ASSOCIATION

Courtesy:
Jiawei Han, Micheline Kamber, and Jian Pei
www.techtarget.com, www.kdnuggets.com
webfocusinfocenter.informationbuilders.com
Agarwal et al 1993
Anjali Jivani

BEER AND DIAPERS!!

- Mark Madsen, a research analyst with information management consultancy Third Nature, traced the beer-and-diapers story to its origin in 1992.
- Karen Heath, then an industry consultant with Teradata, now a senior manager for health analytics with Accenture, was attached to an unspecified Midwest retailer.
- She wrote SQL queries that discovered this correlation between beer and diapers.
- Madsen studied this for almost ten years at different places and different stores – WalMart, Target, 7-Eleven,...

BEER AND DIAPERS!!

11,400 academic papers, 14,000 books, ~1,000,000 web pages

“The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and increased sales on peanuts and pretzels by more that 27%."

“For example, one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer.”

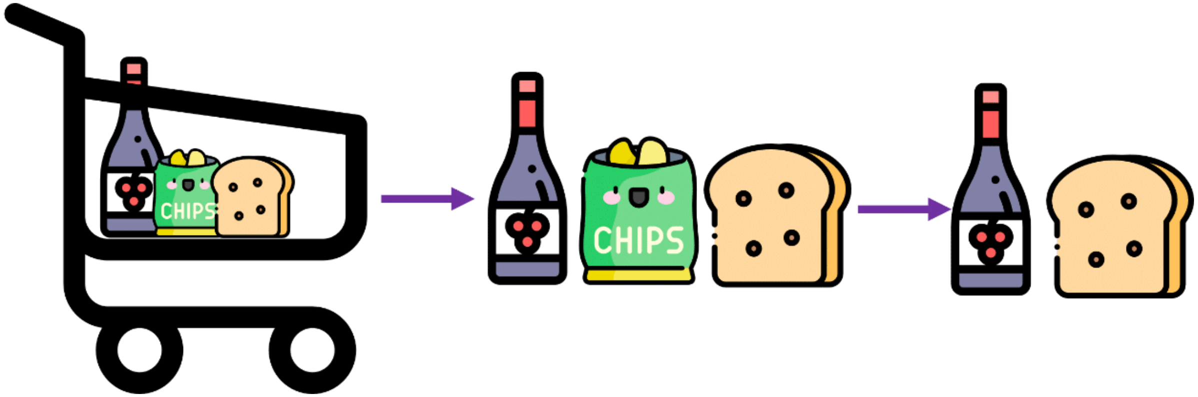
“[UK] Dads with newborns cannot go out to pubs to socialize with friends (me including). So they buy more beers so that they can drink at home!”

“Some time ago, Wal-Mart decided to combine the data from its loyalty card system with that from its PoS…Once combined, the data was mined extensively …On Friday afternoons, young American males who buy diapers (nappies) also have a predisposition to buy beer.”

"Shoppers who buy diapers for the first time at a Tesco store can expect to receive coupons by mail for baby wipes, toys -- and beer. Tesco's analysis showed that new fathers tend to buy more beer because they are home with the baby and can't go to the pub."

"Sometimes the data can throw up surprises: mining of databases held by 7-Eleven stores in the US revealed a link between purchases of beer and nappies. When they were moved together, sales of both increased,”

WHAT IS ASSOCIATION MINING



- Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository.
- The applications of Association Rule Mining are found in Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering and classification.
- The most common approach to find these patterns is Market Basket Analysis, which is a key technique used by large retailers like Amazon, Flipkart, etc to analyze customer buying habits by finding associations between the different items that customers place in their "shopping baskets".

APPLICATIONS OF ASSOCIATION

- The discovery of these associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers. The strategies may include:
  - Changing the store layout according to trends
  - Customer behaviour analysis
  - Catalog design
  - Cross marketing on online stores
  - What are the trending items customers buy
  - Customized emails with add-on sales etc..
- Online retailers and publishers can use this type of analysis to:
  - Inform the placement of content items on their media sites, or products in their catalog
  - Deliver targeted marketing (e.g. emailing customers who bought products specific products with other products and offers on those products that are likely to be interesting to them.)

