Unit 7

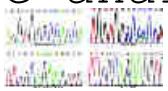
Association

Market Basket Analysis: A Motivating Example

What Is Frequent Pattern Analysis?

- **Frequent pattern:** a pattern (a set of items, subsequences, substructures, etc.) that occurs frequently in a data set
- First proposed by Agrawal, Imielinski, and Swami [AIS93] in the context of frequent itemsets and association rule mining
- **Motivation:** Finding inherent regularities in data
  - What products were often purchased together?— Beer and diapers?!
  - What are the subsequent purchases after buying a PC?
  - What kinds of DNA are sensitive to this new drug?
  - Can we automatically classify web documents?
- **Applications**


Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.



Why Is Freq. Pattern Mining Important?

- Freq. pattern: An intrinsic and important property of datasets
- Foundation for many essential data mining tasks
  - Association, correlation, and causality analysis
  - Sequential, structural (e.g., sub-graph) patterns
  - Pattern analysis in spatiotemporal, multimedia, time-series, and stream data
  - Classification: discriminative, frequent pattern analysis
  - Cluster analysis: frequent pattern-based clustering
  - Data warehousing: iceberg cube and cube-gradient
  - Semantic data compression

Basic Concepts: Frequent Patterns and Association rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk

- Let minsup=50%
- Freq. 1-itemsets:
  - Beer:3(60%);
    Nuts:3(60%);
    Diaper:4(80%);
    Eggs:3(60%)
- Freq. 2-itemsets:
  - {Beer, Diaper}:3(60%)

- itemset: A set of one or more items
- k-itemset $X = \{x_1, \ldots, x_k\}$
- (absolute) support, or, support count of X: Frequency or occurrence of an itemset X
- (relative) support, s, is the fraction of transactions that contains X (i.e., the probability that a transaction contains X)
- An itemset X is frequent if X's support is no less than a minsup threshold

Association rule

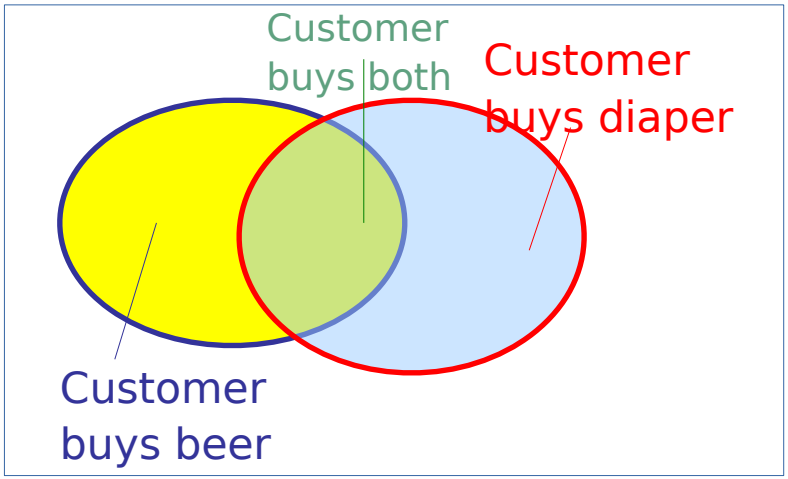
- Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of m distinct attributes, also called **literals**.

  Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$. An association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subset I$, are sets of items called itemsets, and $X \cap Y = \varphi$. Here, X is called antecedent, and Y consequent.

- Two important Rule strength measures for association rules, support (s) and confidence (α), can be defined.

Strength of Association Rules

Tid	Items bought
10	Beer, Nuts, Diaper
20	Beer, Coffee, Diaper
30	Beer, Diaper, Eggs
40	Nuts, Eggs, Milk
50	Nuts, Coffee, Diaper, Eggs, Milk



- Find all the rules X → Y with minimum support and confidence
  - support, s, probability that a transaction contains X ∪ Y

$$support(A \Rightarrow B) = P(A \cup B)$$

  - confidence, c, conditional probability that a transaction having X also contains Y

$$confidence(A \Rightarrow B) = P(B|A).$$

Let minsup = 50%, minconf = 50%

Freq. Pat.: Beer:3, Nuts:3, Diaper:4, Eggs:3, {Beer, Diaper}:3

Association rules: (many more!)

