Κανόνες Συσχέτισης ΙΙΙ

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

Σύντομη Ανακεφαλαίωση

Εισαγωγή

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

συναλλαγή (transaction)

στοιχείο (item)

- Προώθηση προϊόντων
- Τοποθέτηση προϊόντων στα ράφια
- Διαχείριση αποθεμάτων

Το πρόβλημα: Δεδομένου ενός συνόλου συναλλαγών (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}

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 συναλλαγών 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
$X \subseteq I$, $Y \subseteq I$, $X \cap Y = \varnothing$

Παράδειγμα: {Milk, Diaper} → {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 συναλλαγών που περιέχουν και το X και το Y ($X \cup Y$)
- Εμπιστοσύνη - Confidence (c)
  Πόσες από τις συναλλαγές (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 ≥ *minsup*
confidence ≥ *minconf*

Εξόρυξη Κανόνων Συσχέτισης

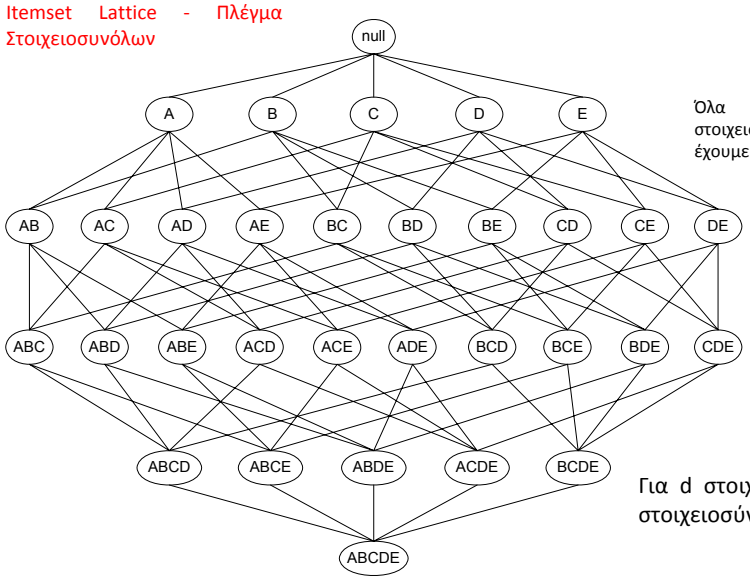
Χωρισμός του προβλήματος σε δύο υπο-προβλήματα:

1. Εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation)
   Εύρεση όλων των στοιχειοσυνόλων με υποστήριξη ≥ minsup

2. Δημιουργία Κανόνων (Rule Generation)
   Για κάθε (συχνό) στοιχειοσύνολο, δημιούργησε κανόνες με μεγάλη υποστήριξη, όπου κάθε κανόνας είναι μια δυαδική διαμέριση (δηλ. χωρισμός στα δύο) του συχνού στοιχειοσυνόλου

Εύρεση Συχνών Στοιχειοσυνόλων

Itemset Lattice - Πλέγμα Στοιχειοσυνόλων



Όλα τα δυνατά στοιχειοσύνολα όταν έχουμε 5 στοιχεία

Για d στοιχεία, 2^d πιθανά στοιχειοσύνολα

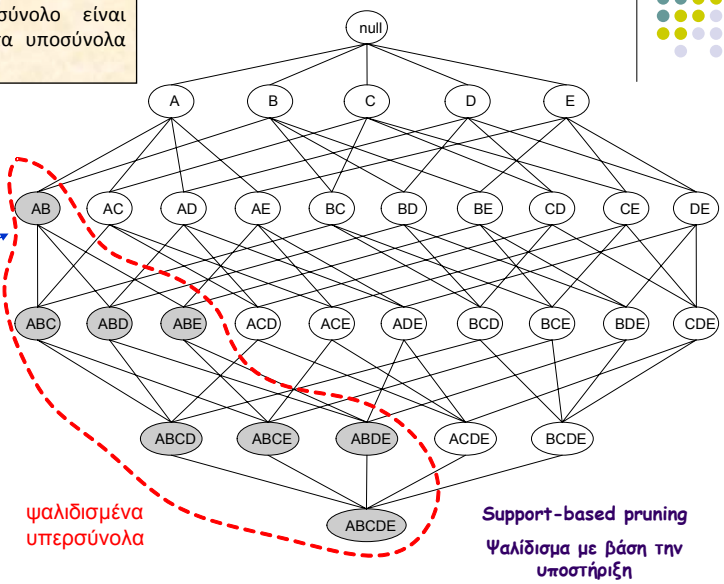
Εύρεση Συχνών Στοιχειοσυνόλων: Στρατηγική apriori

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

Στρατηγική apriori

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 συναλλαγών
4. Σβήσε τα υποψήφια στοιχειοσύνολα που δεν είναι συχνά
5. k = k + 1

Στρατηγική apriori: 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 apriori: Υπολογισμός Υποστήριξης

Για κάθε νέο υποψήφιο k+1-στοιχειοσύνολο, πρέπει να υπολογίσουμε την υποστήριξή 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 συναλλαγή με κάθε υποψήφιο στοιχειοσύνολο,
  - κατακερματίζουμε όλα τα k+1-στοιχειοσύνολα της συναλλαγής και
    - για καθένα, ενημερώνουμε μόνο τους αντίστοιχους κάδους του δέντρου κατακερματισμού των συχνών στοιχειοσυνόλων

Αναπαράσταση Στοιχειοσυνόλων

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

Ποια 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 συχνών):

- Maximal συχνά
- Κλειστά συχνά

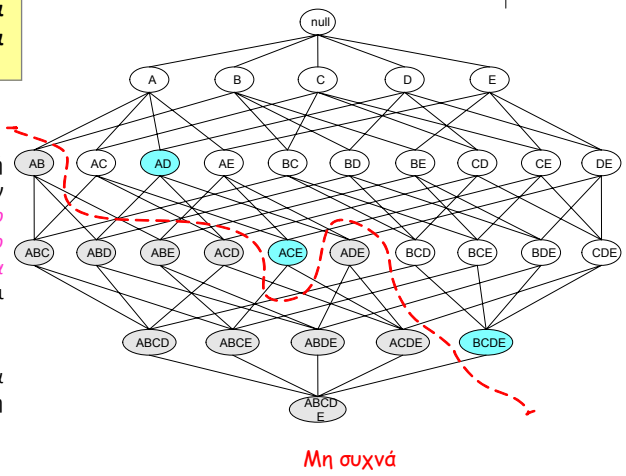
Αναπαράσταση Στοιχειοσυνόλων

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)

