Chapter 04: Pattern Mining: Basic Concepts and Methods

Problem Statement

Transaction Database

The table below lists the 8 transactions in our database.

TID	**Items Bought**
T100	{Laptop, Mouse, Headphone, Printer}
T200	{Laptop, Mouse, Pendrive, Headphone}
T300	{Laptop, iPhone, Charger, Headphone}
T400	{Printer, Ink, Pendrive, Mouse}
T500	{iPhone, Charger, Smartwatch, Headphone}
T600	{Laptop, Printer, Pendrive, Mouse}
T700	{Laptop, iPhone, Mouse, Headphone}
T800	{Laptop, Mouse, Printer, Pendrive}

- **Minimum Support (*min_sup*):** 40%
- **Minimum Confidence (*min_conf*):** 70%

First, we calculate the absolute minimum support count required for an itemset to be considered frequent.

$$\text{Minimum Support Count} = \lceil \mathit{min\_sup} \times \text{Total Transactions} \rceil = \lceil 0.40 \times 8 \rceil = \lceil 3.2 \rceil = 4$$

Any itemset must appear in at least **4 transactions** to be considered frequent.

Execution of the Apriori Algorithm

The algorithm proceeds iteratively, generating candidate itemsets of size k and pruning them to find frequent itemsets of size k.

Pass 1: Finding Frequent 1-Itemsets (L_1)

First, we scan the database to count the occurrences of each individual item. This generates the candidate 1-itemset, C_1.

Item (Candidate C_1)	**Support Count**	**Frequent?**
Laptop	6	Yes
Mouse	6	Yes
Headphone	5	Yes
Printer	4	Yes
Pendrive	4	Yes
iPhone	3	No
Charger	2	No
Ink	1	No
Smartwatch	1	No

We prune C_1 by keeping only the items with a support count ≥ 4. This gives us the frequent 1-itemset, L_1.

$$L_1 = \{\text{Laptop, Mouse, Headphone, Printer, Pendrive}\}$$

Pass 2: Finding Frequent 2-Itemsets (L_2)

Next, we generate candidate 2-itemsets, C_2, by joining L_1 with itself. We then scan the database again to count the support for each candidate pair.

Itemset (Candidate C_2)	**Support Count**	**Frequent?**
{Laptop, Mouse}	5	Yes
{Laptop, Headphone}	4	Yes
{Mouse, Printer}	4	Yes
{Mouse, Pendrive}	4	Yes
{Laptop, Printer}	3	No
{Laptop, Pendrive}	3	No
{Mouse, Headphone}	3	No
{Printer, Pendrive}	3	No
{Headphone, Printer}	1	No
{Headphone, Pendrive}	1	No

After pruning C_2, we obtain the frequent 2-itemset, L_2.

$$L_2 = \{\{\text{Laptop, Mouse}\}, \{\text{Laptop, Headphone}\}, \{\text{Mouse, Printer}\}, \{\text{Mouse, Pendrive}\}\}$$

Pass 3: Finding Frequent 3-Itemsets (L_3)

To generate candidate 3-itemsets (C_3), we join L_2 with itself. The **Apriori principle** states that for an itemset to be frequent, all of its subsets must also be frequent. We use this to prune candidates.

- Join {Laptop, Mouse} and {Laptop, Headphone} → Candidate {Laptop, Mouse, Headphone}.
- **Pruning Check:** The subsets are {Laptop, Mouse}, {Laptop, Headphone}, and {Mouse, Headphone}.
- We check if all subsets are in L_2. Since {Mouse, Headphone} is not in L_2 (its support count is 3), the candidate {Laptop, Mouse, Headphone} is pruned *before* scanning the database.

All other potential candidates generated from L_2 are similarly pruned. Therefore, no candidate 3-itemsets survive.

$$C_3 = \emptyset \implies L_3 = \emptyset$$

The algorithm terminates as no further frequent itemsets can be generated.

Generating Strong Association Rules

We now generate association rules from the frequent itemsets found (L_2) that satisfy the minimum confidence of 70%. The confidence is calculated as:

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support\_Count}(A \cup B)}{\text{Support\_Count}(A)}$$

Rule	**Confidence Calculation**	**Confidence**	**Strong?**
From {Laptop, Mouse}, support=5			
Laptop ⇒ Mouse	5/6	83.3%	**Yes**
Mouse ⇒ Laptop	5/6	83.3%	**Yes**
From {Laptop, Headphone}, support=4			
Laptop ⇒ Headphone	4/6	66.7%	No
Headphone ⇒ Laptop	4/5	80.0%	**Yes**
From {Mouse, Printer}, support=4			
Mouse ⇒ Printer	4/6	66.7%	No
Printer ⇒ Mouse	4/4	100.0%	**Yes**
From {Mouse, Pendrive}, support=4			
Mouse ⇒ Pendrive	4/6	66.7%	No
Pendrive ⇒ Mouse	4/4	100.0%	**Yes**

The algorithm successfully identified all frequent itemsets and derived the strong association rules based on the given thresholds.