Beer → Diaper

Association mining rules in Large Database

- Association rule mining finds interesting association or correlation relationships among a large set of data items.
- With massive amounts of data continuously being collected and stored , many industries are becoming interested in mining association huge amounts of business transaction records which helps in many business decision making processes, such as catalog design, cross-marketing, and loss-leader analysis.

Association mining rules in Large Database

- A typical example of association rule mining is **market basket analysis**. This process analyzes customer buying habits by finding associations between the different items that customers place in their “shopping baskets”.
- The discovery of such associations can help retailers develop marketing strategies by gaining insight into which items are frequently purchased together by customers.
- For instance, if customers are buying milk, how likely are they to also buy bread(and what kind of bread)on the same trip to the supermarket? Such information can lead to increased sales by helping retailers do selective marketing and plan their shelf space. For example, placing milk And bread within close proximity may further encourage the sale of these items together within single visits to the store.

Closed Patterns and Max-Patterns

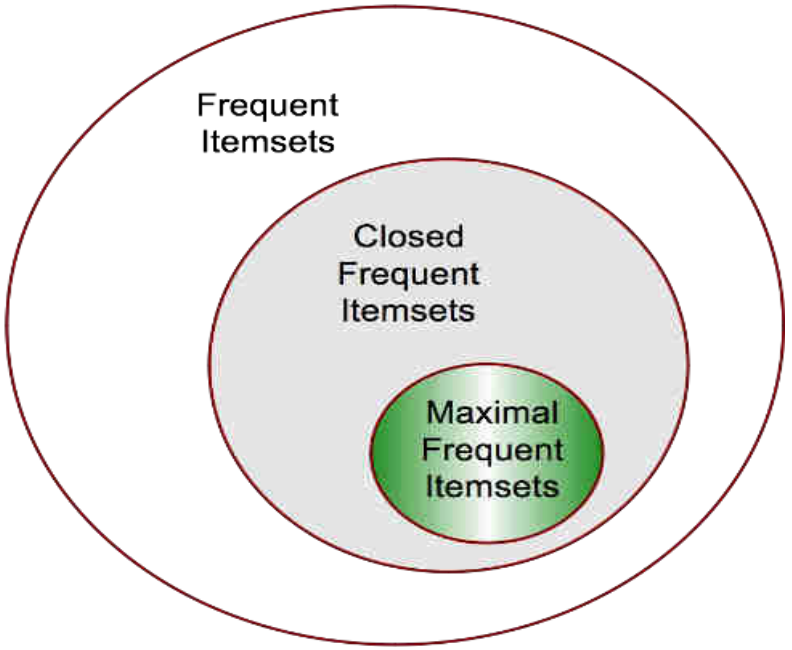
- An itemset X is closed if X is frequent and there exists no super-pattern Y ⊃ X, with the same support as X (proposed by Pasquier, et al. @ ICDT'99)
- An itemset X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ⊃ X (proposed by Bayardo @ SIGMOD'98)

Closed Vs Maximal Itemsets

→ **Closed Frequent Itemsets** are Lossless:
The support for any frequent itemset can be deduced from the closed frequent itemset.

→ **Max-pattern** is a lossy compression:
We only know all subsets of Max-pattern are frequent but not the real support.

Thus in many applications, mining close-patterns is more desirable than mining max-patterns.



