Πανεπιστήμιο Πειραιώς Τμήμα Πληροφορικής

Εξόρυξη Γνώσης από Δεδομένα (Data Mining)

Εξόρυξη Κανόνων Συσχετίσεων

Γιάννης Θεοδωρίδης

Ομάδα Διαχείρισης Δεδομένων
Εργαστήριο Πληροφοριακών Συστημάτων

http://isl.cs.unipi.gr/db

Περιεχόμενα

- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων ή στοιχειοσύνολα *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - Αλγόριθμος δειγματοληψίας
  - Αλγόριθμος διαμερισμού
  - Παράλληλοι αλγόριθμοι
- Ειδικά θέματα

Βασισμένες κατά κύριο λόγο (αλλά όχι αποκλειστικά)
στις διαφάνειες που συνοδεύουν το βιβλίο
M. H. Dunham: Data Mining, Introductory and Advanced Topics

Περιεχόμενα

- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων ή στοιχειοσύνολα *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - Αλγόριθμος δειγματοληψίας
  - Αλγόριθμος διαμερισμού
  - Παράλληλοι αλγόριθμοι
- Ειδικά θέματα

Δεδομένα από το «καλάθι της νοικοκυράς»

- Market-basket data
- Αντικείμενα που αγοράζονται συχνά μαζί:

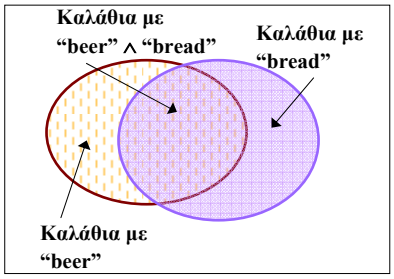
  Bread ⇒ PeanutButter

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Εφαρμογές:
  - Τοποθέτηση προϊόντων στα ράφια
  - Διαφήμιση
  - Πωλήσεις
  - Κουπόνια – Εκπτώσεις
- Σκοπός
  - αύξηση πωλήσεων και μείωση κόστους

Κανόνες Συσχετίσεων: Ορισμοί

- **Σύνολο στοιχείων**: $I=\{I_1, I_2, \ldots, I_m\}$
- **Βάση δοσοληψιών** (transaction database): $D=\{t_1, t_2, \ldots, t_n\}$, $t_j \subseteq I$
- (υπο-) **σύνολο στοιχείων** (itemset): $\{I_{i1}, I_{i2}, \ldots, I_{ik}\} \subseteq I$
- **Υποστήριξη** (support) ενός itemset: το ποσοστό των δοσοληψιών που περιέχουν το συγκεκριμένο itemset.
- **Συχνό** (frequent) itemset: Ένα itemset, η υποστήριξη του οποίου υπερβαίνει ένα συγκεκριμένο κατώφλι.
  - Στη βιβλιογραφία αναφέρεται και ως «**μεγάλο**» (large itemset)



Παράδειγμα

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

I={Beer,Bread,Jelly,Milk,PeanutButter}

Η υποστήριξη του itemset {**Bread**} είναι 80%

Η υποστήριξη του itemset {**Bread**, **PeanutButter**} είναι 60%

Η υποστήριξη του itemset {**Bread**, **Milk**, **PeanutButter**} είναι 20%

Κανόνες Συσχετίσεων: Ορισμοί (συν.)

- **Κανόνας συσχέτισης** (AR): $X \Rightarrow Y$

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

  - Το X ονομάζεται *LHS* (left-hand side) ή *antecedent* (προηγούμενο) ή *head* (κεφαλή) του κανόνα
  - Το Y ονομάζεται *RHS* (right-hand side) ή *consequent* (επακόλουθο) ή *body* (σώμα) του κανόνα
- **Υποστήριξη** (support) του AR (s) $X \Rightarrow Y$:

  το ποσοστό των δοσοληψιών που περιέχουν το $X \cup Y$

  ή αλλιώς η πιθανότητα $P(X \cup Y)$
- **Εμπιστοσύνη** (confidence) του AR (α) $X \Rightarrow Y$:

  η αναλογία του πλήθους των δοσοληψιών που περιέχουν το $X \cup Y$ ως προς το πλήθος των δοσοληψιών που περιέχουν το X.

  ή αλλιώς, η εξαρτημένη πιθανότητα $P(X \cup Y \mid X) = P(X \cup Y) / P(X)$

Παράδειγμα (συν.)

