Association Rules

1

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

2

Association Rule Mining

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction

Market-Basket transactions

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Association Rules

{Diaper} → {Beer},
{Milk, Bread} → {Eggs,Coke},
{Beer, Bread} → {Milk},

Implication means co-occurrence, not causality!

3

Applications

- **Market Basket Analysis:** given a database of customer transactions, where each transaction is a set of items the goal is to find groups of items which are frequently purchased together.
- **Telecommunication** (each customer is a transaction containing the set of phone calls)
- **Credit Cards/ Banking Services** (each card/account is a transaction containing the set of customer's payments)
- **Medical Treatments** (each patient is represented as a transaction containing the ordered set of diseases)
- **Basketball-Game Analysis** (each game is represented as a transaction containing the ordered set of ball passes)

4

Definition: Frequent Itemset

- **Itemset**
  - A collection of one or more items
    - Example: {Milk, Bread, Diaper}
  - k-itemset
    - An itemset that contains k items
- **Support count (σ)**
  - Frequency of occurrence of an itemset
  - E.g. σ({Milk, Bread,Diaper}) = 2
- **Support**
  - Fraction of transactions that contain an itemset
  - E.g. s({Milk, Bread, Diaper}) = 2/5
- **Frequent Itemset**
  - An itemset whose support is greater than or equal to a *minsup* threshold

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

5

Definition: Association Rule

- **Association Rule**
  - An implication expression of the form X $\rightarrow$ Y, where X and Y are itemsets
  - Example:
    {Milk, Diaper} $\rightarrow$ {Beer}

TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

- **Rule Evaluation Metrics**
  - Support (s)
    - Fraction of transactions that contain both X and Y
  - Confidence (c)
    - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

6

Association Rule Mining Task

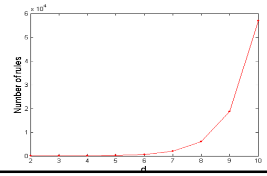
- Given a set of transactions T, the goal of association rule mining is to find all rules having
  - support ≥ *minsup* threshold
  - confidence ≥ *minconf* threshold

7

Association Rule Mining Task

- Brute-force approach:
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the *minsup* and *minconf* thresholds

  $\Rightarrow$ Computationally prohibitive!
- Note that given d unique items:
  - Total number of itemsets = 2^d
  - Total number of possible association rules:



$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right] = 3^d - 2^{d+1} + 1$$

If d=6, R = 602 rules

8

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

9

How to make Efficient Mining Association Rules

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Example of Rules:

{Milk,Diaper} → {Beer} (s=0.4, c=0.67)
{Milk,Beer} → {Diaper} (s=0.4, c=1.0)
{Diaper,Beer} → {Milk} (s=0.4, c=0.67)
{Beer} → {Milk,Diaper} (s=0.4, c=0.67)
{Diaper} → {Milk,Beer} (s=0.4, c=0.5)
{Milk} → {Diaper,Beer} (s=0.4, c=0.5)

Observations:

- All the above rules are binary partitions of the same itemset: {Milk, Diaper, Beer}
- Rules originating from the same itemset have identical support but can have different confidence
- ***Thus, we may decouple the support and confidence requirements!***

10

Mining Association Rules: Problem Decomposition

- Two-step approach:
  1. Frequent Itemset Generation
     - Generate all itemsets whose support ≥ minsup
  2. Rule Generation
     - Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset
- Frequent itemset generation is still computationally expensive

11

Mining Association Rules: Problem Decomposition

Transaction ID	Items Bought
1	Shoes, Shirt, Jacket
2	Shoes,Jacket
3	Shoes, Jeans
4	Shirt, Sweatshirt

If the minimum support is 50%, then {Shoes,Jacket} is the only 2- itemset that satisfies the minimum support.

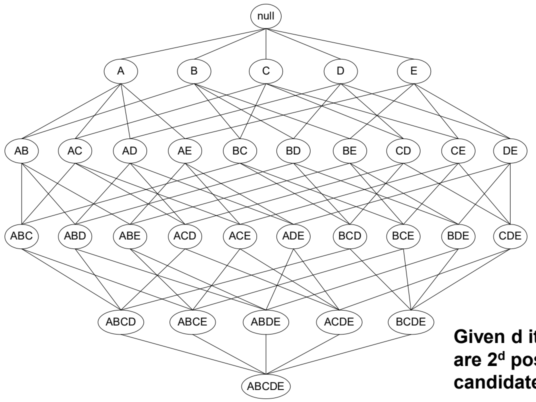
Frequent Itemset	Support
{Shoes}	75%
{Shirt}	50%
{Jacket}	50%
{Shoes, Jacket}	50%

If the minimum confidence is 50%, then the only two rules generated from this 2-itemset, that have confidence greater than 50%, are:

Shoes ⇒ Jacket Support=50%, Confidence=66%
Jacket ⇒ Shoes Support=50%, Confidence=100%

12

Frequent Itemset Generation: Complexity

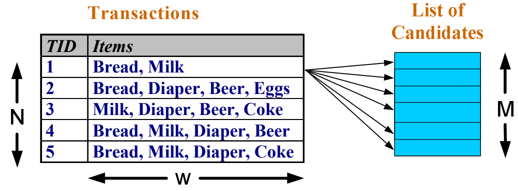


Given d items, there are 2^d possible candidate itemsets

13

Frequent Itemset Generation: Complexity

- Brute-force approach:
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database



TID	*Items*
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => Expensive since $M = 2^d$!!!