Final Frequent Itemsets

- **1-itemsets (L_1):** Laptop (6), Mouse (6), Headphone (5), Printer (4), Pendrive (4)
- **2-itemsets (L_2):** {Laptop, Mouse} (5), {Laptop, Headphone} (4), {Mouse, Printer} (4), {Mouse, Pendrive} (4)

Final Strong Association Rules

1. Laptop ⇒ Mouse (Confidence: 83.3%)
2. Mouse ⇒ Laptop (Confidence: 83.3%)
3. Headphone ⇒ Laptop (Confidence: 80.0%)
4. **Printer ⇒ Mouse (Confidence: 100.0%)**
5. **Pendrive ⇒ Mouse (Confidence: 100.0%)**

The strongest rules indicate that customers who buy a Printer or a Pendrive are certain to also buy a Mouse in this dataset.

Execution of FP-Growth Algorithmn

Transaction Database

TID	Items Bought
T100	{Laptop, Mouse, Headphone, Printer}
T200	{Laptop, Mouse, Pendrive, Headphone}
T300	{Laptop, iPhone, Charger, Headphone}
T400	{Printer, Ink, Pendrive, Mouse}
T500	{iPhone, Charger, Smartwatch, Headphone}
T600	{Laptop, Printer, Pendrive, Mouse}
T700	{Laptop, iPhone, Mouse, Headphone}
T800	{Laptop, Mouse, Printer, Pendrive}

- **Minimum Support (*min_sup*):** 40%
- **Minimum Confidence (*min_conf*):** 70%

The absolute minimum support count required for an itemset to be considered frequent is:

$$\text{Minimum Support Count} = \lceil \mathit{min\_sup} \times \text{Total Transactions} \rceil = \lceil 0.40 \times 8 \rceil = \lceil 3.2 \rceil = 4$$

Any itemset must appear in at least **4 transactions** to be considered frequent.

Step 1: Find Frequent 1-Itemsets

First, we scan the transaction database once to find the support count of each individual item.

Item	Support Count	Frequent?
Laptop	6	Yes
Mouse	6	Yes
Headphone	5	Yes
Printer	4	Yes
Pendrive	4	Yes
iPhone	3	No
Charger	2	No
Ink	1	No
Smartwatch	1	No

The list of frequent items, sorted in descending order of support count, is called the **F-list**.

F-list: {Laptop: 6, Mouse: 6, Headphone: 5, Printer: 4, Pendrive: 4}

Step 2: Reorder Transactions

We remove infrequent items and reorder the remaining frequent items in each transaction according to the F-list.

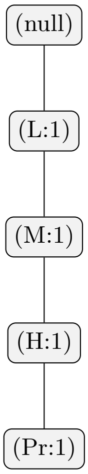
TID	Original Items	Filtered & Ordered Items
T100	{L, M, H, Pr}	{L, M, H, Pr}
T200	{L, M, Pe, H}	{L, M, H, Pe}
T300	{L, i, C, H}	{L, H}
T400	{Pr, I, Pe, M}	{M, Pr, Pe}
T500	{i, C, S, H}	{H}
T600	{L, Pr, Pe, M}	{L, M, Pr, Pe}
T700	{L, i, M, H}	{L, M, H}
T800	{L, M, Pr, Pe}	{L, M, Pr, Pe}

Step 3: Construct the FP-Tree

We build the tree by inserting the ordered transactions one by one, starting with a `null` root. Each node is represented as `(Item:Count)`.

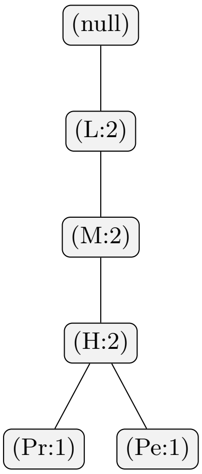
1. Processing Transaction T100: {L, M, H, Pr}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



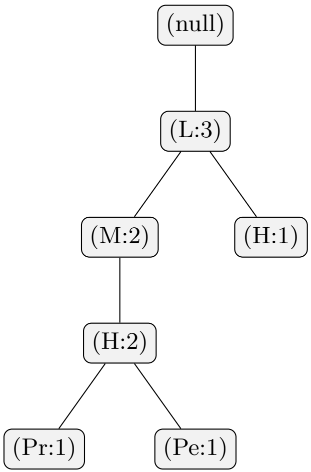
2. Processing Transaction T200: {L, M, H, Pe}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