DIFFERENCE BETWEEN ASSOCIATION AND RECOMMENDATION

- Association rules do not extract an individual's preference, rather find relationships between sets of elements of every distinct transaction. This is what makes them different than Collaborative filtering which is used in recommendation systems.
- To understand it better take a look at next slide snapshot from amazon.com and you notice 2 headings "Frequently Bought Together" and the "Customers who bought this item also bought" on each product's info page.

"Frequently Bought Together" → Association

"Customers who bought this item also bought" → Recommendation

ASSOCIATION AND RECOMMENDATION

THE MARKET BASKET ANALYSIS

- In order to make it easier to understand, think of Market Basket Analysis in terms of shopping at a supermarket.
- Market Basket Analysis takes data at transaction level, which lists all items bought by a customer in a single purchase.
- The technique determines relationships of what products were purchased with which other product(s).
- These relationships are then used to build profiles containing If-Then rules of the items purchased. The rules could be written as: If {A} Then {B}
- The *If* part of the rule (the {A} above) is known as the antecedent and the *THEN* part of the rule is known as the consequent (the {B} above).
- The antecedent is the condition and the consequent is the result.
- The association rule has three measures that express the degree of confidence in the rule, Support, Confidence and Lift.

THE ASSOCIATION RULE

Example Association Rule

90% of transactions that purchase bread and butter also purchase milk

"IF" part = **antecedent**

"THEN" part = **consequent**

"Item set" = the items (e.g., products) comprising the antecedent or consequent

- Antecedent and consequent are *disjoint* (i.e., have no items in common)

Antecedent: bread and butter

Consequent: milk

Confidence factor: 90%

ITEMSETS

- **Itemset**
  A collection of one or more items
  Example: {Milk, Bread, Diaper}
  k-itemset
  An itemset that contains k items
- **Support count (σ)**
  Frequency of occurrence of an itemset:
  E.g. σ({Milk, Bread, Diaper}) = 2
- **Support**
  Fraction of transactions that contain an itemset
  E,g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  An itemset whose support is greater than or equal to a **MinSup** threshold

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

THE MEASURES

Transaction 1	🍎 🍺 🍚 🍗
Transaction 2	🍎 🍺 🍚
Transaction 3	🍎 🍺
Transaction 4	🍎 🍐
Transaction 5	🍼 🍺 🍚 🍗
Transaction 6	🍼 🍺 🍚
Transaction 7	🍼 🍺
Transaction 8	🍼 🍐

Support {🍎} = $\frac{4}{8}$ **i.e. 50%**

Confidence {🍎 → 🍺} = $\frac{\text{Support \{🍎, 🍺\}}}{\text{Support \{🍎\}}} = \frac{3}{4}$ **i.e. 75%**

Lift {🍎 → 🍺} = $\frac{\text{Support \{🍎, 🍺\}}}{\text{Support \{🍎\} x Support \{🍺\}}}$ $= \frac{3/8}{(\frac{4}{8})\mathrm{x}(\frac{6}{8})} = \left(\frac{3}{8}\right) x \left(\frac{8}{4}\right) x \left(\frac{8}{6}\right)$

=1

THE MEASURES

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
  - Example:
    {Milk, Diaper} → {Beer}
- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

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**

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

SUPPORT, CONFIDENCE AND LIFT

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: \; X \Rightarrow Y$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

- Support(A → D) =

$$\frac{Total\ transactions\ with\ A\ and\ D}{Total\ transactions}$$

= 2/5

- Confidence (A → D) =

$$\frac{Total\ transactions\ with\ A\ and\ D}{Total\ transactions\ with\ A}$$

= 2/3

- Lift (A → D) =

$$\frac{Support(A \rightarrow D)}{Support(A) Support(D)}$$

$= (2/5) \,/\, ((3/5)x(3/5)) = 10/9$

SUPPORT, CONFIDENCE AND LIFT

$$Support = \frac{frq(X,Y)}{N}$$

$$Rule: \; X \Rightarrow Y$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

- Confidence (A → D) =

$$\frac{\textit{Total transactions with A and D}}{\textit{Total transactions with A}}$$