Support and Confidence

- **Support of X in D** is count(X)/|D|

  For an association rule X ⇒ Y, we can calculate

  → support (X ⇒ Y) = support (XY)

  → confidence (X ⇒ Y) = support (XY)/support(X)

- **Support Count**

The support count of an itemset X, denoted by X.count, in a dataset T is the number of transactions in T that contain X. Assume T has n transactions. Then,

$$support = \frac{(X \cup Y).count}{n}$$

$$confidence = \frac{(X \cup Y).count}{X.count}$$

Example of support Measure

TID	Items	Support = Occurence / Total Support
1	ABC	
2	ABD	Total Support = 5
3	BC	Support {AB} = 2 / 5 = 40%
4	AC	Support {BC} = 3 / 5 = 60%
5	BCD	Support {ABC} = 1 / 5 = 20%

Example of Confidence Measure

TID	Items	Given $X \Rightarrow Y$ Confidence = Occurrence {Y} / Occurrence {X}
1	ABC	
2	ABD	
3	BC	Confidence $\{A \Rightarrow B\} = 2/3 = 66\%$
4	AC	Confidence $\{B \Rightarrow C\} = 3/4 = 75\%$
5	BCD	Confidence $\{AB \Rightarrow C\} = 1/2 = 50\%$

Example

Database with transactions (customer_# : item_a1, item_a2, …)

1: 3, 5, 8.
2: 2, 6, 8.
3: 1, 4, 7, 10.
4: 3, 8, 10.
5: 2, 5, 8.
6: 1, 5, 6.
7: 4, 5, 6, 8.
8: 2, 3, 4.
9: 1, 5, 7, 8.
10: 3, 8, 9, 10.

Conf ({5} => {8}) ?

supp({5}) = 5 , supp({8}) = 7 , supp({5,8}) = 4,

then **conf({5} => {8}) = 4/5 = 0.8 or 80%**

Example

Database with transactions (customer_# : item_a1, item_a2, …)

1: 3, 5, 8.
2: 2, 6, 8.
3: 1, 4, 7, 10.
4: 3, 8, 10.
5: 2, 5, 8.
6: 1, 5, 6.
7: 4, 5, 6, 8.
8: 2, 3, 4.
9: 1, 5, 7, 8.
10: 3, 8, 9, 10.

Conf ({5} => {8}) ? 80% Done. Conf ({8} => {5}) ?
supp({5}) = 5 , supp({8}) = 7 , supp({5,8}) = 4,
then **conf({8} => {5}) = 4/7 = 0.57 or 57%**

Example

Conf ({5} => {8}) ? 80% Done.

Conf ({8} => {5}) ? 57% Done.

Rule ({5} => {8}) more meaningful then

Rule ({8} => {5})

Practice...

Which is meaningful??

({9}→{3})