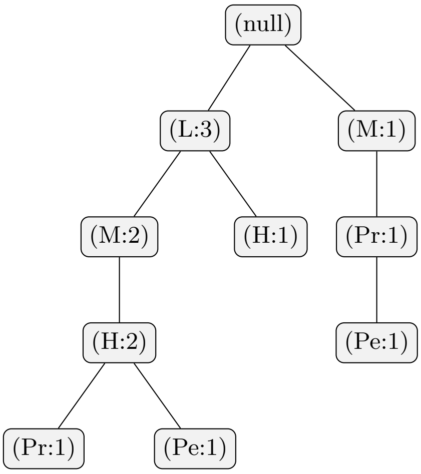
3. Processing Transaction T300: {L, H}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



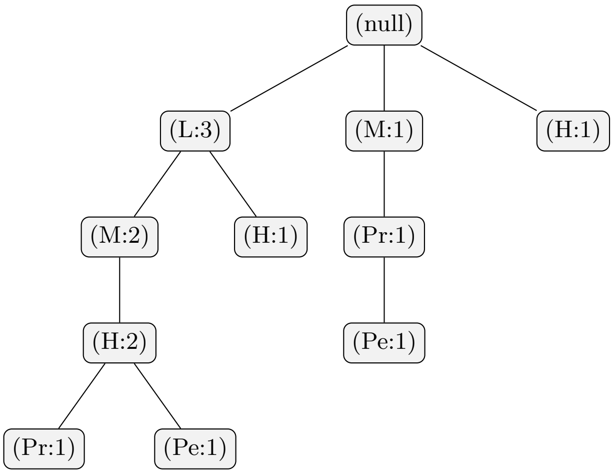
4. Processing Transaction T400: {M, Pr, Pe}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



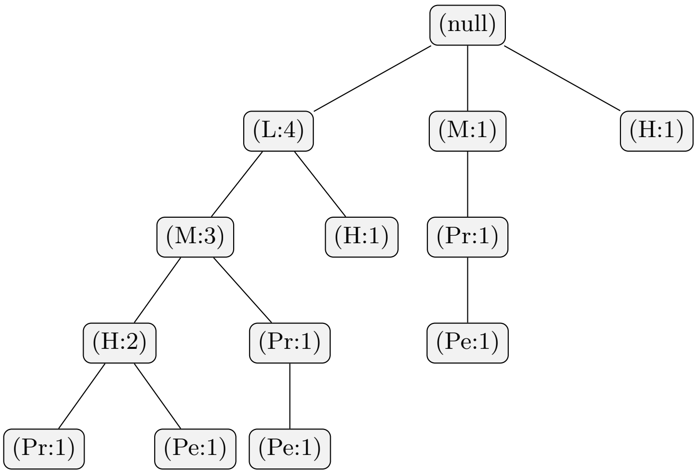
5. Processing Transaction T500: {H}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



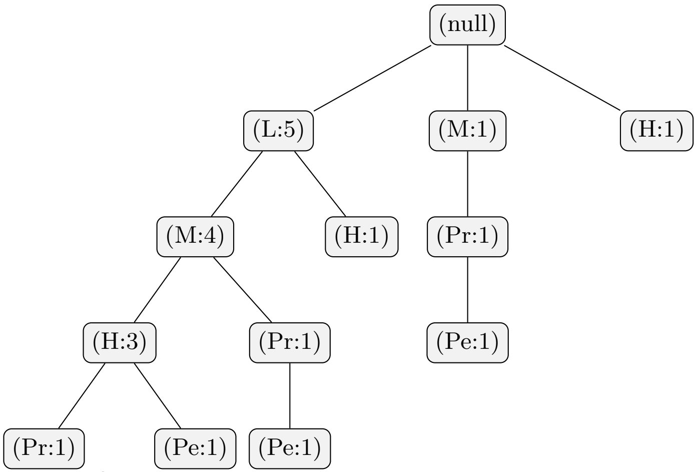
6. Processing Transaction T600: {L, M, Pr, Pe}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



