Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων

Ενότητα 11: Κανόνες Συσχέτισης – Μέρος Α΄

Αναστάσιος Γούναρης, Επίκουρος Καθηγητής
Τμήμα Πληροφορικής ΑΠΘ

BY NC SA

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

- Το παρόν εκπαιδευτικό υλικό υπόκειται σε άδειες χρήσης Creative Commons.
- Για εκπαιδευτικό υλικό, όπως εικόνες, που υπόκειται σε άλλου τύπου άδειας χρήσης, η άδεια χρήσης αναφέρεται ρητώς.

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Κανόνες Συσχέτισης – Μέρος Α΄

Εξόρυξη κανόνων συσχέτισης, αλγόριθμος Apriori

CC BY NC SA

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Περιεχόμενα ενότητας

1. Εξόρυξη κανόνων συσχέτισης.
2. Ο αλγόριθμος Apriori.

Σκοποί ενότητας

- Παρουσίαση της λειτουργίας της εξόρυξης κανόνων συσχέτισης από σύνολα δεδομένων.
- Περιγραφή βασικών ορισμών και παραδειγμάτων κανόνων συσχέτισης.
- Ανάλυση του αλγόριθμου Apriori.

Συσχετίσεις

- Κανόνες για τις σχέσεις μεταξύ των αντικειμένων.
- Παράδειγμα:
  - Δημητριακά, γάλα → φρούτα
  - “τα άτομα που αγόρασαν δημητριακά και γάλα, αγόρασαν επίσης και φρούτα.”
  - Σκοπός: Π.χ., μπορεί να γίνει ειδική προσφορά για όσους αγοράζουν γάλα και δημητριακά, να αγοράσουν και φρούτα σε καλύτερες τιμές.

Ανάλυση καλαθιού αγορών

- Ανάλυση των συναλλαγών
  - Χωρίς την πληροφορία για το πλήθος αντικειμένων.

Person	**Basket**
A	Chips, Salsa, Cookies, Crackers, Coke, Beer
B	Lettuce, Spinach, Oranges, Celery, Apples, Grapes
C	Chips, Salsa, Frozen Pizza, Frozen Cake
D	Lettuce, Spinach, Milk, Butter

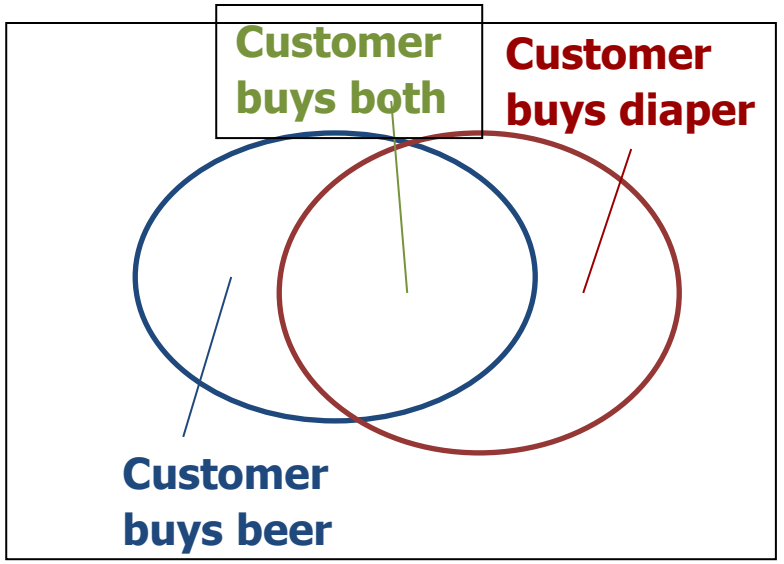
- Πως μπορούμε να ελέγχουμε υποθέσεις;
  - Chips → Salsa Lettuce → Spinach

Ορισμοί

- Έστω $I = \{i_1, \ldots, i_m\}$ το σύνολο των αντικειμένων.
- Έστω **D** το σύνολο των συναλλαγών:
  - Όπου για κάθε $T \in D$, $T \subseteq I$
- Στοιχειοσύνολο (Itemset) **X**: κάθε $X \subseteq I$:
  - k-itemset: $|X| = k$
- Το **X** περιέχεται στη συναλλαγή **T**, αν $X \subseteq T$
- Υποστήριξη **s(X)**: ποσοστό συναλλαγών που περιέχουν το **X.**

Κανόνες Συσχέτισης: Υποστήριξη κι εμπιστοσύνη

- Κανόνας *Beer* → *Diaper*
  - Υποστήριξη (support) s: πιθανότητα μία συναλλαγή να περιέχει {Beer ∪ Diaper}
  - Εμπιστοσύνη (confidence) c: υπο-συνθήκη πιθανότητα μία συναλλαγή που περιέχει {Beer} να περιέχει επίσης και Diaper.



Κωδ. Συναλλαγής	Αντικείμενα
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

- A → C (50%, 66.6%)
- C → A (50%, 100%)

Κανόνες Συσχέτισης: Ορισμός Προβλήματος

- Κανόνας Συσχέτισης (Association rule):

  $\mathbf{X \rightarrow Y,\ X \subset I,\ Y \subset I,\ X \cap Y = \varnothing}$

- Υποστήριξη Κανόνα = $\mathbf{s(X \cup Y)}$
- Εμπιστοσύνη Κανόνα = $\mathbf{s(X \cup Y) / s(X)}$
- Πρόβλημα: Βρες όλους τους κανόνες συσχέτισης με εμπιστοσύνη ίση ή μεγαλύτερη από MINCONF και υποστήριξη ίση ή μεγαλύτερη από MINSUP
- Προσοχή:
  - αν $X \rightarrow Y$, $Y \rightarrow Z$, αυτό δεν σημαίνει ότι $X \rightarrow Z$ καθώς ο $X \rightarrow Z$ μπορεί να μην έχει την απαιτούμενη υποστήριξη ή εμπιστοσύνη.

