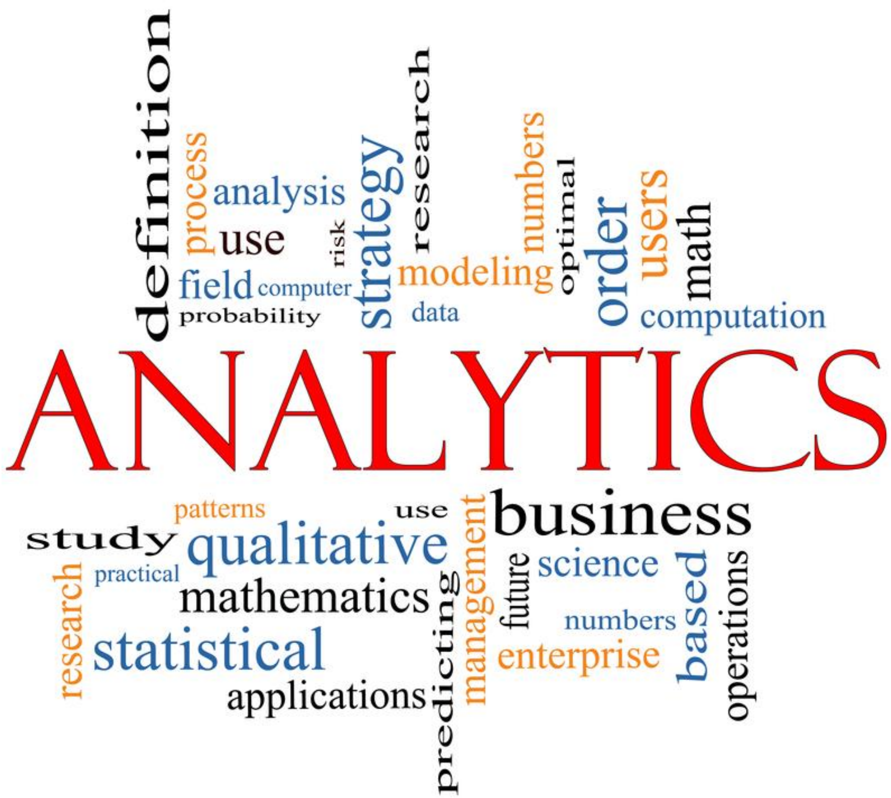
insAnalytics

Consulting. Training. Research & Development.

insAnalytics

Insights.. Through Analytics…



Data Mining

Association Rule Mining

Key Objectives

After successful completion of the topic, participants will be able to:

- Develop awareness of the utility of **Association Rules** in the industry
- Appreciate the **key features** of **Association Rule Mining** and its **applications** across various industries
- Articulate the **key concepts** of Association Rule Mining, such as **Transactions**, **Item Sets** and **Frequent** & **Infrequent** Item Sets
- **Evaluate** Association Rules through various **performance** measures, such as **Support**, **Confidence** and **Lift**
- **Generate** Association Rules with real-world business data using the extremely popular "**apriori**" algorithm

To learn & understand the subject matter better, participants need to be aware of the following areas:

- Descriptive Statistics
- Correlation Analysis
- Statistical Estimation
- Testing of Hypotheses
- Analysis of Variance (ANOVA)
- Linear Regression Analysis
- Data Mining – CRISP DM Methodology
- Data Mining – Data Preparation (Part 1)
- Data Mining – Data Preparation (Part 2)

Association Rule Mining

- What is Association Rule Mining (ARM)?
- Association Rule Mining – Popular Applications
- ARM in action – an illustration from the Retail industry
- Association Rule Mining – Key Concepts & Terminologies
- Popular Algorithms for Association Rule Mining
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Association Rule Mining

- What is Association Rule Mining (ARM)?
- Association Rule Mining – Popular Applications
- ARM in action – an illustration from the Retail industry
- Association Rule Mining – Key Concepts & Terminologies
- Popular Algorithms for Association Rule Mining
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Association Rule Mining

