Κανόνες Συσχέτισης Ι

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

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

δοσοληψία

- Προώθηση προϊόντων
- Τοποθέτηση προϊόντων στα ράφια
- Διαχείριση αποθεμάτων

Το πρόβλημα: Δεδομένου ενός συνόλου δοσοληψιών (transactions), βρες κανόνες που προβλέπουν την εμφάνιση ενός στοιχείου (item) με βάση την εμφάνιση άλλων στοιχείων στις συναλλαγές

Παραδείγματα κανόνων συσχέτισης

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk}

Σημαίνει ότι εμφανίζονται μαζί, όχι ότι η εμφάνιση του ενός είναι η αιτία της εμφάνισης του άλλου (co-occurrence, not causality όχι έννοια χρόνου ή διάταξης)

Εισαγωγή

Δυαδική αναπαράσταση

Γραμμές: δοσοληψίες

Στήλες: Στοιχεία

1 αν το στοιχείο εμφανίζεται στη σχετική δοσοληψία

Μη συμμετρική δυαδική μεταβλητή (1 πιο σημαντικό από το 0)

Παράδειγμα

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

Ορισμοί

- $I = \{i_1, i_2, .., i_k\}$ ένα σύνολο από διακριτά **στοιχεία (items)**
Παράδειγμα: {Bread, Milk, Diapers, Beer, Eggs, Coke}

- **Στοιχειοσύνολο (Itemset):** Ένα υποσύνολο του I
Παράδειγμα: {Milk, Bread, Diaper}
  - **k-στοιχειοσύνολο (k-itemset):** ένα στοιχειοσύνολο με *k* στοιχεία

- $T = \{t_1, i_2, .., t_N\}$ ένα σύνολο από **δοσοληψίες**, όπου κάθε t_i είναι ένα στοιχειοσύνολο

Πλάτος (width) δοσοληψίας: αριθμός στοιχείων ti περιέχει ένα στοιχειοσύνολο X, αν το X είναι υποσύνολο της t_i

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

- **support count (σ) ενός στοιχειοσυνόλου**
Η συχνότητα εμφάνισης του στοιχειοσυνόλου

Παράδειγμα: σ({Milk, Bread, Diaper}) = 2

- **Υποστήριξη (Support (s)) ενός στοιχειοσυνόλου**
Το ποσοστό των δοσοληψιών που περιέχουν ένα στοιχειοσύνολο

Παράδειγμα: s({Milk, Bread, Diaper}) = 2/5

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

- **Frequent Itemset**
Ένα στοιχειοσύνολο του οποίου η υποστήριξη είναι μεγαλύτερη ή ίση από κάποια τιμή κατωφλίου *minsup*

Ορισμοί

Κανόνας Συσχέτισης (Association Rule)
Είναι μια έκφραση της μορφής $X \rightarrow Y$, 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 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

$$s = \frac{\sigma\{\text{Milk, Diaper, Beer}\}}{|T|} = \frac{2}{5} = 0.4 \qquad \{\text{Milk, Diaper}\} \rightarrow \text{Beer}$$

$$c = \frac{\sigma\{\text{Milk, Diaper, Beer}\}}{\sigma\{\text{Milk, Diaper}\}} = \frac{2}{3} = 0.67$$

Εξόρυξη Κανόνων Συσχέτισης

Παρατηρήσεις

- $s(X \rightarrow Y) = s(X \cup Y) = \sigma(X \cup Y)/N$

Ένας κανόνας με μικρή υποστήριξη μπορεί να εμφανίζεται τυχαία
Εξαιρεί κανόνες που δεν έχουν ενδιαφέρον

- $c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$

$c(X \rightarrow Y) = P(Y|X)$ δεσμευμένη πιθανότητα να εμφανίζεται το Y όταν εμφανίζεται το X

Εμπιστοσύνη μετρά την αξιοπιστία

Όσο μεγαλύτερη εμπιστοσύνη τόσο μεγαλύτερη η πιθανότητα εμφάνισης του Y σε κανόνες που περιέχουν το X

Εξόρυξη Κανόνων Συσχέτισης

Εύρεση Κανόνων Συσχέτισης

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

Εξόρυξη Κανόνων Συσχέτισης

Brute-force approach:

- List all possible association rules
- Compute the support and confidence for each rule
- Prune rules that fail the *minsup* and *minconf* thresholds

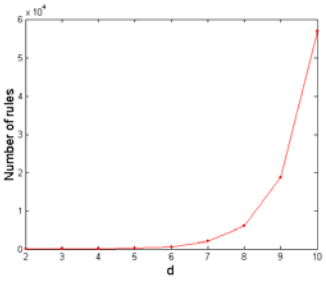
⇒Computationally prohibitive!

Για d στοιχεία, $3^d - 2^{d+1} + 1$

Εύρεση Συχνών Στοιχειοσυνόλων

Υπολογιστική Πολυπλοκότητα

- Given d unique items:
  - Total number of itemsets = 2^d (δυναμοσύνολο)
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right]$$
$$= 3^d - 2^{d+1} + 1$$

If d = 6, R = 602 rules

Εξόρυξη Κανόνων Συσχέτισης

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

Πιθανοί κανόνες με τα στοιχεία Milk, Diaper και Beer (στοιχειοσύνολο {Milk, Diaper, Beer})

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Η υποστήριξη ενός κανόνα $X \rightarrow Y$ εξαρτάται μόνο από την υποστήριξη του $X \cup Y$
Άρα κανόνες που ξεκινούν από το ίδιο στοιχειοσύνολο έχουν την ίδια υποστήριξη (αλλά πιθανών διαφορετική εμπιστοσύνη)
Αν είχαμε minsup = 0.5, θα αποκλείαμε και τους έξη κανόνες

Άρα μπορούμε να θεωρήσουμε τους περιορισμούς για την υποστήριξη και την εμπιστοσύνη ξεχωριστά

Εξόρυξη Κανόνων Συσχέτισης

Χωρισμός του προβλήματος σε <u>δύο υπο-προβλήματα</u>:

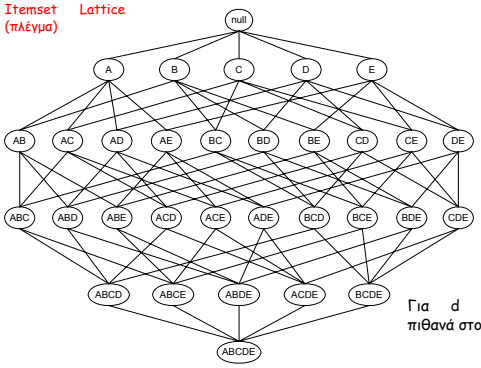
- Εύρεση όλων των συχνών στοιχειοσυνόλων (Frequent Itemset Generation)
  Εύρεση όλων των στοιχειοσυνόλων με υποστήριξη ≥ minsup
- Δημιουργία Κανόνων (Rule Generation)
  Για κάθε στοιχειοσύνολο, δημιούργησε κανόνες με μεγάλη υποστήριξη, όπου κάθε κανόνες είναι μια δυαδική διαμέριση του συχνού στοιχειοσυνόλου