Εξόρυξη Κανόνων Συσχέτισης - Παράδειγμα

- Για τον κανόνα $A \Rightarrow C$:
  - support = support($\{A \cup C\}$) = 50%
  - confidence = support($\{A \cup C\}$)/support($\{A\}$) = 66.6%

Κωδ. Συναλλαγής	Αντικείμενα
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Ελάχιστη υποστήριξη: 50%
Ελάχιστη εμπιστοσύνη: 50%

Συχνά στοιχειοσύνολα	Υποστήριξη
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

Εξόρυξη Κανόνων Συσχέτισης

- Δύο βασικά βήματα – υποπροβλήματα:
  - Βρες όλα τα συχνά στοιχειοσύνολα.
    - Δηλ. τα στοιχειοσύνολα που ικανοποιούν τον περιορισμό ελάχιστης υποστήριξης.
  - Βρες συχνούς και αξιόπιστους κανόνες συσχέτισης.
    - Δημιούργησε κανόνες συσχέτισης από τα συχνά στοιχειοσύνολα.
    - Κράτησε μόνο αυτούς που ικανοποιούν τον περιορισμό ελάχιστης εμπιστοσύνης.

Δημιουργία συχνών στοιχειοσυνόλων

- **Αφελής (naive) αλγόριθμος:**

```
n ← |D|
for each subset s of I do
    m ← 0
    for each transaction T in D do
        if s is a subset of T then
            m ← m + 1
    if minimum support ≤ m/n then
        add s to frequent subsets
```

Ανάλυση αφελούς αλγορίθμου

- Έστω n συναλλαγές από m αντικείμενα.
  - $O(2^m)$ υποσύνολα s του I
  - Διάβασμα n συναλλαγών για κάθε ένα υποσύνολο.
  - $O(2^m n)$ έλεγχοι για το πλήθος εμφανίσεων των υποσυνόλων s.
  - Η αύξηση είναι εκθετική με τον αριθμό των αντικειμένων!
- Μπορούμε να βελτιώσουμε την απόδοση;
  - Μειώνοντας τα υποσύνολα που ελέγχουμε.
  - Μειώνοντας τις συγκρίσεις.

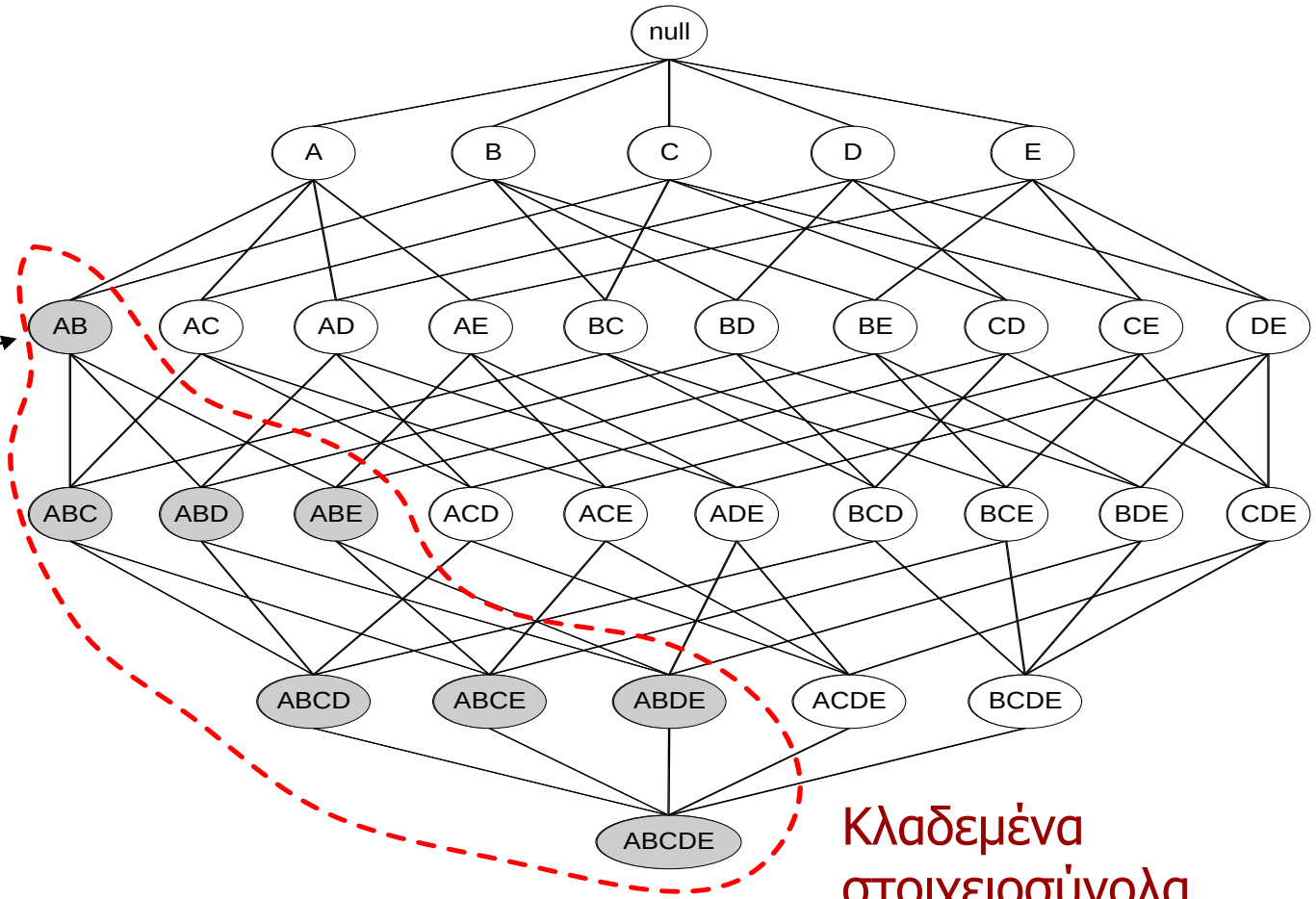
Μείωση υποψηφίων συχνών στοιχειοσυνόλων: Apriori

- Αρχή Apriori:
  - Αν ένα στοιχειοσύνολο είναι συχνό, τότε όλα τα υποσύνολά του είναι επίσης συχνά.
- Ισχύει λόγω της ακόλουθης ιδιότητας σχετικά με την υποστήριξη:

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

  - Η υποστήριξη ενός στοιχειοσυνόλου ποτέ δεν υπερβαίνει την υποστήριξη των υποσυνόλων του.
  - Αυτή η ιδιότητα είναι γνωστή ως η *αντι-μονότονη* ιδιότητα της υποστήριξης.

