Programme : Bachelor of Science

Subject : Computer Science

Semester : V

Course Code : CSD102

Course Title : Data Mining

Unit V : Association Analysis

Module Name : Mining Multilevel Association Rules, Mining Multidimensional Association Rules, Other Applications of Association Rule Mining.

Notes

Mining Multilevel Association Rules

Multilevel associations involve concepts at different abstraction levels. It is interesting to examine how to develop effective methods for mining patterns at multiple abstraction levels, with sufficient flexibility for easy traversal among different abstraction spaces. For many applications, it is difficult to find strong associations among data items at low or primitive levels of abstraction due to the sparsity of data at those levels.

Association rules generated from mining data at multiple levels of abstraction are called multiple-level or multilevel association rules. Multilevel association rules can be mined efficiently using concept hierarchies under a support confidence framework.

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher level, more general concepts. Data can be generalized by replacing low-level concepts within the data by their higher-level concepts, or ancestors, from a concept hierarchy.

In general, a top-down strategy is employed, where counts are accumulated for the calculation of frequent item sets at each concept level, starting at the concept level 1 and working downward in the hierarchy towards the more specific concept levels, until no more frequent item sets can be found. For each level, any algorithm for discovering frequent item sets may be used, such as Apriori or its variations.

Example for Mining multilevel association rules: Suppose we are given the task-relevant set of transactional data for sales in an AllElectronics store, showing the items purchased for each transaction. The concept hierarchy for the items is shown in Figure below:

Task-Relevant Data, *D*

TID	*Items Purchased*
T100	Apple 1700 MacBook Pro Notebook, HP Photosmart Pro b9180
T200	Microsoft Office Professional 2010, MicrosoftWireless Optical Mouse 5000
T300	Logitech VX Nano Cordless Laser Mouse, Fellowes GELWrist Rest
T400	Dell Studio XPS 16 Notebook, Canon PowerShot SD1400
T500	Lenovo ThinkPad X200 Tablet PC, Symantec Norton Antivirus 2010
. . . . . .	

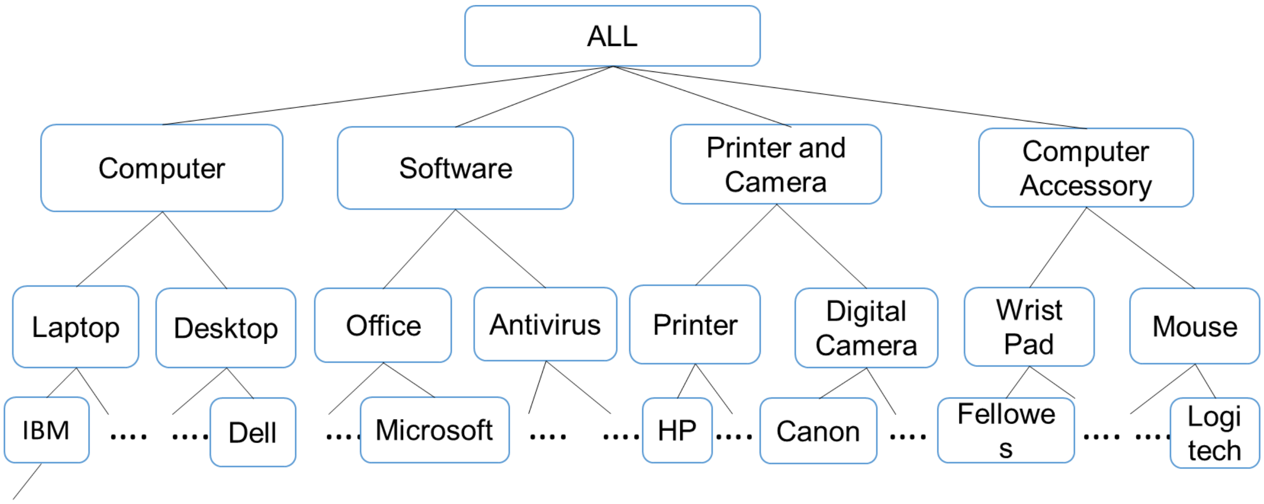

Figure: Concept Hierarchy for AllElectronics Computer Items

Approaches for Mining Multilevel Association Rules

A number of variations to the apriori approach are described next, where each variation involves “playing” around with the support threshold in a slightly different way.

