COMP4433 Data Mining and Data Warehousing

FAQ on Association Rule Mining I

1. A database has four transactions. Let min_sup=60% and min_conf=80%.

TID	Date	Items_bought
100	10/15/99	{K,A,D,B}
200	10/15/99	{D,A,C,E,B}
300	10/19/99	{C,A,B,E}
400	10/22/99	{B,A,D}

a) Find all frequent itemsets using Aprior algorithm.

b) List all of the strong association rules (with support *s* and confidence *c*) matching the following metarule (form), where *X* is a variable representing customers, and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

$$\forall x \in \text{transaction}, \text{buys}(X, item_1) \land \text{buys}(X, item_2) \Rightarrow \text{buys}(X, item_3) \ [s,c]$$

2. Suppose that you are working for a company that provides local courier services. To improve the operations for more speedy delivery, your boss would like to know if there is any association among the originating locations of the parcel to be delivered, the destinations, the type of the parcel and their weight. Given some of the data collected yesterday below, you decide to use the Apriori algorithm to mine the data for interesting association rules.

Parcel ID	Origin	Destination	Type	Weight
1	HK	HK	Parcel	Light
2	Kln	Kln	Letter	Light
3	NT	Kln	Letter	Light
4	HK	HK	Parcel	Heavy
5	Kln	Kln	Parcel	Light
6	NT	NT	Letter	Light
7	HK	HK	Letter	Light
8	Kln	Kln	Parcel	Heavy
9	Kln	Kln	Letter	Light
10	HK	HK	Letter	Light
11	HK	HK	Parcel	Heavy
12	Kln	Kln	Letter	Light
13	HK	HK	Letter	Light
14	Kln	Kln	Parcel	Light
15	HK	NT	Parcel	Heavy
16	NT	Kln	Letter	Light

17	HK	NT	Letter	Light
18	Kln	HK	Parcel	Light
19	HK	NT	Parcel	Heavy
20	HK	HK	Parcel	Light
21	Kln	Kln	Letter	Light
22	Kln	HK	Parcel	Heavy
23	Kln	Kln	Letter	Light
24	Kln	Kln	Letter	Light
25	HK	HK	Parcel	Light

a) By setting the support to 25% and confidence to 50%, use the Apriori algorithm to find all interesting association rules in the above Table.

b) Assume that a user sets the lift ratio to 1.75, which rules you discovered for a) above are still interesting?

3. Given the following web page content database. Let $min\ sup$=60% and $min\ conf$=80%.

URL	Web Page ID	Keywords Found
Jackchan.com	P100	Popstar, Actor, Movie, Hollywood
Nictsz.com	P101	Actor, Popstar, Music, Band, Movie
Faywang.com	P102	Actress, Popstar, Music, Movie, Hollywood
Allantam.com	P103	Actor, Popstar, Music, CEO, Movie
SammyChen.com	P104	Popstar, Actress, Movie, Music

a) Find all frequent 2-itemsets and 3-itemsets using Apriori algorithm.

b) How many scans of the database are required by part (a)?

c) Based on the results of part (a), generate all strong association rules (with sufficient support and confidence) matching the following metarule.
$\forall P \in web\ page$, $contains(P, keyword_1) \wedge contains(P, keyword_2) \Rightarrow contains(P, keyword_3)$

d) Describe how the association rule mining procedure should be modified to discover rules of the following form.
$\forall P \in web\ page$, $contains(P, keyword_1) \wedge$ **NOT contains**$(P, keyword_2) \Rightarrow contains(P, keyword_3)$