Αναπαράσταση της αρχής Apriori



Έστω ότι βρέθηκε μη συχνό

Κλαδεμένα στοιχειοσύνολα

Παράδειγμα εφαρμογής του Apriori

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

Μονά αντικείμενα (1-itemsets)

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

Ζεύγη (2-itemsets)

(Δεν χρειάζεται η δημιουργία ζευγών που περιέχουν Coke ή Eggs)

Minimum Support = 3

Αν ελέγχαμε όλα τα στοιχειοσύνολα:
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$
Με κλάδεμα λόγω της αρχής Apriori:
$6 + 6 + 1 = 13$

Itemset	Count
{Bread,Milk,Diaper}	3

Τριπλέτες (3-itemsets)

Ο Αλγόριθμος Apriori

- Join Step: C_k is generated by joining L_{k-1} with itself.
- Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset.
- Pseudo-code:

C_k: Candidate itemsets of size k
L_k : frequent itemsets of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) **do begin**
  C_{k+1} = candidates generated from L_k; // διαφάνεια 20

  for each transaction *t* in database do // διαφάνεια 22
    increment the count of all candidates in C_{k+1} that are contained in t
    L_{k+1} = candidates in C_{k+1} with min_support
end
return $\cup_k L_k$

Δημιουργία υποψήφιων στοιχειοσυνόλων

- Έστω ότι τα αντικείμενα L_{k-1} ταξινομούνται
  - (π.χ., λεξικογραφικά).
- **Βήμα 1: Σύνδεση του L_{k-1} με τον εαυτό του**
  - για τα στοιχειοσύνολα με κοινά τα πρώτα *k-2* αντικείμενα

  insert into C_k

  select $p.item_1$, $p.item_2$, …, $p.item_{k-1}$, $q.item_{k-1}$

  from L_{k-1} p, L_{k-1} q

  where $p.item_1=q.item_1$, …, $p.item_{k-2}=q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$
- **Βήμα 2: Κλάδεμα**

  for all itemsets c in C_k do

  for all (k-1)-subsets s of c do

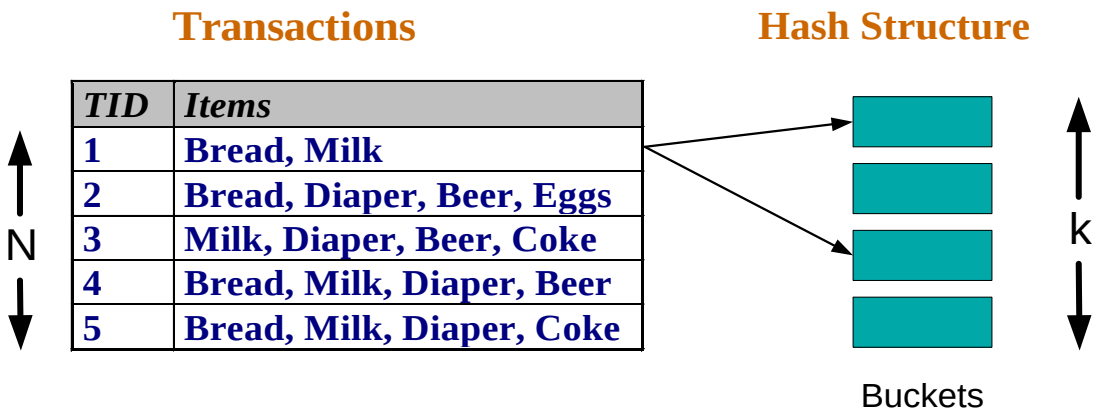
  if (s is not in L_{k-1}) then delete c from C_k

Παράδειγμα

- L_3={abc, abd, acd, ace, bcd}
- Self-joining: L_3*L_3
  - abcd από abc και abd
  - acde από acd και ace
- Κλάδεμα:
  - acde: διαγράφεται γιατί το ade δεν είναι στο L_3
- C_4={abcd}

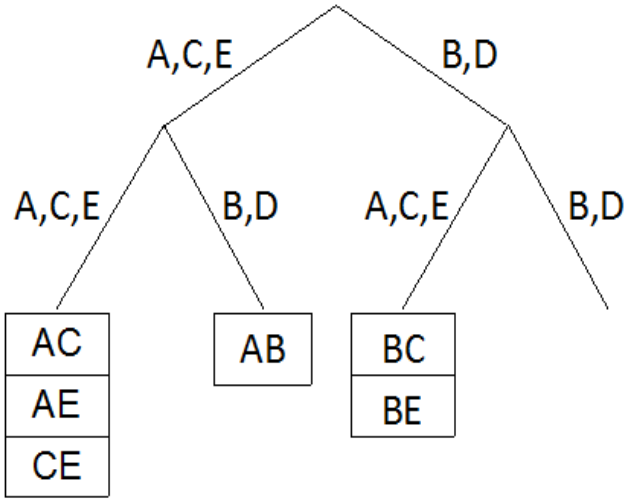
Πως μετράται η υποστήριξη;