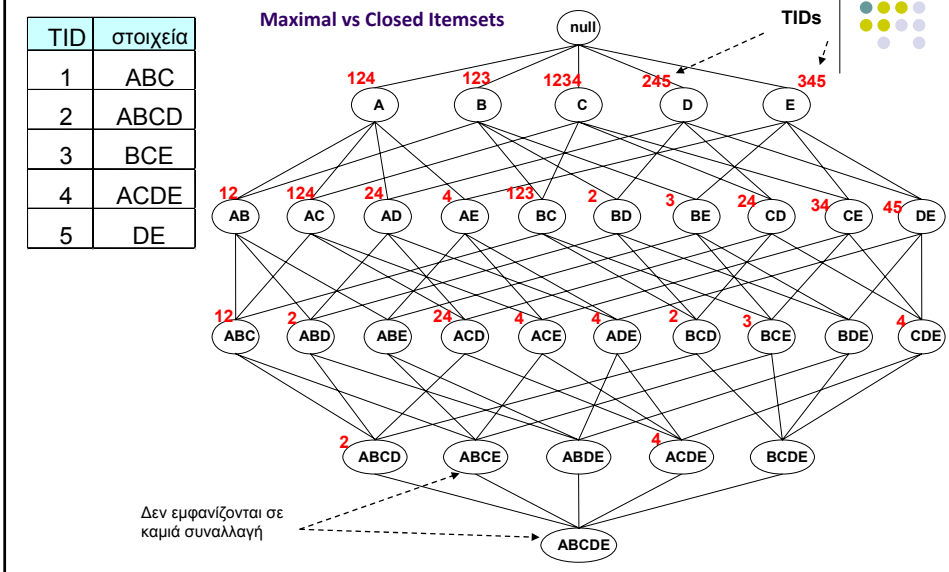
Ένα στοιχειοσύνολο είναι **κλειστό *συχνό* στοιχειοσύνολο** αν είναι κλειστό και συχνό (δηλαδή, η υποστήριξη του είναι μεγαλύτερη ή ίση με minsup)

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

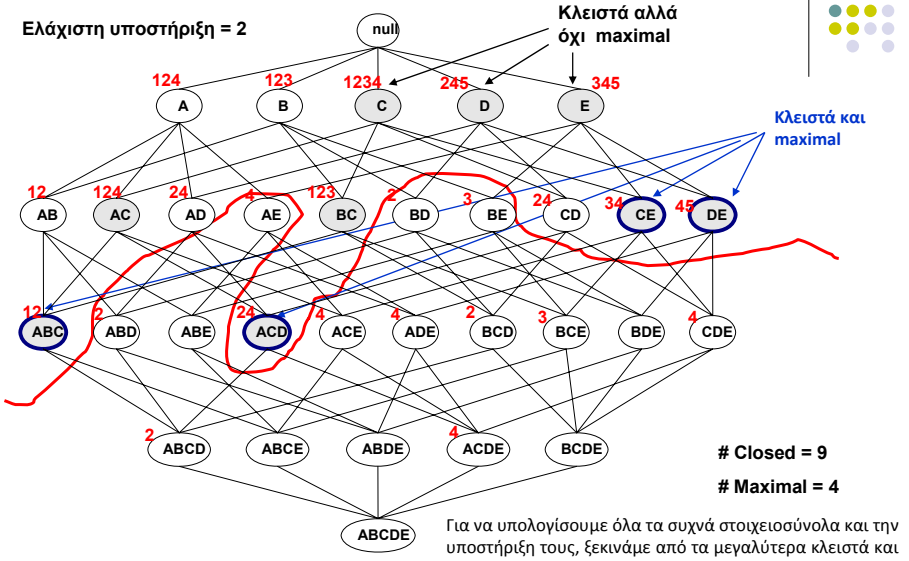
TID	στοιχεία
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Αναπαράσταση Στοιχειοσυνόλων

Maximal vs Closed Itemsets

Ελάχιστη υποστήριξη = 2



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

Εναλλακτικός Υπολογισμός Συχνών Στοιχειοσυνόλων

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 την συναλλαγή σε ένα μονοπάτι του FP-δέντρου
- Μερικά μονοπάτια μπορεί να επικαλύπτονται: όσο περισσότερα μονοπάτια επικαλύπτονται, τόσο καλύτερη συμπίεση

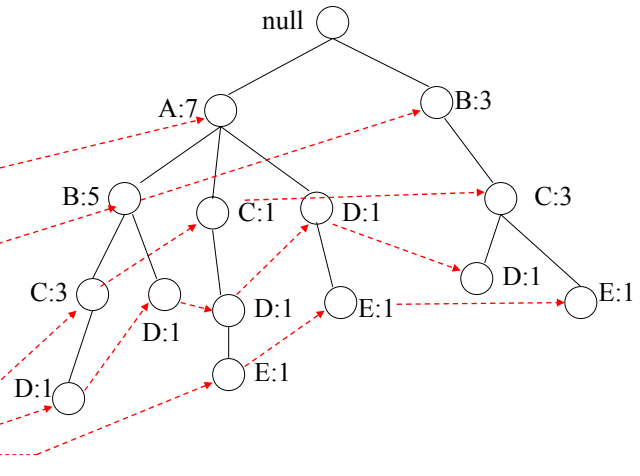
Μόλις κατασκευαστεί το FP-δέντρο, ο αλγόριθμος χρησιμοποιεί μια αναδρομική διαίρει-και-βασίλευε (divide-and-conquer) προσέγγιση για την εξόρυξη των συχνών στοιχειοσυνόλων