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$X \Rightarrow Y$	s	α
Bread $\Rightarrow$ PeanutButter	60%	75%
PeanutButter $\Rightarrow$ Bread	60%	100%
Beer $\Rightarrow$ Bread	20%	50%
PeanutButter $\Rightarrow$ Jelly	20%	33.3%
Jelly $\Rightarrow$ PeanutButter	20%	100%
Jelly $\Rightarrow$ Milk	0%	0%

Κανόνες Συσχετίσεων: το πρόβλημα

- Δοθέντος
  - ενός συνόλου αντικειμένων $I=\{I_1, I_2, \ldots, I_m\}$ και
  - μιας βάσης δοσοληψιών $D=\{t_1, t_2, \ldots, t_n\}$ όπου $t_i=\{I_{i1}, I_{i2}, \ldots, I_{ik}\}$ και $I_{ij} \in I$
  - μιας ελάχιστης υποστήριξης (min_support)
  - μιας ελάχιστης εμπιστοσύνης (min_confidence)

το **Πρόβλημα της εύρεσης Κανόνων Συσχέτισης** ορίζεται ως

- ο προσδιορισμός όλων των κανόνων συσχέτισης $X \Rightarrow Y$, όπου $X, Y \subseteq I$ και $X \cap Y = \emptyset$, οι οποίοι ξεπερνούν το κατώφλι του min_support και του min_confidence.

Κανόνες Συσχετίσεων: Τεχνική

Πρόβλημα εύρεσης κανόνων συσχέτισης:
Προσδιορισμός όλων των κανόνων συσχέτισης $X \Rightarrow Y$, όπου $X,Y \subseteq I$ και $X \cap Y = \emptyset$, οι οποίοι ξεπερνούν τα κατώφλια *min_support* και *min_confidence*

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Μεθοδολογία:

Βήμα 1: Εύρεση του συνόλου L των συχνών itemsets (αυτών δηλαδή που ξεπερνούν το κατώφλι *min_support*).

Βήμα 2: Προσδιορισμός των κανόνων συσχέτισης $X \Rightarrow Y$ από το σύνολο L (και παράλληλα το φιλτράρισμα αυτών με βάση το κατώφλι *min_support*).

1η σημείωση: Η υποστήριξη του κανόνα $X \Rightarrow Y$ είναι ίδια με την υποστήριξη του itemset $X \cup Y$.

2η σημείωση: Το 1ο βήμα δείχνει απλό αλλά κοστίζει πολύ, αφού υπάρχουν μέχρι 2^m-1 «πιθανά» συχνά itemsets (m ο πληθάριθμος του I)

Προσδιορισμός των Κανόνων Συσχετίσεων από τα συχνά itemsets (βήμα 2)

Input:
- D //Database of transactions
- I //Items
- L //Large itemsets
- s //Support
- α //Confidence

Output:
- R //Association Rules satisfying s and α

ARGen Algorithm:

$R = \emptyset$;
for each $l \in L$ **do**
  for each $x \subset l$ **such that** $x \neq \emptyset$ **and** $x \neq l$ **do**
    if $\frac{support(l)}{support(x)} \geq \alpha$ **then**
      $R = R \cup \{x \Rightarrow (l - x)\}$;

Παράδειγμα

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Input:
- D //Database of transactions
- I //Items
- L //Large itemsets
- s //Support
- α //Confidence

Output:
- R //Association Rules satisfying s and α

ARGen Algorithm:

$R = \emptyset$;
for each $l \in L$ **do**
  for each $x \subset l$ **such that** $x \neq \emptyset$ **and** $x \neq l$ **do**
    if $\frac{support(l)}{support(x)} \geq \alpha$ **then**
      $R = R \cup \{x \Rightarrow (l - x)\}$;

$X \Rightarrow Y$	s	α
Bread ⇒ PeanutButter	60%	75%
PeanutButter ⇒ Bread	60%	100%
Beer ⇒ Bread	20%	50%
PeanutButter ⇒ Jelly	20%	33.3%
Jelly ⇒ PeanutButter	20%	100%
Jelly ⇒ Milk	0%	0%

$$support(X \Rightarrow Y) = support(X \cup Y)$$

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)}$$

Περιεχόμενα

- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων *(frequent itemsets)*
- **Αλγόριθμοι εύρεσης συχνών itemsets**
  - **Αλγόριθμος Apriori**
  - Αλγόριθμος δειγματοληψίας
  - Αλγόριθμος διαμερισμού
  - Παράλληλοι αλγόριθμοι
- Ειδικά θέματα

