Contents

6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods 3
   6.1 Basic Concepts .................................................. 4
      6.1.1 Market Basket Analysis: A Motivating Example ...... 4
      6.1.2 Frequent Itemsets, Closed Itemsets, and Association Rules 6
   6.2 Frequent Itemset Mining Methods ................................. 8
      6.2.1 The Apriori Algorithm: Finding Frequent Itemsets by
            Confined Candidate Generation .............................. 9
      6.2.2 Generating Association Rules from Frequent Itemsets .. 13
      6.2.3 Improving the Efficiency of Apriori ....................... 14
      6.2.4 A Pattern-Growth Approach for Mining Frequent Itemsets 17
      6.2.5 Mining Frequent Itemsets Using Vertical Data Format .. 19
      6.2.6 Mining Closed and Max Patterns .......................... 22
   6.3 Which Patterns Are Interesting?—Pattern Evaluation Methods . 25
      6.3.1 Strong Rules Are Not Necessarily Interesting ........... 25
      6.3.2 From Association Analysis to Correlation Analysis ...... 26
      6.3.3 A Comparison of Pattern Evaluation Measures .......... 28
   6.4 Summary ............................................................ 32
   6.5 Exercises ........................................................... 33
   6.6 Bibliographic Notes ............................................... 37
Chapter 6

Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods

Imagine that you are a sales manager in AllElectronics, and you are talking to a customer who recently bought a PC and a digital camera from the store. What should you recommend to her next? Information about which products are frequently purchased by your customers following their purchases of a PC and a digital camera in sequence would be very helpful in making your recommendation. Frequent patterns and association rules are the knowledge that you want to mine in such a scenario.

Frequent patterns are patterns (such as itemsets, subsequences, or substructures) that appear in a data set frequently. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set is a frequent itemset. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as subgraphs, subtrees, or sublattices, which may be combined with itemsets or subsequences. If a substructure occurs frequently, it is called a (frequent) structured pattern. Finding such frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

In this chapter, we introduce the basic concepts of frequent patterns, associations, and correlations (Section 6.1) and study how they can be mined efficiently (Section 6.2). We also discuss how to judge whether the patterns
found are interesting (Section 6.3). In Chapter 7, we extend our discussion to advanced methods of frequent pattern mining, which mine more complex forms of frequent patterns and consider user preferences or constraints to speed up the mining process.

### 6.1 Basic Concepts

Frequent pattern mining searches for recurring relationships in a given data set. This section introduces the basic concepts of frequent pattern mining for the discovery of interesting associations and correlations between itemsets in transactional and relational databases. We begin in Section 6.1.1 by presenting an example of market basket analysis, the earliest form of frequent pattern mining for association rules. The basic concepts of mining frequent patterns and associations are given in Section 6.1.2.

#### 6.1.1 Market Basket Analysis: A Motivating Example

Frequent itemset mining leads to the discovery of associations and correlations among items in large transactional or relational data sets. With massive amounts of data continuously being collected and stored, many industries are becoming interested in mining such patterns from their databases. The discovery of interesting correlation relationships among huge amounts of business transaction records can help in many business decision-making processes, such as catalog design, cross-marketing, and customer shopping behavior analysis.

A typical example of frequent itemset mining is **market basket analysis**. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets” (Figure 6.1). The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. For instance, if customers are buying milk, how likely are they to also buy bread (and what kind of bread) on the same trip to the supermarket?

![Figure 6.1: Market basket analysis.](image_url)
6.1. BASIC CONCEPTS

Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.

Let’s look at an example of how market basket analysis can be useful.

Example 6.1 Market basket analysis. Suppose, as manager of an AllElectronics branch, you would like to learn more about the buying habits of your customers. Specifically, you wonder, “Which groups or sets of items are customers likely to purchase on a given trip to the store?” To answer your question, market basket analysis may be performed on the retail data of customer transactions at your store. You can then use the results to plan marketing or advertising strategies, or in the design of a new catalog. For instance, market basket analysis may help you design different store layouts. In one strategy, items that are frequently purchased together can be placed in proximity in order to further encourage the sale of such items together. If customers who purchase computers also tend to buy antivirus software at the same time, then placing the hardware display close to the software display may help increase the sales of both items. In an alternative strategy, placing hardware and software at opposite ends of the store may entice customers who purchase such items to pick up other items along the way. For instance, after deciding on an expensive computer, a customer may observe security systems for sale while heading toward the software display to purchase antivirus software and may decide to purchase a home security system as well. Market basket analysis can also help retailers plan which items to put on sale at reduced prices. If customers tend to purchase computers and printers together, then having a sale on printers may encourage the sale of printers as well as computers.

If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of values assigned to these variables. The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently associated or purchased together. These patterns can be represented in the form of association rules. For example, the information that customers who purchase computers also tend to buy antivirus software at the same time is represented in the association rule below:

\[ \text{computer} \Rightarrow \text{antivirus software} \quad \text{[support} = 2\%, \text{confidence} = 60\%] \quad (6.1) \]

Rule support and confidence are two measures of rule interestingness. They respectively reflect the usefulness and certainty of discovered rules. A support of 2% for Association Rule (6.1) means that 2% of all the transactions under analysis show that computer and antivirus software are purchased together. A confidence of 60% means that 60% of the customers who purchased a computer also bought the software. Typically, association rules are considered interesting if they satisfy both a minimum support threshold and a minimum confidence threshold. Such thresholds can be set by users or domain experts. Additional analysis can be performed to discover interesting statistical correlations between associated items.
6.1.2 Frequent Itemsets, Closed Itemsets, and Association Rules

Let \( I = \{I_1, I_2, \ldots, I_m\} \) be a set of items. Let \( D \), the task-relevant data, be a set of database transactions where each transaction \( T \) is a nonempty set of items such that \( T \subseteq I \). Each transaction is associated with an identifier, called TID. Let \( A \) be a set of items. A transaction \( T \) is said to contain \( A \) if \( A \subseteq T \). An association rule is an implication of the form \( A \Rightarrow B \), where \( A \subset I \), \( B \subset I \), \( A \neq \emptyset \), \( B \neq \emptyset \), and \( A \cap B = \emptyset \). The rule \( A \Rightarrow B \) holds in the transaction set \( D \) with support \( s \), where \( s \) is the percentage of transactions in \( D \) that contain \( A \cup B \) (i.e., the union of sets \( A \) and \( B \), or say, both \( A \) and \( B \)). This is taken to be the probability, \( P(A \cup B) \).\(^1\) The rule \( A \Rightarrow B \) has confidence \( c \) in the transaction set \( D \), where \( c \) is the percentage of transactions in \( D \) containing \( A \) that also contain \( B \). This is taken to be the conditional probability, \( P(B|A) \). That is,

\[
\begin{align*}
support(A \Rightarrow B) &= P(A \cup B) \quad \text{(6.2)} \\
confidence(A \Rightarrow B) &= P(B|A). \quad \text{(6.3)}
\end{align*}
\]

Rules that satisfy both a minimum support threshold (\( \text{min\_sup} \)) and a minimum confidence threshold (\( \text{min\_conf} \)) are called strong. By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

A set of items is referred to as an itemset.\(^2\) An itemset that contains \( k \) items is a \( k \)-itemset. The set \{computer, antivirus\_software\} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset. Note that the itemset support defined in Equation (6.2) is sometimes referred to as relative support, whereas the occurrence frequency is called the absolute support. If the relative support of an itemset \( I \) satisfies a prespecified minimum support threshold (i.e., the absolute support of \( I \) satisfies the corresponding minimum support count threshold), then \( I \) is a frequent itemset.\(^3\) The set of frequent \( k \)-itemsets is commonly denoted by \( L_k \).\(^4\)

---

\(^{1}\)Notice that the notation \( P(A \cup B) \) indicates the probability that a transaction contains the union of set \( A \) and set \( B \) (i.e., it contains every item in \( A \) and in \( B \)). This should not be confused with \( P(A \text{ or } B) \), which indicates the probability that a transaction contains either \( A \) or \( B \).

\(^{2}\)In the data mining research literature, “itemset” is more commonly used than “item set.”

\(^{3}\)In early work, itemsets satisfying minimum support were referred to as large. This term, however, is somewhat confusing as it has connotations to the number of items in an itemset rather than the frequency of occurrence of the set. Hence, we use the more recent term frequent.

\(^{4}\)Although the term frequent is preferred over large, for historical reasons frequent \( k \)-itemsets are still denoted as \( L_k \).
6.1. BASIC CONCEPTS

From Equation (6.3), we have

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} = \frac{\text{support}_{\text{count}}(A \cup B)}{\text{support}_{\text{count}}(A)}.$$  \hfill (6.4)

Equation (6.4) shows that the confidence of rule $A \Rightarrow B$ can be easily derived from the support counts of $A$ and $A \cup B$. That is, once the support counts of $A$, $B$, and $A \cup B$ are found, it is straightforward to derive the corresponding association rules $A \Rightarrow B$ and $B \Rightarrow A$ and check whether they are strong. Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

1. **Find all frequent itemsets:** By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count, $\text{min}_\text{sup}$.

2. **Generate strong association rules from the frequent itemsets:** By definition, these rules must satisfy minimum support and minimum confidence.

Additional interestingness measures can be applied for the discovery of correlation relationships between associated items, as will be discussed in Section 6.3. Because the second step is much less costly than the first, the overall performance of mining association rules is determined by the first step.

A major challenge in mining frequent itemsets from a large data set is the fact that such mining often generates a huge number of itemsets satisfying the minimum support ($\text{min}_\text{sup}$) threshold, especially when $\text{min}_\text{sup}$ is set low. This is because if an itemset is frequent, each of its subsets is frequent as well. A long itemset will contain a combinatorial number of shorter, frequent sub-itemsets. For example, a frequent itemset of length 100, such as $\{a_1, a_2, \ldots, a_{100}\}$, contains $\binom{100}{1} = 100$ frequent 1-itemsets: $\{a_1\}, \{a_2\}, \ldots, \{a_{100}\}$, $\binom{100}{2} = 4950$ frequent 2-itemsets: $\{a_1, a_2\}, \{a_1, a_3\}, \ldots, \{a_{99}, a_{100}\}$, and so on. The total number of frequent itemsets that it contains is thus:

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}. \hfill (6.5)$$

This is too huge a number of itemsets for any computer to compute or store. To overcome this difficulty, we introduce the concepts of closed frequent itemset and maximal frequent itemset.

An itemset $X$ is **closed** in a data set $D$ if there exists no proper super-itemset $Y$ such that $Y$ has the same support count as $X$ in $D$. An itemset $X$ is a **closed frequent itemset** in set $D$ if $X$ is both closed and frequent in $D$. **

---

5$Y$ is a proper super-itemset of $X$ if $X$ is a proper sub-itemset of $Y$, that is, if $X \subset Y$. In other words, every item of $X$ is contained in $Y$ but there is at least one item of $Y$ that is not in $X$. **
An itemset \( X \) is a **maximal frequent itemset** (or **max-itemset**) in a data set \( D \) if \( X \) is frequent, and there exists no super-itemset \( Y \) such that \( X \subset Y \) and \( Y \) is frequent in \( D \).

Let \( C \) be the set of closed frequent itemsets for a data set \( D \) satisfying a minimum support threshold, \( \text{min\_sup} \). Let \( M \) be the set of maximal frequent itemsets for \( D \) satisfying \( \text{min\_sup} \). Suppose that we have the support count of each itemset in \( C \) and \( M \). Notice that \( C \) and its count information can be used to derive the whole set of frequent itemsets. Thus we say that \( C \) contains complete information regarding its corresponding frequent itemsets. On the other hand, \( M \) registers only the support of the maximal itemsets. It usually does not contain the complete support information regarding its corresponding frequent itemsets. We illustrate these concepts with the following example.

