The Apriori algorithm is a popular rule-generating algorithm, though we still need a way to assess the strength of the rule.

KENT STATE UNIVERSITY

Association Rules - Research

- **Started in 1960s (**Hájek, P., Havel, I. & Chytil, M. Computing (1966) 1: 293. https://doi.org/10.1007/BF02345483)
  - Focused on mathematical representation rather than algorithm
- **Framework for Association Rules (**Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining association rules between sets of items in large databases. SIGMOD Rec. 22, 2 (June 1993), 207-216)
  - Discovering regularities between products in a large database of customer transactions recorded by pos systems in supermarkets

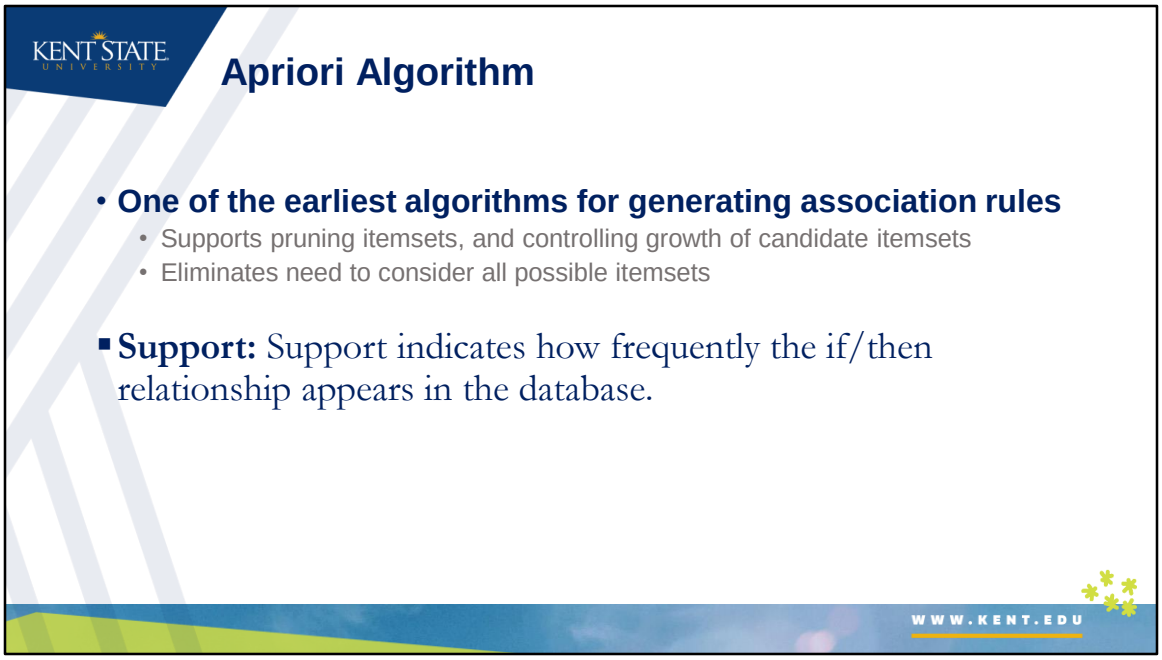
WWW.KENT.EDU

While early research on Association Rules was published in 1960s, it is only with Agarwal et al. paper was a practical implementation of the rules developed. Here, we will concentrate on one such practical implementation called the Apriori Algorithm.

Definitions

- A transaction t contains X, a set of items (itemset) in I, if $X \subseteq t$.
- An association rule is an implication of the form:
  $X \rightarrow Y$, where $X, Y \subset I$, and $X \cap Y = \varnothing$
- An itemset is a set of items.
  - E.g., X = {milk, bread, cereal} is an itemset.
- A k-itemset is an itemset with k items.
  - E.g., {milk, bread, cereal} is a 3-itemset

An association rule is an if-then condition. We refer to the IF as antecedent, and THEN as consequent . In our example above, X is the antecedent, and Y the consequent. In association analysis, the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common). Note that itemsets are not records of what people buy; they are simply possible combinations of items, including single items.