Η δημιουργία των συχνών στοιχειοσυνόλων είναι επίσης υπολογιστικά ακριβή

Εύρεση Συχνών Στοιχειοσυνόλων

Εύρεση Συχνών Στοιχειοσυνόλων

Itemset (πλέγμα) Lattice



Για d στοιχεία, 2^d πιθανά στοιχειοσύνολα

Εύρεση Συχνών Στοιχειοσυνόλων

Brute-force approach:

- Κάθε στοιχειοσύνολο στο lattice είναι ένα υποψήφιο συχνό στοιχειοσύνολο
- Υπολόγισε την υποστήριξη κάθε υποψήφιου στοιχειοσυνόλου διατρέχοντας (scanning) της βάση δεδομένων

Transactions

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

List of Candidates

N

M

w

Ταίριαξε κάθε δοσοληψία με κάθε υποψήφιο
Πολυπλοκότητα ~ O(NMw) => Μεγάλη γιατί M = 2^d !!!

N: αριθμός δοσοληψιών
w: μέγιστο πλάτος δοσοληψίας

Εύρεση Συχνών Στοιχειοσυνόλων

Διαφορετικές Στρατηγικές

Ελάττωση του αριθμού των υποψηφίων στοιχειοσυνόλων (M)
- Πλήρης αναζήτηση: M=2^d
- Χρησιμοποίησε κάποια τεχνική pruning (κλαδέματος - ελάττωσης) για να ελαττωθεί το M (πχ apriori)

Ελάττωση του αριθμού των δοσοληψιών (N)
- Ελάττωση του μεγέθους του N καθώς το μέγεθος του στοιχειοσυνόλου αυξάνεται
- Used by DHP and vertical-based mining algorithms

Ελάττωση του αριθμού των συγκρίσεων (NM)
- Στόχος να αποφύγουμε να ταιριάξουμε κάθε υποψήφιο στοιχειοσύνολο με κάθε δοσοληψία
- Χρήση αποδοτικών δομών δεδομένων για την αποθήκευση των υποψηφίων στοιχειοσυνόλων ή των δοσοληψιών

Αρχή apriori

Ελάττωση συχνών στοιχεισυνόλων

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

Αντιθετοαντιστροφή: Αν ένα στοιχειοσύνολο δεν είναι συχνό, όλα τα υπερσύνολα του δεν είναι συχνά

Η αρχή Apriori ισχύει λόγω της παρακάτω ιδιότητας της υποστήριξης:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Η υποστήριξη ενός στοιχεισύνολου είναι μικρότερη ή ίση της υποστήριξης οποιουδήποτε υποσυνόλου του

Αρχή apriori

Αντι-μονότονη (anti-monotone) ιδιότητα της υποστήριξης

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

s: downwards closed

Μονότονη ιδιότητα ή upwards closed

$$\forall X, Y : (X \subseteq Y) \Rightarrow f(X) \leq f(Y)$$

Αρχή apriori

Όλα τα υποσύνολα του συχνά

(κλειστό από πάνω)

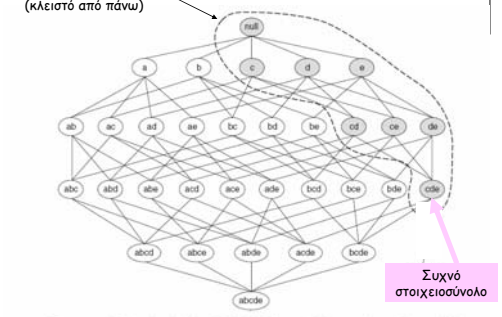
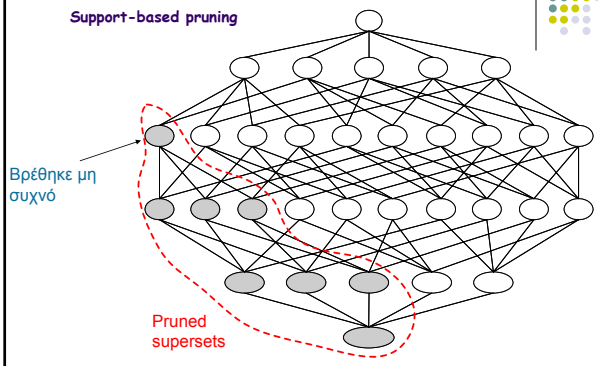


Figure 6.3. An illustration of the Apriori principle. If $\{c, d, e\}$ is frequent, then all subsets of this itemset are frequent.

Στρατηγική apriori

Support-based pruning



Στρατηγική apriori

Παράδειγμα

Minimum Support = 3

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

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Items (1-itemsets)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

Αν όλα τα δυνατά στοιχειοσύνολα:

$$\binom{6}{1}+\binom{6}{2}+\binom{6}{3}=6+15+20=41$$

Μετά την ελάττωση με βάση την υποστήριξη:

$$\binom{6}{1}+\binom{4}{2}+1=1+6+1=13$$

Στρατηγική apriori

Γενικός Αλγόριθμος

Let k=1

Generate frequent itemsets of length 1

Repeat until no new frequent itemsets are identified

- Generate length (k+1) candidate itemsets from length k frequent itemsets
- Prune candidate itemsets containing subsets of length k that are infrequent
- Count the support of each candidate by scanning the DB
- Eliminate candidates that are infrequent, leaving only those that are frequent

Στρατηγική apriori

Σε κάθε βήμα k:

Δημιουργία υποψήφιων k-στοιχειοσυνόλων με βάση τα συχνά k-1 στοιχειοσύνολα

- Όλα τα υποσύνολα του πρέπει να είναι συχνά
- Δεν πρέπει να δημιουργούμε ένα στοιχειοσύνολο πολλές φορές
- complete – δεν πρέπει να χάνουμε κάποιο συχνό

Υπολογισμός της υποστήριξής τους και pruning όσον έχουν μικρή υποστήριξη

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

$F_{k-1} \times F_1$

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

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

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

{Beer, Diaper, Milk}

Δημιουργεί και κάποια περιττά, πχ το παραπάνω δεν είναι συχνό, γιατί το {Beer, Milk} δεν είναι συχνό

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

$F_{k-1} \times F_1$

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

Διάφοροι ευριστικοί για να μειωθεί ο αριθμός των στοιχειοσυνόλων που δημιουργούνται και δεν είναι συχνά

Πχ έστω το $\{i_1, i_2, i_3, i_4\}$ για να είναι συχνό, θα πρέπει να υπάρχουν τουλάχιστον 3 3-στοιχειοσύνολα που περιέχουν πχ το i_4 ($\{i_1, i_2, i_4\}$, $\{i_1, i_3, i_4\}$ και $\{i_2, i_3, i_4\}$

