Whither Data Mining?

Rakesh Agrawal
Ramakrishnan Srikant

Intelligent Information Systems Research
IBM Almaden Research Center

Outline

- Association Rules & the Apriori Algorithm
  - A quick retrospection of VLDB 94 paper
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
  - Beyond Market Basket Analysis
- Whither?

Association Rules
(Agrawal, Imielinski & Swami: SIGMOD '93)

- $I = \{ i_1, i_2 \ldots, i_m \}$: a set of literals, called items.
- Transaction T: a set of items such that T ⊆ I.
- Database D: a set of transactions.
- A transaction T contains X, a set of some items in I, if X ⊆ T.
- An association rule is an implication of the form X ⇒ Y, where X, Y ⊆ I.
  - Support: % of transactions in D that contain X ∪ Y.
  - Confidence: Among transactions that contain X, what % also contain Y.
- Find all rules that have support and confidence greater than user-specified minimum support and minimum confidence.

What Was New About This Formalism?

- Emphasis on finding all rules (Completeness)
  - Strong rules tend to be already known.
- Discovery vs. Hypothesis Testing

Computing Association Rules: Problem Decomposition

- Find all sets of items that have minimum support (frequent itemsets).

- Use the frequent itemsets to generate the desired rules.
  - confidence (X ε Y) = support (X . Y) / support (X)

Computing Association Rules: Problem Decomposition

- Find all sets of items that have minimum support (frequent itemsets).

  What itemsets should you count?

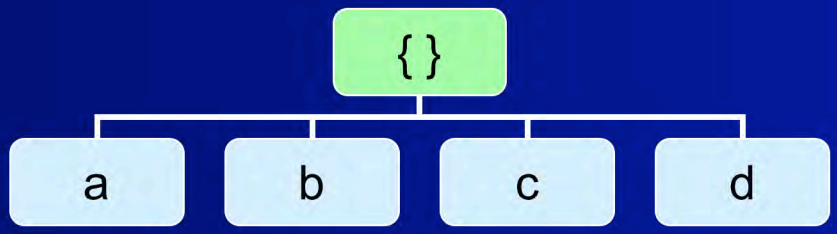
  How do you count them efficiently?

- Use the frequent itemsets to generate the desired rules.
  - confidence (X ε Y) = support (X . Y) / support (X)

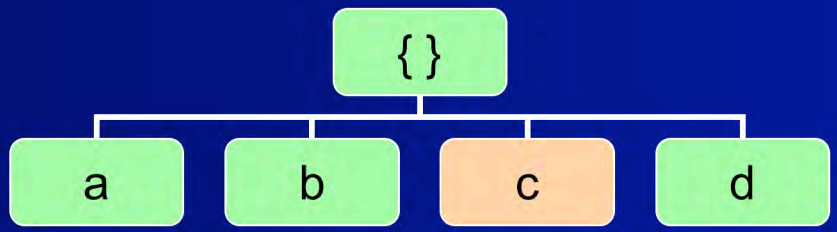
What itemsets do you count?

- Search space is exponential.
  - With n items, ${}^{n}C_{k}$ potential candidates of size k.
- Anti-monotonicity: Any superset of an infrequent itemset is also infrequent (SIGMOD ’93).
  - If an itemset is infrequent, don’t count any of its extensions.
- Flip the property: All subsets of a frequent itemset are frequent.
- Need not count any candidate that has an infrequent subset (VLDB ’94)
  - Simultaneously observed by Mannila et al., KDD ’94
- Broadly applicable to extensions and restrictions.

