

Κανόνες Συσχέτισης II

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



Σύντομη Ανακεφαλαίωση



Market-Basket transactions
(Το καλάθι της νοικοκυράς!)

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

δοσοληψία (transaction) στοιχείο (item)

- Πρώτωση προϊόντων
- Τοποθέτηση προϊόντων στα ράφια
- Διαχείριση αποθεμάτων

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)



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

Ορισμοί

Κανόνας Συσχέτισης (Association Rule)

Είναι μια έκφραση της μορφής $X \rightarrow Y$,
όπου X και Y είναι στοιχειοσύνολα
 $X \subseteq I, Y \subseteq I, X \cap Y = \emptyset$

Παράδειγμα: $\{\text{Milk, Diaper}\} \rightarrow \{\text{Beer}\}$

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

Υποστήριξη Κανόνα Support (s)

Το ποσοστό των δοσοληψιών που περιέχουν και το X και το Y ($X \cup Y$)

Εμπιστοσύνη - Confidence (c)

Πόσες από τις δοσοληψίες (ποσοστό) που περιέχουν το X περιέχουν και το Y

Πρόβλημα

Εύρεση Κανόνων Συσχέτισης

Είσοδος: Ένα σύνολο από δοσοληψίες T
Έξοδος: Όλοι οι κανόνες με
 $support \geq minsup$
 $confidence \geq minconf$

Εξόρυξη Κανόνων Συσχέτισης

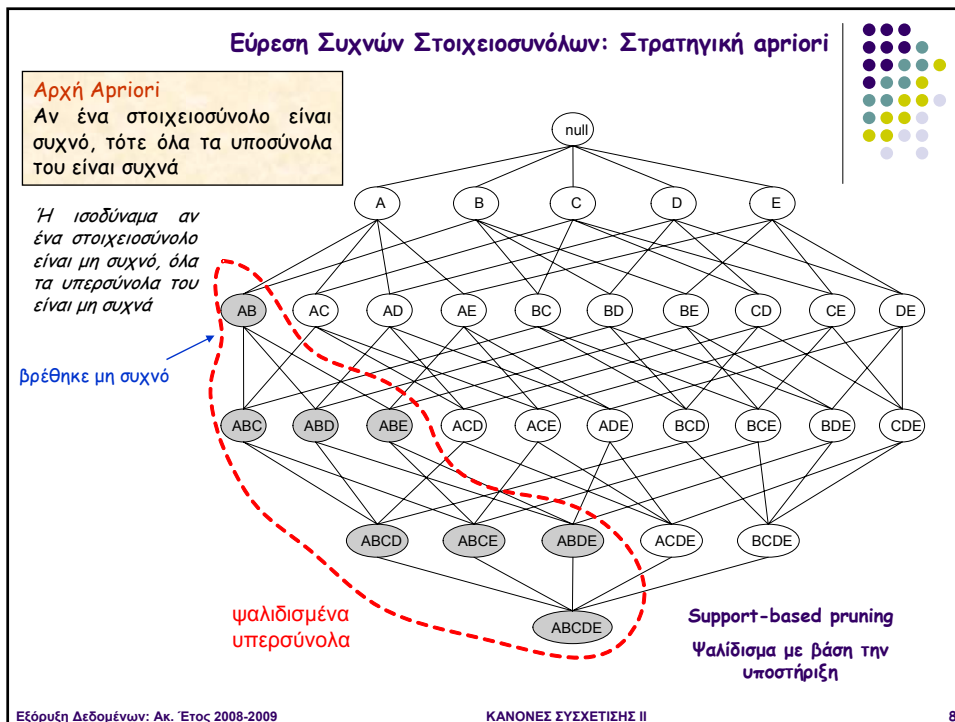
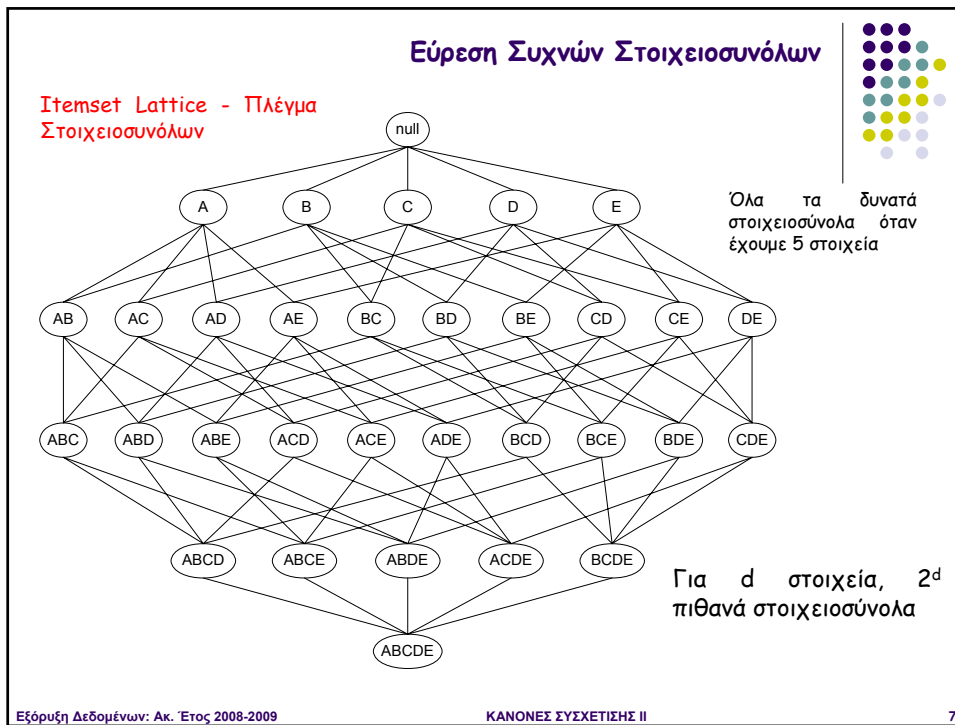
Χωρισμός του προβλήματος σε δύο υπο-προβλήματα:

Εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation)

Εύρεση όλων των στοιχειοσυνόλων με υποστήριξη $\geq minsup$

Δημιουργία Κανόνων (Rule Generation)

Για κάθε (συχνό) στοιχειοσύνολο, δημιουργήσε κανόνες με μεγάλη υποστήριξη, όπου κάθε κανόνας είναι μια δυαδική διαμέριση (δηλ. χωρισμός στα δύο) του συχνού στοιχειοσυνόλου





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. $k = k + 1$

Στρατηγική αρρίορι: Δημιουργία Στοιχειοσυνόλων



Για την παραγωγή υποψήφιων k -στοιχειοσυνόλων

▪ $F_{k-1} \times F_1$

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

▪ $F_{k-1} \times F_{k-1}$

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

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

Ψαλίδισμα

▪ Είναι δυνατόν να γίνουν απλοί έλεγχοι αν τα παραγόμενα πιθανά στοιχειοσύνολα είναι συχνά ελέγχοντας αν τα υποσύνολα τους είναι συχνά και έτσι να αποφύγουμε να υπολογίσουμε την υποστήριξή τους

Στρατηγική αρτιοσι: Υπολογισμός Υποστήριξης



Για κάθε νέο υποψήφιο $k+1$ -στοιχειοσύνολο, πρέπει να υπολογίσουμε την υποστήριξη του

Σε κάθε βήμα $k+1$

- Για να μειώσουμε τον αριθμό των πράξεων, αποθηκεύουμε τα υποψήφια $k+1$ -στοιχειοσύνολα σε ένα **δέντρο κατακερματισμού**
- Αντί να ταιριάζουμε κάθε δοσοληψία με κάθε υποψήφιο στοιχειοσύνολο,
 - κατακερματίζουμε όλα τα $k+1$ -στοιχειοσύνολα της δοσοληψίας και
 - για καθένα, ενημερώνουμε μόνο τους αντίστοιχους κώδους του δέντρου κατακερματισμού των συχνών στοιχειοσυνόλων

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 $L = \{A, B, C, D\}$: $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

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

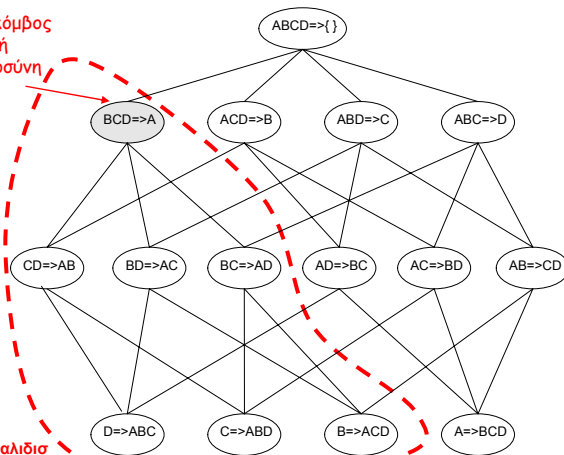
Παραγωγή Κανόνων για τον Αλγόριθμο αρθιορι



Πλέγμα Κανόνων για το
Στοιχειοσύνολο $\{A, B, C, D\}$

Ψαλίδισμα με βάση την εμπιστοσύνη

Έστω κόμβος
με μικρή
εμπιστοσύνη



Ψαλίδισμα
μένον
κανόνες

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

Αναπαράσταση Στοιχειοσυνόλων



Τα στοιχειοσύνολα που παράγονται είναι πολλά, κάποια ίσως περιττά -
οδηγούν σε παραγωγή πολλών κανόνων

Τποια να κρατήσουμε;

Ψάχνουμε για αντιπροσωπευτικά συχνά στοιχειοσύνολα (δηλαδή, να
μπορούμε να πάρουμε από αυτά ακριβώς όλα τα συχνά και ιδεατά να
μπορούμε να υπολογίσουμε και την υποστήριξη όλων των συχρών):

- Maximal συχνά
- Κλειστά συχνά

Αναπαράσταση Στοιχειοσυνόλων

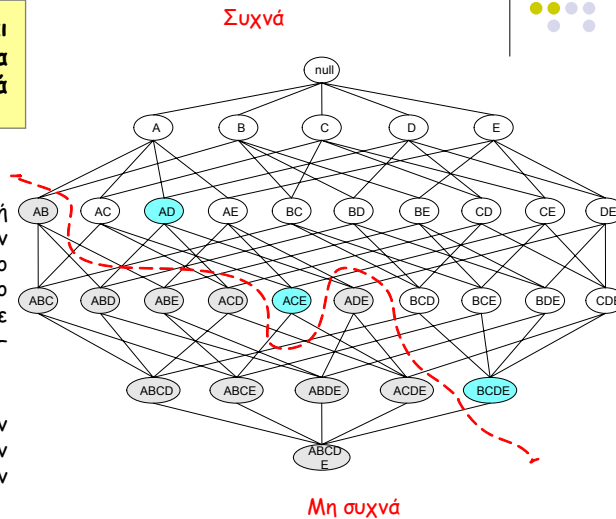


Ένα στοιχειοσύνολο είναι **maximal συχνό** αν κανένα από τα άμεσα υπερσύνολά του δεν είναι συχνό

δηλαδή είναι όλα μη συχνά

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



Αναπαράσταση Στοιχειοσυνόλων



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)

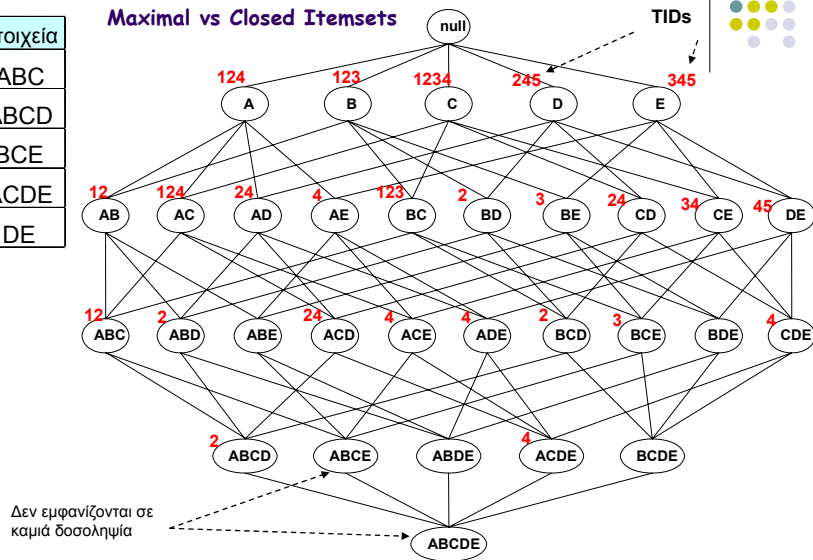
Πάλι τα υποσύνολα τους μας δίνουν όλα τα συχνά υποσύνολα, τώρα όμως μπορούμε να υπολογίσουμε την υποστήριξη των υποσυνόλων τους

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

Αναπαράσταση Στοιχειοσυνόλων

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

Maximal vs Closed Itemsets



Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

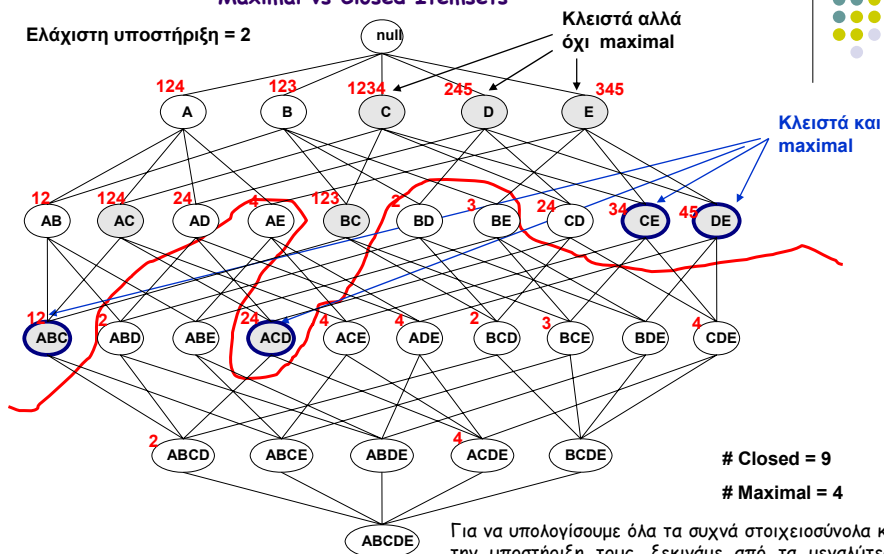
ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

