

Association Rule Mining Review

Florian Verhein
fverhein@it.usyd.edu.au

School of Information Technologies,
The University of Sydney,
Australia

Copyright 2008 Florian Verhein. Figures from [1].

January 10, 2008



Introduction

- ▶ Set of *items* $I = \{i_1, i_2, \dots, i_m\}$
- ▶ Set of *transactions* $T = \{t_1, t_2, \dots, t_n\}$, where each $t_i \subseteq I$
- ▶ **Example:** $I = \{Bread, Milk, Diapers, Beer, Eggs, Cola\}$:

<i>TID</i>	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

- ▶ An *itemset* is a collection of one or more *items*. A “*k*-itemset” has size k
- ▶ The *support* of an *itemset* $I' \subseteq I$ is the number of transactions containing I' :
 - ▶ $support(I') = |\{t \in T : I' \subseteq t\}|$



Association Rule Mining

- ▶ Search for *interesting* rules of the form $X \rightarrow Y$, where X and Y are itemsets
 - ▶ E.g. $\{Bread\} \rightarrow \{Jelly\}$ or $\{Bread, Jelly\} \rightarrow \{PeanutButter\}$
- ▶ **Formally:** Given $I = \{i_1, \dots, i_m\}$ and $T = \{t_1, \dots, t_n\}$ an *association rule* is an implication of the form $X \rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.
- ▶ What association rules are *interesting*?
 - ▶ rules with $support \geq minSup$ and $confidence \geq minConf$
 - ▶ $support(X \rightarrow Y) = support(X \cup Y)$
 - ▶ $confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \sim P(Y|X)$
- ▶ **Problem definition:** find all *interesting* rules and do so *efficiently*.



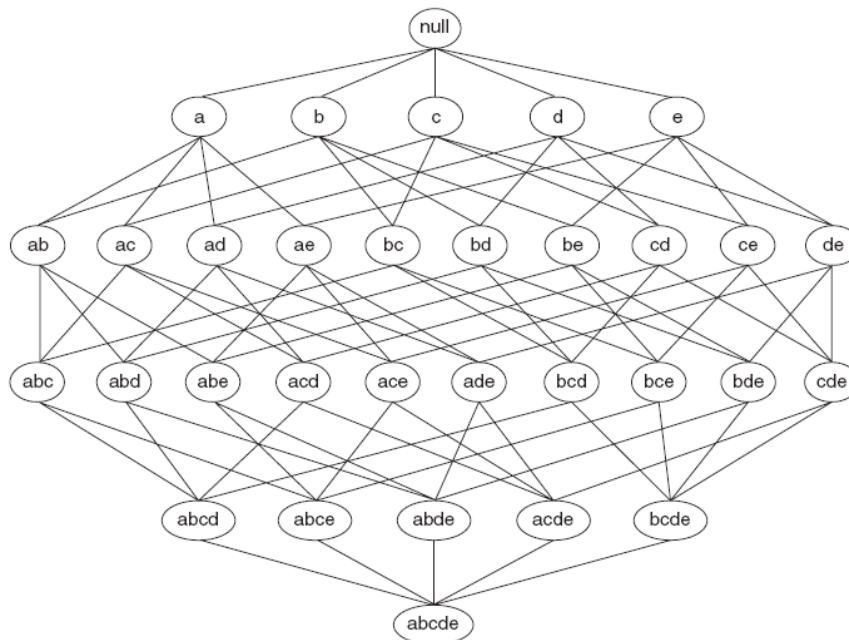
Finding Association Rules

- ▶ **Naive:** Enumerate all possible rules and select those that satisfy $minSup$ and $minConf$ thresholds
 - ▶ Not practical!
 - ▶ For a data-set with m items, the number of possible rules is $3^m - 2^{m+1} + 1$ (why? *hint: use the inclusion exclusion principle*)
 - ▶ most of these will be discarded!
- ▶ **Note:** what do $\{a, b\} \rightarrow \{c\}$, $\{a, c\} \rightarrow \{b\}$ and $\{b, c\} \rightarrow \{a\}$ have in common? – same support.
- ▶ So, find all *frequent itemsets* (itemsets with $support \geq minSup$) and only generate rules from these.
- ▶ 2 step procedure:
 - ▶ Step 1: Frequent Itemset Mining (FIM) (satisfy $minSup$). Most computationally expensive.
 - ▶ Step 2: Association rule generation (satisfy $minConf$)



