

# Principles of Knowledge Discovery in Databases

Fall 1999

## Chapter 6: Mining Association Rules

Dr. Osmar R. Zaiane



Source:  
Dr. Jiawei Han

University of Alberta

## Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- Data mining operations
- Data summarization
- **Association analysis**
- Classification and prediction
- Clustering
- Web Mining
- Similarity Search
- *Other topics if time permits*



## Chapter 6 Objectives

Understand association analysis in large datasets and get a brief introduction to the different types of association rule mining

## Association Rules Outline



- What is association rule mining?
- How do we mine single-dimensional boolean associations?
- How do we mine multilevel associations?
- How do we mine multidimensional associations?
- Can we constrain the association mining?

## What Is Association Mining?

- **Association rule mining searches for relationships between items in a dataset:**
  - Finding association, correlation, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.
  - Rule form: “**Body** → **Head** [support, confidence]”.
- **Examples:**
  - buys(x, “bread”) → buys(x, “milk”) [0.6%, 65%]
  - major(x, “CS”) ^ takes(x, “DB”) → grade(x, “A”) [1%, 75%]



## Association Rule Mining

mining association rules (Agrawal et. al. SIGMOD93)	Fast algorithm (Agrawal et. al. VLDB94)	Partitioning (Navathe et. al. VLDB95)
Hash-based (Park et. al. SIGMOD95)	Multilevel A.R. (Han et. al. VLDB95)	Generalized A.R. (Srikant et. al. VLDB95)
Quantitative A.R. (Srikant et. al. SIGMOD96)	Incremental mining (Cheung et. al. ICDE96)	Parallel mining (Agrawal et. al. TKDE96)
Distributed mining (Cheung et. al. PDIS96)	Meta-ruleguided mining (Kamber et. al. KDD97)	Direct Itemset Counting (Brin et. al. SIGMOD97)
N-dimensional A.R. (Lu et. al. DMKD'98)	Constraint A.R. (Ng et. al. SIGMOD98'98)	A.R. with recurrent items (Zaiane et. al. ICDE'00)

And many many others:  
Spatial AR; Sequence Associations; AR for multimedia; AR in time series; AR with progressive refinement; etc.

## Basic Concepts

A transaction is a set of items:  $T = \{i_a, i_b, \dots, i_n\}$

$T \subset I$ , where  $I$  is the set of all possible items  $\{i_1, i_2, \dots, i_n\}$

$D$ , the task relevant data, is a set of transactions.

An association rule is of the form:

$P \rightarrow Q$ , where  $P \subset I$ ,  $Q \subset I$ , and  $P \cap Q = \emptyset$



## Basic Concepts (con't)

$P \rightarrow Q$  holds in  $D$  with support  $s$

and

$P \rightarrow Q$  has a confidence  $c$  in the transaction set  $D$ .

Support( $P \rightarrow Q$ ) = Probability( $P \cup Q$ )

Confidence( $P \rightarrow Q$ ) = Probability( $Q/P$ )

## Itemsets



A set of items is referred to as itemset.

An itemset containing  $k$  items is called **k-itemset**.

An itemset can also be seen as a conjunction of items (or a predicate)

## Support and Confidence

- **Support** of  $P = P_1 \wedge P_2 \wedge \dots \wedge P_n$  in  $D$ 
  - $\sigma(P/D)$  is the percentage of transactions  $T$  in  $D$  satisfying  $P$ . (number of  $T$  by cardinality of  $D$ ).
- **Confidence** of a rule  $P \rightarrow Q$ 
  - $\phi(P \rightarrow Q/D)$  ratio  $\sigma((P \wedge Q)/D)$  by  $\sigma(P/D)$
- **Thresholds:**
  - *minimum support*  $\sigma'$
  - *minimum confidence*  $\phi'$

## Strong Rules

- **Frequent (or large) predicate**  $P$  in set  $D$ 
  - support of  $P$  larger than minimum support,
- Rule  $P \rightarrow Q$  ( $c\%$ ) is **strong**
  - predicate ( $P \wedge Q$ ) is frequent (or large),
  - $c$  is larger than minimum confidence.