Γενικά, κάθε στοιχείο ενός k-στοιχειοσυνόλου θα πρέπει να περιέχεται σε τουλάχιστον k-1 από το συχνά (k-1)-στοιχειοσύνολα

Στρατηγική apriori: Δημιουργία Στοιχειοσυνόλων

$F_{k-1} \times F_{k-1}$

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

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Itemset	Count
{Bread,Milk}	3
{Beer,Bread}	2
{Bread,Diaper}	3
{Beer, Milk}	2
{Diaper,Milk}	3
{Beer,Diaper}	3

Συγχώνευση δύο συχνών (k-1)-στοιχειοσυνόλων αλλά πρέπει *επιπρόσθετα να ελέγξουμε* ότι και τα υπόλοιπα k-2 υποσύνολα είναι συχνά

Στρατηγική apriori: Υπολογισμός Υποστήριξης

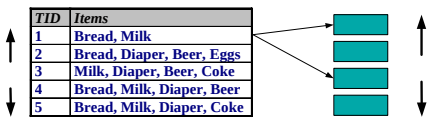
Υπολογισμός υποστήριξης: για κάθε νέο υποψήφιο συχνό στοιχειοσύνολο, πρέπει να υπολογίσουμε την υποστήριξή του

Ελάττωση του αριθμού των συγκρίσεων

- Brute Force: Διαπέρασε τη βάση των δοσοληψιών για τον υπολογισμό της υποστήριξης κάθε υποψήφιου στοιχειοσυνόλου
- Για να *μειώσουμε* τον αριθμό των συγκρίσεων, αποθήκευση των υποψηφίων στοιχειοσυνόλων σε μια δομή κατακερματισμού
  - Αντί να ταιριάζουμε κάθε δοσοληψία με κάθε υποψήφιο στοιχειοσύνολο, **ταίριαξε κάθε δοσοληψία με τα υποψήφια στοιχειοσύνολα που περιέχονται σε κάδους κατακερματισμού**

Transactions

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



Στρατηγική apriori: Υπολογισμός Υποστήριξης

Δημιουργία του δέντρου κατακερματισμού (hash tree)

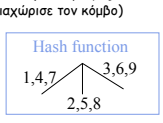
Στο δέντρο κατακερματίζουμε τα υποψήφια στοιχειοσύνολα

Έστω ότι έχουμε 15 υποψήφια 3-στοιχειοσύνολα:

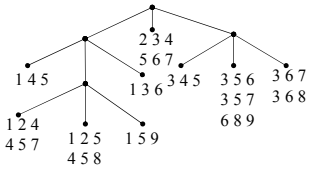
{1 4 5}, {1 2 4}, {4 5 7}, {1 2 5}, {4 5 8}, {1 5 9}, {1 3 6}, {2 3 4}, {5 6 7}, {3 4 5}, {3 5 6}, {3 5 7}, {6 8 9}, {3 6 7}, {3 6 8}

Πρέπει να προσδιορίσουμε:

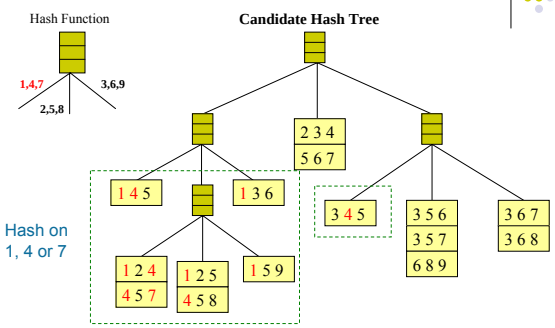
- Συνάρτηση κατακερματισμού
- Μέγιστο Μήκος Φύλλου: μέγιστο αριθμό στοιχειοσυνόλων που θα αποθηκευτούν σε κάθε φύλλο (αν ο αριθμός των στοιχειοσυνόλων υπερβεί το μέγιστο μέγεθος του φύλλου, διαχώρισε τον κόμβο)



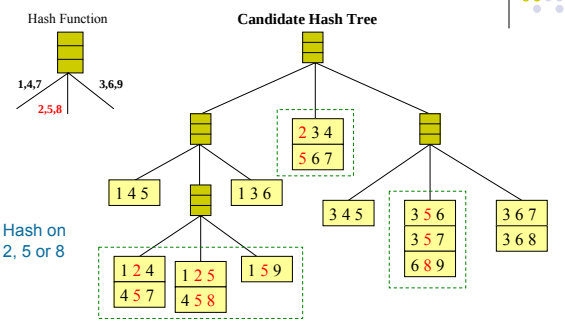
m mod 3



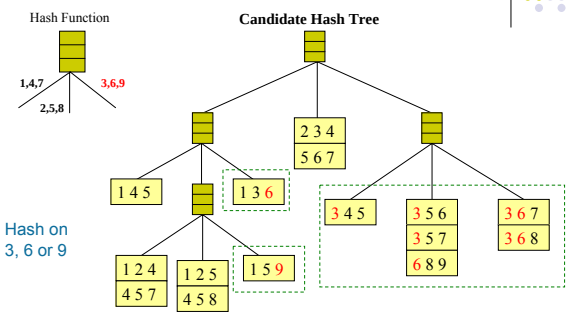
Στρατηγική apriori: Υπολογισμός Υποστήριξης



Στρατηγική apriori: Υπολογισμός Υποστήριξης



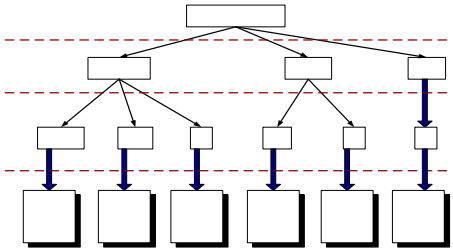
Στρατηγική apriori: Υπολογισμός Υποστήριξης



Στρατηγική apriori: Υπολογισμός Υποστήριξης

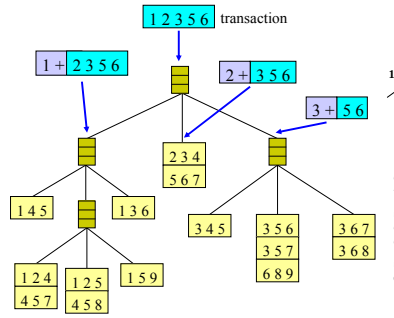
Απαρίθμηση Υποσυνόλων

Έστω μια δοσοληψία t, ποια είναι τα πιθανά υποσύνολο της με τρία στοιχεία;