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}

Πίνακας Δεικτών

Item	Pointer
A	
B	
C	
D	
E	



Αλγόριθμος FP-Growth

Αλγόριθμος εύρεσης συχνών στοιχειοσυνόλων

Είσοδος: FP-δέντρο

Έξοδος: Συχνά στοιχειοσύνολα και η υποστήριξη τους

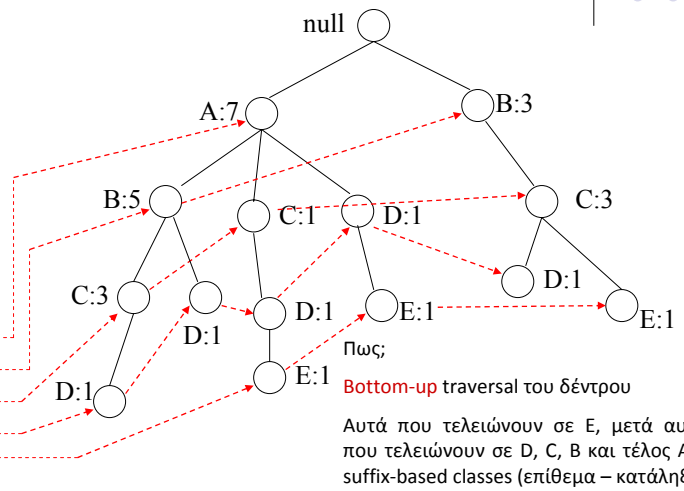
Μέθοδος:

- 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

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}

Αλγόριθμος FP-Growth

Χρήση FP-δέντρου για εύρεση συχνών στοιχειοσυνόλων



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)

Συνοπτικά

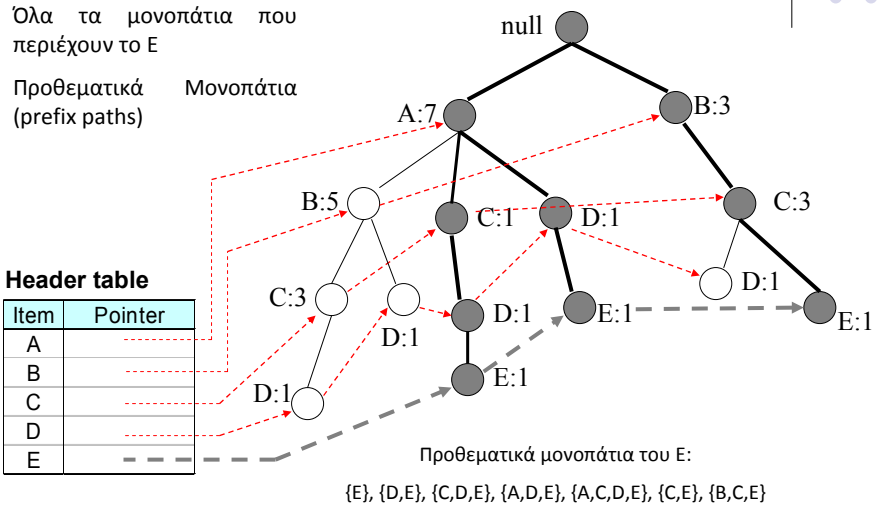
Σε κάθε βήμα, για το suffix (επίθεμα) X

- 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:

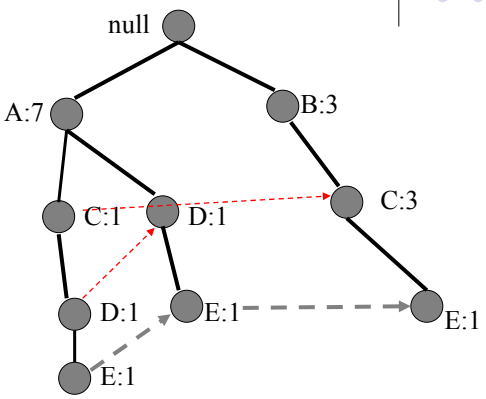
{E}, {D,E}, {C,D,E}, {A,D,E}, {A,C,D,E}, {C,E}, {B,C,E}

Αλγόριθμος FP-Growth

Φάση 1

Όλα τα μονοπάτια που περιέχουν το E

Προθεματικά Μονοπάτια (prefix paths)



Προθεματικά μονοπάτια του E:

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

Αλγόριθμος FP-Growth

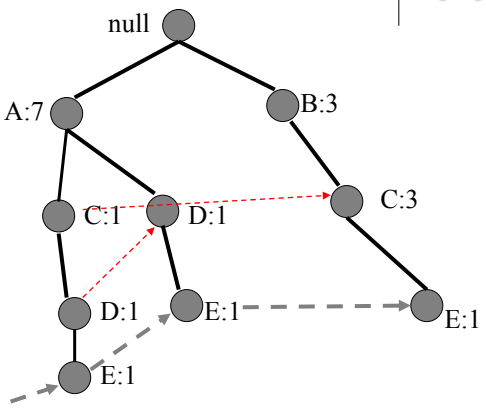
Έστω minsup = 2

Βρες την υποστήριξη του {E}

Πως;

