

Ανάλυση Συσχέτισης ΙΙ

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



Ο Αλγόριθμος FP-Growth



Με λίγα λόγια:

Ο αλγόριθμος χρησιμοποιεί μια *συμπιεσμένη αναπαράσταση της βάσης των συναλλαγών* με τη μορφή ενός **FP-δέντρου**

- Το δέντρο μοιάζει με προθεματικό δέντρο - prefix tree (trie)
- Ο αλγόριθμος κατασκευής διαβάζει μια συναλλαγή τη φορά, απεικονίζει τη συναλλαγή 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** (θα δούμε αργότερα αν κάτι άλλο συμφέρει καλύτερα)

Αρχικά, το δέντρο κενό



null

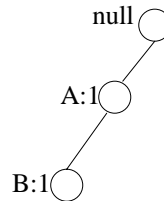
Αλγόριθμος FP-Growth



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 τη συχνότητα εμφάνισης (υποστήριξη) – πόσες δοσοληψίες φτάνουν σε αυτόν

Ετικέτα κόμβου <ΣΤΟΙΧΕΙΟ: ΥΠΟΣΤΗΡΙΞΗ>

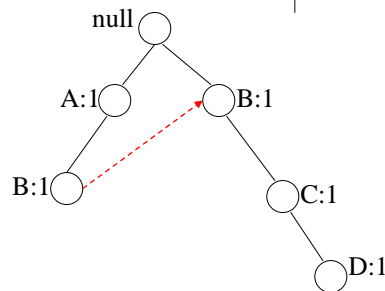
Αλγόριθμος FP-Growth



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 συναλλαγές φτάνουν σε αυτόν

Επίσης, **δείκτες μεταξύ των κόμβων** που αναφέρονται στο ίδιο στοιχείο

Αλγόριθμος FP-Growth

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}

Κατασκευή FP-δέντρου

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

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 7

Αλγόριθμος FP-Growth

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}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3

Πίνακας Δεικτών

Item	Pointer
A	→
B	→
C	→
D	→
E	→

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 8

Αλγόριθμος FP-Growth

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}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

9

Αλγόριθμος FP-Growth

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}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	

10

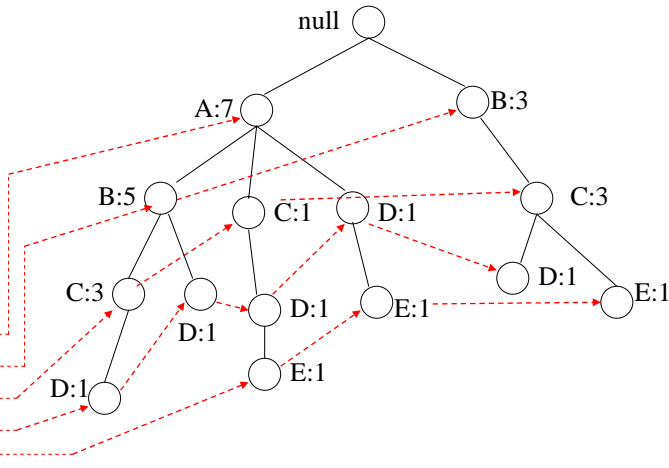


Κατασκευή FP-δέντρου

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

Αλγόριθμος FP-Growth



Κατασκευή FP-δέντρου

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 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,
 $\sigma(A)=7$, $\sigma(B)=8$,
 $\sigma(C)=7$, $\sigma(D)=5$,
 $\sigma(E)=3$

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



Αλγόριθμος εύρεσης συχνών στοιχειοσυνόλων

Είσοδος: FP-δέντρο

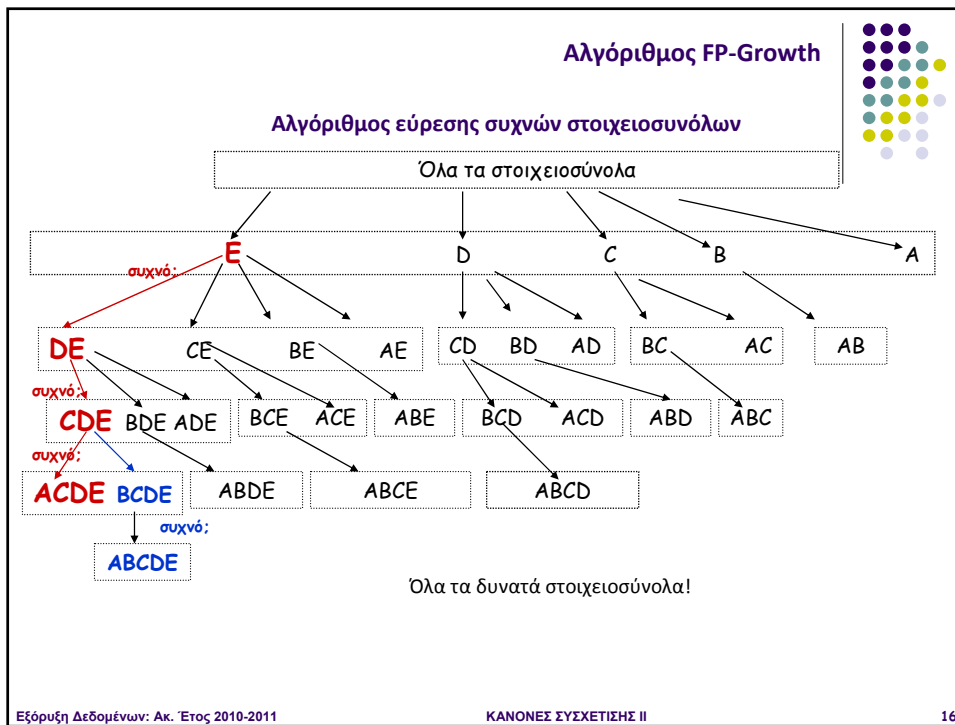
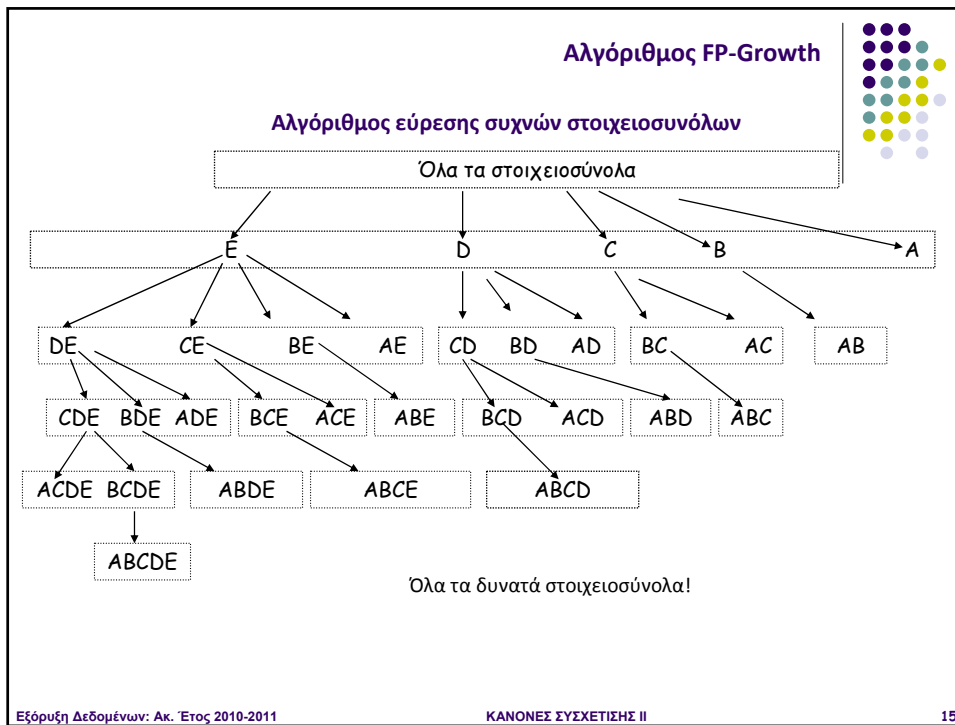
Έξοδος: Συχνά στοιχειοσύνολα και η υποστήριξη τους