## Different Kinds of Association Rules

- **Boolean vs. Quantitative Associations**
  - Association on discrete vs. continuous data
  - Ex. Age( $X, 30-45$ )  $\wedge$  Income( $X, 50K-75K$ )  $\rightarrow$  Buys( $X, SUV$  car)
- **Boolean Association Rules**
- **Quantitative Association Rules**

## Different Kinds of Association Rules



- **Single dimension vs. multiple dimensional associations**
  - Based on the dimensions in data involved.
  - One predicate then single dimension. More predicates then multi-dimensions.
  - Ex. Buys(X, bread) → Buys(X, milk)  
Age(X,30-45) ∧ Income(X, 50K-75K) → Buys(X, SUV car)
- **Single-dimensional Association Rules**
- **Multi-dimensional Association Rules**

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  13

## Different Kinds of Association Rules



- **Single level vs. multiple-level analysis**
    - Based on the level of abstractions involved.
    - Find association rules at different levels of abstraction.
    - Ex. Buys(X, bread) → Buys(X, milk)  
Buys(X, Wheat Bread) → Buys(X, Formost 2% milk)
- 
- 
- **Single-level Association Rules**
  - **Multi-level Association Rules**

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  14

## Different Kinds of Association Rules

- **Single occurrence vs. multiple occurrences**
    - One item may occur more than once in the transaction.
    - Not the presence of the item is important but its frequency.
    - Ex. Buys(X, bread, 2) → Buys(X, milk, 1)
- 
- 
- **Single-occurrence-items Association Rules**
  - **Recurrent-items Association Rules**

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  15

## Different Kinds of Association Rules

- **Simple vs. constraint-based**
    - Constraints can be added on the rules to be discovered
  - **Association vs. correlation analysis**
    - Association does not necessarily imply correlation.
- $$\frac{P(A \wedge B)}{P(A)P(B)} = 1? >1? <1?$$

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  16

## Association Rules Outline



- What is association rule mining?
- How do we mine single-dimensional boolean associations?
- How do we mine multilevel associations?
- How do we mine multidimensional associations?
- Can we constrain the association mining?

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  17

## How do we Mine Association Rules?

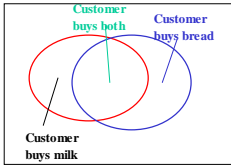
- **Input**
  - A database of transactions
  - Each transaction is a list of items (Ex. purchased by a customer in a visit)
- Find all rules that associate the presence of one set of items with that of another set of items.
  - Example: *98% of people who purchase tires and auto accessories also get automotive services done*
  - There are no restrictions on the number of items in the head or body of the rule.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta  18

## Rule Measures: Support and Confidence



Find all the rules  $X \& Y \rightarrow Z$  with minimum confidence and support

- support,  $s$ , probability that a transaction contains  $\{X, Y, Z\}$
- confidence,  $c$ , conditional probability that a transaction having  $\{X, Y\}$  also contains  $Z$ .

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Let minimum support 50%, and minimum confidence 50%, we have

- $A \rightarrow C$  (50%, 66.6%)
- $C \rightarrow A$  (50%, 100%)

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 19

## Mining Association Rules

Transaction ID	Items Bought
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Min. support 50%  
Min. confidence 50%

Frequent Itemset	Support
{A}	75%
{B}	50%
{C}	50%
{A,C}	50%

For rule  $A \rightarrow C$ :

support = support( $\{A, C\}$ ) = 50%

confidence = support( $\{A, C\}$ )/support( $\{A\}$ ) = 66.6%

The Apriori principle:

Any subset of a frequent itemset must be frequent.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 20

## Mining Frequent Itemsets: the Key Step

① Find the *frequent itemsets*: the sets of items that have minimum support

- ◆ A subset of a frequent itemset must also be a frequent itemset, i.e., if  $\{AB\}$  is a frequent itemset, both  $\{A\}$  and  $\{B\}$  should be frequent itemsets
- ◆ Iteratively find frequent itemsets with cardinality from 1 to  $k$  ( $k$ -itemsets)

② Use the frequent itemsets to generate association rules.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 21

## The Apriori Algorithm

$C_k$ : Candidate itemset of size  $k$

$L_k$ : frequent itemset of size  $k$

```

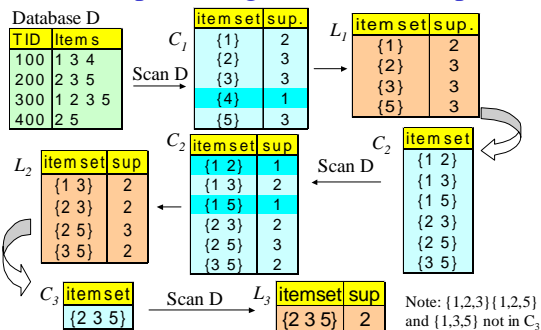
 $L_1 = \{\text{frequent items}\};$ 
for ( $k = 1; L_k \neq \emptyset; k++$ ) do begin
     $C_{k+1} = \text{candidates generated from } L_k;$ 
    for each transaction  $t$  in database do
        increment the count of all candidates in  $C_{k+1}$  that are contained in  $t$ 
     $L_{k+1} = \text{candidates in } C_{k+1} \text{ with min\_support}$ 
    end
return  $\cup_k L_k;$ 
    
```