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

Οπότε {E} συχνό



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

Αλγόριθμος FP-Growth

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

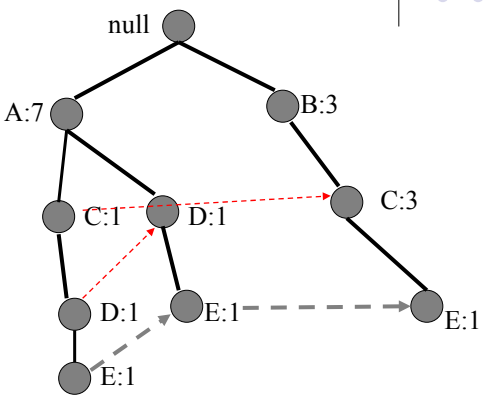
Φάση 2

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

Δύο αλλαγές

(1) Αλλαγή των μετρητών

(2) Περικοπή

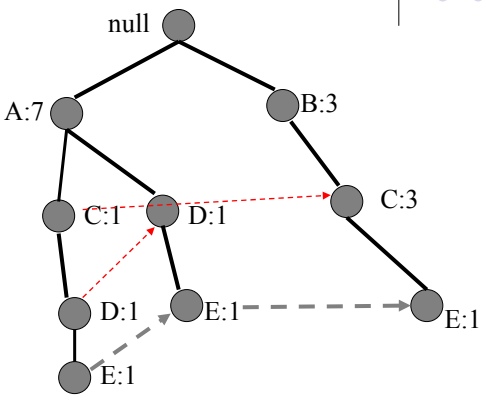


Αλγόριθμος FP-Growth

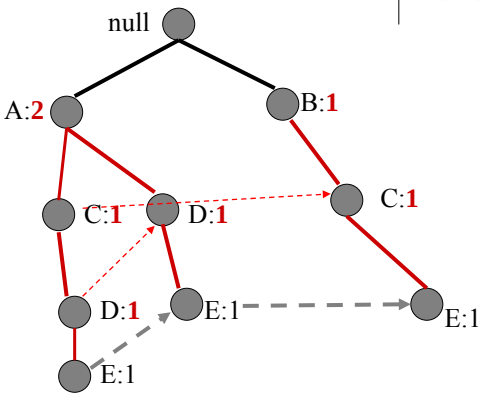
Αλλαγή μετρητών

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

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



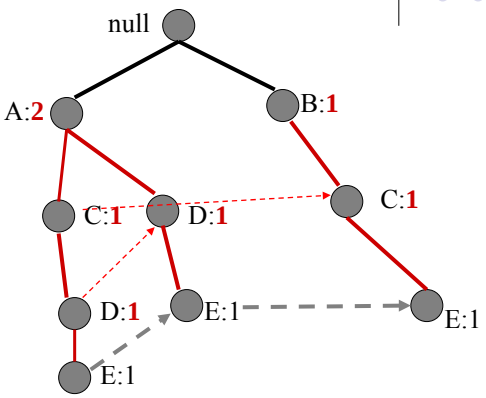
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Περικοπή (truncate)

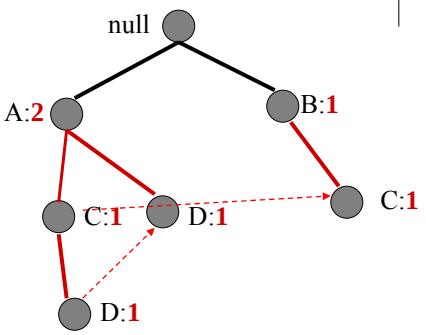
Σβήσε τους κόμβους του E



Αλγόριθμος FP-Growth

Περικοπή (truncate)

Σβήσε τους κόμβους του E



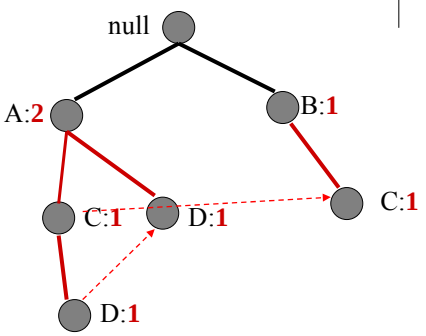
Αλγόριθμος FP-Growth

Πιθανή περαιτέρω περικοπή

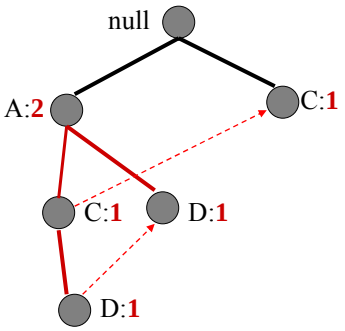
Κάποια στοιχεία μπορεί να έχουν υποστήριξη μικρότερη της ελάχιστης

Πχ το B -> περικοπή

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



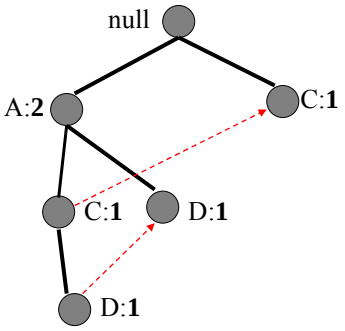
Αλγόριθμος FP-Growth



Αλγόριθμος FP-Growth

Υπο-συνθήκη FP-δέντρο για το E

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



Αλγόριθμος FP-Growth

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)

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

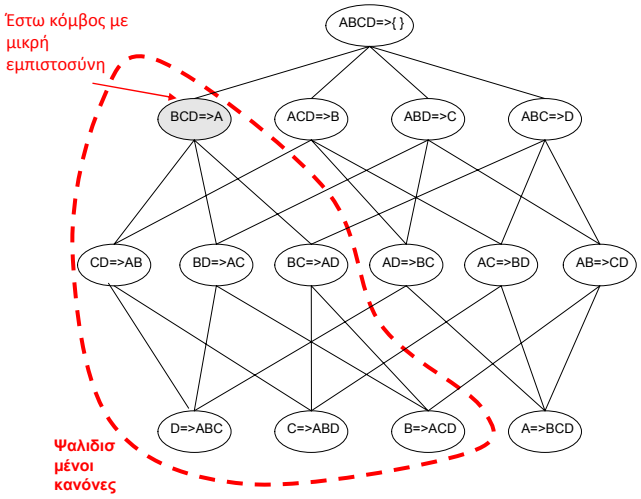
Για παράδειγμα L = {A,B,C,D}: $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

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

Παραγωγή Κανόνων για τον Αλγόριθμο apriori

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

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

Εκτίμηση Κανόνων Συσχέτισης

Εκτίμηση Κανόνων Συσχέτισης

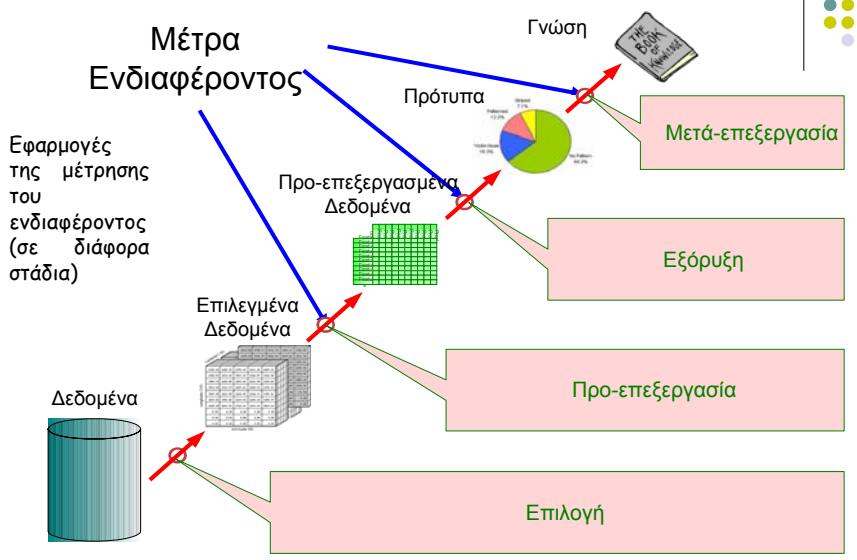
Παράγουν πάρα πολλούς κανόνες που συχνά είναι μη ενδιαφέροντες ή πλεονάζοντες (περιττοί)

