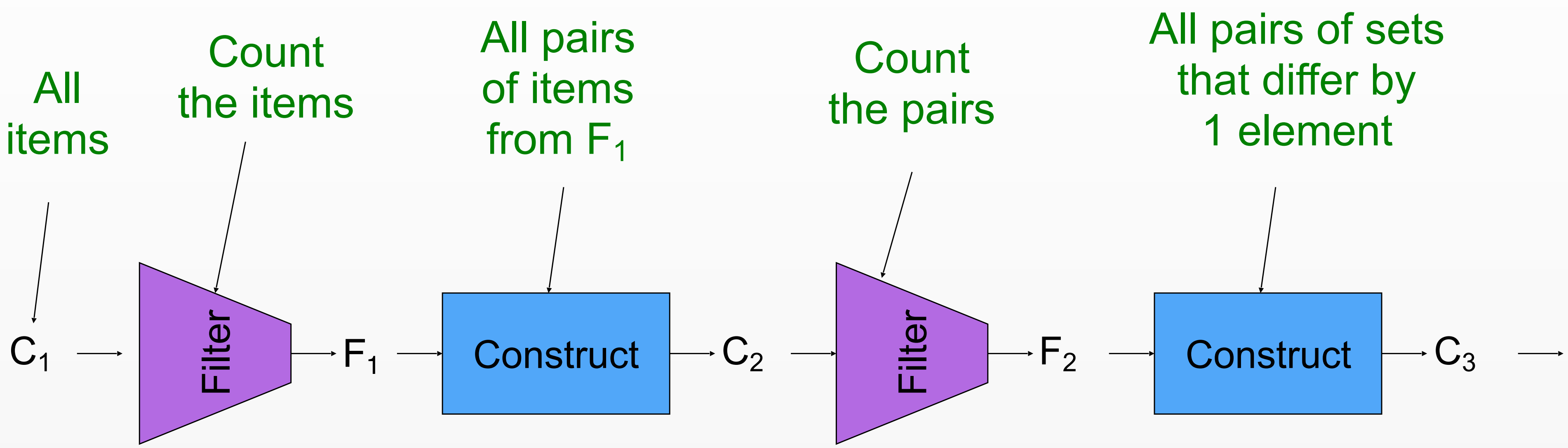
Mining Frequent Itemsets with FP-Growth

A-Priori: Bottlenecks



1. Set $k = 0$
2. Define C_1 as all size 1 item sets
3. ***While C_{k+1} is not empty***
4. Set $k = k + 1$
5. Scan DB to determine subset $F_k \subseteq C_k$ with support $\geq s$ **(I/O limited)** Counting support require reading all transactions from the disk.
6. Construct candidates C_{k+1} by combining sets in F_k that differ by 1 element (Memory limited) Requires additional mai memory to store all the candidates.

FP-Growth Algorithm – Overview

- Apriori requires one pass for each *k*
- Can we find *all* frequent item sets in fewer passes over the data?

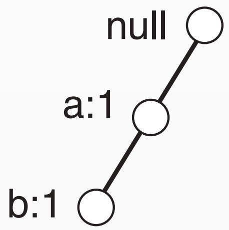
FP-Growth Algorithm:

- *Pass 1*: Count items with support ≥ s
- Sort frequent items in descending order according to count
- *Pass 2*: Store all transactions compressed in a frequent pattern tree (FP-tree)
- Mine patterns from FP-Tree

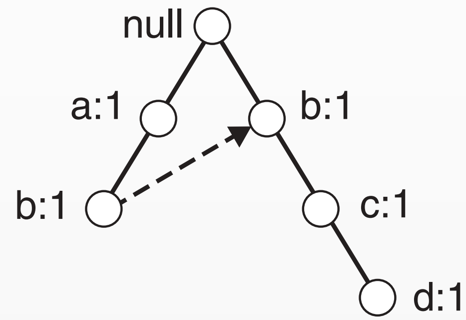
If the transactions can be compressed enough, they might fit into the main memory. No need to reread the transactions repeatedly from the disc.