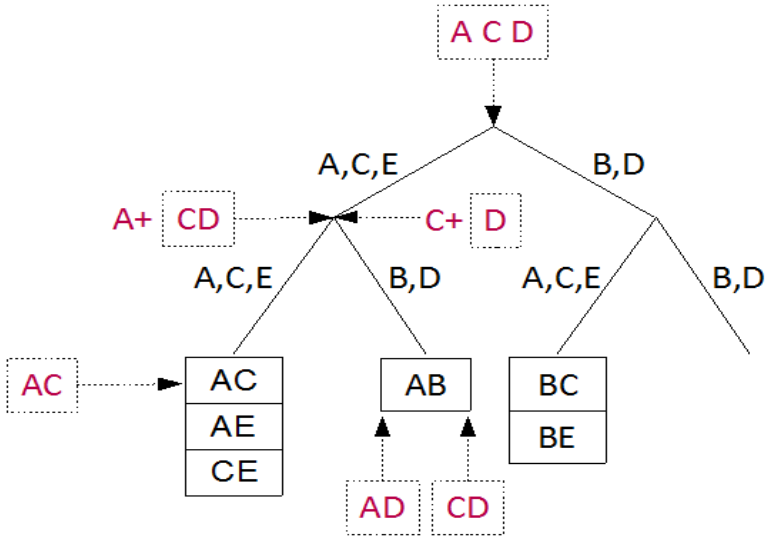
- Μέτρηση υποστήριξης για υποψήφια συχνά στοιχειοσύνολα:
  - Διάβασε τις συναλλαγές για να διαπιστωθεί η υποστήριξη για κάθε ένα υποψήφιο συχνό στοιχειοσύνολο.
  - Για μείωση των συγκρίσεων, αποθήκευσε τα υποψήφια ΣΣ σε μία δομή κατακερματισμού.
    - Αντί να ελέγχεται κάθε συναλλαγή με κάθε υποψήφιο ΣΣ, ελέγχεται μόνο με τα ΣΣ στους κάδους της δομής.



Παράδειγμα

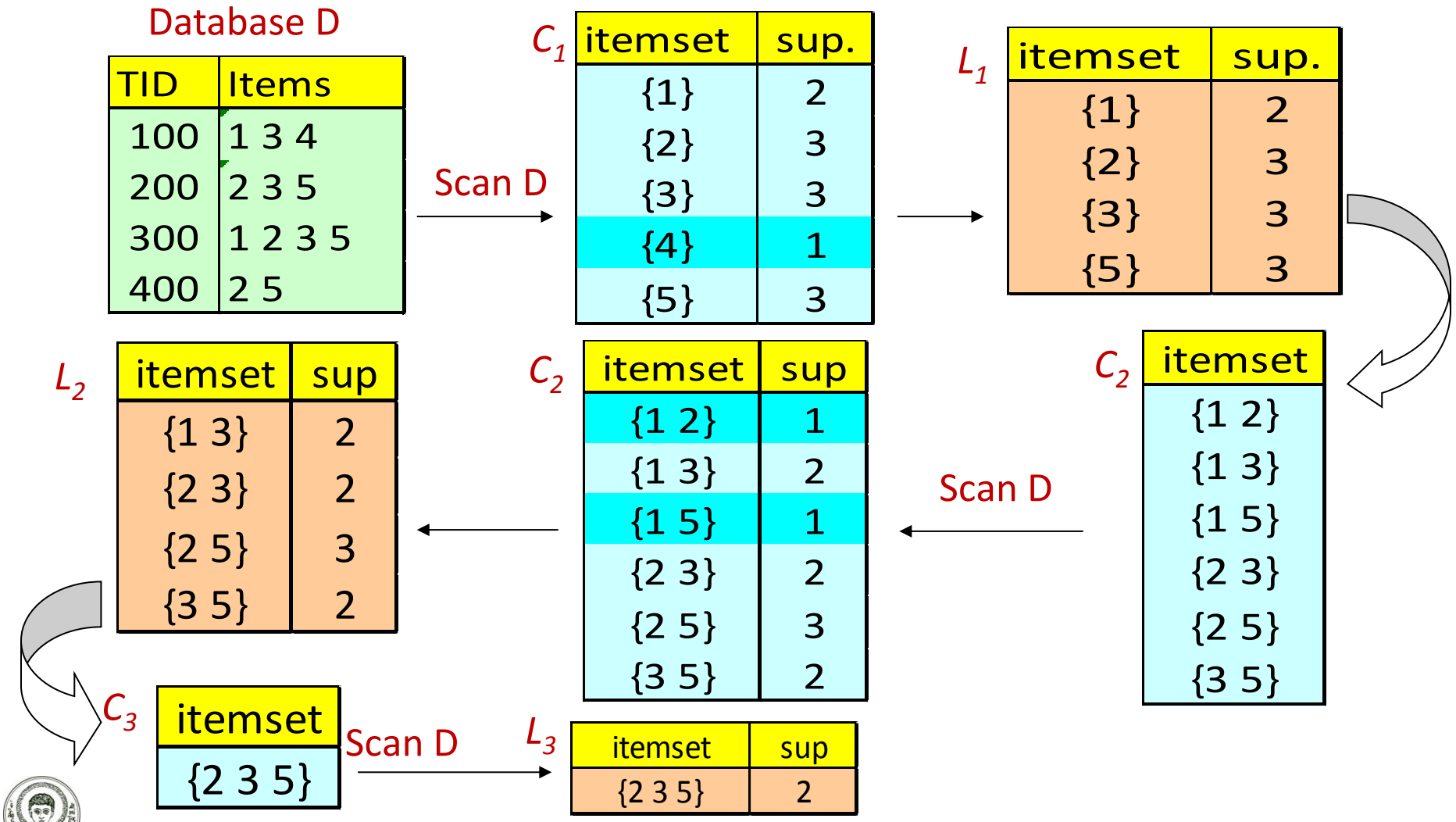
- b (αριθμός παιδιών)=2
- c (μέγεθος φύλλων) =3 στο παράδειγμα.





- **Όφελος: Μία σάρωση όλων των συναλλαγών για όλα τα υποψήφια σύνολα ορισμένου μήκους**.