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



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: Αντικειμενικά Μέτρα

Contingency table (πίνακας συνάφειας/πίνακας ενδεχομένων)

Μέτρηση συχνότητας εμφάνισης

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

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

Έστω ένας κανόνας, X → Y, η πληροφορία που χρειάζεται για τον υπολογισμό της εμπιστοσύνης και της 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	$\overline{\text{Coffee}}$	
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) = 0.75

Ενώ ο κανόνας έχει υψηλή εμπιστοσύνη, ο κανόνας είναι παραπλανητικός

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

P(Coffee) = 0.9

Αγνοεί την υποστήριξη του RHS (στην περίπτωση μας του coffee)

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

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 που λαμβάνουν υπ΄ όψιν τους τη στατιστική εξάρτηση

Για τη συσχέτιση: $X \rightarrow Y$

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

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

= 1, Στατιστική ανεξαρτησία
> 1, θετική συσχέτιση
< 1, αρνητική συσχέτιση

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

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

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 → 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)

Μέτρα βασισμένα στη Στατιστική

Μειονεκτήματα του Lift & Interest

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

	Y	$\overline{\text{Y}}$	
X	90	0	90
$\overline{\text{X}}$	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

ϕ-Coefficient

	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

$$\phi = \frac{0.6 - 0.7 \times 0.7}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238 \qquad \phi = \frac{0.2 - 0.3 \times 0.3}{\sqrt{0.7 \times 0.3 \times 0.7 \times 0.3}} = 0.5238$$

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 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)	$\max\left(P(A,B)\log(\frac{P(B\|A)}{P(B)}) + P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),\right.$ $\left. P(A,B)\log(\frac{P(A\|B)}{P(A)}) + P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\right)$
9	Gini index (G)	$\max\left(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]\right.$ $-P(B)^2-P(\overline{B})^2,$ $\left.P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]\right.$ $\left.-P(A)^2-P(\overline{A})^2\right)$
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)	$\max\left(\frac{P(A)P(\overline{B})}{P(A\overline{B})}, \frac{P(B)P(\overline{A})}{P(B\overline{A})}\right)$
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

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) να 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

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 Male	κ_4 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 $[\kappa_1\kappa_3 f_{11}, \kappa_2\kappa_3 f_{10}; \kappa_1\kappa_4 f_{01}; \kappa_2\kappa_4 f_{00}]$ όπου κ_1, κ_2, κ_3, κ_4 θετικές σταθερές

Ιδιότητες Μέτρων Αποτίμησης

Αντιστροφή (Inversion Operation)

	A	B
Συναλλαγή 1 →	1	0
	0	0
	0	0
	0	0
	0	1
	0	0
	0	0
	0	0
	0	0
Συναλλαγή N →	1	0

(a)

C	D
0	1
1	1
1	1
1	1
1	0
1	1
1	1
1	1
1	1
0	1

(b)

E	F
0	0
1	0
1	0
1	0
1	1
1	0
1	0
1	0
1	0
0	0

(c)

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 – πχ φ το ίδιο για A,B και C,D αλλά μικρότερο για E,F

Ιδιότητες Μέτρων Αποτίμησης

Null Addition (προσθήκη μη σχετιζόμενων στοιχείων)

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

⇒

	B	$\overline{B}$
A	p	q
$\overline{A}$	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} → {EM=Yes})=99/180=55%
c({HDTV=No} → {EM=Yes})=54/120=45%

Students

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

c({HDTV=Yes} → {EM=Yes})=1/10=10%
c({HDTV=No} → {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

c({HTVS=Yes} → {EM=Yes})=98/170=57.7%
c({HTVS=No} → {EM=Yes})=50/86=58.1%

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} → {EM=Yes})=99/180=55%
c({HDTV=No} → {EM=Yes})=54/120=45%

a/b < c/d

p/q < r/s δεν συνεπάγεται ότι

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

Students

c({HDTV=Yes} → {EM=Yes})=1/10=10%
c({HDTVS=No} → {EM=Yes})=4/34=11.8%

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

Working adults

c({HDTV=Yes} → {EM=Yes})=98/170=57.7%
c({HDTV=No} → {EM=Yes})=50/86=58.1%

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

Είναι σημαντικό πως θα γίνει διαχωρισμός (stratification) των δεδομένων

Υποκειμενικά Μέτρα Ενδιαφέροντος

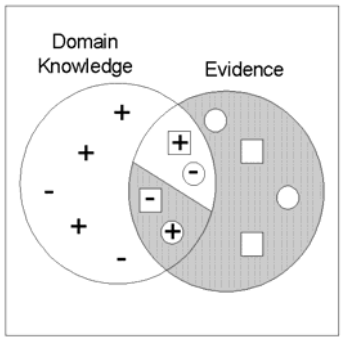
- Αντικειμενικά Μέτρα:
  - Ιεραρχούν τα αποτελέσματα με βάση στατιστικά στοιχεία που υπολογίζονται από τα δεδομένα