Στρατηγική apriori: Υπολογισμός Υποστήριξης

Απαρίθμηση Υποσυνόλων με χρήση του Δέντρου Κατακερματισμού

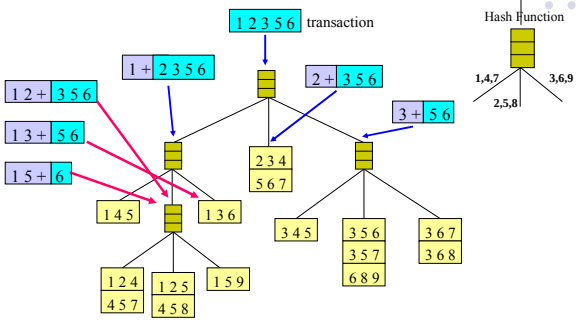


Έχοντας κατασκευάσει το δέντρο κατακερματισμού για τα 3-στοιχειοσύνολα,

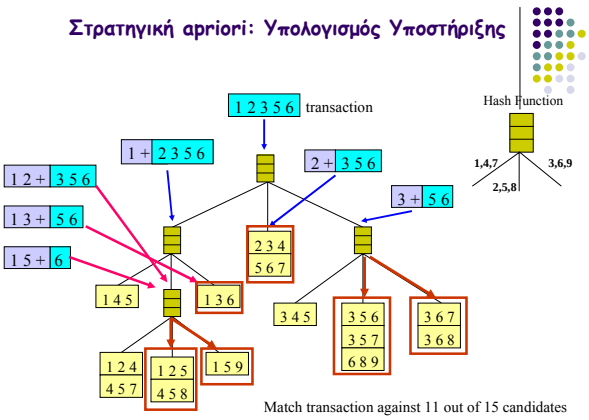
κατακερματίζουμε όλα τα 3-στοιχειοσύνολα της δοσοληψίας στο δέντρο

και αυξάνουμε τον αντίστοιχο μετρητή

Στρατηγική apriori: Υπολογισμός Υποστήριξης

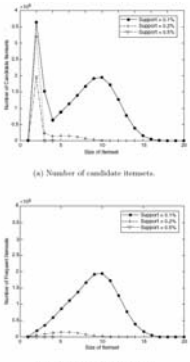


Στρατηγική apriori: Υπολογισμός Υποστήριξης



Match transaction against 11 out of 15 candidates

Στρατηγική apriori: Πολυπλοκότητα



(a) Number of candidate itemsets.

(b) Number of frequent itemsets.

Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.

Επιλογή της τιμής του κατωφλιού για την ελάχιστη υποστήριξη

- Μικρή τιμή => πολλά συχνά στοιχειοσύνολα
- Αύξηση υποψήφιων στοιχειοσυνόλων (πολυπλοκότητα) και το μέγιστο μήκος των συχνών στοιχειοσυνόλων (περισσότερα περάσματα στα δεδομένα)

Στρατηγική apriori: Πολυπλοκότητα

Αριθμός διαστάσεων - Dimensionality (αριθμός στοιχείων) του συνόλου δεδομένων

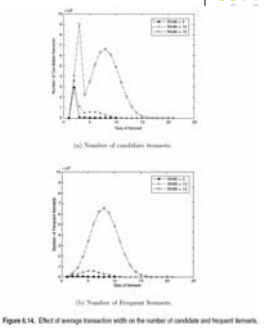
- more space is needed to store support count of each item
- if number of frequent items also increases, both computation and I/O costs may also increase

Μέγεθος της βάσης

Επειδή ο Apriori κάνει πολλαπλά περάσματα, ο χρόνος εκτέλεσης μπορεί να αυξηθεί

Στρατηγική apriori: Πολυπλοκότητα

- Μέσο πλάτος δοσοληψίας
  - transaction width larger with denser data sets
  - This may increase max length of frequent itemsets and traversals of hash tree (number of subsets in a transaction increases with its width)



Στρατηγική apriori: Πολυπλοκότητα

1. Δημιουργία συχνών 1-στοχειοσυνόλων

O(Nw)

2. Δημιουργία υποψήφιων στοιχειοσυνόλων

Έστω $F_{k-1} \times F_{k-1}$

k-2 συγκρίσεις για κοινό prefix

Στη χειρότερη περίπτωση, ταιριάζουν όλα $\Sigma_{k=2,w}|F_{k-1}|^2$

Επίσης κατασκευάζουμε το δέντρο, μέγιστο ύψος k, άρα $\Sigma_{k=2,w}k|F_{k-1}|^2$

Έλεγχος, για τα k-2 υποσύνολα με χρήση του δέντρου

3. Υπολογισμός της Υποστήριξης

Κάθε δοσοληψία έχει k από |t| k-στοιχειοσύνολα

Δημιουργία Κανόνων

Παραγωγή Κανόνων

Παραγωγή Κανόνων (Rule Generation)

- Δοθέντος ενός συχνού στοιχειοσυνόλου L, βρες όλα τα μη κενά υποσύνολα $f \subset L$ τέτοια ώστε ο κανόνας $f \rightarrow L - f$ ικανοποιεί τον περιορισμό της ελάχιστης εμπιστοσύνης
- Παράδειγμα αν {A,B,C,D} υποψήφιοι κανόνες:
  - ABC →D, ABD →C, ACD →B, BCD →A,
    A →BCD, B →ACD, C →ABD, D →ABC
    AB →CD, AC → BD, AD → BC, BC →AD,
    BD →AC, CD →AB,

  Όλοι έχουν την ίδια υποστήριξη, πρέπει να ελέγξουμε την εμπιστοσύνη
- Αν |L| = k, τότε υπάρχουν $2^k - 2$ υποψήφιοι κανόνες συσχέτισης (εξαιρώντας τον $L \rightarrow \emptyset$ και τον $\emptyset \rightarrow L$)

Παραγωγή Κανόνων

Υπολογισμός Εμπιστοσύνης

- Παρατήρηση: Δε χρειάζεται να διαπεράσουμε πάλι τα δεδομένα για να υπολογίσουμε την εμπιστοσύνη ενός κανόνα που προκύπτει από ένα συχνό στοιχειοσύνολο:
  - ABC →D, ABD →C, ACD →B, BCD →A,
    A →BCD, B →ACD, C →ABD, D →ABC
    AB →CD, AC → BD, AD → BC, BC →AD,
    BD →AC, CD →AB

Γιατί; Πχ c(CD →AB) = σ{A,B,C,D}/σ{C, D}

Από την αντι-μονότονη ιδιότητα της υποστήριξης, το {C, D} είναι συχνό στοιχειοσύνολο άρα έχουμε ήδη υπολογίζει την υποστήριξή του

Παραγωγή Κανόνων

Πως να παράξουμε αποδοτικά τους κανόνες από τα συχνά στοιχειοσύνολα;

- Γενικά, η αντι-μονότονη ιδιότητα δεν ισχύει για την εμπιστοσύνη

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Γενικά έστω {p} → {q} με εμπιστοσύνη c_1

- Και {p, r} → {q} με εμπιστοσύνη c_2

  Μπορεί $c_2 > c_1$, $c_2 < c_1$ ή $c_2 = c_1$

- Έστω {p} → {q, r} με εμπιστοσύνη c_3

  $c_3 \leq c_1$

- Επίσης, $c_3 \leq c_2$

Παραγωγή Κανόνων

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

  Για παράδειγμα L = {A,B,C,D}:

