ASSOCIATION RULE MINING

Renato R. Maaliw III, *DIT*
College of Engineering & ITSSO
Southern Luzon State University
Lucban, Quezon, Philippines

In the 1990s, a major retail chain wanted to better understand customer behavior and optimize their store layouts.

They analyzed their sales data, hoping to uncover patterns that could lead to strategic improvements.

During this analysis, they made an intriguing discovery: there was a strong correlation between the purchase of two items.

Cognate/Professional Electives


Initially, this correlation seemed puzzling.

These two items are not typically associated with each other, and their connection was not immediately obvious.

Cognate/Professional Electives

Beer and diapers are not typically associated with each other, and their connection was not immediately obvious.

However, upon further investigation and analysis, the retailer found a compelling explanation.

They examined the shopping patterns of their customers and identified a specific demographic that was responsible for this correlation: young fathers.

These fathers were often responsible for picking up diapers on their way home from work.

Armed with this newfound insight, the retailer recognized a significant business opportunity.

They decided to leverage the correlation between beer and diapers by strategically placing the products in proximity to each other within their stores.

Beer displays were positioned near the diaper section, making it more convenient for customers to find and purchase both items together.

Association Rule Mining (ARM)

- A technique used to uncover interesting relationships, patterns, or associations within large datasets.

- Applied in fields such as retail and e-commerce to help understand customer purchasing patterns, detecting anomalies, and making strategic business decisions.

Basic Concepts of ARM

- For instance, in a supermarket, ARM might reveal that "customers who buy bread and butter often also buy milk"

- If **{bread, butter}** then **{milk}**

- These rules reveal dependencies and correlations among items, often helping organizations in cross-selling, recommendation systems, and inventory management

Key Terminologies

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Itemset:

A collection of one or more items:

{Bread}

{Bread, Milk}

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

K-Itemset:

An itemset containing exactly k-items:

{Bread, Milk} is a 2-itemset

{Bread, Milk, Diaper} is a 3-itemset

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Support Count (σ):

The number of transactions that include a particular itemset. For example, if *{Bread, Milk} appears in 3 out of 5 transactions, its* ***support count*** *is 3.*

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Support (s):

The proportion of transactions containing itemset, calculated as support count / total transactions.

For *{Milk, Bread}* the **support** would be 3 / 5 = 0.6.

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Frequent Itemset:

An itemset whose support meets or exceeds a user-defined minimum support threshold.

If minimum support threshold is set to **0.6**, then *{Milk, Bread}* is a frequent **itemset**.

Associaton Rules

An Association Rule

- It is an implication in the form of ***X*** → ***Y***, where:

  X and ***Y*** are itemsets, with ***X*** as the “antecedent” (if – part) and ***Y*** as the “consequent” (then – part)

- Example: *{Milk, Diaper}* → *{Beer}*

  meaning if *Milk* and *Diaper* are purchased, *Beer* is also likely to be purchased

Metrics for Rule Evaluation

Support (s)

- How often ***X*** and ***Y*** occur together, the probability of both itemsets appearing in the same transactions

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Support for *{Milk, Diaper, Beer}*

***Support** = {Number of transactions containing {Milk, Diaper, Beer} / (Total Transactions)*

From the table, 2 / 5 = ***0.4***

Confidence

- Confidence for the rule *{Milk, Diaper}* → *{Beer}*

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

$$Confidence(\{Milk, Diaper\} \rightarrow \{Beer\}) = \frac{Support(Milk, Diaper, Beer) = (0.4)}{Support(Milk, Diaper) = (0.6)}$$

Support*({Milk, Diaper})* is the number of transactions containing both *Milk and Diaper*, which appears in 3 transactions (3, 4, 5), 3 / 5 = **0.6**

Confidence = 0.4 / 0.6 = 0.67

Lift

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

- Measures how much likely it is to see:

  {Beer} with *{Milk, Diaper}* than it would be to see *{Beer}* itself.

$$Lift(Milk, Diaper \rightarrow Beer) = \frac{Support(Milk, Diaper, Beer)}{Support(Milk, Diaper) \ * Support(Beer)}$$

2 / 5 = 0.4

Lift

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

- Measures how much likely it is to see:

  {Beer} with *{Milk, Diaper}* than it would be to see *{Beer}* itself.

$$Lift(Milk, Diaper \rightarrow Beer) = \frac{Support(Milk, Diaper, Beer)}{Support(Milk, Diaper) \ * Support(Beer)}$$

3 / 5 = 0.6

Lift

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

- Measures how much likely it is to see:

  {Beer} with *{Milk, Diaper}* than it would be to see *{Beer}* itself.

$$Lift(Milk, Diaper \rightarrow Beer) = \frac{Support(Milk, Diaper, Beer)}{Support(Milk, Diaper) \ * Support(Beer)}$$

3 / 5 = 0.6

Lift

- Measures how much likely it is to see:

  {Beer} with *{Milk, Diaper}* than it would be to see *{Beer}* itself.