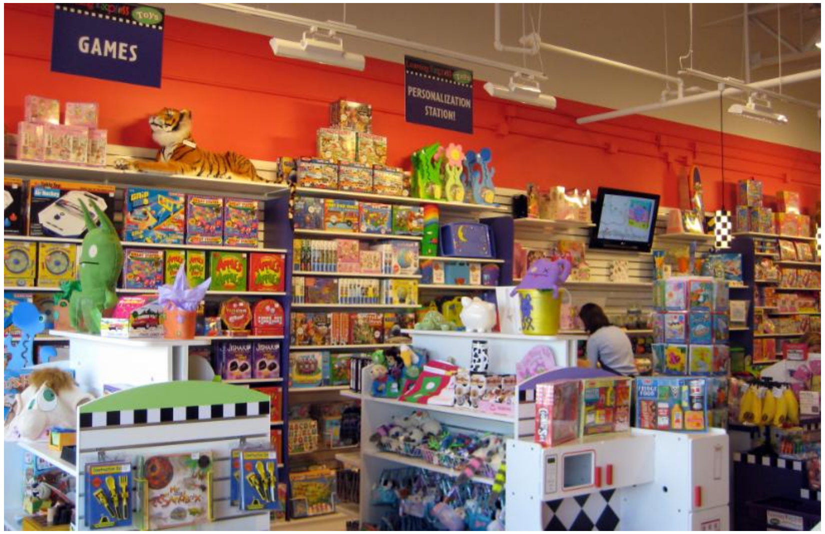
What is Association Rule Mining (ARM)?

- Understanding "Associations" in Data
- Association Rule Mining – the Definition
- Association Rule Mining – finding "patterns" in data
- Association Rule Mining – searching for "co-occurrence" of data

Understanding "Associations" in Data

Have you noticed this in a supermarket?
The "**Kids Toys**" section is just beside the "**Kids Apparel**" section!

Ever wondered why?



Because, analysis has shown that
consumers who buy kids apparel **also** buy toys!!!

In other words, there is an **association** between
purchase of kids apparel and purchase of kids toys

Association Rule Mining – the Definition

The term **“Association Rule”** was first defined in

R Agrawal, T Imielinski, A Swami:
“Mining Association Rules Between Sets of Items in Large Databases”,
SIGMOD Conference 1993: 207-216

Association Rule Mining can be defined as follows:

> *Association Rule Mining is a popular and well-researched class of methods to find patterns in sequential data*

What is Association Rule Mining?

Association Rule Mining – finding "patterns" in data

Association Rule Mining is a popular and well researched **class of methods** to find patterns in sequential data

- Association Rule Mining is a **class of methods**
  Hence, different algorithms are available,
  and many algorithms are **still under development**

What is Association Rule Mining?

Associations – searching for “co-occurrence” of data

Association Rule Mining is a popular and well researched class of methods to find **patterns** in sequential data

- Association Rule Mining is a class of methods
  Hence, different algorithms are available,
  and many algorithms are still under development
- Association Rule Mining finds **patterns**
  of **co-occurrence** of categorical data

Associations – searching for “co-occurrence” of data

Association Rule Mining is a popular and well researched class of methods to find patterns in **sequential** data

- Association Rule Mining is a class of methods
  Hence, different algorithms are available,
  and many algorithms are still under development
- Association Rule Mining finds patterns
  of co-occurrence of categorical data
- The data needs to be **sequential**
  For instance,
  products purchased one after the other,
  web-pages visited one after the other,
  words appearing one after the other, etc

Association Rule Mining

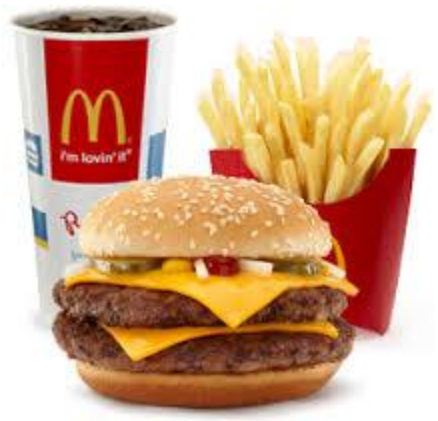
Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- **Association Rule Mining – Popular Applications**
  - Popular ARM Applications – Product Bundling
  - Popular ARM Applications – Retail Store Outlay
  - Popular ARM Applications – Bioinformatics
  - Popular ARM Applications – Text Mining
  - Popular ARM Applications – Web Analytics
- ARM in action – an illu...
- Association Rule Minin...
- Popular Algorithms for...
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Popular ARM Applications – Product Bundling

While snacking at a **Mc Donald's** outlet,
you browse through the menu card

You notice that Mc Donald's offers
a number of "**combo meals**", combining multiple items



How did Mc Donald's arrive at these particular **combinations** of items?

Popular ARM Applications – Retail Store Outlay

In your neighborhood departmental store:

- Breakfast items are kept close to health drinks
- Formal ties are displayed just beside the formal trousers section
- Discounted floor mattress are displayed just beside bedroom items

How did the departmental store management decide on the **product placements**?

