I and all items in J is seldom purchased, whereas a high value tells us that they are purchased together in a large fraction of transactions. Note that the confidence of the association rule $I \to J$ is the conditional probability that a transaction will contain all the items in I and in J given that it contains all the items in I. So, the larger the confidence of $I \to J$, the more likely it is for J to be a subset of a transaction that contains I.

EXAMPLE 15

Extra Examples

What are the support and the confidence of the association rule {cider, donuts} → {apples} for the set of transactions in Example 14?

Solution: The support of this association rule is $\sigma(\{\text{cider, donuts}\} \cup \{\text{apples}\})/|T|$. Because $\sigma(\{\text{cider, donuts}\} \cup \{\text{apples}\}) = \sigma(\{\text{cider, donuts, apples}\}) = 3$ and $|T| = 8$, we see that the support of this rule is $3/8 = 0.375$.

The confidence of this rule is $\sigma(\{\text{cider, donuts}\} \cup \{\text{apples}\})/\sigma(\{\text{cider, donuts}\}) = 3/4 = 0.75$. ◀

An important problem in data mining is to find **strong association rules**, which have support greater than or equal to a minimum support level and confidence greater than or equal to a minimum confidence level. It is important to have efficient algorithms to find strong association rules because the number of available items can be extremely large. For instance, a supermarket may have tens of thousands, or even hundreds of thousands, of items in stock. The brute-force approach of finding association rules with sufficiently large support and confidence by computing the support and confidence of all possible association rules is infeasible because there are an exponential number of such association rules (see Exercise 41). Several widely used algorithms have been developed to solve this problem much more efficiently than brute force. Such algorithms first find frequent itemsets and then turn their attention to finding all the association rules with high confidence from the frequent itemsets that have been found. Consult data mining texts such as [Ag15] for details.

Although we have presented association rules in the context of market baskets, they are useful in a surprisingly wide variety of applications. For instance, association rules can be used to improve medical diagnoses, in which itemsets are collections of test results or symptoms and transactions are the collections of test results and symptoms found on patient records. Association rules, in which itemsets are baskets of key words and transactions are the collections of words on web pages, are used by search engines. Cases of plagiarism can be found using association rules, in which itemsets are collections of sentences and transactions are the contents of documents. Association rules also play helpful roles in various aspects of computer security, including intrusion detection, in which the itemsets are collections of patterns and transactions are the strings transmitted during network attacks. The interested reader will be able to find many more such applications by searching the web.

Exercises

1. List the triples in the relation $\{(a, b, c) \mid a, b,$ and c are integers with $0 < a < b < c < 5\}$.

2. Which 4-tuples are in the relation $\{(a, b, c, d) \mid a, b, c,$ and d are positive integers with $abcd = 6\}$?

3. List the 5-tuples in the relation in Table 8.

4. Assuming that no new n-tuples are added, find all the primary keys for the relations displayed in

a) Table 3. **b)** Table 5.

c) Table 6. **d)** Table 8.

5. Assuming that no new n-tuples are added, find a composite key with two fields containing the *Airline* field for the database in Table 8.

6. Assuming that no new n-tuples are added, find a composite key with two fields containing the Professor field for the database in Table 7.

7. The 3-tuples in a 3-ary relation represent the following attributes of a student database: student ID number, name, phone number.

a) Is student ID number likely to be a primary key?

b) Is name likely to be a primary key?

c) Is phone number likely to be a primary key?

8. The 4-tuples in a 4-ary relation represent these attributes of published books: title, ISBN, publication date, number of pages.

a) What is a likely primary key for this relation?
b) Under what conditions would (title, publication date) be a composite key?
c) Under what conditions would (title, number of pages) be a composite key?

9. The 5-tuples in a 5-ary relation represent these attributes of all people in the United States: name, Social Security number, street address, city, state.

a) Determine a primary key for this relation.
b) Under what conditions would (name, street address) be a composite key?
c) Under what conditions would (name, street address, city) be a composite key?

