Minning Association Rules

HỆ HỖ TRỢ QUYẾT ĐỊNH

Bài 9(a): Association Analysis

Contents

Association Analysis

- Association analysis measures the strength of co-occurrence between one item and another.
- The objective of this class of data science algorithms is not to predict an occurrence of an item but to find usable patterns in the co-occurrences of the items
- Association rules learning is a branch of unsupervised learning processes that discover hidden patterns in data, in the form of easily recognizable rules
- One of the popular applications of this technique is called market basket analysis, which finds co-occurrences of one retail item with another item within the same retail purchase transaction

Method

- **Step 1**: Prepare the data in transaction format. An association algorithm needs input data to be formatted in transaction format $t_x = \{i_1, i_2, i_3\}$.
- **Step 2**: Short-list frequently occurring *itemsets*. Itemsets are combinations of items. An association algorithm limits the analysis to the most frequently occurring items, so that the final rule set extracted in the next step is more meaningful.
- **Step 3**: Generate *relevant* association rules from itemsets. Finally, the algorithm generates and filters the rules based on the interest measure.

Association Analysis

- The model outcome of an association analysis can be represented as a set of rules, like the one below
- {Item A} -> {Item B}
- This rule indicates that based on the history of all the transactions, when Item A is found in a transaction or a basket, there is a strong of the occurrence of Item B within the *same* transaction.
- Item A is the *antecedent* or *premise* of the rule and Item B is the *consequent* or *conclusion* of the rule.
- The antecedent and consequent of the rule can contain more than one item
  - E.g. {News, Finance } - > {Sports }

Example

Table 6.1

Session ID	List of media categories accessed
1	{News, Finance}
2	{News, Finance}
3	{Sports, Finance, News}
4	{Arts}
5	{Sports, News, Finance}
6	{News, Arts, Entertainment}

⇩ Pivot

Table 6.2 Clickstream Data Set

Session ID	News	Finance	Entertainment	Sports	Arts
1	1	1	0	0	0
2	1	1	0	0	0
3	1	1	0	1	0
4	0	0	0	0	1
5	1	1	0	1	0
6	1	0	1	0	1

Itemsets

- An itemset can occur either in the antecedent or in the consequent portion of the rule; however, both sets should be disjointed
- The strength of an association rule is commonly quantified by the *support* and *confidence* measures of a rule
- There are a few more quantifications like *lift* and *conviction* measures that can be used in special cases.
- **Support**: The *support of an item* is simply the relative frequency of occurrence of an itemset in the transaction set.

$$\text{Support}(\{\text{News}\}) = 5/6 = 0.83$$
$$\text{Support}(\{\text{News, Finance}\}) = 4/6 = 0.67$$
$$\text{Support}(\{\text{Sports}\}) = 2/6 = 0.33$$

- The *support of a rule* is a measure of how all the items in a rule are represented in the overall transactions.

Confidence

- The **confidence** of a rule measures the likelihood of the occurrence of the consequent of the rule out of all the transactions that contain the antecedent of the rule.
- Confidence provides the reliability measure of the rule.
- Confidence of the rule ($X -> Y$) is calculated by

$$\text{Confidence}\,(X \rightarrow Y) = \frac{\text{Support}\,(X \cup Y)}{\text{Support}\,(X)}$$

- E.g.

$$\begin{aligned}\text{Confidence}\,(\{\text{News, Finance}\} \rightarrow \{\text{Sports}\}) &= \frac{\text{Support}\,(\{\text{News, Finance, Sports}\})}{\text{Support}\,(\{\text{News, Finance}\})} \\ &= \frac{2/6}{4/6} \\ &= 0.5\end{aligned}$$

Conviction

- The conviction of the rule $X -> Y$ is the ratio of the expected frequency of X occurring in spite of Y and the observed frequency of incorrect predictions.

$$\text{Conviction}\,(X \rightarrow Y) = \frac{1 - \text{Support}\,(Y)}{1 - \text{Confidence}\,(X \rightarrow Y)}$$

- E.g.

$$\text{Conviction}\big(\{\text{News, Finance}\} \rightarrow \{\text{Sports}\}\big) = \frac{1 - 0.33}{1 - 0.5} = 1.32$$

- A conviction of 1.32 means that the rule ({News, Finance}->{Sports}) would be incorrect 32% more often if the relationship between {News, Finance} and {Sports} is purely random.

Lift

- $Lift(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$
- E.g.

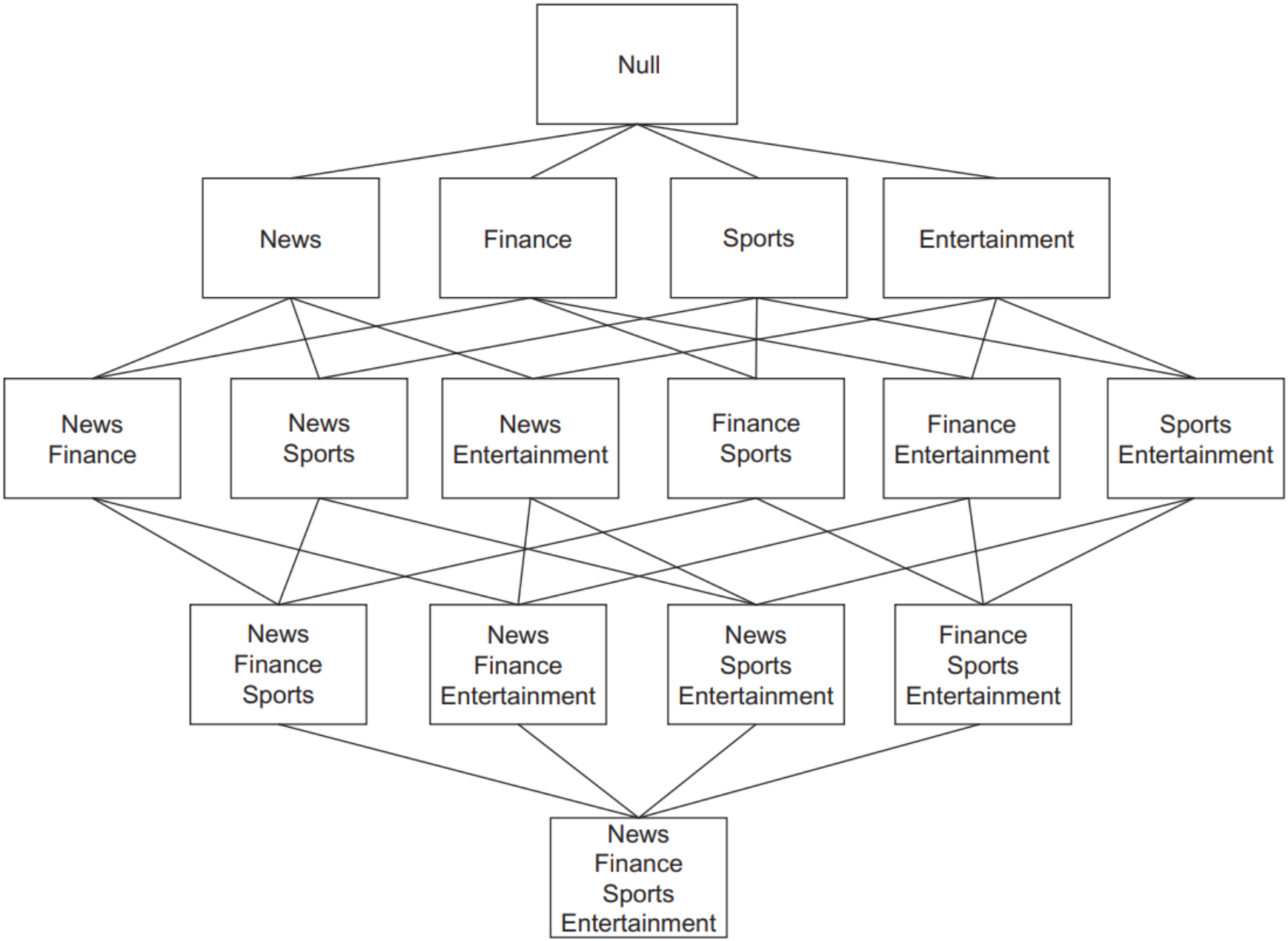
$$Lift(\{News, Finace\} \rightarrow \{Sports\}) = \frac{Support(X \cup Y)}{Support(X) \times Support(Y)}$$