Due to its frequent usage in the **Retail** industry,
Association Rule Mining is commonly referred to as **Market Basket Analysis**

Popular ARM Applications – Bioinformatics

In **Bioinformatics** studies,
scientists often need to profile the **genetic mapping**

Researchers study to detect
which parts of the **genetic sequence** are alike
and which parts are different

Such analyses are known as **Sequence Analysis**

Association Rule Mining can be applied to find the relevance between:

- Two different genetic sequences
- The genetic sequence and medical diseases
- The genetic sequence and the environmental effect, etc

Popular ARM Applications – Text Mining

Maruti Suzuki has recently launched a new car in the sedan segment

Maruti Suzuki wants to understand
the general feedback of the car vis-à-vis its competitors

They want to know how the car is being compared with other sedans

What does Maruti Suzuki have to do?

Maruti Suzuki will study the **co-occurrence**
of the name of the car with other sedan names
in social media, blogs, etc

Popular ARM Applications – Web Analytics

Having watched a certain video clip on **YouTube**,
the next time you log in to the website,
you find a section labelled "**Recommended for You**"

How did YouTube **build** this recommendation for you?

While visiting the profile of a particular individual on **LinkedIn**,
you find a section titled "**People Also Viewed**"

How did LinkedIn **choose** the profiles displayed in this selection?

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- **ARM in action – an illustration from the Retail industry**
- Association Rule Mining – Key Concepts & Terminologies
- Popular Algorithms for Association Rule Mining
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- **ARM in action – an illustration from the Retail industry**
  - Setting the Context – a real-world illustration
  - Setting the Context – Item Sets & Transactions
  - Setting the Context – Defining an Association Rule
- Association Rule Mining – Key Concepts & Terminologies
- Popular Algorithms for ARM
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Setting the Context – a real-world illustration

In a **Retail store**, there are multiple items available
Suppose there are **"m"** items in the Retail store

Let's define **I** as the set of **all items**:
$\mathbf{I = \{i_1, i_2, i_3, \ldots , i_m\}}$

For a Retail store, **I** can be as follows:
I = {milk, bread, jam, biscuits, cookies, butter, cereals, ...}

Setting the Context – Item Sets & Transactions

Every time a customer makes a **transaction**,
he/ she purchases a **subset** of **I**

For instance, Customer 1 purchases {milk, bread, cookies},
while Customer 2 purchases {bread, milk}
These items have to come from **I**

Hence, any transaction **"t"** consists of **one or more items** from **I**
In other words,
$\mathbf{t \subset I}$

Setting the Context – Item Sets & Transactions

A transactional database, "**T**", contains all such transactions "**t**":

$\mathbf{T = \{t_1, t_2, t_3, \ldots., t_n\}}$

Examples of different transactions might look like:

t_1 = {milk, bread, cookies}

t_2 = {egg, milk, bread}

t_3 = {milk, bread, jam, cereals}

…

Setting the Context – Defining an Association Rule

Suppose **X** & **Y** are two different **Item Sets**
For instance, X = {bread, milk} and Y = {jam, cereals}

Now, **X** $\subset$ **I** and **Y** $\subset$ **I**

An **Association Rule** X → Y means:
If a customer purchases **{bread, milk}**,
he/ she is **most likely** to purchase **{jam, cereals}** as well

Setting the Context – Defining an Association Rule

In general, if X & Y are two different **Item Sets**,
i.e., $\mathbf{X \subset I}$ and $\mathbf{Y \subset I}$

then, an **Association Rule** is an **implication** of the form:
$\mathbf{X \rightarrow Y}$, where X & $Y \subset I$ and $X \cap Y = \phi$

In other words, we state that:
the purchase of X "**implies**" the purchase of Y

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- **Association Rule Mining – Key Concepts & Terminologies**
- Popular Algorithms for Association Rule Mining
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- **Association Rule Mining – Key Concepts & Terminologies**
  - Support of an Item Set
  - Confidence of an Association Rule
  - Lift of an Association Rule
  - Support Thresholds & Confidence Thresholds
  - ARM Rule Sets – How many Association Rules do we need?
- Popular Algorithms for
- The Apriori Algorithm
- The Apriori Algorithm

Item Sets & Transactions

Consider the following **10 transactions** in a retail store:

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

From the above table, we find that the **first customer** bought **{milk, cookies}**

Support of an Item Set

The **Support** of an **Item Set** is defined as the **percentage** of transactions where the **Item Set** has **"occurred"**

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

In the adjoining transaction database, **{milk}** has occurred in transaction IDs 1, 2, 3, 5, 8 & 10 (i.e., six out of ten) Hence, the **Support** of {milk} is **60%**

Similarly, **{bread, cookies}** has a **Support** of **30%**, since it has occurred in transaction IDs 2, 6 & 7 (i.e., three out of ten)