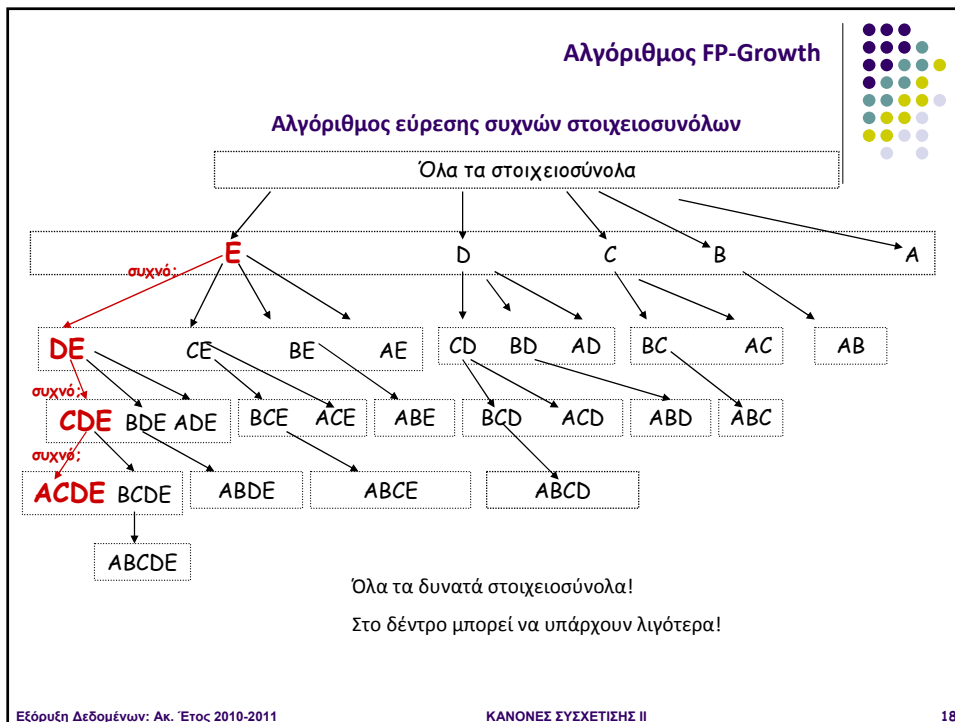
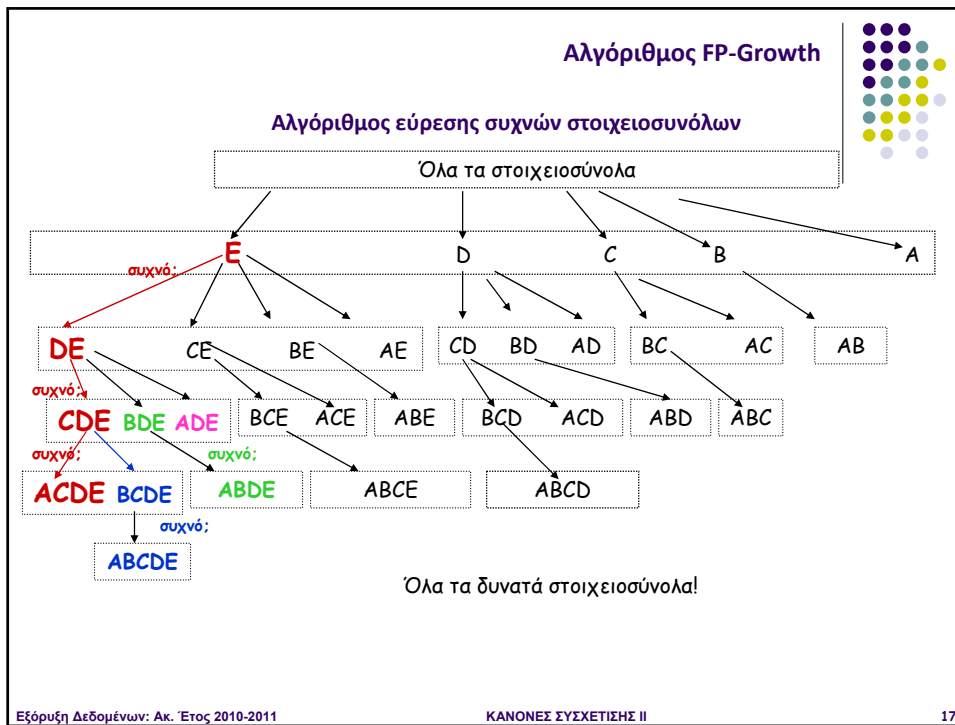
Μέθοδος:

▪ Διαίρει-και-Βασίλευε

ο Χωρίζουμε τα στοιχειοσύνολα σε αυτά που τελειώνουν σε E, D, C, B, A

ο Μετά αυτά που τελειώνουν σε E σε αυτά σε DE, CE, BE, AE κοκ





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	

Bottom-up traversal του δέντρου

Αυτά που τελειώνουν σε E, μετά αυτά που τελειώνουν σε D, C, B και τέλος A – suffix-based classes (επίθεμα – κατάληξη)

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011
ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II
19

Αλγόριθμος FP-Growth

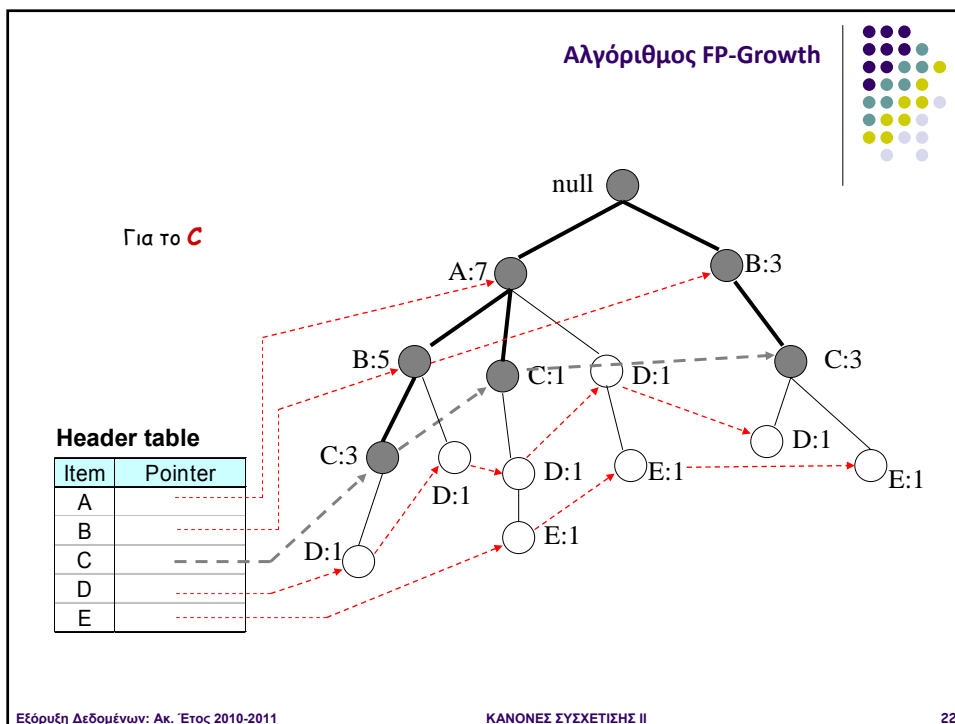
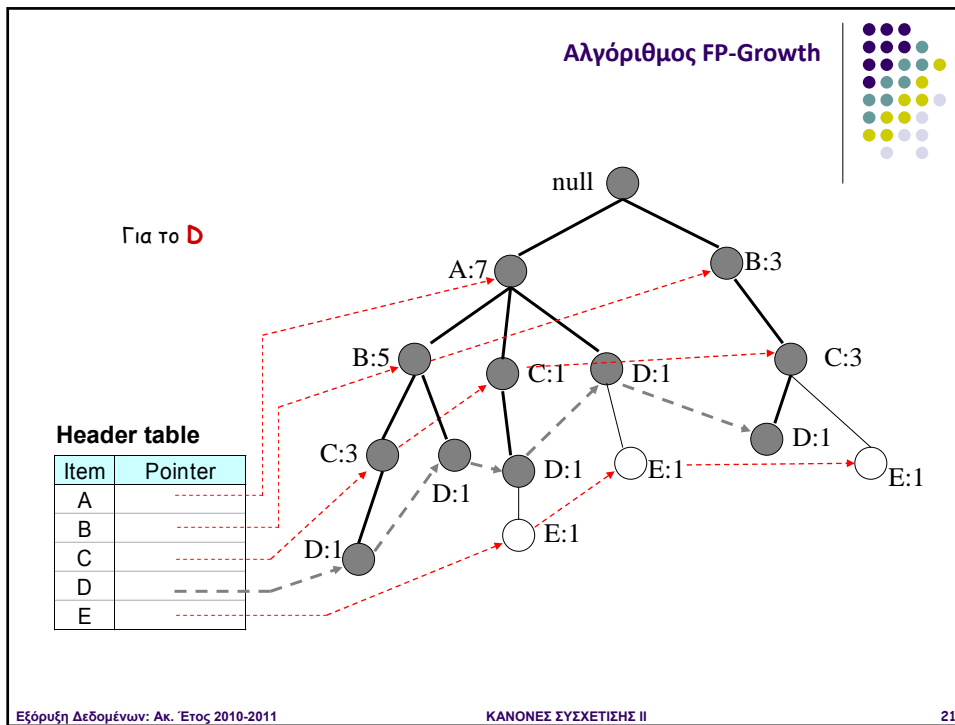
Υποπρόβλημα: Βρες συχνά στοιχειοσύνολα που τελειώνουν σε **E**

