Affinity - Association Rules

Example:

- Market Basket Analysis

Application of Market Basket Analysis:

- Improve the shop layout
- cross selling, cf. Amazon
- designing catalogues

How?

- Establish rules of this kind:

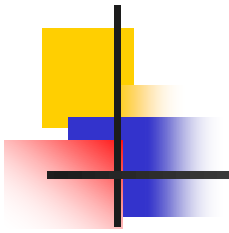
If the items from set A are bought, then the probability that the items from set C are bought is "x" %.

- Short: if A, then P(C) = x%.
- Based on frequencies in a dataset
- A = 'antecedent', C = 'consequent'

Association Rule Mining

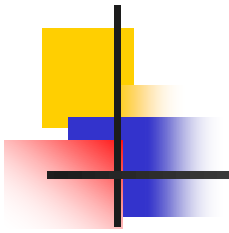
Instructor: Jesse Davis

Slides from: Chris Clifton,
Pedro Domingos, Jeff Ullman



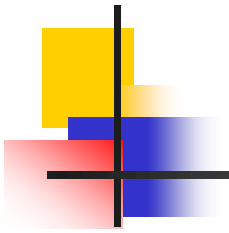
Outline

- Introduction and definitions
- Naïve algorithm
- Apriori
- PCY
- Limiting disk I/O
- FP Growth
- Multi-level association rules
- Incorporating constraints into mining
- Presenting results, other metrics



Outline

- Introduction and definitions
- Naïve algorithm
- Apriori
- PCY
- Limiting disk I/O
- FP Growth
- Multi-level association rules
- Incorporating constraints into mining
- Presenting results, other metrics



Association Rule Mining Task

Given: Set of transactions

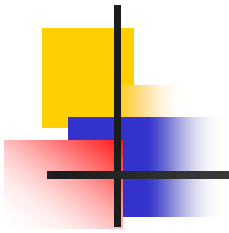
Find: IF-THEN rules that predict the occurrence of an item based on other items in the transaction

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

Association Rules

{Diaper} $\rightarrow$ {Beer},
{Milk, Bread} $\rightarrow$ {Eggs, Coke}
{Beer, Bread} $\rightarrow$ {Milk}

Implication means co-occurrence, not causality!



Why Association Rule Mining

- Motivation: Finding regularities in data
  - What products were often purchased together?
  - What kinds of DNA are sensitive to new drug?
- Foundation for many data mining tasks
  - Association
  - Correlation
- Algorithms do not require labeled data or for a user to specify a predefined target concept

Market Baskets

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

Transaction

Itemset

Item

- General many-many mapping (association) between items and baskets
- Connection among "items," not among "baskets"
- Focuses on common events, not rare events

Definition: Item Set

- **Itemset:** A collection of one or more items
  - Example: {Bread, Milk}
- **k-itemset:** An itemset that contains k items
  - 3-itemset: {Bread, Milk, Diaper}

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

Definition: Support and Frequent Itemsets

- Simplest question: find sets of items that appear "frequently" in the baskets
- *Support count* for itemset I = the number of baskets containing all items in I
- **Support:** Fraction of transactions that contain an itemset
- Given a *support threshold* s, sets of items that appear in at least s baskets are called *frequent itemsets*

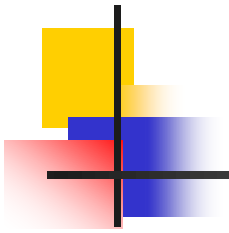
Example: Support

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

Itemset	Freq
{Br,M}	4
{Br,D}	3

Support({Br,M}) = 4/5 = 0.8

Support({Br,D}) = 3/5 = 0.6



Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets:

Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

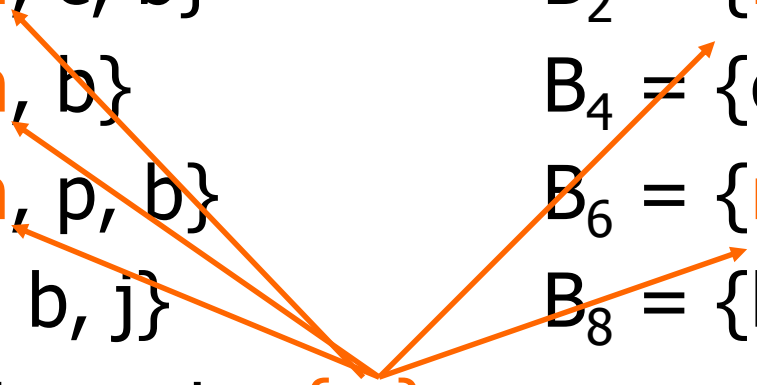
B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}



Example: Frequent Itemsets

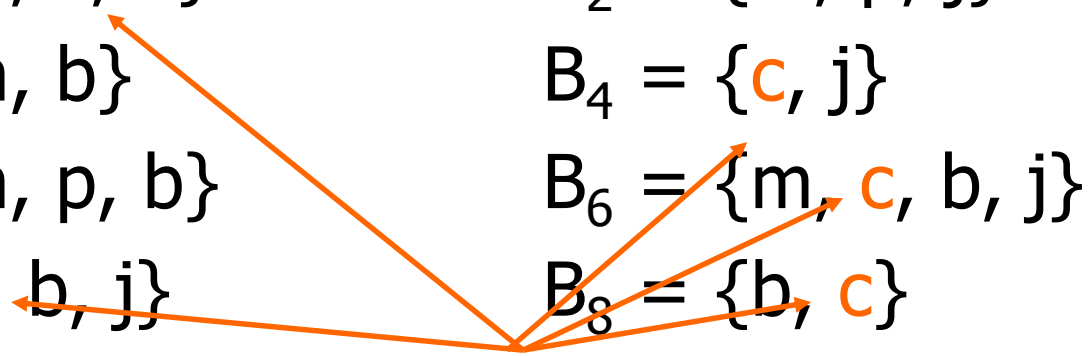
- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}



- Frequent itemsets: {m}, {c}

Example: Frequent Itemsets

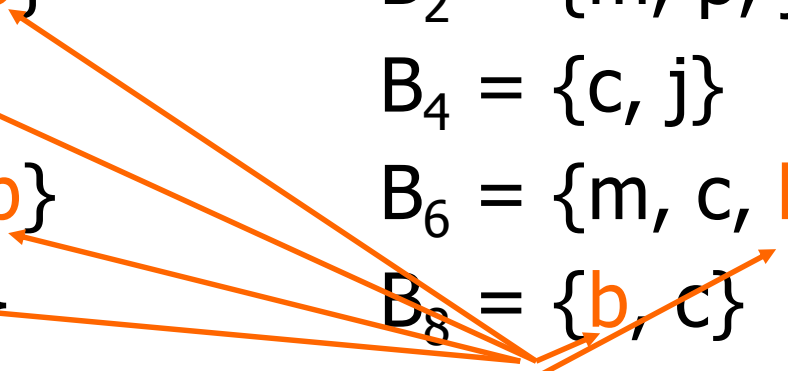
- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}



- Frequent itemsets: {m}, {c}, {b}

Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

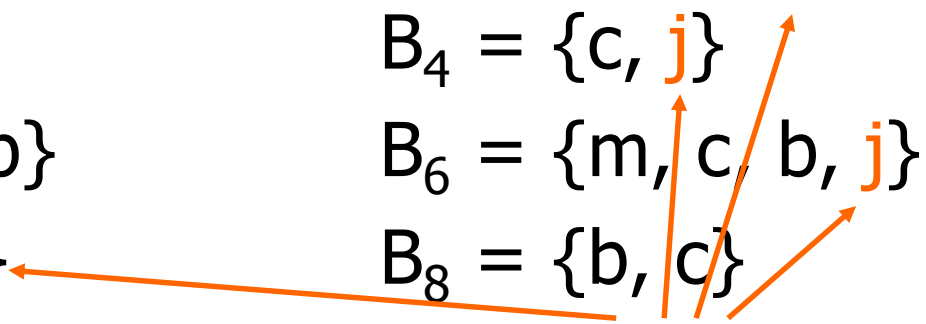
B_1 = {m, c, b} B_2 = {m, p, j}

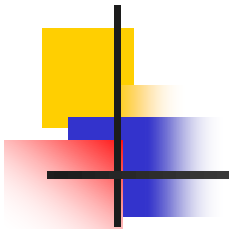
B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}





Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j},

Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

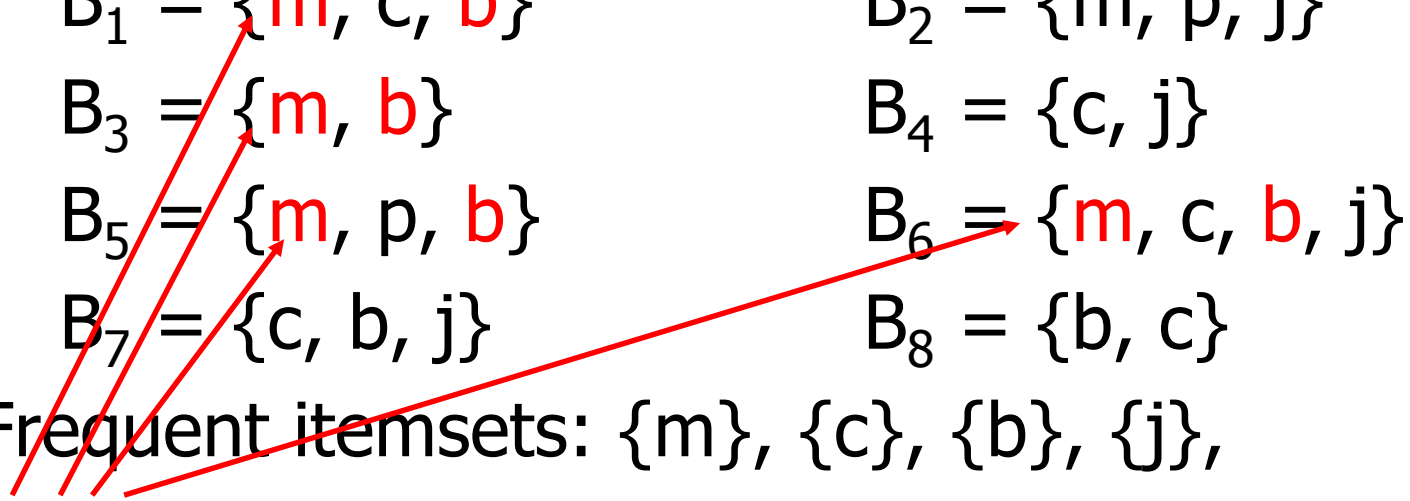
B_1 = {m, c, b} B_2 = {m, p, j}

B_3 = {m, b} B_4 = {c, j}

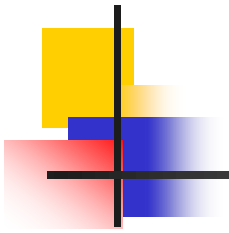
B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}, {m,b}



Example: Frequent Itemsets

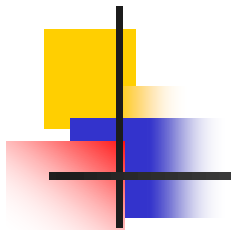


- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b}	B_2 = {m, p, j}
B_3 = {m, b}	B_4 = {c, j}
B_5 = {m, p, b}	B_6 = {m, c, b, j}
B_7 = {c, b, j}	B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}, {m,b}

Example: Frequent Itemsets



- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b} B_2 = {m, p, j}

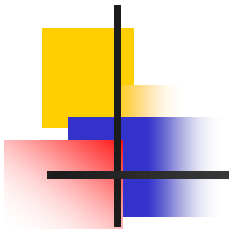
B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}, {m,b}, {b,c}

Example: Frequent Itemsets



- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

B_1 = {m, c, b}	B_2 = {m, p, j}
B_3 = {m, b}	B_4 = {c, j}
B_5 = {m, p, b}	B_6 = {m, c, b, j}
B_7 = {c, b, j}	B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}, {m,b}, {b,c}

Example: Frequent Itemsets

- Items={milk, coke, pepsi, beer, juice}.
- Support = 3 baskets.

