Association Analysis

Data Mining

Bidur Devkota, PhD

Gandaki College of Engineering and Science

Pokhara, Nepal

Association Analysis

Popular story about using data mining to identify a relation between sales of **beer and diapers**?

In the early 1990s, a group of data analysts working for a large grocery chain in the United States stumbled upon an intriguing pattern in their sales data. Here's how it unfolded:

The Discovery:
The team noticed that on ***Friday evenings, there was a significant increase in the sales of both diapers and beer.***

Initially, this seemed counterintuitive. Why would these seemingly unrelated products be purchased together?

https://t.ly/cnHZU

Association Analysis

Popular story about using data mining to identify a relation between sales of **beer and diapers**?

Their Findings was as follows:
New fathers often made late-night diaper runs for their infants.

Leveraging the findings :
Armed with this insight, the grocery chain decided to experiment.

They strategically placed beer near the diaper aisle to encourage this combined purchase behaviour.

The result? A ***boost*** *in sales for both products.*

https://t.ly/cnHZU

Association Analysis

- Uncovering Hidden Patterns: Association Rule Mining & Market Basket Analysis
- What is it?
  - A rule-based machine learning method to discover interesting relationships between variables in large databases.
- Popular Analogy:
  - Market Basket Analysis (MBA) – Figuring out what products customers buy together.
- Core Question:
  - "If a customer buys item X, how likely are they to also buy item Y?"
- Goal:
  - Translate vast transactional data into actionable business intelligence.

Association Analysis

- Association Analysis has been extensively studied in the data mining community.
- What is Association?
  - causal connection (as per the interest, behavior, activity, purpose, etc)
- What is association Analysis/Mining?
- Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories
- Applications
  - Basket basket analysis
  - Cross-marketing, ….

Association Analysis

- **Itemset**: A collection of items (e.g., {Diapers, Beer, Chips}).
- **Association Rule**: An implication of the form X → Y (e.g., "If Diapers, then Beer").
- **Support**(X): How frequently does the itemset appear?
  - Support = Freq(X, Y) / Total Transactions
  - "What % of all transactions contain both diapers and beer?"
- **Confidence**(X → Y): How reliable is the rule?
  - Confidence = Support(X, Y) / Support(X)
  - "When someone buys diapers, what % of the time do they also buy beer?"
- **Lift**(X → Y): How much more likely is Y bought with X?
  - Lift = Support(X, Y) / (Support(X) * Support(Y))
  - Lift = 1: No association. Lift > 1: Positive association. Lift < 1: Negative association.

Association Analysis

- Finding Rules Efficiently: The Apriori Algorithm
- Steps:
  - Step 1: Scan transactions, find Frequent Itemsets (meet min. Support).
  - Step 2: Join frequent itemsets to form candidates for larger itemsets.
  - Step 3: Prune candidates using the Apriori Principle: "All subsets of a frequent itemset must also be frequent."
  - Repeat Steps 2 & 3 until no new frequent itemsets.
  - Generate Rules from frequent itemsets that meet min. Confidence & Lift.
- Basic Apriori Principle:
  - If {Diapers, Beer, Chips} is frequent, then {Diapers, Beer} must also be frequent. This reduces the search space.

Association Analysis

Retail Store Sales Data Analysis

- Question: **Which products are frequently bought together to improve shelf placement and marketing?**

Transaction_ID	Customer	Product	Quantity	Date
1	C001	Bread	1	2025-01-01
2	C002	Bread	1	2025-01-02
3	C002	Butter	1	2025-01-02
4	C003	Milk	?	2025-01-03
5	C004	Bread	1	2025-01-04
6	C004	Butter	1	2025-01-04
7	C004	Jam	1	2025-01-04
8	C002	Butter	1	2025-01-02

Retail Store Sales Data Analysis

- Question: **Which products are frequently bought together to improve shelf placement and marketing?**

Transaction_ID	Customer	Product	Quantity	Date	Remarks
1	C001	Bread	1	2025-01-01	–
2	C002	Bread	1	2025-01-02	–
3	C002	Butter	1	2025-01-02	–
4	C003	Milk	?	2025-01-03	Missing Quantity
5	C004	Bread	1	2025-01-04	–
6	C004	Butter	1	2025-01-04	–
7	C004	Jam	1	2025-01-04	–
8	C002	Butter	1	2025-01-02	Duplicate Entry

Retail Store Sales Data Analysis

- **Data Cleaning:**
  - Remove noise, fix errors, handle missing values
    - Remove duplicate (Transaction 8)
    - Fill missing quantity (Transaction 4 → 1)
  - Output:
    - Cleaned table with consistent, valid data

Transaction_ID	Customer	Product	Quantity	Date
1	C001	Bread	1	2025-01-01
2	C002	Bread	1	2025-01-02
3	C002	Butter	1	2025-01-02
4	C003	Milk	1	2025-01-03
5	C004	Bread	1	2025-01-04
6	C004	Butter	1	2025-01-04
7	C004	Jam	1	2025-01-04

Retail Store Sales Data Analysis

- **Data Integration:**
  - Combine data from multiple sources
    - Merge customer data (demographics, region) with transaction table
  - Output:
    - Unified dataset with fields like Customer_Age, Location