14

Frequent Itemset Generation Strategies

- Reduce the number of candidates (M)
  - Complete search: $M=2^d$
  - Use pruning techniques to reduce M
- Reduce the number of transactions (N)
  - Reduce size of N as the size of itemset increases
  - Used by vertical-based mining algorithms

15

Reducing Number of Candidates

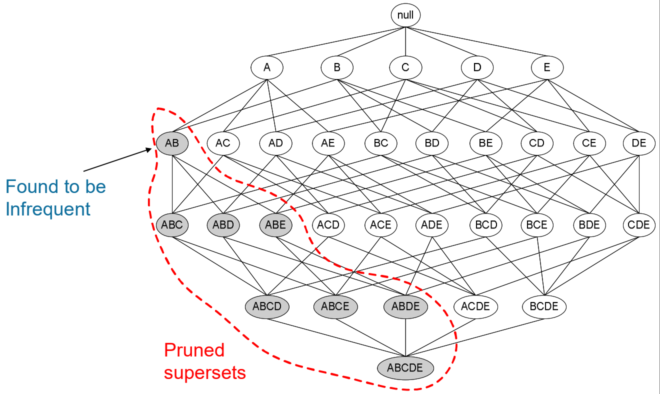
- Apriori principle:
  - If an itemset is frequent, then all of its subsets must also be frequent
- Apriori principle holds due to the following property of the support measure:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

  - Support of an itemset never exceeds the support of its subsets
  - This is known as the anti-monotone property of support

16

Illustrating Apriori Principle



17

Illustrating Apriori Principle

Items (1-itemsets)

Item	Count
Bread	4
Coke	2
Milk	4
Beer	3
Diaper	4
Eggs	1

Minimum Support = 3

Pairs (2-itemsets)

(No need to generate candidates involving Coke or Eggs)

Itemset	Count
{Bread,Milk}	3
{Bread,Beer}	2
{Bread,Diaper}	3
{Milk,Beer}	2
{Milk,Diaper}	3
{Beer,Diaper}	3

Triplets (3-itemsets)

Itemset	Count
{Bread,Milk,Diaper}	3

If every subset is considered,
$^6C_1 + {}^6C_2 + {}^6C_3 = 41$
With support-based pruning,
6 + 6 + 1 = 13

18

Apriori Algorithm

- Let k=1
- Generate frequent itemsets of length 1
- Repeat until no new frequent itemsets are identified
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Prune candidate itemsets containing subsets of length k that are infrequent
  - Count the support of each candidate by scanning the DB
  - Eliminate candidates that are infrequent, leaving only those that are frequent

19

The Apriori Algorithm — Example

Min support =50%

Database D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

20

How to Generate Candidates

Input: L_{i-1} : set of frequent itemsets of size i-1

Output: C_i : set of candidate itemsets of size i

```
Ci = empty set;
for each itemset J in Li-1 do
    for each itemset K in Li-1 s.t. K<> J do
        if i-2 of the elements in J and K are equal then
            if all subsets of {K ∪ J} are in Li-1 then
                Ci = Ci ∪ {K ∪ J}
return Ci;
```

21

Example of Generating Candidates

- *L_3={abc, abd, acd, ace, bcd}*
- Generating *C_4* from *L_3*
  – *abcd* from *abc* and *abd*
  – *acde* from *acd* and *ace*
- Pruning:
  – *acde* is removed because *ade* is not in *L_3*
- *C_4={abcd}*

22

Experiment Results

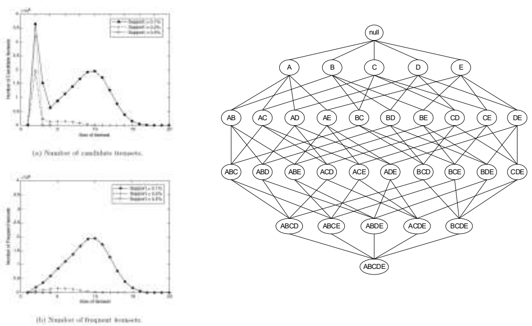


Figure 6.13. Effect of support threshold on the number of candidate and frequent itemsets.

23

Rule Generation

- Given a frequent itemset L, find all non-empty subsets $f \subset L$ such that $f \rightarrow L - f$ satisfies the minimum confidence requirement
  – If {A,B,C,D} is a frequent itemset, candidate rules:

ABC →D,	ABD →C,	ACD →B,	BCD →A,
A →BCD,	B →ACD,	C →ABD,	D →ABC
AB →CD,	AC → BD,	AD → BC,	BC →AD,
BD →AC,	CD →AB,		

- If |L| = k, then there are $2^k - 2$ candidate association rules (ignoring $L \rightarrow \emptyset$ and $\emptyset \rightarrow L$)

24

Rule Generation: Brute Force Approach

for each frequent itemset I **do**
for each subset C of I **do**
if (support(I) / support(I - C) >= minconf) **then**
output the rule (I - C) $\Rightarrow C$,
with confidence = support(I) / support (I - C)
and support = support(I)

25

Rule Generation Example: Brute Force Approach

TID	List of Item_IDs
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Let use consider the 3-itemset {I1, I2, I5} with support of 0.22(2)%. Let generate all the association rules from this itemset:

I1 $\wedge$ I2 $\Rightarrow$ I5 *confidence*= 2/4 = 50%
I1 $\wedge$ I5 $\Rightarrow$ I2 *confidence*= 2/2 = 100%
I2 $\wedge$ I5 $\Rightarrow$ I1 *confidence*= 2/2 = 100%
I1 $\Rightarrow$ I2 $\wedge$ I5 *confidence*= 2/6 = 33%
I2 $\Rightarrow$ I1 $\wedge$ I5 *confidence*= 2/7 = 29%
I5 $\Rightarrow$ I1 $\wedge$ I2 *confidence*= 2/2 = 100%

26

Efficient Rule Generation

- How to efficiently generate rules from frequent itemsets?
  - In general, confidence does not have an anti-monotone property
    c(ABC $\rightarrow$D) can be larger or smaller than c(AB $\rightarrow$CD)
  - But confidence of rules generated from the same itemset has an anti-monotone property
  - e.g., L = {A,B,C,D}:

    c(ABC $\rightarrow$ D) $\geq$ c(AB $\rightarrow$ CD) $\geq$ c(A $\rightarrow$ BCD)

    - Confidence is anti-monotone w.r.t. number of items on the RHS of the rule

27

Efficient Rule Generation

Theorem. Consider a non-empty itemset Y and a non-empty itemset $X \subseteq Y$. Then:

$$c(X \rightarrow Y \setminus X) \geq c(X' \rightarrow Y \setminus X')$$

$$where \qquad X' \subseteq X.$$

Proof:

$$c(X \rightarrow Y \setminus X) = \frac{\sigma(Y)}{\sigma(X)} \qquad and$$

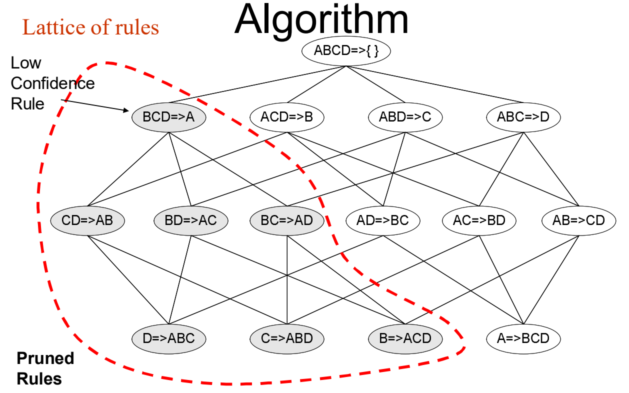
$$c(X' \rightarrow Y \setminus X') = \frac{\sigma(Y)}{\sigma(X')}.$$

$$But, \quad \sigma(X) \leq \sigma(X'). Thus,$$

$$c(X \rightarrow Y \setminus X) \geq c(X' \rightarrow Y \setminus X').$$

28

Rule Generation for Apriori Algorithm



29

Factors Affecting Complexity

- Choice of minimum support threshold
  - lowering support threshold results in more frequent itemsets
  - this may increase number of candidates and max length of frequent itemsets
- Dimensionality (number of items) of the data set
  - more space is needed to store support count
  - if number of frequent items also increases, both computation and I/O costs may also increase
- Size of database
  - since Apriori makes multiple passes, run time of algorithm may increase with number of transactions
- Average transaction width increases max length of frequent itemsets

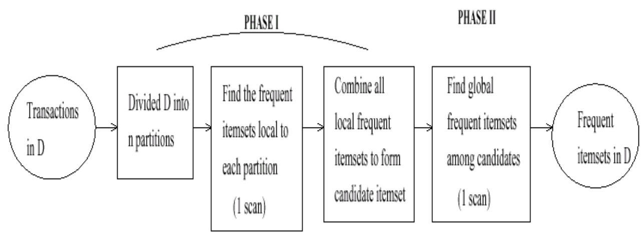
30

Further Improvement of the Apriori Method

- Major computational challenges
  - Multiple scans of transaction database
  - Huge number of candidates
  - Tedious workload of support counting for candidates
- Improving Apriori: general ideas
  - Reduce passes of transaction database scans
  - Shrink number of candidates
  - Reduce data size

31

Partitioning



32

Transaction reduction

A transaction that does not contain any frequent *k*-itemset will not contain frequent *l*-itemset for $l > k$! Thus, it is useless in subsequent scans!

33

Sampling

Mining on a subset of given data, lower support threshold + a method to determine the completeness

34

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

35

Compact Representation of Frequent Itemsets

- Some itemsets are redundant because they have identical support as their supersets

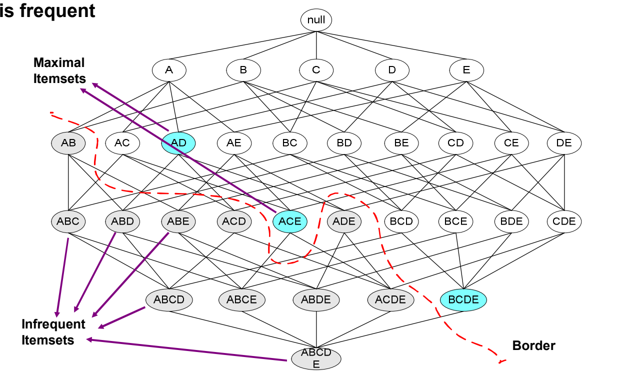
TID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1

- Number of frequent itemsets $= 3 \times \sum_{k=1}^{10} \binom{10}{k}$
- Need a compact representation

