Università di Pisa	A.A. 2007- 2008

Data Mining - *Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda*

PARTE A : Regole Associative, Pattern Sequenziali

Tecniche Data Mining *- Corsi di Laurea Specialistica in Informatica e Tecnologie Informatiche*

PARTE B : Regole Associative, Pattern Sequenziali

Appello del 18 Luglio 2008

Esercizio 1 - Sequential Patterns (6 punti)

Si consideri la seguente sequenza W di input:

$$W = < \{3\}\{1,2,5\}\{1,2,4\}\{1,5\}\{4,6\}>$$

Si indichi quali delle seguenti sequenze sono sotto-sequenze semplici di W (senza vincoli temporali):

	$w_i \leq W$
$w_1 = <\{1\}\{2\}\{3\}>$	
$w_2 = <\{1,2,3,4\}\{5,6\}>$	
$w_3 = <\{2,4\}\{2,4\}\{6\}>$	
$w_4 = <\{1\}\{2,4\}\{6\}>$	
$w_5 = <\{1,2\}\{3,4\}\{5,6\}>$	

Esercizio 2 – Regole associative (**11** punti)

Si consideri il seguente insieme di transazioni:

Transazioni	Item Acquistati
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies }
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bred, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

a) Qual è il massimo numero di regole associative che possono essere estratte da questo database di transazioni (includendo anche regole con supporto uguale a 0)?
b) Qual è l'itemset più grande (di cardinalità maggiore) con supporto non nullo che può essere estratto?
c) Quanti 3-itemset si possono estrarre da questi dati?
d) Estrarre il 2-itemset con il supporto maggiore.
e) Trovare una coppia *a*, *b* tale che la regola {a}→{b} e la regola {b}→{a} abbiano la stessa confidenza

Esercizio 3 – Regole associative (**15** punti)

Dato il seguente insieme di transazioni:

Transazioni	Prodotti Acquistati
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,d,e}
6	{c,d}
7	{a,b,c}
8	{a,d,e}
9	{b,d}
10	{b,c,d,e}

a) Eseguire l'algoritmo *Apriori* per l'estrazione di itemset frequenti con min_sup = 30%, mostrando le varie fasi dell'algoritmo. Si determinino inoltre le regole associative valide rispetto ad una soglia di minima confidenza pari all'80%.
b) Esprimere, inoltre: la percentuale di itemset frequenti trovati (rispetto a tutti gli itemset che si potrebbero generare) e la percentuale di pruning dell'algoritmo (può essere definita come la percentuale di itemset che non sono stati considerati candidati perché (i) non sono stati generati durante la fase di generazione (ii) o sono stati tagliati dalla fase di pruning)