or

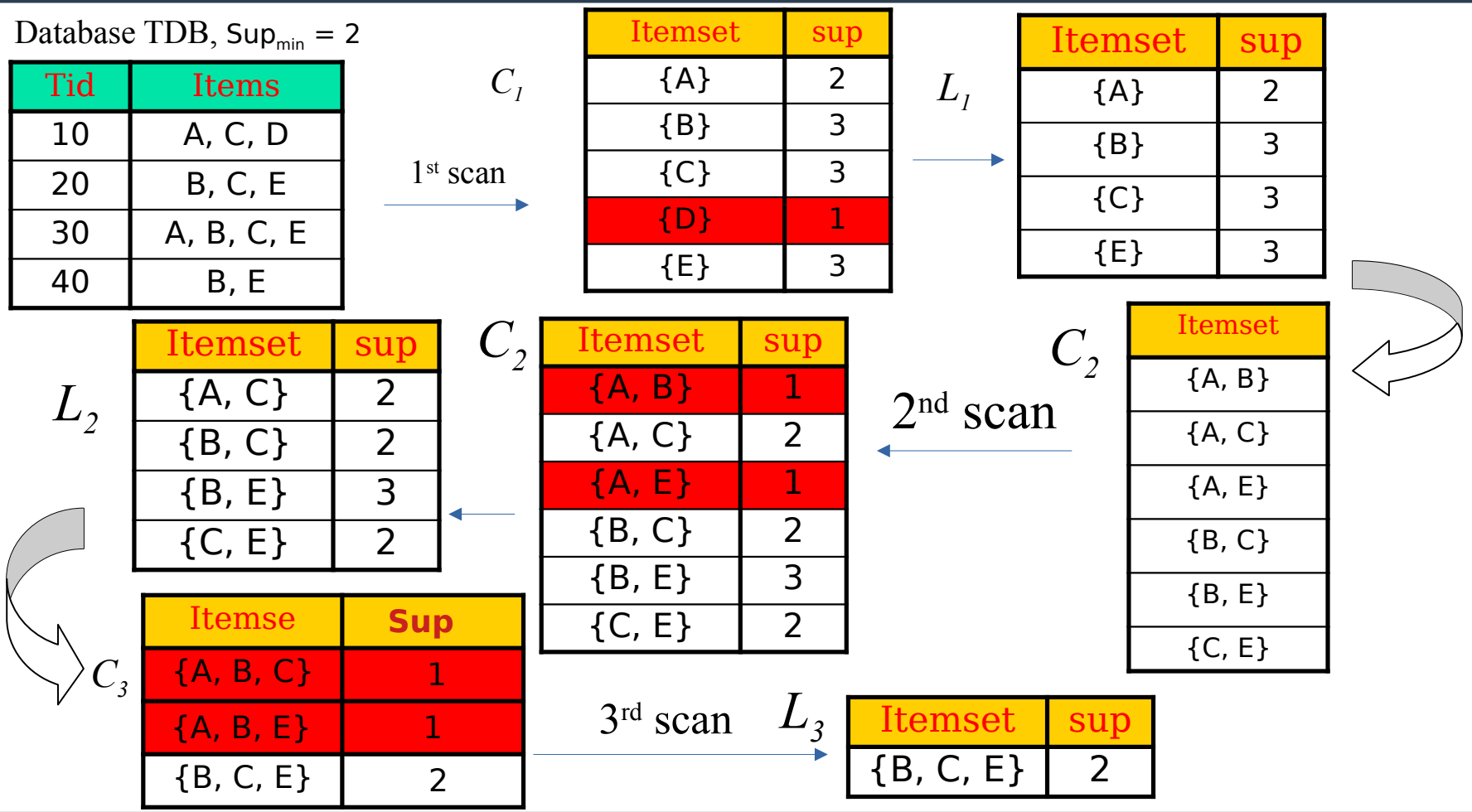
({3}→{9})

Apriori Algorithm

- **Apriori is algorithm** proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties.
- The Apriori algorithm is the most well known association rule algorithm and is used in most commercial products.
- Two steps:
  a) Find all itemsets that have minimum support (frequent itemsets, also called large itemsets).
  b) Use frequent itemsets to generate rules.

Apriori Algorithm → Example

$Sup_{min} = 2$



Apriori Algorithm → Pseudo Code

C_k: Candidate itemset of size k
L_k : frequent itemset of size k

L_1 = {frequent items};
for (k = 1; L_k !=∅; k++) do begin
    C_{k+1} = candidates generated from L_k;
    for each transaction t in database do
        increment the count of all candidates in C_{k+1} that are contained in t
        L_{k+1} = candidates in C_{k+1} with min_support
    end
return L_k;

Apriori: A Candidate Generation & Test Approach

- Outline of Apriori (level-wise, candidate generation and testing)
- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Repeat
    - Generate (k+1) candidate itemsets from length k frequent itemsets
    - Test the candidates against DB to find frequent (k+1) itemsets
    - Set k:=k+1
  - Terminate when no frequent or candidate set can be generated
  - Return all the frequent itemsets derived.