Transaction_ID	Customer	Age	Address	Product	Quantity	Date
1	C001	52	Pokhara	Bread	1	2025-01-01
2	C002	53	Kathmandu	Bread	1	2025-01-02
3	C002	53	Kathmandu	Butter	1	2025-01-02
4	C003	33	Dharan	Milk	**1**	2025-01-03
5	C004	54	Dharan	Bread	1	2025-01-04
6	C004	67	Birgunj	Butter	1	2025-01-04
7	C004	22	Pokhara	Jam	1	2025-01-04

Retail Store Sales Data Analysis

- **Data Selection:**
  - Select relevant subset for analysis
    - Choose only columns: Customer, Product, Quantity
  - Output:
    - Reduced table focused on purchase patterns

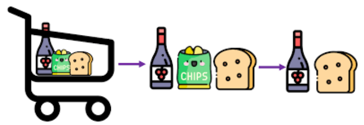
Customer	Product	Quantity
C001	Bread	1
C002	Bread	1
C002	Butter	1
C003	Milk	**1**
C004	Bread	1
C004	Butter	1
C004	Jam	1

Retail Store Sales Data Analysis

- **Data Transformation:**
  - Convert or encode data into useful form
    - Convert transactions into “basket format” per customer:
      - C001 → {Bread}
      - C002 → {Bread, Butter}
      - C004 → {Bread, Butter, Jam}
  - Output:
    - Transaction list ready for pattern mining

Retail Store Sales Data Analysis

- **Data Mining:**
  - Extract useful patterns or associations
    - Apply **Apriori** algorithm → find frequent itemsets
  - Output:
    - Pattern found: {**Bread**, **Butter**} appears frequently
      - C001 → {Bread}
      - C002 → {**Bread, Butter**}
      - C004 → {**Bread, Butter**, Jam}



https://t.ly/T55x7

Retail Store Sales Data Analysis

- **Pattern Evaluation:**
  - Identify meaningful, interesting patterns
    - Evaluate rule confidence:
      - Bread → Butter (Support = 50%, Confidence = 80%)
  - Output:
    - Keep only high-confidence patterns

Retail Store Sales Data Analysis

- **Knowledge Presentation:**
  - Present results in user-friendly format
    - Create bar chart of frequent itemsets or rule list
  - Output:
    - Insights: **“Customers who buy bread often buy butter.”**

Rule Extracted:
If a customer buys **Bread**, they are likely to buy **Butter**.
Support: 50%, **Confidence:** 80%

- **Business Use:**
  - Place bread and butter together on shelves.

Apriori Algorithm - Support and Confidence

Transaction ID	Items Purchased
T1	Bread, Butter, Milk
T2	Bread, Butter
T3	Bread, Milk
T4	Butter, Milk
T5	Bread, Butter, Jam

Find Frequent Itemsets

Items Purchased

Itemset	Occurrence	Support = (Count / Total Txns)
{Bread}	4	4/5 = **0.8**
{Butter}	4	4/5 = **0.8**
{Milk}	3	3/5 = **0.6**
{Bread, Butter}	3	3/5 = **0.6**
{Bread, Milk}	2	2/5 = **0.4**
{Butter, Milk}	2	2/5 = **0.4**
{Bread, Butter, Milk}	**1**	**0.2**

Frequent 2 Itemset

Frequent Itemset

Minimum Support Threshold = 0.4 → keep itemsets ≥ 0.4

Frequent Itemsets → {Bread}, {Butter}, {Milk}, {Bread Butter}, {Bread Milk}, {Butter Milk},

Apriori Algorithm - Support and Confidence

Frequent 2 - Itemsets

Itemset	Occurrence	Support = (Count / Total Txns)
{Bread, Butter}	3	3/5 = **0.6**
{Bread, Milk}	2	2/5 = **0.4**
{Butter, Milk}	2	2/5 = **0.4**

make rules from the frequent 2-itemsets

Rule: Bread → Butter

Support(Bread → Butter) = Support({Bread, Butter}) = **3/5 = 0.6**

Confidence(Bread → Butter) = Support({Bread, Butter}) / Support({Bread})

= (3/5) / (4/5)

= 3/4

= **0.75**

Rule: Butter → Bread

Support({Butter, Bread}) = 3/5 = 0.6

Confidence(Butter → Bread) = Support({Bread, Butter}) / Support({Butter})

= (3/5) / (4/5)

= 3/4

= **0.75**

Apriori Algorithm - Support and Confidence

- **Support** tells *how popular a combination* is in the whole store.
- **Confidence** tells *how reliable the rule* is once the left-hand item is bought.
- With support = 0.6 and confidence = 0.75, the store can:
  - put **Bread** and **Butter** together,
  - or offer a "*buy bread, get butter 30% off*" promo.

References

- Tan, P.N., Steinbach, M. and Kumar, V.,, 2006. Introduction to data mining. Pearson Education, Inc. 3.
- Han, J., Kamber, M. and Mining, D., 2006. Concepts and techniques. Morgan Kaufmann

- Acknowledge the assistance of LLMs like Gemini(Google),DeepSeek, ChatGPT(OpenAI) for helping to refine/generate some content for this work.