**Example 6.2** **Closed and maximal frequent itemsets.** Suppose that a transaction database has only two transactions: \{\langle a_1, a_2, \ldots, a_{100} \rangle; \langle a_1, a_2, \ldots, a_{50} \rangle\}. Let the minimum support count threshold be \( \text{min\_sup} = 1 \). We find two closed frequent itemsets and their support counts, that is, \( C = \{\{a_1, a_2, \ldots, a_{100}\} : 1; \{a_1, a_2, \ldots, a_{50}\} : 2\} \). There is only one maximal frequent itemset: \( M = \{\{a_1, a_2, \ldots, a_{100}\} : 1\} \). Notice that we cannot include \{\{a_1, a_2, \ldots, a_{50}\}\} as a maximal frequent itemset because it has a frequent super-set, \{\{a_1, a_2, \ldots, a_{100}\}\}. Compare this to the above, where we determined that there are \( 2^{100} - 1 \) frequent itemsets, which is too huge a set to be enumerated!

The set of closed frequent itemsets contains complete information regarding the frequent itemsets. For example, from \( C \), we can derive, say, (1) \{\{a_2, a_{45}\} : 2\} since \{\{a_2, a_{45}\}\} is a sub-itemset of the itemset \{\{a_1, a_2, \ldots, a_{50}\} : 2\}; and (2) \{\{a_8, a_{55}\} : 1\} since \{\{a_8, a_{55}\}\} is not a sub-itemset of the previous itemset but of the itemset \{\{a_1, a_2, \ldots, a_{100}\} : 1\}. However, from the maximal frequent itemset, we can only assert that both itemsets (\{\{a_2, a_{45}\}\} and \{\{a_8, a_{55}\}\}) are frequent, but we cannot assert their actual support counts.

### 6.2 Frequent Itemset Mining Methods

In this section, you will learn methods for mining the simplest form of frequent patterns, such as those discussed for market basket analysis in Section 6.1.1. We begin by presenting **Apriori**, the basic algorithm for finding frequent itemsets (Section 6.2.1). In Section 6.2.2, we look at how to generate strong association rules from frequent itemsets. Section 6.2.3 describes several variations to the Apriori algorithm for improved efficiency and scalability. Section 6.2.4 presents pattern-growth methods for mining frequent itemsets that confine the subsequent search space to only the datasets containing the current frequent itemsets. Section 6.2.5 presents methods for mining frequent itemsets that take advantage of vertical data format.
6.2. FREQUENT ITEMSET MINING METHODS

6.2.1 The Apriori Algorithm: Finding Frequent Itemsets by Confined Candidate Generation

Apriori is a seminal algorithm proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties, as we shall see below. Apriori employs an iterative approach known as a level-wise search, where \( k \)-itemsets are used to explore \((k+1)\)-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy minimum support. The resulting set is denoted by \( L_1 \). Next, \( L_1 \) is used to find \( L_2 \), the set of frequent 2-itemsets, which is used to find \( L_3 \), and so on, until no more frequent \( k \)-itemsets can be found. The finding of each \( L_k \) requires one full scan of the database.

To improve the efficiency of the level-wise generation of frequent itemsets, an important property called the Apriori property, presented below, is used to reduce the search space. We will first describe this property, and then show an example illustrating its use.

**Apriori property:** All nonempty subsets of a frequent itemset must also be frequent.

The Apriori property is based on the following observation. By definition, if an itemset \( I \) does not satisfy the minimum support threshold, \( \minsup \), then \( I \) is not frequent; that is, \( P(I) < \minsup \). If an item \( A \) is added to the itemset \( I \), then the resulting itemset \( (I \cup A) \) cannot occur more frequently than \( I \). Therefore, \( I \cup A \) is not frequent either; that is, \( P(I \cup A) < \minsup \).

This property belongs to a special category of properties called antimonotonicity in the sense that if a set cannot pass a test, all of its supersets will fail the same test as well. It is called antimonotonicity because the property is monotonic in the context of failing a test.\(^6\)

“How is the Apriori property used in the algorithm?” To understand this, let us look at how \( L_{k-1} \) is used to find \( L_k \) for \( k \geq 2 \). A two-step process is followed, consisting of join and prune actions.

1. **The join step:** To find \( L_k \), a set of candidate \( k \)-itemsets is generated by joining \( L_{k-1} \) with itself. This set of candidates is denoted \( C_k \). Let \( l_1 \) and \( l_2 \) be itemsets in \( L_{k-1} \). The notation \( l_i[j] \) refers to the \( j \)th item in \( l_i \) (e.g., \( l_1[k-2] \) refers to the second to the last item in \( l_1 \)). For efficient implementation, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. For the \((k-1)\)-itemset, \( l_i \), this means that the items are sorted such that \( l_i[1] < l_i[2] < \ldots < l_i[k-1] \). The join, \( L_{k-1} \Join L_{k-1} \), is performed, where members of \( L_{k-1} \) are joinable if their first \((k-2)\) items are in common. That is, members \( l_1 \) and \( l_2 \) of \( L_{k-1} \) are joined if \( (l_1[1] = l_2[1]) \land (l_1[2] = l_2[2]) \land \ldots \land (l_1[k-2] = l_2[k-2]) \land (l_1[k-1] < l_2[k-1]) \). The condition \( l_1[k-1] < l_2[k-1] \) simply ensures

\(^6\)The Apriori property has many applications. For example, it can also be used to prune search during data cube computation (Chapter 5).
that no duplicates are generated. The resulting itemset formed by joining
$l_1$ and $l_2$ is \{l_1[1], l_1[2], ..., l_1[k-2], l_1[k-1], l_2[k-1]\}.

2. **The prune step**: $C_k$ is a superset of $L_k$, that is, its members may or may
not be frequent, but all of the frequent $k$-itemsets are included in $C_k$. A
scan of the database to determine the count of each candidate in $C_k$ would
result in the determination of $L_k$ (i.e., all candidates having a count no less
than the minimum support count are frequent by definition, and therefore
belong to $L_k$). $C_k$, however, can be huge, and so this could involve heavy
computation. To reduce the size of $C_k$, the Apriori property is used as follows.
Any $(k-1)$-itemset that is not frequent cannot be a subset of a frequent
$k$-itemset. Hence, if any $(k-1)$-subset of a candidate $k$-itemset is not in $L_{k-1}$,
then the candidate cannot be frequent either and so can be removed from $C_k$.
This **subset testing** can be done quickly by maintaining a hash tree of all
frequent itemsets.

Example 6.3 **Apriori**. Let’s look at a concrete example, based on the *AllElectronics*
transaction database, $D$, of Table 6.1. There are nine transactions in this database,
that is, $|D| = 9$. We use Figure 6.2 to illustrate the Apriori algorithm for finding
frequent itemsets in $D$.

1. In the first iteration of the algorithm, each item is a member of the set of
candidate 1-itemsets, $C_1$. The algorithm simply scans all of the transac-
tions in order to count the number of occurrences of each item.

2. Suppose that the minimum support count required is 2, that is, $min\_sup = 2$.
(Here, we are referring to *absolute* support because we are using a
support count. The corresponding relative support is $2/9 = 22\%$). The
set of frequent 1-itemsets, $L_1$, can then be determined. It consists of the
candidate 1-itemsets satisfying minimum support. In our example, all of
the candidates in $C_1$ satisfy minimum support.

3. To discover the set of frequent 2-itemsets, $L_2$, the algorithm uses the join
6.2. FREQUENT ITEMSET MINING METHODS

![Diagram showing the generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.](image)

Figure 6.2: Generation of candidate itemsets and frequent itemsets, where the minimum support count is 2.

$L_1 \times L_1$ to generate a candidate set of 2-itemsets, $C_2^7$. $C_2$ consists of $(L_1^2) - 2$-itemsets. Note that no candidates are removed from $C_2$ during the prune step because each subset of the candidates is also frequent.

4. Next, the transactions in $D$ are scanned and the support count of each candidate itemset in $C_2$ is accumulated, as shown in the middle table of the second row in Figure 6.2.

5. The set of frequent 2-itemsets, $L_2$, is then determined, consisting of those candidate 2-itemsets in $C_2$ having minimum support.

6. The generation of the set of candidate 3-itemsets, $C_3$, is detailed in Figure 6.3. From the join step, we first get $C_3 = L_2 \times L_2 = \{ \{11, 12, 13\}, \{11, 12, 15\}, \{11, 13, 15\}, \{12, 13, 14\}, \{12, 13, 15\}, \{12, 14, 15\} \}$. Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that the four latter candidates are impossible to be frequent. We therefore remove them from $C_3$, thereby saving the effort of unnecessarily obtaining their counts during the subsequent scan of $D$ to determine $L_3$.

---

7$L_1 \times L_1$ is equivalent to $L_1 \times L_1$, since the definition of $L_k \times L_k$ requires the two joining itemsets to share $k - 1 = 0$ items.
procedure generates the candidates and then uses the Apriori property. Figure 6.3: Generation and pruning of candidate 3-itemsets, $C_3$, from $L_2$ using the Apriori property.

$L_3$. Note that when given a candidate $k$-itemset, we only need to check if its $(k-1)$-subsets are frequent since the Apriori algorithm uses a level-wise search strategy. The resulting pruned version of $C_3$ is shown in the first table of the bottom row of Figure 6.2.

7. The transactions in $D$ are scanned in order to determine $L_3$, consisting of those candidate 3-itemsets in $C_3$ having minimum support (Figure 6.2).

8. The algorithm uses $L_3 \times L_3$ to generate a candidate set of 4-itemsets, $C_4$. Although the join results in $\{\{11, 12, 13\}\}$, itemset $\{11, 12, 13, 15\}$ is pruned because its subset $\{12, 13, 15\}$ is not frequent. Thus, $C_4 = \phi$, and the algorithm terminates, having found all of the frequent itemsets.

Figure 6.4 shows pseudo-code for the Apriori algorithm and its related procedures. Step 1 of Apriori finds the frequent 1-itemsets, $L_1$. In steps 2 to 10, $L_{k-1}$ is used to generate candidates $C_k$ in order to find $L_k$ for $k \geq 2$. The apriori_gen procedure generates the candidates and then uses the Apriori property to eliminate those having a subset that is not frequent (step 3). This procedure is described below. Once all of the candidates have been generated, the database is scanned (step 4). For each transaction, a subset function is used to find all subsets of the transaction that are candidates (step 5), and the count for each of these candidates is accumulated (steps 6 and 7). Finally, all of those candidates satisfying minimum support (step 9) form the set of frequent itemsets, $L$ (step 11). A procedure can then be called to generate association rules from the frequent itemsets. Such a procedure is described in Section 6.2.2.

Input:
- \( D \), a database of transactions;
- \( \text{min} \sup \), the minimum support count threshold.

Output: \( L \), frequent itemsets in \( D \).