10. What do you obtain when you apply the selection operator s_C, where C is the condition Room = A100, to the database in Table 7?

11. What do you obtain when you apply the selection operator s_C, where C is the condition Destination = Detroit, to the database in Table 8?

12. What do you obtain when you apply the selection operator s_C, where C is the condition (Project = 2) $\wedge$ (Quantity $\geq$ 50), to the database in Table 10?

13. What do you obtain when you apply the selection operator s_C, where C is the condition (Airline = Nadir) $\vee$ (Destination = Denver), to the database in Table 8?

14. What do you obtain when you apply the projection $P_{2,3,5}$ to the 5-tuple (a, b, c, d, e)?

15. Which projection mapping is used to delete the first, second, and fourth components of a 6-tuple?

16. Display the table produced by applying the projection $P_{1,2,4}$ to Table 8.

17. Display the table produced by applying the projection $P_{1,4}$ to Table 8.

18. How many components are there in the n-tuples in the table obtained by applying the join operator J_3 to two tables with 5-tuples and 8-tuples, respectively?

19. Construct the table obtained by applying the join operator J_2 to the relations in Tables 11 and 12.

20. Show that if C_1 and C_2 are conditions that elements of the n-ary relation R may satisfy, then $s_{C_1 \wedge C_2}(R) = s_{C_1}(s_{C_2}(R))$.

21. Show that if C_1 and C_2 are conditions that elements of the n-ary relation R may satisfy, then $s_{C_1}(s_{C_2}(R)) = s_{C_2}(s_{C_1}(R))$.

22. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R \cup S) = s_C(R) \cup s_C(S)$.

23. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R \cap S) = s_C(R) \cap s_C(S)$.

24. Show that if C is a condition that elements of the n-ary relations R and S may satisfy, then $s_C(R - S) = s_C(R) - s_C(S)$.

25. Show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R \cup S) = P_{i_1,i_2,\ldots,i_m}(R) \cup P_{i_1,i_2,\ldots,i_m}(S)$.

26. Give an example to show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R \cap S)$ may be different from $P_{i_1,i_2,\ldots,i_m}(R) \cap P_{i_1,i_2,\ldots,i_m}(S)$.

27. Give an example to show that if R and S are both n-ary relations, then $P_{i_1,i_2,\ldots,i_m}(R - S)$ may be different from $P_{i_1,i_2,\ldots,i_m}(R) - P_{i_1,i_2,\ldots,i_m}(S)$.

28. a) What are the operations that correspond to the query expressed using this SQL statement?

```
SELECT Supplier
FROM Part_needs
WHERE 1000 ≤ Part_number ≤ 5000
```

b) What is the output of this query given the database in Table 11 as input?

29. a) What are the operations that correspond to the query expressed using this SQL statement?

```
SELECT Supplier, Project
FROM Part_needs, Parts_inventory
WHERE Quantity ≤ 10
```

b) What is the output of this query given the databases in Tables 11 and 12 as input?

30. Determine whether there is a primary key for the relation in Example 2.

31. Determine whether there is a primary key for the relation in Example 3.

32. Show that an n-ary relation with a primary key can be thought of as the graph of a function that maps values of the primary key to $(n - 1)$-tuples formed from values of the other domains.

33. Suppose that the transactions at a convenience store during an evening are {bread, milk, diapers, juice}, {bread, milk, diapers, eggs}, {milk, diapers, beer, eggs}, {bread, beer}, {milk, diapers, eggs, juice}, and {milk, diapers, beer}.

a) Find the count and support of diapers.
b) Find all frequent itemsets if the threshold level is 0.6.