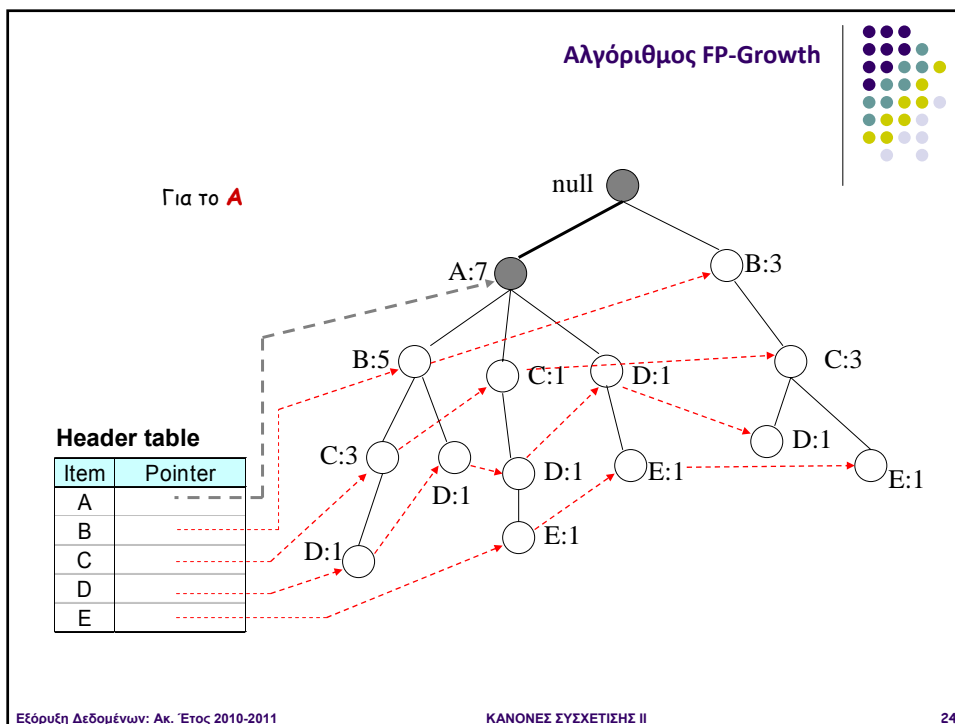
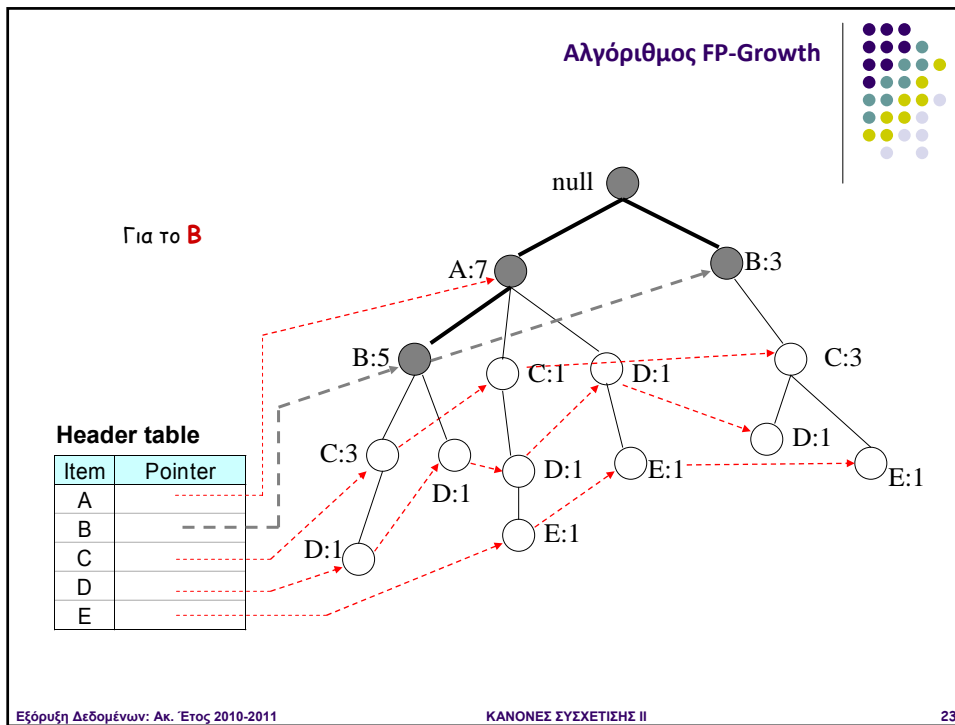
Header table

Item	Pointer
A	
B	
C	
D	
E	

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

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011
ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II
20







Συνοπτικά

Σε κάθε βήμα, για το suffix (επίθεμα) X

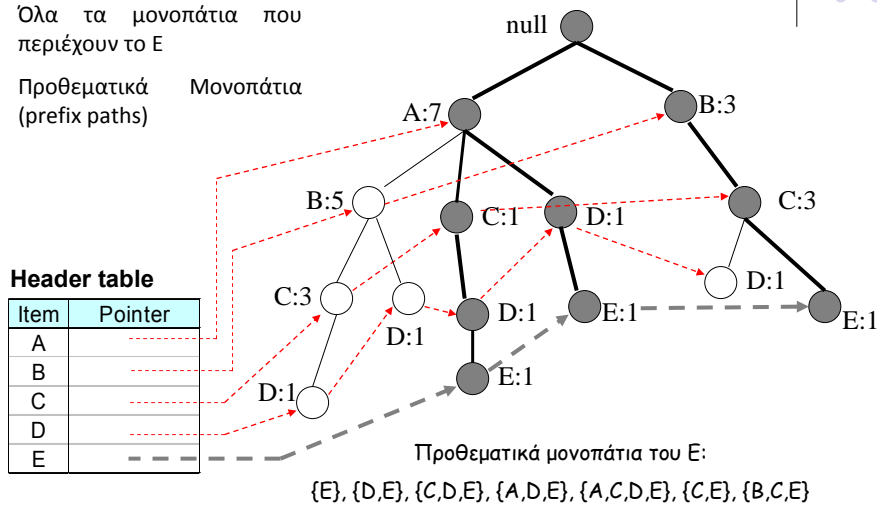
- Φάση 1
 - Κατασκευάζουμε το **προθεματικό δέντρο** για το X και υπολογίζουμε την υποστήριξη χρησιμοποιώντας τον πίνακα

- 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)



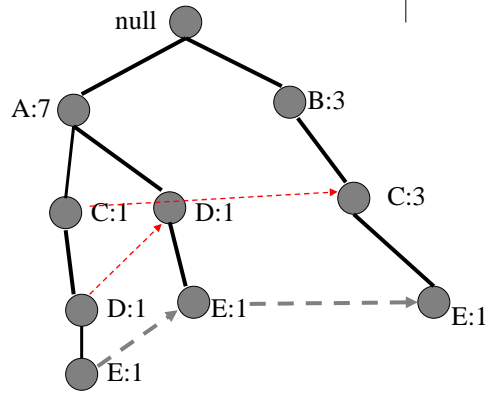
Αλγόριθμος FP-Growth



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



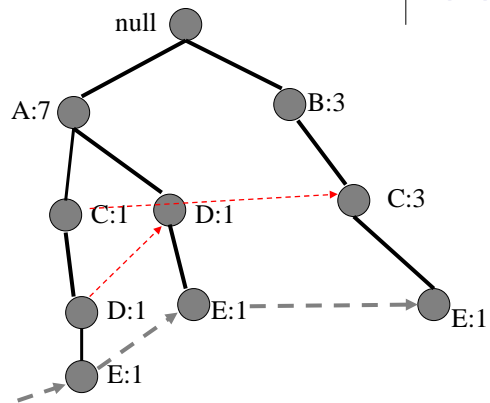
Έστω $\text{minsup} = 2$

Βρες την υποστήριξη του {E}

Πως;

Ακολουθήσε τους συνδέσμους
 αθροίζοντας $1+1+1=3 > 2$

Οπότε {E} συχνό



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

Αλγόριθμος FP-Growth



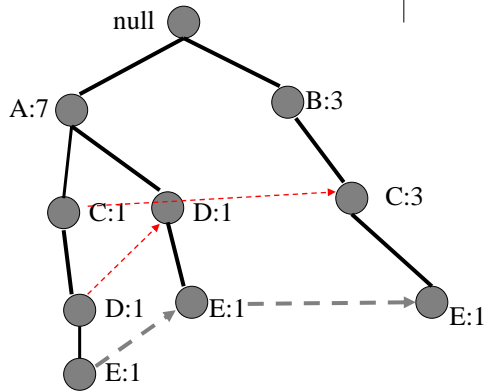
{E} συχνό άρα προχωράμε για DE, CE, BE, AE