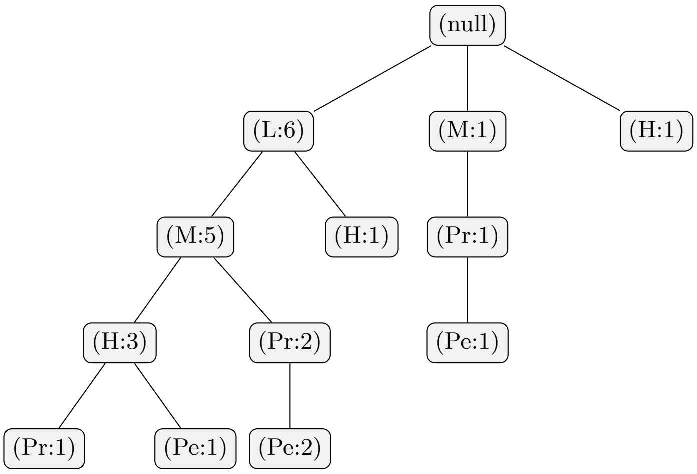
7. Processing Transaction T700: {L, M, H}

TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



8. Processing T800: {L, M, Pr, Pe}

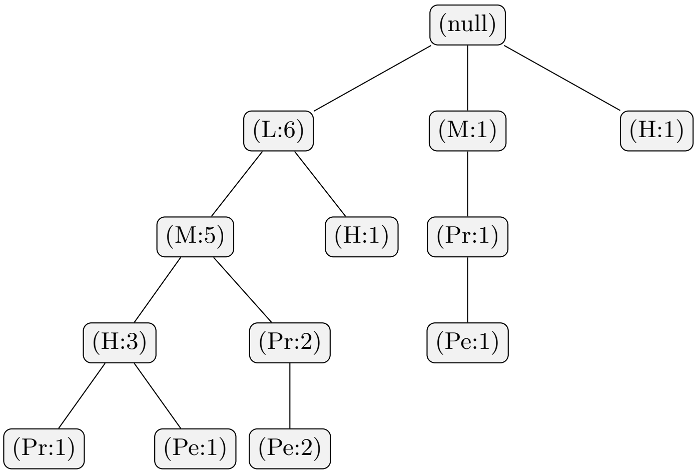
TID	Ordered Items
T100	{L, M, H, Pr}
T200	{L, M, H, Pe}
T300	{L, H}
T400	{M, Pr, Pe}
T500	{H}
T600	{L, M, Pr, Pe}
T700	{L, M, H}
T800	{L, M, Pr, Pe}



Final Tree

Header Table

Item	Count
Laptop (L)	6
Mouse (M)	6
Headphone(H)	5
Printer(Pr)	4
Pendrive (Pe)	4



Step 4: Mine Frequent Patterns from the FP-Tree

We mine the tree by creating conditional pattern bases and conditional FP-trees for each item in the F-list, starting from the bottom.

Mining for 'Pendrive' (Pe)

1. **Conditional Pattern Base for {Pe}:**
   - {L:1, M:1, H:1} from path {L, M, H, Pe}
   - {L:2, M:2, Pr:2} from path {L, M, Pr, Pe}
   - {M:1, Pr:1} from path {M, Pr, Pe}
2. **Conditional FP-Tree for {Pe}:**
   - Item counts: M(1+2+1=4), L(1+2=3), Pr(2+1=3), H(1).
   - Only **Mouse (M)** meets the minimum support of 4.
3. **Generated Frequent Pattern: {Mouse, Pendrive}** with support = 4.

Mining for 'Printer' (Pr)

1. **Conditional Pattern Base for {Pr}:**
   - {L:1, M:1, H:1} from path {L, M, H, Pr}
   - {L:2, M:2} from path {L, M, Pr}
   - {M:1} from path {M, Pr}
2. **Conditional FP-Tree for {Pr}:**
   - Item counts: M(1+2+1=4), L(1+2=3), H(1).
   - Only **Mouse (M)** meets the minimum support of 4.
3. **Generated Frequent Pattern: {Mouse, Printer}** with support = 4.

Mining for 'Headphone' (H)

1. **Conditional Pattern Base for {H}:**
   - {L:3, M:3} from path {L, M, H}
   - {L:1} from path {L, H}
2. **Conditional FP-Tree for {H}:**

- Item counts: L(3+1=4), M(3).
- Only **Laptop (L)** meets the minimum support of 4.

3. **Generated Frequent Pattern: {Laptop, Headphone}** with support = 4.

Mining for 'Mouse' (M)

1. **Conditional Pattern Base for {M}:**
   - {L:5} from path {L, M}
2. **Conditional FP-Tree for {M}:**
   - Item counts: L(5).
   - **Laptop (L)** meets the minimum support of 4.
3. **Generated Frequent Pattern: {Laptop, Mouse}** with support = 5.

Final Frequent Itemsets

Combining the frequent 1-itemsets and the patterns mined from the FP-Tree gives the complete set of frequent itemsets.