Apriori

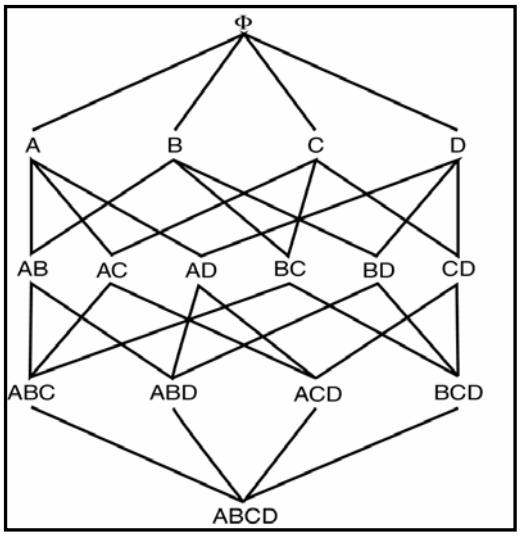
- **Συχνό** ονομάζεται το itemset που έχει υποστήριξη πάνω από ένα κατώφλι
- Παράδειγμα

  (κατώφλι = 40%):
  - {Beer}
  - {Bread}
  - {PeanutButter}
  - {Bread, PeanutButter}

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

- Η **ιδιότητα των συχνών itemsets**:
  - Κάθε υποσύνολο ενός συχνού itemset είναι συχνό.
  - Αντιθέτως, αν ένα itemset δεν είναι συχνό, κανένα από τα υπερσύνολά του δεν μπορεί να είναι συχνό.

Η ιδιότητα των συχνών itemsets



Παράδειγμα Apriori (συν.)

Pass	Candidates	Large Itemsets
1	{Beer},{Bread},{Jelly}, {Milk},{PeanutButter}	{Beer},{Bread}, {Milk},{PeanutButter}
2	{Beer,Bread},{Beer,Milk}, {Beer,PeanutButter},{Bread,Milk}, {Bread,PeanutButter},{Milk,PeanutButter}	{Bread,PeanutButter}

s=30% α = 50%

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

Αλγόριθμος Apriori

```
1.  C1 = Itemsets of size one in I;
2.  Count C1 to determine L1;              // 1st pass
3.  i = 1;
4.  Repeat
5.      i = i + 1;
6.      Ci = Apriori-Gen(Li-1);
7.      Count Ci to determine Li; // 2nd, 3rd, …, pass
8.  until no more frequent itemsets found;
```

Apriori-Gen

- Προσδιορισμός των **υποψηφίων συχνών i**-itemsets από τα **συχνά (i-1)**-itemsets.
- Προσέγγιση που ακολουθείται (2 βήματα):
  - **Βήμα σύνδεσης** (join step): Σύνδεση σε ένα i-itemset των i συχνών (i-1)-itemsets, αν υπάρχουν. $C_i = L_{i-1} \bowtie L_{i-1}$.
  - **Βήμα κλαδέματος** (prune step): Απόρριψη ενός υποψηφίου i-itemset, αν κάποιο υποσύνολο (i-1)-itemset αυτού δεν είναι συχνό.
- Παράδειγμα (s = 30%):

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

$\mathbf{C_1}$: I

$\mathbf{L_1}$: {Beer}, {Bread}, {Milk}, {PeanutButter}

$\mathbf{C_2}$: {Beer, Bread}, {Beer, Milk}, …

$\mathbf{L_2}$: {Bread, PeanutButter}

$\mathbf{C_3}$: ∅

Πώς προσδιορίζουμε τα υποψήφια itemsets

- Θεωρήστε ότι τα στοιχεία του L_{i-1} είναι διατεταγμένα
- Βήμα 1 (σύνδεση): $C_i = L_{i-1} \bowtie L_{i-1}$

```
insert into Ci
select p.item1, p.item2, …, p.itemi-1, q.itemi-1
from Li-1 p, Li-1 q
where p.item1=q.item1, …, p.itemi-1 < q.itemi-1
```

- Βήμα 2 (κλάδεμα)

```
For all itemsets c in Ci do
   For all (k-1)-subsets s of c do
      if (s ∉ Li-1) then delete c from Ci
```