36

Maximal Frequent Itemset

An itemset is maximal frequent if none of its immediate supersets is frequent



37

Closed Itemset

- An itemset is closed if none of its immediate supersets has the same support as the itemset

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

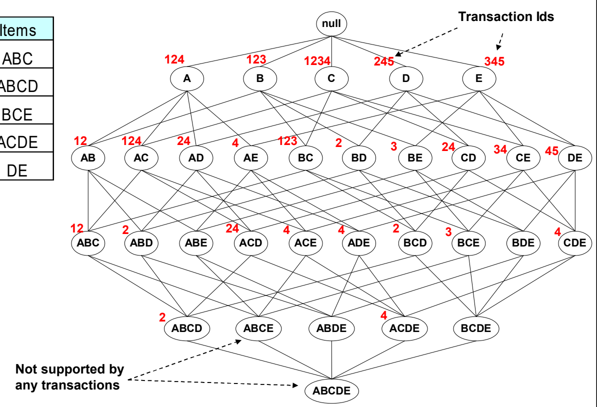
Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	3
{A,B,C,D}	2

38

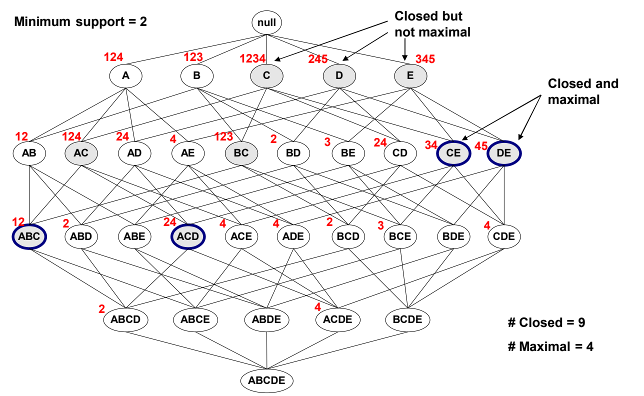
Maximal vs Closed Itemsets

TID	Items
1	ABC
2	ABCD
3	BCE
4	ACDE
5	DE



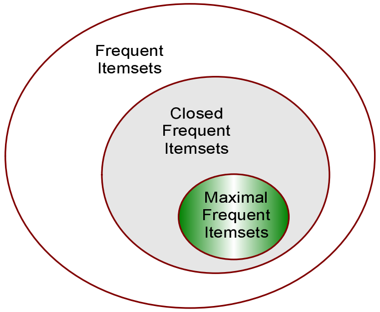
39

Maximal vs Closed Frequent Itemsets



40

Maximal vs Closed Itemsets



41

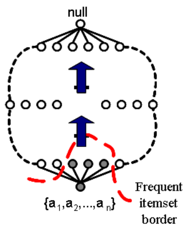
Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

42

Association Rule Problem: second variant

- Given:
  - — a set I of all the items;
  - — a database D of transaction
  - — maximum support M;
  - — minimum confidence c;
- Find:
  - — all association rules $X \Rightarrow Y$ with support smaller than M and confidence greater than c.

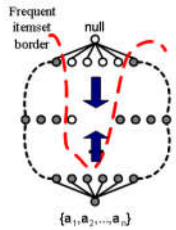


Specific-to-general

43

Association Rule Problem: third variant

- Given:
  - — a set I of all the items;
  - — a database D of transactions;
  - — minimum support m;
  - — maximum support M;
  - — minimum confidence c;
- Find:
  - — all association rules $X \Rightarrow Y$ with support smaller than M and greater than m and confidence greater than c.



Bidirectional

44

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

45

Alternative Measures for Association Rules

- The **confidence** of $X \Rightarrow Y$ in database D is the ratio of the number of transactions containing $X \cup Y$ to the number of transactions that contain X. In other words it is:

$$conf(X \rightarrow Y) = \frac{\frac{\sigma(X \cup Y)}{|D|}}{\frac{\sigma(X)}{|D|}} = \frac{p(X \wedge Y)}{p(X)} = p(Y \mid X)$$

- But, when Y is independent of X: $p(Y) = p(Y \mid X)$. In this case if $p(Y)$ is high we'll have a rule with high confidence that associate independent itemsets! For example, if p("*buy milk*") = 80% and "*buy milk*" is independent from "*buy salmon*", then the rule "*buy salmon*" $\Rightarrow$ "*buy milk*" will have confidence 80%!

46

Alternative Measures for Association Rules

- The **lift** measure indicates the departure from independence of X and Y. The lift of $X \Rightarrow Y$ is :

$$lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{p(Y)} = \frac{\frac{p(X \wedge Y)}{p(X)}}{p(Y)} = \frac{p(X \wedge Y)}{p(X)p(Y)}$$

- But, the lift measure is symmetric; i.e., it does not take into account the direction of implications!
- If lift is greater than 1, then X and Y are ***positively*** correlated; i.e., the occurrence of X (Y) imply occurrence of Y(X).
- If lift is smaller than 1, then X and Y are ***negatively*** correlated; i.e., the occurrence of X (Y) imply absence of Y(X).