$$= \frac{0.333}{0.667 \times 0.33} = 1.5$$

- Lift values closer to 1 mean the antecedent and consequent of the rules are independent and the rule is not interesting. The higher the value of lift, the more interesting the rules are.

Rule Generation

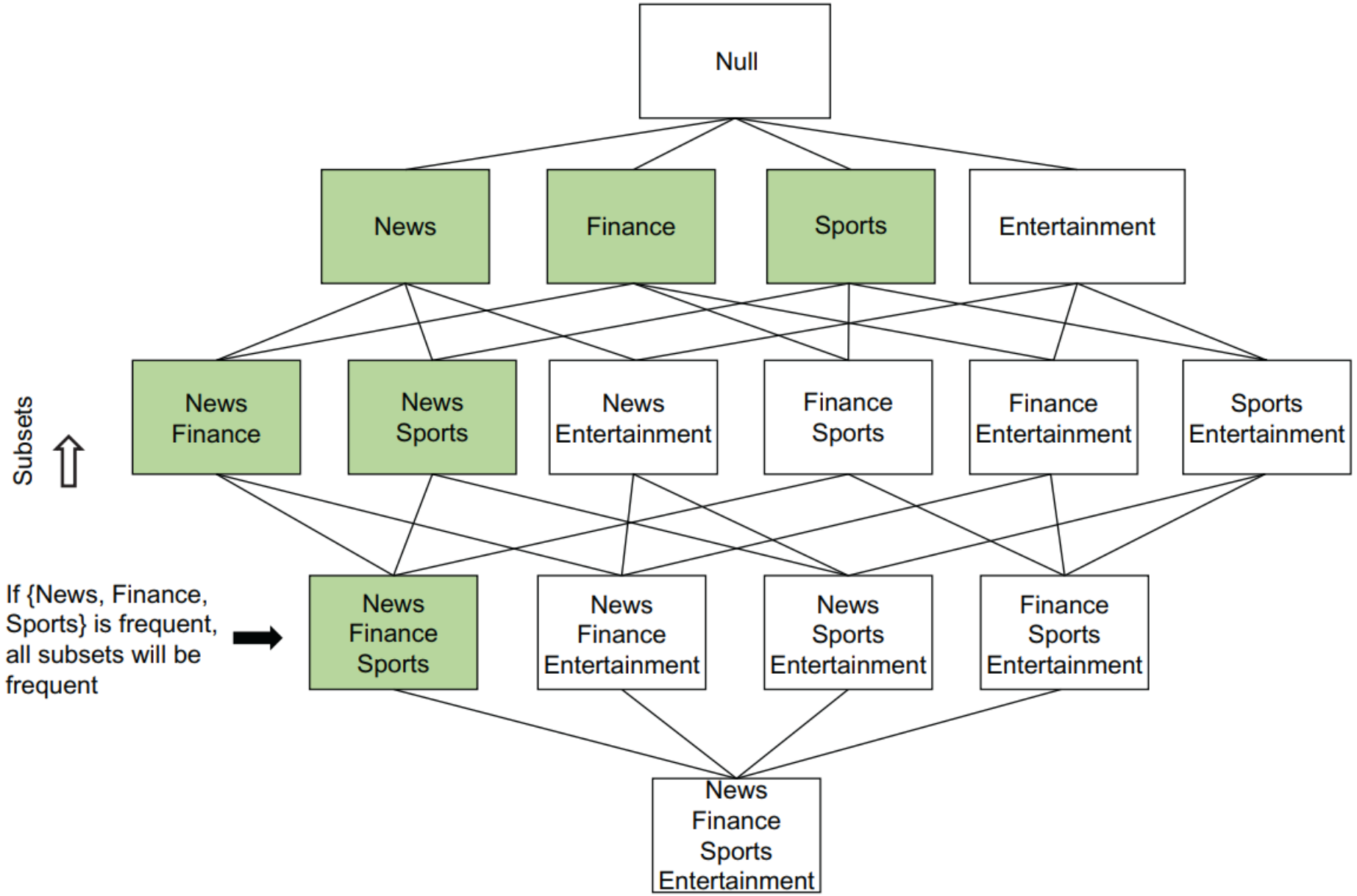
1. Finding all frequent itemsets:
   - For an association analysis of n items it is possible to find $2^n - 1$ itemsets excluding the null itemset
   - It is critical to set a minimal support threshold to discard less frequently occurring itemsets in the transaction universe.
   - It is common to exclude some not interested items
2. Extracting rules from frequent itemsets:
   - For the dataset with n items it is possible to find $3^n - 2^{n+1} + 1$ rules
   - This step extracts all the rules with a confidence higher than a minimum confidence threshold.