= 2/3

- Lift (A → D) =

$$\frac{Support(A \rightarrow D)}{Support(A)Support(D)}$$

$= (2/5) / ((3/5)x(3/5)) = 10/9$

or

$$= \frac{2}{5} \text{ x } \frac{5}{3} \text{ x } \frac{5}{3}$$

$$= \frac{2}{3} \text{ x } \frac{5}{3}$$

= (2/3) / (3/5)

= Conf(A → D) / Supp(D)

SUPPORT AND CONFIDENCE

Support: Its the default popularity of an item. In mathematical terms, the support of item **A** is nothing but the ratio of transactions involving **A** to the total number of transactions.
Support(Grapes) = 4/6 = 0.666

Transaction ID	Grapes	Apple	Mango	Orange
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

Confidence: Likelihood that customer who bought both **A** and **B**. Its divides the number of transactions involving both **A** and **B** by the number of transactions involving **A**.
Confidence({Grapes, Apple}⟶{Mango}) = 2/3, or
Support(Grapes, Apple, Mango)/Support(Grapes, Apple)
= 2/6 / 3/6
= 0.667

LIFT

Transaction ID	Grapes	Apple	Mango	Orange
1	1	1	1	1
2	1	0	1	1
3	0	0	1	1
4	0	1	0	0
5	1	1	1	1
6	1	1	0	1

Lift : Increase in the sale of **A** when you sell **B**.

Lift(A⟶B) = Confidence(A, B) / Support(B)
Lift ({Grapes, Apple}⟶{Mango}) = 1
So, likelihood of a customer buying both **A** and **B** together is 'lift-value' times more than the chance if purchasing alone.
Lift (A⟶B) = 1 means that there is no correlation within the itemset.
Lift (A⟶B) > 1 means that there is a positive correlation within the itemset, i.e., products in the itemset, **A**, and **B**, are more likely to be bought together.
Lift (A⟶B) < 1 means that there is a negative correlation within the itemset, i.e., products in itemset, **A**, and **B**, are unlikely to be bought together.

THE APRIORI ALGORITHM

- **Probably the best known algorithm.**
  Given a set of transactions T, the goal of association rule mining (Apriori) is to find all rules having,
  - **support ≥ *minsup* threshold**
  - **confidence ≥ *minconf* threshold**

Two steps:

1. **Frequent Itemset Generation**
   Generate all itemsets whose support ≥ minsup

2. **Rule Generation**
   Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset. Rules where confidence is ≥ minconf is an interesting rule.

- Frequent itemset generation is still computationally expensive

THE APRIORI ALGORITHM

Frequent Itemset Generation

- Generate all itemsets whose support $\geq$ minsup
- ***if an itemset is frequent, each of its subsets is frequent as well.***
- 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.