Mathematically, the **Support** of an **Item Set** X is defined as:

$$Support\ (X) = \frac{Count\ of\ transactions\ (t)\ where\ X \subseteq t}{Count\ of\ transactions\ (t)}$$

Confidence of an Association Rule

If X and Y are two different **Item Sets**, then the **Confidence** of the **Association Rule** X → Y is defined as the ratio of the **Support** of (XUY) to the **Support** of X

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

Illustration 1:

X = {milk}, Y = {bread}

Confidence (Y | X)

= 20% / 60% = 33%

Illustration 2:

X = {milk}, Y = {cookies}

Confidence (Y | X)

= 30% / 60% = 50%

Mathematically, the **Confidence** of an **Association Rule** X → Y is defined as:

$$Confidence\ (Y \mid X) = \frac{Support\ (X \cup Y)}{Support\ (X)} = \frac{Count\ of\ transactions\ (t)\ where\ (X \cup Y) \subseteq t}{Count\ of\ transactions\ (t)\ where\ X \subseteq t}$$

Confidence of an Association Rule

Confidence (Y|X) is **different** from Confidence (X|Y)
The two values are not necessarily the same

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

Illustration 1:
X = {milk}, Y = {bread}
Confidence (Y|X)
= 20% / 60% = 33%
Confidence (X|Y)
= 20% / 50% = 40%

Illustration 2:
X = {milk}, Y = {cookies}
Confidence (Y|X)
= 30% / 60% = 50%
Confidence (X|Y)
= 30% / 60% = 50%

Lift of an Association Rule

If X → Y be an **Association Rule** with **Confidence (Y|X)**, then the **Lift** of the **Association Rule** X → Y is defined as the ratio of the **Confidence** (Y|X) to the **Support** of Item Set Y

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

Illustration:

X = {milk}

Y = {bread}

Lift (X → Y)

= Confidence (Y|X) / Support (Y)

= 33% / 50%

= 67%

Please note that:

Lift (X → Y) = **Lift (Y → X)**

Mathematically, the **Lift** of an **Association Rule** X → Y is defined as:

$$Lift\ (X \rightarrow Y) = \frac{Confidence\ (Y \mid X)}{Support\ (Y)} = \frac{Support\ (X \cup Y)}{Support\ (X) \times Support\ (Y)}$$

Support Thresholds & Confidence Thresholds

The **ultimate objective** of an **Association Rule Mining** exercise is to identify the "**useful**" Association Rules

We recognize an **Association Rule** X → Y as **significant** (i.e., "**useful**") if:

- ❑ Both X and Y have **Support** greater than a **threshold** value
- ❑ X → Y has **Confidence** greater than a **threshold** value

Rule	Supp(X)	Supp(Y)	Confidence	Lift
Cereal →Bread	20%	60%	0%	0%
Milk → Chocolates	60%	60%	50%	83%
{Bread, Cookies} →Milk	30%	60%	33%	55%
{Bread, Milk} → Chocolates	20%	60%	100%	167%

Support Thresholds & Confidence Thresholds

If we set **30%** as the **threshold** value for **Support** and **Confidence**, then

- **Cereal → Bread** is **not** a useful Association Rule, as both Support and Confidence are **low**
- **{Bread, Milk} → Chocolate** is **not** a useful Association Rule, since Support is **low**, though Confidence is 100%
- **Milk → Chocolate** and **{Bread, Cookies} → Milk** are both **useful** Association Rules

Rule	Supp(X)	Supp(Y)	Confidence	Lift
Cereal →Bread	20%	60%	0%	0%
Milk → Chocolates	60%	60%	50%	83%
{Bread, Cookies} →Milk	30%	60%	33%	55%
{Bread, Milk} → Chocolates	20%	60%	100%	167%

ARM Rule Sets – How many Association Rules do we need?

In the previous illustration, there were **5 items**

How many **Item Sets** can be generated from 5 items? … **31 Item Sets**

How many **Association Rules** can be generated from the 31 Item Sets? ... **210** !!!

As the number of items increases,
the possible set of Association Rules increases **rapidly**

# of Items	# of Item Sets	# of Association Rules
2	3	2
3	7	12
4	15	50
5	31	210

ARM Rule Sets – How many Association Rules do we need?