Null
News
Finance
Sports
Entertainment
News Finance
News Sports
News Entertainment
Finance Sports
Finance Entertainment
Sports Entertainment
News Finance Sports
News Finance Entertainment
News Sports Entertainment
Finance Sports Entertainment
News Finance Sports Entertainment

Apriori

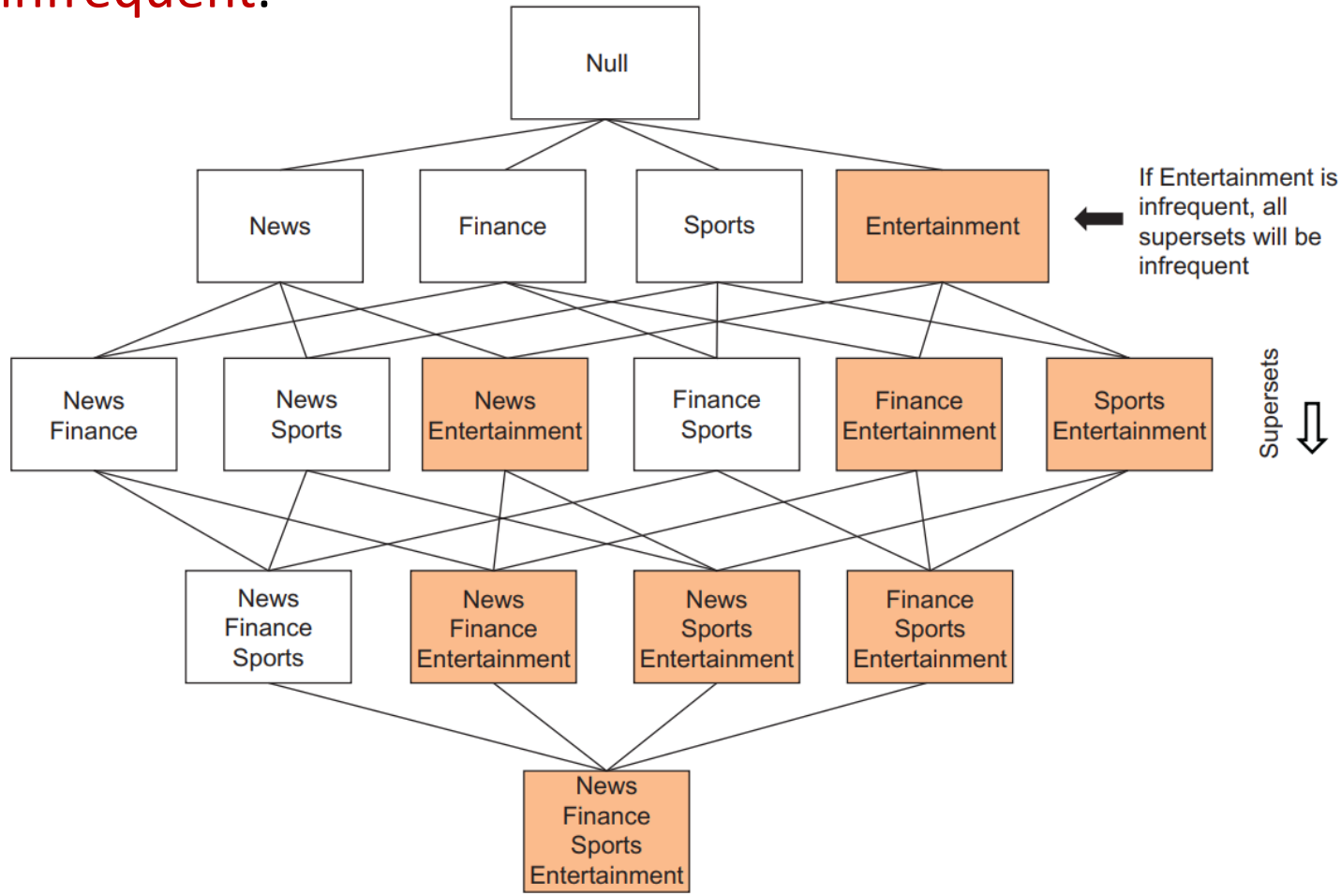
- The Apriori algorithm leverages some simple logical principles on the lattice itemsets to reduce the number of itemsets to be tested for the support measure
- The Apriori principles states that "If an itemset is frequent, then all its subset items will be frequent."
- The itemset is "frequent" if the support for the itemset is more than that of the support threshold.