Apriori: Example

TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

- Consider a database, D , consisting of 9 transactions.
- Suppose min. support count required is 2 (i.e. min_sup = 2/9 = 22 %)
- Let minimum confidence required is 70%.
- We have to first find out the frequent itemset using Apriori algorithm.
- Then, Association rules will be generated using min. support & min. confidence.

Apriori: Example

Step 1: Generating 1-itemset Frequent Pattern

Scan D for count of each candidate

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

C_1

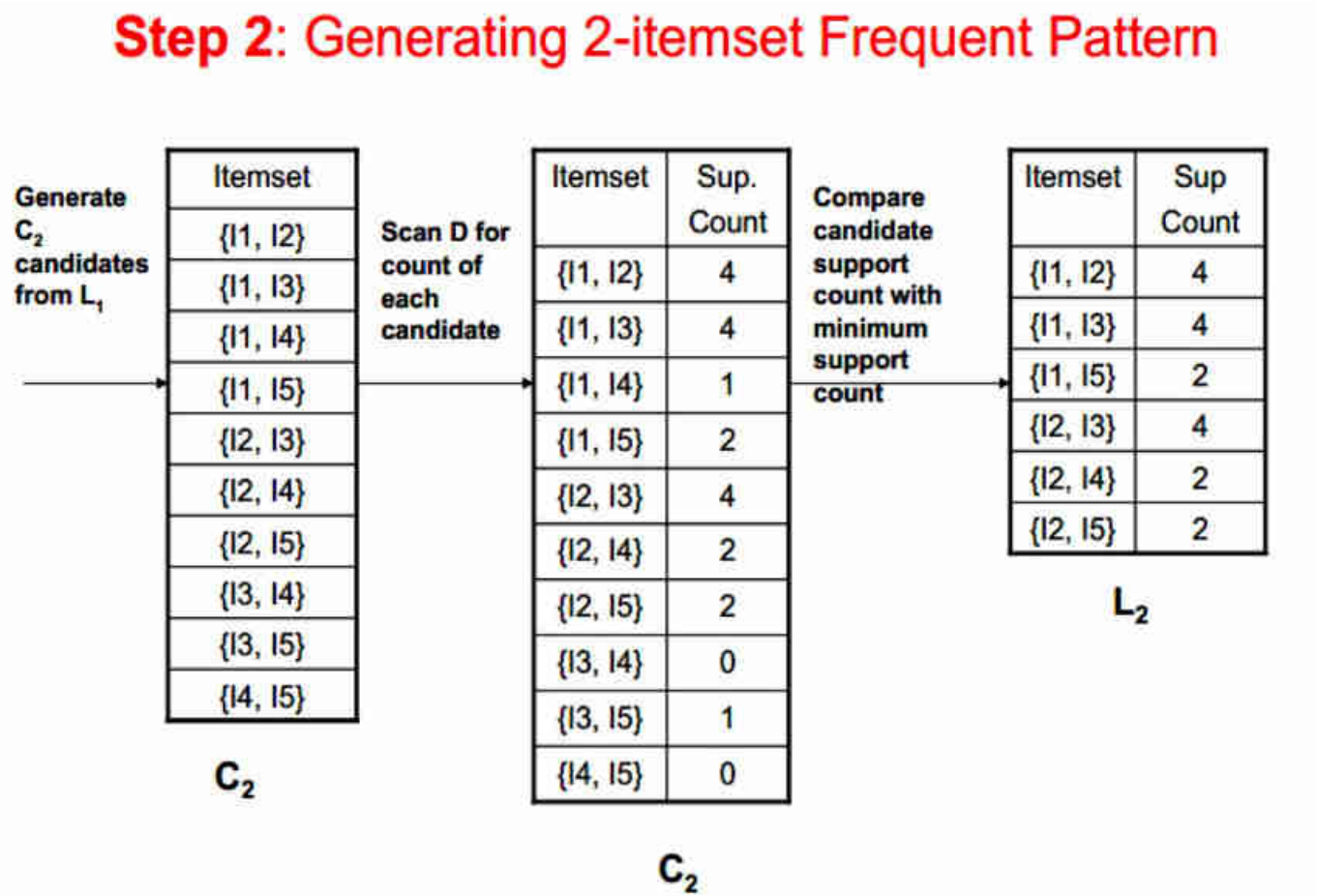
Compare candidate support count with minimum support count

