Università di Pisa	A.A. 2007- 2008

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

Verifica 3 aprile 2008 - **Soluzioni**

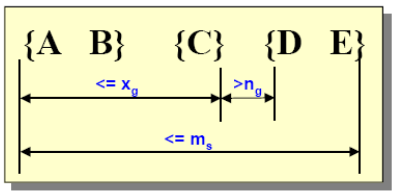
Esercizio 1 - Sequential Patterns (8 punti)

Si consideri la seguente sequenza W di input:

$$W = <\ \{a,b\}\ \{a,c\}\ \{c\}\ \{d,e\}\ \{a\}\ \{b,d\}\ >$$

Si indichi quali delle seguenti sequenze sono sotto-sequenze semplici di W (senza vincoli temporali), quali rispettano il vincolo min-gap=1, quali il vincolo max-gap=2, quali il vincolo max-span=3 e quali li rispettano tutti insieme:

- $w_1 = <\{a\}\ \{e\}\ \{a\}>$
- $w_2 = <\{a,b\}\ \{c\}\ \{d\}>$
- $w_3 = <\{a\}\ \{e\}\ \{d\}>$
- $w_4 = <\{a\}\ \{d\}\ \{c\}>$
- $w_5 = <\{a\}\ \{b\}>$



x_g: max-gap

n_g: min-gap

m_s: maximum span

A tale scopo si utilizzi la seguente tabella:

	$w_i \leq W$	min-gap=1	max-gap=2	max-span=3	Tutti
$w_1 = <\{a\}\{e\}\{a\}>$					
$w_2 = <\{a,b\}\{c\}\{d\}>$					
$w_3 = <\{a\}\{e\}\{d\}>$					
$w_4 = <\{a\}\{d\}\{c\}>$					
$w_5 = <\{a\}\{b\}>$					

SOLUZIONE:

	$w_i \leq W$	min-gap=1	max-gap=2	max-span=3	Tutti
$w_1 = <\{a\}\{e\}\{a\}>$	**X**		**X**	**X**	
$w_2 = <\{a,b\}\{c\}\{d\}>$	**X**	**X**	**X**	**X**	
$w_3 = <\{a\}\{e\}\{d\}>$	**X**	**X**	**X**		
$w_4 = <\{a\}\{d\}\{c\}>$					
$w_5 = <\{a\}\{b\}>$	**X**	**X**	**X**	**X**	

Esercizio 2 – Regole associative (**4** punti)

Si considerino le seguenti regole associative e relativo supporto e confidenza:

1	{beer} □{chips} 50% 100%
2	{beer} □{wine} 25% 50%
3	{chips} □{beer} 50% 66%
4	{pizza} □{chips} 25% 50%
5	{pizza} □{wine} 25% 50%
6	{wine} □{beer} 25% 50%
7	{wine} □{chips} 25% 50%
8	{wine} □{pizza} 25% 50%
9	{beer, chips} □{wine} 25% 50%
10	{beer, wine} □{chips} 25% 100%
11	{chips, wine} □{beer} 25% 100%
12	{beer} □{chips, wine} 25% 50%
13	{wine} □{beer, chips} 25% 50%

1) Determinare le regole valide rispetto al Min_Support = 25%. Min_conf= 60%
2) Determinare le regole valide rispetto al vincolo: "beer" è contenuto nell'antecedente. E' possibile utilizzare tale vincolo nella fase di generazione dei pattern frequenti?
3) Determinare le regole valide rispetto alle seguenti metaregole:

a	W and X => Y and Z
b	X => Y and Z
c	X => Y
d	W and X => Y

SOLUZIONE:

1) Determinare le regole valide rispetto al Min_Support = 25%. Min_conf= 60%
   {beer} =>{chips}
   {chips} =>{beer}
   {beer, wine} =>{chips}
   {chips, wine} =>{beer}

2) Determinare le regole valide rispetto al vincolo: "beer è contenuto nell'antecedente". E' possibile utilizzare tale vincolo nella fase di generazione dei pattern frequenti?