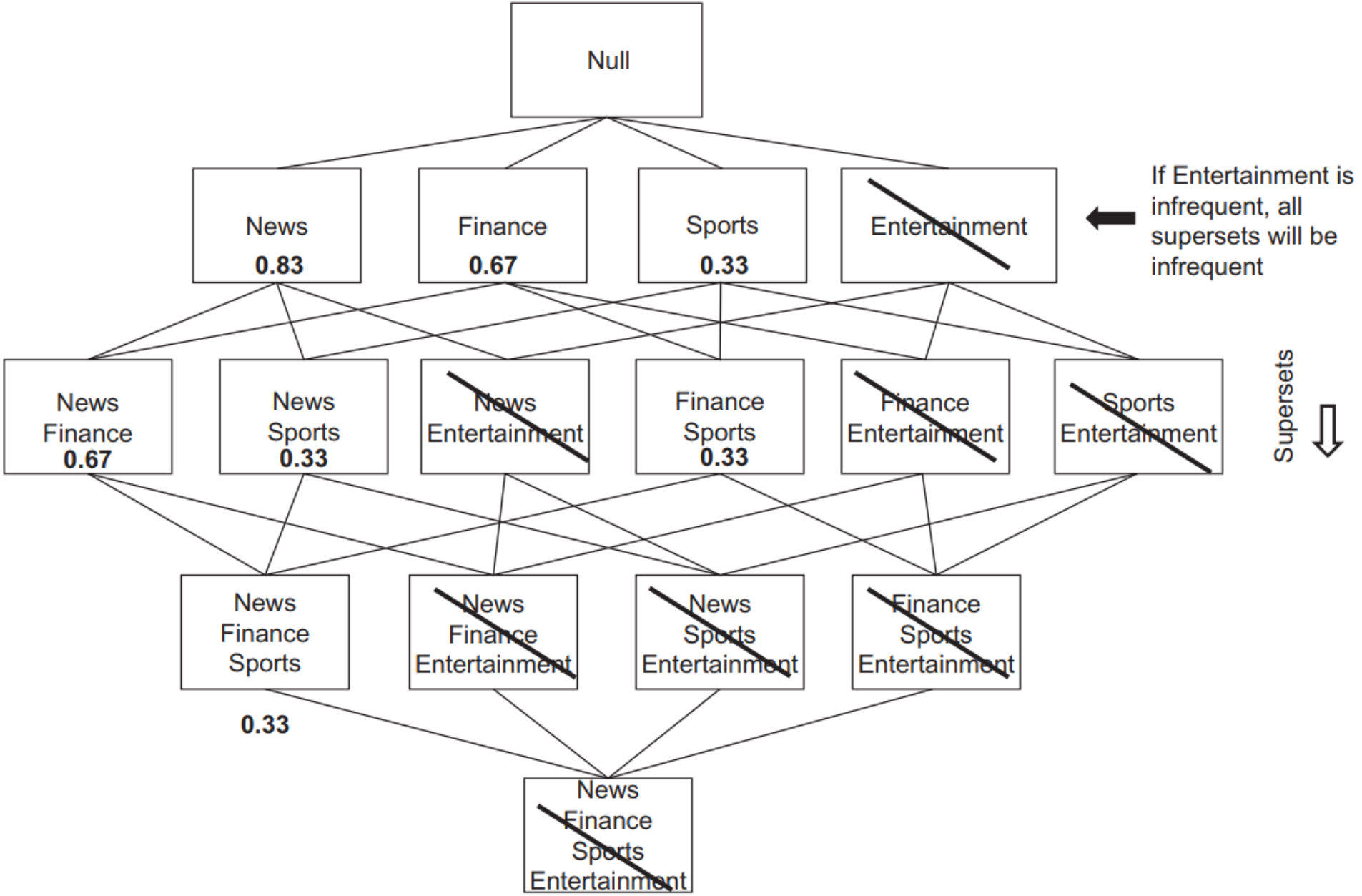
Apriori Algorithm


Null
News
Finance
Sports
Entertainment
News Finance
News Sports
News Entertainment
Finance Sports
Finance Entertainment
Sports Entertainment
Subsets
If {News, Finance, Sports} is frequent, all subsets will be frequent
News Finance Sports
News Finance Entertainment
News Sports Entertainment
Finance Sports Entertainment
News Finance Sports Entertainment

Apriori

- Conversely, if the itemset is infrequent, then all its supersets will be infrequent.




Null
News
0.83
Finance
0.67
Sports
0.33
Entertainment
If Entertainment is infrequent, all supersets will be infrequent
News Finance
0.67
News Sports
0.33
News Entertainment
Finance Sports
0.33
Finance Entertainment
Sports Entertainment
Supersets
News Finance Sports
0.33
News Finance Entertainment
News Sports Entertainment
Finance Sports Entertainment
News Finance Sports Entertainment

Example

- Support threshold is assumed to be 0.25

Table 6.3 Clickstream Dataset: Condensed Version

Session	News	Finance	Entertainment	Sports
1	1	1	0	0
2	1	1	0	0
3	1	1	0	1
4	0	0	0	0
5	1	1	0	1
6	1	0	1	0

Example

Table 6.4 Frequent Itemset Support Calculation

Item	Support Count	Support
{News}	5	0.83
{Finance}	4	0.67
{Entertainment}	1	0.17
{Sports}	2	0.33
Two-Itemsets	**Support Count**	**Support**
{News, Finance}	4	0.67
{News, Sports}	2	0.33
{Finance, Sports}	2	0.33
Three-Itemsets	**Support Count**	**Support**
{News, Finance, Sports}	2	0.33

Generate rules

- Each frequent itemset of *n* items can generate $2^n - 2$ rules
- Example with {News, Sports, Finance}, rules and confidence scores

$$\{\text{News, Sports}\} \rightarrow \{\text{Finance}\} - 0.33/0.33 = 1.0$$
$$\{\text{News, Finance}\} \rightarrow \{\text{Sports}\} - 0.33/0.67 = 0.5$$
$$\{\text{Sports, Finance}\} \rightarrow \{\text{News}\} - 0.33/0.33 = 1.0$$
$$\{\text{News}\} \rightarrow \{\text{Sports, Finance}\} - 0.33/0.83 = 0.4$$
$$\{\text{Sports}\} \rightarrow \{\text{News, Finance}\} - 0.33/0.33 = 1.0$$
$$\{\text{Finance}\} \rightarrow \{\text{News, Sports}\} - 0.33/0.67 = 0.5$$

- It is possible to prune potentially low confidence rules using the same Apriori method
- E.g. if the rule {News, Finance}->{Sports} is a low confidence rule, then it can be concluded that any rules within the subset of the antecedent will be a low confidence rule; all the rules like {News}->{Sports,Finance} and {Finance}->{News, Sports} can be discarded

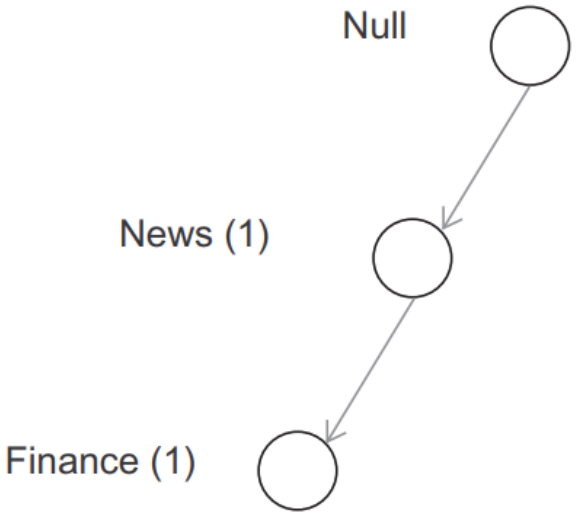
Frequent Pattern-Growth Algorithm

FP-Growth algorithm

- The FP-Growth algorithm provides an alternative way of calculating a frequent itemset by compressing the transaction records using a special graph data structure called *FP-Tree*
- FP-Tree can be thought of as a transformation of the dataset into graph format.
- FP-Growth first generates the FP-Tree and uses this compressed tree to generate the frequent itemsets
- The efficiency of the FP-Growth algorithm depends on how much compression can be achieved in generating the FP-Tree

Steps

1. The first step is to sort all the items in each transaction in descending order of frequency (or support count), e.g. The third transaction of {Sports, News, Finance} has to be rearranged to {News, Finance, Sports}
2. Starting with a null node, the first transaction {News, Finance} can be represented as following



FP-Tree – how to build?

