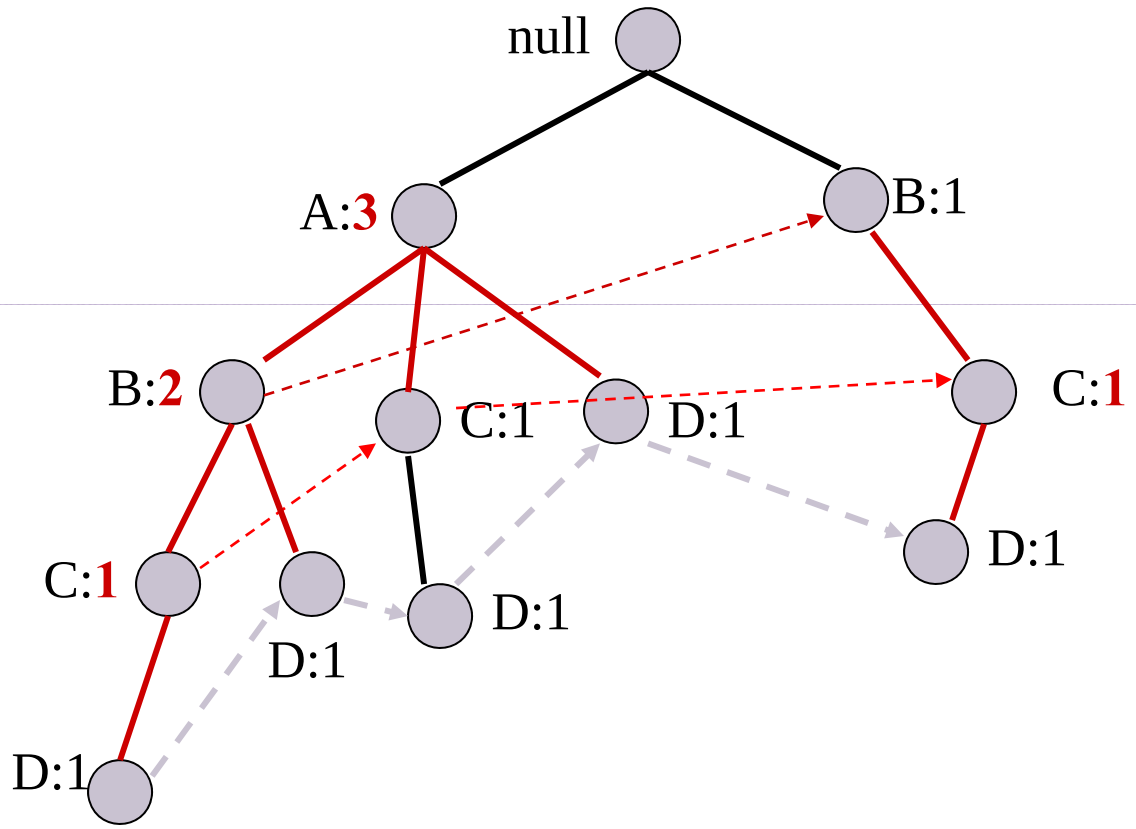
1. Αλλαγή υποστήριξης



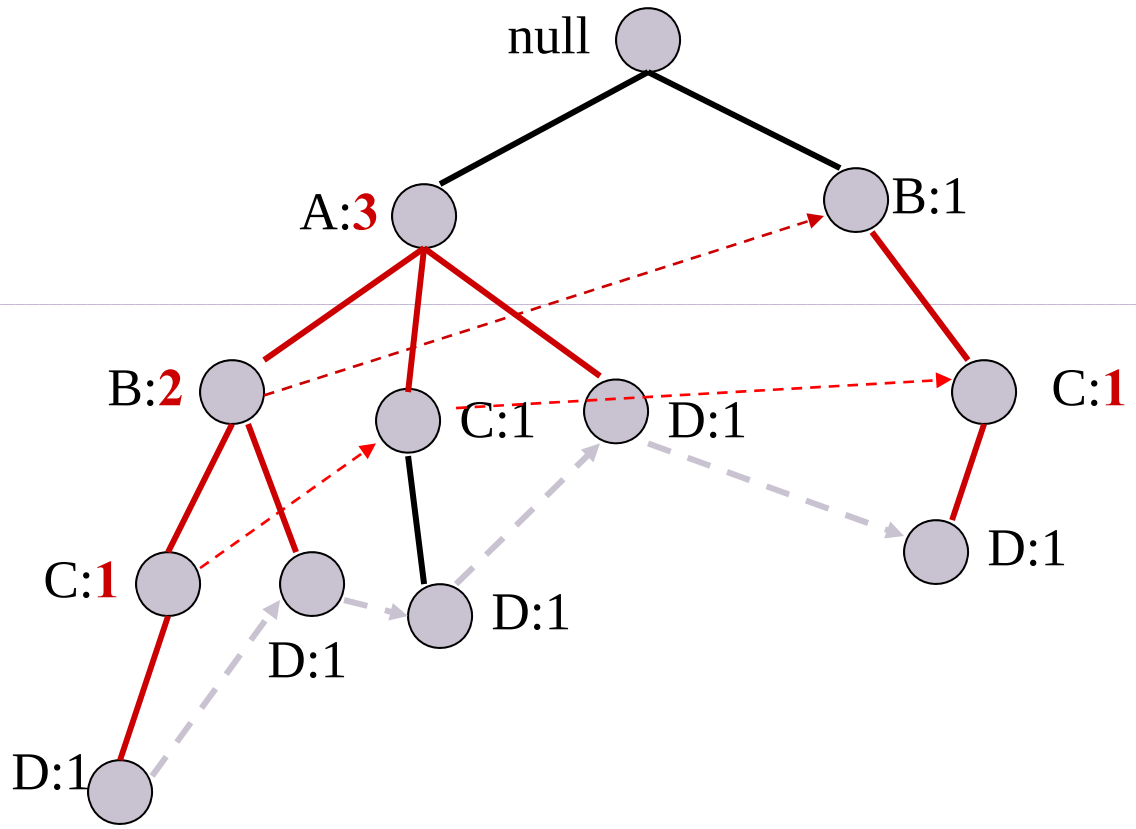
1. Αλλαγή υποστήριξης



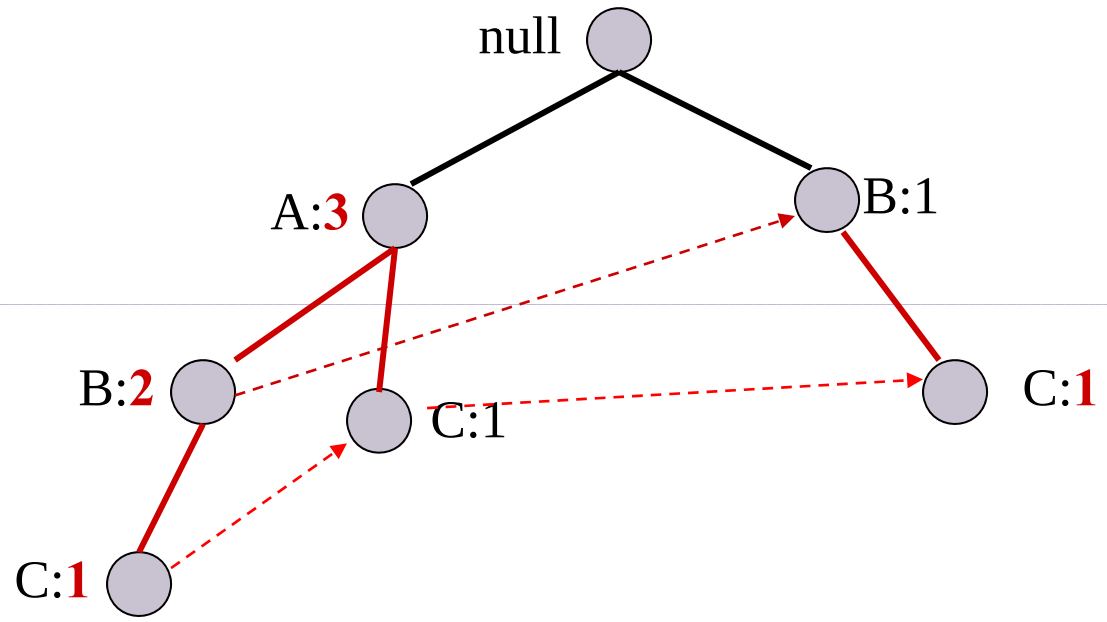
1. Αλλαγή υποστήριξης



2. Περικοπή Κόμβων



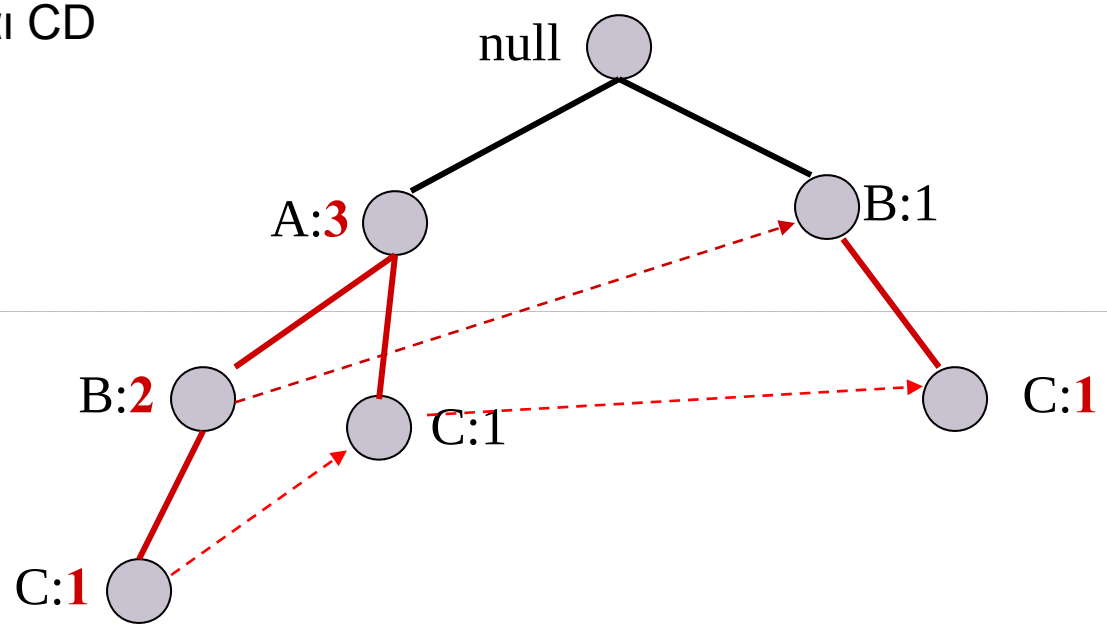
2. Περικοπή Κόμβων



Προθεματικά δέντρα και υποσυνθήκη δέντρα

Για τα AD, BD και CD

κοκ



Παρατηρήσεις

- Παράδειγμα τεχνικής διαίρει-και-βασίλευε
- Σε κάθε αναδρομικό βήμα, λύνεται και ένα υπο-πρόβλημα:
  - Κατασκευάζεται το προθεματικό δέντρο
  - Υπολογίζεται η νέα υποστήριξη για τους κόμβους του
  - Περικόβονται οι κόμβοι με μικρή υποστήριξη
- Επειδή τα υποπροβλήματα είναι ξένα μεταξύ τους, δεν δημιουργούνται τα ίδια συχνά στοιχειοσύνολα δυο φορές
- Ο υπολογισμός της υποστήριξης είναι αποδοτικός – γίνεται ταυτόχρονα με τη δημιουργία των συχνών στοιχειοσυνόλων
- Η απόδοση του FP-Growth εξαρτάται από τον παράγοντα συμπίεσης του συνόλου των δεδομένων (compaction factor)
- Αν τα τελικά δέντρα είναι «θαμνώδη» (bushy) τότε δε δουλεύει καλά, αυξάνεται ο αριθμός των υποπροβλημάτων (οι αναδρομικές κλήσεις)