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

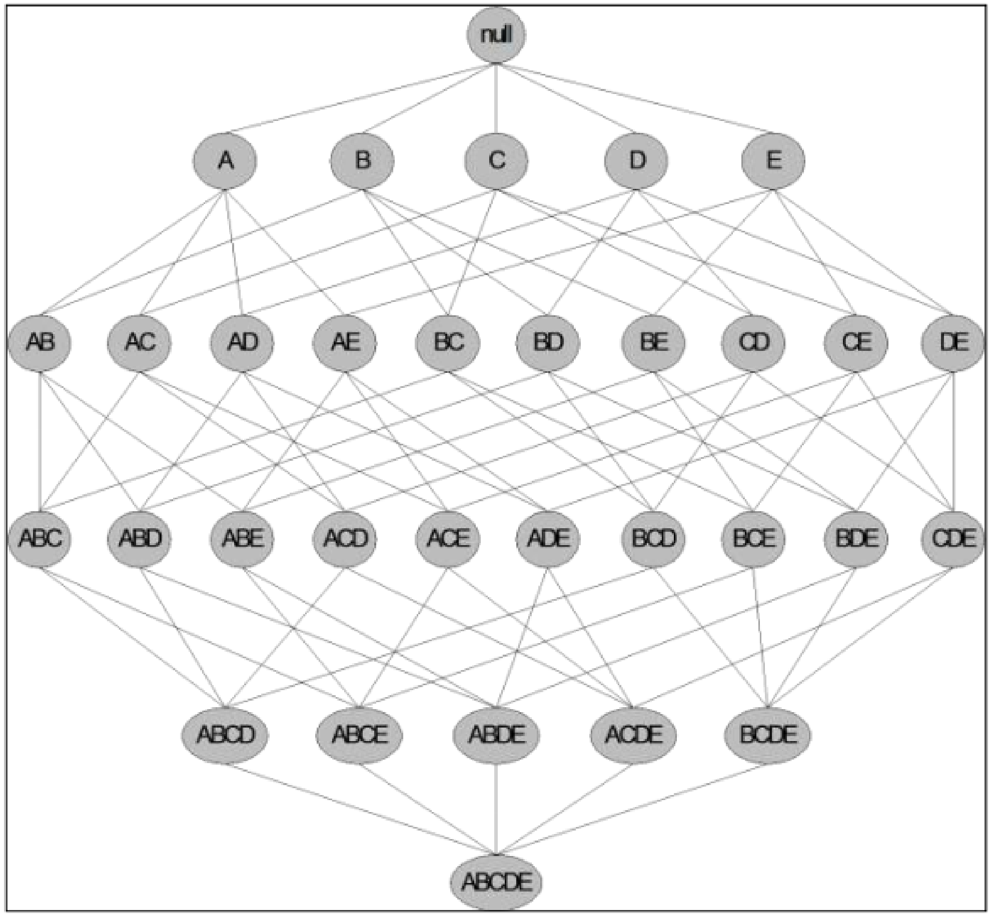
Antimonotonicity:

a) Diapers, Beer – occur 3 times together, then only diaper or only beer will also occur ***atleast 3*** times.

b) Similarly if Milk, Cola – occur only 2 times, then any other item with Milk and Cola will not occur more than 2 times i.e. Milk, Cola, Bread – cannot occur more than 2 times

BRUTE-FORCE

Brute-force approach:



- **Key idea** of the apriori property:
  - any subsets of a frequent itemset are also frequent itemsets (downward closure property).
  - if an itemset is infrequent, all its supersets would also be infrequent.

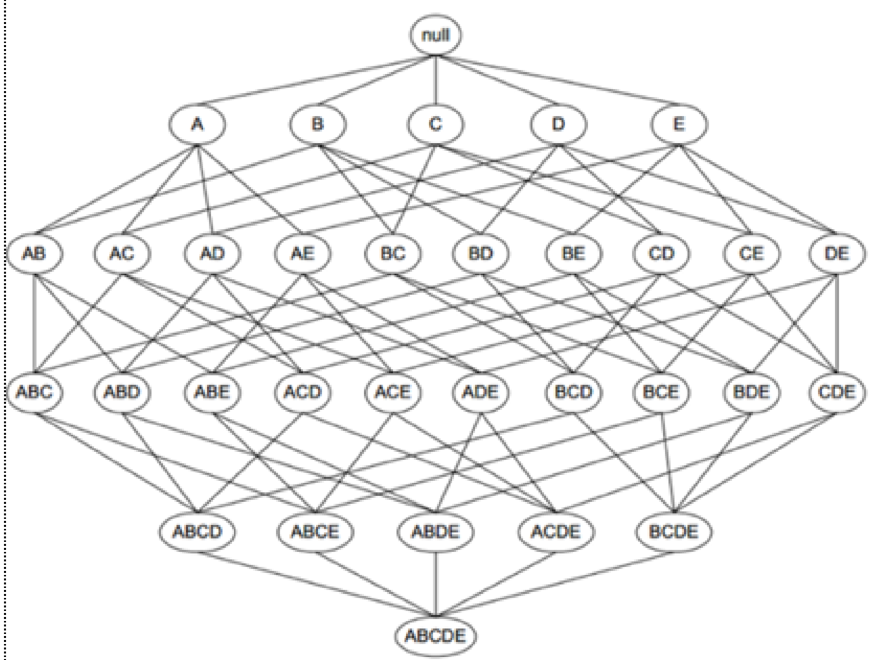
- **Antimonotone property:** If ABC together occur 3 times in a basket, all subsets of ABC will also occur at least 3 times.