In general, if there are "**k**" items in a retail store, this leads to:

Number of possible **Item Sets**:

$(2^k - 1)$

Number of possible **Association Rules**:

$$\sum_{i=1}^{\left[\frac{k}{2}\right]} \sum_{j \leq (k-i)} \left(2 \times \binom{k}{i} \times \binom{k-i}{j}\right)$$

Now, imagine a real-world retail store

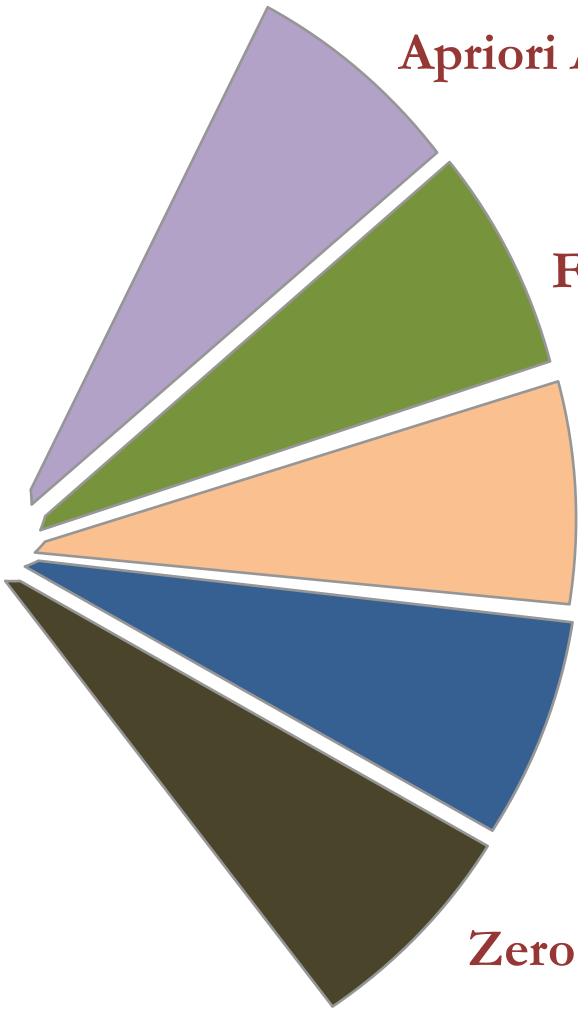
How many items does it keep?

Hence, we need an **algorithm** to identify the **significant** Association Rules

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- Association Rule Mining – Key Concepts & Terminologies ✓
- **Popular Algorithms for Association Rule Mining**
- The Apriori Algorithm for ARM – Key Concepts
- The Apriori Algorithm – Discovering the Association Rules

A few popular ARM Algorithms



- Apriori Algorithm
- Frequent Pattern (FP) Growth Algorithm
- Similis Algorithm
- One Attribute Rule
- Zero Attribute Rule

Exercise

The following table gives the comments of 10 different individuals about a newly launched novel.

S. No.	Comments
1	You would love to read this book, and would be closer to your heart thereafter.
2	From starting till end, the enthusiasm remains the same. Good one and would recommend it.
3	This is a good book.
4	Filmy storyline but flow of writing is at its best as usual.
5	It is a good book. Just buy and enjoy the story.
6	Message of the story is good and also the finishing was touching.
7	Good book. I love this book very much.
8	Okay to read once. Not as interesting as his other books.
9	I recommend this book at least once for my friends It not only explores a true love but honesty as well.
10	A very good and interesting novel to pass time with. Read the whole novel in just two days.

Suppose: X= {good}, Y = {Read}, Z = {Book}, W = {Love}, then find

- Support of X, Y, Z, and W
- Conf(Z|X), Conf(X|Z), Conf(Y|X), Conf(X|Y)

Association Rule Mining

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- Association Rule Mining – Key Concepts & Terminologies ✓
- Popular Algorithms for Association Rule Mining ✓
- **The Apriori Algorithm for ARM – Key Concepts**
  - Frequent Item Sets versus Infrequent Item Sets
  - Apriori Algorithm – the Postulates
  - Implementing the Apriori Algorithm – Process Flow
  - Apriori Algorithm – Process Flow Summary
  - Apriori Algorithm – Merits & Demerits
- The Apriori Algorithm – Discovering the Association Rules

Frequent Item Sets versus Infrequent Item Sets

If an **Item Set** under consideration has a **Support** level that is **more** than a pre-decided **"cut-off"** value (i.e., **"threshold"**), we label the Item Set as a **Frequent Item Set**

Otherwise, we label the Item Set as an **Infrequent Item Set**

In the previous illustration, we set the **"cut-off"** value as **30%**

Hence, the **Frequent Item Sets** and the **Infrequent Item Sets** are as tabulated below:

Item Sets	Support	Frequent/Infrequent
{Bread}	50%	Frequent
{Cereals}	20%	Infrequent
{Cookies}	60%	Frequent
{Bread, Chocolates}	50%	Frequent
{Cereals, Milk}	20%	Infrequent
{Bread, Chocolates, Cookies}	30%	Frequent

Apriori Algorithm – the Postulates

The basic "**postulates**" used in the Apriori Algorithm are the following:

- **Any subset of a Frequent Item Set is itself a Frequent Item Set**
- Any superset of an Infrequent Item Set is also an Infrequent Item Set

Item Sets	Support	Frequent/Infrequent
{Bread}	**50%**	**Frequent**
{Cereals}	20%	Infrequent
{Cookies}	60%	Frequent
{Bread, Chocolates}	**50%**	**Frequent**
{Cereals, Milk}	20%	Infrequent
{Bread, Chocolates, Cookies}	**30%**	**Frequent**

{Bread, Chocolates, Cookies} is frequent.
Hence, all its subsets are frequent.
For example, {Bread}, {Bread, Chocolates} are frequent.