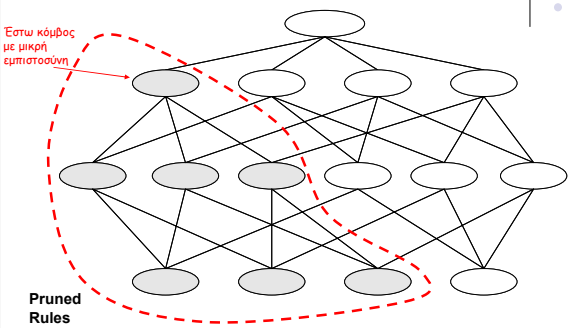
  $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$

- Η εμπιστοσύνη είναι αντι-μονότονη σε σχέση με των αριθμό των στοιχείων στο RHS του κανόνα (ή ισοδύναμα μονότονα στον αριθμό των στοιχείων στο LHS)

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

Παραγωγή Κανόνων για τον Αλγόριθμο apriori

Lattice of rules



Παραγωγή Κανόνων για τον Αλγόριθμο apriori

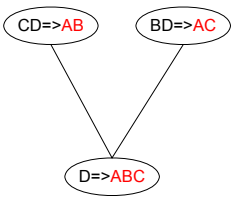
Οι κανόνες παράγονται σε επίπεδα με βάση τα στοιχεία στο RHS

Αρχικά, θεωρούμε όλους τους κανόνες **με ένα στοιχείο** στο RHS

Στη συνέχεια, οι υποψήφιοι κανόνες παράγονται συγχωνεύοντας το RHS δυο υποψηφίων κανόνων
Πχ
Join(ACD=>B, ABD=>C) μας δίνει AD=>BC

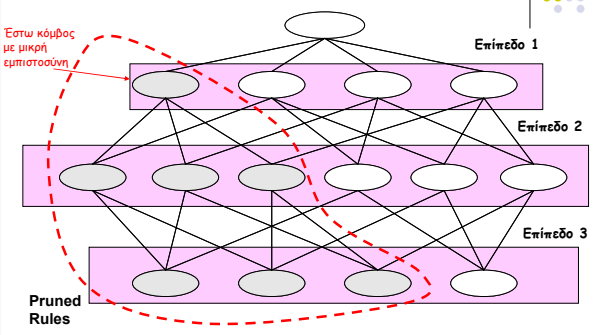
Όπως και στα συχνά στοιχειοσύνολα, στη συνέχεια, με το ίδιο prefix στο RHS
join(CD=>AB,BD=>AC) μας δίνει D => ABC

Prune τον κανόνα D=>ABC, αν το υποσύνολο AD=>BC δεν έχει επαρκή εμπιστοσύνη



Παραγωγή Κανόνων για τον Αλγόριθμο apriori

Lattice of rules



Αναπαράσταση Κανόνων Συσχέτισης

Αναπαράσταση Στοιχειοσυνόλων

Τα στοιχειοσύνολα που παράγονται είναι πολλά, κάποια ίσως περιττά

Ποια να κρατήσουμε;

Αντιπροσωπευτικά συχνά στοιχειοσύνολα

Περιττός κανόνας

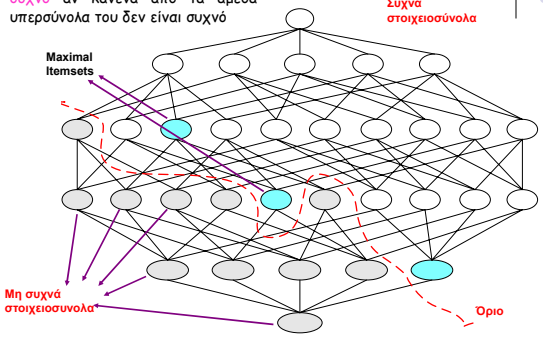
$X \rightarrow Y$, αν υπάρχει ένας κανόνας $X' \rightarrow Y'$, όπου $X \subseteq X'$ και $Y \subseteq Y'$ με την ίδια υποστήριξη και εμπιστοσύνη

{b} → {d, e} περιττός

{b, c} → {d, e}

Αναπαράσταση Στοιχειοσυνόλων

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



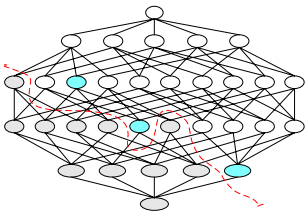
Αναπαράσταση Στοιχειοσυνόλων

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

Βέβαια χρειαζόμαστε έναν αποδοτικό αλγόριθμο για τον υπολογισμό τους που δεν παράγει όλα τα δυνατά υποσύνολα τους

ΜΕΙΟΝΕΚΤΗΜΑ: 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

Δεν είναι κλειστό αν κάποιο άμεσο υπερσύνολό του έχει την ίδια υποστήριξη

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Αναπαράσταση Στοιχειοσυνόλων

Ένα στοιχειοσύνολο είναι **κλειστό συχνό στοιχειοσύνολο** αν είναι κλειστό και η υποστήριξη του είναι μικρότερη ή ίση με minsup

Ο αλγόριθμος υπολογισμού της υποστήριξης βασίζεται στο ότι:

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

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

Αναπαράσταση Στοιχειοσυνόλων

Περιττός κανόνας

$X \rightarrow Y$, αν υπάρχει ένας κανόνας $X' \rightarrow Y'$, όπου $X \subseteq X'$ και $Y \subseteq Y'$ με την ίδια υποστήριξη και εμπιστοσύνη

{b} → {d, e} περιττός

{b, c} → {d, e}

Παρατήρηση: Θα κρατήσουμε μόνο το {b, c, d, e}

Αναπαράσταση Στοιχειοσυνόλων

Some itemsets are redundant because they have identical support as their supersets

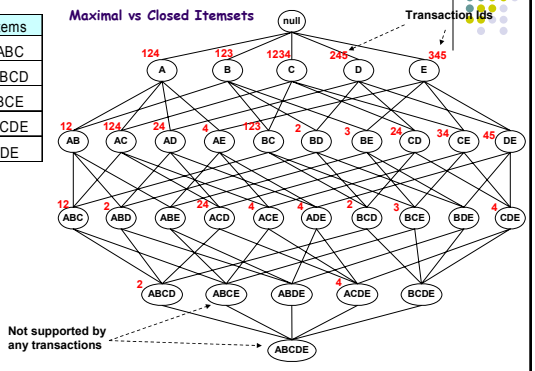
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

$$\text{Number of frequent itemsets} = 3 \times \sum_{k=1}^{10} \binom{10}{k}$$

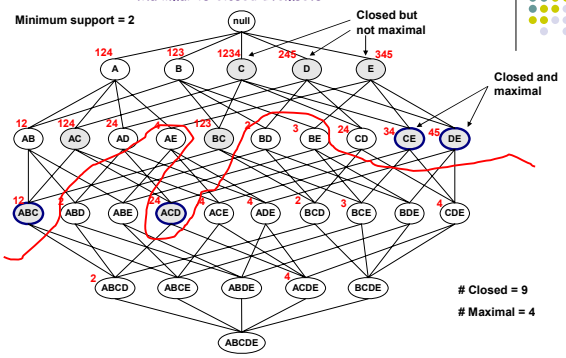
Need a compact representation

Αναπαράσταση Στοιχειοσυνόλων

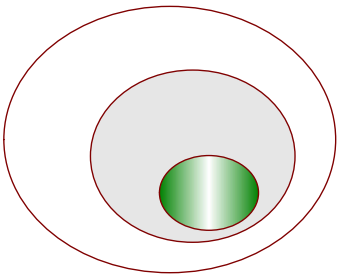
TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