47

Alternative Measures for Association Rules

- The **conviction** measure indicates the departure from independence of X and Y taking into account the implication direction. The conviction of $X \Rightarrow Y$ is :

$$conv(X \rightarrow Y) = \frac{p(X)p(\neg Y)}{p(X \wedge \neg Y)}$$

48

Alternative Measures for Association Rules

#	Measure	Formula
1	ϕ-coefficient	$\frac{P(A,B)-P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$
2	Goodman-Kruskal's (λ)	$\frac{\sum_j \max_k P(A_j,B_k)+\sum_k \max_j P(A_j,B_k)-\max_j P(A_j)-\max_k P(B_k)}{2-\max_j P(A_j)-\max_k P(B_k)}$
3	Odds ratio (α)	$\frac{P(A,B)P(\overline{A},\overline{B})}{P(A,\overline{B})P(\overline{A},B)}$
4	Yule's Q	$\frac{P(A,B)P(\overline{AB})-P(A,\overline{B})P(\overline{A},B)}{P(A,B)P(\overline{AB})+P(A,\overline{B})P(\overline{A},B)} = \frac{\alpha-1}{\alpha+1}$
5	Yule's Y	$\frac{\sqrt{P(A,B)P(\overline{AB})}-\sqrt{P(A,\overline{B})P(\overline{A},B)}}{\sqrt{P(A,B)P(\overline{AB})}+\sqrt{P(A,\overline{B})P(\overline{A},B)}} = \frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
6	Kappa (κ)	$\frac{P(A,B)+P(\overline{A},\overline{B})-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A)P(B)-P(\overline{A})P(\overline{B})}$
7	Mutual Information (M)	$\frac{\sum_i \sum_j P(A_i,B_j)\log\frac{P(A_i,B_j)}{P(A_i)P(B_j)}}{\min(-\sum_i P(A_i)\log P(A_i),-\sum_j P(B_j)\log P(B_j))}$
8	J-Measure (J)	$\max\Big(P(A,B)\log(\frac{P(B\|A)}{P(B)})+P(A\overline{B})\log(\frac{P(\overline{B}\|A)}{P(\overline{B})}),$ $P(A,B)\log(\frac{P(A\|B)}{P(A)})+P(\overline{A}B)\log(\frac{P(\overline{A}\|B)}{P(\overline{A})})\Big)$
9	Gini index (G)	$\max\Big(P(A)[P(B\|A)^2+P(\overline{B}\|A)^2]+P(\overline{A})[P(B\|\overline{A})^2+P(\overline{B}\|\overline{A})^2]$ $-P(B)^2-P(\overline{B})^2,$ $P(B)[P(A\|B)^2+P(\overline{A}\|B)^2]+P(\overline{B})[P(A\|\overline{B})^2+P(\overline{A}\|\overline{B})^2]$ $-P(A)^2-P(\overline{A})^2\Big)$
10	Support (s)	$P(A,B)$
11	Confidence (c)	$\max(P(B\|A),P(A\|B))$
12	Laplace (L)	$\max\Big(\frac{NP(A,B)+1}{NP(A)+2},\frac{NP(A,B)+1}{NP(B)+2}\Big)$
13	Conviction (V)	$\max\Big(\frac{P(A)P(\overline{B})}{P(A\overline{B})},\frac{P(B)P(\overline{A})}{P(B\overline{A})}\Big)$
14	Interest (I)	$\frac{P(A,B)}{P(A)P(B)}$
15	cosine (IS)	$\frac{P(A,B)}{\sqrt{P(A)P(B)}}$
16	Piatetsky-Shapiro's (PS)	$P(A,B)-P(A)P(B)$
17	Certainty factor (F)	$\max\Big(\frac{P(B\|A)-P(B)}{1-P(B)},\frac{P(A\|B)-P(A)}{1-P(A)}\Big)$
18	Added Value (AV)	$\max(P(B\|A)-P(B),P(A\|B)-P(A))$
19	Collective strength (S)	$\frac{P(A,B)+P(\overline{AB})}{P(A)P(B)+P(\overline{A})P(\overline{B})}\times\frac{1-P(A)P(B)-P(\overline{A})P(\overline{B})}{1-P(A,B)-P(\overline{AB})}$
20	Jaccard (ζ)	$\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$
21	Klosgen (K)	$\sqrt{P(A,B)}\max(P(B\|A)-P(B),P(A\|B)-P(A))$

49

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

50

FP-growth Algorithm

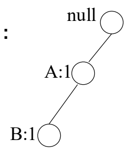
- Use a compressed representation of the database using an FP-tree
- Once an FP-tree has been constructed, it uses a recursive divide-and-conquer approach to mine the frequent itemsets

51

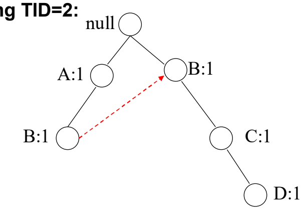
FP-tree construction

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

After reading TID=1:



After reading TID=2:

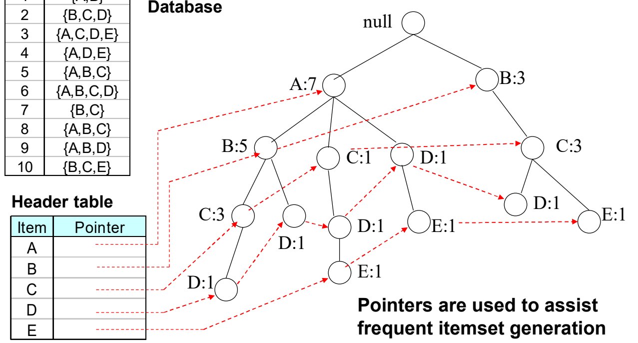


52

FP-Tree Construction

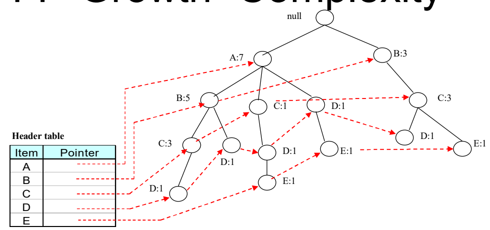
TID	Items
1	{A,B}
2	{B,C,D}
3	{A,C,D,E}
4	{A,D,E}
5	{A,B,C}
6	{A,B,C,D}
7	{B,C}
8	{A,B,C}
9	{A,B,D}
10	{B,C,E}

Transaction Database



Pointers are used to assist frequent itemset generation

FP-Growth Complexity



- Complexity of the PF-Growth algorithm depends very much on the compactness of the data. If the data is (not) compact the complexity gets low (high).