Itemset	**Support Count**
{Laptop}	6
{Mouse}	6
{Headphone}	5
{Printer}	4
{Pendrive}	4
{Laptop, Mouse}	5
{Laptop, Headphone}	4
{Mouse, Printer}	4
{Mouse, Pendrive}	4

Execution of the Eclat

Eclat (Equivalence Class Clustering and bottom-up Lattice Traversal) algorithm for mining frequent itemsets from a transaction database. The algorithm utilizes a vertical data format and TID set intersections to efficiently discover patterns.

Transaction Database

The table below lists the 8 transactions in our database.

TID	**Items Bought**
T100	{Laptop, Mouse, Headphone, Printer}
T200	{Laptop, Mouse, Pendrive, Headphone}
T300	{Laptop, iPhone, Charger, Headphone}
T400	{Printer, Ink, Pendrive, Mouse}
T500	{iPhone, Charger, Smartwatch, Headphone}
T600	{Laptop, Printer, Pendrive, Mouse}
T700	{Laptop, iPhone, Mouse, Headphone}
T800	{Laptop, Mouse, Printer, Pendrive}

- **Total Transactions:** 8
- **Minimum Support Threshold (min_sup):** 40%

First, we calculate the absolute minimum support count required for an itemset to be considered frequent.

$$\text{Minimum Support Count} = \lceil \text{min\_sup} \times \text{Total Transactions} \rceil = \lceil 0.40 \times 8 \rceil = \lceil 3.2 \rceil = 4$$

Therefore, any itemset must appear in at least **4 transactions** to be frequent.

Step 1: Data Transformation

The Eclat algorithm requires the transaction database to be in a vertical format, where each item is mapped to the set of Transaction IDs (TIDs) in which it appears.

The horizontal database is transformed into the vertical format below.

Item	**TID Set**
Laptop	{T100, T200, T300, T600, T700, T800}
Mouse	{T100, T200, T400, T600, T700, T800}
Headphone	{T100, T200, T300, T500, T700}
Printer	{T100, T400, T600, T800}
Pendrive	{T200, T400, T600, T800}
iPhone	{T300, T500, T700}
Charger	{T300, T500}
Ink	{T400}
Smartwatch	{T500}

Step 2: Generate Frequent 1-Itemsets (L_1)

The first pass identifies all frequent 1-itemsets by counting the size of each item's TID set. Items with a support count less than 4 are pruned.

Table 1: Finding Frequent 1-Itemsets

Itemset	**Support Count**	**Frequent?**
{Laptop}	6	✓
{Mouse}	6	✓
{Headphone}	5	✓
{Printer}	4	✓
{Pendrive}	4	✓
{iPhone}	3	×
{Charger}	2	×
{Ink}	1	×
{Smartwatch}	1	×

The set of frequent 1-itemsets, L_1, is: **{Laptop}, {Mouse}, {Headphone}, {Printer}, {Pendrive}**.

Step 3: Generate Frequent 2-Itemsets (L_2)

Candidate 2-itemsets are generated by intersecting the TID sets of all pairs from L_1. The resulting support is the size of the new TID set.

The set of frequent 2-itemsets, L_2, is: **{Laptop, Mouse}**, **{Laptop, Headphone}**, **{Mouse, Printer}**, **{Mouse, Pendrive}**.

Table 2: Generating Frequent 2-Itemsets from L_1

Candidate Itemset	Intersection of TID Sets	Support	Frequent?
{Laptop, Mouse}	{T100, T200, T600, T700, T800}	5	✓
{Laptop, Headphone}	{T100, T200, T300, T700}	4	✓
{Laptop, Printer}	{T100, T600, T800}	3	×
{Laptop, Pendrive}	{T200, T600, T800}	3	×
{Mouse, Headphone}	{T100, T200, T700}	3	×
{Mouse, Printer}	{T100, T400, T600, T800}	4	✓
{Mouse, Pendrive}	{T200, T400, T600, T800}	4	✓
{Headphone, Printer}	{T100}	1	×
{Headphone, Pendrive}	{T200}	1	×
{Printer, Pendrive}	{T400, T600, T800}	3	×

Step 4: Generate Frequent 3-Itemsets (L_3) and Termination

Candidate 3-itemsets are generated by joining members of L_2 that share their first item. The support is calculated by intersecting their respective TID sets.

Candidate: {Laptop, Mouse, Headphone} This candidate is formed by joining {Laptop, Mouse} and {Laptop, Headphone}.

- **Intersection:** $T(\{\text{Laptop, Mouse}\}) \cap T(\{\text{Laptop, Headphone}\})$
- **Result:** {T100, T200, T700}
- **Support Count:** 3. **Not frequent**, since $3 < 4$.

Candidate: {Mouse, Printer, Pendrive} This candidate is formed by joining {Mouse, Printer} and {Mouse, Pendrive}.