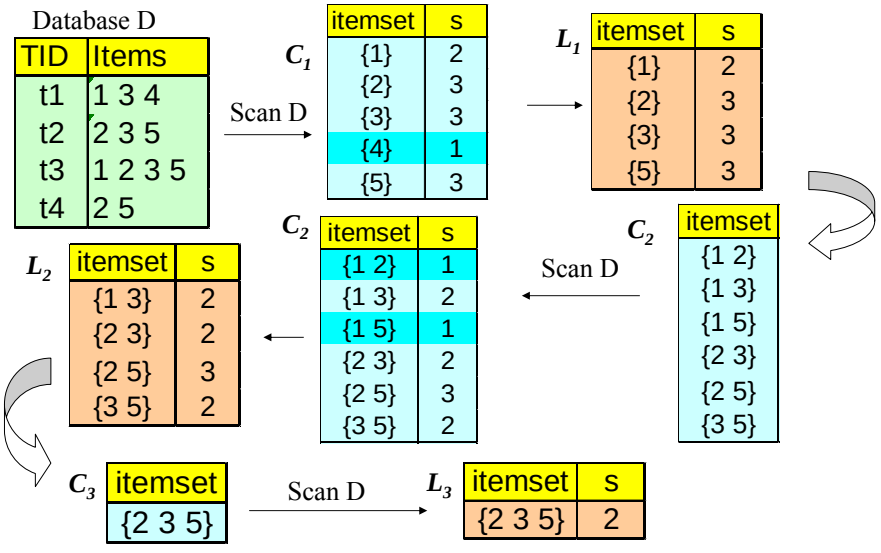
Παράδειγμα:
L_3={abc, abd, acd, ace, bcd}

Σύνδεση: $L_3 \bowtie L_3$ = {abcd, acde} από τα abc-abd και acd-ace, αντίστοιχα

Κλάδεμα: acde, αφού ade ∉ L_3
Άρα, C_4={abcd}

2ο Παράδειγμα Apriori

(Πηγή: "Data Mining: Concepts and Techniques", Han & Kamber)



Υπέρ και Κατά του Apriori

- **Πλεονεκτήματα**:
  - Εκμεταλλεύεται την ιδιότητα των συχνών itemsets.
  - Υλοποιείται εύκολα (και σε παράλληλη μορφή)
- **Μειονεκτήματα**:
  - Υποθέτει ότι η βάση των δοσοληψιών βρίσκεται στη μνήμη.
  - Απαιτεί μέχρι και m σαρώσεις της βάσης (m το πλήθος των items).

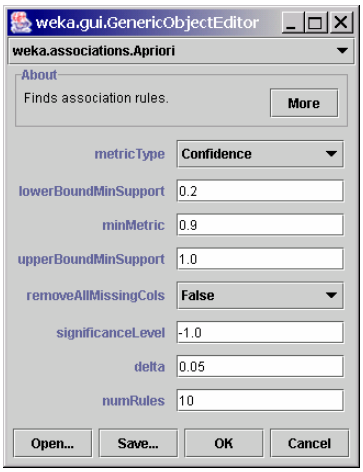
Κανόνες Συσχετίσεων που προκύπτουν από άλλους (πλεονάζοντες)

- Κάποιοι κανόνες προκύπτουν από άλλους, παρατήρηση που μπορεί να επιταχύνει τη διαδικασία εύρεσης του συνόλου L.
- **Απλός πλεονασμός:**
  - Ο κανόνας {a,b} ⇒ {c} προκύπτει από τον {a} ⇒ {b,c}
    - Αποδείξτε το – αν δηλαδή ο δεύτερος ξεπερνά τα όρια που έχουν τεθεί για support και confidence, αναγκαστικά θα ισχύει το ίδιο και για τον πρώτο
- **Αυστηρός πλεονασμός:**
  - Ο κανόνας {a} ⇒ {c} προκύπτει από τον {a} ⇒ {b,c}
    - Αποδείξτε το

Κανόνες Συσχετίσεων στο Weka

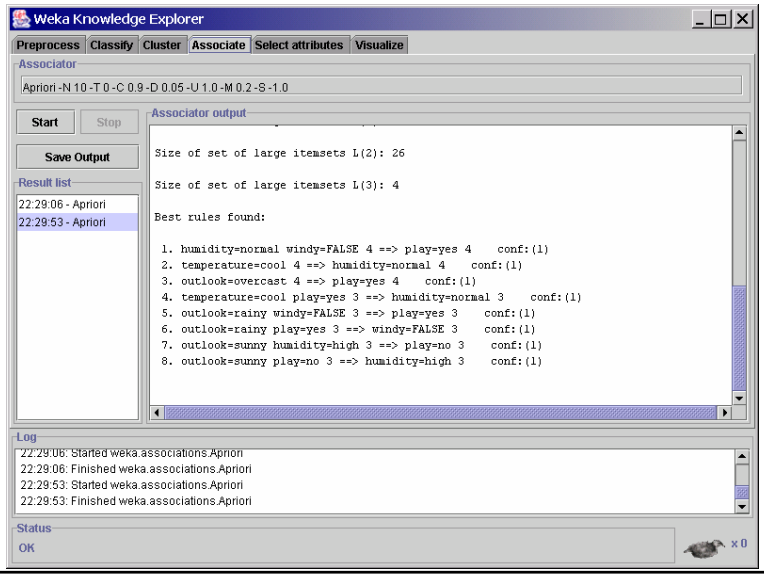
(Πηγή: "Data Mining Course"
http://www.kdnuggets.com/dmcourse)

File: weather.nominal.arff
MinSupport: 0.2



Κανόνες Συσχετίσεων στο Weka (συν.)