Vertical Data Layout

- For each item, store a list of transaction ids (tids)

Horizontal Data Layout

TID	Items
1	A,B,E
2	B,C,D
3	C,E
4	A,C,D
5	A,B,C,D
6	A,E
7	A,B
8	A,B,C
9	A,C,D
10	B

Vertical Data Layout

A	B	C	D	E
1	1	2	2	1
4	2	3	4	3
5	5	4	5	6
6	7	8	9	
7	8	9		
8	10			
9				

TID-list

Vertical Data Layout

- Determine support of any k-itemset by intersecting tid-lists of two of its (k-1) subsets.

A
1
4
5
6
7
8
9

$\wedge$

B
1
2
5
7
8
10

$\rightarrow$

AB
1
5
7
8

- 3 traversal approaches:
  - top-down, bottom-up and hybrid
- Advantage: very fast support counting
- Disadvantage: intermediate tid-lists may become too large for memory

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

57

Continuous and Categorical Attributes

How can we handle continuous and categorical attributes in the context of association rules?

Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
…	…	…	…	…	…	…

Example of Association Rule:

{Number of Pages $\in$[5,10) $\wedge$ (Browser=Mozilla)} $\rightarrow$ {Buy = No}

58

Handling Categorical Attributes

- Transform categorical attribute into asymmetric binary variables

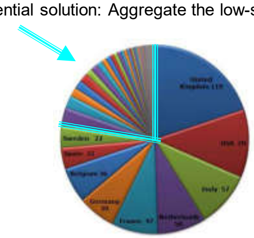
Session Id	Country	Session Length (sec)	Number of Web Pages viewed	Gender	Browser Type	Buy
1	USA	982	8	Male	IE	No
2	China	811	10	Female	Netscape	No
3	USA	2125	45	Female	Mozilla	Yes
4	Germany	596	4	Male	IE	Yes
5	Australia	123	9	Male	Mozilla	No
…	…	…	…	…	…	…

- Introduce a new "item" for each distinct attribute-value pair
  - Example: replace Browser Type attribute with 3 asymmetric binary attributes:
    - Browser Type = IE
    - Browser Type = Mozilla
    - Browser Type = Mozilla

59

Handling Categorical Attributes

- Potential Issues
  - What if attribute has many possible values
    - Example: attribute country has more than 200 possible values
    - Many of the attribute values may have very low support
      - Potential solution: Aggregate the low-support attribute values



60

Handling Categorical Attributes

- Potential Issues
  - What if distribution of attribute values is highly skewed
    - Example: 95% of the visitors have Buy = No
    - Most of the items will be associated with (Buy=No) item
      - Potential solution: drop the highly frequent items

61

Handling Continuous Attributes

- Different kinds of rules:
  - Age∈[21,35) ∧ Salary∈[70k,120k) → Buy
  - Salary∈[70k,120k) ∧ Buy → Age: μ=28, σ=4
- Different methods:
  - Discretization-based
  - Statistics-based

62

Discretization-based Methods

Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...	...	...	...	...	...	...

Male	Female	... ...	Age < 13	Age $\in [13,21)$	Age $\in [21,30)$	... ...	Privacy = Yes	Privacy = No
0	1	...	0	0	1	...	1	0
1	0	...	0	0	0	...	0	1
1	0	...	0	0	1	...	1	0
0	1	...	0	0	0	...	1	0
0	1	...	0	0	0	...	1	0
1	0	...	0	0	1	...	1	0
1	0	...	0	0	0	...	0	1
1	0	...	0	0	0	...	0	1
0	1	...	0	0	1	...	0	1
...	...	...	...	...	...	...	...	...

63

Handling Continuous Attributes

- Use discretization
- Unsupervised:
  - Equal-width binning
  - Equal-depth binning
  - Clustering
- Supervised:

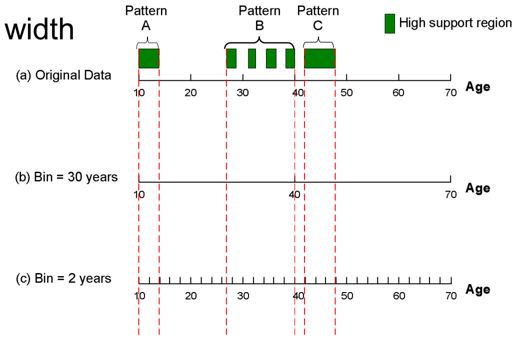
Attribute values, v

Class	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8	v_9
Anomalous	0	0	20	10	20	0	0	0	0
Normal	150	100	0	0	0	100	100	150	100
	$\underbrace{\qquad}_{bin_1}$		$\underbrace{\qquad}_{bin_2}$			$\underbrace{\qquad}_{bin_3}$			