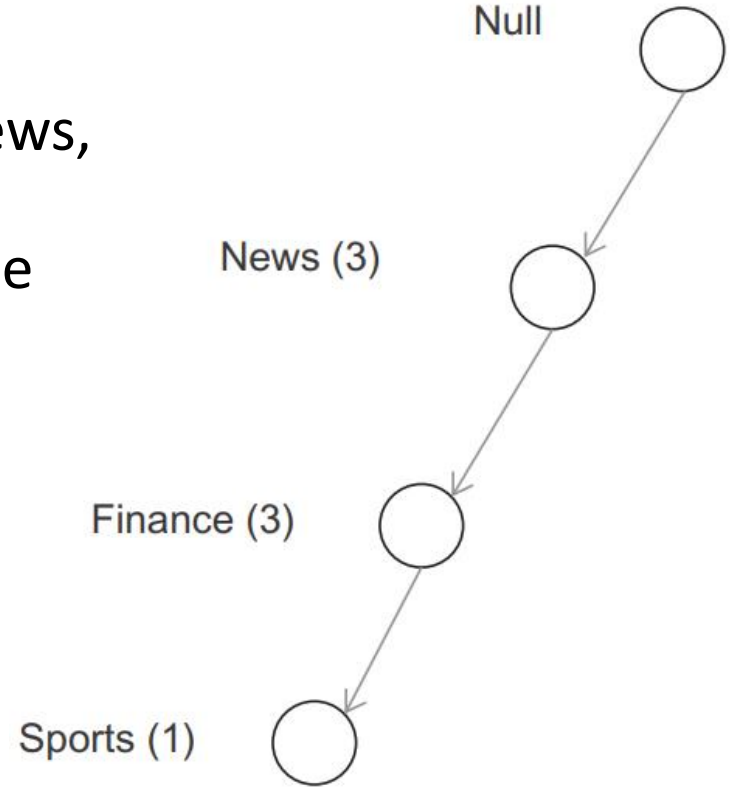
- Consider the dataset containing six transactions of four items—news, finance, sports, and entertainment.

Table 6.5 Transactions List: Session and Items

Session	Items
1	{News, Finance}
2	{News, Finance}
3	{News, Finance, Sports}
4	{Sports}
5	{News, Finance, Sports}
6	{News, Entertainment}

Steps

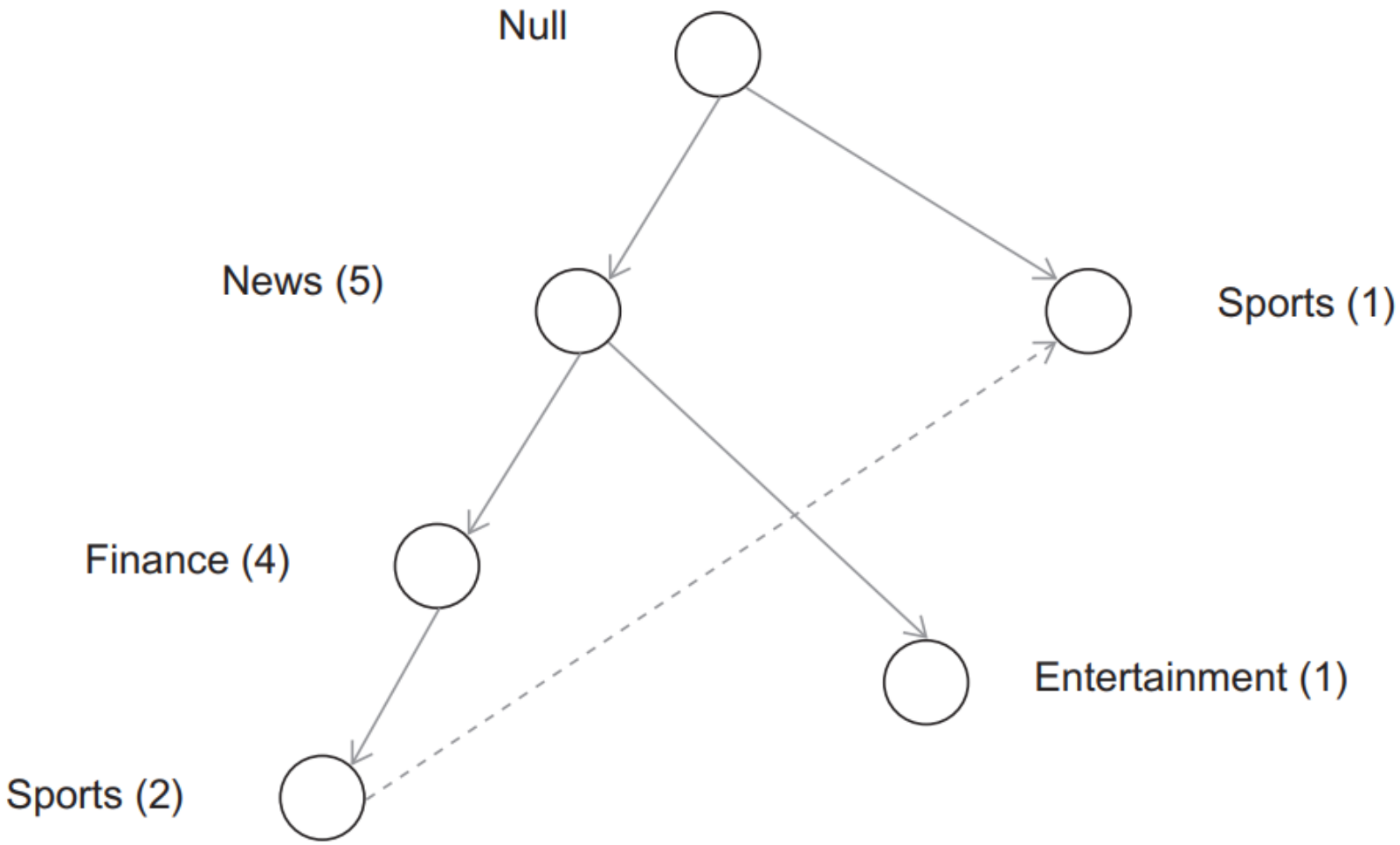
3. Since the second transaction {News, Finance} is the same as the first one, it follows the same path as the first one.
4. The third transaction contains {News, Finance, Sports}. The tree is now extended to include Sports and the item path count is incremented