Φάση 2

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

Δύο αλλαγές

- (1) Αλλαγή των μετρητών
- (2) Περικοπή



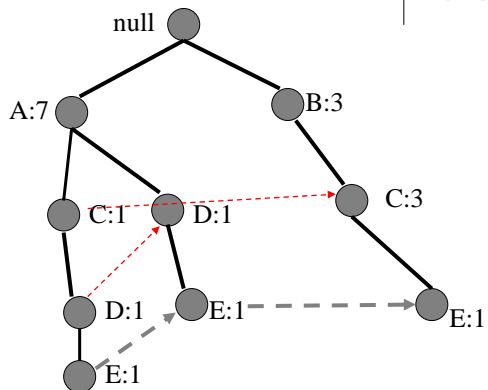
Αλγόριθμος FP-Growth

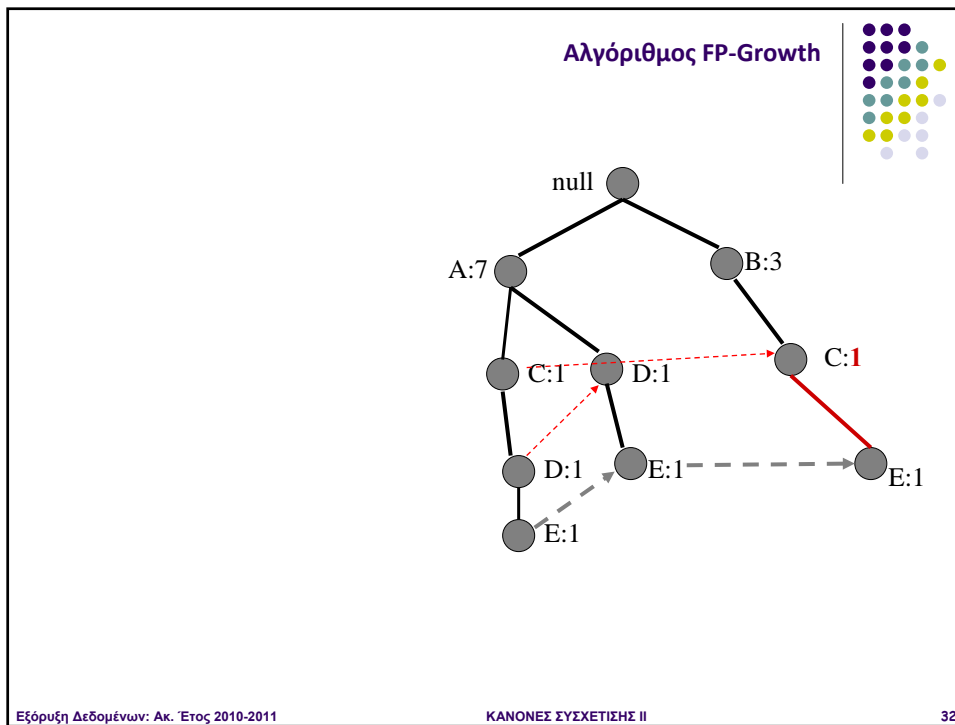
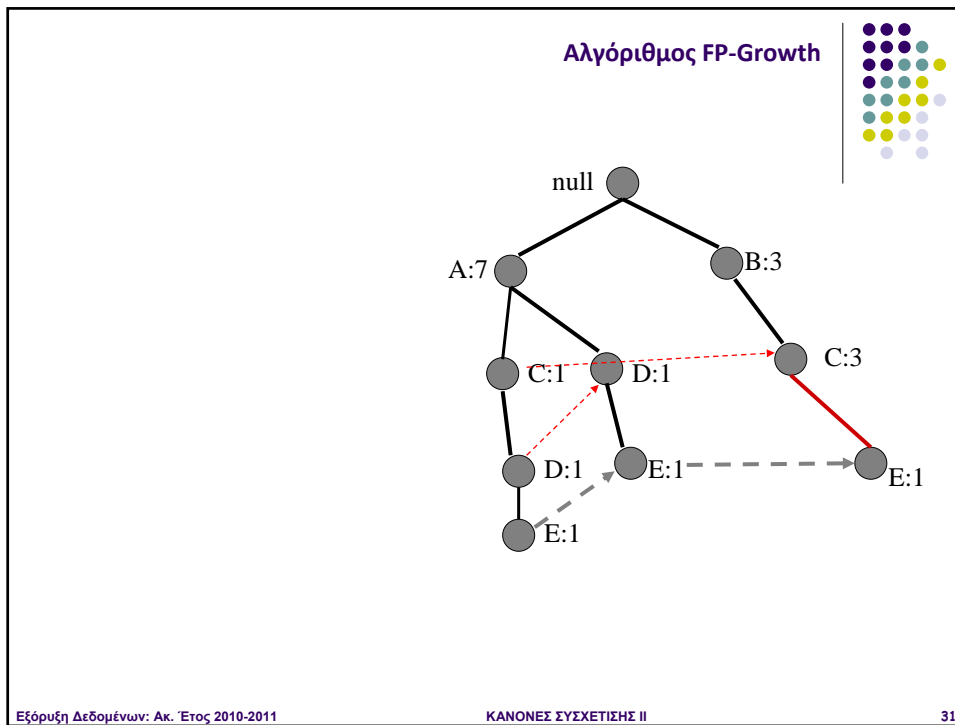


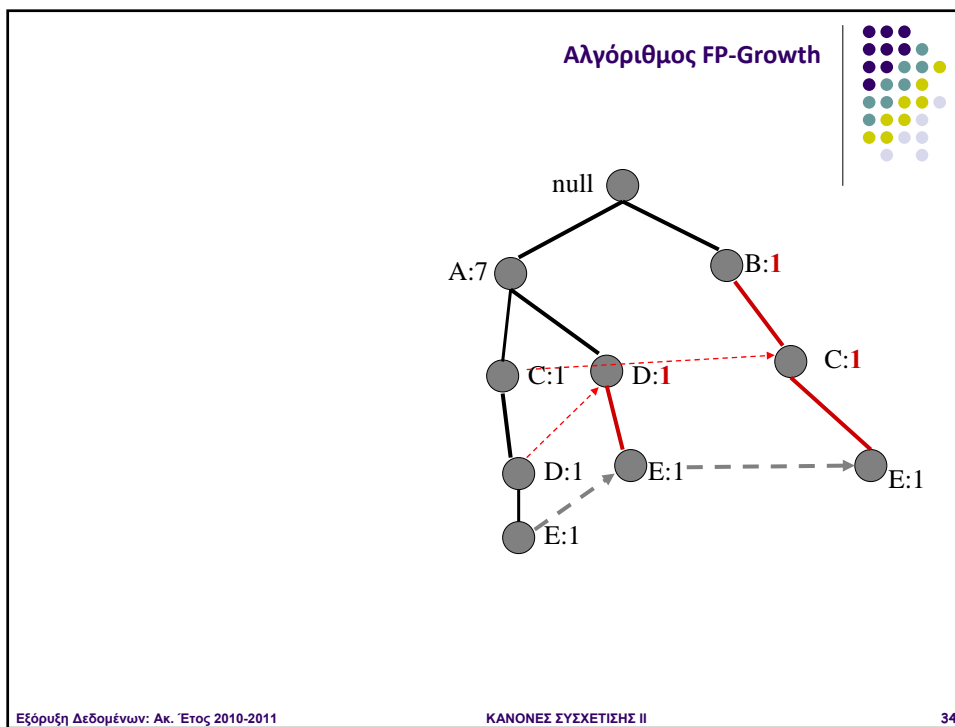
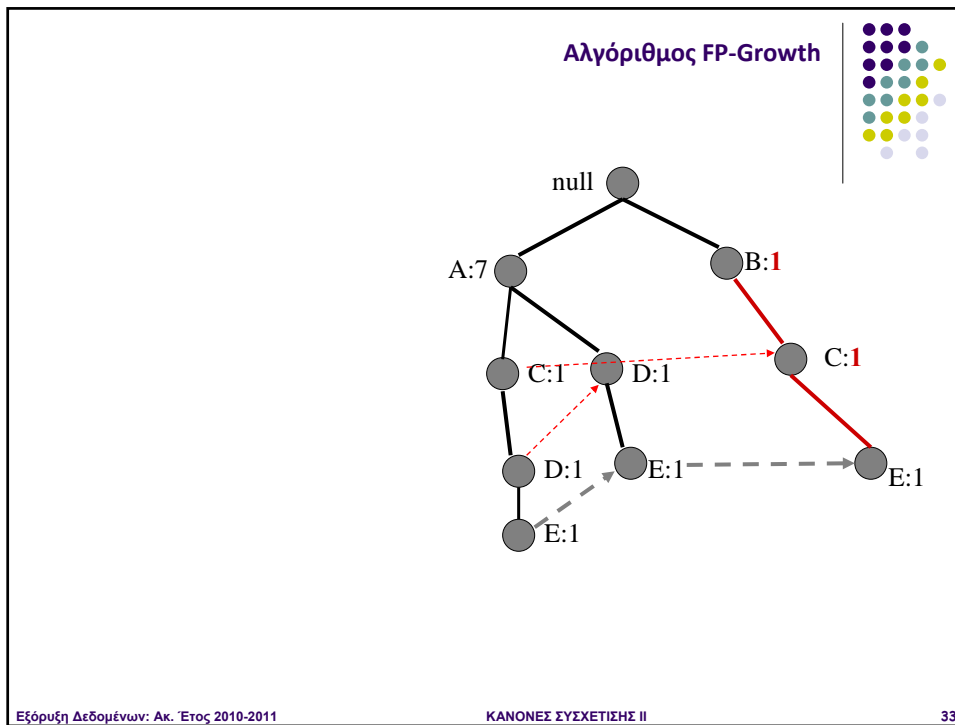
Αλλαγή μετρητών