Αλγόριθμος Apriori: Συνολικό Παράδειγμα



Άλλο ένα παράδειγμα

- Minsup = 2

TID	List of Items
1	1, 2, 5
2	2, 4
3	2, 3
4	1, 2, 4
5	1, 3
6	2, 3
7	1, 3
8	1, 2, 3, 5
9	1, 2, 3

Άλλο ένα παράδειγμα: Λύση (1/3)

TID	List of Items
1	1, 2, 5
2	2, 4
3	2, 3
4	1, 2, 4
5	1, 3
6	2, 3
7	1, 3
8	1, 2, 3, 5
9	1, 2, 3

Scan D for count of each candidate

C_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Compare candidate support count with minimum support count

L_1

Itemset	Sup. Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

Άλλο ένα παράδειγμα: Λύση (2/3)

Generate C_2 candidates from L_1

C_2

Itemset
{I1, I2}
{I1, I3}
{I1, I4}
{I1, I5}
{I2, I3}
{I2, I4}
{I2, I5}
{I3, I4}
{I3, I5}
{I4, I5}

Scan D for count of each candidate

C_2

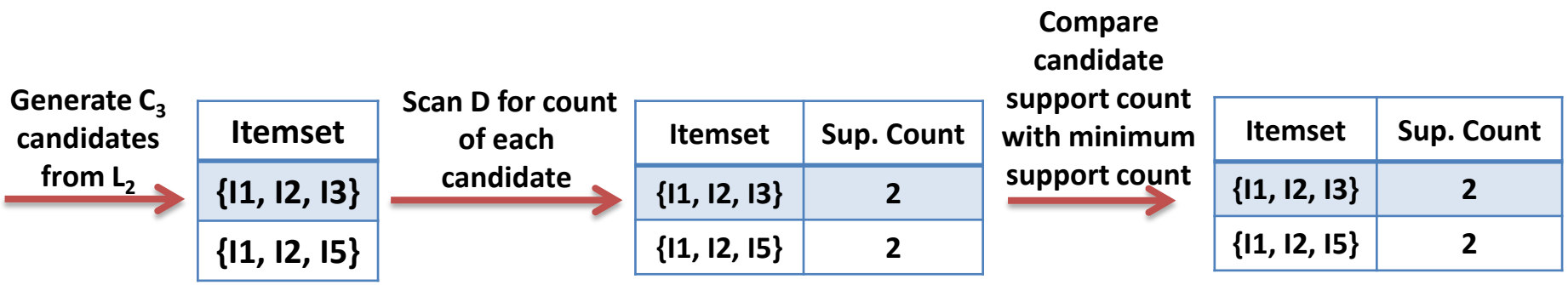
Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I4}	1
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2
{I3, I4}	0
{I3, I5}	1
{I4, I5}	0

Compare candidate support count with minimum support count

L_2

Itemset	Sup. Count
{I1, I2}	4
{I1, I3}	4
{I1, I5}	2
{I2, I3}	4
{I2, I4}	2
{I2, I5}	2

Άλλο ένα παράδειγμα: Λύση (3/3)



Δημιουργία κανόνων (2ο υποπρόβλημα)

- Δεδομένου ενός συχνού στοιχειοσυνόλου ΣΣ **L**, βρες όλα τα μη κενά υποσύνολα **f⊂ L**, ώστε ο κανόνας:

  f → L – f να ικανοποιεί την ελάχιστη εμπιστοσύνη.

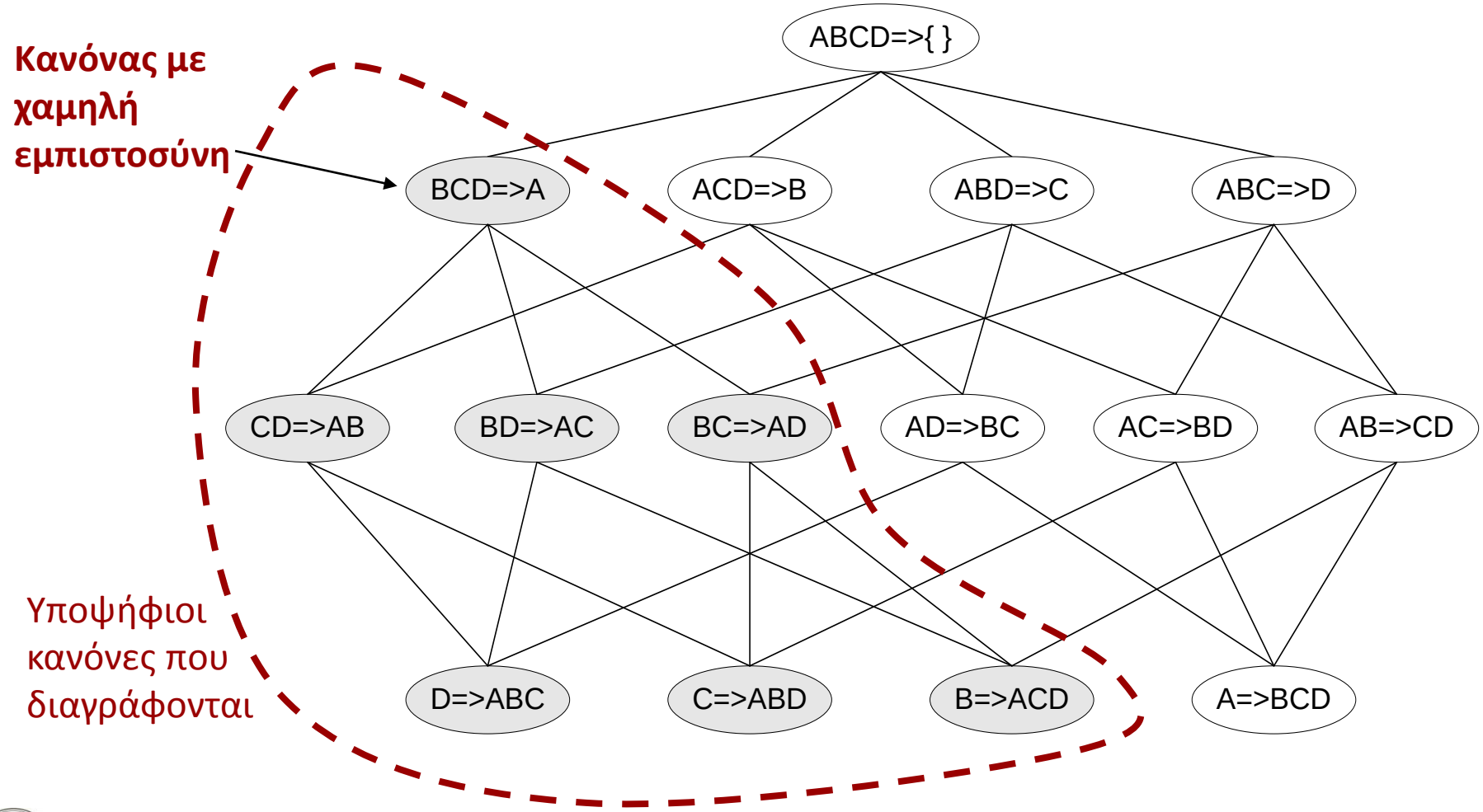
  - Αν {A,B,C,D} είναι ΣΣ, οι υποψήφιοι κανόνες είναι:
    - ABC →D, ABD →C, ACD →B, BCD →A,
      A →BCD, B →ACD, C →ABD, D →ABC
      AB →CD, AC → BD, AD → BC, BC →AD,
      BD →AC, CD →AB,