Method:
1. \( L_1 = \text{find frequent } 1\text{-itemsets}(D) \);
2. for \( (k = 2; L_{k-1} \neq \phi; k++) \) {
   3. \( C_k = \text{apriori} \text{gen}(L_{k-1}) \);
   4. for each transaction \( t \in D \) {
      5. // scan \( D \) for counts
      6. \( C_t = \text{subset}(C_k, t) \); // get the subsets of \( t \) that are candidates
      7. for each candidate \( c \in C_t \) {
         8. \( c.\text{count}++ \);
      }
   } // scan \( D \) for counts
   9. \( L_k = \{ c \in C_k | c.\text{count} \geq \text{min} \sup \} \);
10. } // scan \( D \) for counts
11. return \( L = \cup_k L_k \);

procedure \text{apriori} \text{gen}(L_{k-1}; \text{frequent } (k-1)\text{-itemsets}) {
1. for each itemset \( l_1 \in L_{k-1} \)
2. for each itemset \( l_2 \in L_{k-1} \)
3. if \( (l_1[1] = l_2[1]) \land (l_1[2] = l_2[2]) \land ... \land (l_1[k-2] = l_2[k-2]) \land (l_1[k-1] < l_2[k-1]) \) then {
4. \( c = l_1 \bowtie l_2; \) // join step: generate candidates
5. if \( \text{has\infrequent\subsetset}(c, L_{k-1}) \) then
6. delete \( c \); // prune step: remove unfruitful candidate
7. else add \( c \) to \( C_k \);
8. }
9. return \( C_k \);
10. }

procedure \text{has\infrequent\subsetset}(c: \text{candidate } k\text{-itemset};
11. \( L_{k-1}; \text{frequent } (k-1)\text{-itemsets}); // use prior knowledge
12. for each \( (k-1)\text{-subset } s \) of \( c \)
13. if \( s \notin L_{k-1} \) then
14. return TRUE;
15. return FALSE;
16. }

Figure 6.4: The Apriori algorithm for discovering frequent itemsets for mining Boolean association rules.

The \text{apriori} \text{gen} procedure performs two kinds of actions, namely, join and prune, as described above. In the join component, \( L_{k-1} \) is joined with \( L_{k-1} \) to generate potential candidates (steps 1 to 4). The prune component (steps 5 to 7) employs the Apriori property to remove candidates that have a subset that is not frequent. The test for infrequent subsets is shown in procedure \text{has\infrequent\subsetset}.

6.2.2 Generating Association Rules from Frequent Itemsets

Once the frequent itemsets from transactions in a database \( D \) have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence). This can be done using Equation (6.4) for confidence, which we show
again here for completeness:

\[
\text{confidence}(A \Rightarrow B) = \frac{\text{support}(A \cup B)}{\text{support}(A)}.
\]

The conditional probability is expressed in terms of itemset support count, where \(\text{support}(A \cup B)\) is the number of transactions containing the itemsets \(A \cup B\), and \(\text{support}(A)\) is the number of transactions containing the itemset \(A\). Based on this equation, association rules can be generated as follows:

- For each frequent itemset \(l\), generate all nonempty subsets of \(l\).
- For every nonempty subset \(s\) of \(l\), output the rule “\(s \Rightarrow (l - s)\)” if \(\frac{\text{support}(l)}{\text{support}(s)} \geq \text{min}_\text{conf}\), where \(\text{min}_\text{conf}\) is the minimum confidence threshold.

Because the rules are generated from frequent itemsets, each one automatically satisfies minimum support. Frequent itemsets can be stored ahead of time in hash tables along with their counts so that they can be accessed quickly.

**Example 6.4 Generating association rules.** Let’s try an example based on the transactional data for *AllElectronics* shown in Table 6.1. The data contain frequent itemset \(X = \{I1, I2, I5\}\). What are the association rules that can be generated from \(X\)? The nonempty subsets of \(X\) are \(\{I1, I2\}\), \(\{I1, I5\}\), \(\{I2, I5\}\), \(\{I1\}\), \(\{I2\}\), and \(\{I5\}\). The resulting association rules are as shown below, each listed with its confidence:

- \(\{I1, I2\} \Rightarrow I5\), \(\text{confidence} = \frac{2}{4} = 50\%\)
- \(\{I1, I5\} \Rightarrow I2\), \(\text{confidence} = \frac{2}{2} = 100\%\)
- \(\{I2, I5\} \Rightarrow I1\), \(\text{confidence} = \frac{2}{2} = 100\%\)
- \(I1 \Rightarrow \{I2, I5\}\), \(\text{confidence} = \frac{2}{6} = 33\%\)
- \(I2 \Rightarrow \{I1, I5\}\), \(\text{confidence} = \frac{2}{7} = 29\%\)
- \(I5 \Rightarrow \{I1, I2\}\), \(\text{confidence} = \frac{2}{2} = 100\%\)

If the minimum confidence threshold is, say, 70\%, then only the second, third, and last rules above are output, because these are the only ones generated that are strong. Note that, unlike conventional classification rules, association rules can contain more than one conjunct in the right-hand side of the rule.

### 6.2.3 Improving the Efficiency of Apriori

“How can we further improve the efficiency of Apriori-based mining?” Many variations of the Apriori algorithm have been proposed that focus on improving the efficiency of the original algorithm. Several of these variations are summarized as follows:

**Hash-based technique** (hashing itemsets into corresponding buckets): A hash-based technique can be used to reduce the size of the candidate \(k\)-itemsets, \(C_k\), for \(k > 1\). For example, when scanning each transaction in
6.2. FREQUENT ITEMSET MINING METHODS

Create hash table $H_2$ using hash function $h(x, y) = ((\text{order of } x) \times 10 + (\text{order of } y)) \mod 7$

<table>
<thead>
<tr>
<th>bucket address</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>bucket count</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>bucket contents</td>
<td>{(11, 14), (11, 15), (12, 14), (12, 15)}</td>
<td>{(11, 12), (11, 13)}</td>
<td>{(11, 12), (11, 13)}</td>
<td>{(11, 12), (11, 13)}</td>
<td>{(11, 12), (11, 13)}</td>
<td>{(11, 12), (11, 13)}</td>
<td>{(11, 12), (11, 13)}</td>
</tr>
</tbody>
</table>

Figure 6.5: Hash table, $H_2$, for candidate 2-itemsets: This hash table was generated by scanning the transactions of Table 6.1 while determining $L_1$. If the minimum support count is, say, 3, then the itemsets in buckets 0, 1, 3, and 4 cannot be frequent and so they should not be included in $C_2$.

the database to generate the frequent 1-itemsets, $L_1$, we can generate all of the 2-itemsets for each transaction, hash (i.e., map) them into the different buckets of a hash table structure, and increase the corresponding bucket counts (Figure 6.5). A 2-itemset whose corresponding bucket count in the hash table is below the support threshold cannot be frequent and thus should be removed from the candidate set. Such a hash-based technique may substantially reduce the number of the candidate $k$-itemsets examined (especially when $k = 2$).

Transaction reduction (reducing the number of transactions scanned in future iterations): A transaction that does not contain any frequent $k$-itemsets cannot contain any frequent $(k + 1)$-itemsets. Therefore, such a transaction can be marked or removed from further consideration because subsequent scans of the database for $j$-itemsets, where $j > k$, will not need to consider such a transaction.

Partitioning (partitioning the data to find candidate itemsets): A partitioning technique can be used that requires just two database scans to mine the frequent itemsets (Figure 6.6). It consists of two phases. In Phase I, the algorithm divides the transactions of $D$ into $n$ nonoverlapping partitions. If the minimum relative support threshold for transactions in $D$ is $\text{min\-sup}$, then the minimum support count for a partition is $\text{min\-sup} \times \text{the number of transactions in that partition}$. For each partition, all the local frequent itemsets, i.e., the itemsets frequent within the partition, are found.

A local frequent itemset may or may not be frequent with respect to the entire database, $D$. However, any itemset that is potentially frequent with respect to $D$ must occur as a frequent itemset in at least one of the partitions\(^8\). Therefore, all local frequent itemsets are candidate itemsets with respect to $D$. The collection of frequent itemsets from all partitions forms the global candidate itemsets with respect to $D$. In Phase II, a second scan of $D$ is conducted in which the actual support of each candidate

\(^8\)The proof of this property is left as an exercise (see Exercise 6.3(d)).
is assessed in order to determine the global frequent itemsets. Partition size and the number of partitions are set so that each partition can fit into main memory and therefore be read only once in each phase.

**Sampling** (mining on a subset of the given data): The basic idea of the sampling approach is to pick a random sample \( S \) of the given data \( D \), and then search for frequent itemsets in \( S \) instead of \( D \). In this way, we trade off some degree of accuracy against efficiency. The sample size of \( S \) is such that the search for frequent itemsets in \( S \) can be done in main memory, and so only one scan of the transactions in \( S \) is required overall. Because we are searching for frequent itemsets in \( S \) rather than in \( D \), it is possible that we will miss some of the global frequent itemsets. To reduce this possibility, we use a lower support threshold than minimum support to find the frequent itemsets local to \( S \) (denoted \( L^S \)). The rest of the database is then used to compute the actual frequencies of each itemset in \( L^S \). A mechanism is used to determine whether all of the global frequent itemsets are included in \( L^S \). If \( L^S \) actually contains all of the frequent itemsets in \( D \), then only one scan of \( D \) is required. Otherwise, a second pass can be done in order to find the frequent itemsets that were missed in the first pass. The sampling approach is especially beneficial when efficiency is of utmost importance, such as in computationally intensive applications that must be run frequently.

**Dynamic itemset counting** (adding candidate itemsets at different points during a scan): A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points. In this variation, new candidate itemsets can be added at any start point, unlike in Apriori, which determines new candidate itemsets only immediately before each complete database scan. The technique uses the count-so-far as the lower bound of the actual count. If the count-so-far passes the minimum support, the itemset is added into the frequent itemset collection and can be used to generate longer candidates. This leads to fewer database scans than Apriori for finding all the frequent itemsets.

Other variations are discussed in the next chapter.
6.2.4 A Pattern-Growth Approach for Mining Frequent Itemsets

As we have seen, in many cases the Apriori candidate generate-and-test method significantly reduces the size of candidate sets, leading to good performance gain. However, it can suffer from two nontrivial costs:

- **It may still need to generate a huge number of candidate sets.** For example, if there are $10^4$ frequent 1-itemsets, the Apriori algorithm will need to generate more than $10^7$ candidate 2-itemsets.

- **It may need to repeatedly scan the whole database and check a large set of candidates by pattern matching.** It is costly to go over each transaction in the database to determine the support of the candidate itemsets.

"Can we design a method that mines the complete set of frequent itemsets without such a costly candidate generation process?" An interesting method in this attempt is called frequent-pattern growth, or simply FP-growth, which adopts a divide-and-conquer strategy as follows. First, it compresses the database representing frequent items into a frequent-pattern tree, or FP-tree, which retains the itemset association information. It then divides the compressed database into a set of conditional databases (a special kind of projected database), each associated with one frequent item or "pattern fragment," and mines each such database separately. This approach may substantially reduce the sizes of datasets to be searched along with pattern growth. You’ll see how it works with the following example.

**Example 6.5 FP-growth (finding frequent itemsets without candidate generation).**

We re-examine the mining of transaction database, $D$, of Table 6.1 in Example 6.3 using the frequent-pattern growth approach.

The first scan of the database is the same as Apriori, which derives the set of frequent items (1-itemsets) and their support counts (frequencies). Let the minimum support count be 2. The set of frequent items is sorted in the order of descending support count. This resulting set or list is denoted by $L$. Thus, we have $L = \{\{I2: 7\}, \{I1: 6\}, \{I3: 6\}, \{I4: 2\}, \{I5: 2\}\}$.

An FP-tree is then constructed as follows. First, create the root of the tree, labeled with "null." Scan database $D$ a second time. The items in each transaction are processed in $L$ order (i.e., sorted according to descending support count), and a branch is created for each transaction. For example, the scan of the first transaction, “T100: I1, I2, I5," which contains three items (I2, I1, I5 in $L$ order), leads to the construction of the first branch of the tree with three nodes, (I2: 1), (I1:1), and (I5: 1), where I2 is linked as a child to the root, I1 is linked to I2, and I5 is linked to I1. The second transaction, T200, contains the items I2 and I4 in $L$ order, which would result in a branch where I2 is linked to the root and I4 is linked to I2. However, this branch would share a common prefix, I2, with the existing path for T100. Therefore, we instead increment the count of the I2 node by 1, and create a new node, (I4: 1), which is linked as a child to (I2: 2). In general, when considering the branch to be added
for a transaction, the count of each node along a common prefix is incremented by 1, and nodes for the items following the prefix are created and linked accordingly.