34. Suppose that the key words on eight different web pages are {evolution, primate, Human, Neanderthal, DNA, fossil}, {evolution, Neanderthal, Denisovan, Human, DNA}, {cave, fossil, primate}, {Human, Neanderthal, Denisovan, evolution}, {DNA, genome, evolution, fossil}, {DNA, Human, Neanderthal, Denisovan, genome}, {evolution, primate, cave, fossil}, and {Human, Neanderthal, genome}.

a) Find the count and support of Neanderthal.
b) Find all frequent itemsets if the threshold level is 0.6.

35. Find the support and confidence of the association rule {beer} → {diapers} for the set of transactions in Exercise 33. (This association rule has played an important role in the development of the subject.)

36. Find the support and confidence of the association rule {human, DNA} → {Neanderthal} for the set of transactions in Exercise 34.

TABLE 11 **Part_needs.**

Supplier	*Part_number*	*Project*
23	1092	1
23	1101	3
23	9048	4
31	4975	3
31	3477	2
32	6984	4
32	9191	2
33	1001	1

TABLE 12 **Parts_inventory.**

Part_number	*Project*	*Quantity*	*Color_code*
1001	1	14	8
1092	1	2	2
1101	3	1	1
3477	2	25	2
4975	3	6	2
6984	4	10	1
9048	4	12	2
9191	2	80	4

37. Suppose that I is an itemset with positive count in a set of transactions. Find the confidence of the association rule $I \to \emptyset$.

38. Suppose that I, J, and K are itemsets. Show that the six association rules $\{I, J\} \to K$, $\{J, K\} \to I$, $\{I, K\} \to J$, $I \to \{J, K\}$, $J \to \{I, K\}$, and $K \to \{I, J\}$ all have the same support.

39. The **lift** of the association rule $I \to J$, where I and J are itemsets with positive support in a set of transactions, equals support($I \cup J$)/(support(I)support(J)).

a) Show that the lift of $I \to J$, when support(I) and support(J) are both positive, equals 1 if and only if the occurrence of I in a transaction and the occurrence of J in a transaction are independent events.

b) Find the lift of the association rule $\{\text{beer}\} \to \{\text{diapers}\}$ for the set of transactions in Exercises 33.

c) Find the lift of the association rule $\{\text{evolution}\} \to \{\text{Neanderthals, Denisovans}\}$ for the set of transactions in Exercise 34.

40. Show that if an itemset is frequent in a set of transactions, then all its subsets are also frequent itemsets in this set of transactions.

41. Given n unique items, show that there are 3^n possible association rules of the form $I \to J$, where I and J are disjoint subsets of the set of all items. Be sure to allow the association rules where I or J, or both, are empty.

9.3 Representing Relations

9.3.1 Introduction

In this section, and in the remainder of this chapter, all relations we study will be binary relations. Because of this, in this section and in the rest of this chapter, the word relation will always refer to a binary relation. There are many ways to represent a relation between finite sets. As we have seen in Section 9.1, one way is to list its ordered pairs. Another way to represent a relation is to use a table, as we did in Example 3 in Section 9.1. In this section we will discuss two alternative methods for representing relations. One method uses zero–one matrices. The other method uses pictorial representations called directed graphs, which we will discuss later in this section.

Generally, matrices are appropriate for the representation of relations in computer programs. On the other hand, people often find the representation of relations using directed graphs useful for understanding the properties of these relations.

9.3.2 Representing Relations Using Matrices

A relation between finite sets can be represented using a zero–one matrix. Suppose that R is a relation from $A = \{a_1, a_2, \ldots, a_m\}$ to $B = \{b_1, b_2, \ldots, b_n\}$. (Here the elements of the sets A and B have been listed in a particular, but arbitrary, order. Furthermore, when $A = B$ we use the same ordering for A and B.) The relation R can be represented by the matrix $\mathbf{M}_R = [m_{ij}]$, where

$$m_{ij} = \begin{cases} 1 \text{ if } (a_i, b_j) \in R, \\ 0 \text{ if } (a_i, b_j) \notin R. \end{cases}$$