FP-Tree Construction

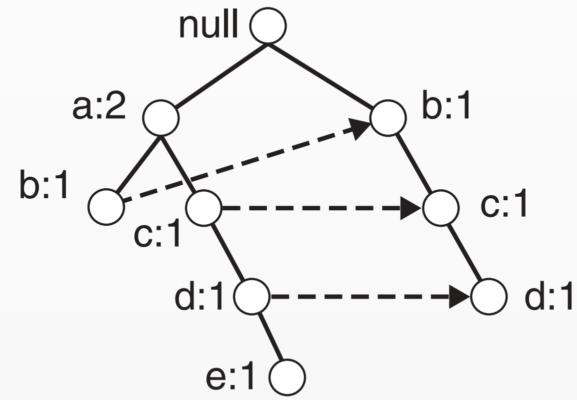
TID = 1



TID = 2

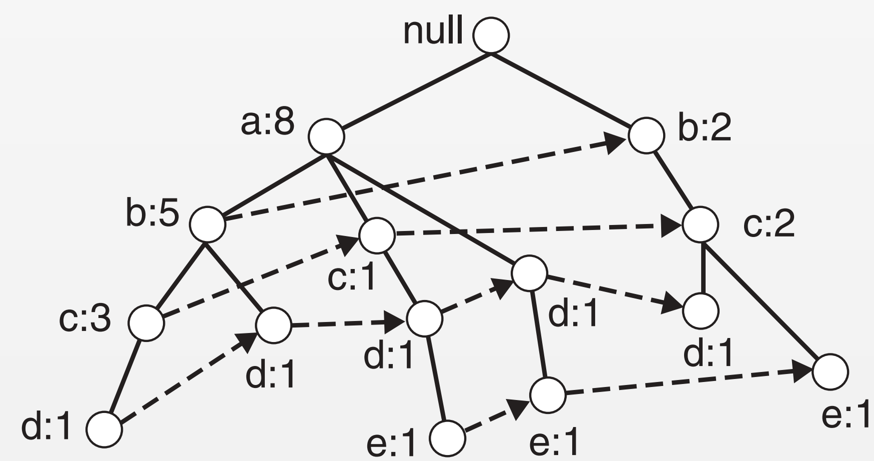


TID = 3



TID	Items Bought	Frequent Items
1	{a,b,f}	{a,b}
2	{b,g,c,d}	{b,c,d}
3	{h, a,c,d,e}	{a,c,d,e}
4	{a,d, p,e}	{a,d,e}
5	{a,b,c}	{a,b,c}
6	{a,b,q,c,d}	{a,b,c,d}
7	{a}	{a}
8	{a,m,b,c}	{a,b,c}
9	{a,b,n,d}	{a,b,d}
10	{b,c,e}	{b,c,e}

TID = 10



a: 8, b: 7, c: 6, d: 5, e: 3,

~~f: 1~~, ~~g: 1~~, ~~h: 1~~, ~~m: 1~~, ~~n: 1~~ ~~p: 1~~ ~~q: 1~~

adapted from: Tan, Steinbach & Kumar, "Introduction to Data Mining", http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Suboptimal FP-Tree