To facilitate tree traversal, an item header table is built so that each item points to its occurrences in the tree via a chain of **node-links**. The tree obtained after scanning all of the transactions is shown in Figure 6.7 with the associated node-links. In this way, the problem of mining frequent patterns in databases is transformed to that of mining the FP-tree.

The FP-tree is mined as follows. Start from each frequent length-1 pattern (as an initial **suffix pattern**), construct its **conditional pattern base** (a “sub-database,” which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then construct its (conditional) FP-tree, and perform mining recursively on such a tree. The pattern growth is achieved by the concatenation of the suffix pattern with the frequent patterns generated from a conditional FP-tree.

Mining of the FP-tree is summarized in Table 6.2 and detailed as follows. We first consider I5, which is the last item in \( L \), rather than the first. The reason for starting at the end of the list will become apparent as we explain the FP-tree mining process. I5 occurs in two branches of the FP-tree of Figure 6.7. (The occurrences of I5 can easily be found by following its chain of node-links.) The paths formed by these branches are \( \langle I2, I1, I5: 1 \rangle \) and \( \langle I2, I1, I3, I5: 1 \rangle \). Therefore, considering I5 as a suffix, its corresponding two prefix paths are \( \langle I2, I1: 1 \rangle \) and \( \langle I2, I1, I3: 1 \rangle \), which form its conditional pattern base. Using this conditional pattern base as a transaction database, we build an I5-conditional FP-tree, which contains only a single path, \( \langle I2, I1: 2 \rangle \); I3 is not included because its support count of 1 is less than the minimum support count. The single path generates all the combinations of frequent patterns: \{I2, I5: 2\}, \{I1, I5: 2\}, \{I2, I1, I5: 2\}.

For I4, its two prefix paths form the conditional pattern base, \{\{I2 I1: 1\}, \{I2: 1\}\}, which generates a single-node conditional FP-tree, \( \langle I2: 2 \rangle \), and derives one frequent pattern, \{I2, I4: 2\}. 

Figure 6.7: An FP-tree registers compressed, frequent pattern information.
6.2. FREQUENT ITEMSET MINING METHODS

Table 6.2: Mining the FP-tree by creating conditional (sub-)pattern bases.

<table>
<thead>
<tr>
<th>Item</th>
<th>Conditional Pattern Base</th>
<th>Conditional FP-tree</th>
<th>Frequent Patterns Generated</th>
</tr>
</thead>
<tbody>
<tr>
<td>I5</td>
<td>{I2, I1: 1}, {I2, I1, I3: 1}</td>
<td>(I2: 2, I1: 2)</td>
<td>{I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2}</td>
</tr>
<tr>
<td>I4</td>
<td>{I2, I1: 1}, {I2: 1}</td>
<td>(I2: 2)</td>
<td>{I2, I4: 2}</td>
</tr>
<tr>
<td>I3</td>
<td>{I2, I1: 2}, {I2: 2}, {I1: 2}</td>
<td>(I2: 4, I1: 2), (I1: 2)</td>
<td>{I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2}</td>
</tr>
<tr>
<td>I1</td>
<td>{I2: 4}</td>
<td>(I2: 4)</td>
<td>{I2, I1: 4}</td>
</tr>
</tbody>
</table>

Figure 6.8: The conditional FP-tree associated with the conditional node I3.

Similar to the above analysis, I3’s conditional pattern base is \{\{I2, I1: 2\}, \{I2: 2\}, \{I1: 2\}\}. Its conditional FP-tree has two branches, (I2: 4, I1: 2) and (I1: 2), as shown in Figure 6.8, which generates the set of patterns \{\{I2, I3: 4\}, \{I1, I3: 4\}, \{I2, I1, I3: 2\}\}. Finally, I1’s conditional pattern base is \{\{I2: 4\}\}, whose FP-tree contains only one node, (I2: 4), which generates one frequent pattern, \{I2, I1: 4\}. This mining process is summarized in Figure 6.9.

The FP-growth method transforms the problem of finding long frequent patterns to searching for shorter ones in much smaller conditional databases recursively and then concatenating the suffix. It uses the least frequent items as a suffix, offering good selectivity. The method substantially reduces the search costs.

When the database is large, it is sometimes unrealistic to construct a main memory-based FP-tree. An interesting alternative is to first partition the database into a set of projected databases, and then construct an FP-tree and mine it in each projected database. Such a process can be recursively applied to any projected database if its FP-tree still cannot fit in main memory.

A study on the performance of the FP-growth method shows that it is efficient and scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm.

6.2.5 Mining Frequent Itemsets Using Vertical Data Format

Both the Apriori and FP-growth methods mine frequent patterns from a set of transactions in TID-itemset format (that is, \{TID : itemset\}), where TID is a transaction-id and itemset is the set of items bought in transaction TID.
Algorithm: **FP\_growth**. Mine frequent itemsets using an FP-tree by pattern fragment growth.

**Input:**
- \( D \), a transaction database;
- \( \text{min\_sup} \), the minimum support count threshold.

**Output:** The complete set of frequent patterns.

**Method:**

1. The FP-tree is constructed in the following steps:
   
   (a) Scan the transaction database \( D \) once. Collect \( F \), the set of frequent items, and their support counts. Sort \( F \) in support count descending order as \( L \), the list of frequent items.
   
   (b) Create the root of an FP-tree, and label it as “null.” For each transaction \( \text{Trans} \) in \( D \) do the following.
      
      Select and sort the frequent items in \( \text{Trans} \) according to the order of \( L \). Let the sorted frequent item list in \( \text{Trans} \) be \( [p|P] \), where \( p \) is the first element and \( P \) is the remaining list. Call \( \text{insert\_tree}([p|P], T) \), which is performed as follows. If \( T \) has a child \( N \) such that \( N.item\_name = p.item\_name \), then increment \( N\)’s count by 1; else create a new node \( N \), and let its count be 1, its parent link be linked to \( T \), and its node-link to the nodes with the same \( item\_name \) via the node-link structure. If \( P \) is nonempty, call \( \text{insert\_tree}(P, N) \) recursively.

2. The FP-tree is mined by calling \( \text{FP\_growth}(FP\_tree, \text{null}) \), which is implemented as follows.

   **procedure** \( \text{FP\_growth}(Tree, \alpha) \)
   
   (1) if \( Tree \) contains a single path \( P \) then
   (2) for each combination (denoted as \( \beta \)) of the nodes in the path \( P \)
   (3) generate pattern \( \beta \cup \alpha \) with \( \text{support\_count} = \text{minimum support count of nodes in } \beta \);
   (4) else for each \( a_i \) in the header of \( Tree \) { 
   (5) generate pattern \( \beta = a_i \cup \alpha \) with \( \text{support\_count} = a_i . \text{support\_count} \);
   (6) construct \( \beta \)'s conditional pattern base and then \( \beta \)'s conditional FP-tree \( Tree_\beta \);
   (7) if \( Tree_\beta \neq \emptyset \) then
   (8) call \( \text{FP\_growth}(Tree_\beta, \beta) \); }

Figure 6.9: The FP-growth algorithm for discovering frequent itemsets without candidate generation.
6.2. FREQUENT ITEMSET MINING METHODS

Table 6.3: The vertical data format of the transaction data set \( D \) of Table 6.1.

<table>
<thead>
<tr>
<th>itemset</th>
<th>( TID_{set} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I1</td>
<td>{T100, T400, T500, T700, T800, T900}</td>
</tr>
<tr>
<td>I2</td>
<td>{T100, T200, T300, T400, T600, T800, T900}</td>
</tr>
<tr>
<td>I3</td>
<td>{T300, T500, T600, T700, T800, T900}</td>
</tr>
<tr>
<td>I4</td>
<td>{T200, T400}</td>
</tr>
<tr>
<td>I5</td>
<td>{T100, T800}</td>
</tr>
</tbody>
</table>

This data format is known as the **horizontal data format**. Alternatively, data can also be presented in \( \text{item-}TID_{set} \) format (that is, \{\text{item} : TID_{set}\}), where \text{item} is an item name, and \( TID_{set} \) is the set of transaction identifiers containing the item. This format is known as the **vertical data format**.

In this section, we look at how frequent itemsets can also be mined efficiently using vertical data format, which is the essence of the **ECLAT** (Equivalence CLASS Transformation) algorithm.

Example 6.6 **Mining frequent itemsets using vertical data format.** Consider the horizontal data format of the transaction database, \( D \), of Table 6.1 in Example 6.3. This can be transformed into the vertical data format shown in Table 6.3 by scanning the data set once.

Mining can be performed on this data set by intersecting the \( TID_{set} \)s of every pair of frequent single items. The minimum support count is 2. Because every single item is frequent in Table 6.3, there are 10 intersections performed in total, which lead to 8 nonempty 2-itemsets as shown in Table 6.4. Notice that because the itemsets \{I1, I4\} and \{I3, I5\} each contain only one transaction, they do not belong to the set of frequent 2-itemsets.

Based on the Apriori property, a given 3-itemset is a candidate 3-itemset only if every one of its 2-itemset subsets is frequent. The candidate generation process here will generate only two 3-itemsets: \{I1, I2, I3\} and \{I1, I2, I5\}. By intersecting the \( TID_{set} \)s of any two corresponding 2-itemsets of these candidate

Table 6.4: The 2-itemsets in vertical data format.

<table>
<thead>
<tr>
<th>itemset</th>
<th>( TID_{set} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>{I1, I2}</td>
<td>{T100, T400, T800, T900}</td>
</tr>
<tr>
<td>{I1, I3}</td>
<td>{T500, T700, T800, T900}</td>
</tr>
<tr>
<td>{I1, I4}</td>
<td>{T400}</td>
</tr>
<tr>
<td>{I1, I5}</td>
<td>{T100, T800}</td>
</tr>
<tr>
<td>{I2, I3}</td>
<td>{T300, T600, T800, T900}</td>
</tr>
<tr>
<td>{I2, I4}</td>
<td>{T200, T400}</td>
</tr>
<tr>
<td>{I2, I5}</td>
<td>{T100, T800}</td>
</tr>
<tr>
<td>{I3, I5}</td>
<td>{T800}</td>
</tr>
</tbody>
</table>
3-itemsets, it derives Table 6.5, where there are only two frequent 3-itemsets: 
{I1, I2, I3: 2} and {I1, I2, I5: 2}.

Example 6.6 illustrates the process of mining frequent itemsets by exploring 
the vertical data format. First, we transform the horizontally formatted data 
to the vertical format by scanning the data set once. The support count of an 
itemset is simply the length of the TID set of the itemset. Starting with \( k = 1 \), 
the frequent \( k \)-itemsets can be used to construct the candidate \((k + 1)\)-itemsets 
based on the Apriori property. The computation is done by intersection of the 
TID sets of the frequent \( k \)-itemsets to compute the TID sets of the correspond-
ing \((k + 1)\)-itemsets. This process repeats, with \( k \) incremented by 1 each time, 
until no frequent itemsets or no candidate itemsets can be found.

Besides taking advantage of the Apriori property in the generation of candidate 
\((k + 1)\)-itemset from frequent \( k \)-itemsets, another merit of this method is 
that there is no need to scan the database to find the support of \((k + 1)\) itemsets 
(for \( k \geq 1 \)). This is because the TID set of each \( k \)-itemset carries the complete 
information required for counting such support. However, the TID sets can be 
quite long, taking substantial memory space as well as computation time for 
intersecting the long sets.