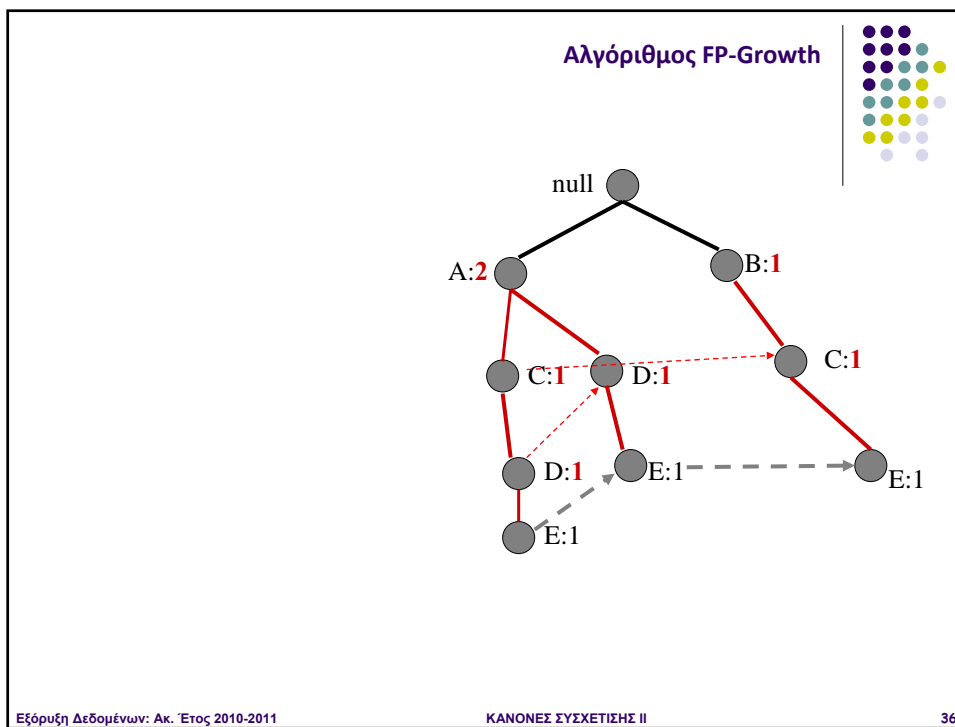
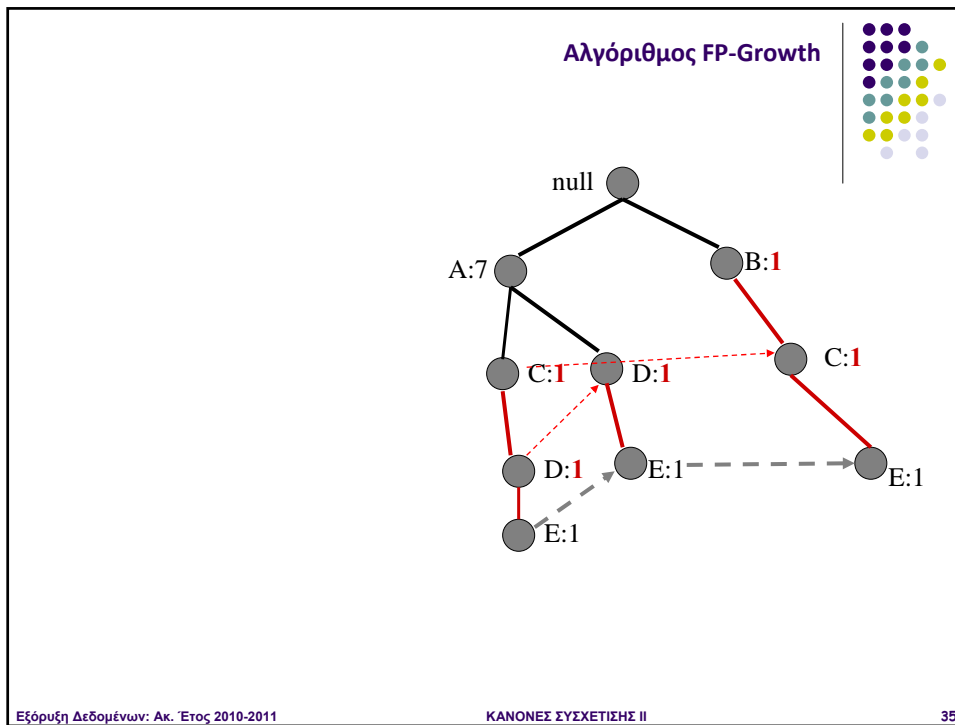
Οι μετρητές σε κάποιους κόμβους περιλαμβάνουν συναλλαγές που δεν έχουν το E

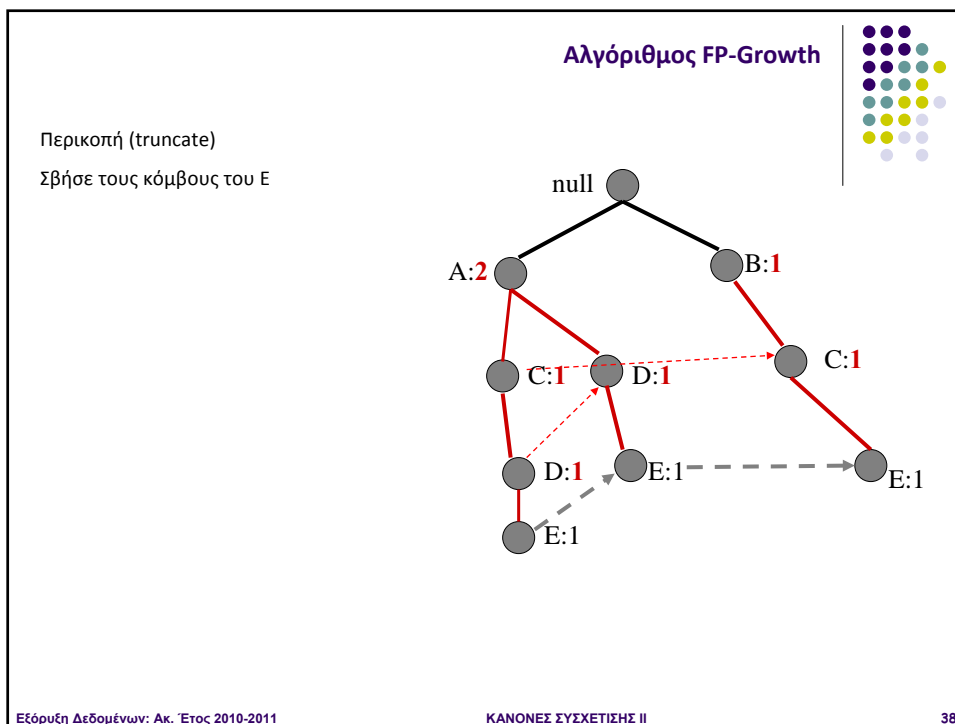
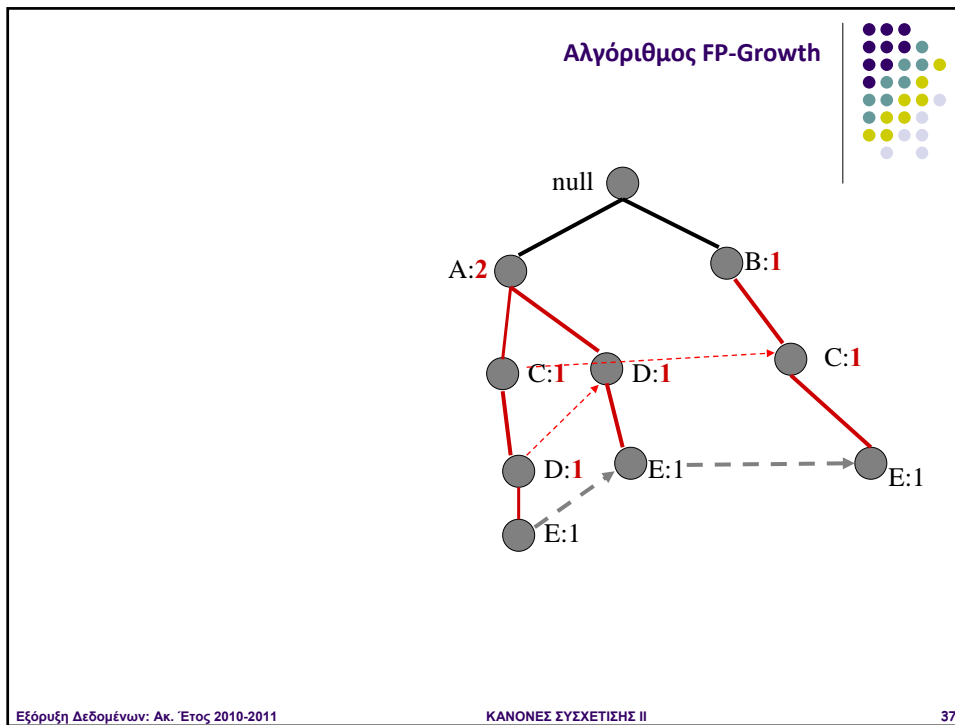
Πχ στο null->B->C->E μετράμε και την {B, C}