- Given d items, there are 2^d-1 possible candidate itemsets.

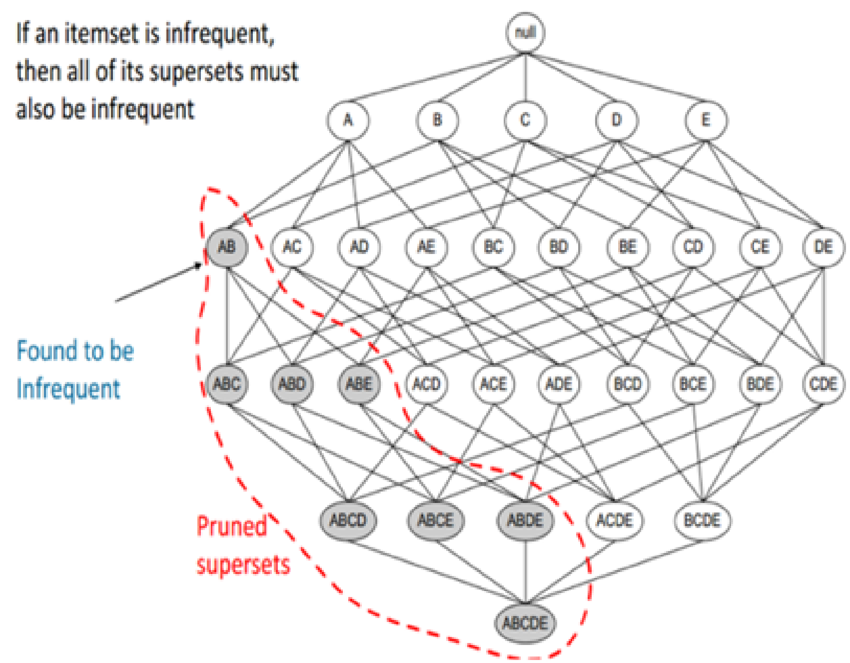
APRIORI - PRUNING

Optimizing Apriori algorithm

The combinations of 5items



The Apriori Algorithm



GENERATING FREQUENT ITEMSETS

minSup count = 3 (60%)

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

Red - Pruning

C1

Item	Count
Bread	4
Milk	4
Diaper	4
Beer	3
Eggs	1
Coke	2

F1

Bread	4
Milk	4
Diaper	4
Beer	3

C2

Bread, Milk	3
Bread, Diaper	3
Bread, Beer	2
Milk, Diaper	3
Milk, Beer	2
Diaper, Beer	3

F2

Bread, Milk	3
Bread, Diaper	3
Milk, Diaper	3
Diaper, Beer	3

C3

Bread, Milk, Diaper	2

GENERATING FREQUENT ITEMSETS

minSup count = 2 (40%)

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

Red - Pruning

C1

Item	Count
Bread	4
Milk	4
Diaper	4
Beer	3
Eggs	1
Coke	2

F1

Bread	4
Milk	4
Diaper	4
Beer	3
Coke	2

C2

Bread, Milk	3
Bread, Diaper	3
Bread, Beer	2
Bread, Coke	1
Milk, Diaper	3
Milk, Beer	2
Milk, Coke	2
Diaper, Beer	3
Diaper, Coke	2
Beer, Coke	1