Apriori Algorithm – the Postulates

The basic "**postulates**" used in the Apriori Algorithm are the following:

- Any subset of a Frequent Item Set is itself a Frequent Item Set
- **Any superset of an Infrequent Item Set is also an Infrequent Item Set**

Item Sets	Support	Frequent/Infrequent
{Bread}	50%	Frequent
{Cereals}	**20%**	**Infrequent**
{Cookies}	60%	Frequent
{Bread, Chocolates}	50%	Frequent
{Cereals, Milk}	**20%**	**Infrequent**
{Bread, Chocolates, Cookies}	30%	Frequent

{Cereals} is infrequent. Hence, all its supersets are frequent.
For example, {Milk, Cereals} is infrequent.

Process Flow – Frequent Item Sets & Candidate Sets

1. Decide on the **cut-off** values (i.e., **thresholds**) for **Support** and **Confidence**
   Let's label these **MinSupport** and **MinConfidence**, respectively

2. Define **L_k**: Set of **Frequent Item Sets** of **size k**
   (i.e., with **Support > MinSupport**)

3. Define **C_k**: Set of **Candidate Item Sets** of **size k**
   (i.e., **potentially** Frequent Item Sets)

Process Flow – Getting started (the first Frequent Item Set)

1. Decide on the cut-off values (i.e., thresholds) for Support and Confidence
   Let's label these MinSupport and MinConfidence, respectively
2. Define L_k: Set of Frequent Item Sets of size k
   (i.e., with Support > MinSupport)
3. Define C_k: Set of Candidate Item Sets of size k
   (i.e., potentially Frequent Item Sets)
4. Start with **k=1**, and find $\mathbf{L_1}$ from $\mathbf{C_1}$

Process Flow – the "Join" step

1. Decide on the cut-off values (i.e., thresholds) for Support and Confidence
   Let's label these MinSupport and MinConfidence, respectively
2. Define L_k: Set of Frequent Item Sets of size k
   (i.e., with Support > MinSupport)
3. Define C_k: Set of Candidate Item Sets of size k
   (i.e., potentially Frequent Item Sets)
4. Start with k=1, and find L_1 from C_1
5. Create $\mathbf{C_{k+1}}$ from $\mathbf{L_k}$ (this step is called the "**Join**" step)
6. **Increment** k by **one**, i.e., **k = k+1**

Process Flow – the "Prune" step

1. Decide on the cut-off values (i.e., thresholds) for Support and Confidence
   Let's label these MinSupport and MinConfidence, respectively
2. Define L_k: Set of Frequent Item Sets of size k
   (i.e., with Support > MinSupport)
3. Define C_k: Set of Candidate Item Sets of size k
   (i.e., potentially Frequent Item Sets)
4. Start with k=1, and find L_1 from C_1
5. Create C_{k+1} from L_k (this step is called the "Join" step)
6. Increment k by one, i.e., k = k+1
7. If C_k is **non-empty**, then calculate L_k (this step is called the **"Prune"** step),
   and **go back** to **Step 5**

Process Flow – Stopping Rule (empty Candidate Set)

1. Decide on the cut-off values (i.e., thresholds) for Support and Confidence
   Let's label these MinSupport and MinConfidence, respectively
2. Define L_k: Set of Frequent Item Sets of size k
   (i.e., with Support > MinSupport)
3. Define C_k: Set of Candidate Item Sets of size k
   (i.e., potentially Frequent Item Sets)
4. Start with k=1, and find L_1 from C_1
5. Create C_{k+1} from L_k (this step is called the "Join" step)
6. Increment k by one, i.e., k = k+1
7. If C_k is non-empty, then calculate L_k (this step is called the "Prune" step), and go back to Step 5
8. If **C_k** is **empty**, then **stop** and **proceed** to **Step 9**