17

Αναπαράσταση Στοιχειοσυνόλων

Maximal vs Closed Itemsets

Ελάχιστη υποστήριξη = 2



Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

18



Άλλοι Μέθοδοι Υπολογισμού Συχνών Στοιχειοσυνόλων

Άλλοι Μέθοδοι Υπολογισμού Συχνών Στοιχειοσυνόλων

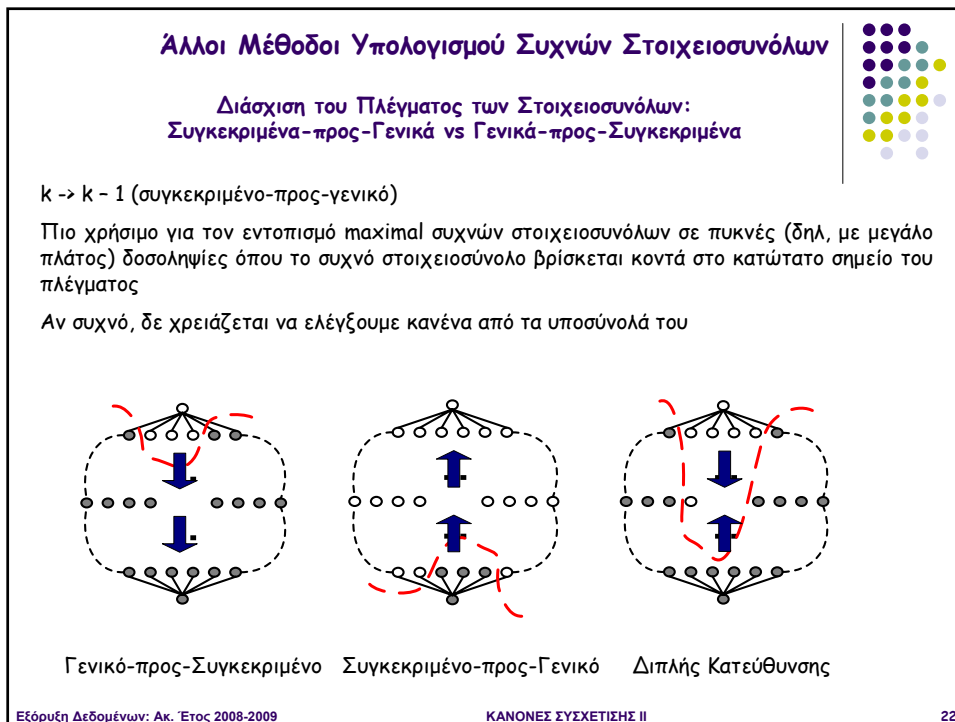
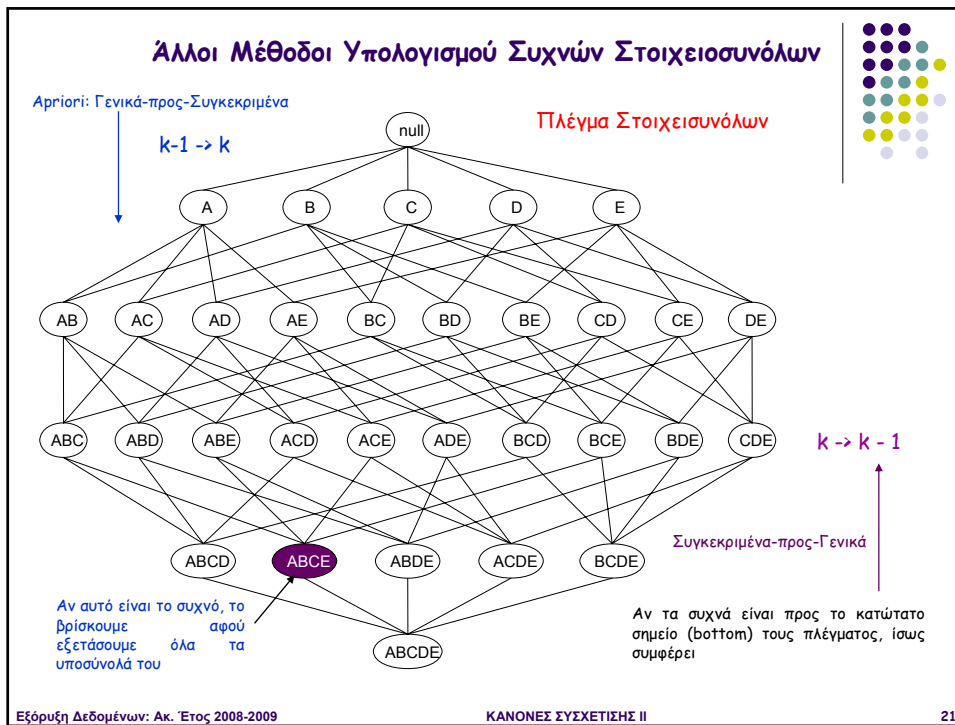


Ο Αρριορι από τους παλιότερους, αλλά:

- 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



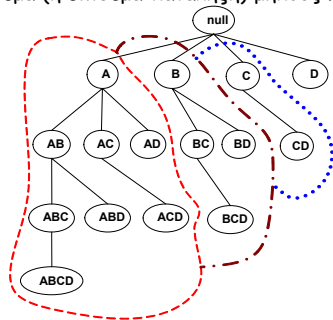
Άλλοι Μέθοδοι Υπολογισμού Συχνών Στοιχειοσυνόλων

Διάσχιση του Πλέγματος των Στοιχειοσυνόλων: Κλάσεις Ισοδυναμίας

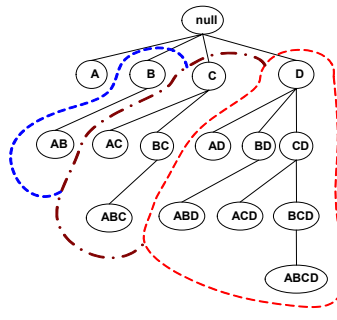
Χωρισμός των στοιχειοσυνόλων του πλέγματος σε ξένες μεταξύ τους ομάδες (κλάσεις ισοδυναμίας) και εξέταση των στοιχειοσυνόλων ανά κλάσεις

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

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



(a) Prefix tree

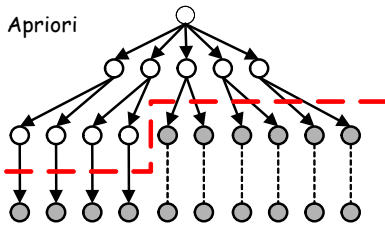


(b) Suffix tree

Άλλοι Μέθοδοι Υπολογισμού Συχνών Στοιχειοσυνόλων

Διάσχιση του Πλέγματος των Στοιχειοσυνόλων: BFS vs DFS

Apriori

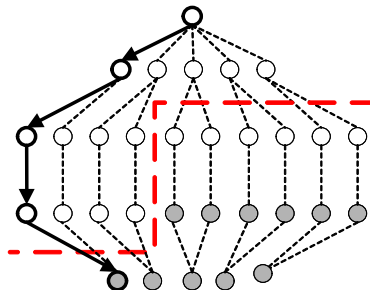


BFS: Breadth-First-Search
Διάσχιση κατά Πλάτος

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

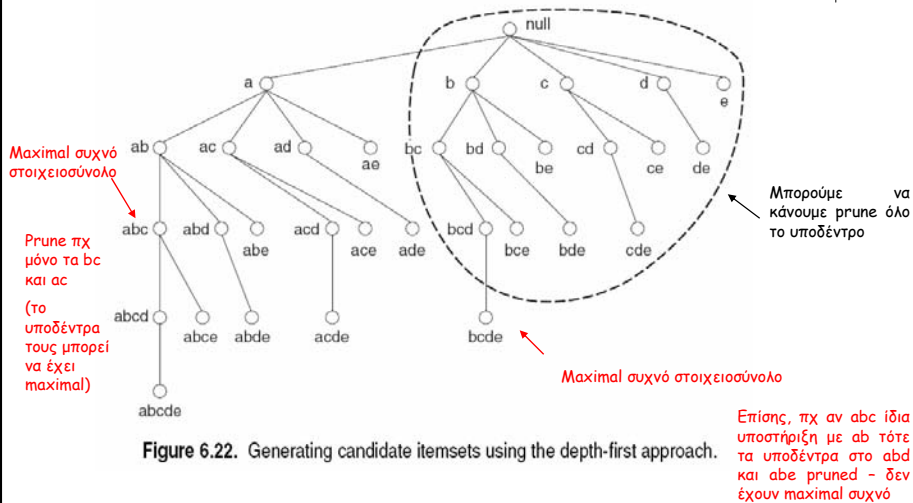
Μόλις εντοπιστεί το maximal, είναι δυνατόν να κλαδευτούν πολλά υποσύνολα του

DFS: Depth-First-Search
Διάσχιση κατά Βάθος



Άλλοι Μέθοδοι Υπολογισμού Συχνών Στοιχειοσυνόλων

Διάσχιση του Πλέγματος των Στοιχειοσυνόλων: BFS vs DFS



Άλλοι Τρόποι Υπολογισμού

Αναπαράσταση της Βάσης Δεδομένων των Δοσοληψιών: Οριζόντια vs Κάθετη

Αυτό χρησιμοποιεί ο αργιορί

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

Εναλλακτικά:

Για κάθε στοιχείο σε ποιες δοσοληψίες εμφανίζεται

Κάθετη Διάρθρωση Δεδομένων

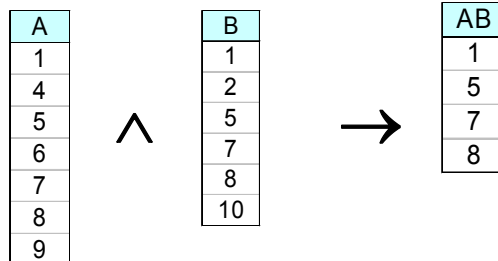
	A	B	C	D	E
1	1		2	2	1
4		2	3	4	3
5	5	5	4	5	6
6		7	8	9	
7		8	9		
8		10			
9					

Η υποστήριξη υπολογίζεται παίρνοντας τις τομές των TID-λιστών

Άλλοι Τρόποι Υπολογισμού



Η υποστήριξη υπολογίζεται παίρνοντας τις τομές των TID-λυστών



- Η υποστήριξη ενός k -στοιχειοσύνολου υπολογίζεται παίρνοντας τις τομές των TID-λυστών 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

Θα δούμε τον FP-Growth που χρησιμοποιεί μια prefix-based αναπαράσταση των δοσοληψιών

Ο Αλγόριθμος FP-Growth





Με λίγα λόγια:

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** (θα δούμε αργότερα αν κάτι άλλο συμφέρει καλύτερα)

Αρχικά, το δέντρο κενό



null

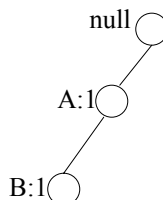
Αλγόριθμος FP-Growth



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 **ΕΤΙΚΕΤΑ**: ποιο στοιχείο και τη συχνότητα εμφάνισης (υποστήριξη) - πόσες δοσοληψίες φτάνουν σε αυτόν

Ετικέτα κόμβου <ΣΤΟΙΧΕΙΟ: ΥΠΟΣΤΗΡΙΞΗ>

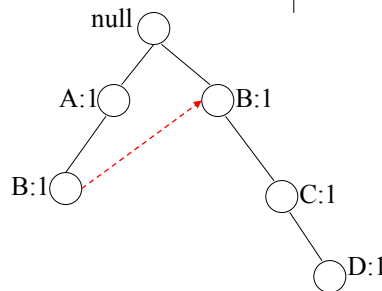
Αλγόριθμος FP-Growth



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 δοσοληψίες φτάνουν σε αυτόν

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

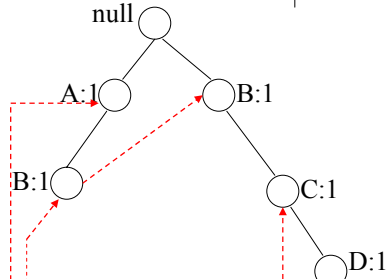
Αλγόριθμος FP-Growth



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

Αλγόριθμος FP-Growth

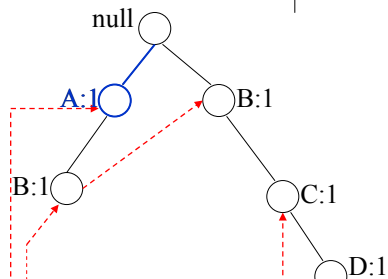


TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3



Πίνακας Δεικτών

Item	Pointer
A	----->
B	----->
C	----->
D	----->
E	----->

Αλγόριθμος FP-Growth

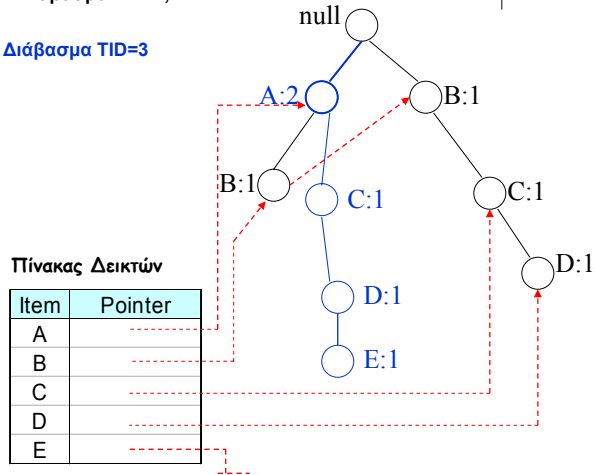


TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3



Αλγόριθμος FP-Growth

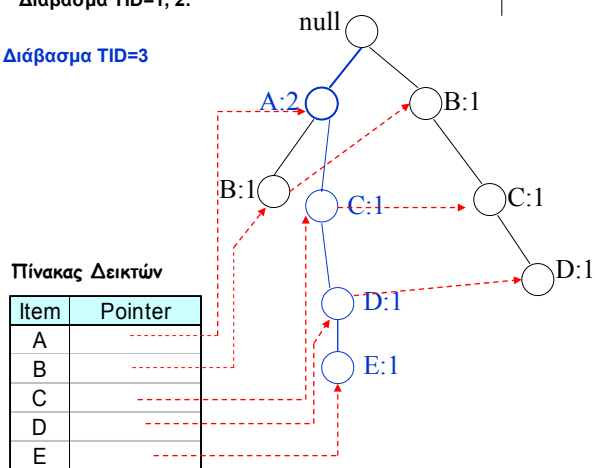


TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Κατασκευή FP-δέντρου

Διάβασμα TID=1, 2:

Διάβασμα TID=3



Αλγόριθμος FP-Growth

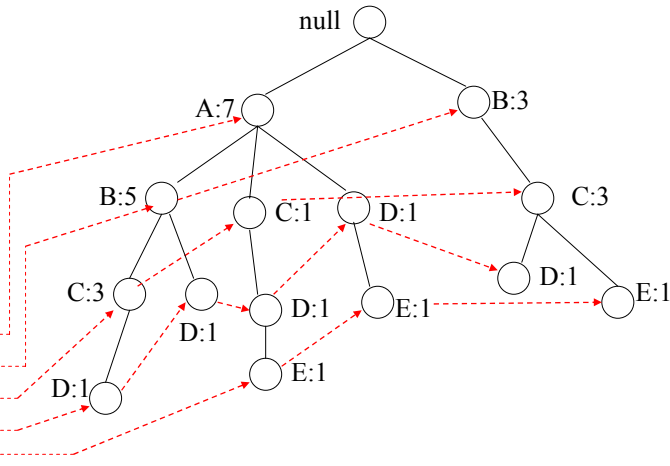


Κατασκευή FP-δέντρου

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}

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	



Αλγόριθμος FP-Growth



Μέγεθος FP-δέντρου

- 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)

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

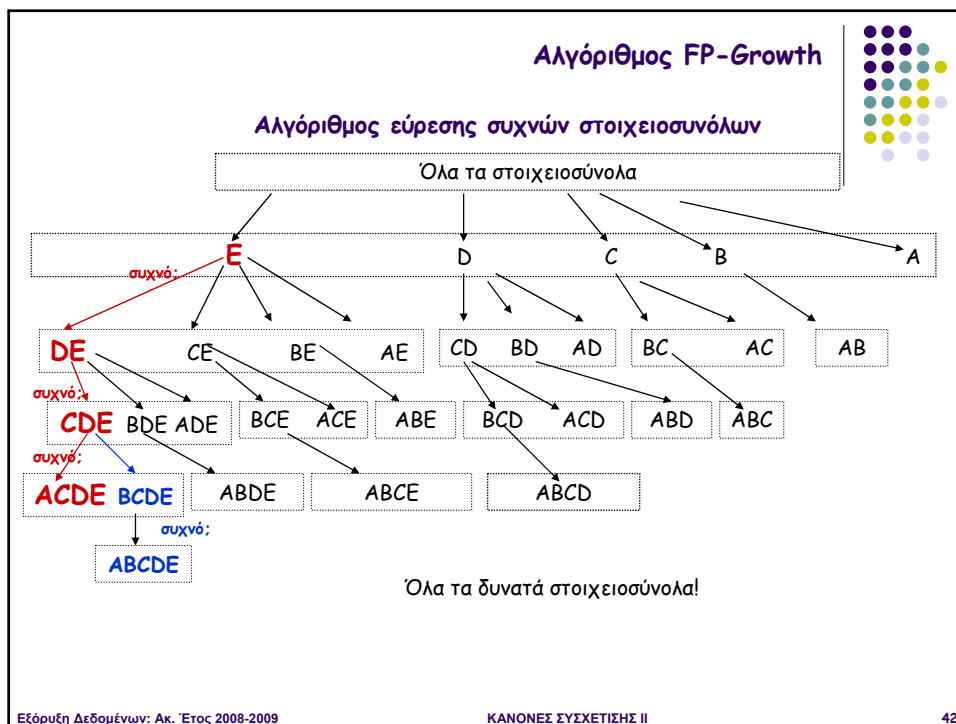
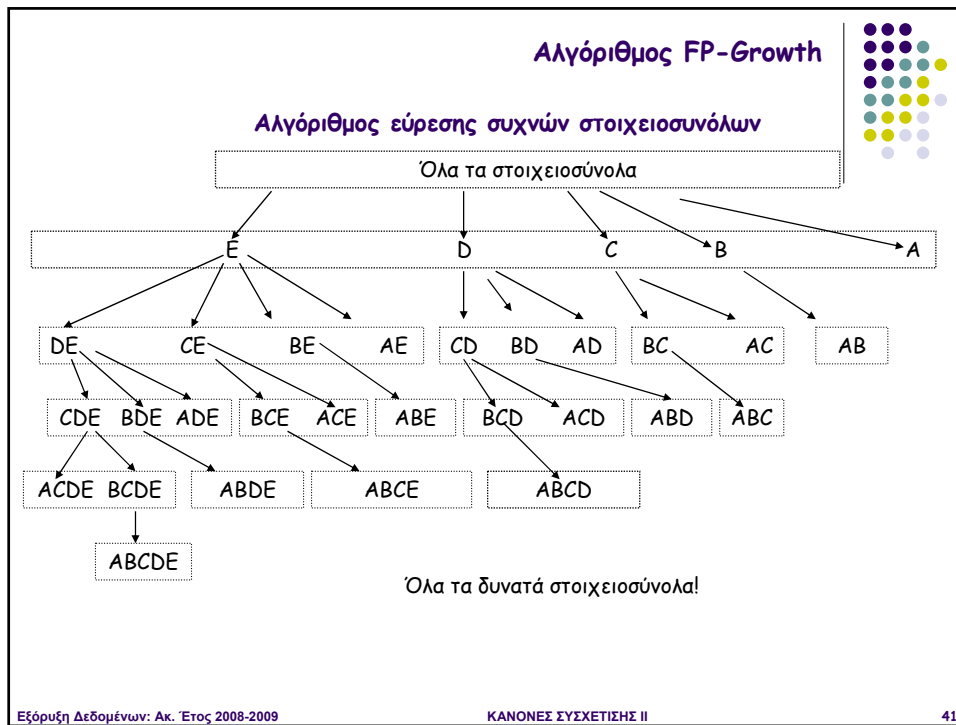
Έξοδος: Συχνά στοιχειοσύνολα και η υποστήριξη τους

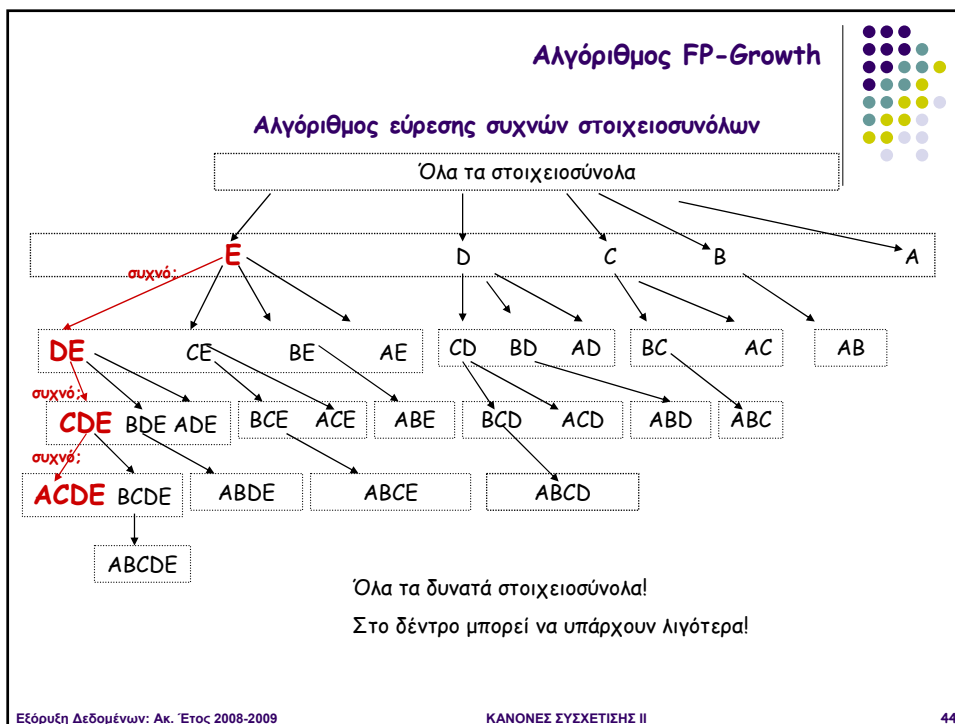
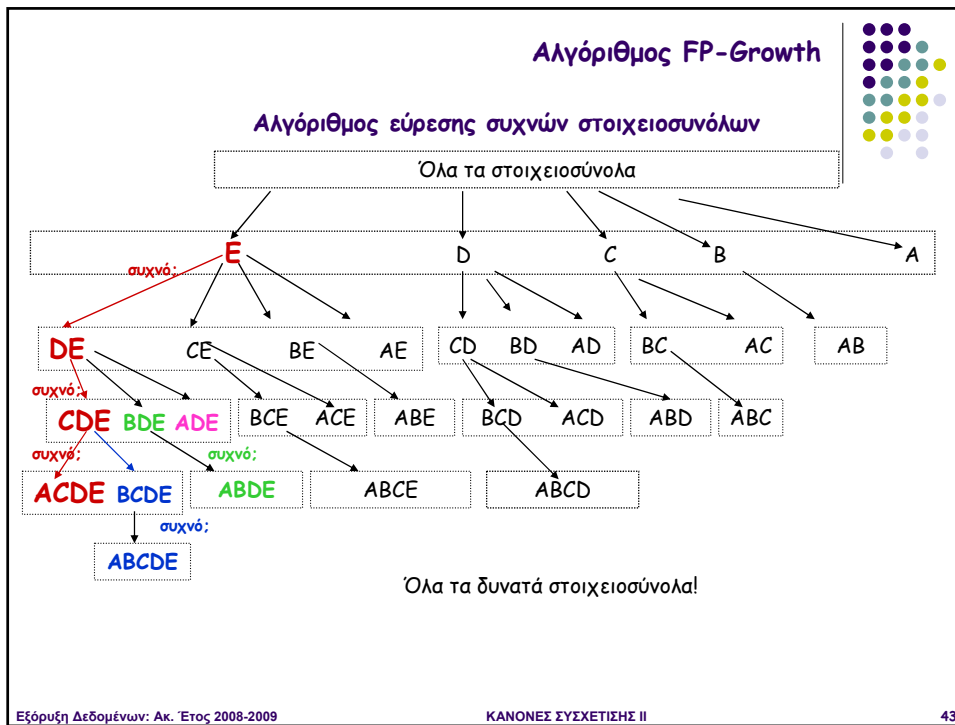
Μέθοδος:

▪ Διαιρεί-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



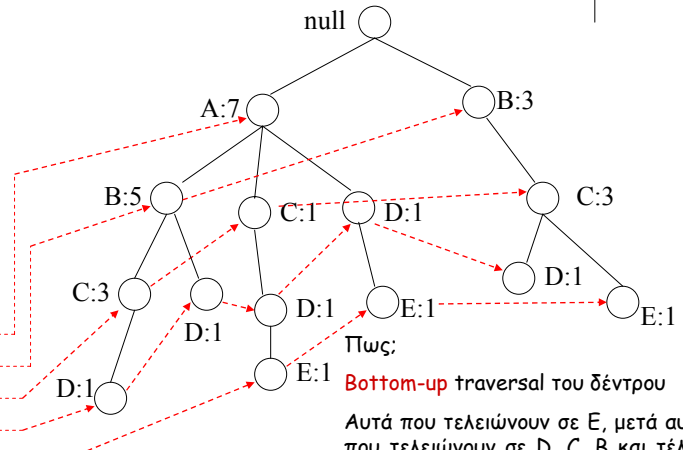


Αλγόριθμος FP-Growth



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	

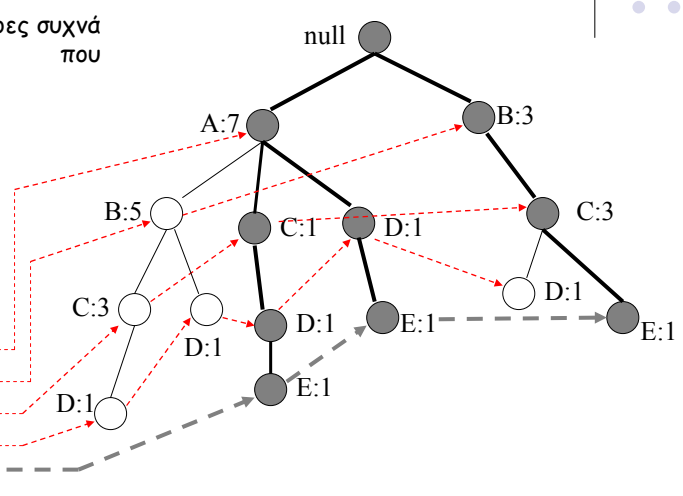
Πως:
Bottom-up traversal του δέντρου
 Αυτά που τελειώνουν σε E, μετά αυτά που τελειώνουν σε D, C, B και τέλος A - suffix-based classes (επίθεμα - κατάληξη)