Itemset	Sup.Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

L_1

- The set of frequent 1-itemsets, L_1, consists of the candidate 1-itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

Apriori: Example



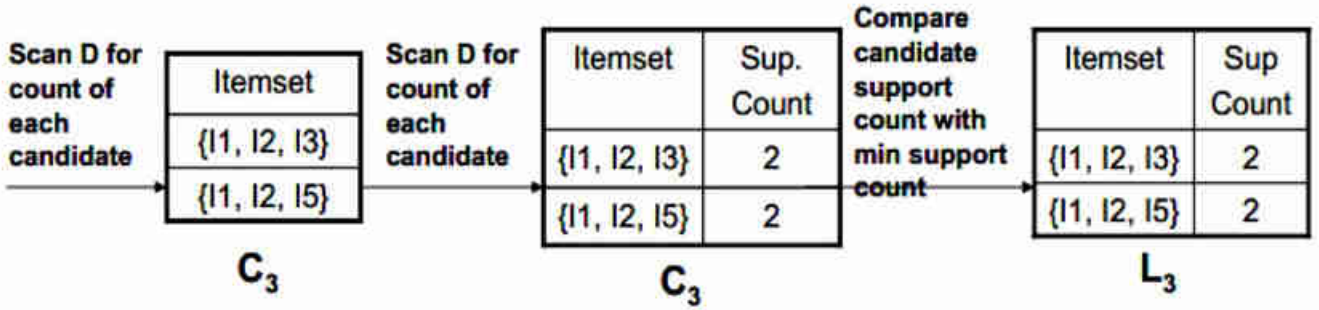
Apriori: Example

Step 2: Generating 2-itemset Frequent Pattern

- To discover the set of frequent 2-itemsets, L_2 , the algorithm uses L_1 *Join* L_1 to generate a candidate set of 2-itemsets, C_2.
- Next, the transactions in D are scanned and the support count for each candidate itemset in C_2 is accumulated (as shown in the middle table).
- The set of frequent 2-itemsets, L_2 , is then determined, consisting of those candidate 2-itemsets in C_2 having minimum support.

Apriori: Example

Step 3: Generating 3-itemset Frequent Pattern



- The generation of the set of candidate 3-itemsets, C_3 , involves use of the Apriori Property.
- In order to find C_3, we compute L_2 *Join* L_2.
- C_3 = L2 *Join* L2 = {{I1, I2, I3}, {I1, I2, I5}, {I1, I3, I5}, {I2, I3, I4}, {I2, I3, I5}, {I2, I4, I5}}.
- Now, Join step is complete and Prune step will be used to reduce the size of C_3. Prune step helps to avoid heavy computation due to large C_k.

Apriori: Example

Step 3: Generating 3-itemset Frequent Pattern

- Based on the Apriori property that all subsets of a frequent itemset must also be frequent, we can determine that four latter candidates cannot possibly be frequent. How ?
- For example , lets take {I1, I2, I3}. The 2-item subsets of it are {I1, I2}, {I1, I3} & {I2, I3}. Since all 2-item subsets of {I1, I2, I3} are members of L_2, We will keep {I1, I2, I3} in C_3.
- Lets take another example of {I2, I3, I5} which shows how the pruning is performed. The 2-item subsets are {I2, I3}, {I2, I5} & {I3,I5}.
- BUT, {I3, I5} is not a member of L_2 and hence it is not frequent violating Apriori Property. Thus We will have to remove {I2, I3, I5} from C_3.
- Therefore, C_3 = {{I1, I2, I3}, {I1, I2, I5}} after checking for all members of result of Join operation for Pruning.
- Now, the transactions in D are scanned in order to determine L_3, consisting of those candidates 3-itemsets in C_3 having minimum support.