- Αν |L| = k, τότε υπάρχουν 2^k – 2 υποψήφιοι κανόνες συσχέτισης:
  - (εξαιρώντας τους L → ∅ και ∅ → L)

Δημιουργία Κανόνων - Αντιμονοτονικότητα

- Αποδοτική δημιουργία κανόνων από ΣΣ.
- Γενικά, δεν ισχύει η αντιμονοτονική ιδιότητα στην εμπιστοσύνη.
  - $c(ABC \rightarrow D) <> c(AB \rightarrow D)$
  - ΑΛΛΑ ισχύει όταν οι κανόνες έχουν δημιουργηθεί από το ΙΔΙΟ συχνό στοιχειοσύνολο.
  - Π.χ., L = {A,B,C,D}:
    $c(ABC \rightarrow D) \geq c(AB \rightarrow CD) \geq c(A \rightarrow BCD)$
  - Η εμπιστοσύνη είναι αντιμονότονη, όσον αφορά τον αριθμό των στοιχείων το δεξί κομμάτι του κανόνα.

Παράδειγμα


Κανόνας με χαμηλή εμπιστοσύνη
ABCD=>{ }
BCD=>A
ACD=>B
ABD=>C
ABC=>D
CD=>AB
BD=>AC
BC=>AD
AD=>BC
AC=>BD
AB=>CD
Υποψήφιοι κανόνες που διαγράφονται
D=>ABC
C=>ABD
B=>ACD
A=>BCD

Μέθοδοι βελτίωσης της αποδοτικότητας του Apriori

- **Μέτρηση υποστήριξης στοιχειοσυνόλων βασισμένη σε κατακερματισμό**:
  - Ένα k-στοιχειοσύνολο με μέγεθος του αντίστοιχου κάδου κάτω από το κατώφλι, δεν μπορεί να είναι συχνό.
- **Μείωση συναλλαγών**:
  - Μία συναλλαγή που δεν περιέχει κανένα συχνό k-στοιχειοσύνολο μπορεί να αγνοηθεί σε μεταγενέστερες σαρώσεις.
- **Κατάτμηση**:
  - Ένα υποψήφιο στοιχειοσύνολο πρέπει να είναι συχνό σε τουλάχιστον ένα τμήμα της ΒΔ.
- **Δειγματοληψία**:
  - Εξόρυξη σε ένα δείγμα των δεδομένων, μικρότερο κατώφλι υποστήριξης, μέθοδος για επίτευξη πληρότητας.
- **Δυναμική μέτρηση υποστήριξης**:
  - Προσθήκη νέων υποψήφιων στοιχειοσυνόλων όταν όλα τα υποσύνολά τους είναι συχνά.

Μέθοδος βασισμένη σε κατακερματισμό

- $|C2| = comb(|L1|, 2) = O(|L1|^2)$
  - $|L1|$ φθάνει τις χιλιάδες, $|C2|$ bottleneck.
- Καθώς μετράται η υποστήριξη του C1:
  - Βρες τα 2-στοιχειοσύνολα σε κάθε συναλλαγή.
  - Εφάρμοσε κατακερματισμό μετρώντας το μέγεθος των κάδων.
  - Αν ένα 2-στοιχειοσύνολο ανήκει σε κάδο με μέγεθος μικρότερο από MINSUP, τότε δεν είναι συχνό.

Παράδειγμα

100	ACD
200	BCE
300	ABCE
400	BE

C_1

{A} (2)

{B} (3)

{C} (3)

{D} (1)

{E} (3)

L_1

{A}

{B}

{C}

{E}

L_1 x L_1

{A,B}

{A,C}

{A,E}

{B,C}

{B,E}

{C,E}

C_2

{A,B}

{A,C}

{A,E}

{B,C}

{B,E}

{C,E}

Min sup = 2

3	*3*	*0*	*3*	*1*	*3*
{C,E}	{B,C}		{B,E}	{A,B}	{A,C}
{C,E}	{B,C}		{B,E}		{C,D}
{A,D}	{A,E}		{B,E}		{A,C}

Μείωση συναλλαγών

- Καθώς υπολογίζεται το C_k (k φάση)
  - Ένα αντικείμενο σε μια συναλλαγή t μπορεί ν' αφαιρεθεί, αν δεν εμφανίζεται σε τουλάχιστον k στοιχεία του C_k .
  - Παράδειγμα: τα ABC, ABD, BCD είναι υποψήφια στοιχειοσύνολα, ενώ το ACD δεν είναι.
    - Αν μία συναλλαγή t περιέχει το A, τότε το A μπορεί να αφαιρεθεί από το t, επειδή δεν θα συμμετέχει σε μελλοντικούς ελέγχους (το ABCD δεν θα δημιουργηθεί).
    - Αν η συναλλαγή t δεν περιέχει περισσότερα από k αντικείμενα, αφαιρείται.