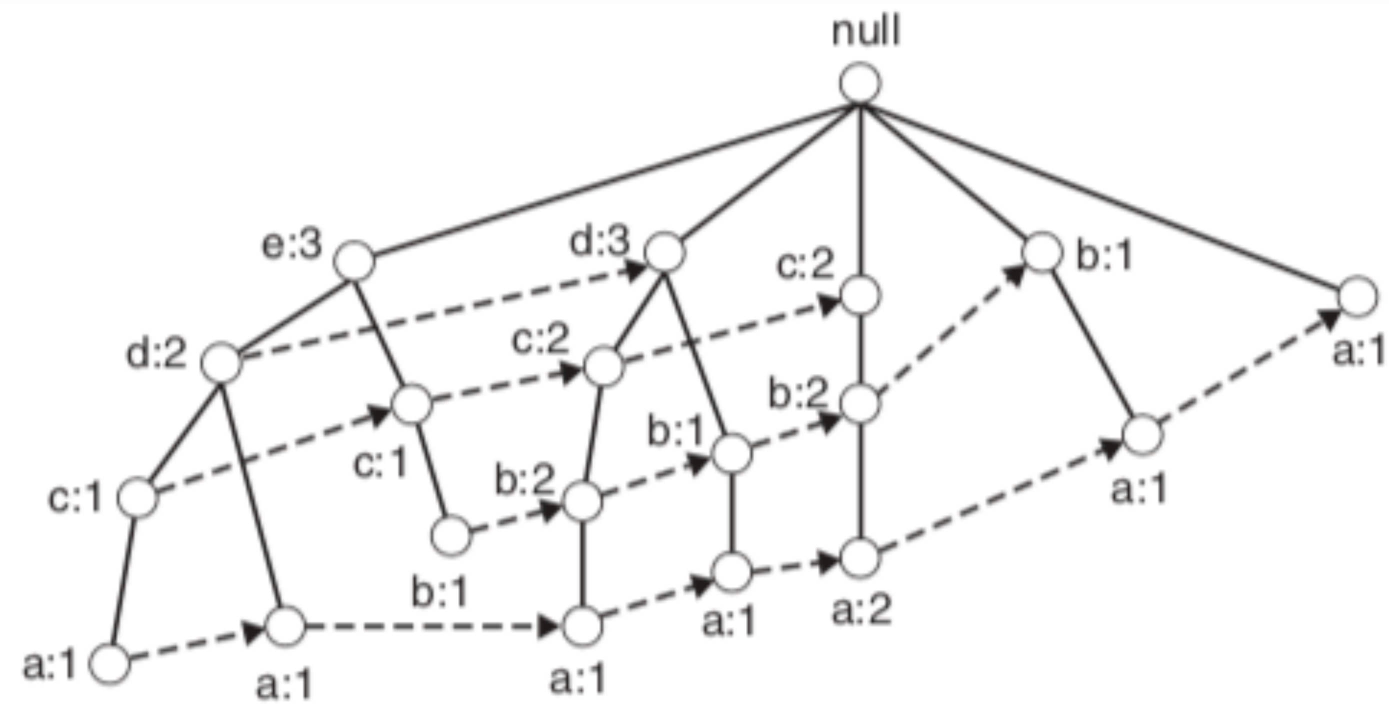


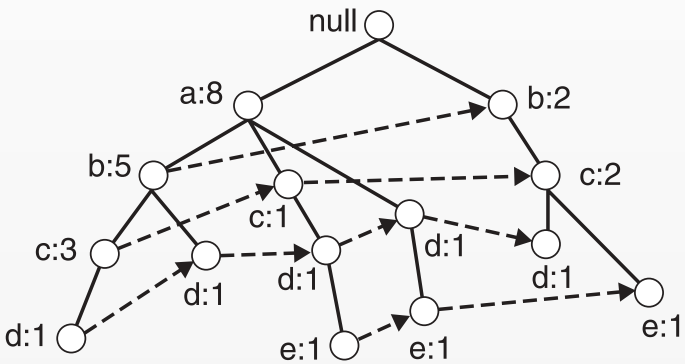
Figure 6.25. An FP-tree representation for the data set shown in Figure 6.24 with a different item ordering scheme.

Less compression: items were ordered from less frequent to more frequent.

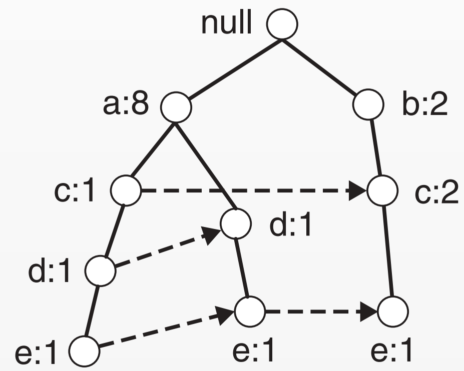
Mining Patterns from the FP-Tree

Step 1: Extract subtrees ending in each item

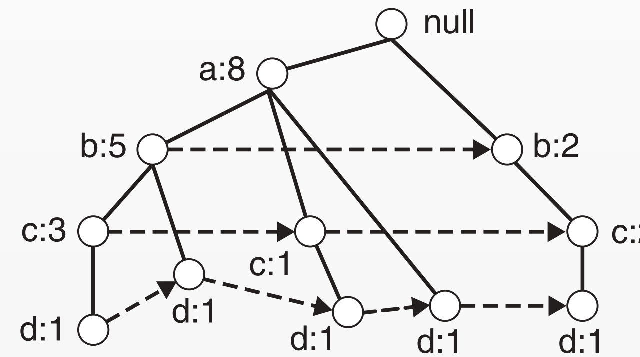
Full Tree



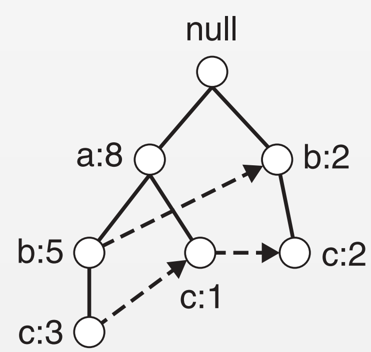
Subtree *e*



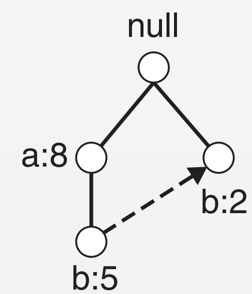
Subtree *d*



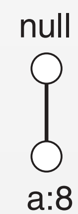
Subtree *c*



Subtree *b*



Subtree *a*



a: 8, b: 7, c: 6, d: 5, e: 3, ~~f: 1~~, ~~g: 1~~, ~~h: 1~~, ~~m: 1~~, ~~n: 1~~ ~~p: 1~~ ~~q: 1~~

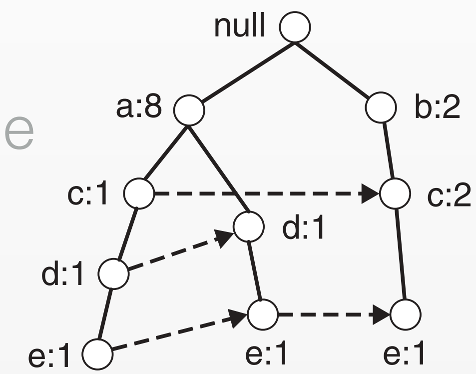
adapted from: Tan, Steinbach & Kumar, “Introduction to Data Mining”, http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Mining Patterns from the FP-Tree