Steps

5. The fourth transaction only contains the {Sports} item. Since Sports is not preceded by News and Finance, a new path should be created from the null item and the item count should be noted. This node for Sports is different from the Sports node next to Finance (the latter cooccurs with News and Finance). However, since both nodes indicate the same item, they should be linked by a dotted line.
6. This process is continued until all the transactions are scanned. All of the transaction records can be now represented by a compact FP-Tree

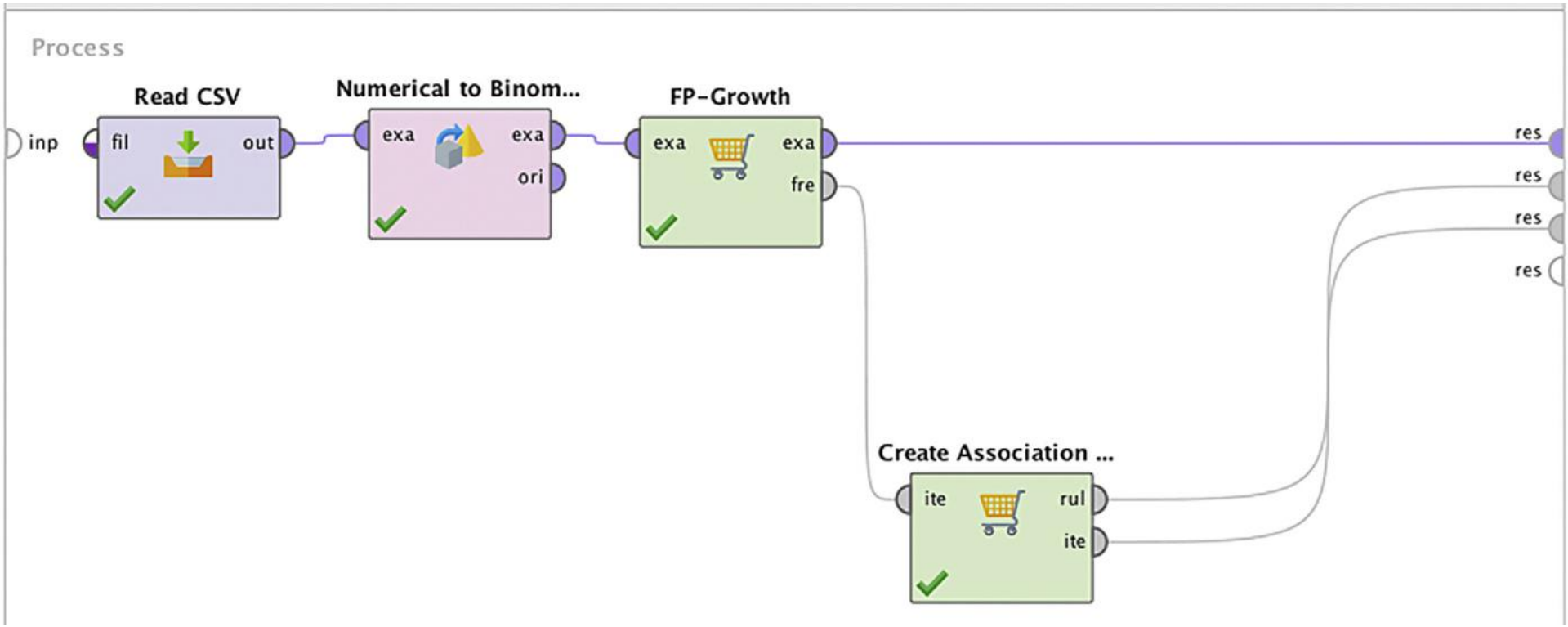
Steps



Frequent Itemset Generation

- Once the transaction set is expressed by a compact FP-Tree, the most frequent itemset can be generated from the FP-Tree effectively.
- To generate the frequent itemset, the FP-Growth algorithm adopts a bottom-up approach of generating all the itemsets starting with the least frequent items
- Since the structure of the tree is ordered by the support count, the least frequent items can be found in the leaves of the tree
- If {Entertainment} is indeed a frequent item, because the support exceeds the threshold, the algorithm will find all the itemsets ending with entertainment, like {Entertainment} and {News, Entertainment}, by following the path from the bottom-up.