(Πηγή: "Data Mining Course" http://www.kdnuggets.com/dmcourse)



Περιεχόμενα

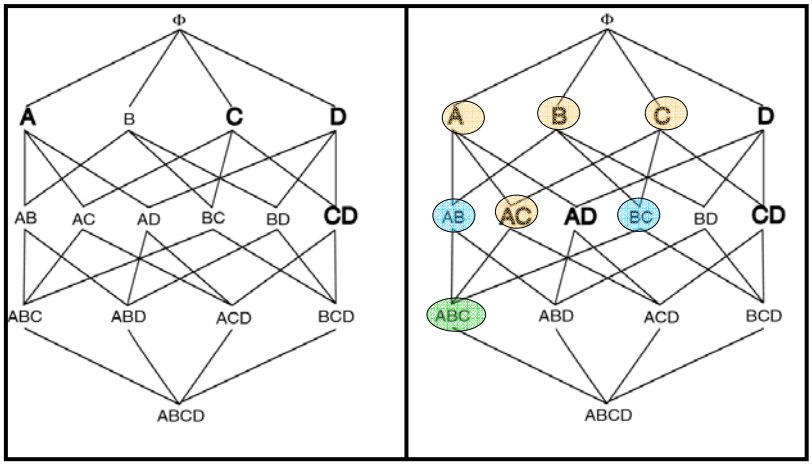
- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - **Αλγόριθμος δειγματοληψίας**
  - Αλγόριθμος διαμερισμού
  - Παράλληλοι αλγόριθμοι
- Ειδικά θέματα

Δειγματοληψία

- Πρόβλημα: Μεγάλες ΒΔ
- Λύση: Δειγματοληψία στη ΒΔ και εφαρμογή του Apriori στο επιλεγμένο δείγμα.

- **Ενδεχομένως Μεγάλα Itemsets (Potentially Large Itemsets -PL):** μεγάλα itemsets από το επιλεγμένο δείγμα
- **Αρνητικό Όριο (Negative Border - BD^{-}):**
  - Γενίκευση του Apriori-Gen που εφαρμόζεται σε itemsets διαφόρων μεγεθών.
  - Το ελάχιστο σύνολο των itemsets που δεν ανήκουν στο PL, ενώ όλα τα υποσύνολά τους ανήκουν στο PL.

Αρνητικό Όριο: Παράδειγμα



PL $\quad$ PL $\cup BD^{-}(PL)$

Αλγόριθμος δειγματοληψίας

```
1.  D_s = sample of Database D;
2.  PL = Large itemsets in D_s (using smalls);
3.  C = PL ∪ BD⁻(PL);
4.  L = Count C in Database (using s);  // 1st pass
5.  ML = itemsets in BD⁻(PL) found to be frequent;
6.  If ML = ∅ then
      done
    else
      C = L;
      repeat
            C = C ∪ BD⁻(C);
      until BD⁻(C) = ∅
7.  L = Count C in Database (using s);  // 2nd pass
```

Παράδειγμα δειγματοληψίας

Θέλουμε να βρούμε τους κανόνες συσχετίσεων, με s = 40%

Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

1. D_s = { t_1, t_2}; smalls = 10%
2. PL = { {Bread}, {Jelly}, {PeanutButter}, {Bread, Jelly}, {Bread, PeanutButter}, {Jelly, PeanutButter}, {Bread, Jelly, PeanutButter} }
3. $BD^{-}(PL)$ = { {Beer}, {Milk} } ; C = PL ∪ $BD^{-}(PL)$
4. (1ο πέρασμα στη ΒΔ) Μέτρηση του C απ' όπου προκύπτει ότι L = { {Bread}, {PeanutButter}, {Bread, PeanutButter}, {Beer}, {Milk} }
5. ML = $BD^{-}(PL)$ ∩ L = { {Beer}, {Milk} } ≠ ∅
6. C = L;
   επαναληπτικά: C = C ∪ $BD^{-}(C)$

   // στο παράδειγμα, θα γίνουν 3 επαναλήψεις: για 2-, 3-, 4-itemsets
7. (2ο πέρασμα στη ΒΔ) Μέτρηση του C απ' όπου προκύπτει ότι L = L (που ήδη βρέθηκαν στο βήμα 4) ∪ ∅ (από το βήμα 7)