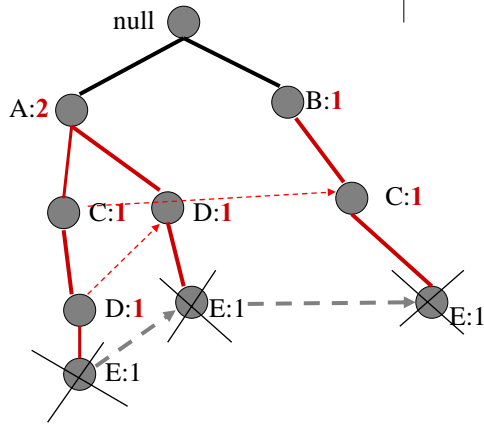




Αλγόριθμος FP-Growth



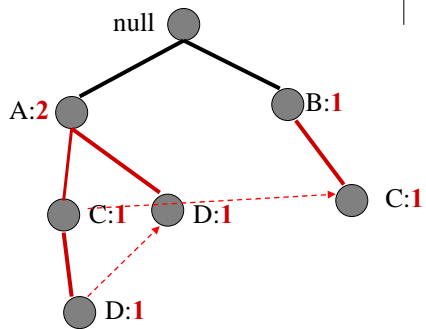
Περικοπή (truncate)
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth



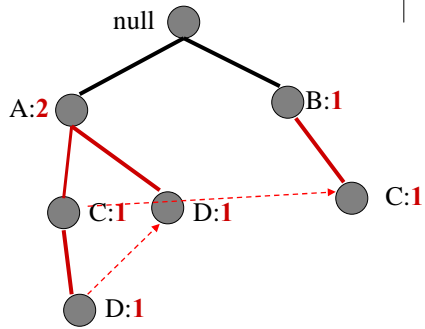
Περικοπή (truncate)
Σβήσε τους κόμβους του E



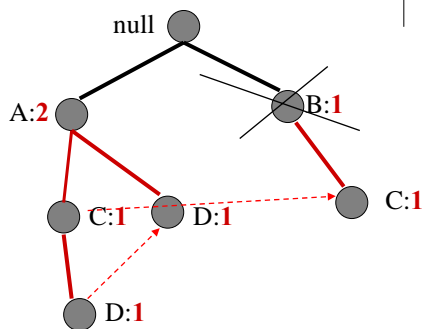
Αλγόριθμος FP-Growth



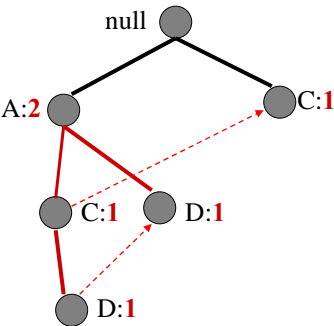
Πιθανή περαιτέρω περικοπή
Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης
Πχ το B -> περικοπή
Αυτό σημαίνει ότι το B εμφανίζεται μαζί με το E λιγότερο από minsup φορές




Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth



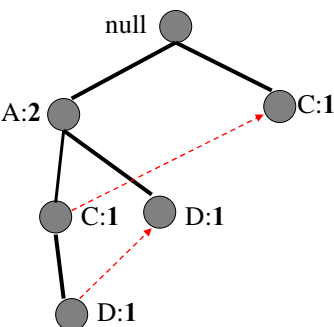



Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 43

Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

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





Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 44

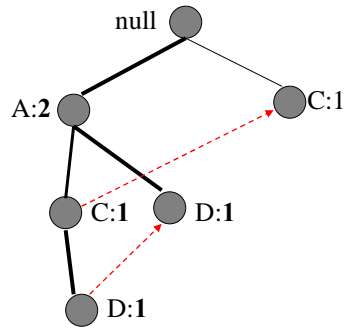
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



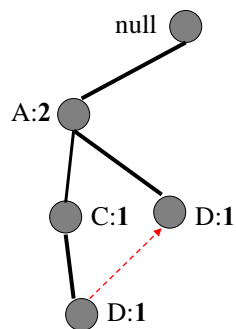
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



Αλγόριθμος FP-Growth

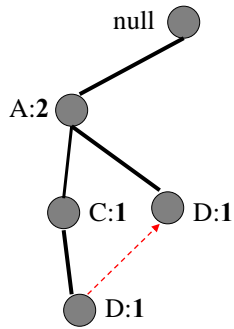


Βρες την υποστήριξη του {D, E}

Πως;

Ακολούθησε τους συνδέσμους
αθροίζοντας $1+1=2 \geq 2$

Οπότε {D, E} συχνό



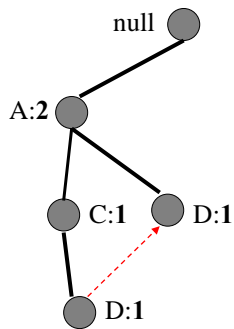
Αλγόριθμος FP-Growth



Φάση 2

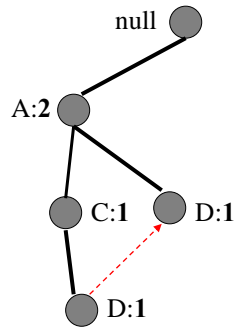
Κατασκεύασε το υπο-συνθήκη FP-δέντρο
για το {D, E}

1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων

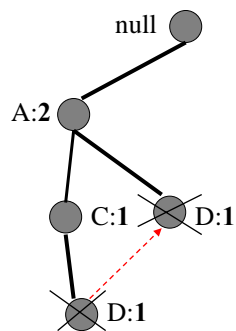




1. Αλλαγή υποστήριξης



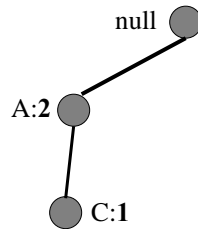
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



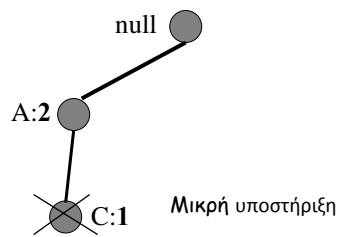
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



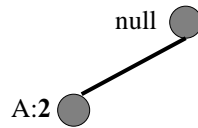
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



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

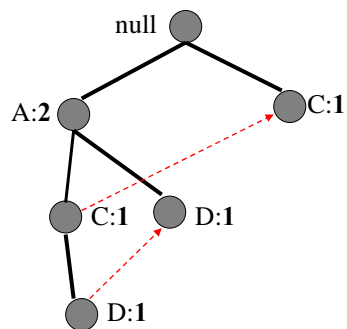


Υποστήριξη του A είναι $\geq \text{minsup}$ \rightarrow {A, D, E} συχνό
Αφού μόνο έναν κόμβο, επιστροφή στο επόμενο υποπρόβλημα

Αλγόριθμος FP-Growth



Υπο-συνθήκη FP-δέντρο για το E
Ο αλγόριθμος επαναλαμβάνεται για το $\{D, E\}$, **{C, E}**, {A, E}



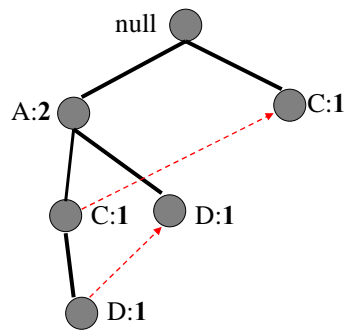
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