Transactions
Bread, Milk
Bread, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

$$Lift(Milk, Diaper \rightarrow Beer) = \frac{Support(Milk, Diaper, Beer)}{Support(Milk, Diaper) * Support(Beer)}$$

2 / 5 = 0.4

3 / 5 = 0.6 **3 / 5 = 0.6**

Lift = 0.4 / (0.6) * (0.6) = **1.11**

Interpretation:

- **Support** of 0.40 indicates that *{Milk, Diaper, Beer}* appears 40% of all transactions.

- **Confidence** of 0.67 suggests that 67% of transactions containing *{Milk, Diaper}* also contain *{Beer}*

- **Lift** of 1.11 means that transactions with *{Milk, Diaper}* are 1.11 times more likely to include *{Beer}* than by random chance.

Interpreting Lift Values (L > 1)

- A lift of **greater than 1** indicates that items in the antecedent and consequent appear together more frequently than would be expected by random chance.

- **Higher values signify a stronger association:**
  1.1 to 1.5: Weak positive association
  1.6 to 2.0: Moderate positive association
  > 2.0: Strong positive association

Interpreting Lift Values (L = 1)

- A lift of **exactly 1** indicates that the antecedent and consequent are statistically independent. The presence of one item does not affect the likelihood of the other appearing.

Interpreting Lift Values (L < 1)

- A lift of **less than 1** indicates that the presence of the antecedent actually makes the consequent less likely to occur in the same transaction.

- **Values further below 1 imply a stronger negative association:**
  0.75 to 1.0: Weak negative association
  0.5 to 0.74: Moderate negative association
  > 0.5: Strong negative association

The Apriori Principle

To reduce compuational effort, the Apriori principle states:

- “If an itemset is frequent, all its subsets must also be frequent”

- This allows us to prune (ignore) itemsets that have infrequent subsets, reducing thee number of itemsets we need to consider

Rule Generation

Once **frequent itemsets** are identified, rules are generated by partitioning the itemset:

For example, from the itemset *{Milk, Diaper, Beer},* possible rules include:

- *{Milk, Diaper}* → *{Beer}*
- *{Diaper, Beer}* → *{Milk}*

**** Only rules that meet the confidence threshold are retained***

Apriori Principle in Action

Suppose we have a transaction database with five transactions and a minimum support threshold of **3** (i.e., an itemset needs to appear in at least 3 transactions to be considered **frequent**).

Step 1: Identify Frequent 1-Itemsets

Count each item individually across transactions

Bread appears in 4 transactions (1, 2, 4, 5) ➔ **frequent**
Milk appears in 4 transactions (1, 3, 4, 5) ➔ **frequent**
Diaper appears in 4 transactions (2, 3, 4, 5) ➔ **frequent**
Beer appears in 3 transactions (2, 3, 4) ➔ **frequent**
Coke appears in 2 transactions (3, 5) ➔ **not frequent**
Eggs appears in 1 transaction (2) ➔ **not frequent**

Based on the minimum support threshold, only *{Bread, Milk, Diaper, Beer}* are frequent 1-itemsets.
We discard {Coke} and {Eggs} from further consideration.11

Step 2: Identify Frequent 2-Itemsets

Form 2-itemsets using only the frequent 1-itemsets: {Bread, Milk, Diaper, Beer}.

{Bread, Milk} appears in 3 transactions (1, 4, 5) ➔ **frequent**
{Bread, Diaper} appears in 3 transactions (2, 4, 5) ➔ **frequent**
{Bread, Beer} appears in 2 transactions (2, 4) ➔ **not frequent** (pruned)
{Milk, Diaper} appears in 3 transactions (3, 4, 5) ➔ **frequent**
{Milk, Beer} appears in 2 transactions (3, 4) ➔ **not frequent** (pruned)
{Diaper, Beer} appears in 3 transactions (2, 3, 4) ➔ **frequent**

Using the Apriori Principle, we ignore any 3-itemsets that include pruned 2-itemsets *{Bread, Beer}* and *{Milk, Beer}* because they contain infrequent subsets.

Step 3: Generate Candidate 3-Itemsets

We form 3-itemsets only from combinations of the **frequent 2-itemsets**.

{Bread, Milk, Diaper} appears in 3 transactions (4, 5) ➔ **frequent**
{Milk, Diaper, Beer} appears in 2 transactions (3, 4) ➔ **not frequent** (pruned)
{Bread, Diaper, Beer} appears in 2 transactions (2, 4) ➔ **not frequent** (pruned)

Because *{Milk, Diaper, Beer}* and *{Bread, Diaper, Beer}* are infrequent, we don't consider any further supersets of these itemsets.