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 45

Αλγόριθμος FP-Growth



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

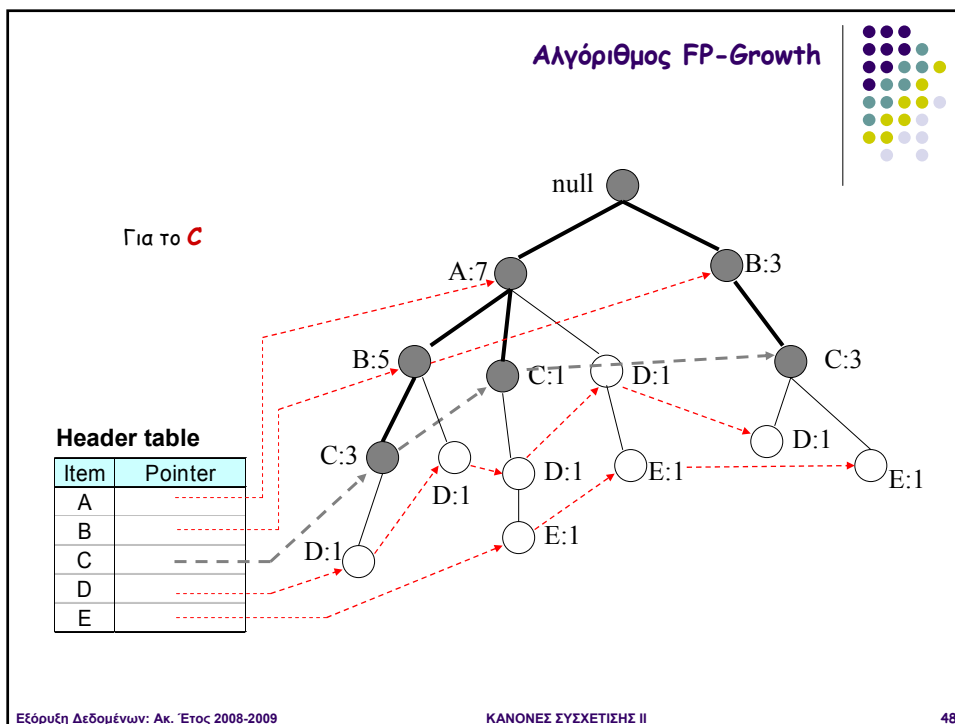
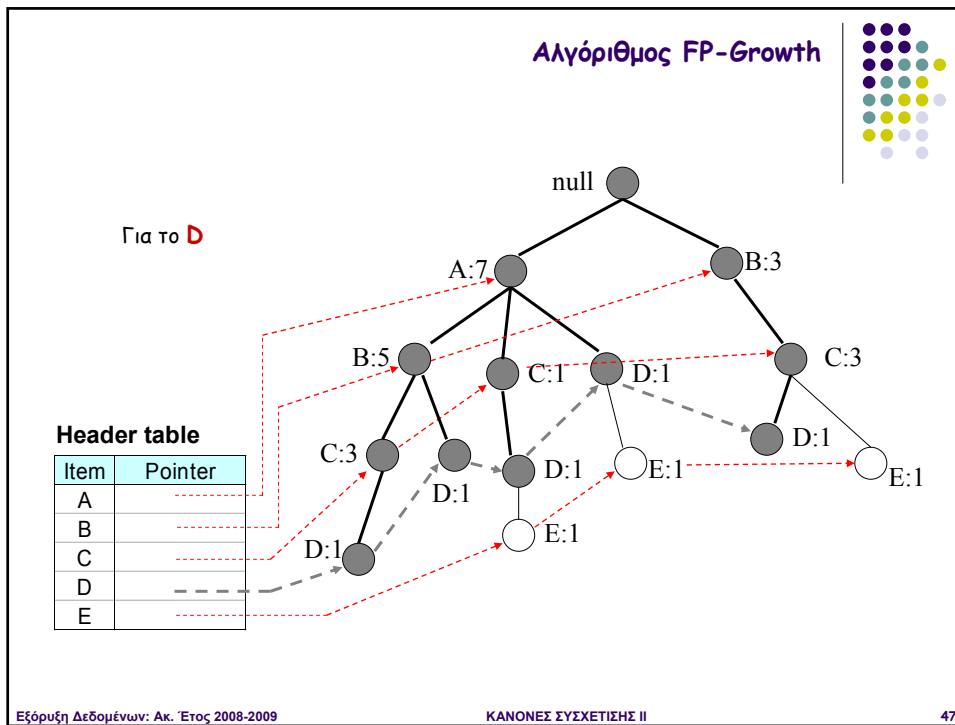


Header table

Item	Pointer
A	
B	
C	
D	
E	

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

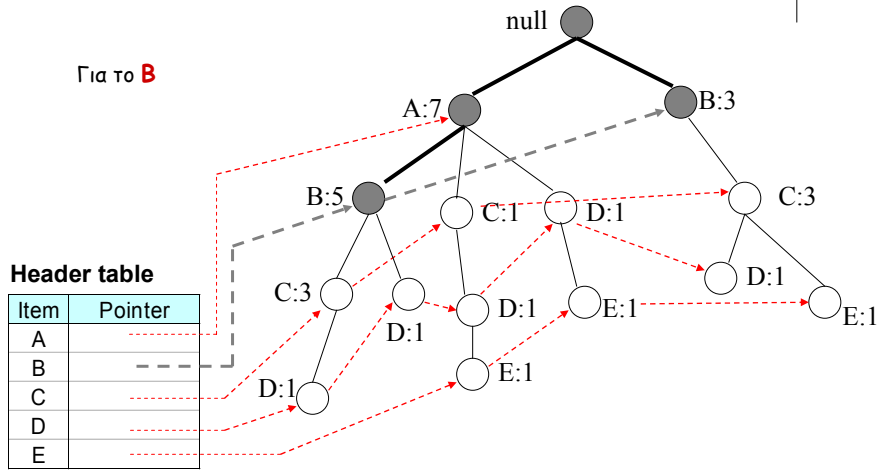
Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 46



Αλγόριθμος FP-Growth



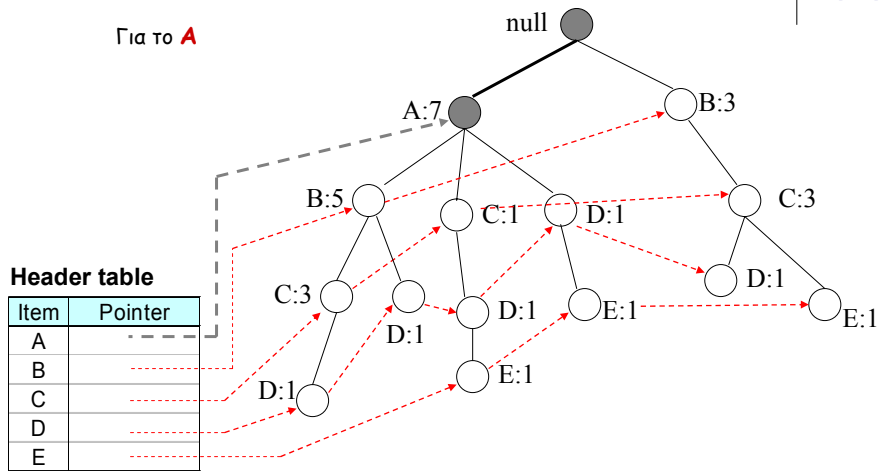
Για το **B**



Αλγόριθμος FP-Growth



Για το **A**





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



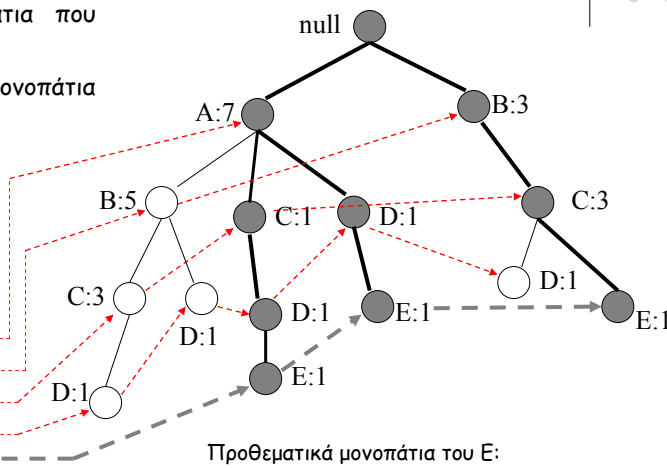
Φάση 1 - κατασκευή προθεματικού δέντρου

Όλα τα μονοπάτια που περιέχουν το E

Προθεματικά Μονοπάτια (prefix paths)

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}

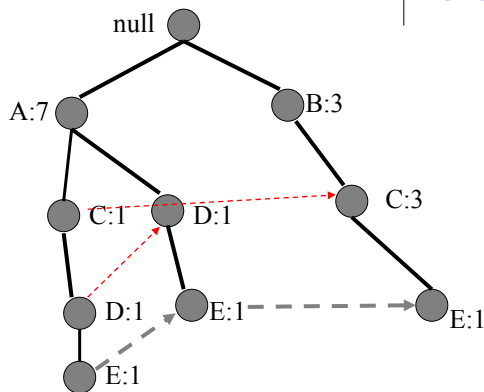
Αλγόριθμος FP-Growth



Ψάξη 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



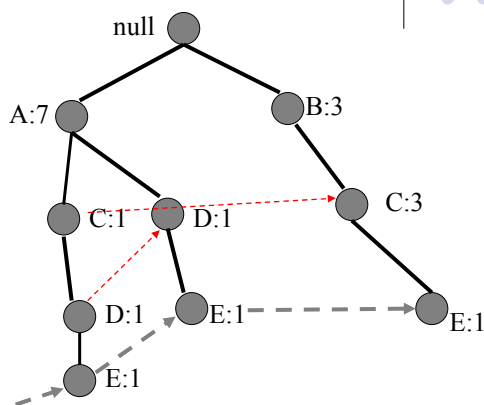
Έστω $\text{minsup} = 2$

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



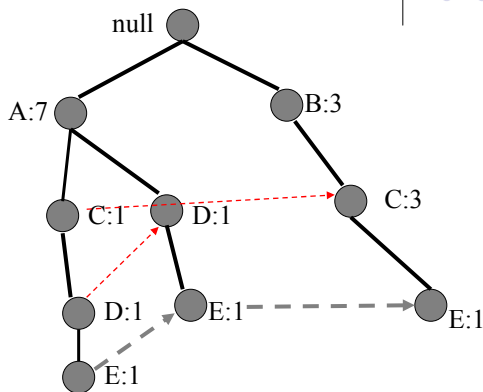
{E} συχνό άρα προχωράμε για DE, CE, BE, AE

Φάση 2

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

Δύο αλλαγές

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



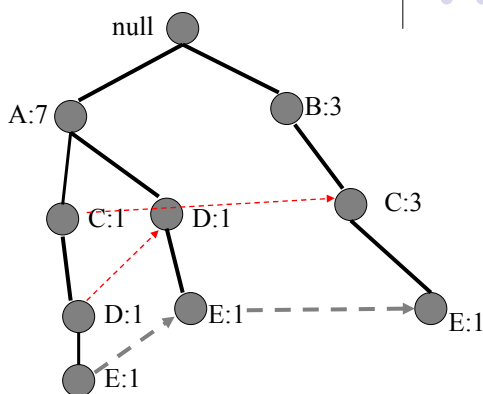
Αλγόριθμος FP-Growth

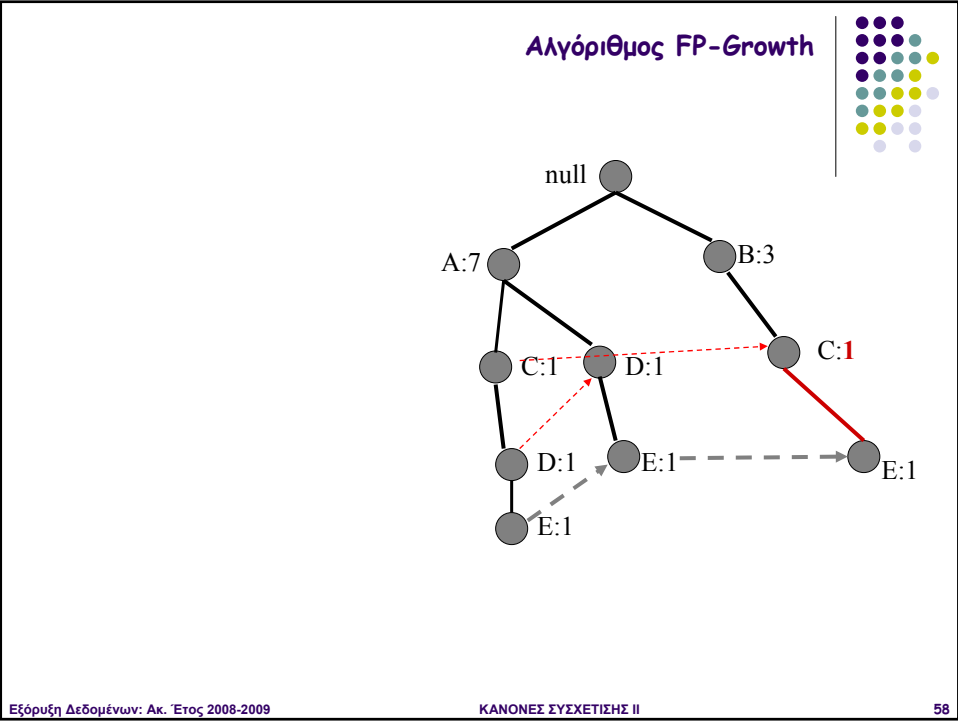
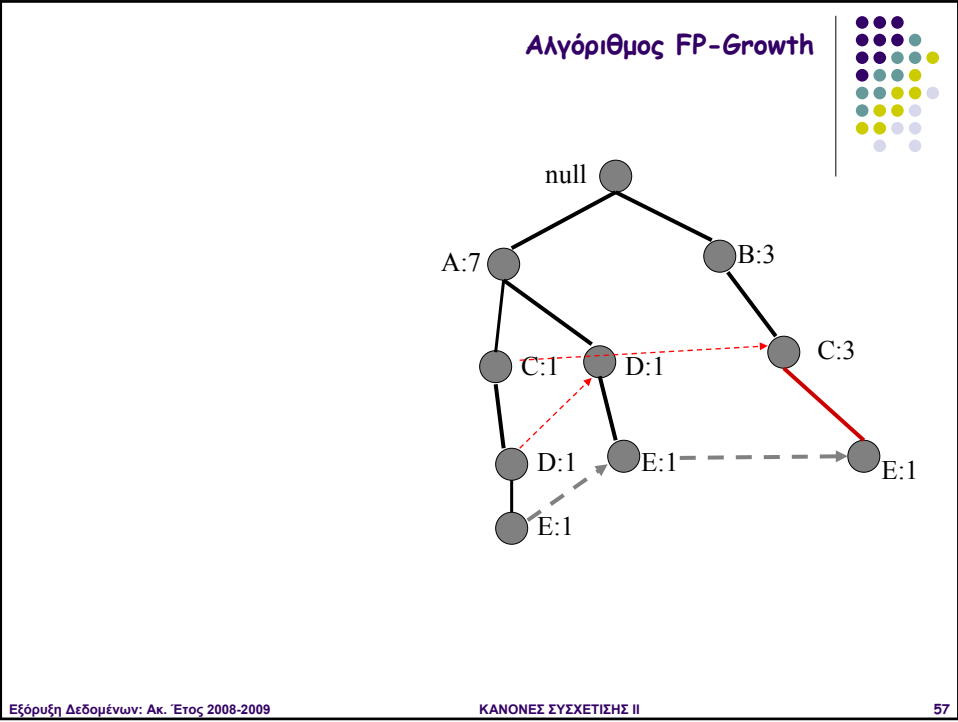


