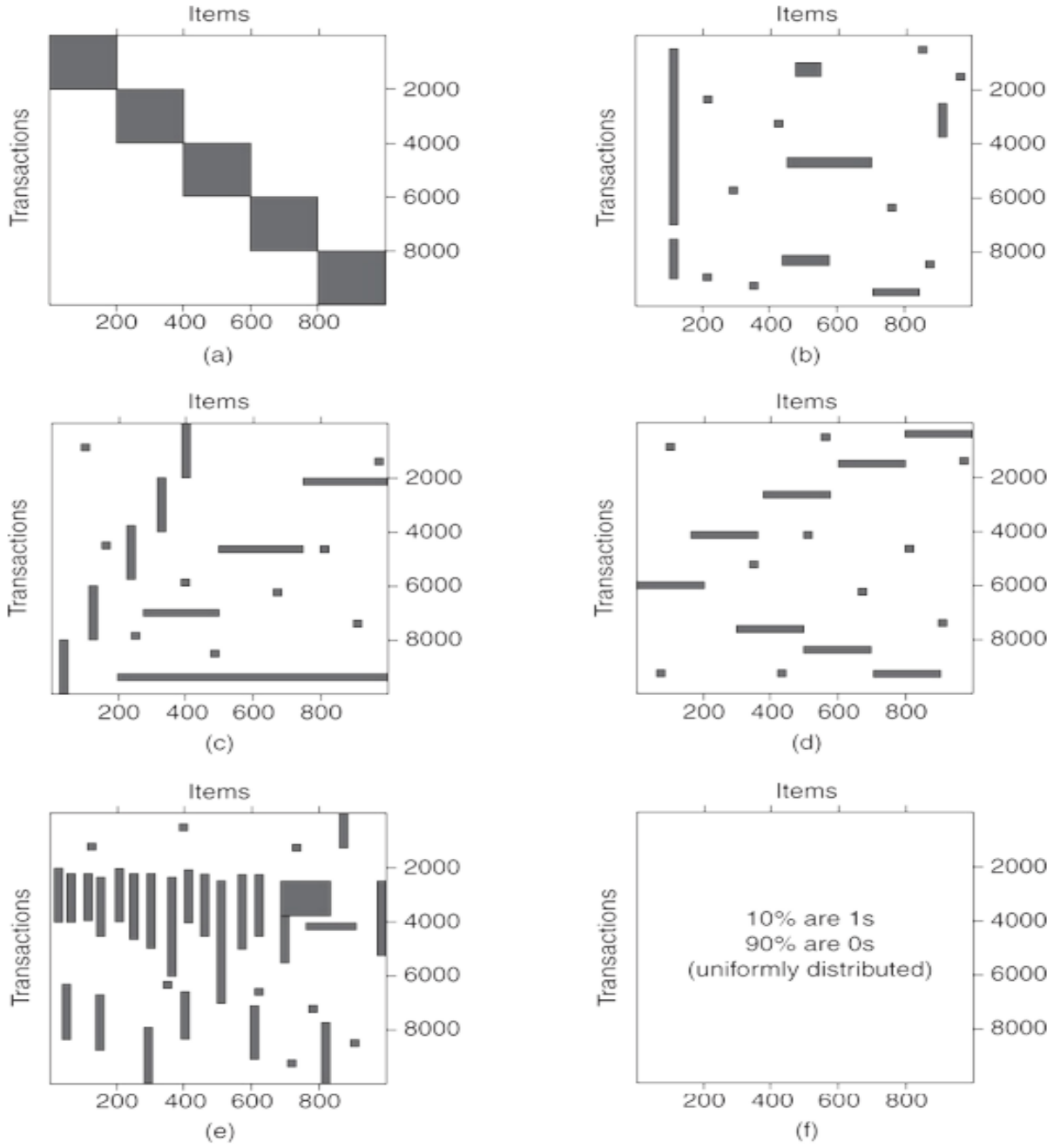
Exercise 1.1

Tan, Ch. 5 (Association Analysis)

15. Answer the following questions using the data sets shown in Figure 5.34. Note that each data set contains 1000 items and 10,000 transactions. Dark cells indicate the presence of items and white cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent itemsets with minsup=10% (i.e., itemsets must be contained in at least 1000 transactions).