Pruning Summary

Apriori Principle allows us to skip evaluating itemsets with infrequent subsets. For example, we didn't evaluate *{Bread, Beer}* further because *{Bread, Beer}* itself was infrequent.

Resulting Frequent Itemsets:

1-itemsets: *{Bread}, {Milk}, {Diaper}, {Beer}*
2-itemsets: *{Bread, Milk}, {Bread, Diaper}, {Milk, Diaper}, {Diaper, Beer}*
3-itemsets: *{Bread, Milk, Diaper}*

This example clearly shows how the Apriori Principle helps reduce the number of calculations by **eliminating candidates with infrequent subsets** early, making the algorithm more efficient.

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

01. What is the support of itemset {A, B}?

A. 2/5 = 40%
B. 3/5 = 60%
C. 4/5 = 80%
D. 5/5 = 100%

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

01. What is the support of itemset {A, B}?

A. 2/5 = 40%
B. 3/5 = 60%
C. 4/5 = 80%
D. 5/5 = 100%

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

1. What is the support of itemset {A, B}?

A. 2/5 = 40%
B. 3/5 = 60%
C. 4/5 = 80%
D. 5/5 = 100%

*** {A, B} appears in T1, T2, T5 → 3 occurrences; support = 3/5 = 60%**

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

02. What is the confidence of rule {A} → {C}

A. 50%
B. 60%
C. 70%
D. 75%

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

02. What is the confidence of rule {A} → {C}

A. 50%
B. 60%
C. 70%
D. 75%

- Support(A, C) = 3 (T1, T3, T5)
- Support(A) = 4 (T1, T2, T3, T5)
- Confidence = ¾ = 75%

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

03. What is the lift of rule {A} → {C}

A. 0.75
B. 0.80
C. 0.94
D. 1.20

Given the following dataset:

T1: {A, B, C}
T2: {A, B}
T3: {A, C}
T4: {B, C}
T5: {A, B, C}

03. What is the lift of rule {A} → {C}

A. 0.75
B. 0.80
C. 0.94
D. 1.20

$$Lift(A \rightarrow C) = \frac{Support(A, C) = \frac{3}{5} = 0.6}{Support(A) * Support(C)}$$

0.8   0.8

0.9375

04. Given supports:

s(A,B) = 0.50

s(A,C) = 0.40

s(B,C) = 0.45

s(A,B,C) = 0.35

With min_conf = 0.80, which rule(s) from {A, B, C} pass?

A. Only {A,B} → C
B. Only {A,C} → B
C. Only {B,C} → A
D. All three

04. Given supports:

s(A,B) = 0.50

s(A,C) = 0.40

s(B,C) = 0.45

s(A,B,C) = 0.35

Compute each confidence:

1. {A, B} → C = conf({A,B} → {C}) = s(A,B,C) = 0.35 / 0.50 = 0.70
2. {A,C} → B = conf({A,C} → {B}) = s(A,B,C) = 0.35 / 0.40 = 0.875
3. {B,C} → A = conf({B,C} → {A}) = s(A,B,C) = 0.35 / 0.45 = 0.77

05. What is the primary goal of the Apriori algorithm

A. To cluster data points
B. To calculate the distances between data points
C. To find frequent item sets
D. To find associated items based on preset labels

05. What is the primary goal of the Apriori algorithm

A. To cluster data points
B. To calculate the distances between data points
C. To find frequent item sets
D. To find associated items based on preset labels

07. What is the main input required for the Apriori algorithm

A. Boolean data
B. Transactional dataset
C. Matrix data
D. Predefined data

07. What is the main input required for the Apriori algorithm

A. Boolean data
B. Transactional dataset
C. Matrix data
D. Predefined data

08. What is the confidence metric in Apriori

A. The probability that an item appears in the dataset
B. The difference between support and lift
C. The overall quality of association in the dataset
D. The likelihood of B item occurring given that A has occured

08. What is the confidence metric in Apriori

A. The probability that an item appears in the dataset
B. The difference between support and lift
C. The overall quality of association in the dataset
D. The likelihood of B item occurring given that A has occurred

09. In which area is the Apriori algorithm most commonly applied

A. Natural language processing
B. Image processing
C. Cluster processing
D. Market basket analysis

09. In which area is the Apriori algorithm most commonly applied

A. Natural language processing
B. Image processing
C. Cluster processing
D. Market basket analysis

10. If the minimum support threshold is set too high, what is likely to happen?

A. Too many frequent itemset will be generated
B. No frequent itemsets will be generated
C. Only high frequent itemsets will be identified
D. The algorithm will stop working

10. If the minimum support threshold is set too high, what is likely to happen?

A. Too many frequent itemset will be generated
B. No frequent itemsets will be generated
C. Only high frequent itemsets will be identified
D. The algorithm will stop working

Thank you very much for listening.