Apriori: Example

Step 4: Generating 4-itemset Frequent Pattern

- The algorithm uses L_3 *Join* L_3 to generate a candidate set of 4-itemsets, C_4. Although the join results in {{I1, I2, I3, I5}}, this itemset is pruned since its subset {{I2, I3, I5}} is not frequent.
- Thus, $C_4 = \varphi$, and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm.
- What's Next ?

  These frequent itemsets will be used to generate strong association rules (where strong association rules satisfy both minimum support & minimum confidence).

Apriori: Example

- Lets take *l* = {I1,I2,I5}.
- Its all nonempty subsets are {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}.

Step 5: Generating Association Rules from Frequent Itemsets

- Let minimum confidence threshold is , say 70%.
- The resulting association rules are shown below, each listed with its confidence.
  - R1: I1 ^ I2 → I5
    - Confidence = sc{I1,I2,I5}/sc{I1,I2} = 2/4 = 50%
    - R1 is Rejected.
  - R2: I1 ^ I5 → I2
    - Confidence = sc{I1,I2,I5}/sc{I1,I5} = 2/2 = 100%
    - R2 is Selected.
  - R3: I2 ^ I5 → I1
    - Confidence = sc{I1,I2,I5}/sc{I2,I5} = 2/2 = 100%
    - R3 is Selected.

Apriori: Example

- Lets take *l* = {I1,I2,I5}.
- Its all nonempty subsets are {I1,I2}, {I1,I5}, {I2,I5}, {I1}, {I2}, {I5}.

Step 5: Generating Association Rules from Frequent Itemsets

- R4: I1 → I2 ^ I5
  - Confidence = sc{I1,I2,I5}/sc{I1} = 2/6 = 33%
  - R4 is Rejected.
- R5: I2 → I1 ^ I5
  - Confidence = sc{I1,I2,I5}/{I2} = 2/7 = 29%
  - R5 is Rejected.
- R6: I5 → I1 ^ I2
  - Confidence = sc{I1,I2,I5}/ {I5} = 2/2 = 100%
  - R6 is Selected.

  In this way, We have found three strong association rules.

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
  - Facilitate support counting of candidates

Application of Association Rule Mining

- Market Basket Analysis
- Medical diagnosis
- Census data
- CRM of credit card business

Application of Association Rule Mining → Market Basket Analysis

Refer Slide 11

Application of Association Rule Mining → Medical Diagonsis

- Association rules can be used in medical analysis for assisting physicians to cure patients.
- The common problem of the induction of reliable analytic rules is hard as theoretically no induction process by itself can guarantee the accuracy of induced hypotheses.
- Basically diagnosis is not an easy process because of unreliable diagnosis tests and the presence of noise in training examples.
- This may result in hypotheses with insufficient prediction correctness which is too unreliable for critical medical applications

Application of Association Rule Mining → Census Data

- Censuses make a huge variety of general statistical information on society available to both researchers and the general community .
- The information related to population and economic census can be forecasted in planning public services(education, health, transport, funds) as well as in public business(for setup new factories, shopping malls or banks and even marketing particular products).
- The application of data mining techniques to census data and more generally to official data has great potential in supporting good community policy and in underpinning the effective functioning of a democratic society.

Application of Association Rule Mining → CRM of credit card business

- Customer Relationship Management (CRM), through which, banks expect to identify the preference of different customer groups, products and services adapted to their liking to enhance the cohesion between credit card customers and the bank, has become a topic of great interest .
- The collective application of association rule techniques reinforces the knowledge management process and allows marketing personnel to know their customers well to provide better quality services.

FPGrowth Approach

Assignment