Several algorithms have been proposed for generating frequent itemsets, but the classic algorithm is the Apriori algorithm of Agrawal et al. (1993). **It pioneered the use of support for pruning the itemsets and controlling the exponential growth of candidate itemsets. Shorter candidate itemsets, which are known to be frequent itemsets, are combined and pruned to generate longer frequent itemsets. This approach eliminates the need for all possible itemsets to be enumerated within the algorithm, since the number of all possible itemsets can become exponentially large.**

What is Support that is used to generate frequent itemsets?

KENT STATE UNIVERSITY

Support: Formal Definition

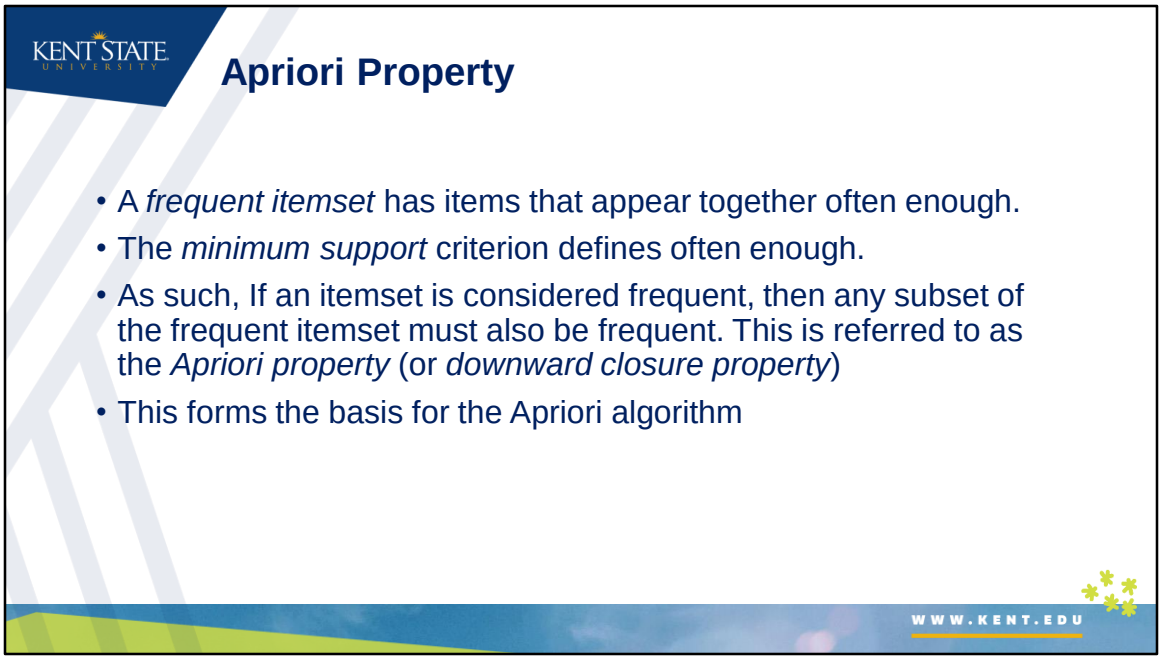
- **Support:** The rule holds with support sup in T (the transaction data set) if sup% of transactions contain X ∩ Y.
  - sup = Pr(X ∩ Y).

$$\text{Support} = \frac{\text{Count}(X \cap Y)}{n}$$

- An association rule is a pattern that states when X occurs, Y occurs with certain probability.

WWW.KENT.EDU

The support of a rule is simply the number of transactions that include both the antecedent and consequent itemsets. It is called a support because it measures the degree to which the data "support" the validity of the rule. The support is sometimes expressed as a percentage of the total number of records in the database.