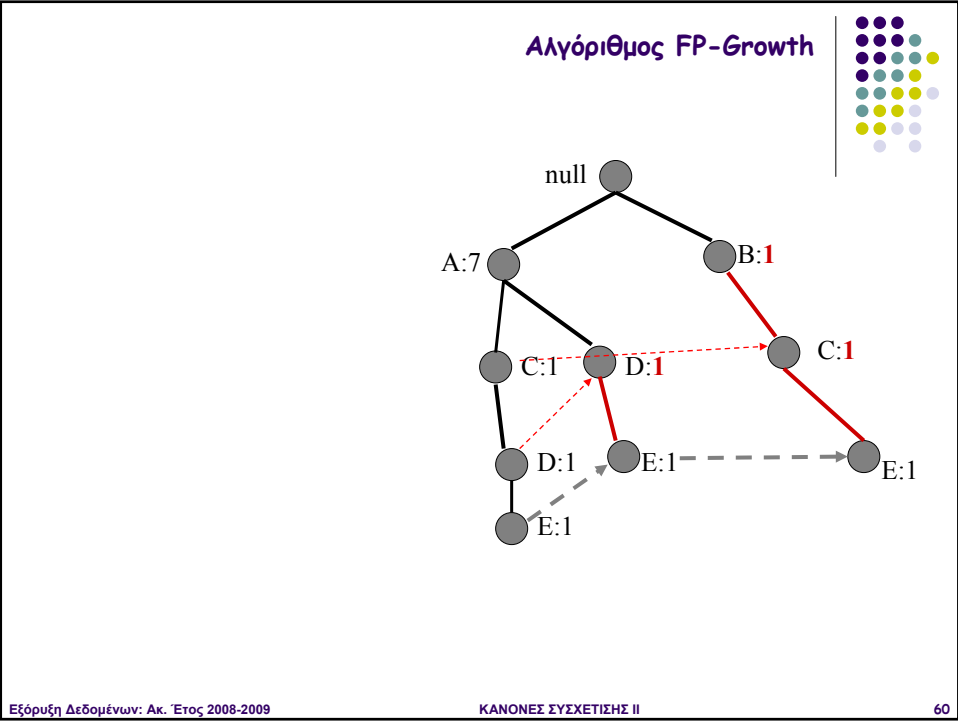
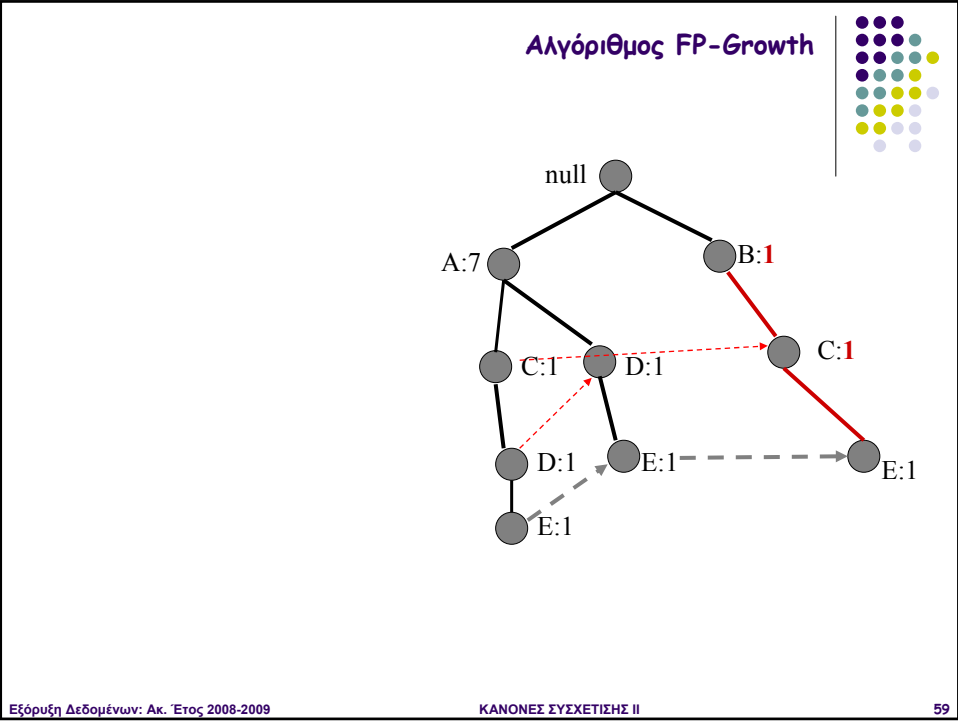
Αλλαγή μετρητών

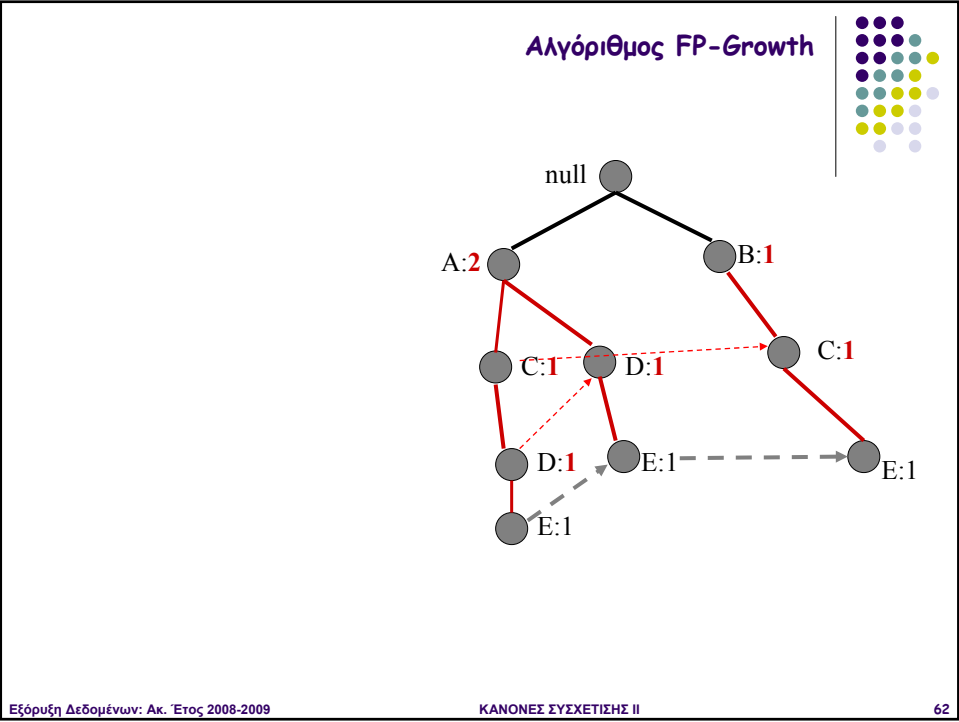
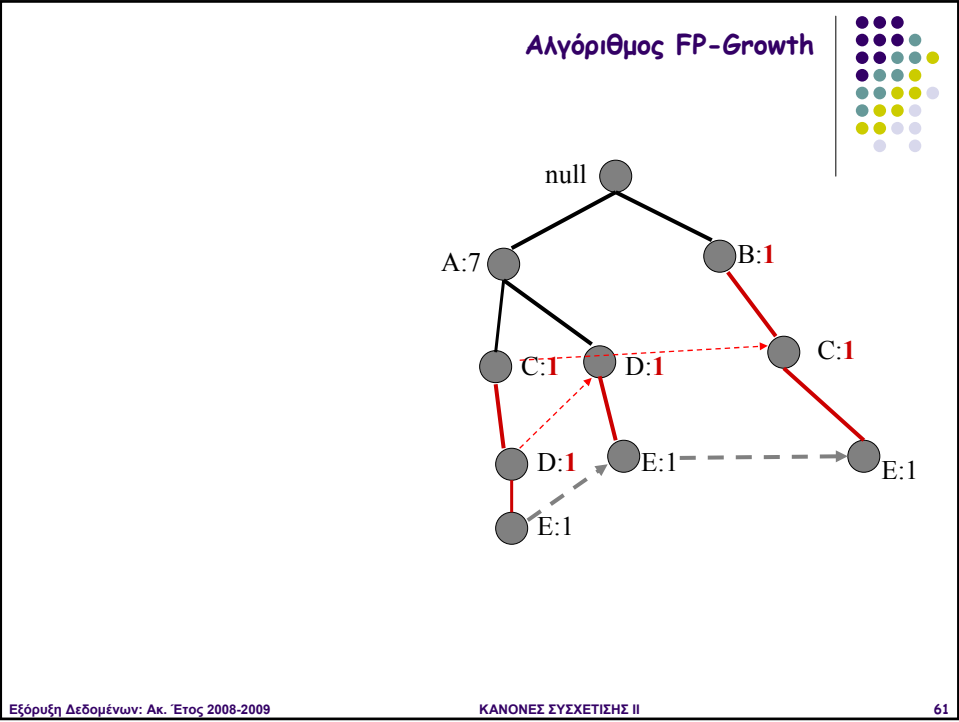
Οι μετρητές σε κάποιους κόμβους περιλαμβάνουν ποσοληψίες που δεν έχουν το E

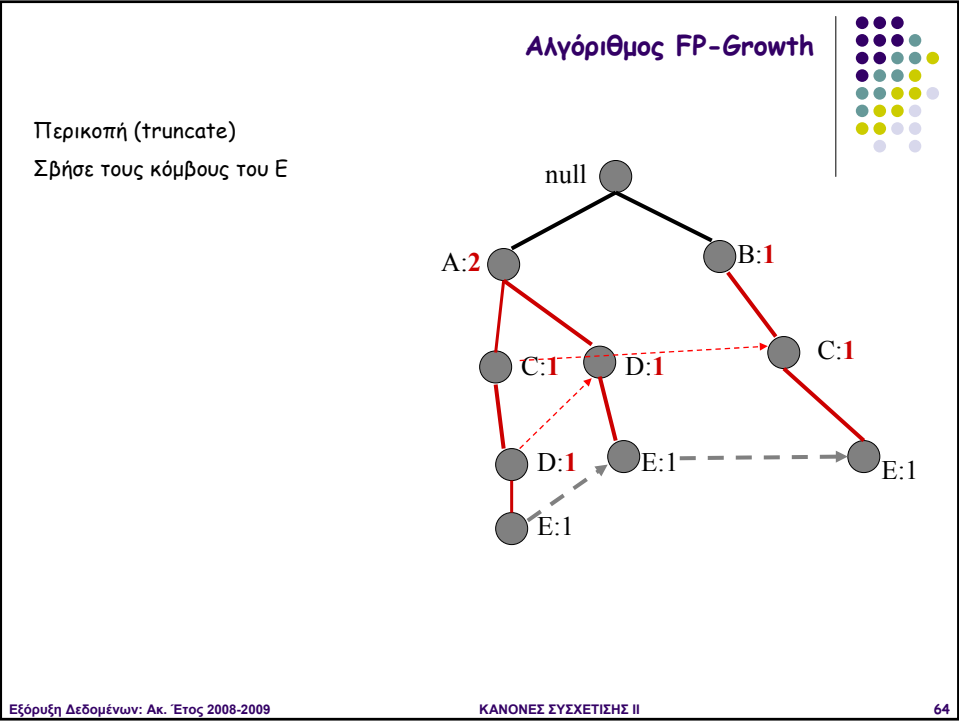
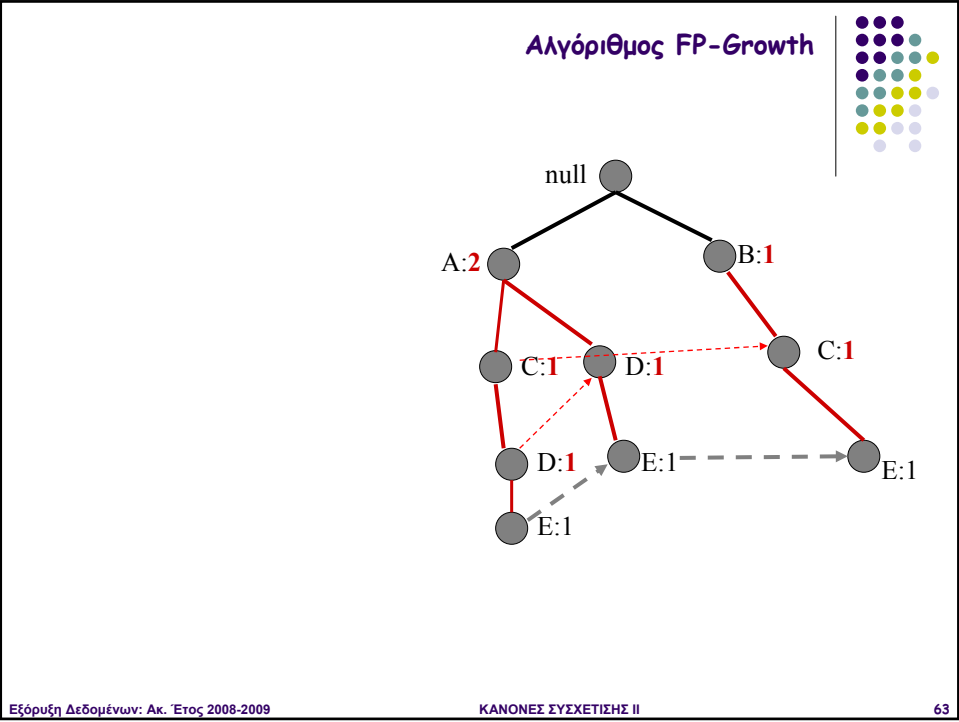
Πχ στο null->B->C->E μετράμε και την {B, C}