Υπέρ και Κατά της Δειγματοληψίας

- **Πλεονεκτήματα**:
  - Μειώνει το πλήθος των σαρώσεων της ΒΔ σε 1 (στην καλύτερη περίπτωση) ή 2 (στη χειρότερη περίπτωση).
  - Καλύτερη κλιμάκωση (αποδοτικό σε μικρές αλλά και μεγάλες ΒΔ).
- **Μειονεκτήματα**:
  - Ενδεχομένως μεγάλος αριθμός υποψηφίων itemsets στη δεύτερη σάρωση (πολλά υποψήφια λόγω επαναληπτικού υπολογισμού του αρνητικού ορίου)

Περιεχόμενα

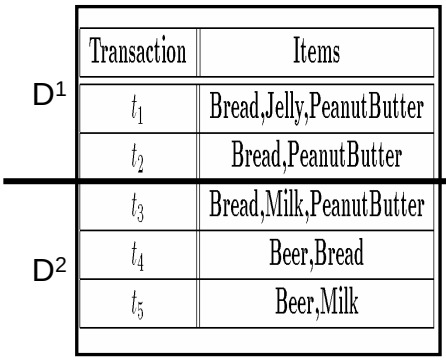
- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - Αλγόριθμος δειγματοληψίας
  - **Αλγόριθμος διαμερισμού**
  - Παράλληλοι αλγόριθμοι
- Ειδικά θέματα

Διαμερισμός

- Διαμερίζουμε τη ΒΔ σε τμήματα (partitions) D^1, D^2, …, D^p
- Εφαρμόζουμε τον Apriori σε κάθε τμήμα
- Κάθε συχνό itemset θα πρέπει να είναι συχνό σε ένα τουλάχιστον από τα τμήματα.
- Αλγόριθμος:

```
1.  Divide D into partitions D^1,D^2,…,D^p;
2.  For i = 1 to p do
3.      L^i = Apriori(D^i);                  // 1st pass
4.  C = L^1 ∪ … ∪ L^p;
5.  Count C on D to generate L;              // 2nd pass
```

Παράδειγμα διαμερισμού



Transaction	Items
t_1	Bread,Jelly,PeanutButter
t_2	Bread,PeanutButter
t_3	Bread,Milk,PeanutButter
t_4	Beer,Bread
t_5	Beer,Milk

D1

D2

L^1 ={{Bread}, {Jelly}, {PeanutButter}, {Bread,Jelly}, {Bread,PeanutButter}, {Jelly, PeanutButter}, {Bread,Jelly,PeanutButter}}

L^2 ={{Bread}, {Milk}, {PeanutButter}, {Bread,Milk}, {Bread,PeanutButter}, {Milk, PeanutButter}, {Bread,Milk,PeanutButter}, {Beer}, {Beer,Bread}, {Beer,Milk}}

s =10%

Υπέρ και Κατά του Διαμερισμού

- **Πλεονεκτήματα**:
  - Προσαρμόζεται στη διαθέσιμη κύρια μνήμη
  - Υλοποιείται εύκολα σε παράλληλη μορφή
  - Μέγιστο πλήθος σαρώσεων της ΒΔ = 2.
    - 1η σάρωση: διαμερισμός σε τμήματα που χωρούν στην κύρια μνήμη (και εκτέλεση Apriori στην κύρια μνήμη)
    - 2η σάρωση: για να εντοπιστούν τα πραγματικά συχνά itemsets από τα υποψήφια που προέκυψαν από την 1η σάρωση
- **Μειονεκτήματα**:
  - Μπορεί να έχει πολλά υποψήφια itemsets κατά την διάρκεια της δεύτερης σάρωσης.

Περιεχόμενα

- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - Αλγόριθμος δειγματοληψίας
  - Αλγόριθμος διαμερισμού
  - **Παράλληλοι αλγόριθμοι**
- Ειδικά θέματα

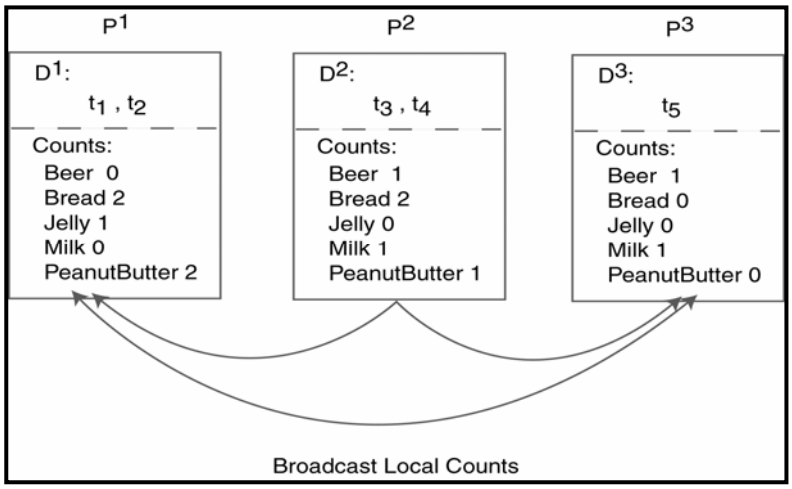
Παράλληλοι Αλγόριθμοι Κανόνων Συσχετίσεων