To further reduce the cost of registering long TID sets, as well as the subse-
quent costs of intersections, we can use a technique called diffset, which keeps 
track of only the differences of the TID sets of a \((k + 1)\)-itemset and a corre-
sponding \( k \)-itemset. For instance, in Example 6.6 we have \{I1\} = \{T100, T400, 
T500, T700, T800, T900\} and \{I1, I2\} = \{T100, T400, T800, T900\}. The diffset 
between the two is \( \text{diffset} \{\{I1, I2\}, \{I1\}\} = \{T500, T700\} \). Thus, rather than 
recording the four TIDs that make up the intersection of \{I1\} and \{I2\}, we can 
instead use diffset to record just two TIDs indicating the difference between \{I1\} 
and \{I1, I2\}. Experiments show that in certain situations, such as when the 
data set contains many dense and long patterns, this technique can substantially 
reduce the total cost of vertical format mining of frequent itemsets.

6.2.6 Mining Closed and Max Patterns

In Section 6.1.2 we saw how frequent itemset mining may generate a huge num-
ber of frequent itemsets, especially when the \( \text{min} \sup \) threshold is set low or 
when there exist long patterns in the data set. Example 6.2 showed that closed

<table>
<thead>
<tr>
<th>itemset</th>
<th>TID set</th>
</tr>
</thead>
<tbody>
<tr>
<td>{I1, I2, I3}</td>
<td>{T800, T900}</td>
</tr>
<tr>
<td>{I1, I2, I5}</td>
<td>{T100, T800}</td>
</tr>
</tbody>
</table>
frequent itemsets can substantially reduce the number of patterns generated in frequent itemset mining while preserving the complete information regarding the set of frequent itemsets. That is, from the set of closed frequent itemsets, we can easily derive the set of frequent itemsets and their support. Thus in practice, it is more desirable to mine the set of closed frequent itemsets rather than the set of all frequent itemsets in most cases.

“How can we mine closed frequent itemsets?” A naïve approach would be to first mine the complete set of frequent itemsets and then remove every frequent itemset that is a proper subset of, and carries the same support as, an existing frequent itemset. However, this is quite costly. As shown in Example 6.2, this method would have to first derive $2^{100} - 1$ frequent itemsets in order to obtain a length-100 frequent itemset, all before it could begin to eliminate redundant itemsets. This is prohibitively expensive. In fact, there exist only a very small number of closed frequent itemsets in the data set of Example 6.2.

A recommended methodology is to search for closed frequent itemsets directly during the mining process. This requires us to prune the search space as soon as we can identify the case of closed itemsets during mining. Pruning strategies include the following:

**Item merging:** If every transaction containing a frequent itemset $X$ also contains an itemset $Y$ but not any proper superset of $Y$, then $X \cup Y$ forms a frequent closed itemset and there is no need to search for any itemset containing $X$ but no $Y$.

For example, in Table 6.2 of Example 6.5, the projected conditional database for prefix itemset $\{I5:2\}$ is $\{\{I2, I1\}, \{I2, I1, I3\}\}$, from which we can see that each of its transactions contains itemset $\{I2, I1\}$ but no proper superset of $\{I2, I1\}$. Itemset $\{I2, I1\}$ can be merged with $\{I5\}$ to form the closed itemset $\{I5, I2, I1: 2\}$, and we do not need to mine for closed itemsets that contain $I5$ but not $\{I2, I1\}$.

**Sub-itemset pruning:** If a frequent itemset $X$ is a proper subset of an already found frequent closed itemset $Y$ and $\text{support}(X) = \text{support}(Y)$, then $X$ and all of $X$’s descendants in the set enumeration tree cannot be frequent closed itemsets and thus can be pruned.

Similar to Example 6.2, suppose a transaction database has only two transactions: $\{(a_1, a_2, \ldots, a_{100}), (a_1, a_2, \ldots, a_{50})\}$, and the minimum support count is $\text{minsup} = 2$. The projection on the first item, $a_1$, derives the frequent itemset $\{a_1, a_2, \ldots, a_{50}: 2\}$, based on the itemset merging optimization. Because $\text{support}\{\{a_2\}\} = \text{support}\{\{a_1, a_2, \ldots, a_{50}\}\} = 2$, and $\{a_2\}$ is a proper subset of $\{a_1, a_2, \ldots, a_{50}\}$, there is no need to examine $a_2$ and its projected database. Similar pruning can be done for $a_3, \ldots, a_{50}$ as well. Thus the mining of closed frequent itemsets in this data set terminates after mining $a_1$’s projected database.

---

9Remember that $X$ is a closed frequent itemset in a data set $S$ if there exists no proper super-itemset $Y$ such that $Y$ has the same support count as $X$ in $S$, and $X$ satisfies minimum support.
**Item skipping:** In the depth-first mining of closed itemsets, at each level, there will be a prefix itemset $X$ associated with a header table and a projected database. If a local frequent item $p$ has the same support in several header tables at different levels, we can safely prune $p$ from the header tables at higher levels.

Consider, for example, the transaction database above having only two transactions: \{\langle a_1, a_2, \ldots, a_{100} \rangle, \langle a_1, a_2, \ldots, a_{50} \rangle\}, where $\minsup = 2$. Because $a_2$ in $a_1$'s projected database has the same support as $a_2$ in the global header table, $a_2$ can be pruned from the global header table. Similar pruning can be done for $a_3, \ldots, a_{50}$. There is no need to mine anything more after mining $a_1$’s projected database.

Besides pruning the search space in the closed itemset mining process, another important optimization is to perform efficient checking of a newly derived frequent itemset to see whether it is closed, because the mining process cannot ensure that every generated frequent itemset is closed.

When a new frequent itemset is derived, it is necessary to perform two kinds of closure checking: (1) superset checking, which checks if this new frequent itemset is a superset of some already found closed itemsets with the same support, and (2) subset checking, which checks whether the newly found itemset is a subset of an already found closed itemset with the same support.

If we adopt the item merging pruning method under a divide-and-conquer framework, then the superset checking is actually built-in and there is no need to explicitly perform superset checking. This is because if a frequent itemset $X \cup Y$ is found later than itemset $X$, and carries the same support as $X$, it must be in $X$’s projected database and must have been generated during itemset merging.

To assist in subset checking, a compressed pattern-tree can be constructed to maintain the set of closed itemsets mined so far. The pattern-tree is similar in structure to the FP-tree except that all of the closed itemsets found are stored explicitly in the corresponding tree branches. For efficient subset checking, we can use the following property: If the current itemset $S_c$ can be subsumed by another already found closed itemset $S_a$, then (1) $S_c$ and $S_a$ have the same support, (2) the length of $S_c$ is smaller than that of $S_a$, and (3) all of the items in $S_c$ are contained in $S_a$. Based on this property, a two-level hash index structure can be built for fast accessing of the pattern-tree: The first level uses the identifier of the last item in $S_c$ as a hash key (since this identifier must be within the branch of $S_c$), and the second level uses the support of $S_c$ as a hash key (since $S_c$ and $S_a$ have the same support). This will substantially speed up the subset checking process.

The above discussion illustrates methods for efficient mining of closed frequent itemsets. “Can we extend these methods for efficient mining of maximal frequent itemsets?” Because maximal frequent itemsets share many similarities with closed frequent itemsets, many of the optimization techniques developed here can be extended to mining maximal frequent itemsets. However, we leave this method as an exercise for interested readers.
6.3 Which Patterns Are Interesting?—Pattern Evaluation Methods

Most association rule mining algorithms employ a support-confidence framework. Although minimum support and confidence thresholds help weed out or exclude the exploration of a good number of uninteresting rules, many of the rules generated are still not interesting to the users. Unfortunately, this is especially true when mining at low support thresholds or mining for long patterns. This has been a major bottleneck for successful application of association rule mining.

In this section, we first look at how even strong association rules can be uninteresting and misleading (Section 6.3.1). We then discuss how the support-confidence framework can be supplemented with additional interestingness measures based on correlation analysis (Section 6.3.2). Section 6.3.3 presents additional pattern evaluation measures. It then provides an overall comparison of all of the measures discussed here. By the end, you will learn which pattern evaluation measures are most effective for the discovery of only interesting rules.

6.3.1 Strong Rules Are Not Necessarily Interesting

Whether or not a rule is interesting can be assessed either subjectively or objectively. Ultimately, only the user can judge if a given rule is interesting, and this judgment, being subjective, may differ from one user to another. However, objective interestingness measures, based on the statistics “behind” the data, can be used as one step toward the goal of weeding out uninteresting rules from presentation to the user.

“How can we tell which strong association rules are really interesting?” Let’s examine the following example.

Example 6.7 A misleading “strong” association rule. Suppose we are interested in analyzing transactions at AllElectronics with respect to the purchase of computer games and videos. Let game refer to the transactions containing computer games, and video refer to those containing videos. Of the 10,000 transactions analyzed, the data show that 6,000 of the customer transactions included computer games, while 7,500 included videos, and 4,000 included both computer games and videos. Suppose that a data mining program for discovering association rules is run on the data, using a minimum support of, say, 30% and a minimum confidence of 60%. The following association rule is discovered:

\[
\text{buys}(X, \text{“computer games”}) \Rightarrow \text{buys}(X, \text{“videos”}) \quad \text{[support} = \text{40%}, \text{confidence} = \text{66%]} 
\]

(6.6)

Rule (6.6) is a strong association rule and would therefore be reported, since its support value of \(\frac{4,000}{10,000} = 40\%\) and confidence value of \(\frac{4,000}{6,000} = 66\%\) satisfy the minimum support and minimum confidence thresholds, respectively. However, Rule (6.6) is misleading because the probability of purchasing videos is 75\%, which is even larger than 66\%. In fact, computer games and videos are negatively associated because the purchase of one of these items actually decreases
the likelihood of purchasing the other. Without fully understanding this phenomenon, we could easily make unwise business decisions based on Rule (6.6).

The above example also illustrates that the confidence of a rule \( A \Rightarrow B \) can be deceiving. It does not measure the real strength (or lack of strength) of the correlation and implication between \( A \) and \( B \). Hence, alternatives to the support-confidence framework can be useful in mining interesting data relationships.

### 6.3.2 From Association Analysis to Correlation Analysis

As we have seen above, the support and confidence measures are insufficient at filtering out uninteresting association rules. To tackle this weakness, a correlation measure can be used to augment the support-confidence framework for association rules. This leads to correlation rules of the form

\[
A \Rightarrow B \text{ [support, confidence, correlation].} \tag{6.7}
\]

That is, a correlation rule is measured not only by its support and confidence but also by the correlation between itemsets \( A \) and \( B \). There are many different correlation measures from which to choose. In this section, we study several correlation measures to determine which would be good for mining large data sets.

**Lift** is a simple correlation measure that is given as follows. The occurrence of itemset \( A \) is independent of the occurrence of itemset \( B \) if \( P(A \cup B) = P(A)P(B) \); otherwise, itemsets \( A \) and \( B \) are dependent and correlated as events. This definition can easily be extended to more than two itemsets. The lift between the occurrence of \( A \) and \( B \) can be measured by computing

\[
lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}. \tag{6.8}
\]

If the resulting value of Equation (6.8) is less than 1, then the occurrence of \( A \) is negatively correlated with the occurrence of \( B \), meaning that the occurrence of one likely leads to the absence of the other one. If the resulting value is greater than 1, then \( A \) and \( B \) are positively correlated, meaning that the occurrence of one implies the occurrence of the other. If the resulting value is equal to 1, then \( A \) and \( B \) are independent and there is no correlation between them.