FP-Growth in RapidMiner



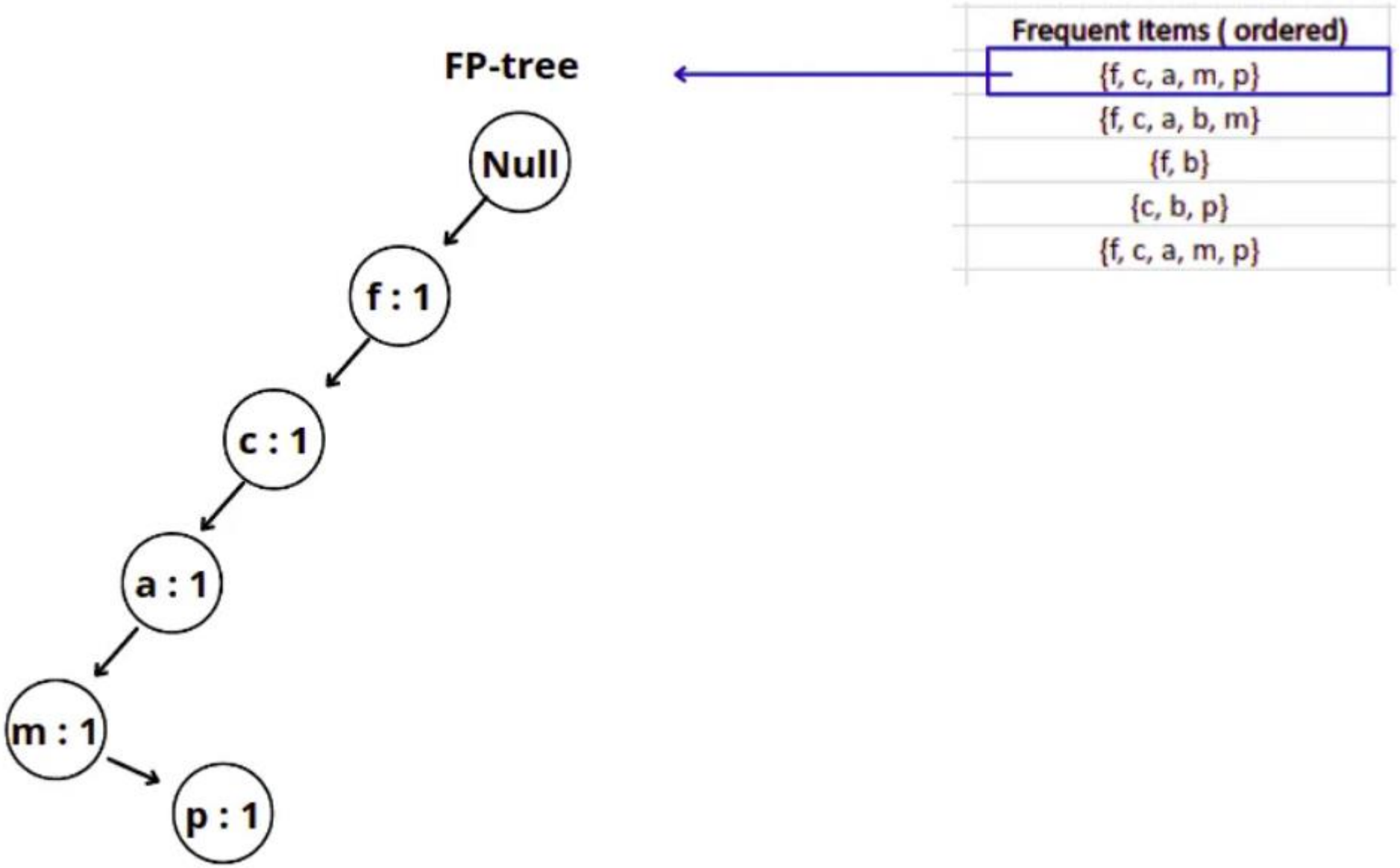
Thí dụ

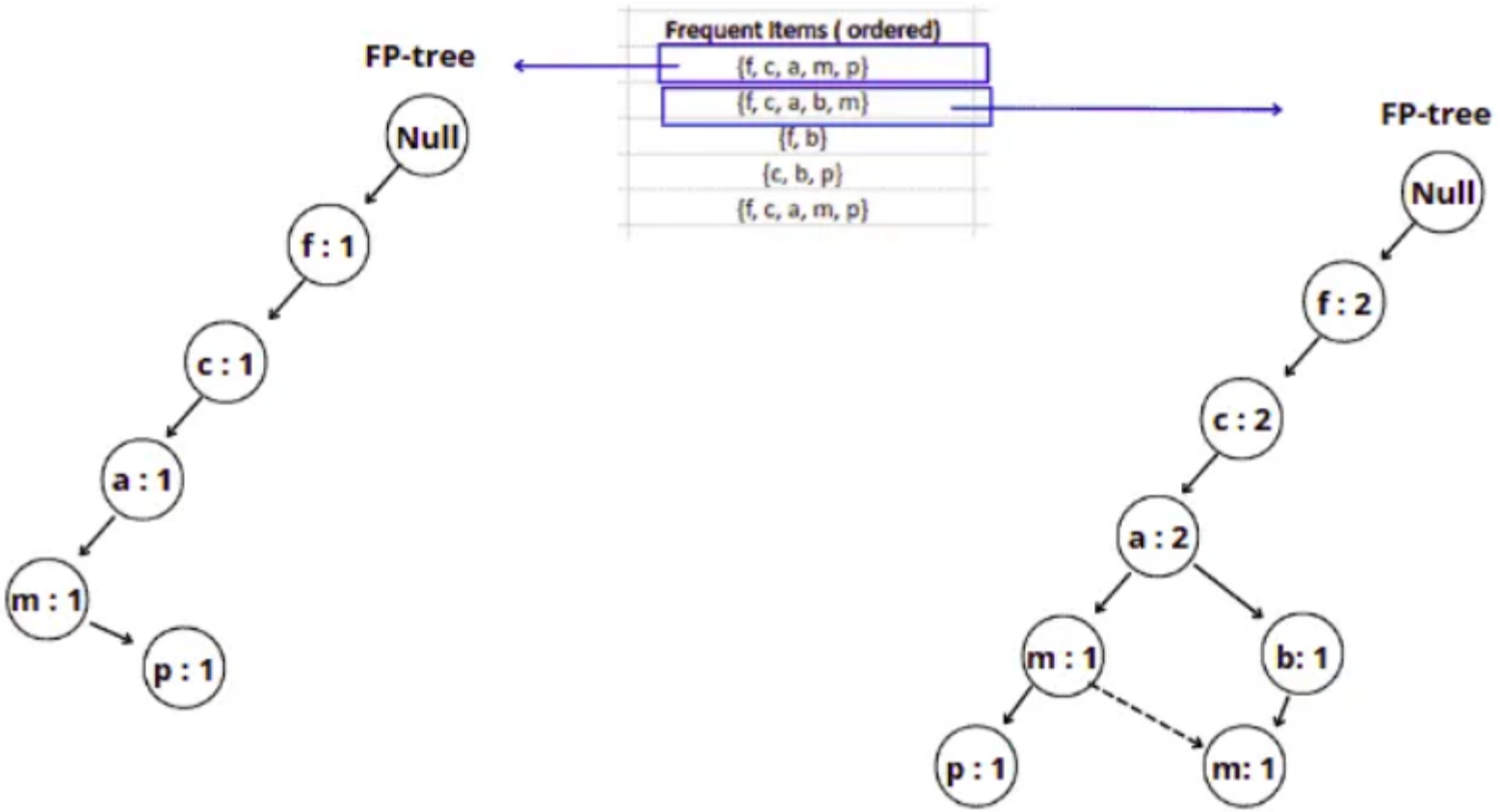
ID	Items bought
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