Αποτίμηση Κανόνων Συσχέτισης

Αποτίμηση Κανόνων Συσχέτισης

Παράγουν πάρα πολλούς κανόνες που συχνά είναι μη ενδιαφέροντες ή πλεονάζοντες (περιττοί)

Πλεονάζοντες αν {A, B, C} → {D} και {A,B} → {D} έχουν την ίδια υποστήριξη & εμπιστοσύνη

Μέτρα ενδιαφέροντος (interestingness) χρησιμοποιούνται για να ελαττώσουν (prune) ή να ιεραρχήσουν (rank) τα παραγόμενα πρότυπα

Χρησιμοποιούνται σε διάφορα στάδια της διαδικασίας ανάκτησης γνώσης

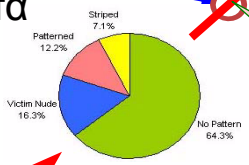
Μέτρηση Ενδιαφέροντος

Μέτρα Ενδιαφέροντος

Εφαρμογές της μέτρησης του ενδιαφέροντος(σε διάφορα στάδια)

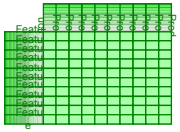
Γνώση

Πρότυπα



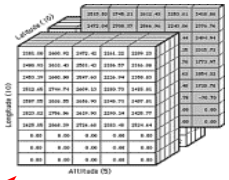
Μετά-επεξεργασία

Προ-επεξεργασμένα Δεδομένα



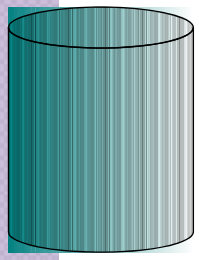
Εξόρυξη

Επιλεγμένα Δεδομένα



Προ-επεξεργασία

Δεδομένα



Επιλογή

Αποτίμηση Κανόνων Συσχέτισης

Γενικά: αντικειμενικά (objective) και υποκειμενικά (subjective) μέτρα ενδιαφέροντος

Ας δούμε πρώτα μερικά αντικειμενικά κριτήρια:

Στην αρχική διατύπωση του προβλήματος της εξόρυξης κανόνων συσχέτισης χρησιμοποιήθηκαν ως μέτρα μόνο η *υποστήριξη* και η *εμπιστοσύνη*

Γενικά συνήθως βασίζονται σε μετρήσεις της συχνότητας εμφάνισης που δίνονται μέσω ενός πίνακα “contingency” (συνάφειας)

Μέτρηση Ενδιαφέροντος

Υπολογισμός του Μέτρου Ενδιαφέροντος (αντικειμενικά μέτρα)

Έστω ένας κανόνας, $X \rightarrow Y$, η πληροφορία που χρειάζεται για τον υπολογισμό της εμπιστοσύνης και υποστήριξης του κανόνα μπορεί να υπολογιστεί από τον **contingency table**

Contingency table (πίνακας συνάφειας)

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

Μέτρηση συχνότητας εμφάνισης

f_{11}: support of X and Y

f_{10}: support of X and $\overline{Y}$

f_{01}: support of $\overline{X}$ and Y

f_{00}: support of $\overline{X}$ and $\overline{Y}$

f_{11} πόσο συχνά εμφανίζεται το X και το Y (support count)

f_{+1} μετρητής υποστήριξης (support count) του Y

Χρησιμοποιείται για τον ορισμό διαφόρων μέτρων

Μειονεκτήματα της Εμπιστοσύνης

Μεγάλες τιμές υποστήριξης μπορεί να «διώξουν» ενδιαφέροντες κανόνες. Τι γίνεται με την εμπιστοσύνη;

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

Ποια είναι μια καλή τιμή για την εμπιστοσύνη;

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

Κανόνας Συσχέτισης: Tea → Coffee

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

Ενώ ο κανόνας έχει υψηλή εμπιστοσύνη, ο κανόνας είναι παραπλανητικός

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

P(Coffee) =90/100= 0.9

Αγνοεί την υποστήριξη του RHS

Μέτρηση Ενδιαφέροντος

Εξαιτίας τέτοιων προβλημάτων της υποστήριξης/εμπιστοσύνης,

Έχουν προταθεί πολλά αντικειμενικά μέτρα για τη μέτρηση του ενδιαφέροντος των κανόνων, που στηρίζονται κυρίως στην έννοια της στατιστικής ανεξαρτησίας