- **Intersection:** $T(\{\text{Mouse, Printer}\}) \cap T(\{\text{Mouse, Pendrive}\})$
- **Result:** {T400, T600, T800}
- **Support Count:** 3. **Not frequent**, since $3 < 4$.

Since no frequent 3-itemsets (L_3) were found, the algorithm terminates.

Summary of All Frequent Itemsets

The Eclat algorithm successfully identified all itemsets meeting the minimum support count of 4. The final results are summarized below.

Frequent 1-Itemsets (L_1)

- {Laptop}: support 6
- {Mouse}: support 6
- {Headphone}: support 5
- {Printer}: support 4
- {Pendrive}: support 4

Frequent 2-Itemsets (L_2)

- {Laptop, Mouse}: support 5
- {Laptop, Headphone}: support 4
- {Mouse, Printer}: support 4
- {Mouse, Pendrive}: support 4

Rule Evaluation Measures

A detailed analysis of the association rule for the itemset **{Laptop, Mouse}** based on the following transaction database.

TID	Items Bought
T100	{Laptop, Mouse, Headphone, Printer}
T200	{Laptop, Mouse, Pendrive, Headphone}
T300	{Laptop, iPhone, Charger, Headphone}
T400	{Printer, Ink, Pendrive, Mouse}
T500	{iPhone, Charger, Smartwatch, Headphone}
T600	{Laptop, Printer, Pendrive, Mouse}
T700	{Laptop, iPhone, Mouse, Headphone}
T800	{Laptop, Mouse, Printer, Pendrive}

Fundamental Probabilities and Support

From the database of $N = 8$ transactions, we derive the following support counts and probabilities:

- Support count for {Laptop, Mouse}: $\text{supp}(L \cap M) = 5$
- Support count for {Laptop}: $\text{supp}(L) = 6$
- Support count for {Mouse}: $\text{supp}(M) = 6$

The corresponding probabilities are:

- $P(L \cap M) = 5/8 = 0.625$
- $P(L) = 6/8 = 0.75$
- $P(M) = 6/8 = 0.75$

Confidence

The confidence for the two directional rules is calculated as:

$$\text{conf}(L \to M) = \frac{\text{supp}(L \cap M)}{\text{supp}(L)} = \frac{5}{6} \approx 0.833$$
$$\text{conf}(M \to L) = \frac{\text{supp}(L \cap M)}{\text{supp}(M)} = \frac{5}{6} \approx 0.833$$

All-Confidence

All-Confidence is defined as the minimum of the two directional confidence values.

$$\text{all\_conf}(L, M) = \min(\text{conf}(L \to M), \text{conf}(M \to L)) = \min\left(\frac{5}{6}, \frac{5}{6}\right) = \frac{5}{6} \approx 0.833$$

Max-Confidence

Max-Confidence is the maximum of the two directional confidence values, highlighting the stronger of the two potential implications.

$$\text{max\_conf}(L, M) = \max(\text{conf}(L \to M), \text{conf}(M \to L)) = \max\left(\frac{5}{6}, \frac{5}{6}\right) = \frac{5}{6} \approx 0.833$$

Kulczynski

The Kulczynski measure provides a balanced view of the association by calculating the arithmetic mean of the two directional confidence values.

$$\text{Kulczynski}(L, M) = \frac{1}{2}\left(\text{conf}(L \to M) + \text{conf}(M \to L)\right) = \frac{1}{2}\left(\frac{5}{6} + \frac{5}{6}\right) = \frac{1}{2}\left(\frac{10}{6}\right) = \frac{5}{6} \approx 0.833$$

Cosine Measure

Cosine measures the similarity between the item vectors.

$$\text{cosine}(L, M) = \frac{\text{supp}(L \cap M)}{\sqrt{\text{supp}(L) \times \text{supp}(M)}} = \frac{5}{\sqrt{6 \times 6}} = \frac{5}{6} \approx 0.833$$

Jaccard Coefficient

The Jaccard coefficient measures similarity as the ratio of the intersection to the union of the itemsets.

$$\text{Jaccard}(L, M) = \frac{\text{supp}(L \cap M)}{\text{supp}(L \cup M)} = \frac{\text{supp}(L \cap M)}{\text{supp}(L) + \text{supp}(M) - \text{supp}(L \cap M)} = \frac{5}{6 + 6 - 5} = \frac{5}{7} \approx 0.714$$

Lift

Lift measures how many times more often L and M occur together than if they were statistically independent.

$$\text{lift}(L, M) = \frac{P(L \cap M)}{P(L)P(M)} = \frac{0.625}{0.75 \times 0.75} = \frac{0.625}{0.5625} \approx 1.111$$

A lift value greater than 1 indicates a positive correlation.