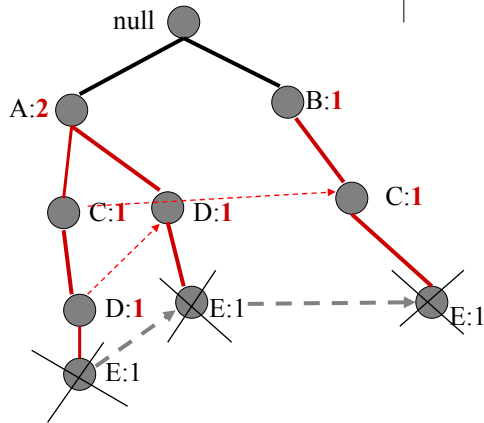




Αλγόριθμος FP-Growth



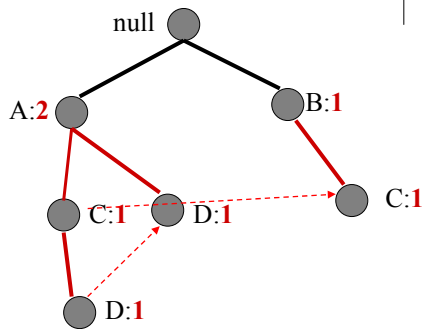
Περικοπή (truncate)
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth



Περικοπή (truncate)
Σβήσε τους κόμβους του E

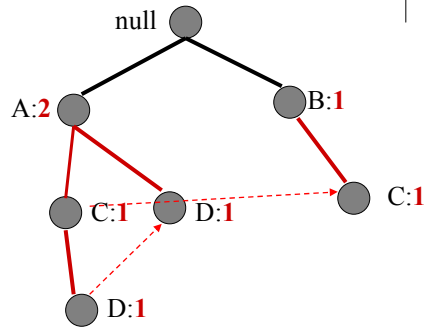


Αλγόριθμος FP-Growth

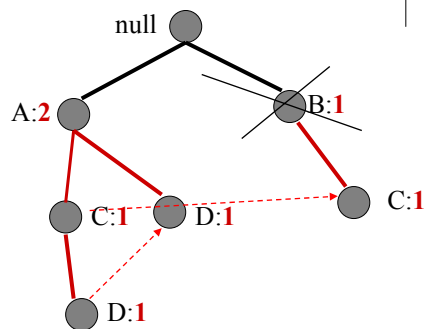


Πιθανή περαιτέρω περικοπή
Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης
Πχ το B → περικοπή

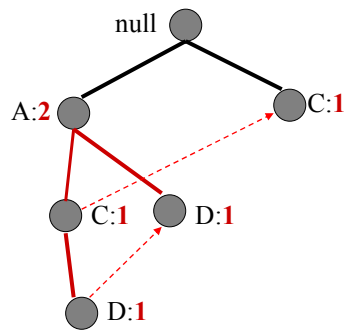
Αυτό σημαίνει ότι το B εμφανίζεται μαζί με το E λιγότερο από minsup φορές



Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

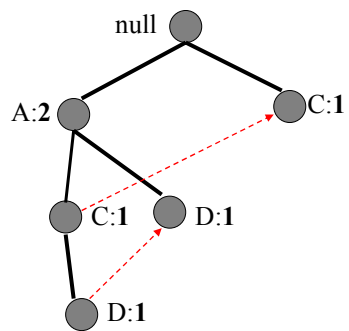


Αλγόριθμος FP-Growth



Υπο-συνθήκη FP-δέντρο για το E

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



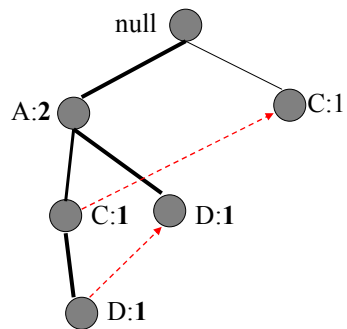
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



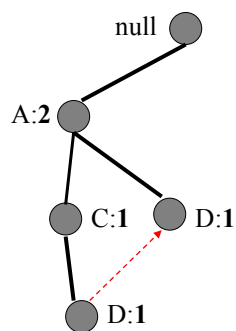
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το D (DE)

Προθεματικά Μονοπάτια (prefix paths)



Αλγόριθμος FP-Growth

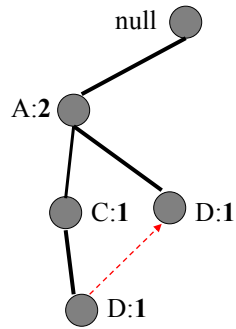


Βρες την υποστήριξη του {D, E}

Πως;

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

Οπότε {D, E} συχνό



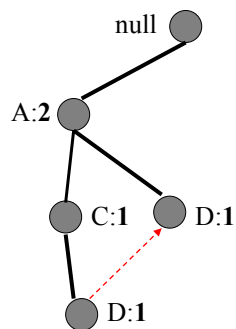
Αλγόριθμος FP-Growth



Φάση 2

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

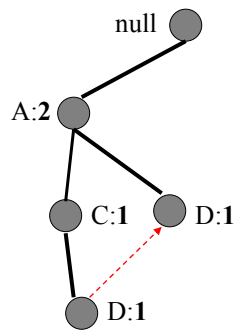
1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



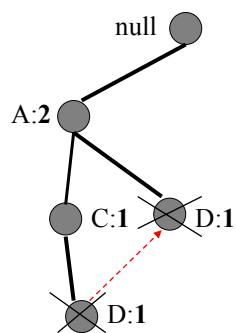
1. Αλλαγή υποστήριξης



Αλγόριθμος FP-Growth



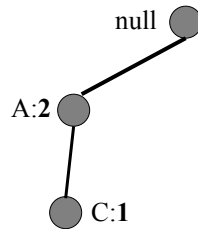
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



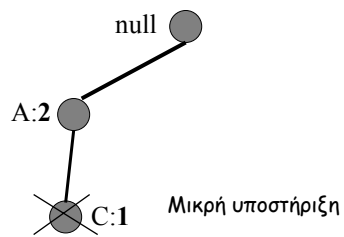
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



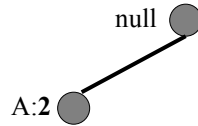
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



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

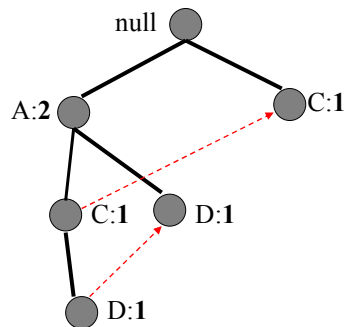


Υποστήριξη του A είναι $\geq \text{minsup}$ \rightarrow {A, D, E} συχνό
Αφού μόνο έναν κόμβο, επιστροφή στο επόμενο υποπρόβλημα

Αλγόριθμος FP-Growth



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



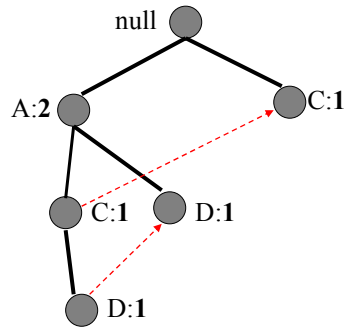
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



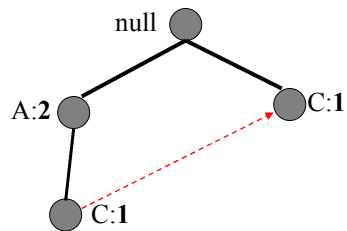
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το C (CE)

Προθεματικά Μονοπάτια (prefix paths)



Αλγόριθμος FP-Growth

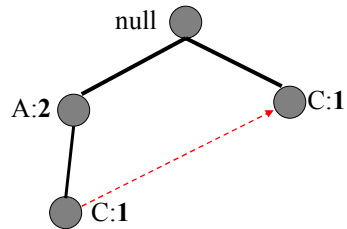


Βρες την υποστήριξη του {C, E}

Πως;

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

Οπότε {C, E} συχνό

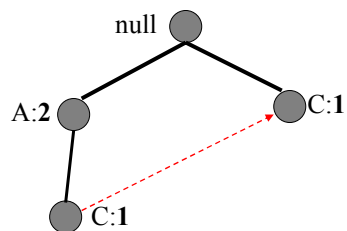


Αλγόριθμος FP-Growth



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

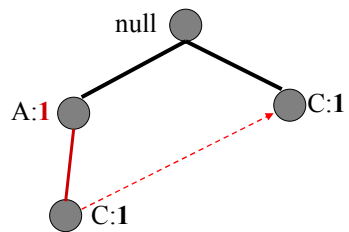
1. Αλλαγή υποστήριξης
2. Περικοπές κόμβων



Αλγόριθμος FP-Growth



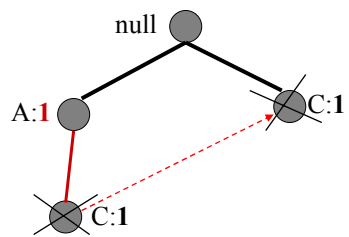
1. Αλλαγή υποστήριξης



Αλγόριθμος FP-Growth

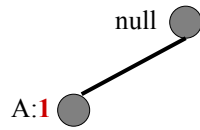


2. Περικοπή Κόμβων

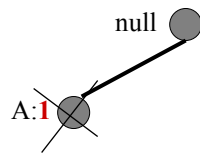




2. Περικοπή Κόμβων



2. Περικοπή Κόμβων





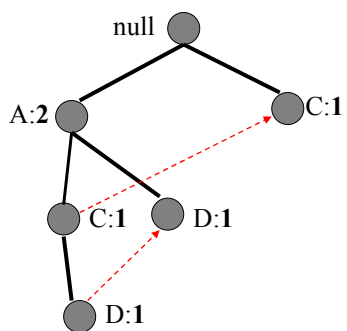
2. Περικοπή Κόμβων

null 

Άρα, επιστροφή στο επόμενο υποπρόβλημα



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



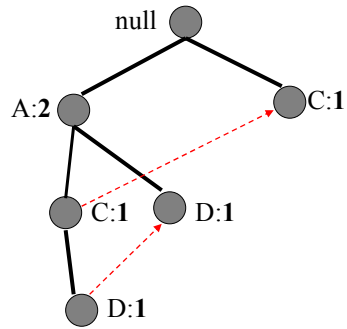
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)



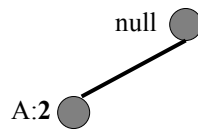
Αλγόριθμος FP-Growth



Φάση 1

Όλα τα μονοπάτια που περιέχουν το A (AE)

Προθεματικά Μονοπάτια (prefix paths)



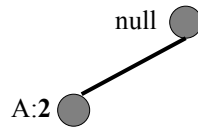
Αλγόριθμος FP-Growth



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

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

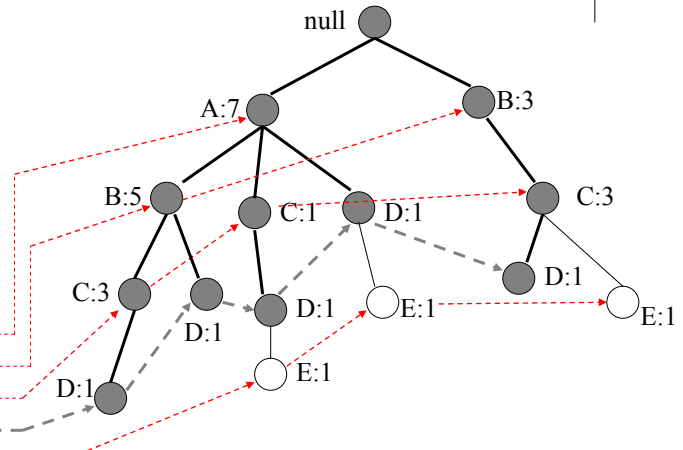
Συνεχίζουμε για το D

Αλγόριθμος FP-Growth




Για το **D**

Item	Pointer
A	
B	
C	
D	
E	



Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 95

Αλγόριθμος FP-Growth

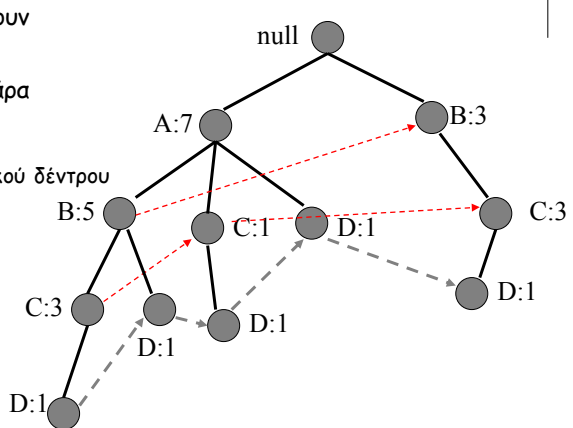


Φάση 1

Όλα τα προθεματικά μονοπάτια που περιέχουν το D

Υποστήριξη $5 > 2$ → άρα συχνό

Μετατροπή του προθεματικού δέντρου σε FP-δέντρο υπό συνθήκη

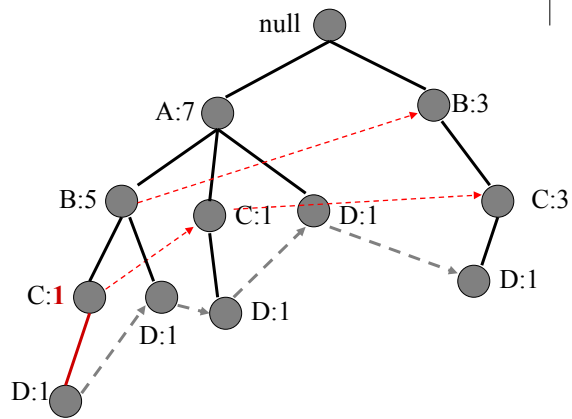


Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009 ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II 96