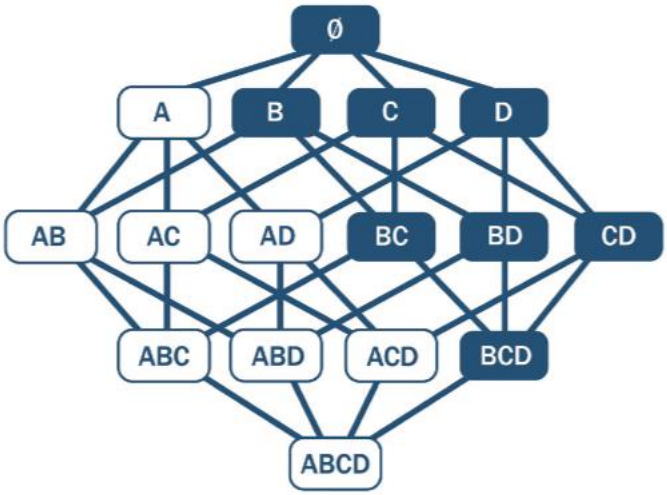
A *frequent itemset* has items that appear together often enough. The term "often enough" is formally defined with a *minimum support* criterion. If the minimum support is set at 0.5, any itemset can be considered a frequent itemset if at least 50% of the transactions contain this itemset. In other words, the support of a frequent itemset should be greater than or equal to the minimum support.

If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent. This is referred to as the *Apriori property* (or *downward closure property*). For example, if 60% of the transactions contain {bread,jam}**, then at least 60% of all the transactions will contain** {bread} **or** {jam}**. In other words, when the support of** {bread,jam} **is 0.6, the support of** {bread} **or** {jam} **is at least 0.6.**

The next diagram illustrates this Apriori property.

KENT STATE UNIVERSITY

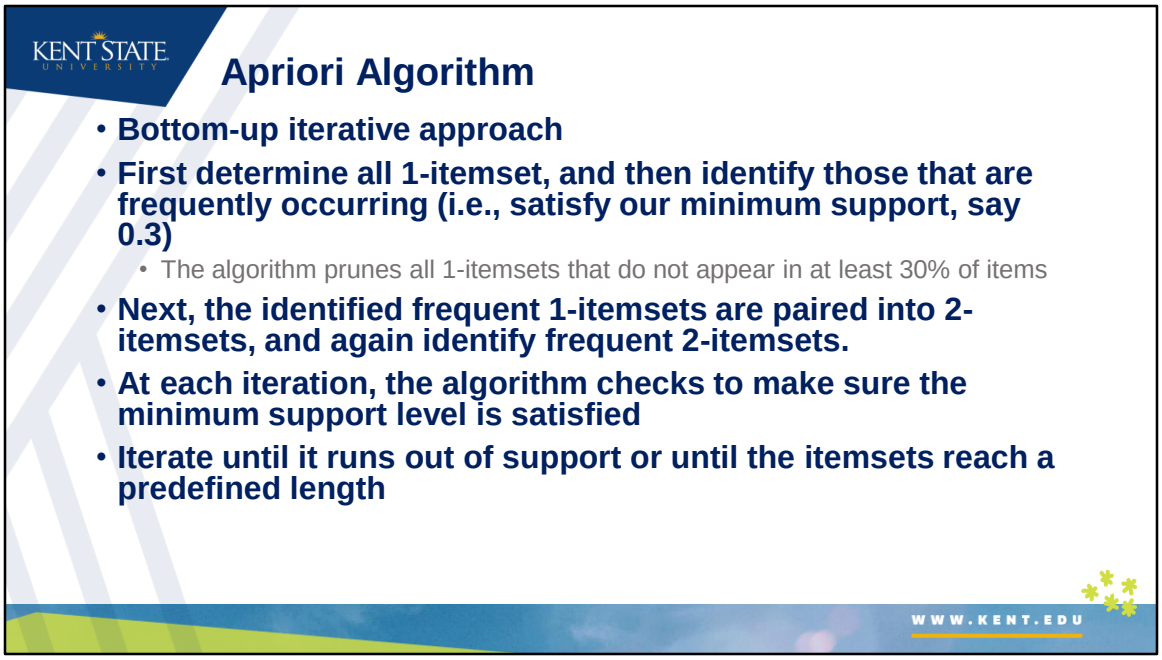
Apriori Property Diagram



Ref: Data Science and Big Data Analytics

WWW.KENT.EDU

The above figure illustrates how the Apriori property works. If itemset {B,C,D} **is frequent, then all the subsets of this itemset, shaded, must also be frequent itemsets. The Apriori property provides the basis for the Apriori algorithm.**