Correlation (ϕ-Coefficient)

The Phi Coefficient is a measure of association for two binary variables.

$$\phi = \frac{P(L, M) - P(L)P(M)}{\sqrt{P(L)P(M)(1 - P(L))(1 - P(M))}} = \frac{0.625 - (0.75 \times 0.75)}{\sqrt{0.75 \times 0.75 \times 0.25 \times 0.25}} = \frac{0.0625}{0.1875} = \frac{1}{3} \approx 0.333$$

Chi-Squared (χ^2) Test of Independence

This test assesses if the observed association is statistically significant. The null hypothesis (H_0) is that the two items are independent.

Contingency Table

First, we construct a 2x2 contingency table of observed frequencies.

	Mouse	**Not Mouse**	**Total**
Laptop	5	1	6
Not Laptop	1	1	2
Total	6	2	8

Expected Frequencies and Calculation

The χ^2 statistic is calculated as $\chi^2 = \sum \frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$.

$$\chi^2 = \frac{(5 - 4.5)^2}{4.5} + \frac{(1 - 1.5)^2}{1.5} + \frac{(1 - 1.5)^2}{1.5} + \frac{(1 - 0.5)^2}{0.5} = 0.89$$

Interpretation

For a 2x2 table, the degrees of freedom (df) is 1. The critical value for χ^2 at a significance level of $\alpha = 0.05$ with df=1 is 3.841. Since our calculated value $\chi^2 = 0.89 < 3.841$, we fail to reject the null hypothesis. The association is **not statistically significant**.

Summary and Conclusion

The results from all measures are summarized below.

Measure	Value (Decimal)	Interpretation
Confidence-Based (All/Max/Kulczynski)	≈ 0.833	High directional predictability.
Cosine	≈ 0.833	High similarity in purchase patterns.
Jaccard	≈ 0.714	Strong overlap between the sets of transactions.
Lift	≈ 1.111	Slight positive dependence (11% more likely than chance).
Correlation (ϕ)	≈ 0.333	Weak to moderate positive statistical association.
Chi-Squared (χ^2)	0.89	The association is not statistically significant.

Conclusion: The analysis reveals a dichotomy. On one hand, the confidence-based measures, Cosine (≈ 0.833), and Jaccard (≈ 0.714) suggest a strong and reliable association based on co-occurrence. However, the more rigorous statistical measures provide a nuanced perspective. Lift (1.111) and Correlation (0.333) indicate that while the relationship is positive, it is relatively weak. Most importantly, the Chi-Squared test ($\chi^2 = 0.89$) concludes that the observed co-occurrence is not statistically significant and is likely due to random chance, a common outcome when the constituent items are individually frequent. Therefore, despite high confidence and similarity scores, the rule {Laptop, Mouse} should not be considered a strong or actionable association from a statistical standpoint.

Data Summary and Contingency Table

The transaction data is summarized in the 2x2 contingency table below. This table forms the basis for all subsequent calculations.

Table 3: Contingency table for singara and cha purchases.

	cha (C)	**¬ cha**	**Row Total**
singara (S)	1,800	700	2,500
¬ singara	1,200	1,300	2,500
Column Total	3,000	2,000	**5,000**

From the table, we extract the following primary counts:

- Total transactions, $N = 5000$
- Transactions with singara, $count(S) = 2500$
- Transactions without singara, $count(\neg S) = 2500$
- Transactions with cha, $count(C) = 3000$
- Transactions without cha, $count(\neg C) = 2000$
- Transactions with both singara and cha, $count(S \cap C) = 1800$

Calculation of Association Rule Measures

Support

Support measures the frequency of an itemset in the dataset.

$$\text{Support}(S) = \frac{\text{count}(S)}{N} = \frac{2500}{5000} = 0.50$$
$$\text{Support}(C) = \frac{\text{count}(C)}{N} = \frac{3000}{5000} = 0.60$$
$$\text{Support}(S, C) = \frac{\text{count}(S \cap C)}{N} = \frac{1800}{5000} = 0.36$$

Confidence

Confidence indicates the probability of seeing the consequent in a transaction that also contains the antecedent.

$$\text{Confidence}(S \to C) = \frac{\text{Support}(S, C)}{\text{Support}(S)} = \frac{0.36}{0.50} = 0.72$$
$$\text{Confidence}(C \to S) = \frac{\text{Support}(S, C)}{\text{Support}(C)} = \frac{0.36}{0.60} = 0.60$$

Lift

Lift measures how much more likely two items are to be purchased together than if they were independent.