Αλγόριθμος FP-Growth



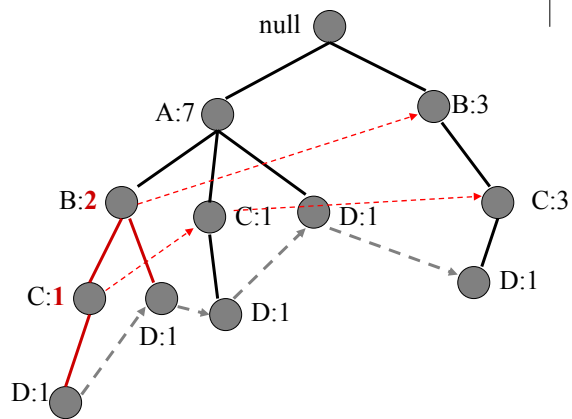
1. Αλλαγή υποστήριξης

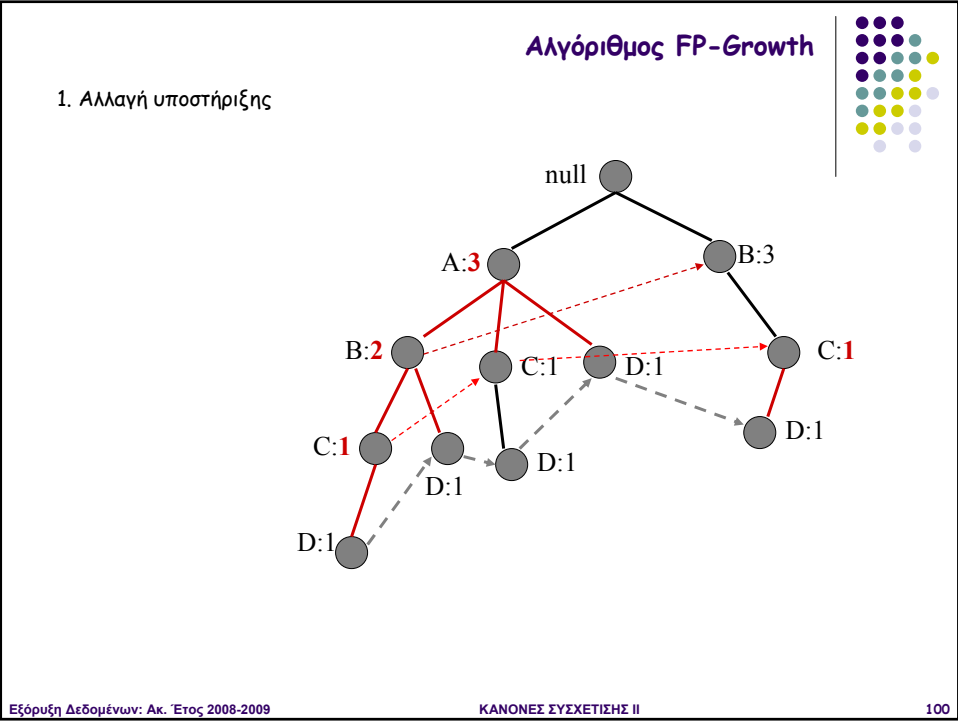
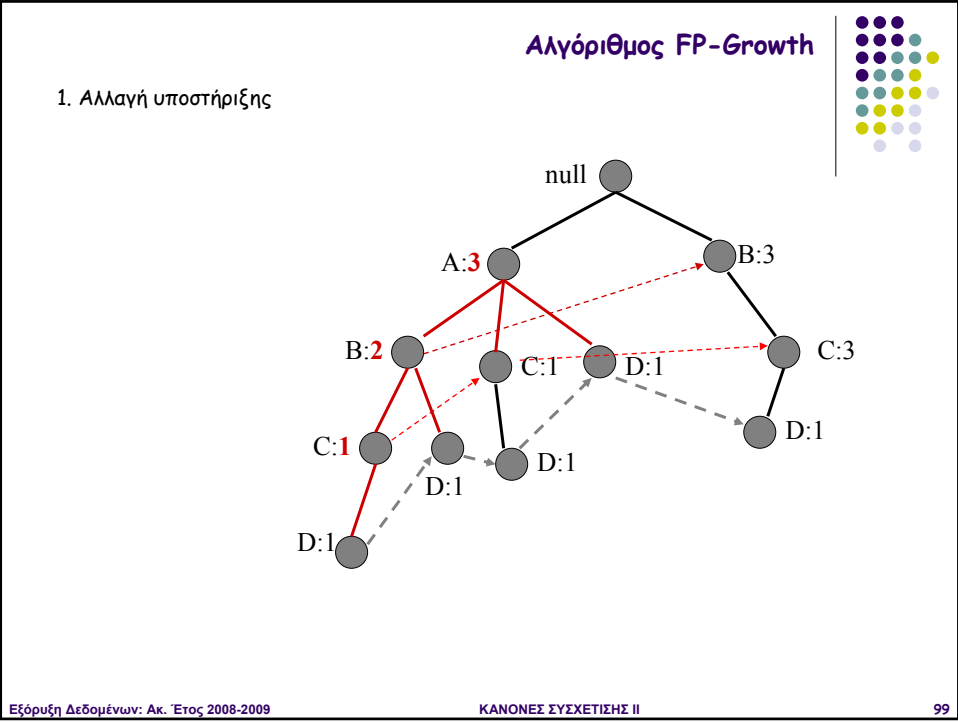


Αλγόριθμος FP-Growth



1. Αλλαγή υποστήριξης

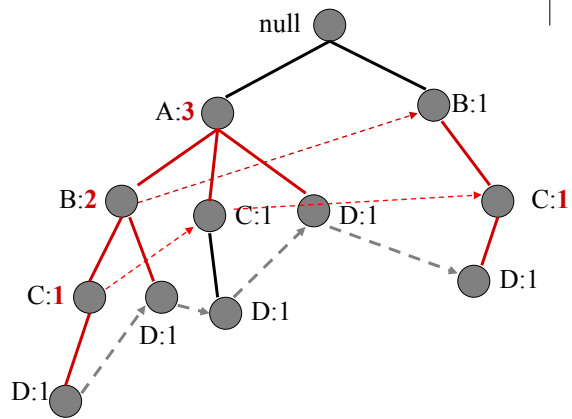




Αλγόριθμος FP-Growth



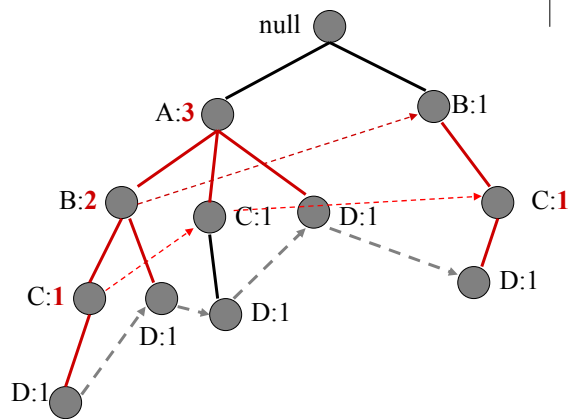
1. Αλλαγή υποστήριξης



Αλγόριθμος FP-Growth



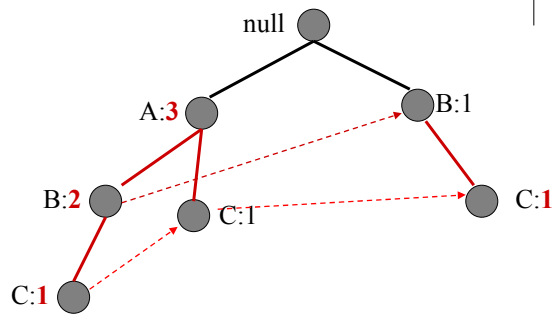
2. Περικοπή Κόμβων



Αλγόριθμος FP-Growth



2. Περικοπή Κόμβων

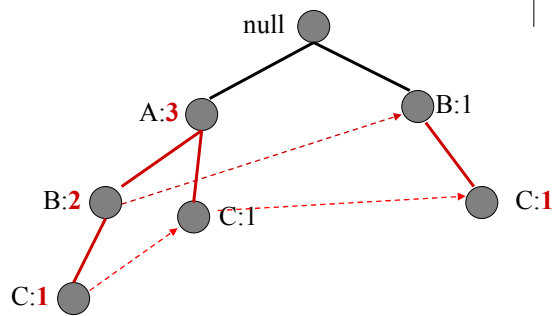


Αλγόριθμος FP-Growth



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



Παρατηρήσεις

Η απόδοση του FP-Growth εξαρτάται από τον παράγοντα συμπίεσης του συνόλου των δεδομένων (compression 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

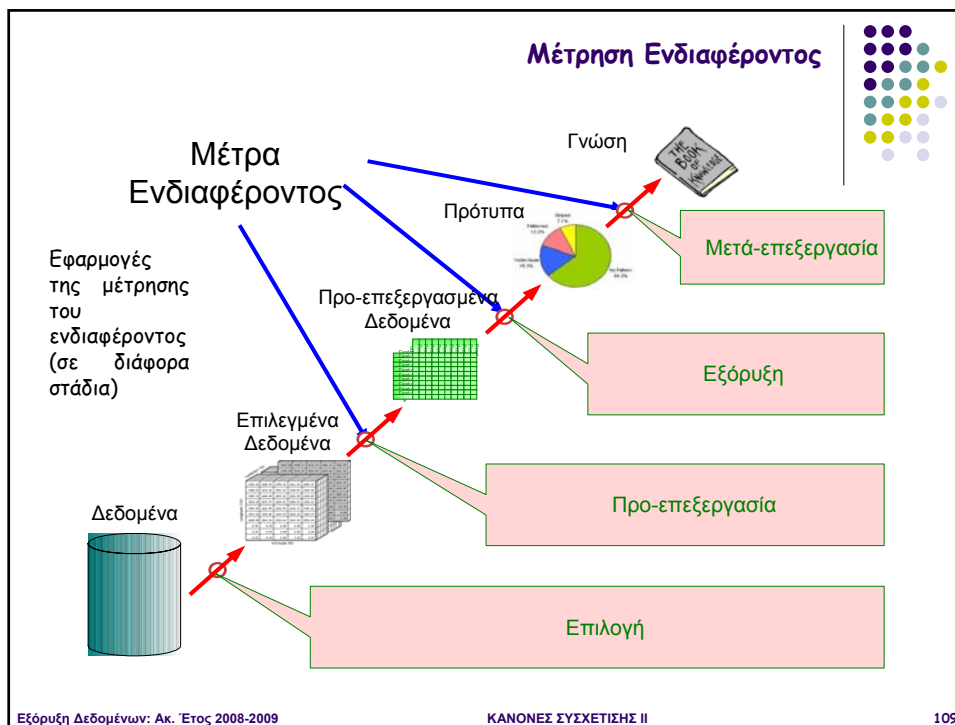


Παράγουν πάρα πολλούς κανόνες που συχνά είναι μη ενδιαφέροντες ή πλεονάζοντες (περιττοί)

Πλεονάζοντες αν $\{A, B, C\} \rightarrow \{D\}$ και $\{A, B\} \rightarrow \{D\}$
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



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)

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

110

Μέτρηση Ενδιαφέροντος



Υπολογισμός του Μέτρου Ενδιαφέροντος (αντικειμενικά μέτρα)

Contingency table (πίνακας συνάφειας)

	Y	\bar{Y}	
X	f_{11}	f_{10}	f_{1+}
\bar{X}	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	T

Μέτρηση συχνότητας εμφάνισης

f_{11} : support of X and Y

f_{10} : support of \bar{X} and \bar{Y}

f_{01} : support of X and \bar{Y}

f_{00} : support of \bar{X} and 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

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

Μέτρηση Ενδιαφέροντος



Μειονεκτήματα της Εμπιστοσύνης

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

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

Τι είναι μια καλή τιμή για την εμπιστοσύνη;

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

Κανόνας Συσχέτισης: Tea \rightarrow Coffee

Εμπιστοσύνη = $P(\text{Coffee}|\text{Tea}) = 0.75$

Ενώ ο κανόνας έχει υψηλή εμπιστοσύνη, ο κανόνας είναι παραπλανητικός

$P(\text{Coffee}|\bar{\text{Tea}}) = 0.9375$

$P(\text{Coffee}) = 0.9$

Αγνοεί την υποστήριξη του RHS



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



Στατιστική Ανεξαρτησία

Πληθυσμός 1000 σπουδαστών

- 600 σπουδαστές ξέρουν κολύμπι (S)
- 700 σπουδαστές ξέρουν ποδήλατο (B)
- 420 σπουδαστές ξέρουν κολύμπι και ποδήλατο (S, B)

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

- $P(S \wedge B) = P(S) \times P(B) \Rightarrow$ Στατιστική ανεξαρτησία
- $P(S \wedge B) > P(S) \times P(B) \Rightarrow$ Positively correlated (Θετική συσχέτιση)
- $P(S \wedge B) < P(S) \times P(B) \Rightarrow$ Negatively correlated (αρνητική συσχέτιση)

Μέτρα βασισμένα στη Στατιστική



Μέτρα που λαμβάνουν υπ' όψιν τους τη στατιστική εξάρτηση

Για τη συσχέτιση: $X \rightarrow Y$

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

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

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

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

Μέτρα βασισμένα στη Στατιστική



Παράδειγμα: Lift/Interest

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

Κανόνας συσχέτιση: Tea \rightarrow Coffee

Εμπιστοσύνη= $P(\text{Coffee}|\text{Tea}) = 0.75$

αλλά $P(\text{Coffee}) = 0.9$

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

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)