    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

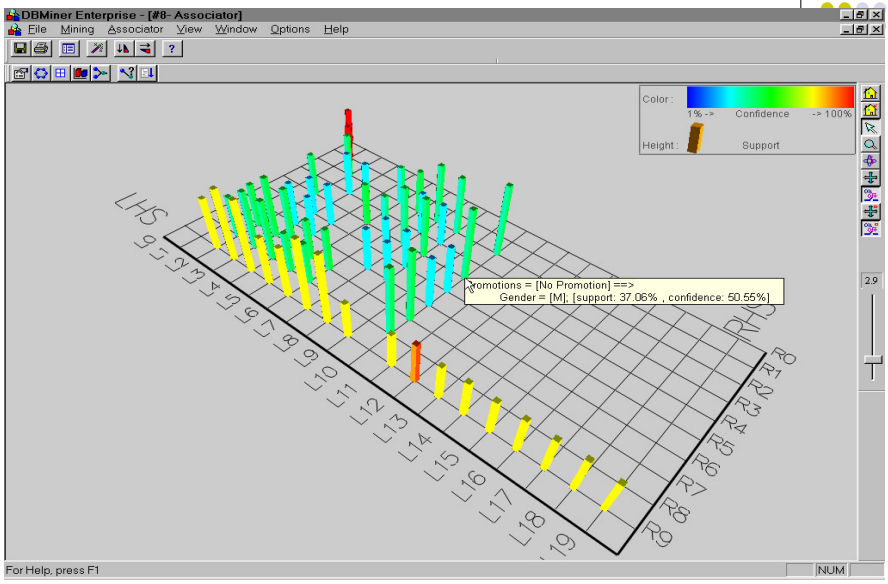
Pattern found to be infrequent

Expected Patterns

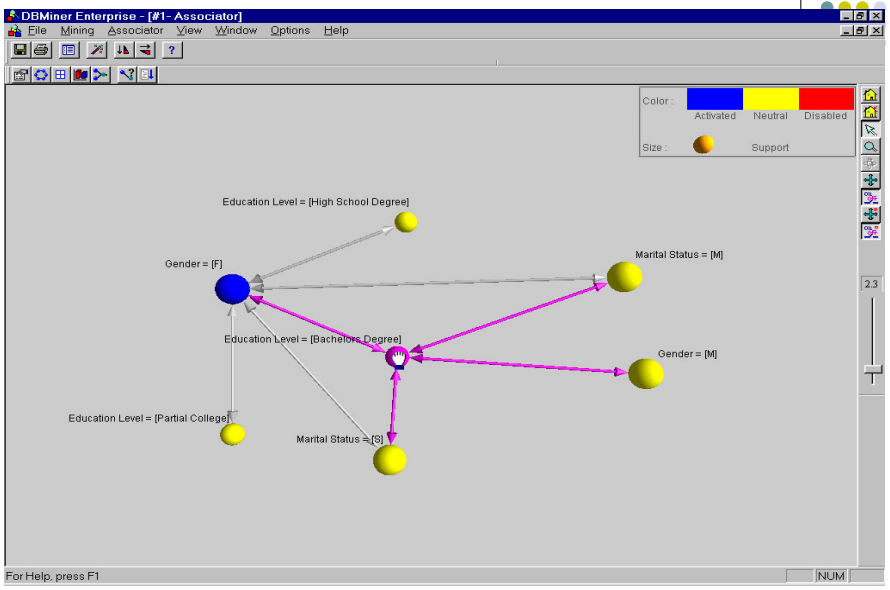
Unexpected Patterns

- 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)

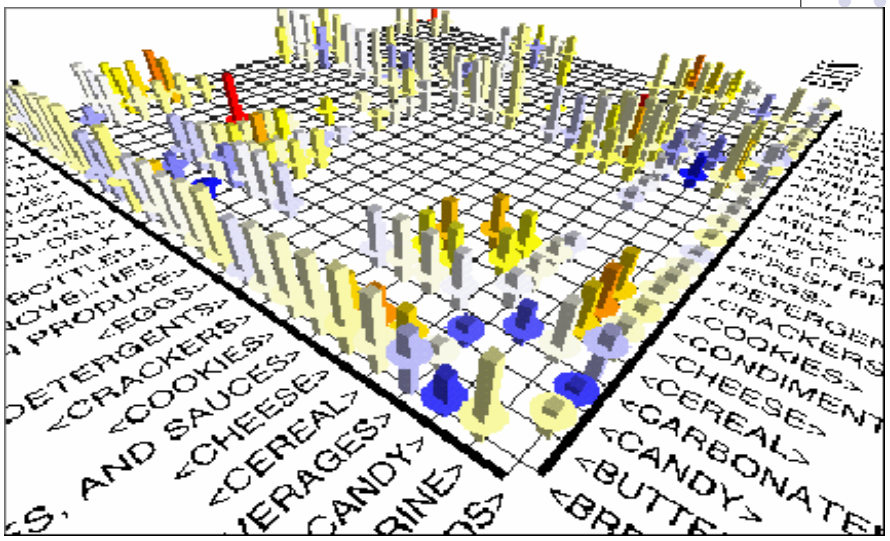
Οπτικοποίηση: Απλός Γράφος



Οπτικοποίηση: Γράφος Κανόνων



Οπτικοποίηση: (SGI/MineSet 3.0)



Επίδραση της «Λοξής Κατανομής» της Υποστήριξης

Κατανομή Υποστήριξης

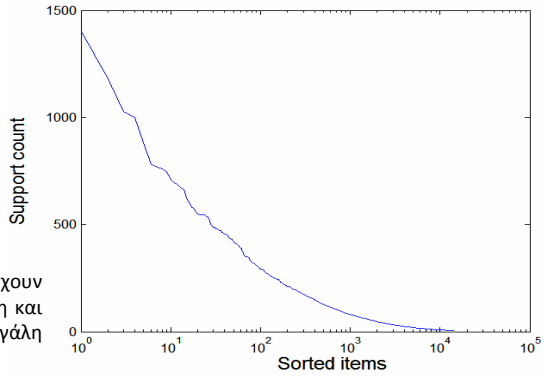
- Η απόδοση των αλγορίθμων εξαρτάται από τα δεδομένα εισόδου, πχ ο apriori από τον αριθμό των στοιχείων, το πλάτος των δοσοληψιών, ο FP-Growth από την τομή (κοινά στοιχεία) των δοσοληψιών
- Επίσης, από την τιμή της ελάχιστης υποστήριξης (minsup). Πως θα προσδιοριστεί μια κατάλληλη τιμή για το *minsup;*
  - Αν η τιμή είναι πολύ υψηλή, μπορεί να χαθούν στοιχειοσύνολα που περιέχουν ενδιαφέροντα σπάνια στοιχεία (πχ ακριβά προϊόντα)
  - Αν η τιμή είναι πολύ χαμηλή, οι μέθοδοι γίνονται ακριβοί γιατί ο αριθμός των υποψήφιων στοιχειοσυνόλων είναι πολύ μεγάλος και ο αριθμός των συχνών στοιχειοσυνόλων γίνεται πολύ μεγάλος

