**Topic:**Association Rules In Data Mining in large databases

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is Market Based Analysis.

Market Based Analysis is one of the key techniques used by large relations to show associations between items.It allows retailers to identify relationships between the items that people buy together frequently.

Given a set of transactions, we can find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction.

| TID | Items |
| --- | --- |
| 1 | Bread, Milk |
| 2 | Bread, Diaper, Beer, Eggs |
| 3 | Milk, Diaper, Beer, Coke |
| 4 | Bread, Milk, Diaper, Beer |
| 5 | Bread, Milk, Diaper, Coke |

Before we start defining the rule, let us first see the basic definitions.

**Support Count() –** Frequency of occurrence of a itemset.

Here ({Milk, Bread, Diaper})=2

**Frequent Itemset –** An itemset whose support is greater than or equal to minsup threshold.

**Association Rule –** An implication expression of the form X -> Y, where X and Y are any 2 itemsets.

Example: {Milk, Diaper}->{Beer}

**Rule Evaluation Metrics –**

* **Support(s) –**  
  The number of transactions that include items in the {X} and {Y} parts of the rule as a percentage of the total number of transaction.It is a measure of how frequently the collection of items occur together as a percentage of all transactions.
* **Support = (X+Y)  total –**  
  It is interpreted as fraction of transactions that contain both X and Y.
* **Confidence(c) –**  
  It is the ratio of the no of transactions that includes all items in {B} as well as the no of transactions that includes all items in {A} to the no of transactions that includes all items in {A}.
* **Conf(X=>Y) = Supp(XY)  Supp(X) –**  
  It measures how often each item in Y appears in transactions that contains items in X also.
* **Lift(l) –**  
  The lift of the rule X=>Y is the confidence of the rule divided by the expected confidence, assuming that the itemsets X and Y are independent of each other.The expected confidence is the confidence divided by the frequency of {Y}.
* **Lift(X=>Y) = Conf(X=>Y)  Supp(Y) –**  
  Lift value near 1 indicates X and Y almost often appear together as expected, greater than 1 means they appear together more than expected and less than 1 means they appear less than expected.Greater lift values indicate stronger association.

**Example –** From the above table, {Milk, Diaper}=>{Beer}

s= ({Milk, Diaper, Beer}) |T|

= 2/5

= 0.4

c= (Milk, Diaper, Beer) (Milk, Diaper)

= 2/3

= 0.67

l= Supp({Milk, Diaper, Beer}) Supp({Milk, Diaper})\*Supp({Beer})

= 0.4/(0.6\*0.6)

= 1.11

One of the best examples of association rule mining is **market basket analysis**.  
  
This process analyzes the customer's buying habits by finding associations between different items that customers place in their shopping habits.   
  
The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by the customers.  
  
For instance, if customers are buying soap, how likely are they to also buy shampoo(and which kind of shampoo) on the same trip to the supermarket.  
  
Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space.  
  
For example, placing milk and bread within close proximity may further encourage the sale of these items together within single visits to the store.

## Classification of Association Rules

### Boolean Association Rule

* It is based on the types of values handled in the rule, If a rule concerns associations between the presence or absence of items, it is a Boolean Association Rule.
* Example : laptop=> system\_management\_software  
                  [ support = 2%, confidence = 60%]

### Quantitative Association Rule

* If a rule describes associations between quantitative items or attributes, then it is a quantitative association rule.
* In these rules, quantitative values for items or attributes are partitioned into intervals. The following rule is an ex of a quantitative association rule, where X is a variable representing a customer
* Example: age (x, “30..39”) ^ income (x, “42..48K”) - >buys (x, bike)

### Single Dimension Association Rules

* It is based on the dimensions of data involved in the rule, If the items or attributes in an association rule reference only one dimension, then it is a single-dimensional association rule.

### Note the Rule

**laptop=> system\_management\_software**[ support = 2%, confidence = 60%] is a single-dimensional association rule since it refers to only one dimension, buys.

If a rule references two or more dimensions, such as the dimensions buys, time\_of\_transaction, and customer\_category, then it is a multidimensional association rule.  
   
age (x, “30..39”) ^ income (x, “42..48K”) - > buys (x, bike)  
      
The above rule is considered a multidimensional association rule since it involves three dimensions, age, income, and buys.

# Topic: Analytical Characterization In Data Mining - Attribute Relevance Analysis

Let's consider a situation where  
  
"What if we are not sure which attribute to include for class characterization and class comparison? We may end up specifying too many attributes, which could slow down the system considerably”.  
  
To overcome this situation we need to perform analytical characterization.  
  
It is the measure of attribute relevance analysis that can be used to help identify irrelevant or weakly relevant attributes that can be excluded from the concept description process.  
   
The incorporation of this processing step into class characterization or comparison is referred to as analytical characterization or analytical comparison.

## Why Analytical Characterization

It is used because,

The first limitation of the OLAP tool is the handling of complex objects.  
  
The second limitation is the lack of an automated generalization process, the user must explicitly tell the system which dimensions should be included in the class characterization and how high a level each dimension should be generalized.  
  
Actually, each step of generalization or specialization on any dimension must be specified by the user.  
  
Usually, it is not difficult for a user to instruct a data mining system regarding how high a level each dimension should be generalized.   
  