The first step of the Apriori algorithm is to identify the frequent itemsets by starting with each item in the transactions that meets the predefined minimum support threshold. These itemsets are 1-itemsets denoted as *L*1, as each 1-itemset contains only one item. Next, the algorithm grows the itemsets by joining *L*1 onto itself to form new, grown 2-itemsets denoted as *L*2 and determines the support of each 2-itemset in *L*2. Those itemsets that do not meet the minimum support threshold are pruned away. The growing and pruning process is repeated until no itemsets meet the minimum support threshold. Optionally, a threshold *N* can be set up to specify the maximum number of items the itemset can reach or the maximum number of iterations of the algorithm. Once completed, output of the Apriori algorithm is the collection of all the frequent *k*-itemsets.

Next, a collection of candidate rules is formed based on the frequent itemsets uncovered in the iterative process described earlier.

KENT STATE UNIVERSITY

Example

- Transaction data →

t1: Beef, Chicken, Milk
t2: Beef, Cheese
t3: Cheese, Boots
t4: Beef, Chicken, Cheese
t5: Beef, Chicken, Clothes, Cheese, Milk
t6: Chicken, Clothes, Milk
t7: Chicken, Milk, Clothes

- Assume:
  minsup = 30%
- An example frequent itemset:
  {Chicken, Clothes, Milk} [sup = 3/7]
- Association rules from the itemset:
  Clothes → Milk, Chicken [sup = 3/7]
  … …
  Clothes, Chicken → Milk, [sup = 3/7]

WWW.KENT.EDU

This illustrates the result of the Apriori algorithm and a set of candidate rules. Note that Chicken, Clothes, Milk would each satisfy the support level as 1-itemsets.

Measures to Evaluate Candidate Rules

Frequent itemsets can form candidate rules such as X implies Y (X → Y). There are three measures we can use to evaluate candidate rules

1. Confidence
2. Lift
3. Leverage

WWW.KENT.EDU

From the many rules generated, the goal is to find only the rules that indicate a strong dependence between the antecedent and consequent itemsets. To measure the strength of association implied by a rule, we can use several measures to evaluate our rules. Three of them are Confidence, Lift, and Leverage.

KENT STATE UNIVERSITY

Confidence

Confidence is defined as the measure of certainty or trustworthiness associated with each discovered rule. Confidence is the percent of transactions that contain both X and Y out of all the transactions that contain X

$$\text{Confidence}(X \rightarrow Y) = \frac{\text{Support}(X \cap Y)}{\text{Support}\, X}$$

WWW.KENT.EDU

Another measure that expresses the degree of uncertainty about the if–then rule is known as the confidence of the rule. This measure compares the co-occurrence of the antecedent and consequent itemsets in the database to the occurrence of the antecedent itemsets. Confidence is defined as the ratio of the number of transactions that include all antecedent and consequent itemsets (namely, the support) to the number of transactions that include all the antecedent itemsets

For example, assume that we have 1000 transactions in our dataset. Of these, 200 of them include Beer and Diapers, and of them, 50 transactions also include ice cream. The association rule "IF Beer and Diapers, THEN Ice cream" has the support of 50 transactions (or, alternatively 5% = 50/1000), and a confidence of 25% = 50/200

A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold. This predefined threshold is called the *minimum confidence*. A higher confidence indicates that the rule (X → Y) is more interesting or more trustworthy, based on the sample dataset.