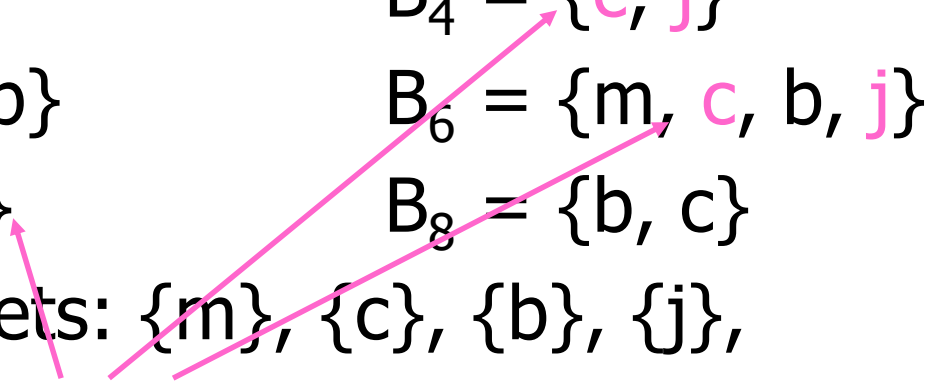
B_1 = {m, c, b} B_2 = {m, p, j}

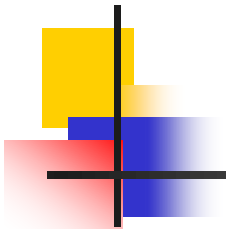
B_3 = {m, b} B_4 = {c, j}

B_5 = {m, p, b} B_6 = {m, c, b, j}

B_7 = {c, b, j} B_8 = {b, c}

- Frequent itemsets: {m}, {c}, {b}, {j}, {m,b}, {b,c}, {c,j}





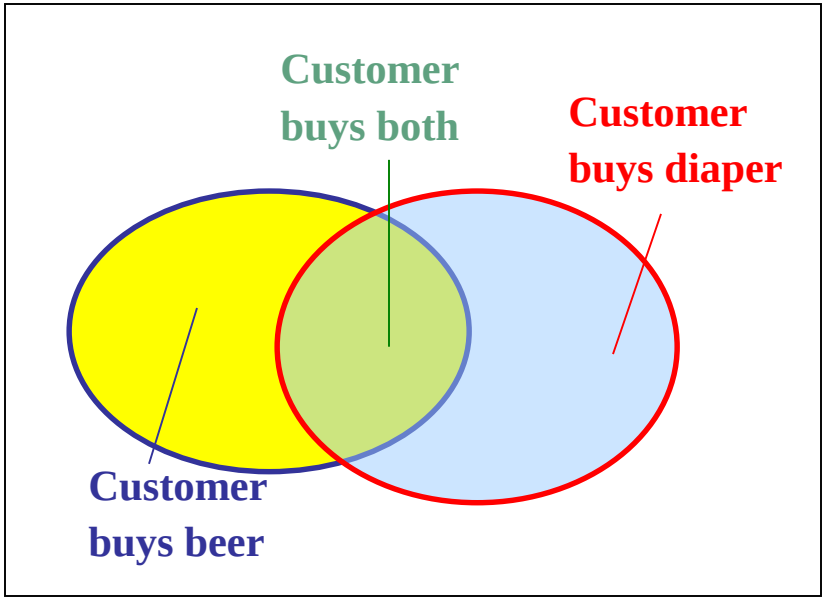
Definition: Association Rules

- If-then rules about the contents of baskets
- Given:
  - Set of *items*: $I = \{i_1, i_2, \ldots, i_m\}$
  - Set of *transactions*: $D = \{d_1, d_2, \ldots, d_n\}$
- An *association rule*: $A \Rightarrow B$, where
  - $A \subset I$
  - $B \subset I$
  - $A \cap B = \varnothing$
- $\{i_1, i_2, \ldots, i_k\} \rightarrow j$ means: "if a basket contains all of $i_1, \ldots, i_k$ then it is *likely* to contain j."