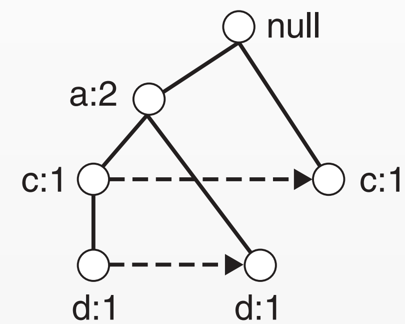
Step 2: Construct Conditional FP-Tree for each item

Subtree *e*

Prefix paths ending in e



Conditional *e*



{d,e}:2, {c,e}:2, {a,e}:2

Spawn 3 sub problems:

- Find FIP ending in {d,e}
- Find FIP ending in {c,e}
- Find FIP ending in {a,e}

Conditional Pattern Base for e

acd: 1, ad: 1, bc: 1

Conditional Node Counts

a: 2, ~~b: 1~~, c: 2, d: 2

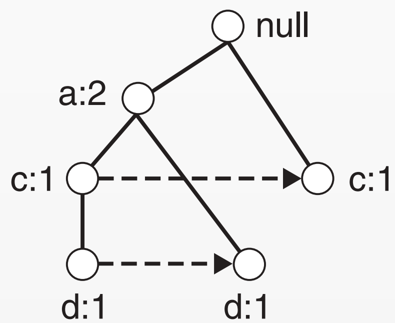
- Calculate counts for paths ending in *e*
- Remove leaf nodes
- Prune items with cumulative count < *s*

adapted from: Tan, Steinbach & Kumar, "Introduction to Data Mining", http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

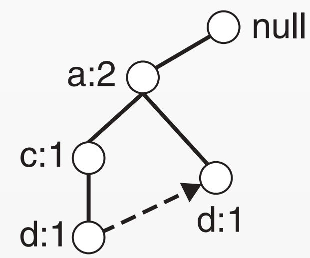
Mining Patterns from the FP-Tree

Step 3: Recursively mine conditional FP-Tree for each item

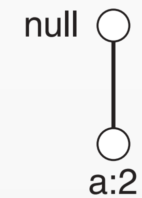
Conditional *e*



Subtree *de*

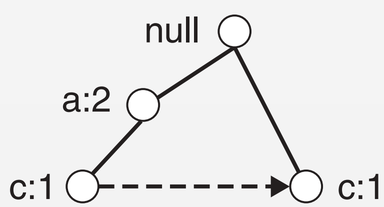


Conditional *de*



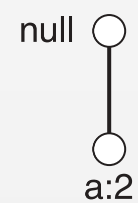
{a,d,e}:2

Subtree *ce*



Conditional *ce*

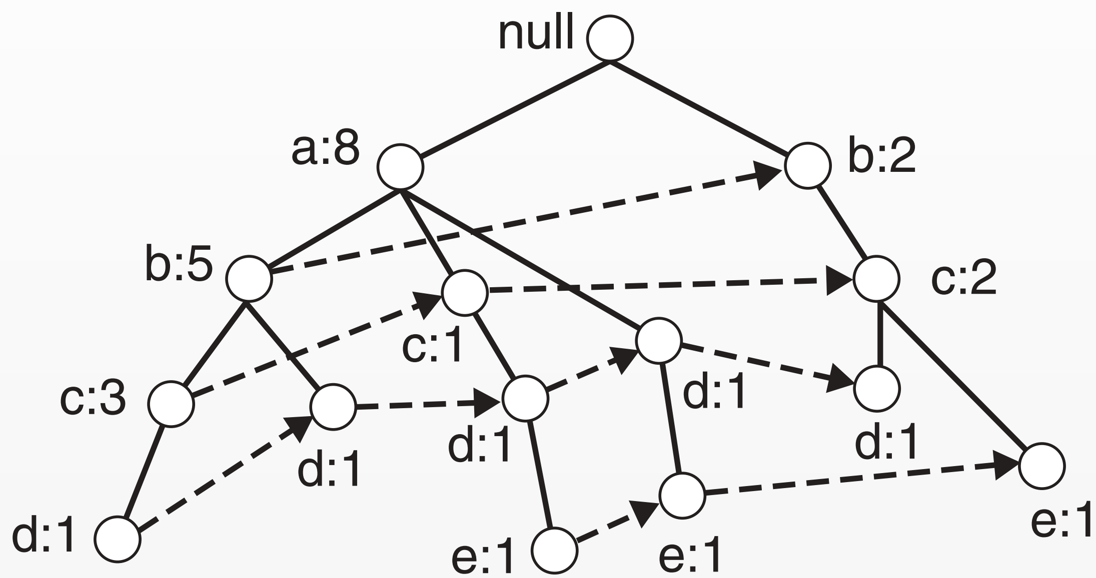
Subtree *ae*



Conditional *ae*

adapted from: Tan, Steinbach & Kumar, "Introduction to Data Mining", http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