For example, users can set attribute generalization thresholds for this, or specify which level a given dimension should reach, such as with the command “generalize dimension location to the country level”.

Even without explicit user instructions, a default value such as 2 to 8 can be set by the data mining system, which would allow each dimension to be generalized to a level that contains only 2 to 8 unique values.  
  
On the other hand, normally a user may include too few attributes in the analysis, causing incomplete mining results or a user may introduce too many attributes for analysis e.g “in relevance to \*”.  
  
Methods should be introduced to perform attribute relevance analysis to filter out statistically irrelevant or weakly relevant attributes.  
  
The class characterization that includes the analysis of attribute/dimension relevance is called analytical characterization.  
  
The class comparison that includes such analysis is called analytical comparison.

## Attribute Relevance Analysis

### 1. Data Collection:

* It is collecting the data for both the target class and the contrasting class by query processing.

### 2. Preliminary relevance analysis using conservative AOI:

* This step identifies a set of dimensions and attributes on which the selected relevance measure is to be applied.
* The relation obtained by such an application of Attribute Oriented Induction is called the candidate relation of the mining task.

### 3. Remove irrelevant and weakly relevant attributes using the selected relevance analysis:

* We evaluate each attribute in the candidate relation using the selected relevance analysis measure.
* This step results in an initial target class working relation and initial contrasting class working relation.

### 4. Generate the concept description using AOI:

* We need to perform the Attribute Oriented Induction process using a less conservative set of attribute generalization thresholds.

If descriptive mining is

* Class characterization, only ITCWR is included.
* Class Comparison both ITCWR and ICCWR are included.

## Relevance Measures

Quantitative relevance measure determines the classifying power of an attribute within a set of data.

Some of the methods of quantitative relevance measure are:

* Information Gain (ID3)
* Gain Ratio (C4.5)
* Gini Index
* Chi^2 contingency table statistics
* Uncertainty Coefficient

## Entropy & Information Gain

S contains si tuples of class Ci for i = {1, …, m}.

Information measures info required to classify any arbitrary tuple.

Text

Description automatically generated with low confidence  
  
Entropy of attribute A with values {a1,a2,…,av} can be used to partition S into the susets {S1,S2,..Sv} where Sj contain sij samples of class Ci.

Chart

Description automatically generated

Gain(A) = I(S1,S2,...,Sm)- E(A)

The information gained by branching on attribute A

## Example: Analytical Characterization

### Task

* To mine general characteristics describing graduate students using analytical characterization.

### Given

* Attributes name, gender, major, birth\_place, birth\_date, phone#, and GPA.
* Gen(ai) = concept hierarchies on ai.
* Ui = attribute analytical thresholds for ai.
* Ti = attribute generalization thresholds for ai.
* R = attribute relevance threshold.

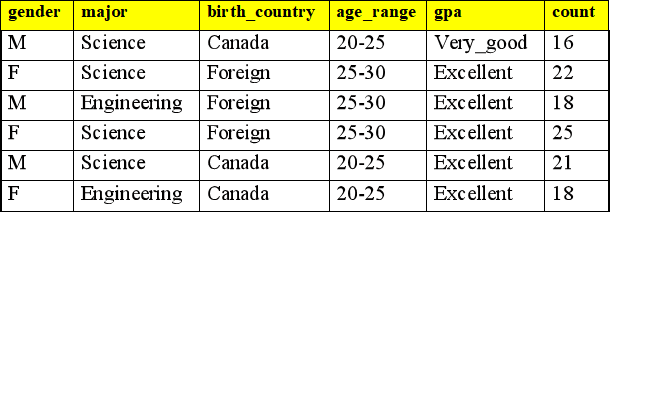
### 1. Data collection

* Target Class: graduate student
* Contrasting Class: undergraduate student

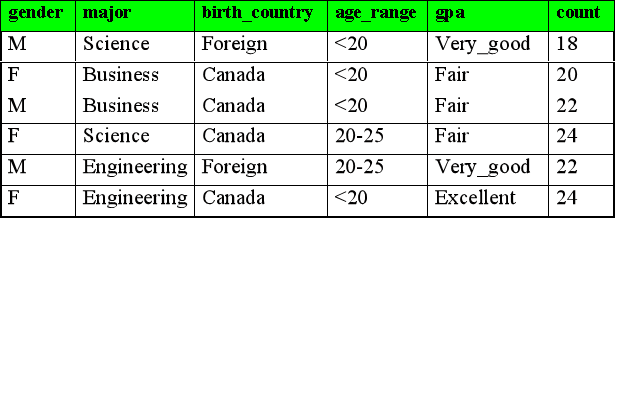
### 2. Analytical generalization using Ui

* Attribute Removal -> to remove the name and phone#
* Attribute Generalization -> to generalize major, birth\_place, birth\_date and GPA, accumulate counts
* Candidate Relation(large attribute generalization threshold): gender, major, birth\_country, age\_range, and GPA

Candidate relation for Target class: Graduate students (summation = 120):



Candidate relation for Contrasting class: Undergraduate students (Summation=130):



### 3. Relevance analysis

* We need to calculate the expected info required to classify an arbitrary tuple.
* Similarly, we need to calculate the entropy of each attribute: e.g. major
* And also calculate information gain for each attribute.