Definition: Confidence

- *Confidence* of this association rule is the conditional probability of j given $i_1,\ldots,i_k$
  - This gives a measure of how accurate the rule is
  - confidence(A $\Rightarrow$ B) = P(B|A) = sup({A,B}) /sup(A)

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

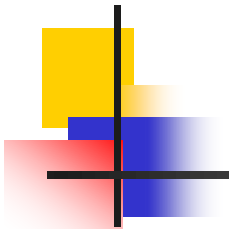


Example: Confidence

+ B_1 = {m, c, b}
− B_3 = {m, b}
− B_5 = {m, p, b}
B_7 = {c, b, j}

B_2 = {m, p, j}
B_4 = {c, j}
+ B_6 = {m, c, b, j}
B_8 = {b, c}

- An association rule: {m, b} → c
  - Confidence = 2/4 = 50%



Interestingness

- Given association rule: $I \rightarrow j$

  Interest = Confidence(j | I) - Support(j)

- Interest = 0: I has no influence on j
- Interest > 0: I may cause the presence of j
- Interest < 0: I discourages presence of j

Example: Interestingness

Items
Bread, Milk
Bread, Milk, Diaper, Beer, Eggs
Milk, Diaper, Beer, Coke
Bread, Milk, Diaper, Beer
Bread, Milk, Diaper, Coke

Itemset	Freq
{Br,M}	4
{Br,D}	3

Support({M}) = 5/5 = 1.0

Confidence(Br → M) = 4/4 = 1.0

Interest(Br → M) = Conf(Br → M) – Supp(M) = 0.0

Types of Associations

- Boolean:

  Bread ^ Milk → Diapers

  Items are either purchased or not

- Quantitative:

  age in 30..39 ^ income in 42..48K → buys PC

  Look at a range of value

Number of Predicates Captured

- Single attribute:

  Bread ^ Milk → Diapers

  Just purchases

- Multiple attributes:

  age in 30..39 ^ income in 42..48K → buys PC

  ≠

- Multi-relational:

  buys(x, PC) ^ friends(x,y) → buys(y, PC)

  Look at relationship between individuals

Single or Multiple Level

- Single level:

  Beer → Diapers

  Generic item types

- Multiple level:

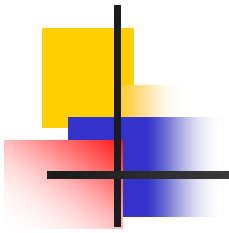
  Jupiler → Happy Baby

  Stella → Care

  Westmalle → Huggies

  Specific beer

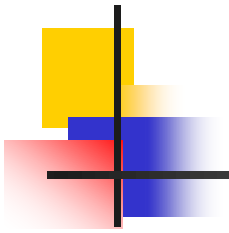
  Specific diaper brand



Applications: Retail

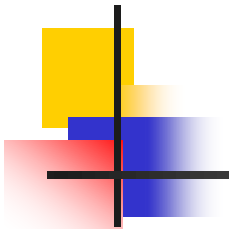
- Baskets = sets of products someone bought in one trip to the store

  Items = products
- Example application: given that many people buy beer and diapers together:
  - Run a sale on diapers; raise price of beer
  - Only useful if many buy diapers and beer
- Example application: What items should store stock up on



Application: Plagarism

- Baskets = sentences
  Items = documents containing those sentences
- Items that appear together too often could represent plagiarism
- Notice items do not have to be "in" baskets



Application: Web Pages

- Baskets = Web pages
  Items = words

- Unusual words appearing together in a large number of documents, e.g., "Brad" and "Angelina," may indicate an interesting relationship