Mining Patterns from the FP-Tree



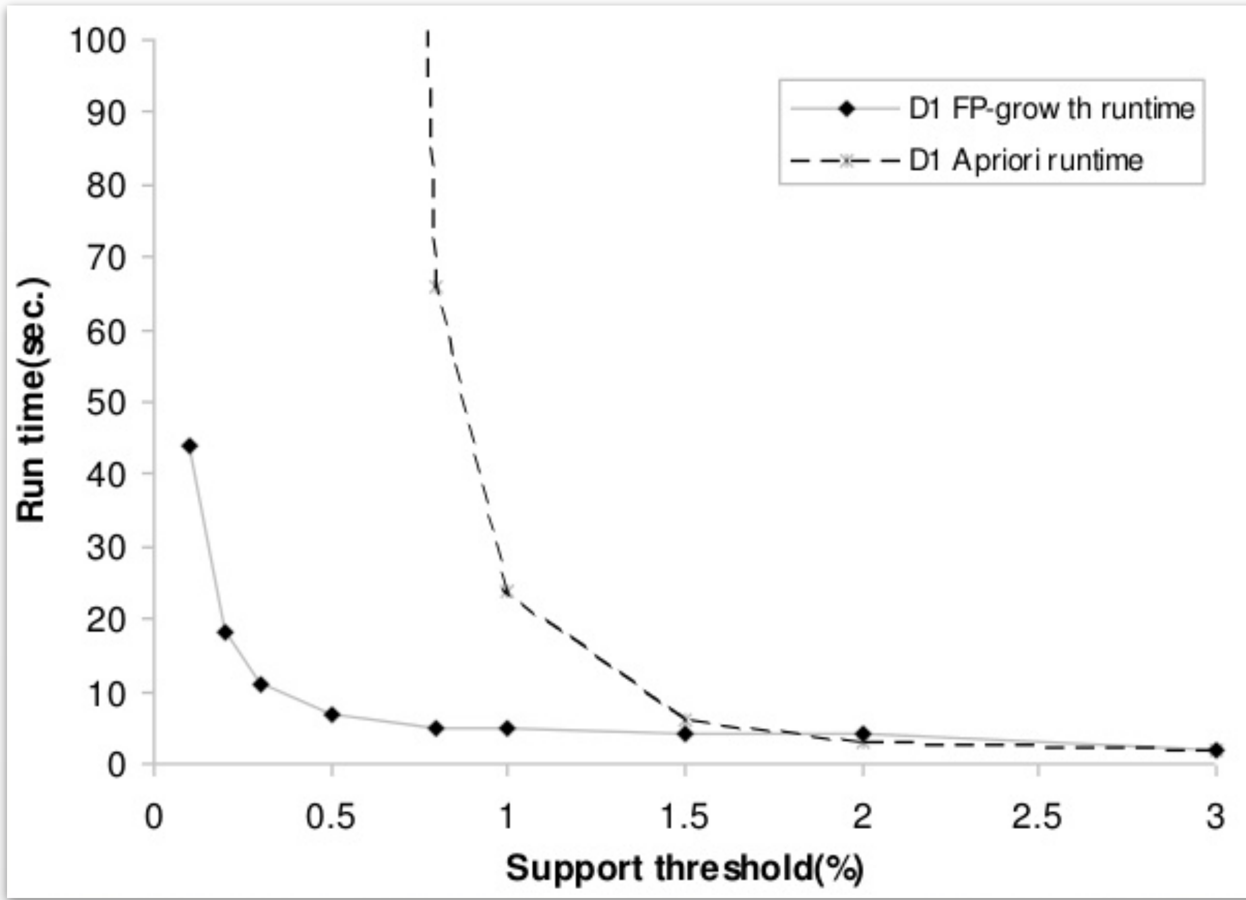
Suffix	Conditional Pattern Base
e	acd:1; ad:1; bc:1
d	abc:1; ab:1; ac:1; a:1; bc:1
c	ab:3; a:1; b:2
b	a:5
a	ϕ

Suffix	Frequent Itemsets
e	{e}, {d,e}, {a,d,e}, {c,e}, {a,e}
d	{d}, {c,d}, {b,c,d}, {a,c,d}, {b,d}, {a,b,d}, {a,d}
c	{c}, {b,c}, {a,b,c}, {a,c}
b	{b}, {a,b}
a	{a}

adapted from: Tan, Steinbach & Kumar, "Introduction to Data Mining", http://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf

FP-Growth vs Apriori

Simulated data 10k baskets, 25 items on average



(from: Han, Kamber & Pei, Chapter 6)

FP-Growth vs Apriori

File	Apriori	FP-Growth
Simple Market Basket test file	3.66 s	3.03 s
"Real" test file (1 Mb)	8.87 s	3.25 s
"Real" test file (20 Mb)	34 m	5.07 s
Whole "real" test file (86 Mb)	4+ hours (Never finished, crashed)	8.82 s

http://singularities.com/blog/2015/08/apriori-vs-fpgrowth-for-frequent-item-set-mining

FP-Growth vs Apriori

Advantages of FP-Growth

- Only 2 passes over dataset
- Stores “compact” summary of data
- No candidate generation
- Faster than A-priori

Disadvantages of FP-Growth

- The FP-Tree may not be “compact” enough to fit in memory
- *Used in practice*: PFP (distributed version of FP-growth)

Objective measures of interestingness

Limitations of confidence

Support and confidence look reasonable

$$s(\{\text{Tea}\} \rightarrow \{\text{Coffee}\}) = 150/1000 = 0.15$$
$$c(\{\text{Tea}\} \rightarrow \{\text{Coffee}\}) = 150/200 = 0.75$$

Table 6.8. Beverage preferences among a group of 1000 people.

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Inverse relationship

$$s(\{\text{Coffee}\}) = 0.8 > c(\{\text{Tea}\} \rightarrow \{\text{Coffee}\})$$

Since Coffee is very popular it occurs in many baskets with Tea just by chance.

Support of the consequent isn't accounted for!

Lift or Interest Factor

$$Lift(A,B) = \frac{c(A \rightarrow B)}{s(A)} = \frac{s(A \cup B)}{s(A)s(B)} = \frac{c(B \rightarrow A)}{s(B)}$$

Probabilistic Interpretation

$$Lift(A,B) = \frac{P(A,B)}{P(A)P(B)} \quad = \frac{P(A \subseteq T, B \subseteq T)}{P(A \subseteq T)P(B \subseteq T)}$$

Is a pure measure of statistical dependence

$$\begin{cases} = 1 & A \text{ and } B \text{ are independent} \\ > 1 & A \text{ and } B \text{ are positively correlated} \\ < 1 & A \text{ and } B \text{ are negativley correlated} \end{cases}$$

Table 6.8. Beverage preferences among a group of 1000 peopl

	$Coffee$	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

Inverse relationship detected by Lift

$$Lift = \frac{0.15}{0.2 \times 0.8} = 0.9375$$

Limitations of Lift

Cannot tell if an association is infrequent

Table 6.9. Contingency tables for the word pairs ($\{p,q\}$ and $\{r,s\}$.

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

$$Lift(p,q) = \frac{.88}{.93 \times .93} = 1.02$$

$$Lift(r,s) = \frac{.02}{.07 \times .07} = 4.08$$

Correlation Coefficient

$$\rho = \frac{\mathbf{E}[X_A X_B] - \mathbf{E}[X_A]\mathbf{E}[X_B]}{\sqrt{\mathbf{V}[X_A]\mathbf{V}[X_B]}}$$

Here, X_A represents random variables that take value 1 when a transaction contains A and 0 when it does not.

$$\rho = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

Table 6.7. A 2-way contingency table for variables A and B.

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

$$\begin{cases} = 1 & A \text{ and } B \text{ are uncorrelated} \\ > 1 & A \text{ and } B \text{ are positively correlated} \\ < 1 & A \text{ and } B \text{ are negativley correlated} \end{cases}$$

$$\rho = -0.0625$$

For {Tea}-> {Coffee}

Limitation of Correlation Coefficient

Treats absence of an item as important as its presence

Table 6.9. Contingency tables for the word pairs ($\{p,q\}$ and $\{r,s\}$.

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

Not useful for asymmetric binary data, where presence of an item is more important than its absence

$$\rho = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$$

$$\rho(p,q) = \rho(r,s) = 0.236$$

Cosine Similarity

$$cosine(A,B) = \frac{\overrightarrow{x}_A \cdot \overrightarrow{x}_B}{||\overrightarrow{x}_A|| \times ||\overrightarrow{x}_B||} \leq 1$$

$$= \frac{s(A \cup B)}{\sqrt{s(A)s(B)}}$$

$$= Lift(A,B) \times \sqrt{s(A)s(B)}$$

Here, $\overrightarrow{x}_A$ is a binary vector whose i[th] entry is 1, when the i[th] basket contains A, otherwise 0

Table 6.9. Contingency tables for the word pairs ($\{p,q\}$ and $\{r,s\}$.

	p	$\overline{p}$	
q	880	50	930
$\overline{q}$	50	20	70
	930	70	1000

	r	$\overline{r}$	
s	20	50	70
$\overline{s}$	50	880	930
	70	930	1000

It is expected to be large when both statistical dependence and supports of A and B are large. However, that might not always happen in practice.