Αναπαράσταση Στοιχειοσυνόλων



Αναπαράσταση Στοιχειοσυνόλων



Αποτίμηση Κανόνων Συσχέτισης

Αποτίμηση Κανόνων Συσχέτισης

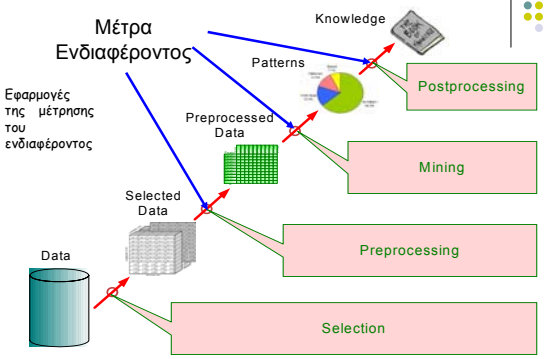
Παράγουν πάρα πολλούς κανόνες που συχνά είναι μη ενδιαφέροντες ή πλεονάζοντες (περιττοί)

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

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

Μέτρηση Ενδιαφέροντος



Μέτρηση Ενδιαφέροντος

Υπολογισμός του Μέτρου Ενδιαφέροντος

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

Contingency table for $X \rightarrow Y$

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

Μέτρηση συχνότητας εμφάνισης

Χρησιμοποιείται για τον ορισμό διαφόρων μέτρων

Μέτρηση Ενδιαφέροντος

Μειονεκτήματα της Εμπιστοσύνης

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

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

Κανόνας Συσχέτισης: Tea → Coffee

Εμπιστοσύνη = P(Coffee|Tea) = 0.75

αλλά P(Coffee) = 0.9

⇒ Ενώ ο κανόνας έχει υψηλή εμπιστοσύνη, ο κανόνας είναι παραπλανητικός

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

Αγνοεί την υποστήριξη του RHS

Μέτρα βασισμένα στη Στατιστική

Στατιστική Ανεξαρτησία

- Πληθυσμός 1000 σπουδαστών
  - 600 σπουδαστές ξέρουν κολύμπι (S)
  - 700 σπουδαστές ξέρουν ποδήλατο (B)
  - 420 σπουδαστές ξέρουν κολύμπι και ποδήλατο (S,B)
  - P(S∧B) = 420/1000 = 0.42
  - P(S) × P(B) = 0.6 × 0.7 = 0.42
  - P(S∧B) = P(S) × P(B) => Στατιστική ανεξαρτησία
  - P(S∧B) > P(S) × P(B) => Positively correlated (θετική συσχέτιση)
  - P(S∧B) < P(S) × P(B) => Negatively correlated (αρνητική συσχέτιση)

Μέτρα βασισμένα στη Στατιστική

Μέτρα που λαμβάνουν υπ΄ όψιν τους στη στατιστική εξάρτηση

$X \rightarrow Y$

$$Lift = \frac{P(Y \mid X)}{P(Y)}$$

$$Interest = \frac{P(X,Y)}{P(X)P(Y)}$$

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

$$\phi - coefficient = \frac{P(X,Y) - P(X)P(Y)}{\sqrt{P(X)[1-P(X)]P(Y)[1-P(Y)]}}$$

Μέτρα βασισμένα στη Στατιστική

Παράδειγμα: Lift/Interest

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

Κανόνας συσχέτιση: Tea → Coffee

Εμπιστοσύνη= P(Coffee|Tea) = 0.75

αλλά P(Coffee) = 0.9

⇒ Lift = 0.75/0.9= 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

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

Μεγαλύτερο αν και σπάνια εμφανίζονται μαζί

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

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

Στατιστική ανεξαρτησία:

If P(X,Y)=P(X)P(Y) => Lift = 1

Example: ϕ-Coefficient

ϕ-coefficient is analogous to correlation coefficient for continuous variables

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

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

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

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

ϕ Coefficient is the same for both tables

Στη βιβλιογραφία έχουν προταθεί πολλές 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)	$\frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$
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)	$\frac{\sum_i \sum_j P(A_i,B_j) \log \frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i), -\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\left(P(A,B)\log\left(\frac{P(B\|A)}{P(B)}\right) + P(A\bar{B})\log\left(\frac{P(\bar{B}\|A)}{P(\bar{B})}\right), P(A,B)\log\left(\frac{P(A\|B)}{P(A)}\right) + P(\bar{A}B)\log\left(\frac{P(\bar{A}\|B)}{P(\bar{A})}\right)\right)$
9	Gini index (G)	$\max\left(P(A)[P(B\|A)^2 + P(\bar{B}\|A)^2] + P(\bar{A})[P(B\|\bar{A})^2 + P(\bar{B}\|\bar{A})^2] - P(B)^2 - P(\bar{B})^2, P(B)[P(A\|B)^2 + P(\bar{A}\|B)^2] + P(\bar{B})[P(A\|\bar{B})^2 + P(\bar{A}\|\bar{B})^2] - P(A)^2 - P(\bar{A})^2\right)$
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)	$\max\left(\frac{P(A)P(\bar{B})}{P(A\bar{B})}, \frac{P(B)P(\bar{A})}{P(B\bar{A})}\right)$
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)	$\frac{P(A,B)+P(\bar{A}\bar{B})}{P(A)P(B)+P(\bar{A})P(\bar{B})} \times \frac{1-P(A)P(B)-P(\bar{A})P(\bar{B})}{1-P(A,B)-P(\bar{A}\bar{B})}$
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

Ιδιότητες ενός Καλού Μέτρου

Piatetsky-Shapiro:

3 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

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	8	5	5	5	9	9	7	9	8	3	3	3	7	9	9	3	7	9	9	8
E10	10	10	6	6	6	10	5	9	10	10	6	6	5	1	10	10	5	1	10	10	7

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, etc

Μη συμμετρικά (asymmetric) μέτρα:

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

Ιδιότητες Μέτρων Αποτίμησης

Κλιμάκωση Γραμμής/Στήλης (Row/Column Scaling)

Παράδειγμα Βαθμός-Φύλο (Mosteller, 1968):