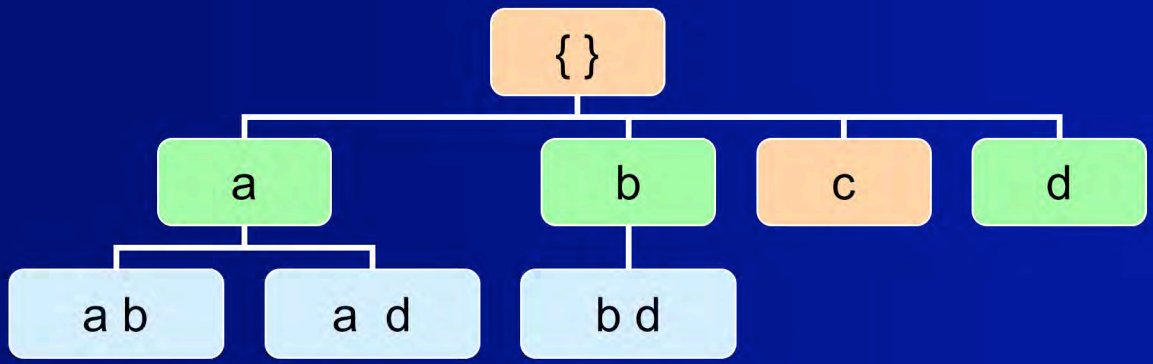
Apriori Algorithm: Breadth First Search



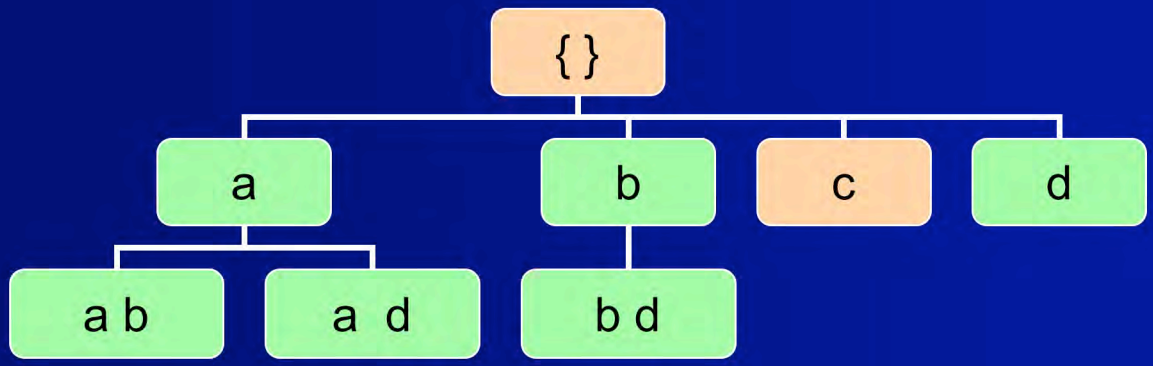
Apriori Algorithm: Breadth First Search



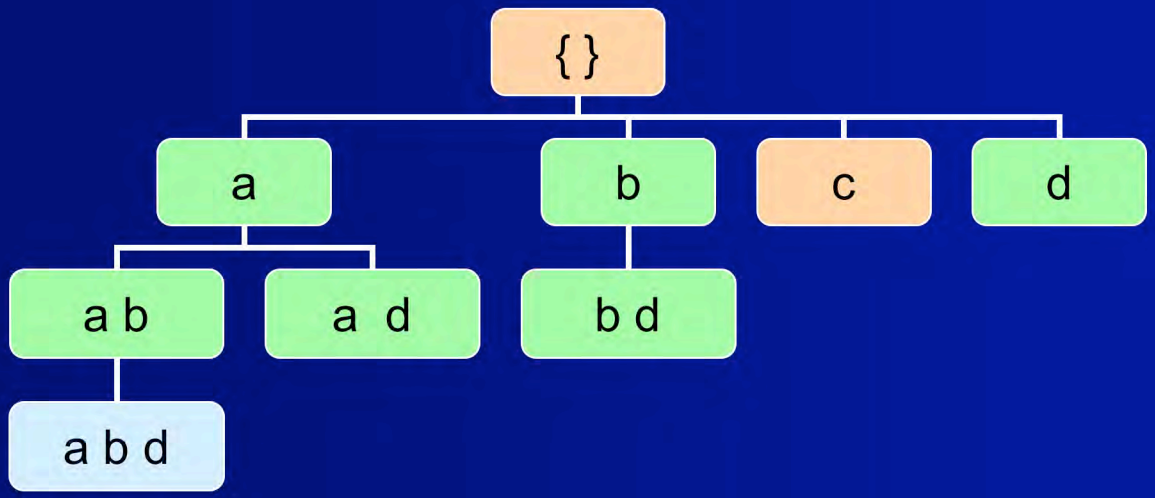
Apriori Algorithm: Breadth First Search



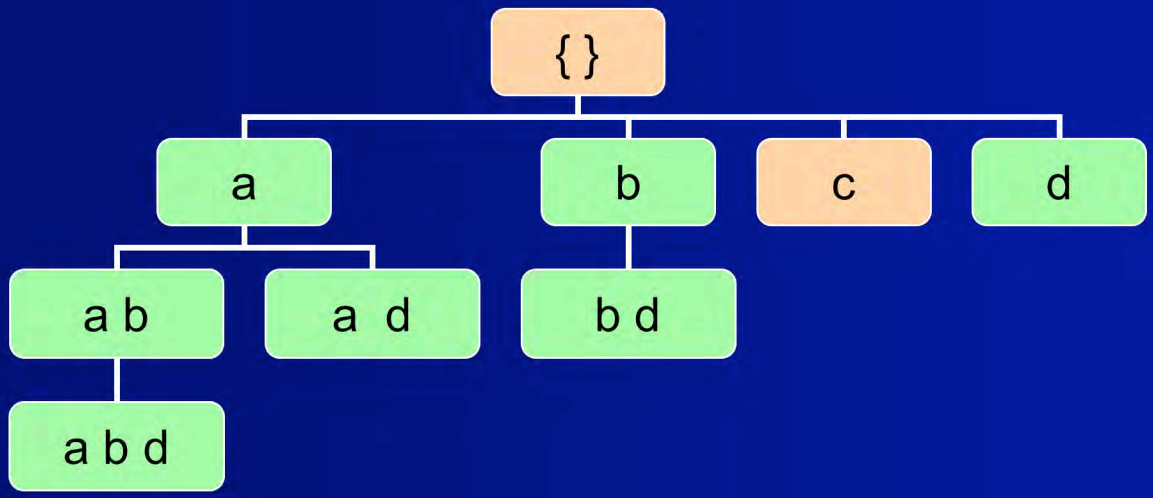
Apriori Algorithm: Breadth First Search



Apriori Algorithm: Breadth First Search



Apriori Algorithm: Breadth First Search



APRIORI Candidate Generation (VLDB 94)

- L_k : Frequent itemsets of size k, C_k : Candidate itemsets of size k
- Given L_k, generate C_{k+1} in two steps:
  - Join Step : Join L_k with L_k, with the join condition that the first k-1 items should be the same and $l^1[k] < l^2[k]$.

L_3
{ a b c }
{ a b d }
{ a c d }
{ a c e }
{ b c d }

C_4
{ a b c d }
{ a c d e }

APRIORI Candidate Generation (VLDB 94)