F2

Bread, Milk	3
Bread, Diaper	3
Bread, Beer	2
Milk, Diaper	3
Milk, Beer	2
Milk, Coke	2
Diaper, Beer	3
Diaper, Coke	2

C3

Bread, Milk, Diaper	2
Bread, Milk, Beer	1
Bread, Diaper, Beer	2
Milk, Coke, Diaper	2
Milk, Diaper, Beer	2

F3

Bread, Milk, Diaper	2
Bread, Diaper, Beer	2
Milk, Coke, Diaper	2
Milk, Diaper, Beer	2

C4 not possible

RULE GENERATION

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

- **Rule Generation**
- Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset.
- Rules where confidence is ≥ minconf is an interesting rule.
- For each frequent itemset *X*,

  For each proper nonempty subset *A* of *X*,

  Let *B* = X - *A*

  A → B is an association rule if,

  Confidence(A → B) ≥ minconf,

  Confidence(A → B) = support(A ∪ B) / support(A)

e.g. X = Milk, Beer, Diaper
Let B = X – Beer
Beer → Milk, Diaper can be a rule

- **Let minConf = 60%**

{Milk, Diaper, Beer} – frequent itemset when minSup = 2 (40%)

Conf({Milk, Diaper} → {Beer}) = (Supp. count of {Milk, Diaper, Beer}) / (Supp. of {Milk, Diaper})

= 2/3

= 0.67

∴ **Conf({Milk, Diaper} → {Beer}) > minConf (Rule is interesting)**

RULE GENERATION

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

MinSup = 40%
MinConf = 60%

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Rules not interesting

Observations:

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- Thus, we may decouple the support and confidence requirements

EXAMPLE

TID	ITEMS
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

MinSup count = 2
MinConf = 70%

Min_Sup_count = 2

Database ***D***

TID	Items
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

Scan ***D*** for count of each candidate →

C1

Itemset	S.C
I1	6
I2	7
I3	6
I4	2
I5	2

Compare candidate support count with minimum support count →

L1

Itemset	S.C
I1	6
I2	7
I3	6
I4	2
I5	2

Generate C2 candidates form L1 →

C2

Itemset
I1,I2
I1,I3
I1,I4
I1,I5
I2,I3
I2,I4
I2,I5
I3,I4
I3,I5
I4,I5

Scan ***D*** for count of each candidate →

Itemset	S.C
I1,I2	4
I1,I3	4
I1,I4	1
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I3,I4	0
I3,I5	1
I4,I5	0