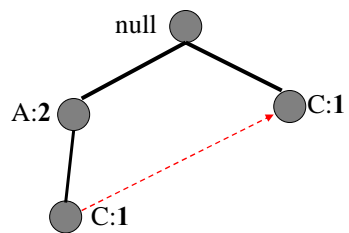
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



Αλγόριθμος FP-Growth

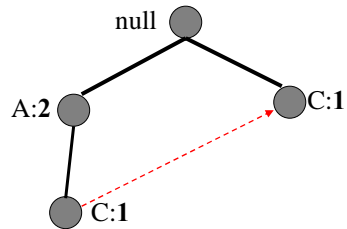


Βρες την υποστήριξη του {C, E}

Πως;

Ακολούθησε τους συνδέσμους
αθροίζοντας $1+1=2 \geq 2$

Οπότε {C, E} συχνό

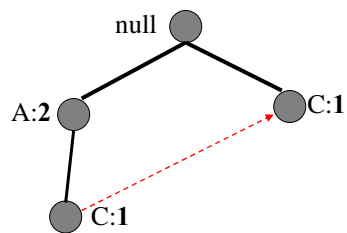


Αλγόριθμος FP-Growth



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

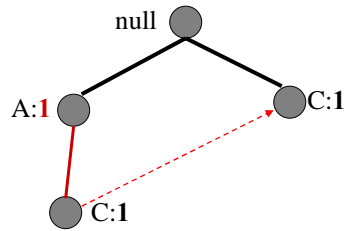
1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



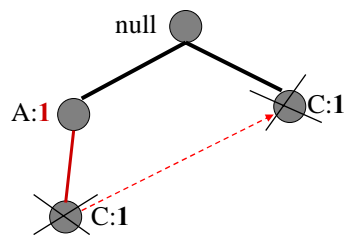
1. Αλλαγή υποστήριξης



Αλγόριθμος FP-Growth

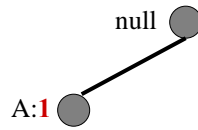


2. Περικοπή Κόμβων

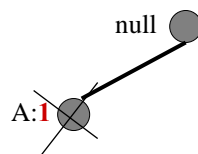




2. Περικοπή Κόμβων



2. Περικοπή Κόμβων





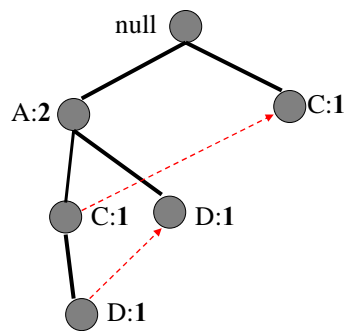
2. Περικοπή Κόμβων

null

Άρα, επιστροφή στο επόμενο υποπρόβλημα



Υπο-συνθήκη FP-δέντρο για το E
Ο αλγόριθμος επαναλαμβάνεται για το
~~{D, E}~~, {C, E}, {A, E}



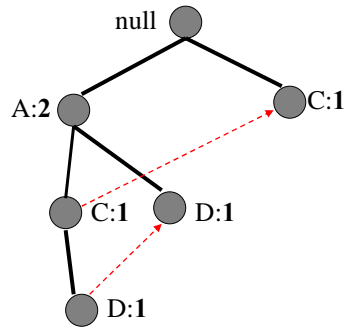
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)



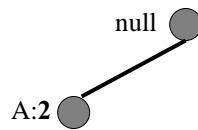
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)