$$\text{Lift}(S, C) = \frac{\text{Support}(S, C)}{\text{Support}(S) \times \text{Support}(C)} = \frac{0.36}{0.50 \times 0.60} = \frac{0.36}{0.30} = 1.20$$

Correlation (Phi Coefficient)

The Phi Coefficient (ϕ) measures the linear association for a 2x2 table, normalized between -1 and +1.

$$\phi = \frac{N \cdot \text{count}(S \cap C) - \text{count}(S) \cdot \text{count}(C)}{\sqrt{\text{count}(S) \cdot \text{count}(C) \cdot \text{count}(\neg S) \cdot \text{count}(\neg C)}}$$
$$\phi = \frac{5000 \cdot 1800 - 2500 \cdot 3000}{\sqrt{2500 \cdot 3000 \cdot 2500 \cdot 2000}} = \frac{9,000,000 - 7,500,000}{\sqrt{3.75 \times 10^{13}}} = \frac{1,500,000}{6,123,724.35} \approx 0.245$$

Chi-Squared (χ^2)

The Chi-Squared test assesses the independence of the two items by comparing observed frequencies to expected frequencies under the null hypothesis of independence.

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Expected$(S, C) = \frac{2500 \times 3000}{5000} = 1500$.

$$\begin{aligned}\chi^2 &= \frac{(1800 - 1500)^2}{1500} + \frac{(700 - 1000)^2}{1000} + \frac{(1200 - 1500)^2}{1500} + \frac{(1300 - 1000)^2}{1000} \\ &= \frac{90000}{1500} + \frac{90000}{1000} + \frac{90000}{1500} + \frac{90000}{1000} \\ &= 60 + 90 + 60 + 90 = 300\end{aligned}$$

With 1 degree of freedom, $\chi^2 = 300$ is highly statistically significant ($p \ll 0.001$), allowing us to reject the hypothesis of independence.

Other Standard Measures

- **All-Confidence:** The minimum of the two confidence values.

$$\text{All-Conf}(S, C) = \min(\text{Conf}(S \to C), \text{Conf}(C \to S)) = \min(0.72, 0.60) = 0.60$$

- **Max-Confidence:** The maximum of the two confidence values.

$$\text{Max-Conf}(S, C) = \max(\text{Conf}(S \to C), \text{Conf}(C \to S)) = \max(0.72, 0.60) = 0.72$$

- **Kulczynski:** The arithmetic mean of the confidence values.

$$\text{Kulczynski}(S, C) = \frac{1}{2}(\text{Conf}(S \to C) + \text{Conf}(C \to S)) = \frac{1}{2}(0.72 + 0.60) = 0.66$$

- **Cosine Similarity:** The geometric mean of the confidence values.

$$\text{Cosine}(S, C) = \frac{\text{Support}(S, C)}{\sqrt{\text{Support}(S) \times \text{Support}(C)}} = \frac{0.36}{\sqrt{0.50 \times 0.60}} \approx 0.657$$

- **Jaccard Similarity:** The ratio of the intersection to the union of the itemsets.

$$\text{Jaccard}(S, C) = \frac{\text{count}(S \cap C)}{\text{count}(S) + \text{count}(C) - \text{count}(S \cap C)} = \frac{1800}{2500 + 3000 - 1800} = \frac{1800}{3700} \approx 0.486$$

Comparative Analysis of Key Measures

While all measures point towards a positive association, they each offer a unique perspective on the relationship between purchasing singara and cha. The table below summarizes the key metrics for comparison.

Table 4: Key Association Measures for Singara (S) and Cha (C)

Measure	**Value**	**Interpretation**
Support(S, C)	0.36	Occurs in 36% of all transactions.
Confidence(S → C)	0.72	72% of singara purchases also include cha.
Lift	1.20	20% more likely to be co-purchased than by chance.
Correlation (ϕ)	0.245	Weak positive linear association.

Conclusion The relationship between singara and cha purchases exhibits a statistically significant positive association, as confirmed by a high Chi-Squared value and a Lift greater than 1. While symmetric measures like Lift (1.20) and Correlation (0.245) indicate that the overall strength of this association is modest, the directional measures provide actionable business insights. The confidence of the rule $S \rightarrow C$ (72%) is notably higher than that of $C \rightarrow S$ (60%). This suggests that while the items are often bought together, a singara purchase is a stronger trigger for a co-purchase of cha. Therefore, marketing strategies such as promotions or product placement could be effectively designed by leveraging singara as the driver item to increase sales of cha.

Prepared By:
Md. Atikuzzaman
Lecturer
Department of Computer Science and Engineering
Green University of Bangladesh
Email: atik@cse.green.edu.bd