Process Flow – Frequent Item Sets & Confidence Thresholds

1. Decide on the cut-off values (i.e., thresholds) for Support and Confidence
   Let's label these MinSupport and MinConfidence, respectively
2. Define L_k: Set of Frequent Item Sets of size k
   (i.e., with Support > MinSupport)
3. Define C_k: Set of Candidate Item Sets of size k
   (i.e., potentially Frequent Item Sets)
4. Start with k=1, and find L_1 from C_1
5. Create C_{k+1} from L_k (this step is called the "Join" step)
6. Increment k by one, i.e., k = k+1
7. If C_k is non-empty, then calculate L_k (this step is called the "Prune" step),
   and go back to Step 5
8. If C_k is empty, then stop and proceed to Step 9
9. From the set of Frequent Item Sets (i.e., $\mathbf{L_{k-1}}$),
   select the **Association Rules** that have **Confidence > MinConfidence**

Apriori Algorithm – Process Flow Summary

Step 1
- Set k = 1
- Find the Support (S) for each Item Set of the Candidate Set C_1
- Arrive at L_1

Step 2: <u>The "Join" Step</u>
- Join L_k with itself to generate the Candidate Set C_{k+1} with Item Sets of (k+1) items
- Set k = k+1

Step 3: <u>The "Prune" Step</u>
- Find the Support for each Item Set of the Candidate Set C_k
- Arrive at L_k where Support is greater than MinSupport

Step 4
Is the Candidate Set C_k empty?

NO

YES

Step 5: <u>Generate Association Rules</u>
- For each Frequent Item Set F, generate all non-empty subsets of F
- Label the set of these subsets as G

Step 6
- For each element of G, calculate the Confidence (C) of the Association Rule "G => (F-G)"
- If Confidence (C) > MinConfidence, then "G => (F-G)" is a significant Association Rule

The Apriori Algorithm for ARM – Key Concepts

Apriori Algorithm – Merits & Demerits

- The "**oldest**" ARM algorithm
- The **most popular** ARM algorithm
  - Implemented in most **Data Mining** tools
- Easy to **implement**, and easy to **understand** & **explain**
- Probably the "**best known**" ARM algorithm
- Implemented in all leading **software**

- Generates a **huge** number of **Candidate Sets**
- **Scans** through the transactional database a **number of times**
- Too many **iterations** & **scans** makes the algorithm **expensive**
- Often turns out to be **less efficient** in identifying **Association Rules** involving a **large** number of items

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- Association Rule Mining – Key Concepts & Terminologies ✓
- Popular Algorithms for Association Rule Mining ✓
- The Apriori Algorithm for ARM – Key Concepts ✓
- **The Apriori Algorithm – Discovering the Association Rules**

InsAnalytics

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- Association Rule Mining – Key Concepts & Terminologies ✓
- Popular Algorithms for Association Rule Mining ✓
- The Apriori Algorithm for ARM – Key Concepts ✓
- **The Apriori Algorithm – Discovering the Association Rules**
  - Getting started – the first Frequent Item Set
  - Apply "Join" & "Prune" – the 2nd & the 3rd Frequent Item Sets
  - Stopping Rule – empty Candidate Set
  - Selecting the Frequent Item Sets – Confidence Thresholds
  - Association Rules uncovered – the final "significant" list

Getting started – the first Frequent Item Set

Let's revisit the previous illustration with 10 transactions and with **MinSupport**=30%

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

First, we create $\mathbf{C_1}$ and $\mathbf{L_1}$

C_1

1 Item Set	Support Count
Cereals (Cr)	2
Milk (M)	6
Breads(B)	5
Cookies(Ck)	6
Chocolates (Ch)	6

L_1

1 Item Set
Milk (M)
Breads(B)
Cookies(Ck)
Chocolates (Ch)

Apply "Join" & "Prune" – the second Frequent Item Set

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

From $\mathbf{L_1}$, we apply the "**Join**" step to create $\mathbf{C_2}$

Then, we apply the "**Prune**" step to create $\mathbf{L_2}$

1 Item Set
Milk (M)
Breads(B)
Cookies(Ck)
Chocolates (Ch)

L_1

Join

2 Item Set	Support Count
M,B	2
M, Ck	3
M, Ch	3
B, Ck	3
B, Ch	4
Ck, Ch	3

C_2

Prune

2 Item Set
M, Ck
M, Ch
B, Ck
B, Ch
Ck, Ch

L_2

Apply “Join” & “Prune” – the third Frequent Item Set

Transaction ID	Cereals	Milk	Breads	Cookies	Chocolates
1	0	1	0	1	0
2	0	1	1	1	1
3	0	1	0	1	0
4	0	0	1	0	1
5	1	1	0	0	0
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	0	0	1
9	0	0	0	1	0
10	0	1	1	0	1