64

Discretization Issues

- Interval width



Pattern A: Age $\in [10, 15) \longrightarrow$ Chat Online = Never
Pattern B: Age $\in [26, 41) \longrightarrow$ Chat Online = Never
Pattern C: Age $\in [42, 48) \longrightarrow$ Online Banking = Yes

65

Discretization Issues

- Interval too wide (e.g., Bin size= 30)
  - May merge several disparate patterns
    - Patterns A and B are merged together
  - May lose some of the interesting patterns
    - Pattern C may not have enough confidence
- Interval too narrow (e.g., Bin size = 2)
  - Pattern A is broken up into two smaller patterns
    - Can recover the pattern by merging adjacent subpatterns
  - Pattern B is broken up into smaller patterns
    - Cannot recover the pattern by merging adjacent subpatterns
- Potential solution: use all possible intervals
  - Start with narrow intervals
  - Consider all possible mergings of adjacent intervals

66

Statistics-based Methods

- Example:
  {Income > 100K, Online Banking=Yes} → Age: μ=34
- Rule consequent consists of a continuous variable, characterized by their statistics
  - mean, median, standard deviation, etc.
- Approach:
  - Withhold the target attribute from the rest of the data
  - Extract frequent itemsets from the rest of the attributes
    - Binarized the continuous attributes (except for the target attribute)
  - For each frequent itemset, compute the corresponding descriptive statistics of the target attribute
    - Frequent itemset becomes a rule by introducing the target variable as rule consequent
  - Apply statistical test to determine interestingness of the rule

67

Statistics-based Methods

Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...	...	...	...	...	...	...

Frequent Itemsets:

{Male, Income > 100K}
{Income < 40K, No hours ∈[10,15)}
{Income > 100K, Online Banking = Yes}
....

Association Rules:

{Male, Income > 100K} → Age: μ = 30
{Income < 40K, No hours ∈[10,15)} → Age: μ = 24
{Income > 100K,Online Banking = Yes} → Age: μ = 34
....

68

Statistics-based Methods

- How to determine whether an association rule is interesting?
  - Compare the statistics for segment of population covered by the rule vs segment of population not covered by the rule:

    A ⇒ B: μ versus $\bar{A}$ ⇒ B: μ'

  - Statistical hypothesis testing:

    $$Z = \frac{\mu' - \mu - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

    - Null hypothesis: H0: μ' = μ + Δ
    - Alternative hypothesis: H1: μ' > μ + Δ
    - Z has zero mean and variance 1 under null hypothesis
    - Note that s_1 (s_2) is standard deviation for B among the transaction that support A ($\bar{A}$).

69

Statistics-based Methods

- Example:

  r: Browser=Mozilla ∧ Buy=Yes → Age: μ=23
  - Rule is interesting if difference between μ and μ' is greater than 5 years (i.e., Δ = 5)
  - For r, suppose n1 = 50, s1 = 3.5
  - For r' (complement): n2 = 250, s2 = 6.5, and average age is 30.

$$Z = \frac{\mu' - \mu - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{30 - 23 - 5}{\sqrt{\frac{3.5^2}{50} + \frac{6.5^2}{250}}} = 3.11$$

  - For 1-sided test at 95% confidence level, critical Z-value for rejecting null hypothesis is 1.64.
  - Since Z is greater than 1.64, r is an interesting rule

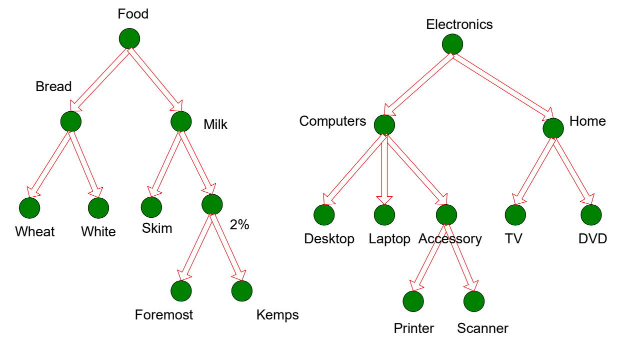
70

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

71

Multi-level Association Rules



72

Multi-level Association Rules

- Why should we incorporate concept hierarchy?
  – Rules at lower levels may not have enough support to appear in any frequent itemsets

  – Rules at lower levels of the hierarchy are overly specific
    - e.g., skim milk → white bread, milk → wheat bread,
      skim milk → wheat bread, etc.
      are indicative of association between milk and bread

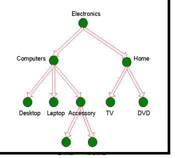
73

Multi-level Association Rules

- How do support and confidence vary as we traverse the concept hierarchy?
  – If X is the parent item for both X1 and X2, then $\sigma(X) \geq \sigma(X1) + \sigma(X2)$

  – If $\sigma(X1 \cup Y1) \geq minsup$,
    and X is parent of X1, Y is parent of Y1
    then $\sigma(X \cup Y1) \geq minsup$, $\sigma(X1 \cup Y) \geq minsup$
    $\sigma(X \cup Y) \geq minsup$

  – If $conf(X1 \Rightarrow Y1) \geq minconf$,
    then $conf(X1 \Rightarrow Y) \geq minconf$



74

Multi-level Association Rules

- Approach 1:
  – Extend current association rule formulation by augmenting each transaction with higher level items