	Male	Female	
High	2	3	5
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	C	D	E	F
Transaction 1	1	0	0	1	0	0
	0	0	1	1	1	0
	0	0	1	1	1	0
	0	0	1	1	1	0
	0	1	1	0	1	1
	0	0	1	1	1	0
	0	0	1	1	1	0
	0	0	1	1	1	0
	0	0	1	1	1	0
Transaction N	1	0	0	1	0	0
	(a)		(b)		(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}

Ιδιότητες Μέτρων Αποτίμησης

Null Addition (προσθήκη μη σχετιζόμενων στοιχείων)

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

⇒

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

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

Invariant measures:
- support, cosine, Jaccard, etc

Non-invariant measures:
- correlation, Gini, mutual information, odds ratio, etc

Ιδιότητες Μέτρων Αποτίμησης

Symbol	Measure	Range	P1	P2	P3	O1	O2	O3	O3'	O4
Φ	Correlation	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	Yes	Yes	No
λ	Lambda	0 … 1	Yes	No	No	Yes	No	No*	Yes	No
α	Odds ratio	0 … 1 … ∞	Yes*	Yes	Yes	Yes	Yes	Yes*	Yes	No
Q	Yule's Q	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Y	Yule's Y	-1 … 0 … 1	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
κ	Cohen's	-1 … 0 … 1	Yes	Yes	Yes	Yes	No	No	Yes	No
M	Mutual Information	0 … 1	Yes	Yes	Yes	Yes	No	No*	Yes	No
J	J-Measure	0 … 1	Yes	No	No	No	No	No	No	No
G	Gini Index	0 … 1	Yes	No	No	No	No	No*	Yes	No
s	Support	0 … 1	No	Yes	No	Yes	No	No	No	No
c	Confidence	0 … 1	No	Yes	No	Yes	No	No	No	Yes
L	Laplace	0 … 1	No	Yes	No	Yes	No	No	No	No
V	Conviction	0.5 … 1 … ∞	No	Yes	No	Yes**	No	No	Yes	No
I	Interest	0 … 1 … ∞	Yes*	Yes	Yes	Yes	No	No	No	No
IS	IS (cosine)	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
PS	Piatetsky-Shapiro's	-0.25 … 0 … 0.25	Yes	Yes	Yes	Yes	No	Yes	Yes	No
F	Certainty factor	-1 … 0 … 1	Yes	Yes	Yes	No	No	No	Yes	No
AV	Added value	0.5 … 1 … 1	Yes	Yes	Yes	No	No	No	No	No
S	Collective strength	0 … 1 … ∞	No	Yes	Yes	Yes	No	Yes*	Yes	No
ζ	Jaccard	0 .. 1	No	Yes	Yes	Yes	No	No	No	Yes
K	Klosgen's	$\left(\sqrt{\frac{2}{\sqrt{3}}}-1\right)\left(2-\sqrt{3}-\frac{1}{\sqrt{3}}\right)\ldots 0 \ldots \frac{2}{3\sqrt{3}}$	Yes	Yes	Yes	No	No	No	No	No

Αποτίμηση Κανόνων Συσχέτισης

Παράδοξο του Simpson

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

Students

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

Working adults

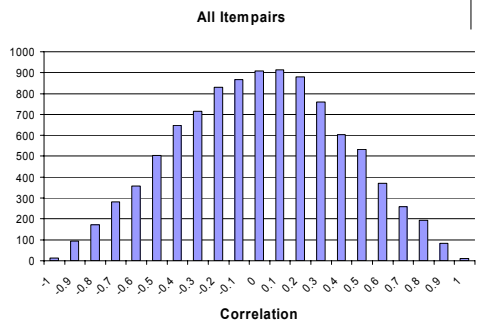
Buy HDTV	Buy Exercise Machine		
	Yes	No	
Yes	98	72	170
No	50	36	86

Ελάττωση με Βάση την Υποστήριξη

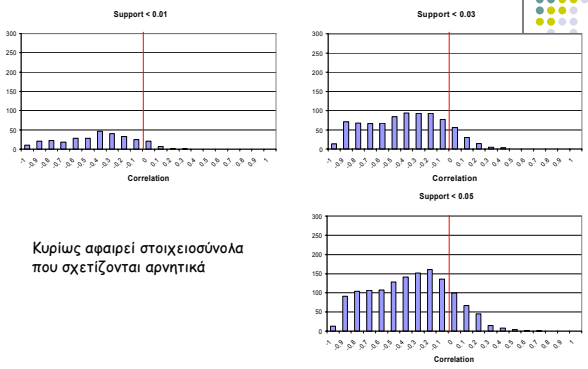
Support-based Pruning

- Οι περισσότεροι αλγόριθμοι για την εξόρυξη κανόνων συσχέτισης χρησιμοποιούν την υποστήριξη για να μειώσουν (prune) κανόνες και στοιχειοσύνολα
- Μελέτη του αποτελέσματος της μείωσης στη συσχέτιση των στοιχειοσυνόλων
  - Δημιουργία 10000 τυχαίων contingency tables
  - Υπολογισμός της υποστήριξης και της ανα-δύο συσχέτισης των πινάκων
  - Εφαρμογή της ελάττωσης με βάση την υποστήριξη και μελέτη των πινάκων που αφαιρέθηκαν

Ελάττωση με Βάση την Υποστήριξη



Ελάττωση με Βάση την Υποστήριξη



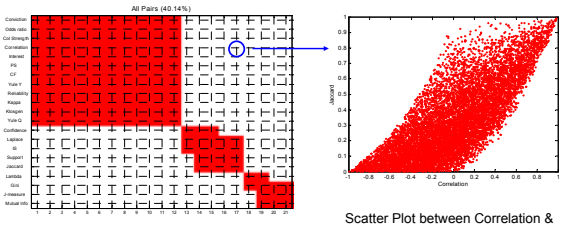
Κυρίως αφαιρεί στοιχειοσύνολα που σχετίζονται αρνητικά

Ελάττωση με Βάση την Υποστήριξη

- Μελέτη του πως επηρεάζει τα άλλα μέτρα
- Βήματα:
  - Δημιουργία 10000 contingency tables
  - Ιεράρχηση κάθε πίνακα με βάση τα διαφορετικά μέτρα
  - Υπολογισμός της ανά-δύο συσχέτισης μεταξύ των μέτρων

Ελάττωση με Βάση την Υποστήριξη

- Χωρίς ελάττωση (Όλα τα ζεύγη)

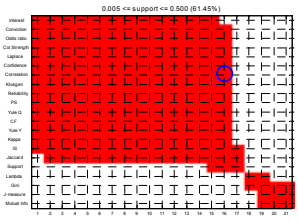


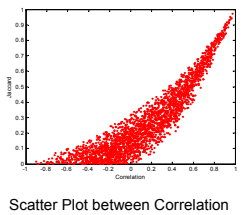
Scatter Plot between Correlation & Jaccard Measure