Compare candidate support count with minimum support count →

L2

Itemset	S.C
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2

Generate C3 candidates form L2 →

C3

Itemset
I1,I2,I3
I1,I2,I5

Scan ***D*** for count of each candidate →

C3

Itemset	S.C
I1,I2,I3	2
I1,I2,I5	2

Compare candidate support count with minimum support count →

L3

Itemset	S.C
I1,I2,I3	2
I1,I2,I5	2

RULE GENERATION

TID	ITEMS
T100	I1,I2,I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

MinConf = 70%

Frequent Itemsets:
{I1, I2, I3} and {I1, I2, I5}

1. {I1} → {I2,I3}
   Conf = 2/6 = 33% Rule not interesting

2. {I1,I2} → {I3}
   Conf = 2/4 = 50% Rule not interesting

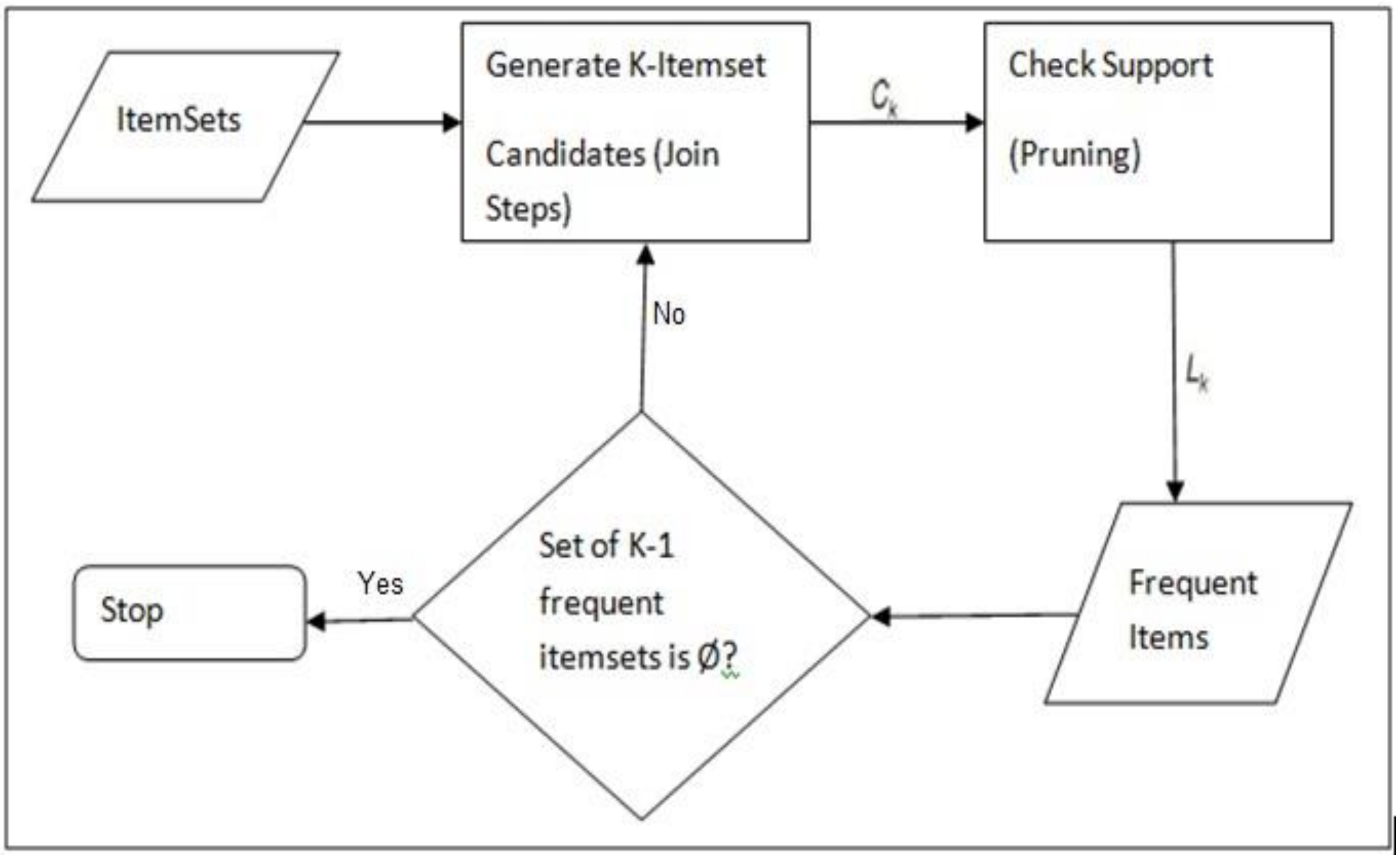
3. {I5} → {I1,I2}
   Conf = 2/2 = 100% **Rule is interesting**

THE MEASURES

Can minimum support and minimum confidence can be automatically determined in mining association rules?

- For the **minimum support**, it all depends on the dataset. Usually, may start with a high value, and then decrease the values until you find a value that will generate enough patterns.
- For the **minimum confidence**, it is a little bit easier because it represents the confidence that you want in the rules. So usually, use something like 60 % . But it also depends on the data.
- In terms of performance, when ***minsup*** is higher you will find **less patterns** and the algorithm is faster.
- For ***minconf***, when it is set higher, there will be less patterns but it may not be faster because many algorithms don't use minconf to prune the search space. So obviously, setting these parameters also depends on how many rules you want.