Ας δούμε ένα παράδειγμα

Μέτρα βασισμένα στη Στατιστική

Στατιστική Ανεξαρτησία

Πληθυσμός 1000 σπουδαστών

- 600 σπουδαστές ξέρουν κολύμπι (S)
- 700 σπουδαστές ξέρουν ποδήλατο (B)
- 420 σπουδαστές 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
- $P(S \wedge B) > P(S) \times P(B)$ => Positively correlated (θετική συσχέτιση)
- $P(S \wedge B) < P(S) \times P(B)$ => Negatively correlated (αρνητική συσχέτιση)

Μέτρα βασισμένα στη Στατιστική

Μέτρα που λαμβάνουν υπ΄ όψιν τους τη στατιστική εξάρτηση

Για τη συσχέτιση: X → Y

	Y	Y	
X	f_{11}	f_{10}	f_{1+}
X	f_{01}	f_{00}	f_{o+}
	f_{+1}	f_{+0}	|T|

$$Lift = \frac{P(Y \mid X)}{P(Y)} = \frac{f_{11}/f_{+1}}{f_{+1}} = \frac{\sup(XUY)}{\sup(X)\sup(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)} = \frac{|T|\, f_{11}}{f_{1+}\, f_{+1}}$$

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

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}} = \frac{f_{11} f_{00} - f_{01} f_{10}}{\sqrt{f_{1+}\, f_{+1} f_{0+}\, f_{+0}}}$$

Παράδειγμα: Lift/Interest

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

Κανόνας συσχέτιση: Tea → Coffee

Εμπιστοσύνη= P(Coffee|Tea) = 0.75

αλλά P(Coffee) = 0.9

⇒ Interest = 0.15/(0.9*0.2)= 0.8333 (< 1, άρα αρνητικά συσχετιζόμενα)

$$Interest = \frac{P(X,Y)}{P(X)P(Y)} = \frac{|T|\, f_{11}}{f_{1+}\, f_{+1}}$$

Μέτρα βασισμένα στη Στατιστική

Μειονεκτήματα του Lift & Interest

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

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

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

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

Μεγαλύτερο αν και σπάνια εμφανίζονται μαζί

c = 10/100 = 0.1

s = 1

c = 90/100 = 0.9

s = 1

c (confidence – εμπιστοσύνη)

s (support – υποστήριξη)

ϕ-Coefficient

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}} = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

Κανονικοποιημένη τιμή μεταξύ του -1 και 1

Δυαδική εκδοχή του Pearson's coefficient

- 0: στατιστική ανεξαρτησία
- -1: τέλεια αρνητική συσχέτιση
- 1: τέλεια θετική συσχέτιση

- Είναι κατάλληλο για μη συμμετρικές (η απουσία και η παρουσία μετρούν το ίδιο)
- Λόγω κανονικοποίησης, αγνοεί το μέγεθος του δείγματος

Μέτρα βασισμένα στη Στατιστική

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}} = \frac{f_{11}f_{00} - f_{01}f_{10}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

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

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

$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

$$\phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

ϕ Coefficient ίδιος και για τους δύο πίνακες

Μέτρα βασισμένα στη Στατιστική

IS-measure

$$IS(X,Y) = \frac{s(X,Y)}{\sqrt{s(X)s(Y)}} = \frac{f_{11}}{\sqrt{f_{1+} f_{+1}}} = \sqrt{I(X,Y)s(x,Y)}$$

- είναι το συνημίτονο αν θεωρηθούν δυαδικές μεταβλητές
- γεωμετρικός μέσος της εμπιστοσύνης του X $\rightarrow$Y και Y $\rightarrow$ X

Στη βιβλιογραφία έχουν προταθεί πολλά μέτρα ανάλογα με την εφαρμογή

Με ποια κριτήρια θα επιλέξουμε ένα καλό μέτρο;