a. Which data set(s) will produce the most number of frequent itemsets?
Answer: The data set (e) will produce the most number of frequent itemsets because it has a lot of itemsets with minsup of >=10% and it is the data set which has to produce the longest frequent itemset along with its subsets.

b. Which data set(s) will produce the fewest number of frequent itemsets?
Answer: The data set (d) will produce the fewest number of frequent itemsets because it does not have any itemset with a minsup of >=10%.

c. Which data set(s) will produce the longest frequent itemset?
Answer: The data set (e) will produce the longest frequent itemset because there are a lot of itemsets with a minsup of >=10%.

d. Which data set(s) will produce frequent itemsets with highest maximum support?
Answer: The data set (b) will produce the frequest itemset with highest maximum support because it has an itemset which starts almost at the lower end of the transaction set and goes up to almost at the upper end of the transaction set.

e. Which data set(s) will produce frequent itemsets containing items with wide-varying support levels (i.e., items with mixed support, ranging from less than 20% to more than 70%)?
Answer: The data set (e) will produce frequent itemsets containing items with wide-varying support levels because it has lot of itemset strips which are of various lengths transaction wise than any other data set.

Exercise 1.2

Zaki, Chapter 8 (Frequent Pattern Mining)

Q1. Given the database in Table 8.2.

(a) Using minsup = 3/8, show how the Apriori algorithm enumerates all frequent patterns from this dataset.

tid	itemset
t_1	$ABCD$
t_2	$ACDF$
t_3	$ACDEG$
t_4	$ABDF$
t_5	BCG
t_6	DFG
t_7	ABG
t_8	$CDFG$

Table 8.2: Transaction Database for Q1

Answer:

We have a minsup of 3/8, i.e., a minimum support count of 3.

From the above transaction database, when we find all the possible itemsets along with their respective support, we get the following:

Set	Count
{A}	5
{B}	4
{C}	4
{D}	6
{E}	1
{G}	5
{F}	4
{A,B}	3
{A,C}	3
{A,D}	4
{A,E}	1
{A,F}	2
{A,G}	2
{B,C}	2
{B,D}	2
{B,F}	1
{B,G}	2
{C,D}	4
{C,E}	1
{C,F}	2
{C,G}	3
{D,E}	1
{D,F}	3
{D,G}	3
{E,G}	1
{F,G}	2
{A,B,C}	1
{A,B,D}	2
{A,C,D}	3
{A,C,E}	1
{A,D,E}	1
{A,E,G}	1
{C,D,E}	1
{C,D,G}	2

{D,E,G}	1
{D,F,G}	2

Thus, frequent itemsets with minsup of 3/8 from the dataset are: {A}, {B}, {C}, {D}, {G}, {F}, {A,B}, {A,C}, {A,D}, {C,D}, {C,G}, {D,F}, {D,G}, {A,C,D}.

Q4. Given the database in Table 8.4. Show all rules that one can generate from the set ABE.

tid	itemset
t_1	ACD
t_2	BCE
t_3	$ABCE$
t_4	BDE
t_5	$ABCE$
t_6	$ABCD$

Table 8.4: Dataset for Q4

Answer:

Possible rules with set ABE:

{A} -> {B, E}

{B} -> {A, E}

{E} -> {A, B}

{A,B} -> {E}

{A,E} -> {B}

{B,E} -> {A}

{A} -> {B}

{A} -> {E}

{B} -> {A}

{B} -> {E}

{E} -> {A}

{E} -> {B}

{A} -> {B,E}: Support = 2/6 = 1/3; Confidence = 2/4 = 1/2

{B} -> {A,E}: Support = 1/3; Confidence = 2/5

{E} -> {A,B}: Support = 1/3; Confidence = 2/4 = 1/2

{A,B} -> {E}: Support = 1/3; Confidence = 2/3

{A,E} -> {B}: Support = 1/3; Confidence = 2/2 = 1

{B,E} -> {A}: Support = 1/3; Confidence = 2/4 = /2

{A} -> {B}: Support = 3/6 = 1/2; Confidence = 3/4

{A} -> {E}: Support = 2/6 = 1/3; Confidence = 2/4 = 1/2

{B} -> {A}: Support = 1/2; Confidence = 3/5

{B} -> {E}: Support = 4/6 = 2/3; Confidence = 4/5

{E} -> {A}: Support = 1/3; Confidence = 2/4 = 1/2

{E} -> {B}: Support = 2/3; Confidence = 4/4 = 1