$cosine(p,q) = 0.946$ $\quad$ $cosine(r,s) = 0.286$

Limitations of Cosine Similarity

Under independence
of A and B

$$cosine_{ind}(A,B) = \sqrt{s(A)s(B)}$$

$$\approx 1$$

If $s(A)$ and $s(B)$ are
close to 1.

$$cosine(A,B) = \frac{A \cdot B}{||A|| \times ||B||}$$

$$= \frac{s(A \cup B)}{\sqrt{s(A)s(B)}}$$

$$= Lift(A,B) \times \sqrt{s(A)s(B)}$$

Support alone can drive the value
close to 1, even under independence.

Table 6.10. Example of a contingency table for items p and q.

	q	$\overline{q}$	
p	800	100	900
$\overline{p}$	100	0	100
	900	100	1000

$$cosine(p,q) = 0.889$$

$$cosine_{ind}(p,q) = 0.9$$

Other measures

Table 6.7. A 2-way contingency table for variables A and B.

	B	$\overline{B}$	
A	f_{11}	f_{10}	f_{1+}
$\overline{A}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

$$M(A \to B) = M(B \to A)$$

$$M(A \to B) \neq M(B \to A)$$

Table 6.11. Examples of symmetric objective measures for the itemset $\{A, B\}$.

Measure (Symbol)	Definition
Correlation (ϕ)	$\frac{Nf_{11}-f_{1+}f_{+1}}{\sqrt{f_{1+}f_{+1}f_{0+}f_{+0}}}$
Odds ratio (α)	$(f_{11}f_{00})/(f_{10}f_{01})$
Kappa (κ)	$\frac{Nf_{11}+Nf_{00}-f_{1+}f_{+1}-f_{0+}f_{+0}}{N^2-f_{1+}f_{+1}-f_{0+}f_{+0}}$
Interest (I)	$(Nf_{11})/(f_{1+}f_{+1})$
Cosine (IS)	$(f_{11})/(\sqrt{f_{1+}f_{+1}})$
Piatetsky-Shapiro (PS)	$\frac{f_{11}}{N} - \frac{f_{1+}f_{+1}}{N^2}$
Collective strength (S)	$\frac{f_{11}+f_{00}}{f_{1+}f_{+1}+f_{0+}f_{+0}} \times \frac{N-f_{1+}f_{+1}-f_{0+}f_{+0}}{N-f_{11}-f_{00}}$
Jaccard (ζ)	$f_{11}/(f_{1+}+f_{+1}-f_{11})$
All-confidence (h)	$\min\left[\frac{f_{11}}{f_{1+}}, \frac{f_{11}}{f_{+1}}\right]$

Table 6.12. Examples of asymmetric objective measures for the rule $A \longrightarrow B$.

Measure (Symbol)	Definition
Goodman-Kruskal (λ)	$(\sum_j \max_k f_{jk} - max_k f_{+k})/(N - \max_k f_{+k})$
Mutual Information (M)	$(\sum_i \sum_j \frac{f_{ij}}{N} \log \frac{Nf_{ij}}{f_{i+}f_{+j}})/(-\sum_i \frac{f_{i+}}{N} \log \frac{f_{i+}}{N})$
J-Measure (J)	$\frac{f_{11}}{N} \log \frac{Nf_{11}}{f_{1+}f_{+1}} + \frac{f_{10}}{N} \log \frac{Nf_{10}}{f_{1+}f_{+0}}$
Gini index (G)	$\frac{f_{1+}}{N} \times (\frac{f_{11}}{f_{1+}})^2 + (\frac{f_{10}}{f_{1+}})^2] - (\frac{f_{+1}}{N})^2$ $+ \frac{f_{0+}}{N} \times [(\frac{f_{01}}{f_{0+}})^2 + (\frac{f_{00}}{f_{0+}})^2] - (\frac{f_{+0}}{N})^2$
Laplace (L)	$(f_{11}+1)/(f_{1+}+2)$
Conviction (V)	$(f_{1+}f_{+0})/(Nf_{10})$
Certainty factor (F)	$(\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N})/(1 - \frac{f_{+1}}{N})$
Added Value (AV)	$\frac{f_{11}}{f_{1+}} - \frac{f_{+1}}{N}$

Effect of skewed support distribution

High support threshold
- might miss many infrequent, but strong associations.

Low support threshold
- Significant computational needs
- Number extracted patterns might be too huge to analyze individually and take actions on.
- Cross-support patterns: Many spurious patterns relating high-frequency items like milk to low-frequency items like caviar.