- Red cells indicate correlation between the pair of measures > 0.85
- 40.14% pairs have correlation > 0.85

Ελάττωση με Βάση την Υποστήριξη

- 0.5% ≤ support ≤ 50%



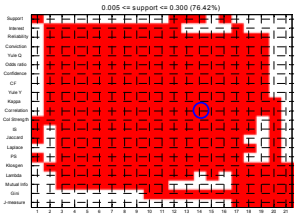


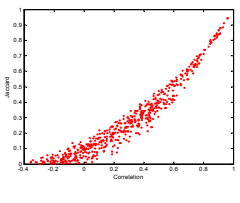
Scatter Plot between Correlation & Jaccard Measure:

- 61.45% pairs have correlation > 0.85

Ελάττωση με Βάση την Υποστήριξη

- 0.5% ≤ support ≤ 30%





Scatter Plot between Correlation & Jaccard Measure

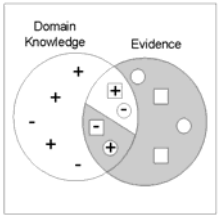
- 76.42% pairs have correlation > 0.85

Υποκειμενικά Μέτρα Ενδιαφέροντος

- Αντικειμενικά Μέτρα:
  - Rank patterns based on statistics computed from data
  - e.g., 21 measures of association (support, confidence, Laplace, Gini, mutual information, Jaccard, etc).
- Υποκειμενικά Μέτρα:
  - Ιεράρχηση των προτύπων με βάση την ερμηνεία του χρήστη
    - A pattern is subjectively interesting if it contradicts the expectation of a user (Silberschatz & Tuzhilin)
    - A pattern is subjectively interesting if it is actionable (Silberschatz & Tuzhilin)

Υποκειμενικά Μέτρα Ενδιαφέροντος

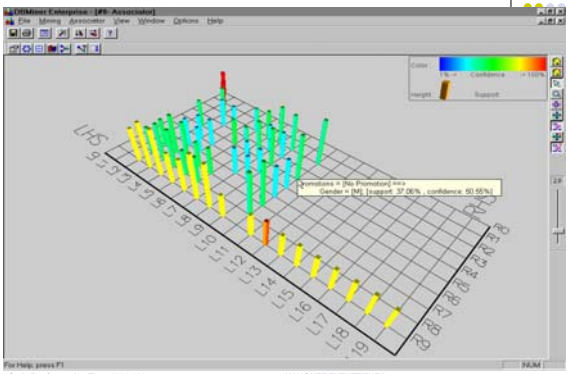
Interestingness via Unexpectedness



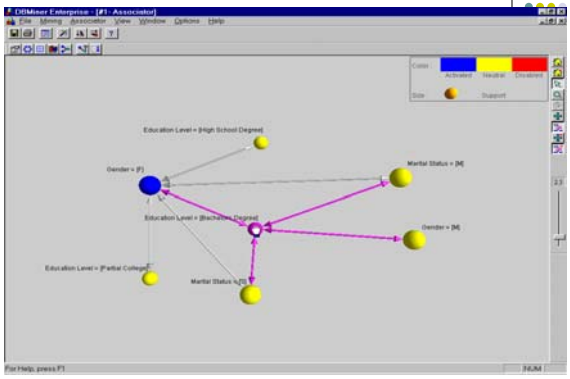
+ Pattern expected to be frequent
- Pattern expected to be infrequent
Pattern found to be frequent
Pattern found to be infrequent
Expected Patterns
Unexpected Patterns

- Need to model expectation of users (domain knowledge)
- Need to combine expectation of users with evidence from data (i.e., extracted patterns)

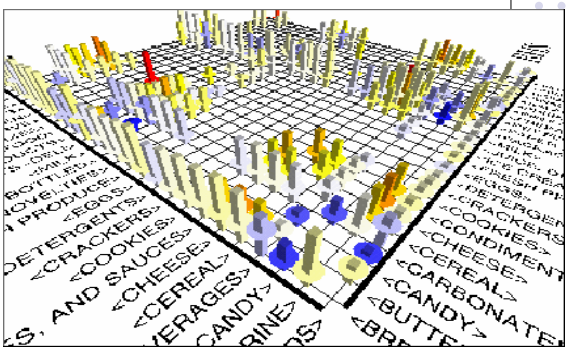
Οπτικοποίηση: Απλός Γράφος



Οπτικοποίηση: Γράφος Κανόνων

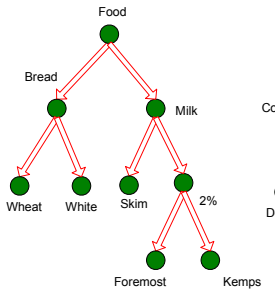


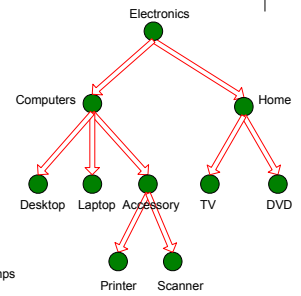
Οπτικοποίηση: (SGI/MineSet 3.0)



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 στοχειοσύνολο
- Οι κανόνες στα χαμηλότερα επίπεδα είναι πάρα πολύ συγκεκριμένοι
  - π.χ., skim milk → white bread, 2% milk → wheat bread, skim milk → wheat bread, etc.

είναι ενδεικτικοί της συσχέτισης μεταξύ γάλατος και ψωμιού

Κανόνες Συσχέτισης Πολλών Επιπέδων

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

    Αρχική Δοσοληψία: {skim milk, wheat bread}

    Επαυξημένη Δοσοληψία: {skim milk, wheat bread, milk, bread, food}
- Θέματα:
  - Τα στοιχεία στα υψηλότερα επίπεδα θα εμφανίζονται πολύ συχνά, μεγάλους μετρητές υποστήριξης
    - if support threshold is low, too many frequent patterns involving items from the higher levels
  - Increased dimensionality of the data

Κανόνες Συσχέτισης Πολλών Επιπέδων

- How do support and confidence vary as we traverse the concept hierarchy?
  - If X is the parent item for both X1 and X2, then $\sigma(X) \leq \sigma(X1) + \sigma(X2)$
  - If $\sigma(X1 \cup Y1) \geq minsup$,
    and X is parent of X1, Y is parent of Y1
    then $\sigma(X \cup Y1) \geq minsup$, $\sigma(X1 \cup Y) \geq minsup$
    $\sigma(X \cup Y) \geq minsup$
  - If $conf(X1 \Rightarrow Y1) \geq minconf$,
    then $conf(X1 \Rightarrow Y) \geq minconf$

Κανόνες Συσχέτισης Πολλών Επιπέδων

- Approach 2:
  - Generate frequent patterns at highest level first
  - Then, generate frequent patterns at the next highest level, and so on
- Issues:
  - I/O requirements will increase dramatically because we need to perform more passes over the data
  - May miss some potentially interesting cross-level association patterns