The approaches are:

1. **Using uniform minimum support for all levels (referred to as uniform support):**

The same minimum support threshold is used when mining at each level of abstraction. When a uniform minimum support threshold is used, the search procedure is simplified. The method is also simple in that users are required to specify only one minimum support threshold.

The uniform support approach, however, has some difficulties. It is unlikely that items at lower levels of abstraction will occur as frequently as those at higher levels of abstraction. If the minimum support threshold is set too high, it could miss some meaningful associations occurring at low abstraction levels. If the threshold is set too low, it may generate many uninteresting associations occurring at high abstraction levels.

Figure below shows the representation of uniform min_sup=5% at all levels. The nodes in the concept hierarchy indicate an item or itemset that has been examined, and nodes with thick borders indicate that an examined item or itemset is frequent.

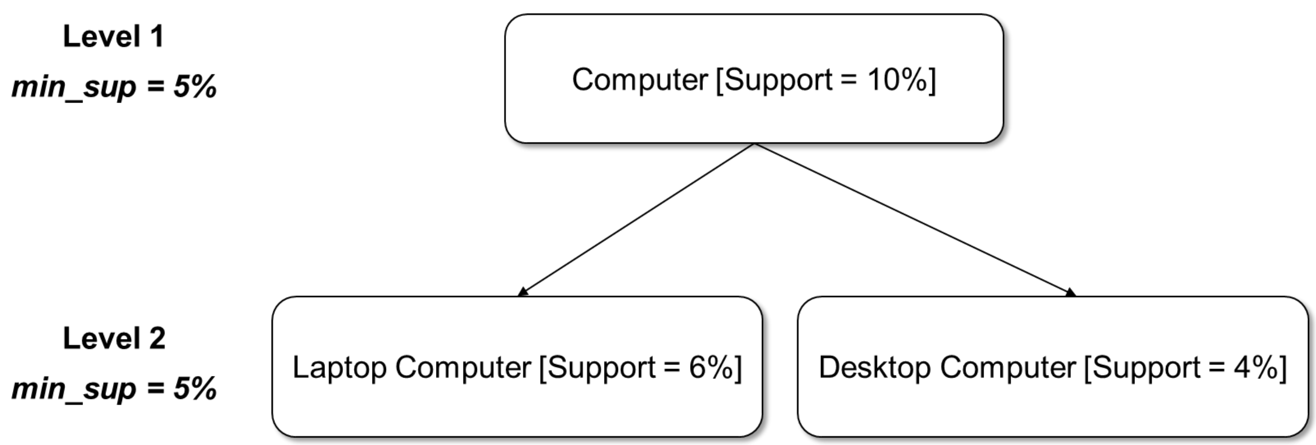

Figure : Multilevel Mining with Uniform Support

2. Using reduced minimum support at lower levels (referred to as Reduced Support)

Each level of abstraction has its own minimum support threshold. The deeper the level of abstraction, the smaller the corresponding threshold is. For example, in the below figure ,the minimum support thresholds for levels 1 and 2 are 5% and 3%, respectively. In this way, —computer, —laptop computer, and —desktop computer are all considered frequent.