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 22

## The Apriori Algorithm -- Example



© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 23

## Generating Association Rules from Frequent Itemsets

- Only strong association rules are generated.
- Frequent itemsets satisfy minimum support threshold.
- Strong AR satisfy minimum confidence threshold.

$$\text{Confidence}(A \rightarrow B) = \text{Prob}(B/A) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

**For each** frequent itemset,  $f$ , generate all non-empty subsets of  $f$ .  
**For every** non-empty subset  $s$  of  $f$  **do**  
 output rule  $s \rightarrow (f-s)$  if support( $f$ )/support( $s$ )  $\geq$  min\_confidence  
**end**

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 24

## Improving efficiency of Apriori

- **Reducing the number of scans**  
(there are  $k$  DB scans for  $k$ -itemsets)
- **Eliminating scans by indexing** (Hashing)
- **Reducing sizes and number of transactions**  
(no need for non frequent items)
- **Partitioning**

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

25

## Optimization: Direct Hash and Pruning

- **DHP: Direct Hash and Pruning** (Park, Chen and Yu, SIGMOD'95).
  - Reduce the size of candidate sets to minimize the cost
  - Reduce the size of the transaction database as well
- Using a hash table to keep track the counts of 2-itemset. Using the counts to prune  $C_2$  ( $C_2$  is usually the largest)
- An item in transaction  $t$  can be trimmed if it does not appear in at least  $k$  of the candidate  $k$ -itemsets in  $t$ .

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

26

## Optimization: The Partitioning Algorithm

- **Partition** (Savasere, Omiecinski, & Navathe, VLDB'95).
  - Divide database into  $n$  partitions.
  - A frequent item must be frequent in at least one partition.
  - Process one partition in main memory at a time:
    - For each partition, generate frequent itemsets using the Apriori algorithm
    - also form *tidlist* for all item sets to facilitate counting in the merge phase
  - After all partitions are processed, the local frequent itemsets are merged into global frequent sets; support can be computed from the *tidlists*.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

27

## Optimization: Sampling and Itemset Counting

- **Sampling** (Toivonen, VLDB'96).
  - A probabilistic approach finds association rules in about one pass.
- **Dynamic Itemset Counting** (Brin et. al. SIGMOD'97)
  - Reducing the number of scans over the transactions by starting to count itemsets dynamically during scans
  - Using data structure to keep track of counters and reordering items to reduce increment costs

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

28

## Incremental Update of Discovered Rules

- Partitioned derivation and incremental updating.
- A fast updating algorithm, FUP (Cheung et al.'96)
  - View a database: original  $DB \cup$  incremental  $db$ .
  - A  $k$ -itemset (for any  $k$ ),
    - \* **frequent** in  $DB \cup db$  if frequent in both  $DB$  and  $db$ .
    - \* **non frequent** in  $DB \cup db$  if also in both  $DB$  and  $db$ .
  - For those only frequent in  $DB$ , merge corresponding counts in  $db$ .
  - For those only frequent in  $db$ , search  $DB$  to update their itemset counts.
- Similar methods can be adopted for data removal and update.
- Principles applicable to distributed/parallel mining.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

29

## Parallel and Distributed Mining

- **PDM** (Park et al.'95):
  - Use a hashing technique (DHP-like) to identify candidate  $k$ -itemsets from the local databases.
- **Count Distribution** (Agrawal & Shafer'96):
  - An extension of the Apriori algorithm.
  - May require a lot of messages in count exchange.
- **FDM** (Cheung et al.'96).
  - Observation: If an itemset  $X$  is globally large, there exists partition  $Di$  such that  $X$  and all its subsets are locally large at  $Di$ .
  - Candidate sets are those which are also local candidates in some component database, plus some message passing optimizations.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta

30

## Association Rules Outline



- What is association rule mining?
- How do we mine single-dimensional boolean associations?
- How do we mine multilevel associations?
- How do we mine multidimensional associations?
- Can we constrain the association mining?

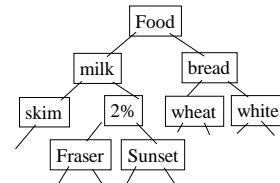
© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 31

## Multiple-Level Association Rules

- Items often form hierarchy.
- Items at the lower level are expected to have lower support.
- Rules regarding itemsets at appropriate levels could be quite useful.
- Transaction database can be encoded based on dimensions and levels
- It is smart to explore shared multi-level mining (Han & Fu, VLDB'95).