   {beer} =>{chips}
   {beer, wine} =>{chips}
   Tale vincolo può essere usato in fase di generazione dei pattern frequenti

3) Determinare le regole valide rispetto alle seguenti metaregole:
   a) Nessuna
   b) Nessuna
   c) {beer} =>{chips}
      {chips} =>{beer}
   d) {beer, wine} =>{chips}
      {chips, wine} =>{beer}

Esercizio 4 - Pattern Frequenti (12 punti)

Si consideri il seguente dataset:

	A	B	C	D	E
1	1	0	1	1	0
2	1	1	0	1	1
3	1	0	1	1	0
4	0	1	0	1	0
5	1	0	1	0	1
6	0	1	1	1	0

Assumendo un supporto del 50% e una confidenza dell'80%,
(a) Trasformare il dataset in formato transazionale e applicare l'algoritmo Apriori per il calcolo degli itemset frequenti mostrando le varie iterazioni
(b) Calcolare le regole associative a partire dagli itemset frequenti calcolati nel punto (a).
(c) Calcolare gli itemset frequenti multidimensionali sul dataset originario.
(d) Costruire la matrice di contingenza della regola con confidenza più alta e calcolare il lift.

SOLUZIONE:

a) Trasformare il dataset in formato transazionale e applicare l'algoritmo Apriori per il calcolo degli itemset frequenti mostrando le varie iterazioni

<1, {A, C, D}>
<2, {A, B, D, E}>
<3, {A, C, D}>
<4, {B, D}>
<5, {A, C, E}>
<6, {B, C, D}>

1-itemset	2-itemset	3-itemset
{A} 66%	~~{A,B} 16%~~	~~{A,C,D} 33%~~
{B} 50%	{A,C} 50%	
{C} 66%	{A,D} 50%	
{D} 83%	~~{B, C} 16%~~	
~~{E} 33%~~	{B, D} 50%	
	{C, D} 50%	

b) Calcolare le regole associative a partire dagli itemset frequenti calcolati nel punto (a).

A → C 76%
C → A 76%
A → D 76%
D → A 60%
B → D 100%
D → B 60%
C → D 76%
D → C 60%

c) Calcolare gli itemset frequenti multidimensionali sul dataset originario

<1, {A/1, B/0, C/1, D/1, E/0}>
<2, {A/1, B/1, C/0, D/1, E/1}>
<3, {A/1, B/0, C/1, D/1, E/0}>
<4, {A/0, B/1, C/0, D/1, E/0}>
<5, {A/1, B/0, C/1, D/0, E/1}>
<6, {A/0, B/1, C/1, D/1, E/0}>

1-itemset	2-itemset	3-itemset
{A/1} 66%	~~{A/1,B/1} 16%~~	~~{A/1, C/1, D/1} 33%~~
{B/1} 50%	{A/1,C/1} 50%	{A/1, B/0, C/1} 50%
{C/1} 66%	{A/1,D/1} 50%	~~{A/1, B/0, D/1} 33%~~
{D/1} 83%	~~{B~~/1~~, C~~/1~~} 16%~~	~~{A/1, B/0, E/0} 33%~~
~~{E/1} 33%~~	{B/1, D/1} 50%	~~{A/1, C/1, D/1} 33%~~
~~{A/0} 33%~~	{C, /1 D/1} 50%	~~{A/1, C/1, E/0} 33%~~
{B/0} 50%	{A/1, B/0} 50%	~~{A/1, D/1, E/0} 33%~~
~~{C/0} 33%~~	~~{A/1, E/0} 33%~~	{C/1, D/1, E/0} 50%
~~{D/0} 16%~~	~~{B/1, C/1} 16%~~	
{E/0} 66%	{B/1, D/1} 50%	
	~~{B/1, E/0} 33%~~	
	{B/0, C/1} 50%	
	~~{B/0, D/1} 33%~~	
	~~{B/0, E/0} 33%~~	
	{C/1, D/1} 50%	
	{C/1, E/0} 50%	
	{D/1, E/0} 66%	

(d) Costruire la matrice di contingenza della regola con confidenza più alta e calcolare il lift.

	B	~~B~~	
D	3	2	5
~~D~~	0	1	1
	3	3	6

LIFT: 6/5 = 1.2 > 1, positivamente correlati

Esercizio 2 - Indici di similarità (8 punti)

Si consideri il seguente insieme di transazioni:

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

- Si calcoli la matrice di similarità utilizzando l'indice di Jaccard.
- Si calcoli la medesima matrice utilizzando l'indice Simple Matching (Nota: il dataset contiene solo 5 items distinti: Beer, Bread, Coke, Diaper, Milk).

SOLUZIONE:

NOTA: nell'esercizio ci si riferisce (implicitamente) alla similarità tra le transazioni.

Jaccard

1	1/4	2/5	2/5	2/4
	1	1/5	2/4	0
		1	3/5	3/4
			1	2/5
				1

Simple Matching

1	2/5	2/5	2/5	3/5
	1	1/5	2/5	0
		1	3/5	4/5
			1	2/5
				1