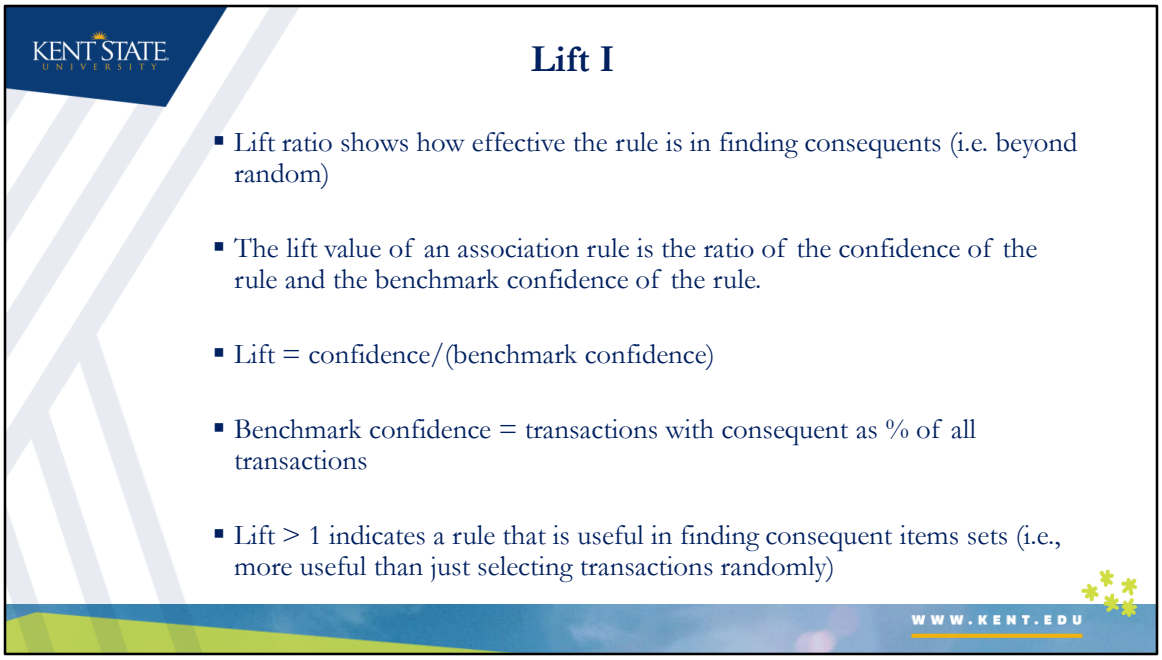
Example Confidence Values

- Transaction data →

t1: Beef, Chicken, Milk
t2: Beef, Cheese
t3: Cheese, Boots
t4: Beef, Chicken, Cheese
t5: Beef, Chicken, Clothes, Cheese, Milk
t6: Chicken, Clothes, Milk
t7: Chicken, Milk, Clothes

- Assume:
  minsup = 30%
  minconf = 80%
- An example frequent itemset:
  {Chicken, Clothes, Milk} [sup = 3/7]
- Association rules from the itemset:
  Clothes → Milk, Chicken [sup = 3/7, conf = 3/3]
  … …
  Clothes, Chicken → Milk, [sup = 3/7, conf = 3/3]

Here is our earlier example now including confidence values. Make sure you are able to calculate the confidence values depicted in the slide.



Even though confidence can identify the interesting rules from all the candidate rules, it comes with a problem. Given rules in the form of X → Y, confidence considers only the antecedent (X) and the co- occurrence of X and Y; it does not take the consequent of the rule (Y) into concern. For example, if nearly all customers buy bananas and nearly all customers buy ice cream, the confidence level of a rule such as "IF bananas THEN ice-cream" will be high regardless of whether there is an association between the items.

Therefore, confidence cannot tell if a rule contains true implication of the relationship or if the rule is purely coincidental. X and Y can be statistically independent yet still receive a high confidence score. Other measures such as lift and leverage are designed to address this issue.

***Lift* measures how many times more often X and Y occur together than expected if they are statistically independent of each other. Lift is a measure of how X and Y are really related rather than coincidentally happening together . This is illustrated in the next slide.**

KENT STATE UNIVERSITY

Lift II

- *Lift* measures how many times more often X and Y occur together than expected if they are statistically independent of each other. Lift is a measure of how X and Y are really related rather than coincidentally happening together

$$\text{Lift}(X \rightarrow Y) = \frac{\text{Support}(X \cap Y)}{\text{Support}(X) * \text{Support}(Y)}$$

- Lift greater than 1 indicates that there is some usefulness to the rule. A larger value of lift suggests a greater strength of the association between X and Y.