Κατανομή Υποστήριξης

Επιπρόσθετα, η χρήση μόνο μίας ελάχιστης υποστήριξης μπορεί να μην αρκεί

Για πολλά πραγματικά δεδομένα η κατανομή της υποστήριξης δεν είναι ομοιόμορφη (skewed support distribution)

Κατανομή υποστήριξης για δεδομένα λιανικών πωλήσεων



Τα περισσότερα στοιχεία έχουν μικρή ή μέτρια υποστήριξη και μόνο λίγα έχουν μεγάλη υποστήριξη

Κατανομή Υποστήριξης

Ομάδα	G1	G2	G3
Υποστήριξη	<1%	1%-90%	>90%
Αριθμός στοιχείων	1735	358	20

Πως θα βρούμε κανόνες με «σπάνια» αλλά ενδιαφέροντα στοιχεία;

Πολύ μικρή υποστήριξη;

- πολυπλοκότητα (πολλά υποψήφια στοιχειοσύνολα + πολλά συχνά στοιχειοσύνολα άρα και κανόνες)
- παράξενοι κανόνες μεταξύ G1 και G3 (χαβιάρι και γάλα) πχ support = 0.05 -> 18,847 συχνά ζεύγη (από τα οποία μεικτά (διασταυρωμένης υποστήριξης το 93%)

Cross-support patterns (υποδείγματα διασταυρωμένης υποστήριξης) – ανάμιξη στοιχείων πολύ συχνών με στοιχεία που είναι σπάνια!

$$\min\{s(i_1), s(i_2), \ldots, s(i_k)\}/\max\{s(i_1), s(i_2), \ldots, s(i_k)\}$$

Κατανομή Υποστήριξης

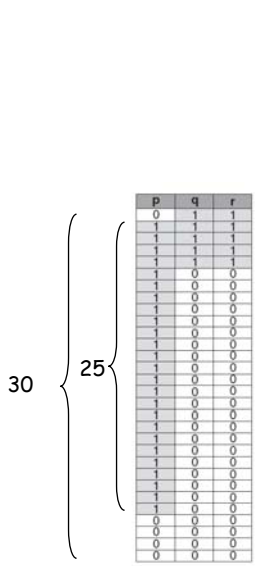


Figure 6.30. A transaction data set containing three items, p, q, and r, where p is a high support item and q and r are low support items.

υποστήριξη

{p, q, r} s = 4/30

{p, q} s = 4/30

{p, r} s = 4/30

{q, r} s = 5/30

εμπιστοσύνη

{p, q, r}

{p, q} $p \rightarrow q$, c = 4/25 $q \rightarrow p$ c = 4/5

{p, r} 4/30

{q, r} $q \rightarrow r$ c = 5/5 $r \rightarrow q$ c = 5/5

Υπάρχει ένας κανόνας με μικρή εμπιστοσύνη – ο εντοπισμός του δηλώνει ότι πρόκειται για cross-support

Κατανομή Υποστήριξης

Cross-support patterns – ανάμιξη στοιχείων πολύ συχνών με στοιχεία που είναι σπάνια!

$\min\{s(i_1), s(i_2), \ldots, s(i_k)\}/\max\{s(i_1), s(i_2), \ldots, s(i_k)\}$

Πως να απαλλαγούμε

Να θεωρήσουμε τον κανόνα με τη μικρότερη δυνατή εμπιστοσύνη ανάμεσα στους κανόνες με στοιχεία από το $\{i_1, i_2, \ldots, i_k\}$

Ποιος είναι αυτός

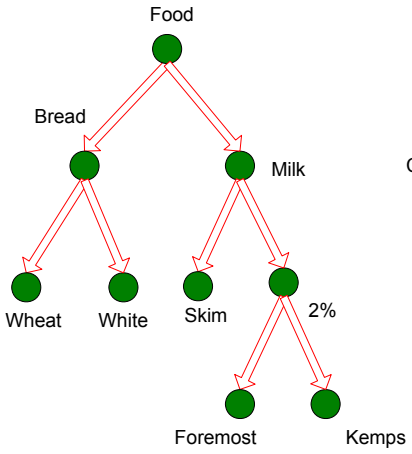
ένα στοιχείο στο LHS

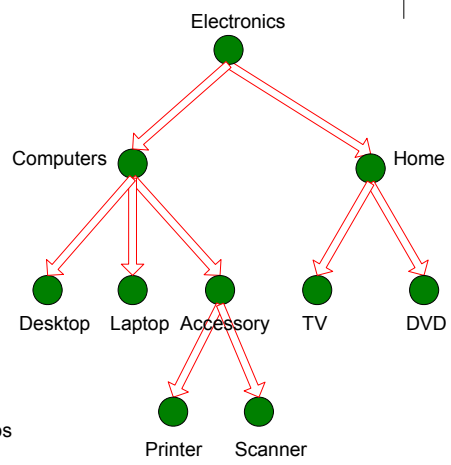
ποιο στοιχείο: αυτό με τη μεγαλύτερη υποστήριξη!

$h_c = s\{i_1, i_2, \ldots, i_k\}/\max_k\{s(i1), s(i2), \ldots, s(ik)\}$

Κανόνων Συσχέτισης Πολλαπλών Επιπέδων

Κανόνες Συσχέτισης Πολλών Επιπέδων





Κανόνες Συσχέτισης Πολλών Επιπέδων

Γιατί είναι χρήσιμοι;

- Οι κανόνες στα χαμηλότερα επίπεδα δεν έχουν αρκετή υποστήριξη σε κανένα στοιχειοσύνολο
- Οι κανόνες στα χαμηλότερα επίπεδα είναι πάρα πολύ συγκεκριμένοι και στα υψηλότερα επίπεδα πολύ γενικοί!
  - π.χ., skim milk → white bread, 2% milk → wheat bread, skim milk → wheat bread, κλπ.