TID	Items
T1	{111, 121, 211, 221}
T2	{111, 211, 222, 323}
T3	{112, 122, 221, 411}
T4	{111, 121}
T5	{111, 122, 211, 221, 413}

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 32

## Mining Multi-Level Associations

- A **top\_down, progressive deepening approach**.
  - First find high-level strong rules:  
milk  $\rightarrow$  bread [20%, 60%].
  - Then find their lower-level “weaker” rules:  
2% milk  $\rightarrow$  wheat bread [6%, 50%].
- **Variations at mining multiple-level association rules.**
  - Level-crossed association rules:  
2% milk  $\rightarrow$  *Wonder wheat bread*
  - Association rules with multiple, alternative hierarchies:  
2% milk  $\rightarrow$  *Wonder bread*

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 33

## Multi-Level Mining: Progressive Deepening

- A top-down, progressive deepening approach:
  - First mine high-level frequent items:  
milk (15%), bread (10%)
  - Then mine their lower-level frequent itemsets:  
2% milk (5%), wheat bread (4%)

When one threshold set for all levels; if support too high the possible miss meaningful associations at low level; if support too low the possible generation of uninteresting rules

- Different minimum support threshold across multi-levels lead to different algorithms.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 34

## Approaches to Mining Multi-level Association Rules



- Uniform minimum support for all levels
  - Same support  $\sigma$  for all levels
  - Avoid examining itemsets containing items whose ancestor is not frequent.
  - Simpler, but it is unlikely that lower level items are as frequent as higher level items.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 35

## Approaches to Mining Multi-level Association Rules



- Reduced minimum support at lower levels

Examine only those descendents whose ancestor's support is frequent or non-negligible (controlled).

### – Level-by-level independent

Full depth search

### – Level-cross filtering by single item

A specific association is examined from a more general one  $\Rightarrow$  items are examined only if parents are frequent.

### – Level-cross filtering by k-itemsets

Frequency of ancestry examined for k-itemsets and not just items

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 36

## Association Rules Outline



- What is association rule mining?
- How do we mine single-dimensional boolean associations?
- How do we mine multilevel associations?
- How do we mine multidimensional associations?
- Can we constrain the association mining?

## Mining Multi-Dimensional Associations

- **Multi-dimensional vs. transaction-based associations**
  - Multi-dimensional (linking different attributes)
    - $\text{major}(x, \text{"cs"}) \wedge \text{region}(x, \text{"oxford"}) \rightarrow \text{gpa}(x, \text{"good"})$ .
  - Transaction-based (linking the same kind of attributes)
    - $\text{takes}(x, \text{"chemistry"}) \wedge \text{takes}(x, \text{"biology"}) \rightarrow \text{takes}(x, \text{"bio-chemistry"})$ .
- **Multi-level association (drilling on any dimension)**
  - Lower levels often adopt lower *min\_support* thresholds.
- **Method:**
  - Construct data cube (with count/frequency aggregated)
  - Perform level-wise/dimension-wise search in the data cube (Kamber et al., KDD'97).

## Categorical and Quantitative

In a multidimensional context there are:

- Categorical dimensions (attributes)
  - Ex. Occupation, Location, etc.
- Quantitative dimensions (attributes)
  - Ex. Price, Age, etc.



Apriori, as it is, does not handle quantitative data.

## Quantitative Association Rules

RecordID	Age	Married	NumCars
100	23	No	1
200	25	Yes	1
300	29	No	0
400	34	Yes	2
500	38	Yes	2

Sample Rules	Support	Confidence
$\langle \text{age}:30..39 \rangle \text{ and } \langle \text{married}: \text{yes} \rangle \Rightarrow \langle \text{numCars}:2 \rangle$	40%	100%
$\langle \text{NumCars}: 0..1 \rangle \Rightarrow \langle \text{Married}: \text{No} \rangle$	40%	66.70%

## Mapping Quantitative to Boolean

- One possible solution is to map the problem to the Boolean association rules:
  - discretizing a non-categorical attribute to intervals
    - Age [20,29], [30,39],...
  - forming Boolean records
    - categorical attributes: each value becomes one item
    - non-categorical attributes: each interval becomes one item

RecordID	Age	Married	NoCars
100	23	No	1
500	38	Yes	2



RecID	Age: 20..29	Age: 30..39	Married: Yes	Married: No	Cars: 0	Cars: 1	Cars: 2
100	1	0	0	1	0	1	0
500	0	1	1	0	0	0	1

## Mining Quantitative Association Rules

- Problems with the mapping
  - too few intervals: lost information
  - too low support: too many rules
- Solutions
  - using the supports of an itemset and its generalizations to determine the intervals
  - Binning (equi-width, equi-dept, distance based)
  - using interest measure to control the number of association rules

## Association Rules Outline



- What is association rule mining?
- How do we mine single-dimensional boolean associations?
- How do we mine multilevel associations?
- How do we mine multidimensional associations?
- Can we constrain the association mining?

