| Università di Pisa | A.A. 2009- 2010 |
|---|---|

**Data Mining**

II verifica intermedia – 25 maggio 2010

**Soluzioni**

Esercizio 1 - Sequential Patterns (**6 punti**)

Si consideri la seguente sequenza di input:

| < {A} | {A,B,C} | {D} | {H} | {B,E} | {A,B,D} > |
|---|---|---|---|---|---|
| t=0 | t=1 | t=2 | t=3 | t=4 | t=5 |

Si indichi quali sono le occorrenze delle seguenti sotto-sequenze nella sequenza di input, senza considerare vincoli temporali (colonna sinistra) e considerando il vincolo temporale *max-gap = 1* (colonna destra). Per brevità, si rappresenti ogni occorrenza tramite la corrispondente ennupla di tempi nella sequenza di input, es.: <0,2,3> = <t=0, t=2, t=3>.

| | *Occorrenze* | *Occorrenze con max-gap=1* |
|---|---|---|
| *es.:* <{A}{D}{H}> | **<0,2,3>** **<1,2,3>** | **<1,2,3>** |
| $w_1$ = <{A} {B} {D} > | **<0,1,2>** **<0,1,5>** **<0,4,5>** **<1,4,5>** | **<0,1,2>** |
| $w_2$ = <{A}{H}{B}> | **<0,3,4>** **<0,3,5>** **<1,3,4>** **<1,3,5>** | **nessuna** |
| $w_2$ = <{A}{C}{E}> | **<0,1,4>** | **nessuna** |

Esercizio 2 – Itemset Frequenti (**12 punti**)

Considerare la seguente tabella di transazioni:

| ID | ITEMS |
|---|---|
| 1 | A B C |
| 2 | B |
| 3 | A B C D |
| 4 | C |
| 5 | A D |

| ID | ITEMS |
|---|---|
| 6 | B C E |
| 7 | C D E |
| 8 | A B |
| 9 | A B C D |
| 10 | A B D |

A) Elencare gli itemset frequenti nel caso di un supporto minimo min_sup = 25% ed indicare il loro supporto.

B) Quali itemset frequenti sono anche closed? Quali sono massimali?

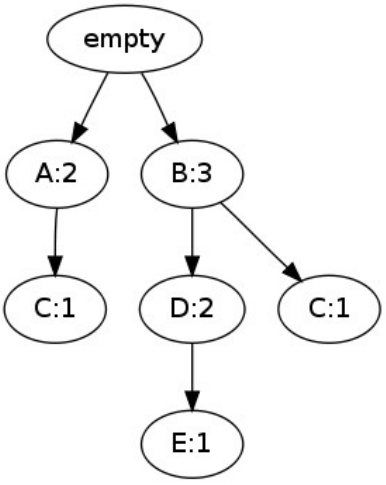
**Soluzione:** Elenco itemset frequenti

| | | |
|---|---|---|
| (c=closed, m=massimale) | | |
| c | | A (60.0) |
| c | | B (70.0) |
| c | | C (60.0) |
| c | | D (50.0) |
| c | | A B (50.0) |
| | | A C (30.0) |
| c | | A D (40.0) |
| c | | B C (40.0) |
| | | B D (30.0) |
| c | m | C D (30.0) |
| c | m | A B C (30.0) |
| c | m | A B D (30.0) |

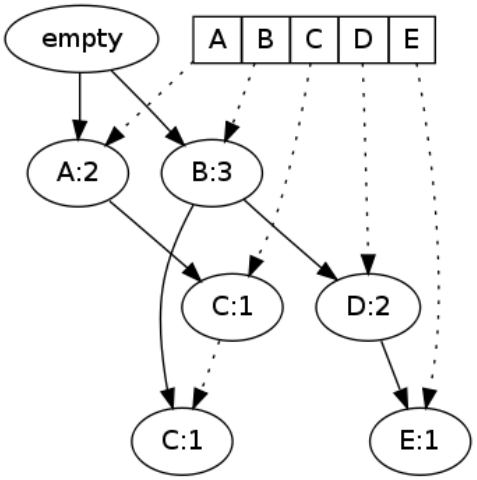
Esercizio 3 – FP-tree (**2 punti**)

Si disegni l' FP-tree corrispondente al seguente dataset (in figura a sinistra). **Soluzione a destra**.

| ID | Itemset |
|---|---|
| 1 | A |
| 2 | B D E |
| 3 | A C |
| 4 | B D |
| 5 | B C |



Semplice



Con link di collegamento tra items

Esercizio 4 - Clustering **(12 punti)**

Sul seguente dataset (figura a destra):

A) Si utilizzi l'algoritmo di clustering density-based DBSCAN, con raggio (ε) pari a 1.9, e minPts pari a 4 → **soluzione a destra**

B) Si disegni il dendogramma ottenuto con un algoritmo di clustering agglomerativo MIN-link (o *Single linkage*). → **soluzione in basso**