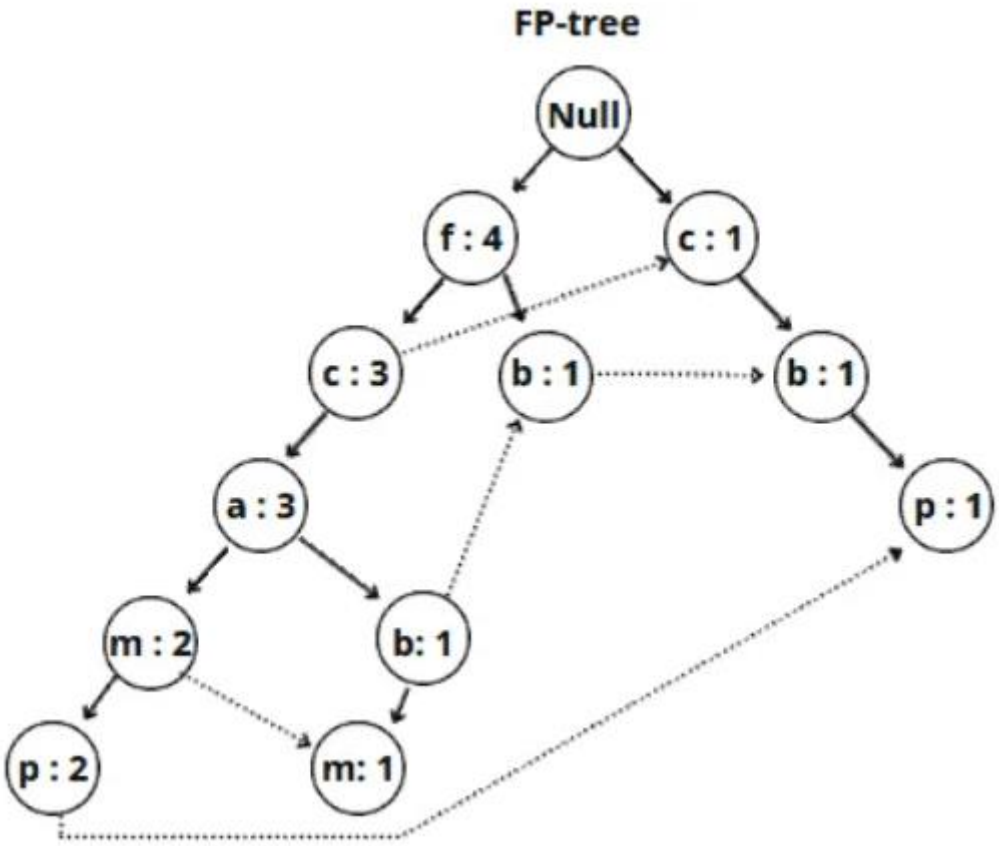
Item	Frequecy
{f}	4
{c}	3
{a}	3
{b}	3
{m}	3
{p}	3

ID	Items bought
100	{f, a, c, d, g, i, m, p}
200	{a, b, c, f, l, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, f, c, e, l, p, m, n}

Frequent Items (ordered)
{f, c, a, m, p}
{f, c, a, b, m}
{f, b}
{c, b, p}
{f, c, a, m, p}

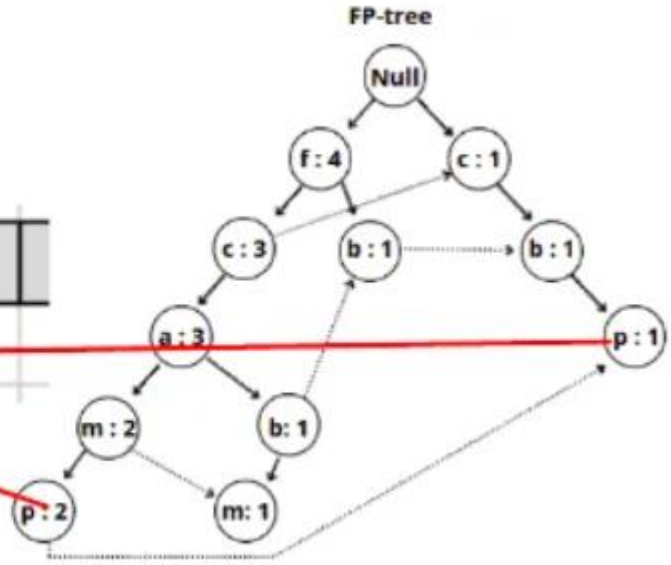

FP-tree
Null
f : 1
c : 1
a : 1
m : 1
p : 1
Frequent Items (ordered)
{f, c, a, m, p}
{f, c, a, b, m}
{f, b}
{c, b, p}
{f, c, a, m, p}


FP-tree
Null
f : 1
c : 1
a : 1
m : 1
p : 1
Frequent Items (ordered)
{f, c, a, m, p}
{f, c, a, b, m}
{f, b}
{c, b, p}
{f, c, a, m, p}
FP-tree
Null
f : 2
c : 2
a : 2
m : 1
b: 1
p : 1
m: 1


FP-tree
Null
f : 4
c : 1
c : 3
b : 1
b : 1
a : 3
p : 1
m : 2
b: 1
p : 2
m: 1

Xây dựng luật liên kết

Item	Conditional Pattern Base
p	{{ f, c, a, m, : 2}, { c, b : 1}}



Item	Conditional Pattern Base
p	{{ f, c, a, m : 2}, { c, b : 1}}
m	{{ f, c, a : 2}, {f, c, a, b : 1}}
b	{{f :1}, { c:1}, {f, c, a : 1}}
a	{{f, c : 3}}
c	{{ f : 3}}

{ f, c, a, m : 2 }, { c, b : 1 } - > { f: 2, c:3, a:2, m:2, b:1 }

Item	Conditional Pattern Base	Conditional FP-tree
p	{{f, c, a, m : 2 }, {c, b : 1 }}	{c:3 }
m	{{f, c, a : 2}, {f, c, a, b : 1 }}	{f:3, c:3, a:3 }
b	{{f : 1 }, {c : 1}, {f, c, a : 1 }}	--
a	{{f, c : 3 }}	{f:3, c:3 }
c	{{f : 3 }}	{f:3 }

Item	Conditional Pattern Base	Conditional FP-tree	Generated Frequent Patterns
p	{{f, c, a, m : 2 }, {c, b : 1 }}	{c:3 }	{c, p : 3 }
m	{{f, c, a : 2}, {f, c, a, b : 1 }}	{f:3, c:3, a:3 }	{f, m : 3 }, {c, p : 3 }, {a, m : 3 }, {f, c, m : 3}, {f, a, m : 3}, {c, a, m : 3 }, {f, c, a, m : 3 }
b	{{f : 1 }, {c : 1}, {f, c, a : 1 }}	--	--
a	{{f, c : 3 }}	{f:3, c:3 }	{f, a : 3 }, {c, a : 3 }, {f, c, a : 3 }
c	{{f : 3 }}	{f:3 }	{f, c : 3 }