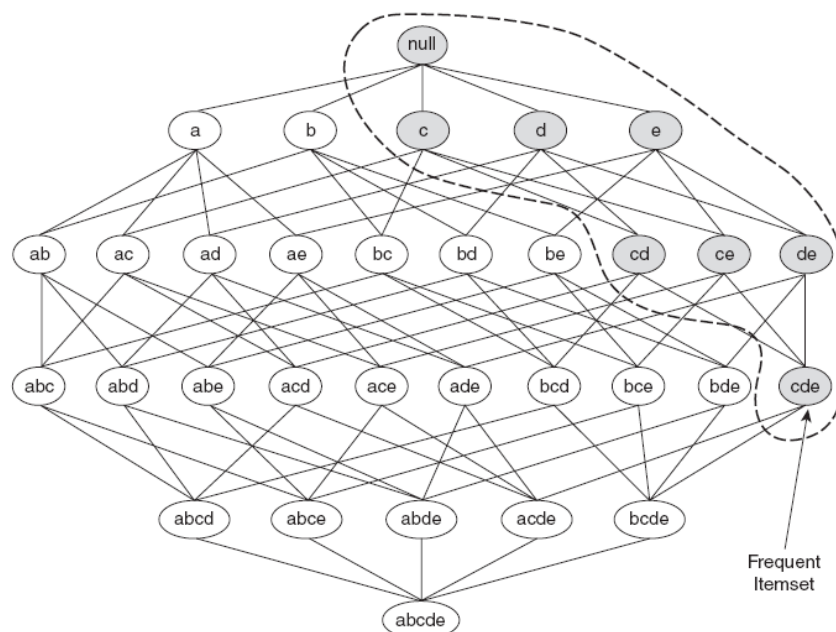
Frequent Itemset Mining

- ▶ Search space: $2^m - 1$ (enumeration not practical)
- ▶ **Example:** $I = \{a, b, c, d, e\}$. Lattice showing superset/subset relationships.



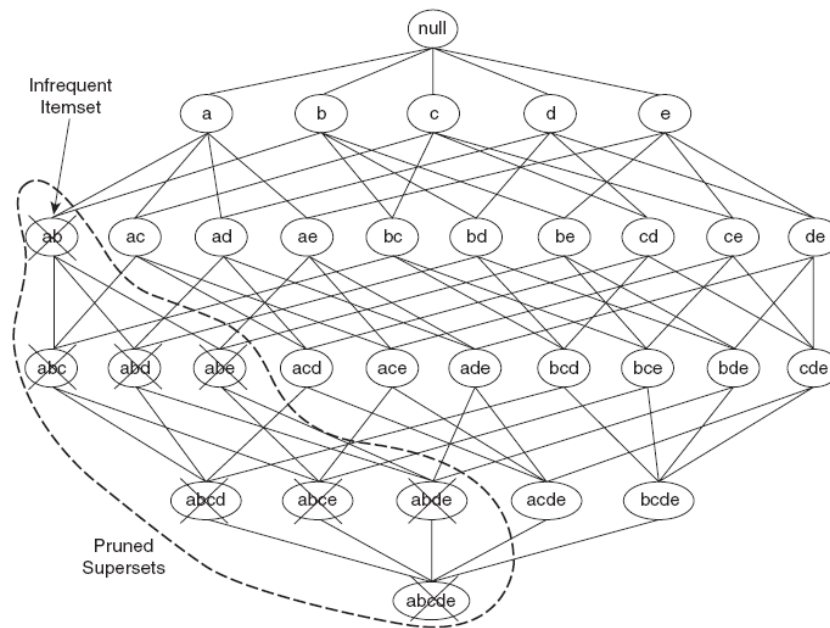
Anti-monotonicity of Support

- ▶ All subsets of a frequent itemset are also frequent



Anti-monotonicity of Support

- ▶ All *supersets* of an *infrequent* itemset are also *infrequent*.
- ▶ “Apriori principle”, used for pruning the search space.



Some Frequent Itemset Mining Algorithms

- ▶ Apriori
- ▶ FP-Growth
- ▶ GLIMIT



Apriori

Algorithm 6.1 Frequent itemset generation of the *Apriori* algorithm.

```
1:  $k = 1$ .
2:  $F_k = \{ i \mid i \in I \wedge \sigma(\{i\}) \geq N \times \text{minsup} \}$ .   {Find all frequent 1-itemsets}
3: repeat
4:    $k = k + 1$ .
5:    $C_k = \text{apriori-gen}(F_{k-1})$ .   {Generate candidate itemsets}
6:   for each transaction  $t \in T$  do
7:      $C_t = \text{subset}(C_k, t)$ .   {Identify all candidates that belong to  $t$ }
8:     for each candidate itemset  $c \in C_t$  do
9:        $\sigma(c) = \sigma(c) + 1$ .   {Increment support count}
10:    end for
11:  end for
12:   $F_k = \{ c \mid c \in C_k \wedge \sigma(c) \geq N \times \text{minsup} \}$ .   {Extract the frequent  $k$ -itemsets}
13: until  $F_k = \emptyset$ 
14: Result =  $\bigcup F_k$ .
```

Various possibilities for *apriori-gen*(\cdot) – $F_k \times F_1, F_{k-1} \times F_{k-1}$.

Various possibilities for *subset*(\cdot). See [1].



References

- ▶ [1] Pang-Ning Tan, Michael Steinbach, Vipin Kumar:
Introduction to Data Mining, Addison-Wesley.
 - ▶ Chapter 6: *Association Analysis: Basic Concepts and Algorithms*.
 - ▶ Available from

http:

[//www-users.cs.umn.edu/~kumar/dmbook/index.php](http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