Παράδειγμα

100	ACD
200	BCE
300	ABCE
400	BE

C_1

{A} (2)

{B} (3)

{C} (3)

{D} (1)

{E} (3)

L_2

{A,C}

{B,C}

{B,E}

{C,E}

100	ACD
200	BCE
300	ABCE
400	BE

{C} → Αφαίρεση A (και D ως συχνό)

{B,C,E} → Καμία αλλαγή

{B,C,E} → Αφαίρεση A

{B,E} → Αφαίρεση εγγραφής

Το A αφαιρείται γιατί δεν μπορεί να συμμετάσχει σε 3-ΣΣ.

Τμηματοποίηση (Partitioning)

- Λογική κατάτμηση της ΒΔ σε μη αλληλο-επικαλυπτόμενα τμήματα.
  - Κάθε τμήμα χωράει εξ ολοκλήρου στην μνήμη.
- Εύρεση των συχνών στοιχειοσυνόλων σε κάθε τμήμα.
  - Όλες οι φάσεις επιτελούνται στην κύρια μνήμη, μόνο μία φορά το κάθε τμήμα διαβάζεται από τον δίσκο.
- Καθορισμός ποια σύνολα είναι και συνολικά συχνά.
  - Μία σάρωση της ΒΔ.
  - Κανένα συχνό στοιχειοσύνολο δε χάνεται.

Αλγόριθμος

partition_database(*D*)

n = Number of partitions

//Phase I

for *i*=1 to *n* do

read *i* -th partition

L^i = gen_all_large_itemsets(*i*-th partition)

$C = L^1 \cup \ldots \cup L^n$

for *i*=1 to *n* do

read *i* -th partition

for each candidate $c \in C$ update *c*.count

$L = \{c \mid c \in C, c.\text{count} \geq MINSUP\}$

Παράδειγμα

	Transaction	Items
D¹	t_1	Bread, Jelly, PeanutButter
	t_2	Bread, PeanutButter
D²	t_3	Bread, Milk, PeanutButter
	t_4	Beer, Bread
	t_5	Beer, Milk

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}}

Ορθότητα

- Κανένα συχνό στοιχειοσύνολο δεν πρέπει να χάνεται.
- Ένα συνολικά συχνό στοιχειοσύνολο X δε χάνεται αν είναι υποψήφιο στη 2η φάση, δηλ. αν είναι συχνό σε τουλάχιστον ένα τμήμα.
  - Έστω X ένα συνολικά συχνό σύνολο που δεν είναι τοπικά συχνά πουθενά. Θα δούμε ότι αυτό είναι αδύνατο να συμβεί.
  - Έστω $t(X,P^i)$ το πλήθος των εμφανίσεων του X στο i-οστό τμήμα.
  - **Απόδειξη**:

    $t(X, P^i) < MINSUP \times |P^i| \quad \forall\ 1 \leq i \leq n \Rightarrow$

    $\Sigma\ t(X, P^i) < MINSUP \times \Sigma\ |P^i| \Rightarrow$

    $t(X, D) < MINSUP \times |D| \Rightarrow$

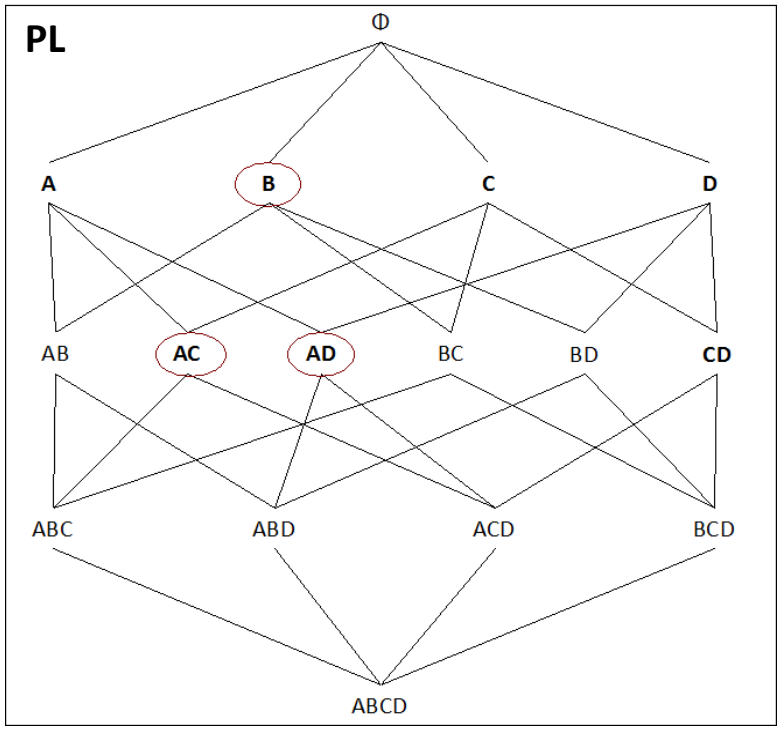
    $s(X) < MINSUP$ (contradiction)