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 43

## Restricting Association Rules



- Useful for interactive and ad-hoc mining
- Reduces the the set of association rules discovered and confines them to more relevant rules.
- **Before mining**
  - ✓ Knowledge type constraints: classification, etc.
  - ✓ Data constraints: SQL-like queries (DMQL)
  - ✓ Dimension/level constraints: relevance to some dimensions and some concept levels.
- **While mining**
  - ✓ Rule constraints: form, size, and content.
  - ✓ Interestingness constraints: support, confidence, correlation.
- **After mining**
  - ✓ Querying association rules

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 44

## Rule Constraints in Association Mining

- Two kind of rule constraints:
  - Rule form constraints: meta-rule guided mining.
    - $P(x, y) \wedge Q(x, w) \rightarrow \text{takes}(x, \text{"database systems"})$ .
  - Rule content constraint: constraint-based query optimization (where and having clauses)(Ng, et al., SIGMOD'98).
    - $\text{sum}(\text{LHS}) < 100 \wedge \text{min}(\text{LHS}) > 20 \wedge \text{count}(\text{LHS}) > 3 \wedge \text{sum}(\text{RHS}) > 1000$
- **1-variable vs. 2-variable constraints** (Lakshmanan, et al. SIGMOD'99):
  - 1-var: A constraint confining only one side (L/R) of the rule, e.g., as shown above.
  - 2-var: A constraint confining both sides (L and R).
    - $\text{sum}(\text{LHS}) < \text{min}(\text{RHS}) \wedge \text{max}(\text{RHS}) < 5 * \text{sum}(\text{LHS})$

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 45

## Constrained Association Query Optimization Problem

- Given a set of constraints C, the algorithm should:
  - Find only the frequent sets that satisfy the given constraints C
  - Find all frequent sets that satisfy the given constraints C
- A naïve solution:
  - Apply Apriori for finding all frequent sets, and then test them for constraint satisfaction one by one.
- Better approach:
  - Comprehensive analysis of the properties of constraints and try to **push them as deeply as possible inside** the frequent set computation.

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 46

## Presentation of Association Rules (Table Form)

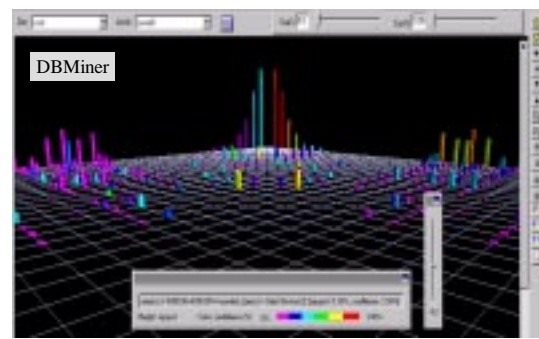
LHS	RHS	support	confidence	lift
beer, bread	beer, bread, butter	0.001	0.001	0.001
beer, bread	beer, bread, milk	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham, cheese	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham, cheese, wine	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham, cheese, wine, coffee	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham, cheese, wine, coffee, juice	0.001	0.001	0.001
beer, bread	beer, bread, butter, milk, eggs, ham, cheese, wine, coffee, juice, tea	0.001	0.001	0.001

© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 47

## Visualization of Association Rule in Plane Form



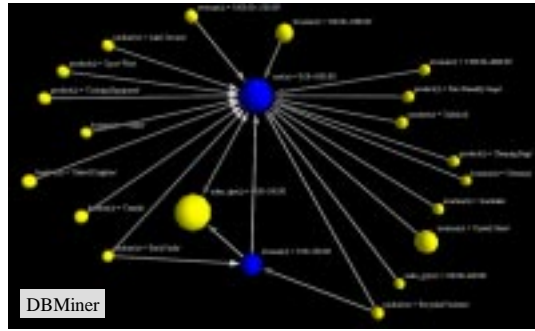
© Dr. Omar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta 48



### Visualization of Association Rule Using Rule Graph



© Dr. Ouzar R. Zaiane, 1999

Principles of Knowledge Discovery in Databases

University of Alberta