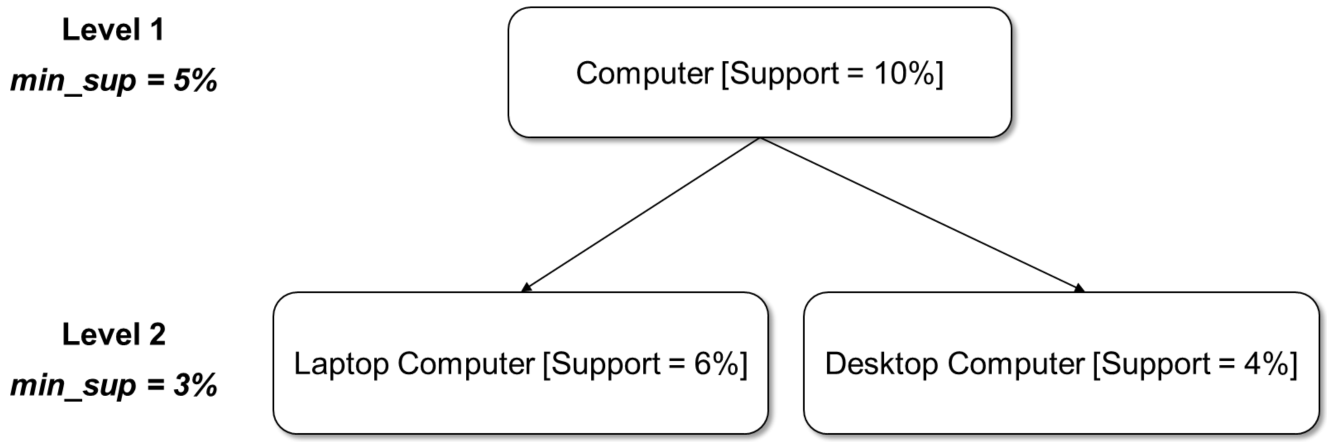

Figure: Multilevel Mining with Reduced Support

3. Using item or group-based minimum support (referred to as group-based support)

Because users or experts often have insight as to which groups are more important than others, it is sometimes more desirable to set up user-specific, item, or group based minimal support thresholds when mining multilevel rules. For example, a user could set up the minimum support thresholds based on product price, or on items of interest, such as by setting particularly low support thresholds for laptop computers and flash drives in order to pay particular attention to the association patterns containing items in these categories.

Mining Multidimensional Association Rules

For instance, in mining ***AllElectronics*** database, we may discover the Boolean association rule as follows:

buys(**X**, ’’**digital camera**”) -> **buys**(**X**, ”**HP printer”**)

Following the terminology used in multidimensional databases, we refer to each distinct predicate in a rule as a dimension. Hence, we can refer to Rule above as a single-dimensional or intra dimensional association rule because it contains a single distinct predicate (e.g., buys) with multiple occurrences (i.e., the predicate occurs more than once within the rule). Such rules are commonly mined from transactional data.

Instead of considering transactional data only, sales and related information are often linked with relational data or integrated into a data warehouse. Such data stores are multidimensional in nature. For instance, other in addition to keeping track of the items purchased in sales transactions, a relational database may record attributes associated with the items and/or transactions such as the item description or the branch location of the sale. Additional relational information regarding the customers who purchased the items (e.g., customer age, occupation, credit rating, income, and address) may also be stored.

Considering each database attribute or warehouse dimension as a predicate, we can therefore mine association rules containing multiple predicates such as *age*.

age(X, “20.....29') ^ occupation(X, "student”) => buys(X, "laptop”)

Single dimensional or intra dimensional association rule contains a single distinct predicate (e.g., buys) with multiple occurrences i.e., the predicate occurs more than once within the rule.

buys(X, ’’digital camera”)->buys(X, ”HP printer”)

Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules.

age(X, “20…29”)^occupation(X, “student”)=>buys(X, “laptop”)

Above Rule contains three predicates (age, occupation, and buys), each of which occurs only once in the rule. Hence, we say that it has no repeated predicates.

Multidimensional association rules with no repeated predicates are called inter dimensional association rules. We can also mine multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates. These rules are called hybrid- dimensional association rules.

Other Applications of Association Rule Mining

1] Medical Diagnosis:

Association rules in medical diagnosis can be useful for assisting physicians for curing patients. Diagnosis is not an easy process and has a scope of errors which may result in unreliable end-results. Using relational association rule mining, we can identify the probability of the occurrence of an illness concerning various factors and symptoms. Further, using learning techniques, this interface can be extended by adding new symptoms and defining relationships between the new signs and the corresponding diseases.

2] Census Data:

Every government has tonnes of census data. This data can be used to plan efficient public services (education, health, transport) as well as help public businesses (for setting up new factories, shopping malls, and even marketing particular products). This application of association rule mining and data mining has immense potential in supporting sound public policy and bringing forth an efficient functioning of a democratic society.

3] Protein Sequence:

Proteins are sequences made up of twenty types of amino acids. Each protein bears a unique 3D structure which depends on the sequence of these amino acids.

A slight change in the sequence can cause a change in structure which might change the functioning of the protein. This dependency of the protein functioning on its amino acid sequence has been a subject of great research.

The nature of associations between different amino acids that are present in a protein can be identified by generating association rules.