  Original Transaction: {skim milk, wheat bread}
  Augmented Transaction:
  {skim milk, wheat bread, milk, bread, food}

- Issues:
  – Items that reside at higher levels have much higher support counts
    - if support threshold is low, too many frequent patterns involving items from the higher levels
  – Increased dimensionality of the data

75

Multi-level Association Rules

- Approach 2:
  – Generate frequent patterns at highest level first

  – Then, generate frequent patterns at the next highest level, and so on

- Issues:
  – I/O requirements will increase dramatically because we need to perform more passes over the data
  – May miss some potentially interesting cross-level association patterns

76

Association Analysis: Basic Concepts and Algorithms

- Association Rule Problem and Complexity
- Apriori Algorithm and Rule Generation
- Compact Representations
- Alternative Association Rule Problems
- Quality Measures for Association Rules
- Alternative Frequent Itemset Algorithms: FP-Growth and Vertical Data Layout
- Handling Categorical and Numeric Data
- Multi-Level Association Rules

77

Summary

- Basic concepts: association rules, support-confident framework, closed and max-patterns
- Scalable frequent pattern mining methods
  - Apriori (Candidate generation & test)
  - Projection-based (FPgrowth, CLOSET+, ...)
  - Vertical format approach (ECLAT, CHARM, ...)
- Which patterns are interesting?
  - Pattern evaluation methods

78

Ref: Basic Concepts of Frequent Pattern Mining

- (Association Rules) R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93.
- (Max-pattern) R. J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98.
- (Closed-pattern) N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. ICDT'99.
- (Sequential pattern) R. Agrawal and R. Srikant. Mining sequential patterns. ICDE'95

79

Ref: Apriori and Its Improvements

- R. Agrawal and R. Srikant. Fast algorithms for mining association rules. VLDB'94.
- H. Mannila, H. Toivonen, and A. I. Verkamo. Efficient algorithms for discovering association rules. KDD'94.
- A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. VLDB'95.
- J. S. Park, M. S. Chen, and P. S. Yu. An effective hash-based algorithm for mining association rules. SIGMOD'95.
- H. Toivonen. Sampling large databases for association rules. VLDB'96.
- S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket analysis. SIGMOD'97.
- S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database systems: Alternatives and implications. SIGMOD'98.

80

Ref: Depth-First, Projection-Based FP Mining

- R. Agarwal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent itemsets. J. Parallel and Distributed Computing:02.
- J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. SIGMOD' 00.
- J. Liu, Y. Pan, K. Wang, and J. Han. Mining Frequent Item Sets by Opportunistic Projection. KDD'02.
- J. Han, J. Wang, Y. Lu, and P. Tzvetkov. Mining Top-K Frequent Closed Patterns without Minimum Support. ICDM'02.
- J. Wang, J. Han, and J. Pei. CLOSET+: Searching for the Best Strategies for Mining Frequent Closed Itemsets. KDD'03.
- G. Liu, H. Lu, W. Lou, J. X. Yu. On Computing, Storing and Querying Frequent Patterns. KDD'03.
- G. Grahne and J. Zhu, Efficiently Using Prefix-Trees in Mining Frequent Itemsets, Proc. ICDM'03 Int. Workshop on Frequent Itemset Mining Implementations (FIMI'03), Melbourne, FL, Nov. 2003

81

Ref: Vertical Format and Row Enumeration Methods

- M. J. Zaki, S. Parthasarathy, M. Ogihara, and W. Li. Parallel algorithm for discovery of association rules. DAMI:97.
- Zaki and Hsiao. CHARM: An Efficient Algorithm for Closed Itemset Mining, SDM'02.
- C. Bucila, J. Gehrke, D. Kifer, and W. White. DualMiner: A Dual-Pruning Algorithm for Itemsets with Constraints. KDD'02.
- F. Pan, G. Cong, A. K. H. Tung, J. Yang, and M. Zaki , CARPENTER: Finding Closed Patterns in Long Biological Datasets. KDD'03.
- H. Liu, J. Han, D. Xin, and Z. Shao, Mining Interesting Patterns from Very High Dimensional Data: A Top-Down Row Enumeration Approach, SDM'06.

82

Ref: Mining Correlations and Interesting Rules

- M. Klemettinen, H. Mannila, P. Ronkainen, H. Toivonen, and A. I. Verkamo. Finding interesting rules from large sets of discovered association rules. CIKM'94.
- S. Brin, R. Motwani, and C. Silverstein. Beyond market basket: Generalizing association rules to correlations. SIGMOD'97.
- C. Silverstein, S. Brin, R. Motwani, and J. Ullman. Scalable techniques for mining causal structures. VLDB'98.
- P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the Right Interestingness Measure for Association Patterns. KDD'02.
- E. Omiecinski. Alternative Interest Measures for Mining Associations. TKDE'03.
- T. Wu, Y. Chen and J. Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", PKDD'07

83

Ref: Freq. Pattern Mining Applications

- Y. Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen. Efficient Discovery of Functional and Approximate Dependencies Using Partitions. ICDE'98.
- H. V. Jagadish, J. Madar, and R. Ng. Semantic Compression and Pattern Extraction with Fascicles. VLDB'99.
- T. Dasu, T. Johnson, S. Muthukrishnan, and V. Shkapenyuk. Mining Database Structure; or How to Build a Data Quality Browser. SIGMOD'02.
- K. Wang, S. Zhou, J. Han. Profit Mining: From Patterns to Actions. EDBT'02.

84