  είναι ενδεικτικοί της συσχέτισης μεταξύ γάλατος και ψωμιού

Κανόνες Συσχέτισης Πολλών Επιπέδων

- Προσέγγιση 1:
  - Επέκταση κάθε συναλλαγής με στοιχεία από τα υψηλότερα επίπεδα της ιεραρχίας

    Αρχική Συναλλαγή: {skim milk, wheat bread}

    ΕπαυξημένηΣυναλλαγή: {skim milk, wheat bread, milk, bread, food}

- Θέματα:
  - Τα στοιχεία στα υψηλότερα επίπεδα θα εμφανίζονται πολύ συχνά, μεγάλους μετρητές υποστήριξης
    - μικρή υποστήριξη, θα οδηγούσε σε πολλά συχνά στοιχειοσύνολα από τα υψηλότερα επίπεδα
  - Αύξηση της διάστασης των δεδομένων
  - Πλεονάζοντες κανόνες

Κανόνες Συσχέτισης Πολλών Επιπέδων

- Πως τροποποιούνται η υποστήριξη και η εμπιστοσύνη στην ιεραρχία;
  - Αν X ο γονέας των X1 and X2, τότε
    $\sigma(X) \leq \sigma(X1) + \sigma(X2)$
  - Αν $\sigma(X1 \cup Y1) \geq$ minsup,
    και X γονέας του X1, Y γονέας του Y1
    τότε $\sigma(X \cup Y1) \geq$ minsup, $\sigma(X1 \cup Y) \geq$ minsup
    $\sigma(X \cup Y) \geq$ minsup
  - Αν conf($X1 \Rightarrow Y1$) ≥ minconf,
    τότε conf($X1 \Rightarrow Y$) ≥ minconf

Κανόνες Συσχέτισης Πολλών Επιπέδων

- Προσέγγιση 2:
  - Δημιούργησε συχνά στοιχειοσύνολα πρώτα για τα υψηλότερα επίπεδα
  - Μετά, δημιούργησε στοιχειοσύνολα για το αμέσως επόμενο επίπεδο κοκ
- Θέματα:
  - I/O απαιτήσεις αυξάνουν, γιατί απαιτούνται πολλαπλά περάσματα
  - Μπορεί να χαθούν συσχετίσεις ανάμεσα στα επίπεδα

Πολλαπλές Τιμές Υποστήριξης

Πολλαπλές Τιμές Υποστήριξης

Πως θα Μπορούμε να έχουμε πολλές Ελάχιστες Τιμές Υποστήριξης

Ορισμός MS(i): ελάχιστη υποστήριξη για το στοιχείο i

- Π.χ.: MS(Milk)=5%, MS(Coke) = 3%,
  MS(Broccoli)=0.1%, MS(Salmon)=0.5%
- MS({Milk, Broccoli}) = min (MS(Milk), MS(Broccoli))
  = 0.1%

Πρόβλημα: Η υποστήριξη παύει να είναι αντιμονότονη:

- Έστω: Support(Milk, Coke) = 1.5% and
  Support(Milk, Coke, Broccoli) = 0.5%
- {Milk, Coke} είναι μη συχνό αλλά το {Milk, Coke, Broccoli} είναι συχνό

Λόγω του Broccoli που κατεβάζει το minsup

Πολλαπλές Τιμές Υποστήριξης

Multiple Minimum Support (Liu 1999)

- Ταξινόμησε τα στοιχεία με βάση την ελάχιστη τιμή υποστήριξης (σε αύξουσα διάταξη)
  - πχ.: MS(Milk)=5%, MS(Coke) = 3%, MS(Broccoli)=0.1%, MS(Salmon)=0.5%
  - Διάταξη: Broccoli, Salmon, Coke, Milk
- Τροποποίηση του Apriori έτσι ώστε:
  - L_1 : σύνολο συχνών στοιχειοσυνόλων
  - F_1 : σύνολο στοιχείων που η υποστήριξη τους είναι ≥ MS(1) όπου MS(1) είναι $min_i(MS(i))$
  - C_2 : τα υποψήφια στοιχειοσύνολα μεγέθους 2 παράγονται από το F_1 αντί του L_1

Πολλαπλές Τιμές Υποστήριξης

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

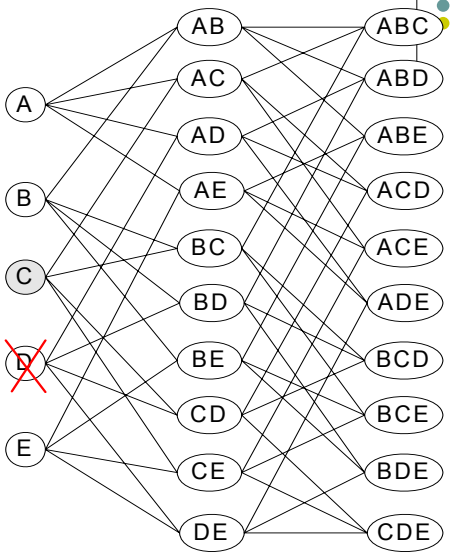
    πχ Candidate={Broccoli, Coke, Milk} (διατεταγμένα με βάση την μικρότερη ελάχιστη υποστήριξη)

    {Broccoli, Coke} και {Broccoli, Milk} είναι συχνά αλλά {Coke, Milk} είναι μη συχνό
      - Δε σβήνεται γιατί το {Coke, Milk} δεν περιέχει το πρώτο στοιχείο, δηλαδή, Broccoli.

Πολλαπλές Τιμές Υποστήριξης

επιθυμητό

Item	MS(I)	Sup(I)
A	0.10%	0.25%
B	0.20%	0.26%
C	0.30%	0.29%
D	0.50%	0.05%
E	3%	4.20%



Πολλαπλές Τιμές Υποστήριξης

Item	MS(I)	Sup(I)
A	0.10%	0.25%
B	0.20%	0.26%
C	0.30%	0.29%
D	0.50%	0.05%
E	3%	4.20%