Πως έναν Apriori-style support based pruning επηρεάζει αυτά τα μέτρα;

#	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)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
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)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B), P(A\|B)-P(A))$

Αποτίμηση Κανόνων Συσχέτισης

Σύγκριση Μέτρων

10 παραδείγματα contingency πινάκων:

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 ο πιο ενδιαφέρον, 10 ο 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

Αποτίμηση Κανόνων Συσχέτισης

Ιδιότητες ενός Καλού Μέτρου

Piatetsky-Shapiro:

3 γενικές ιδιότητες που πρέπει να ικανοποιεί ένα καλό μέτρο M:

- M(A, B) = 0 αν τα A και B είναι 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

Ιδιότητες Μέτρων Αποτίμησης

Αλλαγή Διάταξης Μεταβλητών (variable permutation)

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

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

Ισχύει M(A, B) = M(B, A)?

Γενικά χρησιμοποιούμε συμμετρικά μέτρα για στοιχειοσύνολα και μη συμμετρικά για κανόνες

Συμμετρικά (symmetric) μέτρα:

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

Μη συμμετρικά (asymmetric) μέτρα:

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

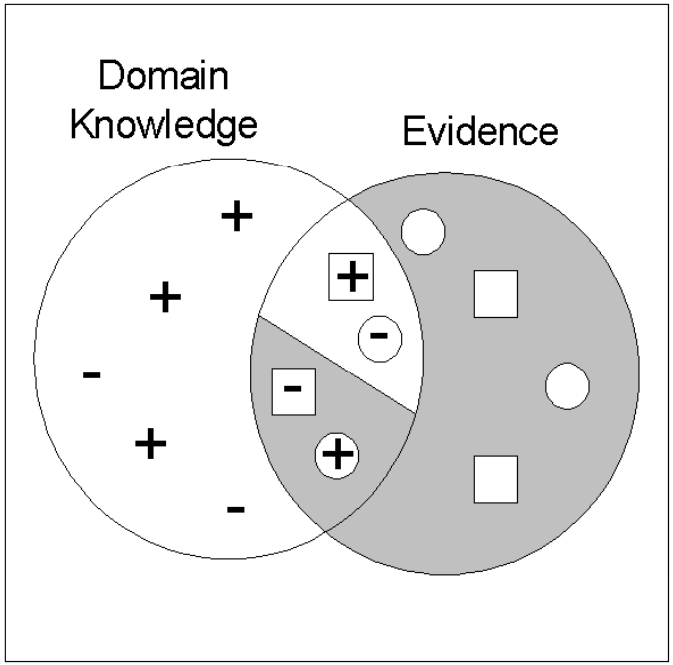
- Αντικειμενικά Μέτρα:
  - Ιεραρχούν τα αποτελέσματα με βάση στατιστικά στοιχεία που υπολογίζονται από τα δεδομένα

    πχ., 21 μετρήσεις συσχέτισης (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).

- Υποκειμενικά Μέτρα:
  - Ιεράρχηση των προτύπων με βάση την <u>ερμηνεία του χρήστη</u>
    - Ένα πρότυπο είναι υποκειμενικά ενδιαφέρον αν είναι σε αντίθεση με αυτό που αναμένει ο χρήστης (Silberschatz & Tuzhilin)
    - Ένα πρότυπο είναι υποκειμενικά ενδιαφέρον αν μπορεί να χρησιμοποιηθεί (Silberschatz & Tuzhilin)

Υποκειμενικά Μέτρα Ενδιαφέροντος

Interestingness (ενδιαφέρον) via Unexpectedness (μη αναμονή)



+ Pattern expected to be frequent

- Pattern expected to be infrequent

Pattern found to be frequent

Pattern found to be infrequent

+ - Expected Patterns

- + Unexpected Patterns

- Χρειάζεται να μοντελοποιήσουμε τι αναμένει ο χρήστης (domain knowledge)
- Χρειάζεται να συνδυάσουμε το τι αναμένεται από τους χρήστες με το τι δίνουν τα δεδομένα (δηλαδή τα πρότυπα που παίρνουμε - evidence)