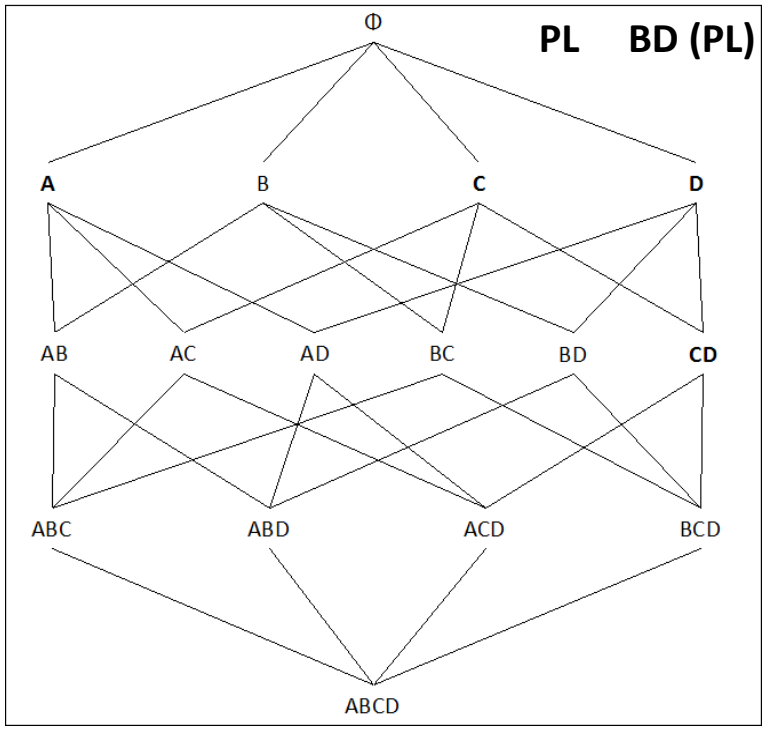
Μέγεθος υποψηφίων συνόλων στη 2η φάση

- Μπορούμε να κάνουμε την παραδοχή ότι το μέγεθος |C | από όλα τα τοπικά ΣΣ είναι αρκετά μικρό.
  - και συγκρίσιμο με το |L| (global large).
- Το ίδιο MINSUP χρησιμοποιείται, άρα κανείς θα περίμενε να είναι ίδιο με το |L|.
  - Στη χειρότερη περίπτωση n × |L|.
  - Στην πράξη, υπάρχει επικάλυψη (το ίδιο ΣΣ σε διαφορετικά τμήματα).
- Για μικρό n, |C | κοντά στο |L|.
- Για μεγαλύτερα n, |C | πιο μεγάλο αλλά όχι n × |L|.

Αρνητικό όριο - Negative border

- Στοιχειοσύνολα των οποίων καταμετρήθηκε η υποστήριξη και δεν βρέθηκαν συχνά.
- Ή αυτά των οποίων όλα τα υποσύνολα είναι συχνά.





Ορθότητα

- Ο στόχος είναι να μην χαθούν στοιχειοσύνολα που είναι συχνά σε όλη τη βάση:
  - Αν το X είναι συχνό στη D, αλλά όχι στο δείγμα S, τότε $\exists$ Y $\subseteq$ X, Y$\in$ Bd-(S).
- Ελέγχουμε αν κάποιο σύνολο στο Bd-(S) είναι συχνό σε όλη τη βάση D.
  - Επιλογή του Minsup στο S, ώστε η πιθανότητα αρχικής αποτυχίας να είναι μικρή αλλά τα αρχικά στοιχειοσύνολα να χωρούν στην μνήμη.

Παράδειγμα

- Domain = {A, B, C, D, E, F}, minsup = 25%
- Δείγμα = {A,B,C}, {A,C,F}, {A,D}, {B,D}
- Bd-(S) = {B,F}, {C,D}, {D,F}, {E}
  - Υποψήφια σύνολα που δεν βρέθηκαν συχνά στο δείγμα.
- L = {A,B}, {B,F}, {A,C,F} συχνά σε όλη τη ΒΔ D.
  - Με όλα τα υποσύνολά τους.
- {B,F} είναι μια “αποτυχία”.
  - {A,B},{B,F} και {A,F} δημιουργούν το {A,B,F} που πιθανώς να είναι συχνό αλλά δεν έχει μετρηθεί η υποστήριξή του.

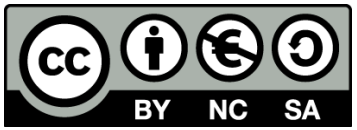
Σημείωμα Αναφοράς

Copyright Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Αναστάσιος Γούναρης. «Αποθήκες Δεδομένων και Εξόρυξη Δεδομένων. Ενότητα 11. Κανόνες Συσχέτισης – Μέρος Α΄». Έκδοση: 1.0. Θεσσαλονίκη 2014.

Διαθέσιμο από τη δικτυακή διεύθυνση:http://eclass.auth.gr/courses/OCRS182/

Σημείωμα Αδειοδότησης

Το παρόν υλικό διατίθεται με τους όρους της άδειας χρήσης Creative Commons Αναφορά - Μη Εμπορική Χρήση - Παρόμοια Διανομή 4.0 [1] ή μεταγενέστερη, Διεθνής Έκδοση. Εξαιρούνται τα αυτοτελή έργα τρίτων π.χ. φωτογραφίες, διαγράμματα κ.λ.π., τα οποία εμπεριέχονται σε αυτό και τα οποία αναφέρονται μαζί με τους όρους χρήσης τους στο «Σημείωμα Χρήσης Έργων Τρίτων».



Ο δικαιούχος μπορεί να παρέχει στον αδειοδόχο ξεχωριστή άδεια να χρησιμοποιεί το έργο για εμπορική χρήση, εφόσον αυτό του ζητηθεί.

Ως **Μη Εμπορική** ορίζεται η χρήση:

- που δεν περιλαμβάνει άμεσο ή έμμεσο οικονομικό όφελος από την χρήση του έργου, για το διανομέα του έργου και αδειοδόχο
- που δεν περιλαμβάνει οικονομική συναλλαγή ως προϋπόθεση για τη χρήση ή πρόσβαση στο έργο
- που δεν προσπορίζει στο διανομέα του έργου και αδειοδόχο έμμεσο οικονομικό όφελος (π.χ. διαφημίσεις) από την προβολή του έργου σε διαδικτυακό τόπο

[1] http://creativecommons.org/licenses/by-nc-sa/4.0/

ΑΡΙΣΤΟΤΕΛΕΙΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ
ΑΚΑΔΗΜΑΪΚΑ
ΜΑΘΗΜΑΤΑ

Τέλος ενότητας

Επεξεργασία: Ανδρέας Κοσματόπουλος

Θεσσαλονίκη, Χειμερινό Εξάμηνο 2013-2014

CC BY NC SA

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Σημειώματα

Διατήρηση Σημειωμάτων

Οποιαδήποτε αναπαραγωγή ή διασκευή του υλικού θα πρέπει να συμπεριλαμβάνει:

- το Σημείωμα Αναφοράς
- το Σημείωμα Αδειοδότησης
- τη δήλωση Διατήρησης Σημειωμάτων
- το Σημείωμα Χρήσης Έργων Τρίτων (εφόσον υπάρχει)

μαζί με τους συνοδευόμενους υπερσυνδέσμους.