WWW.KENT.EDU

For example, **assuming 1,000 transactions, with** {milk,eggs} **appearing in 300 of them,** {milk} **appearing in 500, and** {eggs} **appearing in 400, then *Lift*(*milk* →*eggs*)=0.3/(0.5*0.4)=1.5.**

A lift ratio greater than 1.0 suggests that there is some usefulness to the rule. In other words, the level of association between the antecedent and consequent itemsets is higher than would be expected if they were independent. The larger the lift ratio, the greater the strength of the association.

KENT STATE UNIVERSITY

Example Lift

- An example itemset: {Chicken, Clothes, Milk} [sup = 3/7]

Association rules from the itemset:

1. Clothes → Milk, Chicken [Lift =]
2. Clothes, Chicken → Milk, [Lift =]
3. Chicken → Clothes, Milk [Lift =]

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

WWW.KENT.EDU

Calculate Lift for our previous example.

KENT STATE UNIVERSITY

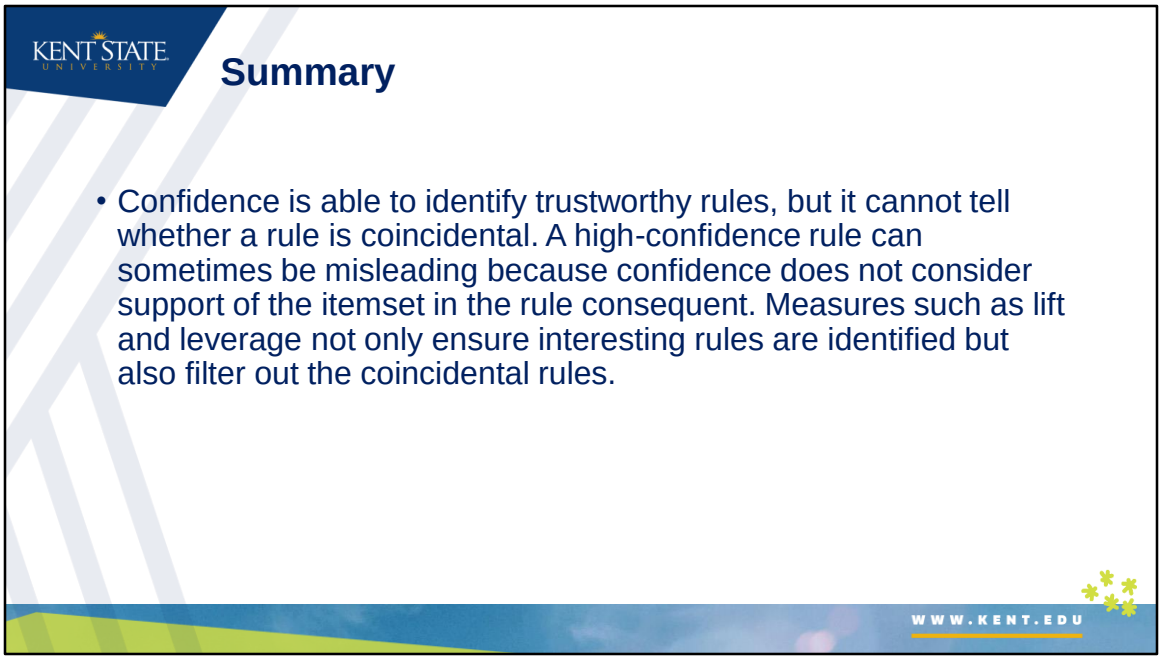
Leverage

- *Leverage* is a similar notion to Lift, but instead of using a ratio, leverage uses the difference. Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.

$$\text{Leverage}(X \rightarrow Y) = \text{Support}(X \cap Y) - \text{Support}(X) * \text{Support}(Y)$$

WWW.KENT.EDU

***Leverage* is a similar notion, but instead of using a ratio, leverage uses the difference. In theory, leverage is 0 when X and Y are statistically independent of each other. If X and Y have some kind of relationship, the leverage would be greater than zero. A larger leverage value indicates a stronger relationship between X and Y. For the previous example, *Leverage*(*milk*→*eggs*)=0.3−(0.5*0.4)=0.1**



This summarizes our measures that can be used to identify rules that provide a strong dependence between antecedent and consequent itemsets.