Μέτρα βασισμένα στη Στατιστική



φ-Coefficient

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

Κανονικοποιημένη τιμή μεταξύ του -1 και 1

Διαδική εκδοχή του Pearson's coefficient

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



φ-Coefficient

	Y	\bar{Y}	
X	60	10	70
\bar{X}	10	20	30
	70	30	100

	Y	\bar{Y}	
X	20	10	30
\bar{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



φ-Coefficient

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

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



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

#	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)	$\frac{\sum_j \max_k P(A_j, B_k) - \max_j P(A_j) - \max_k P(B_k)}{2 - \max_j P(A_j) - \max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A, B)P(\bar{A}, \bar{B})}{P(\bar{A}, B)P(A, \bar{B})}$
4	Yule's Q	$\frac{P(A, B)P(\bar{A}\bar{B}) - P(A, \bar{B})P(\bar{A}, B)}{P(A, B)P(\bar{A}\bar{B}) + P(A, \bar{B})P(\bar{A}, B)} = \frac{\alpha - 1}{\alpha + 1}$
5	Yule's Y	$\frac{\sqrt{P(A, B)P(\bar{A}\bar{B})} - \sqrt{P(A, \bar{B})P(\bar{A}, B)}}{\sqrt{P(A, B)P(\bar{A}\bar{B})} + \sqrt{P(A, \bar{B})P(\bar{A}, B)}} = \frac{\sqrt{\alpha - 1}}{\sqrt{\alpha + 1}}$
6	Kappa (κ)	$\frac{P(A, B) + P(\bar{A}, \bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}{\sum_i \sum_j P(A_i, B_j) \log P(A_i, B_j)}$
8	J-Measure (J)	$\min(-\sum_i P(A_i) \log P(A_i), -\sum_j P(B_j) \log P(B_j))$ $\max\left(P(A, B) \log\left(\frac{P(A, B)}{P(A)P(B)}\right) + P(\bar{A}\bar{B}) \log\left(\frac{P(\bar{A}\bar{B})}{P(\bar{A})P(\bar{B})}\right),\right.$ $\left.P(A, B) \log\left(\frac{P(A, B)}{P(A)P(B)}\right) + P(\bar{A}\bar{B}) \log\left(\frac{P(\bar{A}\bar{B})}{P(\bar{A})P(\bar{B})}\right)\right)$
9	Gini index (G)	$\max\left(P(A)[P(B A)^2 + P(\bar{B} A)^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2,\right.$ $\left.P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] - P(A)^2 - P(\bar{A})^2\right)$
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)	$\max\left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{A}B)}\right)$
14	Interest (I)	$\frac{P(A, B)}{P(\bar{A})P(\bar{B})}$
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)	$\frac{P(A, B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A, B) - P(\bar{A}\bar{B})}$
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

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

123

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

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

124

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 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

Ιδιότητες Μέτρων Αποτίμησης



Κλιμάκωση Γραμμής/Στήλης (Row/Column Scaling)

Παράδειγμα Βαθμός-Φύλο (Mosteller, 1968):

		κ_3	κ_4	
		Male	Female	
κ_1	High	2	3	5
κ_2	Low	1	4	5
		3	7	10

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

↓ $2x$
↓ $10x$

Mosteller:

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

Invariant under the row/column scaling operation αν $M(T) = M(T')$ όπου T ο πίνακας contingency με μετρητές συχνότητας $[f_{11}, f_{10}, f_{01}, f_{00}]$ και T' ο πίνακας contingency με μετρητές συχνότητας $[\kappa_1\kappa_3f_{11}, \kappa_2\kappa_3f_{10}, \kappa_1\kappa_4f_{01}, \kappa_2\kappa_4f_{00}]$ όπου $\kappa_1, \kappa_2, \kappa_3, \kappa_4$ θετικές σταθερές

Ιδιότητες Μέτρων Αποτίμησης



Αντιστροφή (Inversion Operation)

Δοσοληψία 1 →	1	0		0	1		0	0
⋮	0	0		1	1		1	0
⋮	0	0		1	1		1	0
⋮	0	0		1	0		1	1
⋮	0	1		1	1		1	0
⋮	0	0		1	1		1	0
⋮	0	0		1	1		1	0
⋮	0	0		1	1		1	0
⋮	0	0		1	1		1	0
Δοσοληψία N →	1	0		0	1		0	0

Invariant under the inversion operation αν η τιμή της παραμένει η ίδια αν ανταλλάξουμε τις τιμές f_{11} και f_{00} και τις τιμές f_{10} και f_{01}

Χρήσιμο για συμμετρικές μεταβλητές

Ιδιότητες Μέτρων Αποτίμησης



Null Addition (προσθήκη μη σχετιζόμενων στοιχείων)

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

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

Invariant measures:

- ◆ support, cosine, Jaccard, κλπ

Non-invariant measures:

- ◆ correlation, Gini, mutual information, odds ratio, κλπ

Αποτίμηση Κανόνων Συσχέτισης



Παράδοξο του Simpson

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

$$c(\{HDTV=Yes\} \rightarrow \{EM=Yes\})=99/180=55\%$$

$$c(\{HDTV=No\} \rightarrow \{EM=Yes\})=54/120=45\%$$

$$c(\{HTVS=Yes\} \rightarrow \{EM=Yes\})=98/170=57.7\%$$

$$c(\{HTVS=No\} \rightarrow \{EM=Yes\})=50/86=58.1\%$$

Students

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	1	9	10
No	4	30	34
	5	39	44

$$c(\{HDTV=Yes\} \rightarrow \{EM=Yes\})=1/10=10\%$$

$$c(\{HDTV=No\} \rightarrow \{EM=Yes\})=4/34=11.8\%$$

Working adults

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	98	72	170
No	50	36	86
	148	108	256

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

129

Αποτίμηση Κανόνων Συσχέτισης



Παράδοξο του Simpson

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	99 a+p	81	180 b+q
No	54 c+r	66	120 d+s
	153	147	300

$$c(\{HDTV=Yes\} \rightarrow \{EM=Yes\})=99/180=55\%$$

$$c(\{HDTV=No\} \rightarrow \{EM=Yes\})=54/120=45\%$$

$$a/b < c/d$$

$p/q < r/s$ δεν συνεπάγεται ότι

$$(a+p)/(b+q) < (c+r)/(d+s)!$$

Είναι σημαντικό πως θα γίνει διαχωρισμός (stratification) των δεδομένων

Students

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	1 a	9	10 b
No	4 c	30	34 d
	5	39	44

$$c(\{HDTV=Yes\} \rightarrow \{EM=Yes\})=1/10=10\%$$

$$c(\{HDTV=No\} \rightarrow \{EM=Yes\})=4/34=11.8\%$$

Working adults

Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	98 p	72	170 q
No	50 r	36	86 s
	148	108	256

$$c(\{HDTV=Yes\} \rightarrow \{EM=Yes\})=98/170=57.7\%$$

$$c(\{HDTV=No\} \rightarrow \{EM=Yes\})=50/86=58.1\%$$

Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

130

Υποκειμενικά Μέτρα Ενδιαφέροντος

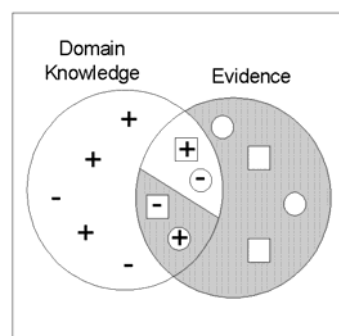


- Αντικειμενικά Μέτρα:
 - Ιεραρχούν τα αποτελέσματα με βάση στατιστικά στοιχεία που υπολογίζονται από τα δεδομένα
πχ., 21 μετρήσεις συσχέτισης (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).
- Υποκειμενικά Μέτρα:
 - Ιεράρχηση των προτύπων με βάση την ερμηνεία του χρήστη
 - Ένα 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



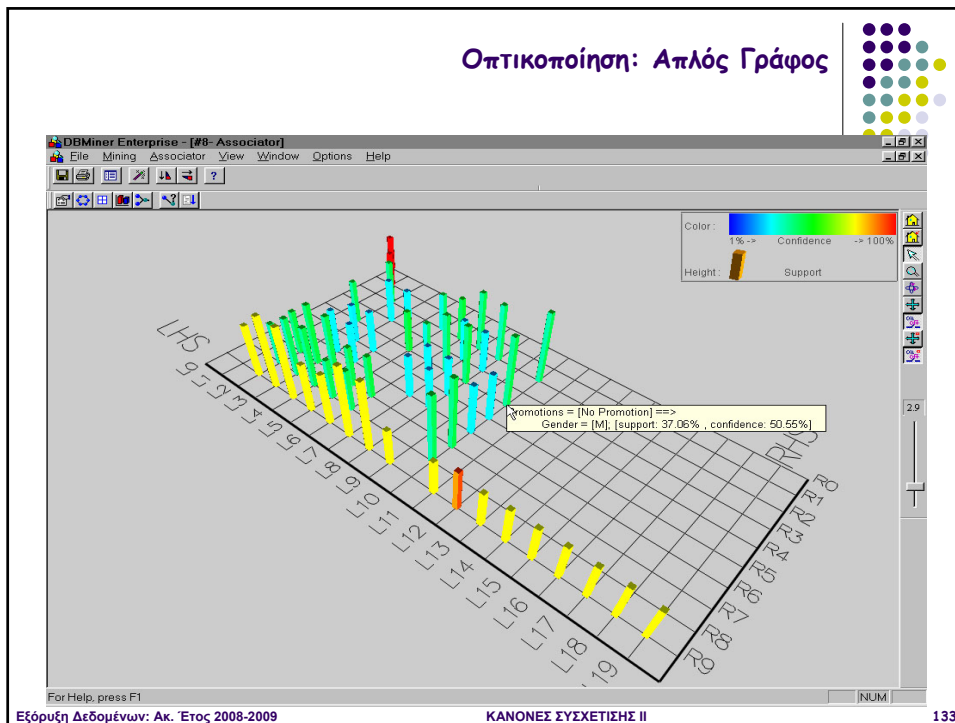
Interestingness (ενδιαφέρον) via Unexpectedness (μη αναμονή)



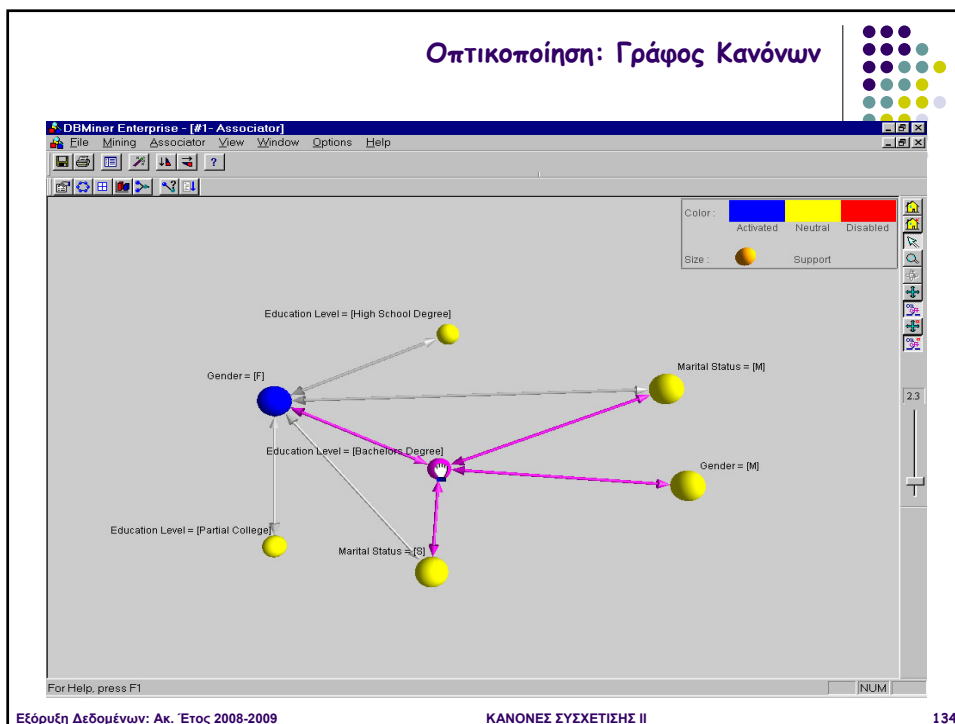
- + Pattern expected to be frequent
- Pattern expected to be infrequent
- Pattern found to be frequent
- Pattern found to be infrequent
- ⊕ ⊖ Expected Patterns
- ⊖ ⊕ Unexpected Patterns

- 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)

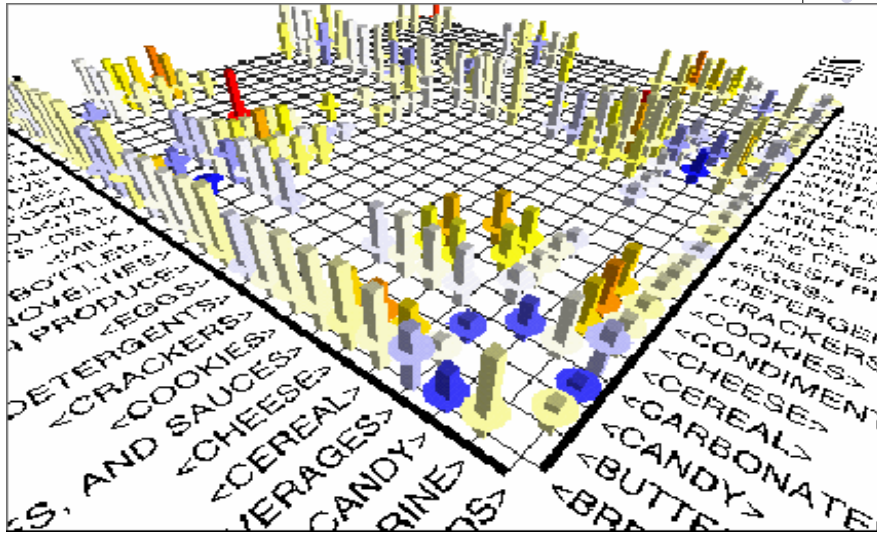
Οπτικοποίηση: Απλός Γράφος



Οπτικοποίηση: Γράφος Κανόνων



Οπτικοποίηση: (SGI/MineSet 3.0)



Εξόρυξη Δεδομένων: Ακ. Έτος 2008-2009

ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ II

135