From $\mathbf{L_2}$, we apply the **“Join”** step to create $\mathbf{C_3}$

Then, we apply the **“Prune”** step to create $\mathbf{L_3}$

L_2

2 Item Set
M, Ck
M, Ch
B, Ck
B, Ch
Ck, Ch

Join

C_3

2 Item Set	Support Count
M, Ch, Ck	1
M, B, Ch	2
B, Ck, Ch	3

Prune

L_3

2 Item Set
B, Ck, Ch

Stopping Rule – empty Candidate Set

Since L_3 has just **one element**, therefore C_4 is **empty**
Hence, we **stop** the iteration

We calculate all the **subsets** of L_3, i.e., all the subsets of **{B, Ck, Ch}**:
{B}, {Ck}, {Ch}, {B, Ck}, {B, Ch}, {Ck, Ch}, {B, Ck, Ch}

As L_3=**{B, Ck, Ch}** is a **Frequent Item Set**,
therefore, all the above **subsets** are **Frequent Item Sets** as well

Now, we can proceed to calculate the **Confidence** levels for
all the possible **Association Rules** obtained from these **Frequent Item Sets**

Selecting the Frequent Item Sets – Confidence Thresholds

With **MinConfidence**=60%, we select the following **Association Rules**:

Rule	Support	Confidence	Action	Lift
Bread →Chocolates	**50%**	**100%**	**Accept**	**167%**
Chocolates → Bread	**60%**	**83%**	**Accept**	**167%**
Bread →Cookies	50%	60%	Accept	100%
Cookies→ Bread	60%	50%	Reject	100%
Bread → {Chocolates, Cookies}	**50%**	**60%**	**Accept**	**200%**
{Chocolates, Cookies} →Bread	**30%**	**100%**	**Accept**	**200%**
Chocolates→ {Bread , Cookies}	60%	50%	Reject	167%
{Bread, Cookies} →Chocolates	**30%**	**100%**	**Accept**	**167%**
Cookies→ {Bread , Chocolates}	60%	50%	Reject	100%
{Bread, Chocolates} →Cookies	**50%**	**60%**	**Accept**	**100%**

Selecting the Frequent Item Sets – Confidence Thresholds

Consider the following two **Association Rules**:

- Bread → Chocolates
- Chocolates → Bread

The former has higher **Confidence**
Hence, we **accept** Bread → Chocolates as an Association Rule
and **drop** Chocolates → Bread

Similarly,
{Chocolates, Cookies} → Bread is **accepted** as an **Association Rule**,
while Bread → {Chocolates, Cookies} is **dropped**

Association Rules uncovered – the final “significant” list

Therefore, we finally end up with the following **Association Rules**:

1. Bread → Chocolates
2. {Chocolates, Cookies} → Bread
3. {Bread, Cookies} → Chocolates
4. {Bread, Chocolates} → Cookies

Association Rule Mining

- What is Association Rule Mining (ARM)? ✓
- Association Rule Mining – Popular Applications ✓
- ARM in action – an illustration from the Retail industry ✓
- Association Rule Mining – Key Concepts & Terminologies ✓
- Popular Algorithms for Association Rule Mining ✓
- The Apriori Algorithm for ARM – Key Concepts ✓
- The Apriori Algorithm – Discovering the Association Rules ✓

Recapitulation & Key Takeaways

- **Association Rule Mining** is a popular and well researched class of methods to find **patterns** of **"co-occurrence"** in sequential data
- Association Rule Mining is **used widely across industries** for product bundling, designing retail-store outlays, bioinformatics, text mining, web analytics, etc
- Due to its **frequent usage** in the **Retail** industry, Association Rule Mining is commonly referred to as **Market Basket Analysis**
- A typical **Retail outlet** will store a number of items
  - A **purchase transaction** by a customer involves buying a subset of these items
  - A set of items purchased by a customer constitutes an **Item Set**
- If X and Y be two Item Sets purchased by a customer, then an **Association Rule** of the form **X → Y** means that if a customer **buys X**, then she is **very likely** to **buy Y** as well
- While a **large number** of Association Rules can be generated for a typical retail outlet, only a **few** of these eventually turn out to be **"significant"**, i.e., useful
- Real-world businesses would deploy an **algorithm** to identify the useful (i.e., significant) Association Rules based on **performance measures** like **Support** of an Item Set, **Confidence** of an Association Rule, and **Lift** of an Association Rule
- **Significance** of the Association Rules generated would be based on the levels of these performance measures against **pre-defined benchmarks** called **Support thresholds** and **Confidence thresholds**
- **Various algorithms** are available for Association Rule Mining, among which a very popular one is the **"apriori"** algorithm

Data Mining

Association Rule Mining

insAnalytics

Consulting. Training. Research & Development.

Insights.. Through Analytics…