For a support threshold of 0.05%:
- 18K patterns involving items from G1 and G3
- 93% are cross-support patterns.
- Maximum correlation for cross-support patterns is 0.029
- Max Correlation for items from same group as high as 1.0

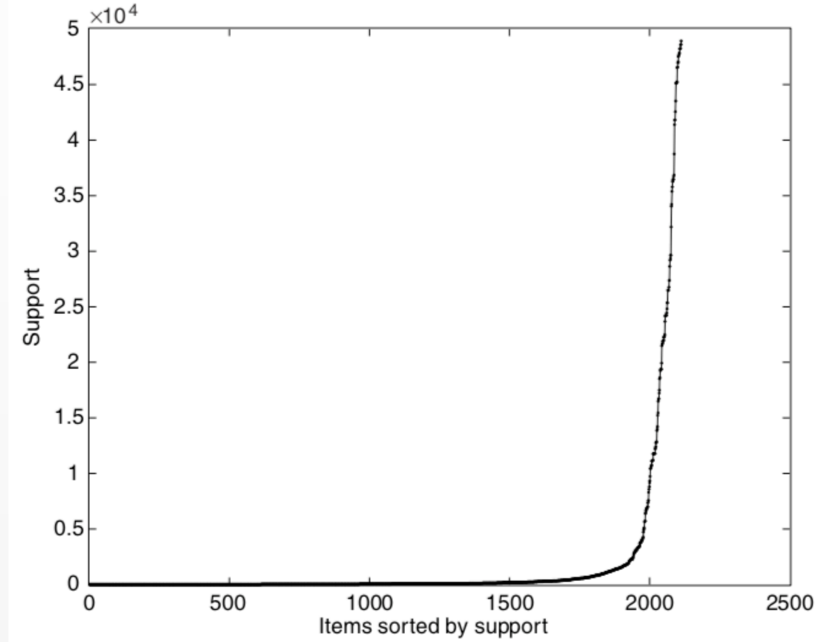


Figure 6.29. Support distribution of items in the census data set.

Table 6.21. Grouping the items in the census data set based on their support values.

Group	G_1	G_2	G_3
Support	$< 1\%$	$1\% - 90\%$	$> 90\%$
Number of Items	1735	358	20

Effect of skewed support distribution

Definition 6.9 (Cross-Support Pattern). A cross-support pattern is an itemset $X = \{i_1, i_2, \ldots, i_k\}$ whose support ratio

$$r(X) = \frac{\min\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}{\max\left[s(i_1), s(i_2), \ldots, s(i_k)\right]}, \qquad (6.13)$$

is less than a user-specified threshold h_c.

- {p,q}, {p,r}, {p,q,r} (support ratio = 0.2) are cross-support items at $h_c = 0.3$
- Using a support threshold to 0.2 would eliminate the cross-support items, but also eliminate interesting pattern {q,r} (support = 0.167)
- Confidence pruning doesn't help either. {q} -> {p} has high confidence (=0.8)
- However, confidence of inverse association {p} -> {q} is low. Can this observation be exploited to prune {p,q}?
- Solution: Prune itemsets based on their lowest confident rule.

p	q	r
0	1	1
1	1	1
1	1	1
1	1	1
1	1	1
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
1	0	0
0	0	0
0	0	0
0	0	0
0	0	0

Effect of skewed support distribution

Extension to larger itemset

- For a given itemset $X = \{i_1, i_2 \ldots i_k\}$, which association rule has the lowest confidence?
- Anti-monotone property of confidence

$$conf(\{i_1 i_2\} \longrightarrow \{i_3, i_4, \ldots, i_k\}) \leq conf(\{i_1 i_2 i_3\} \longrightarrow \{i_4, i_5, \ldots, i_k\}).$$

- Association rule with the smallest confidence will have a singleton antecedent.

$$\{i_j\} \longrightarrow \{i_1, i_2, \ldots, i_{j-1}, i_{j+1}, \ldots, i_k\}$$

- It is the antecedent with the highest support.

$$s(i_j) = \max\left[s(i_1), s(i_2), \ldots, s(i_k)\right].$$

- Lowest attainable confidence is

$$\text{h-confidence}(X) = \frac{s(\{i_1, i_2, \ldots, i_k\})}{\max[s(i_1), s(i_2) \ldots s(i_k)]} \leq \frac{\min[s(i_1), s(i_2) \ldots s(i_k)]}{\max[s(i_1), s(i_2) \ldots s(i_k)]}$$

- Anti-monotone property: Can be directly incorporated in the algorithm to assist pruning.

$$\text{h-confidence}(\{i_1, i_2, \ldots, i_k\}) \geq \text{h-confidence}(\{i_1, i_2, \ldots, i_{k+1}\}),$$

- Strongly associated patterns: h-confidence of 80% implies that if one of the transactions is in the basket the probability that the others are in the basket too is at least 80%