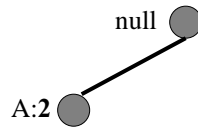
Αλγόριθμος FP-Growth



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

$\{E\}$ $\{D, E\}$ $\{A, D, E\}$ $\{C, E\}$ $\{A, E\}$

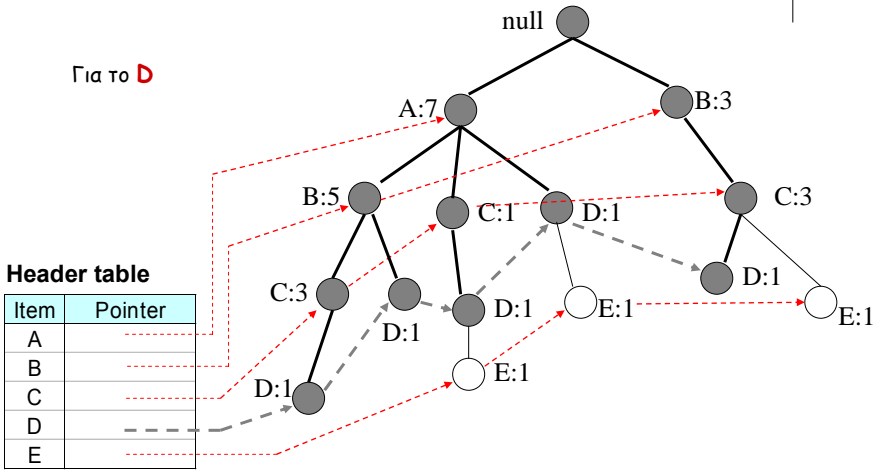
Συνεχίζουμε για το D

Αλγόριθμος FP-Growth




Για το **D**

Item	Pointer
A	
B	
C	
D	
E	



Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 69

Αλγόριθμος FP-Growth

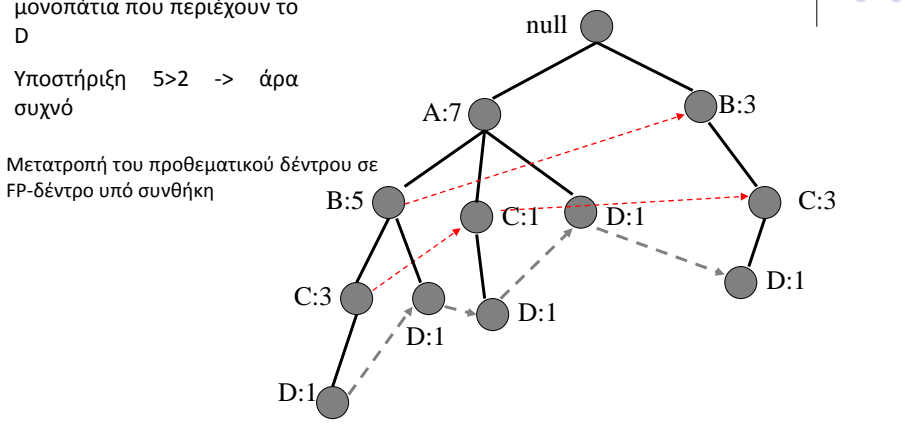


Φάση 1

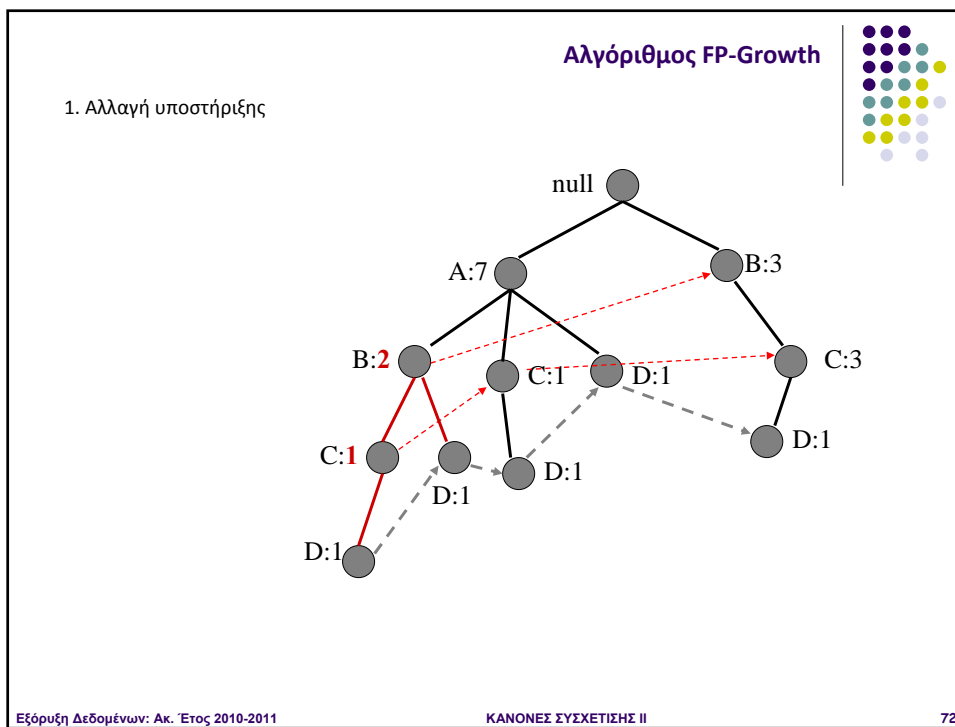
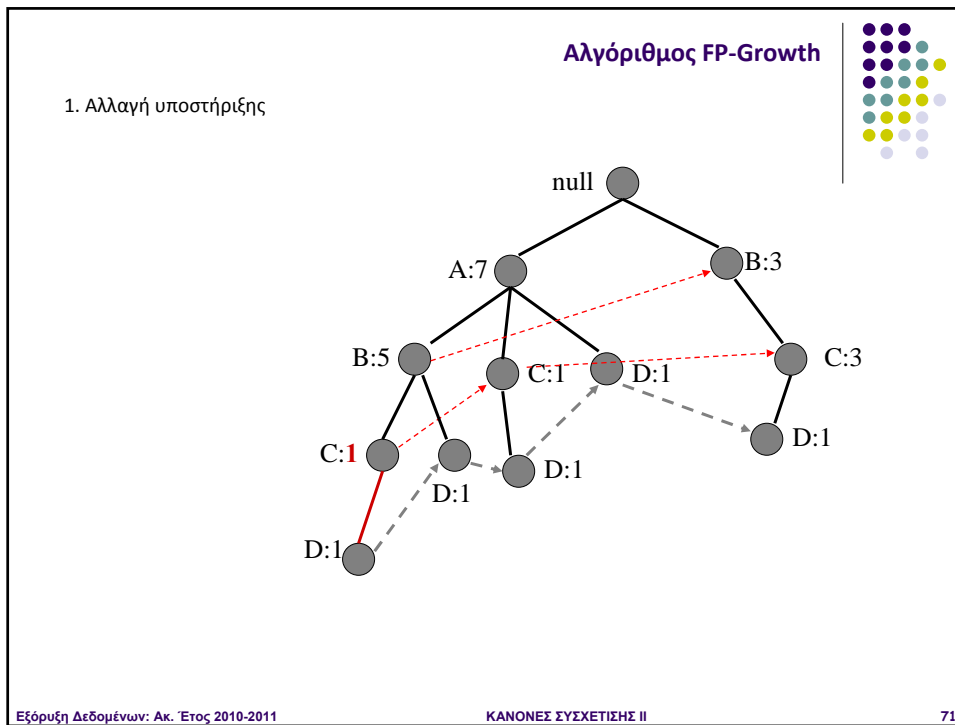
Όλα τα προθεματικά μονοπάτια που περιέχουν το D

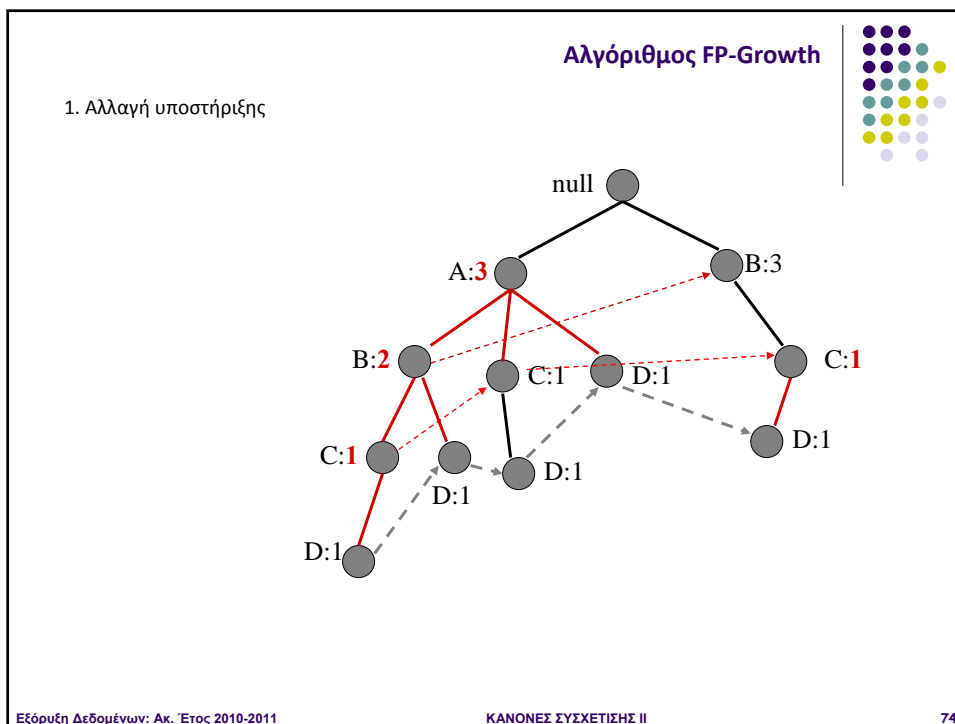
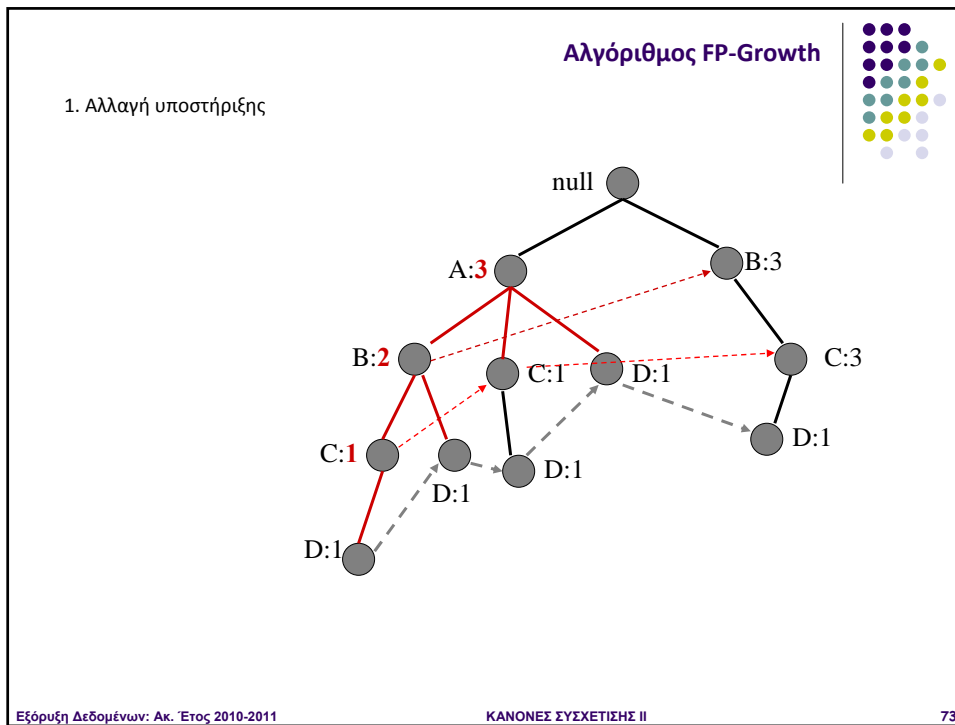
Υποστήριξη $5 > 2 \rightarrow$ άρα συχνό

Μετατροπή του προθεματικού δέντρου σε FP-δέντρο υπό συνθήκη



Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 70





Αλγόριθμος FP-Growth

1. Αλλαγή υποστήριξης

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 75

Αλγόριθμος FP-Growth

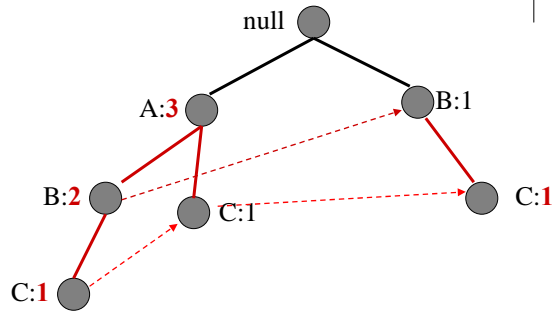
2. Περικοπή Κόμβων

Εξόρυξη Δεδομένων: Ακ. Έτος 2010-2011 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 76

Αλγόριθμος FP-Growth



2. Περικοπή Κόμβων



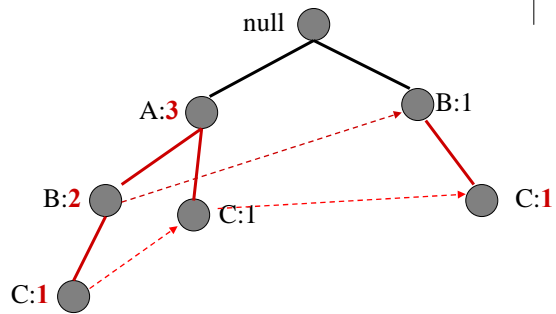
Αλγόριθμος FP-Growth



Προβληματικά 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



Παρατηρήσεις

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)