APRIORI FLOWCHART



PRACTICAL APPLICATIONS OF MBA

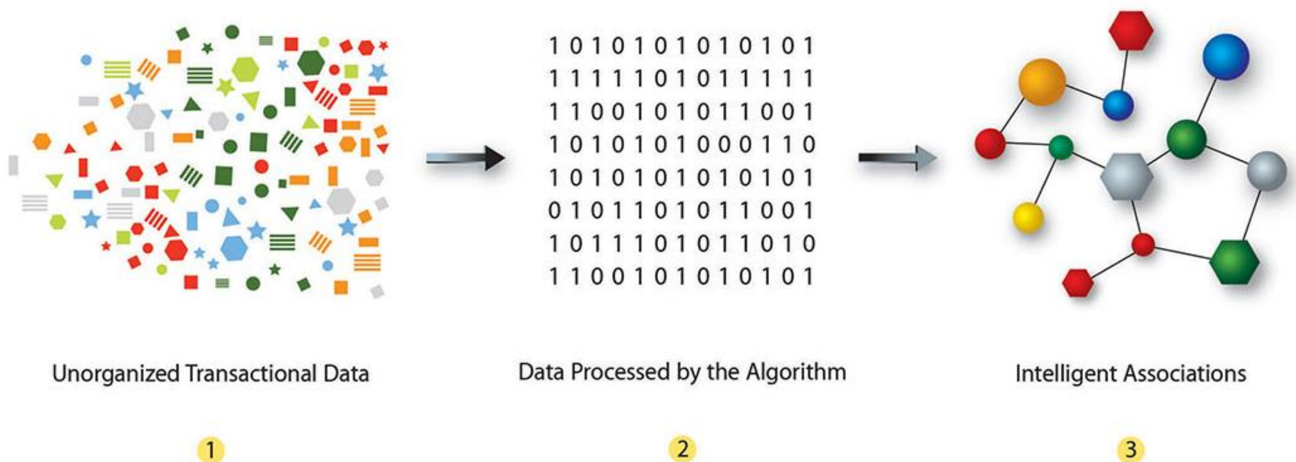
- **Retail.** In Retail, Market Basket Analysis can help determine what items are purchased together, purchased sequentially, and purchased by season. This can assist retailers to determine product placement and promotion optimization (for instance, combining product incentives). Does it make sense to sell soda and chips or soda and crackers?
- **Telecommunications.** In Telecommunications, where high churn rates continue to be a growing concern, Market Basket Analysis can be used to determine what services are being utilized and what packages customers are purchasing. They can use that knowledge to direct marketing efforts at customers who are more likely to follow the same path.
- For instance, Telecommunications these days is also offering TV and Internet. Creating bundles for purchases can be determined from an analysis of what customers purchase, thereby giving the company an idea of how to price the bundles. This analysis might also lead to determining the capacity requirements.

PRACTICAL APPLICATIONS OF MBA

- **Banks.** In Financial (banking for instance), Market Basket Analysis can be used to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities.
- **Insurance.** In Insurance, Market Basket Analysis can be used to build profiles to detect medical insurance claim fraud. By building profiles of claims, you are able to then use the profiles to determine if more than 1 claim belongs to a particular claimee within a specified period of time.
- **Medical.** In Healthcare or Medical, Market Basket Analysis can be used for comorbid conditions and symptom analysis, with which a profile of illness can be better identified. It can also be used to reveal biologically relevant associations between different genes or between environmental effects and gene expression.

Summary

- Association rule mining has been extensively studied in the data mining community.
- There are many efficient algorithms and model variations.
- Other related work includes
- Multi-level or generalized rule mining
- Constrained rule mining
- Incremental rule mining
- Maximal frequent item set mining
- Numeric association rule mining
- Rule interestingness and visualization
- Parallel algorithms
- …



END