Equation (6.8) is equivalent to \( P(B|A)/P(B) \), or \( \text{conf}(A \Rightarrow B)/\text{sup}(B) \), which is also referred as the lift of the association (or correlation) rule \( A \Rightarrow B \). In other words, it assesses the degree to which the occurrence of one “lifts” the occurrence of the other. For example, if \( A \) corresponds to the sale of computer games and \( B \) corresponds to the sale of videos, then given the current market conditions, the sale of games is said to increase or “lift” the likelihood of the sale of videos by a factor of the value returned by Equation (6.8).

Let’s go back to the computer game and video data of Example 6.7.
6.3. WHICH PATTERNS ARE INTERESTING?—PATTERN EVALUATION METHODS

Table 6.6: A 2 × 2 contingency table summarizing the transactions with respect to game and video purchases.

<table>
<thead>
<tr>
<th></th>
<th>game</th>
<th>game</th>
<th>Σrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>video</td>
<td>4,000</td>
<td>3,500</td>
<td>7,500</td>
</tr>
<tr>
<td>video</td>
<td>2,000</td>
<td>500</td>
<td>2,500</td>
</tr>
<tr>
<td>Σcol</td>
<td>6,000</td>
<td>4,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Table 6.7: The above contingency table, now shown with the expected values.

<table>
<thead>
<tr>
<th></th>
<th>game</th>
<th>game</th>
<th>Σrow</th>
</tr>
</thead>
<tbody>
<tr>
<td>video</td>
<td>4,000 (4,500)</td>
<td>3,500 (3,000)</td>
<td>7,500</td>
</tr>
<tr>
<td>video</td>
<td>2,000 (1,500)</td>
<td>500 (1,000)</td>
<td>2,500</td>
</tr>
<tr>
<td>Σcol</td>
<td>6,000</td>
<td>4,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Example 6.8 **Correlation analysis using lift.** To help filter out misleading “strong” associations of the form \( A \Rightarrow B \) from the data of Example 6.7, we need to study how the two itemsets, \( A \) and \( B \), are correlated. Let \( \text{game} \) refer to the transactions of Example 6.7 that do not contain computer games, and \( \text{video} \) refer to those that do not contain videos. The transactions can be summarized in a contingency table, as shown in Table 6.6. From the table, we can see that the probability of purchasing a computer game is \( P(\{\text{game}\}) = 0.60 \), the probability of purchasing a video is \( P(\{\text{video}\}) = 0.75 \), and the probability of purchasing both is \( P(\{\text{game, video}\}) = 0.40 \). By Equation (6.8), the lift of Rule (6.6) is \( P(\{\text{game, video}\})/(P(\{\text{game}\}) \times P(\{\text{video}\})) = 0.40/(0.60 \times 0.75) = 0.89 \). Because this value is less than 1, there is a negative correlation between the occurrence of \( \{\text{game}\} \) and \( \{\text{video}\} \). The numerator is the likelihood of a customer purchasing both, while the denominator is what the likelihood would have been if the two purchases were completely independent. Such a negative correlation cannot be identified by a support-confidence framework.

The second correlation measure that we study is the \( \chi^2 \) measure, which was introduced in Chapter 3 (Equation 3.1). To compute the \( \chi^2 \) value, we take the squared difference between the observed and expected value for a slot (A and B pair) in the contingency table, divided by the expected value. This amount is summed for all slots of the contingency table. Let’s perform a \( \chi^2 \) analysis of the above example.

Example 6.9 **Correlation analysis using \( \chi^2 \).** To compute the correlation using \( \chi^2 \) analysis for nominal data, we need the observed value and expected value (displayed in parenthesis) for each slot of the contingency table, as shown in Table 6.7. From
CHAPTER 6. MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS: BASIC

the table, we can compute the $\chi^2$ value as follows:

$$
\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(4,000 - 4,500)^2}{4,500} + \frac{(3,500 - 3,000)^2}{3,000} + \frac{(2,000 - 1,500)^2}{1,500} + \frac{(500 - 1,000)^2}{1,000} = 555.6.
$$

Because the $\chi^2$ value is greater than one, and the observed value of the slot (game, video) = 4,000, which is less than the expected value 4,500, buying game and buying video are negatively correlated. This is consistent with the conclusion derived from the analysis of the lift measure in Example 6.9.

6.3.3 A Comparison of Pattern Evaluation Measures

The above discussion shows that instead of using the simple support-confidence framework to evaluate frequent patterns, other measures, such as lift and $\chi^2$, often discloses more intrinsic pattern relationships. How effective are these measures? Should we also consider other alternatives?

Researchers have studied many pattern evaluation measures even before the start of in-depth research on scalable methods for mining frequent patterns. Recently, several other pattern evaluation measures have attracted interest. In this section, we present four such measures: all confidence, max confidence, Kulczynski, and cosine. We’ll then compare their effectiveness with respect to one another and with respect to the lift and $\chi^2$ measures.

Given two itemsets $A$ and $B$, the all confidence measure of $A$ and $B$ is defined as

$$
\text{allconf}(A, B) = \frac{\text{sup}(A \cup B)}{\max\{\text{sup}(A), \text{sup}(B)\}} = \min\{P(A|B), P(B|A)\}, \quad (6.9)
$$

where $\max\{\text{sup}(A), \text{sup}(B)\}$ is the maximum support of the itemsets $A$ and $B$. Thus allconf$(A, B)$ is also the minimum confidence of the two association rules related to $A$ and $B$, namely, “$A\Rightarrow B$” and “$B\Rightarrow A$”.

Given two itemsets $A$ and $B$, the max confidence measure of $A$ and $B$ is defined as

$$
\text{maxconf}(A, B) = \max\{P(A|B), P(B|A)\}. \quad (6.10)
$$

The maxconf measure is the maximum confidence of the two association rules, “$A\Rightarrow B$” and “$B\Rightarrow A$”.

Given two itemsets $A$ and $B$, the Kulczynski measure of $A$ and $B$ (abbreviated as Kulc) is defined as

$$
\text{Kulc}(A, B) = \frac{1}{2}(P(A|B) + P(B|A)). \quad (6.11)
$$

It was proposed in 1927 by Polish mathematician S. Kulczynski. It can be viewed as an average of two confidence measures. That is, it is the average of
two conditional probabilities: the probability of itemset \( B \) given itemset \( A \), and the probability of itemset \( A \) given itemset \( B \).

Finally, given two itemsets \( A \) and \( B \), the cosine measure of \( A \) and \( B \) is defined as

\[
\text{cosine}(A, B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{\sup(A \cup B)}{\sqrt{\sup(A) \times \sup(B)}} = \sqrt{P(A|B) \times P(B|A)}.
\]

The cosine measure can be viewed as a harmonized lift measure: the two formulae are similar except that for cosine, the square root is taken on the product of the probabilities of \( A \) and \( B \). This is an important difference, however, because by taking the square root, the cosine value is only influenced by the supports of \( A \), \( B \), and \( A \cup B \), and not by the total number of transactions.

Each of the four measures defined above has the following property: Its value is only influenced by the supports of \( A \), \( B \), and \( A \cup B \), or more exactly, by the conditional probabilities of \( P(A|B) \) and \( P(B|A) \), but not by the total number of transactions. Another common property is that each measure ranges from 0 to 1, and the higher the value, the closer the relationship between \( A \) and \( B \).

Now, together with lift and \( \chi^2 \), we have introduced in total six pattern evaluation measures. You may wonder, “Which is the best in assessing the discovered pattern relationships?” To answer this question, we examine their performance on some typical data sets.

Table 6.8: A \( 2 \times 2 \) contingency table for two items.

<table>
<thead>
<tr>
<th></th>
<th>milk</th>
<th>( \overline{\text{milk}} )</th>
<th>( \Sigma_{\text{row}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>coffee</td>
<td>( \text{mc} )</td>
<td>( \overline{\text{mc}} )</td>
<td>( \bar{c} )</td>
</tr>
<tr>
<td>coffee</td>
<td>( \text{mc} )</td>
<td>( \overline{\text{mc}} )</td>
<td>( \bar{c} )</td>
</tr>
<tr>
<td>( \Sigma_{\text{col}} )</td>
<td>( m )</td>
<td>( \overline{m} )</td>
<td>( \Sigma )</td>
</tr>
</tbody>
</table>

Example 6.10 Comparison of six pattern evaluation measures on typical data sets.

The relationships between the purchases of two items, milk and coffee, can be examined by summarizing their purchase history in Table 6.8, a \( 2 \times 2 \) contingency table, where an entry such as \( \text{mc} \) represents the number of transactions containing both milk and coffee.

Table 6.9 shows a set of transactional data sets with their corresponding contingency tables and the associated values for each of the six evaluation measures. Let’s first examine the first four data sets, \( D_1 \) through \( D_4 \). From the table, we see that \( m \) and \( c \) are positively associated in \( D_1 \) and \( D_2 \), negatively associated in \( D_3 \), and neutral in \( D_4 \). For \( D_1 \) and \( D_2 \), \( m \) and \( c \) are positively associated because \( mc \) (10,000) is considerably greater than \( \overline{mc} \) (1,000) and \( m\overline{c} \) (1,000). Intuitively, for people who bought milk (\( m = 10,000 + 1,000 = 11,000 \)), it is very likely that they also bought coffee (\( mc/m = 10/11 = 91\% \)), and vice versa. The results of the four newly introduced measures show that \( m \) and \( c \) are strongly
Table 6.9: Comparison of six pattern evaluation measures using contingency tables for a variety of data sets.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>mc</th>
<th>m̄c</th>
<th>n̄c</th>
<th>m̄c</th>
<th>χ²</th>
<th>lift</th>
<th>all_conf</th>
<th>max_conf</th>
<th>Kulc.</th>
<th>cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>D₁</td>
<td>10,000</td>
<td>1,000</td>
<td>1,000</td>
<td>100,000</td>
<td>90557</td>
<td>9.26</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>D₂</td>
<td>10,000</td>
<td>1,000</td>
<td>1,000</td>
<td>100</td>
<td>1</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
</tr>
<tr>
<td>D₃</td>
<td>100</td>
<td>1,000</td>
<td>1,000</td>
<td>100,000</td>
<td>670</td>
<td>8.44</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
<td>0.09</td>
</tr>
<tr>
<td>D₄</td>
<td>1,000</td>
<td>1,000</td>
<td>1,000</td>
<td>100,000</td>
<td>24740</td>
<td>25.75</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>D₅</td>
<td>1,000</td>
<td>100</td>
<td>100,000</td>
<td>100,000</td>
<td>8173</td>
<td>9.18</td>
<td>0.99</td>
<td>0.91</td>
<td>0.5</td>
<td>0.29</td>
</tr>
<tr>
<td>D₆</td>
<td>1,000</td>
<td>10</td>
<td>100,000</td>
<td>100,000</td>
<td>965</td>
<td>1.97</td>
<td>0.01</td>
<td>0.99</td>
<td>0.5</td>
<td>0.10</td>
</tr>
</tbody>
</table>

positively associated in both data sets by producing a measure value of 0.91. However, lift and χ² generate dramatically different measure values for D₁ and D₂ due to their sensitivity to m̄c. In fact, in many real-world scenarios m̄c is usually huge and unstable. For example, in a market basket database, the total number of transactions could fluctuate on a daily basis and overwhelmingly exceed the number of transactions containing any particular itemset. Therefore, a good interestingness measure should not be affected by transactions that do not contain the itemsets of interest; otherwise, it would generate unstable results as illustrated in D₁ and D₂.

Similarly, in D₃, the four new measures correctly show that m and c are strongly negatively associated because the ratio of m to c equals the ratio of m̄c to m̄c, that is 100/1100 = 9.1%. However, lift and χ² both contradict this in an incorrect way: their values for D₂ are between those for D₁ and D₃.