- Βασίζονται στον Apriori
- Οι τεχνικές διαφέρουν ανάλογα με το:
  - τι μετράμε σε κάθε τόπο (site)
  - πως είναι κατανεμημένα τα δεδομένα (δοσοληψίες)
- Παραλληλισμός δεδομένων
  - διαμέριση δεδομένων, εκπομπή μετρητών: Count Distribution Algorithm – CDA
- Παραλληλισμός εργασιών
  - διαμέριση δεδομένων και υποψηφίων, εκπομπή μετρητών και δεδομένων: Data Distribution Algorithm - DDA

Παραλληλισμός Δεδομένων
Count Distribution Algorithm (CDA)

```
1.   Place data partition at each site.
2.   In Parallel at each site do
3.     C1 = Itemsets of size one in I;
4.     Count C1;
5.     Broadcast counts to all sites;
6.     Determine global large itemsets of size 1, L1;
7.     i = 1;
8.     Repeat
9.       i = i + 1;
10.      Ci = Apriori-Gen(Li-1);
11.      Count Ci;
12.      Broadcast counts to all sites;
13.      Determine global large itemsets of size i,Li;
14.    until no more large itemsets found;
```

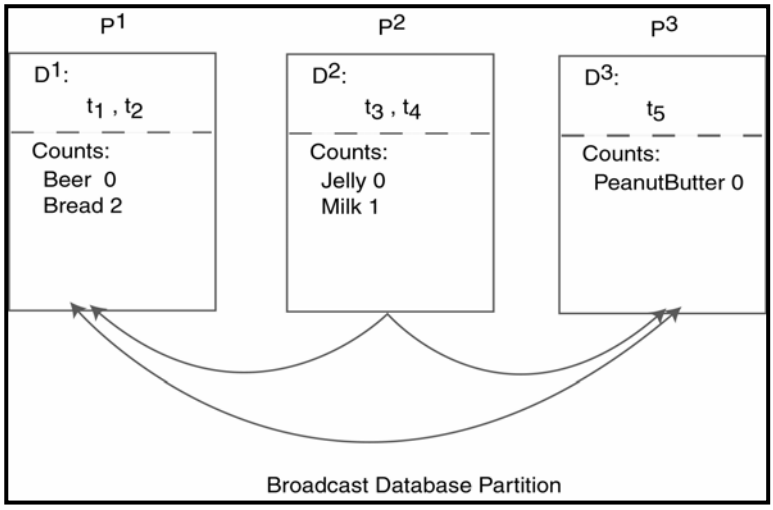
Παράδειγμα CDA



Παραλληλισμός Εργασιών
Data Distribution Algorithm(DDA)

```
1.   Place data partition at each site.
2.   In Parallel at each site do
3.    Determine local candidates of size 1 to count;
4.    Broadcast local transactions to other sites;
5.    Count local candidates of size 1 on all data;
6.    Determine large itemsets of size 1 for local candidates;
7.    Broadcast large itemsets to all sites;
8.    Determine L1;
9.    i = 1;
10.   Repeat
11.    i = i + 1;
12.    Ci = Apriori-Gen(Li-1);
13.    Determine local candidates of size i to count;
14.    Count, broadcast, and find  Li;
15.   until no more large itemsets found;
```

Παράδειγμα DDA

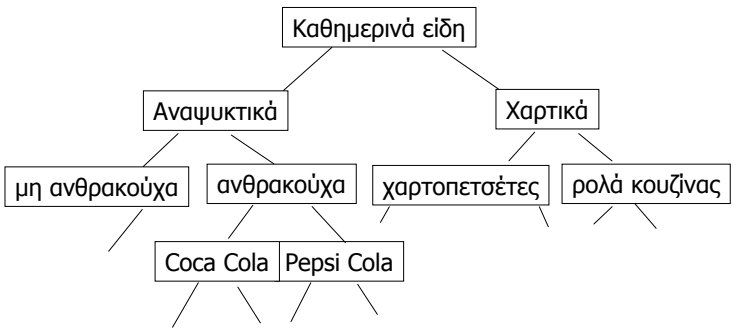


Περιεχόμενα

- Παρουσίαση του προβλήματος εύρεσης Κανόνων Συσχετίσεων
  - Συχνά σύνολα αντικειμένων *(frequent itemsets)*
- Αλγόριθμοι εύρεσης συχνών itemsets
  - Αλγόριθμος Apriori
  - Αλγόριθμος δειγματοληψίας
  - Αλγόριθμος διαμερισμού
  - Παράλληλοι αλγόριθμοι
- **Ειδικά θέματα**