For data set D₄, both lift and χ² indicate a highly positive association between m and c, whereas the others indicate a “neutral” association because the ratio of m̄c to m̄c equals the ratio of m̄c to m̄c, which is 1. This means that if a customer buys coffee (or milk), the probability that she will also purchase milk (or coffee) is exactly 50%.

“Why are lift and χ² so poor at distinguishing pattern association relationships in the above transactional data sets?” To answer this, we have to consider the null-transactions. A null-transaction is a transaction that does not contain any of the itemsets being examined. In our example, m̄c represents the number of null-transactions. Lift and χ² have difficulty distinguishing interesting pattern association relationships because they are both strongly influenced by m̄c. Typically, the number of null-transactions can outweigh the number of individual purchases, because many people may buy neither milk nor coffee. On the other hand, the other four measures are good indicators of interesting pattern associations because their definitions remove the influence of m̄c (that is, they are not influenced by the number of null-transactions).

The above discussion shows that it is highly desirable to have a measure whose value is independent of the number of null-transactions. A measure is null-invariant if its value is free from the influence of null-transactions. Null-invariance is an important property for measuring association patterns in large transaction databases. Among the six discussed measures in this section, only lift and χ² are not null-invariant measures.

“Among the all_confidence, max_confidence, Kulczynski, and cosine measures, which is best at indicating interesting pattern relationships?”
To answer this question, we introduce the **imbalance ratio** (IR), which assesses the imbalance of two itemsets \(A\) and \(B\) in rule implications. It is defined as

\[
IR(A, B) = \frac{|\text{sup}(A) - \text{sup}(B)|}{\text{sup}(A) + \text{sup}(B) - \text{sup}(A \cup B)}, \tag{6.13}
\]

where the numerator is the absolute value of the difference between the support of the itemsets \(A\) and \(B\), and the denominator is the number of transactions containing \(A\) or \(B\). If the two directional implications between \(A\) and \(B\) are the same, \(IR(A, B)\) will be zero. Otherwise, the larger the difference between the two is, the larger the imbalance ratio is. This ratio is independent of the number of null-transactions and independent of the total number of transactions.

Let’s continue examining the remaining datasets of Example 6.10.

**Example 6.11** Comparing null-invariant measures in pattern evaluation. Although the four measures introduced in this section are null-invariant, they may present dramatically different values on some subtly different datasets. Let’s examine data sets \(D_5\) and \(D_6\) in Table 6.9, where the two events \(m\) and \(c\) have unbalanced conditional probabilities. That is, the ratio of \(mc\) to \(c\) is greater than 0.9. This means that knowing that \(c\) occurs should strongly suggest that \(m\) occurs also. The ratio of \(mc\) to \(m\) is less than 0.1, indicating that \(m\) implies that \(c\) is quite unlikely to occur. The **all-confidence** and **cosine** measures view both cases as negatively associated and the **Kulc** measure views both as neutral. The **max-confidence** measure claims strong positive associations for these cases. The measures give very diverse results!

“Which measure intuitively reflects the true relationship between the purchase of milk and coffee?” Due to the “balanced” skewness of the data, it is difficult to argue whether the two data sets have positive or negative association. From one point of view, only \(mc/(mc + m\bar{c}) = 1,000/(1,000 + 10,000) = 9.09\%\) of milk-related transactions contain coffee in \(D_5\) and this percentage is \(1,000/(1,000 + 100,000) = 0.99\%\) in \(D_6\), both indicating a negative association. On the other hand, 90.9\% of transactions in \(D_5\) (that is, \(mc/(mc + m\bar{c}) = 1,000/(1,000 + 100))\) and 9\% in \(D_6\) (that is, \(1,000/(1,000 + 10))\) containing coffee contain milk as well, which indicates a positive association between milk and coffee. These draw very different conclusions. For such “balanced” skewness, it could be fair to treat it as neutral, as **Kulc** does, and in the meantime indicate its skewness using the **imbalance ratio** (IR). According to Equation (6.13), for \(D_4\) we have \(IR(m, c) = 0\), a perfectly balanced case; for \(D_5\), \(IR(m, c) = 0.89\), a rather imbalanced case; whereas for \(D_6\), \(IR(m, c) = 0.99\), a very skewed case. Therefore, the two measures, **Kulc** and **IR**, work together, presenting a clear picture for all three datasets, \(D_4\) through \(D_6\).

In summary, the use of only support and confidence measures to mine associations may generate a large number of rules, many of which can be uninteresting to users. Instead, we can augment the support-confidence framework with a pattern interestingness measure, which helps focus the mining towards rules with strong pattern relationships. The added measure substantially reduces
the number of rules generated and leads to the discovery of more meaningful rules. Besides those introduced in this section, many other interestingness measures have been studied in the literature. Unfortunately, most of them do not have the null-invariance property. Because large data sets typically have many null-transactions, it is important to consider the null-invariance property when selecting appropriate interestingness measures for pattern evaluation. Among the four null-invariant measures studied here, namely \textit{all-confidence}, \textit{max-confidence}, \textit{Kulc}, and \textit{cosine}, we recommend using \textit{Kulc} in conjunction with the imbalance ratio.

6.4 Summary

- The discovery of frequent patterns, association, and correlation relationships among huge amounts of data is useful in selective marketing, decision analysis, and business management. A popular area of application is \textit{market basket analysis}, which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence).

- \textbf{Association rule mining} consists of first finding \textbf{frequent itemsets} (set of items, such as $A$ and $B$, satisfying a \textit{minimum support threshold}, or percentage of the task-relevant tuples), from which \textbf{strong association rules} in the form of $A \Rightarrow B$ are generated. These rules also satisfy a \textit{minimum confidence threshold} (a prespecified probability of satisfying $B$ under the condition that $A$ is satisfied). Associations can be further analyzed to uncover \textbf{correlation rules}, which convey statistical correlations between itemsets $A$ and $B$.

- Many efficient and scalable algorithms have been developed for \textbf{frequent itemset mining}, from which association and correlation rules can be derived. These algorithms can be classified into three categories: (1) \textit{Apriori-like algorithms}, (2) \textit{frequent-pattern growth}-based algorithms, such as \textit{FP-growth}, and (3) \textit{algorithms that use the vertical data format}.

- The \textbf{Apriori algorithm} is a seminal algorithm for mining frequent itemsets for Boolean association rules. It explores the level-wise mining Apriori property that \textit{all nonempty subsets of a frequent itemset must also be frequent}. At the $k$th iteration (for $k \geq 2$), it forms frequent $k$-itemset candidates based on the frequent $(k - 1)$-itemsets, and scans the database once to find the \textit{complete} set of frequent $k$-itemsets, $L_k$.

  Variations involving hashing and transaction reduction can be used to make the procedure more efficient. Other variations include partitioning the data (mining on each partition and then combining the results) and sampling the data (mining on a subset of the data). These variations can reduce the number of data scans required to as little as two or one.

- \textbf{Frequent pattern growth (FP-growth)} is a method of mining frequent itemsets without candidate generation. It constructs a highly compact data
structure (an FP-tree) to compress the original transaction database. Rather than employing the generate-and-test strategy of Apriori-like methods, it focuses on frequent pattern (fragment) growth, which avoids costly candidate generation, resulting in greater efficiency.

- **Mining frequent itemsets using vertical data format (ECLAT)** is a method that transforms a given data set of transactions in the horizontal data format of TID-itemset into the vertical data format of item-TID-set. It mines the transformed data set by TID-set intersections based on the Apriori property and additional optimization techniques, such as diffset.

- Not all strong association rules are interesting. Therefore, the support-confidence framework should be augmented with a pattern evaluation measure, which promotes the mining of interesting rules. A measure is null-invariant if its value is free from the influence of null-transactions, i.e., the transactions that do not contain any of the itemsets being examined. Among many pattern evaluation measures, we examined lift, $\chi^2$, all-confidence, max-confidence, Kulczynski, and cosine, and showed that only the latter four are null-invariant. We suggest using the Kulczynski measure, together with the imbalance ratio, to present pattern relationships among itemsets.

### 6.5 Exercises

1. Suppose you have the set $C$ of all frequent closed itemsets on a data set $D$, as well as the support count for each frequent closed itemset. Describe an algorithm to determine whether a given itemset $X$ is frequent or not, and the support of $X$ if it is frequent.

2. An itemset $X$ is called a *generator* on a data set $D$ if there does not exist a proper sub-itemset $Y \subset X$ such that $support(X) = support(Y)$. A generator $X$ is a frequent generator if $support(X)$ passes the minimum support threshold. Let $G$ be the set of all frequent generators on a data set $D$.

   (a) Can you determine whether an itemset $A$ is frequent and the support of $A$, if it is frequent, using only $G$ and the support counts of all frequent generators? If yes, present your algorithm. Otherwise, what other information is needed? Can you give an algorithm assuming the information needed is available?

   (b) What is the relationship between closed itemsets and generators?

3. The Apriori algorithm makes use of prior knowledge of subset support properties.

   (a) Prove that all nonempty subsets of a frequent itemset must also be frequent.
(b) Prove that the support of any nonempty subset $s'$ of itemset $s$ must be at least as great as the support of $s$.

(c) Given frequent itemset $l$ and subset $s$ of $l$, prove that the confidence of the rule $s' \Rightarrow (l - s')$ cannot be more than the confidence of $s \Rightarrow (l - s)$, where $s'$ is a subset of $s$.

(d) A partitioning variation of Apriori subdivides the transactions of a database $D$ into $n$ nonoverlapping partitions. Prove that any itemset that is frequent in $D$ must be frequent in at least one partition of $D$.

4. Let $c$ be a candidate itemset in $C_k$ generated by the Apriori algorithm. How many length-$(k - 1)$ subsets do we need to check in the prune step? According to your answer to the above question, can you give an improved version of procedure \texttt{hasInfrequentSubset} in Figure 6.4?

5. Section 6.2.2 describes a method for generating association rules from frequent itemsets. Propose a more efficient method. Explain why it is more efficient than the one proposed there. (Hint: Consider incorporating the properties of Exercise 6.3(b) and 6.3(c) into your design.)

6. A database has 5 transactions. Let $\text{min sup} = 60\%$ and $\text{min conf} = 80\%$.

<table>
<thead>
<tr>
<th>TID</th>
<th>items bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>T100</td>
<td>{M, O, N, K, E, Y}</td>
</tr>
<tr>
<td>T200</td>
<td>{D, O, N, K, E}</td>
</tr>
<tr>
<td>T300</td>
<td>{M, A, K}</td>
</tr>
<tr>
<td>T400</td>
<td>{M, U, C, K}</td>
</tr>
<tr>
<td>T500</td>
<td>{C, O, O, K, I, E}</td>
</tr>
</tbody>
</table>

(a) Find all frequent itemsets using Apriori and FP-growth, respectively. Compare the efficiency of the two mining processes.

(b) List all of the \textit{strong} association rules (with support $s$ and confidence $c$) matching the following metarule, where $X$ is a variable representing customers, and \textit{item}_i denotes variables representing items (e.g., “A”, “B”, etc.):

$$\forall x \in \text{transaction, } \text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3) \quad [s, c]$$

7. (Implementation project) Implement three frequent itemset mining algorithms introduced in this chapter: (1) Apriori [AS94b], (2) FP-growth [HPY00], and (3) Eclat [Zak00] (mining using vertical data format), using a programming language that you are familiar with, such as C++ or Java. Compare the performance of each algorithm with various kinds of large data sets. Write a report to analyze the situations (such as data size, data distribution, minimal support threshold setting, and pattern density) where one algorithm may perform better than the others, and state why.
8. A database has four transactions. Let \( \minsup = 60\% \) and \( \minconf = 80\% \).

<table>
<thead>
<tr>
<th>cust_ID</th>
<th>TID</th>
<th>items_bought (in the form of brand-item_category)</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>T100</td>
<td>{King’s-Crab, Sunset-Milk, Dairyland-Cheese, Best-Bread}</td>
</tr>
<tr>
<td>02</td>
<td>T200</td>
<td>{Best-Cheese, Dairyland-Milk, Goldenfarm-Apple, Tasty-Pie, Wonder-Bread}</td>
</tr>
<tr>
<td>01</td>
<td>T300</td>
<td>{Westcoast-Apple, Dairyland-Milk, Wonder-Bread, Tasty-Pie}</td>
</tr>
<tr>
<td>03</td>
<td>T400</td>
<td>{Wonder-Bread, Sunset-Milk, Dairyland-Cheese}</td>
</tr>
</tbody>
</table>

(a) At the granularity of \( \text{item\_category} \) (e.g., \( i_k \) could be “Milk”), for the following rule template,

\[
\forall X \in \text{transaction}, \text{buys}(X, i_1) \land \text{buys}(X, i_2) \Rightarrow \text{buys}(X, i_3) \quad [s, c]
\]

list the frequent \( k \)-itemset for the largest \( k \), and all of the strong association rules (with their support \( s \) and confidence \( c \)) containing the frequent \( k \)-itemset for the largest \( k \).

(b) At the granularity of \( \text{brand\_item\_category} \) (e.g., \( i_k \) could be “Sunset-Milk”), for the following rule template,

\[
\forall X \in \text{customer}, \text{buys}(X, i_1) \land \text{buys}(X, i_2) \Rightarrow \text{buys}(X, i_3)
\]

list the frequent \( k \)-itemset for the largest \( k \) (but do not print any rules).

9. Suppose that a large store has a transaction database that is \textit{distributed} among four locations. Transactions in each component database have the same format, namely \( T_j : \{i_1, \ldots, i_m\} \), where \( T_j \) is a transaction identifier, and \( i_k (1 \leq k \leq m) \) is the identifier of an item purchased in the transaction. Propose an efficient algorithm to mine global association rules (without considering multilevel associations). You may present your algorithm in the form of an outline. Your algorithm should not require shipping all of the data to one site and should not cause excessive network communication overhead.

10. Suppose that frequent itemsets are saved for a large transaction database, \( DB \). Discuss how to efficiently mine the (global) association rules under the same minimum support threshold, if a set of new transactions, denoted as \( \Delta DB \), is (incrementally) added in?

11. Most frequent pattern mining algorithms consider only distinct items in a transaction. However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transaction data analysis. How can one mine frequent itemsets efficiently considering multiple occurrences of items? Propose modifications to the well-known algorithms, such as Apriori and FP-growth, to adapt to such a situation.
12. **(Implementation project)** Many techniques have been proposed to further improve the performance of frequent-itemset mining algorithms. Taking FP-tree-based frequent pattern-growth algorithms (such as FP-growth) as an example, implement one of the following optimization techniques. Compare the performance of your new implementation with the unoptimized version.

(a) The frequent pattern mining method of Section 6.2.4) uses an FP-tree to generate conditional pattern bases using a bottom-up projection technique, i.e., project on the prefix path of an item \( p \). However, one can develop a *top-down projection* technique, that is, project on the suffix path of an item \( p \) in the generation of a conditional pattern-base. Design and implement such a top-down FP-tree mining method and compare your performance with the bottom-up projection method.

(b) Nodes and pointers are used uniformly in an FP-tree in the design of the FP-growth algorithm. However, such a structure may consume a lot of space when the data are sparse. One possible alternative design is to explore *array-and pointer-based hybrid implementation*, where a node may store multiple items when it contains no splitting point to multiple sub-branches. Develop such an implementation and compare it with the original one.

(c) It is time- and space-consuming to generate numerous conditional pattern bases during pattern-growth mining. An interesting alternative is to *push right* the branches that have been mined for a particular item \( p \), that is, to push them to the remaining branch(es) of the FP-tree. This is done so that fewer conditional pattern bases have to be generated and additional sharing can be explored when mining the remaining branches of the FP-tree. Design and implement such a method and conduct a performance study on it.

13. Give a short example to show that items in a strong association rule may actually be *negatively correlated*.

14. The following contingency table summarizes supermarket transaction data, where *hot dogs* refers to the transactions containing hot dogs, *hotdogs* refers to the transactions that do not contain hot dogs, *hamburgers* refers to the transactions containing hamburgers, and *hamburgers* refers to the transactions that do not contain hamburgers.

<table>
<thead>
<tr>
<th></th>
<th>hot dogs</th>
<th>hotdogs</th>
<th>( \Sigma_{row} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>hamburgers</td>
<td>2000</td>
<td>500</td>
<td>2500</td>
</tr>
<tr>
<td>hamburgers</td>
<td>1000</td>
<td>1500</td>
<td>2500</td>
</tr>
<tr>
<td>( \Sigma_{col} )</td>
<td>3000</td>
<td>2000</td>
<td>5000</td>
</tr>
</tbody>
</table>
(a) Suppose that the association rule “hot dogs $\Rightarrow$ hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong?

(b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?

(c) Compare the use of the all\textunderscore confidence, max\textunderscore confidence, Kulczynski, and cosine measures with lift and correlation on the given data.

15. (Implementation project) The DBLP dataset (http://www.informatik.uni-trier.de/~ley/db/) consists of over one million entries of research papers published in computer science conferences and journals. Among these entries, there are a good number of authors that have coauthor relationships.

(a) Propose a method that can mine efficiently a set of coauthor relationships that are closely correlated (e.g., often coauthoring papers together).

(b) Based on the mining results and the pattern evaluation measures discussed in this chapter, discuss which measure may convincingly uncover close collaboration patterns better than others.

(c) Based on the study above, can you develop a method that can roughly predict advisor and advisee relationships and the approximate period for such advisory supervision?

### 6.6 Bibliographic Notes

Association rule mining was first proposed by Agrawal, Imielinski, and Swami [AIS93]. The Apriori algorithm discussed in Section 6.2.1 for frequent itemset mining was presented in Agrawal and Srikant [AS94b]. A variation of the algorithm using a similar pruning heuristic was developed independently by Mannila, Tiovonen, and Verkamo [MTV94]. A joint publication combining these works later appeared in Agrawal, Mannila, Srikant, Toivonen, and Verkamo [AMS+96]. A method for generating association rules from frequent itemsets is described in Agrawal and Srikant [AS94a].

References for the variations of Apriori described in Section 6.2.3 include the following. The use of hash tables to improve association mining efficiency was studied by Park, Chen, and Yu [PCY95a]. The partitioning technique was proposed by Savasere, Omiecinski, and Navathe [SON95]. The sampling approach is discussed in Toivonen [Toi96]. A dynamic itemset counting approach is given in Brin, Motwani, Ullman, and Tsur [BMUT97]. An efficient incremental updating of mined association rules was proposed by Cheung, Han, Ng, and Wong [CHNW96]. Parallel and distributed association data mining under the Apriori framework was studied by Park, Chen, and Yu [PCY95b], Agrawal and
Shafer [AS96], and Cheung, Han, Ng, et al. [CHN+96]. Another parallel association mining method, which explores itemset clustering using a vertical database layout, was proposed in Zaki, Parthasarathy, Ogihara, and Li [ZPOL97].

Other scalable frequent itemset mining methods have been proposed as alternatives to the Apriori-based approach. FP-growth, a pattern-growth approach for mining frequent itemsets without candidate generation, was proposed by Han, Pei, and Yin [HPY00] (Section 6.2.4). An exploration of hyper-structure mining of frequent patterns, called H-Mine, was proposed by Pei, Han, Lu, et al. [PHMA+01]. A method that integrates top-down and bottom-up traversal of FP-trees in pattern-growth mining, was proposed by Liu, Pan, Wang, and Han [LPWH02]. An array-based implementation of prefix-tree-structure for efficient pattern growth mining was proposed by Grahne and Zhu [GZ03]. Eclat, an approach for mining frequent itemsets by exploring the vertical data format, was proposed by Zaki [Zak00]. A depth-first generation of frequent itemsets by a tree projection technique was proposed by Agarwal, Aggarwal, and Prasad [AAP01]. An integration of association mining with relational database systems was studied by Sarawagi, Thomas, and Agrawal [? ].

The mining of frequent closed itemsets was proposed in Pasquier, Bastide, Taouil, and Lakhal [? ], where an Apriori-based algorithm called A-Close for such mining was presented. CLOSESET, an efficient closed itemset mining algorithm based on the frequent pattern growth method, was proposed by Pei, Han, and Mao [? ]. CHARm by Zaki and Hsiao [? ] develops a compact vertical TID list structure called diffset which only records the difference in the TID list of a candidate pattern from its prefix pattern. A fast hash-based approach is also used in CHARm to prune non-closed patterns. CLOSESET+ by Wang, Han and Pei [? ] integrates previously proposed effective strategies as well as newly developed techniques such as hybrid tree-projection and item skipping. AFOPT, a method that explores a right push operation on FP-trees during the mining process, was proposed by Liu, Lu, Lou and Yu [? ]. A prefix tree-based algorithm integrated with array representation, called FPClose, for mining closed itemsets using pattern-growth approach, was proposed by Grahne and Zhu [GZ03]. Pan, Cong, Tung, et al. [? ] proposed CARPENTER, a method for finding closed patterns in long biological datasets, which integrates the advantages of vertical data formats and pattern growth methods. Mining max-patterns was first studied by Bayardo [? ], where MaxMiner, an Apriori-based, level-wise, breadth-first search method was proposed to find max-itemset by performing superset frequency pruning and subset infrequency pruning for search space reduction. Another efficient method MAFIA, proposed by Burdick, Calimlim, and Gehrke [? ], uses vertical bitmap to compress transaction id list, thus improving the counting efficiency. A FIMI (Frequent Itemset Mining Implementation) workshop dedicated to the implementation methods of frequent itemset mining was reported by Goethals and Zaki [? ].

The problem of mining interesting rules has been studied by many researchers. The statistical independence of rules in data mining was studied by Piatetski-Shapiro [PS91]. The interestingness problem of strong association rules is discussed in Chen, Han, and Yu [CHY96], Brin, Motwani, and Silverstein [BMS97],
and Aggarwal and Yu [AY99], which cover several interestingness measures including lift. An efficient method for generalizing associations to correlations is given in Brin, Motwani, and Silverstein [BMS97]. Other alternatives to the support-confidence framework for assessing the interestingness of association rules are proposed in Brin, Motwani, Ullman, and Tsur [BMUT97] and Ahmed, El-Makky, and Taha [AEMT00]. A method for mining strong gradient relationships among itemsets was proposed by Imielinski, Khachiyan, and Abdulghani [IKA02]. Silverstein, Brin, Motwani, and Ullman [SBMU98] studied the problem of mining causal structures over transaction databases. Some comparative studies of different interestingness measures were done by Hilderman and Hamilton [HH01]. The notion of null transaction invariance was introduced, together with a comparative analysis of interestingness measures, by Tan, Kumar and Srivastava [TKS02]. The use of all confidence as a correlation measure for generating interesting association rules was studied by Omiecinski [Omi03] and by Lee, Kim, Cai and Han [LKCH03]. Wu, Chen and Han [WCH10] introduced the Kulczynski measure for associative patterns and performed a comparative analysis of a set of measures for pattern evaluation.
CHAPTER 6. MINING FREQUENT PATTERNS, ASSOCIATIONS, AND CORRELATIONS: BASI
Bibliography