Ειδικά θέματα Κανόνων Συσχετίσεων

- Πολυεπίπεδοι (multiple-level) κανόνες συσχετίσεων
  - Στόχος: να ληφθεί υπόψη πιθανή «ιεραρχία» των items
  - π.χ. **{Αναψυκτικά} ⇒ {Χαρτικά}**
  - Σχετικό με τις ιεραρχίες εννοιών (concept hierarchies) των Αποθηκών Δεδομένων

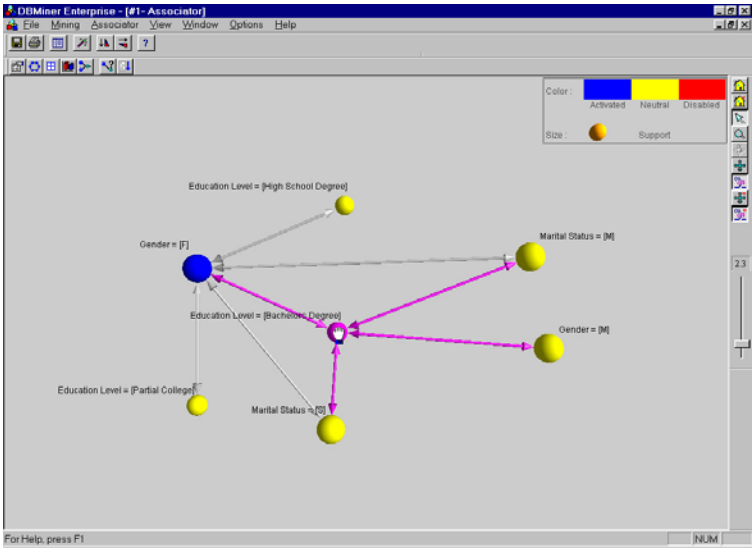


Ειδικά θέματα Κανόνων Συσχετίσεων (συν.)

- Αυξητικοί (incremental) αλγόριθμοι
  - Πρόβλημα: οι αλγόριθμοι υποθέτουν στατικές ΒΔ
  - Αντικειμενικός σκοπός: Αν γνωρίζουμε τα συχνά itemsets του D να βρούμε τα συχνά itemsets του D′ = D ∪ {ΔD}
  - Το itemset που είναι συχνό στο D′ πρέπει να είναι συχνό είτε στο D είτε στο ΔD
  - Χρειάζεται να αποθηκευτούν τα συχνά itemsets αλλά και οι μετρήσεις τους (για να είναι διαθέσιμα για την επόμενη φορά)

Οπτικοποίηση Κανόνων Συσχετίσεων

(Πηγή: "Data Mining: Concepts and Techniques", Han & Kamber)



Εφαρμογές Κανόνων Συσχετίσεων

- Ανάλυση καλαθιού αγοράς
  - Τοποθέτηση προϊόντων στα ράφια
  - Διαφήμιση
  - Πωλήσεις
  - Κουπόνια – Εκπτώσεις
- Μερικές φορές δεν είναι προφανής η εφαρμογή:

Wal-Mart knows that customers who buy Barbie dolls (it sells one every 20 seconds) have a 60% likelihood of buying one of three types of candy bars.
What does Wal-Mart do with information like that? '
I don't have a clue,' says Wal-Mart's chief of merchandising, Lee Scott.

(Πηγή: C. Palmeri. Believe in yourself, believe in the merchandise. Forbes 160(5), 8-Sep-1997. Also at: http://www.kdnuggets.com/news/98/n01.html)

Σύνοψη

- Εύρεση Κανόνων Συσχετίσεων: η εύρεση κανόνων της μορφής $X \Rightarrow Y$ μέσα από μια βάση δοσοληψιών
  - με βάση 2 κατώφλια (ελάχιστης υποστήριξης και ελάχιστης εμπιστοσύνης)
- Ο πιο δημοφιλής αλγόριθμος: Apriori
  - βασίζεται στην ιδιότητα των συχνών στοιχειοσυνόλων (frequent itemsets property)
- Άλλοι αλγόριθμοι: δειγματοληψίας, διαμερισμού, παράλληλοι