- L_k : Frequent itemsets of size k, C_k : Candidate itemsets of size k
- Given L_k, generate C_{k+1} in two steps:
  - Join Step : Join L_k with L_k, with the join condition that the first k-1 items should be the same and $l^1[k] < l^2[k]$.
  - Prune Step : Delete all candidates which have a non-frequent subset.

L_3
{ a b c }
{ a b d }
{ a c d }
{ a c e }
{ b c d }

C_4
{ a b c d }
{ a c d e }

APRIORI Candidate Generation (VLDB 94)

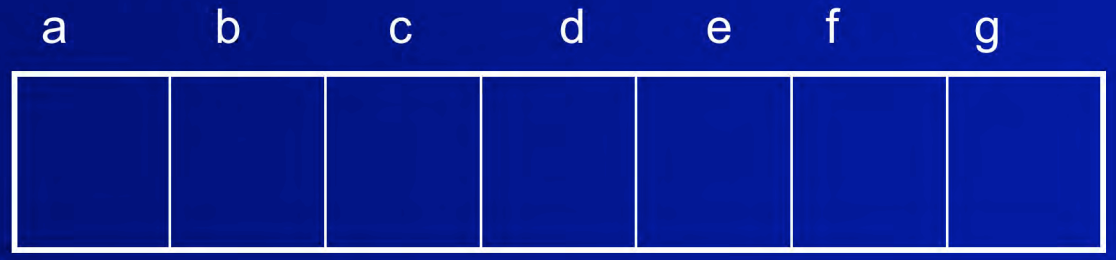
- L_k : Frequent itemsets of size k, C_k : Candidate itemsets of size k
- Given L_k, generate C_{k+1} in two steps:
  - Join Step : Join L_k with L_k, with the join condition that the first k-1 items should be the same and $l^1[k] < l^2[k]$.
  - Prune Step : Delete all candidates which have a non-frequent subset.

L_3
{ a b c }
{ a b d }
{ a c d }
{ a c e }
{ b c d }

C_4
{ a b c d }
~~{ a c d e }~~

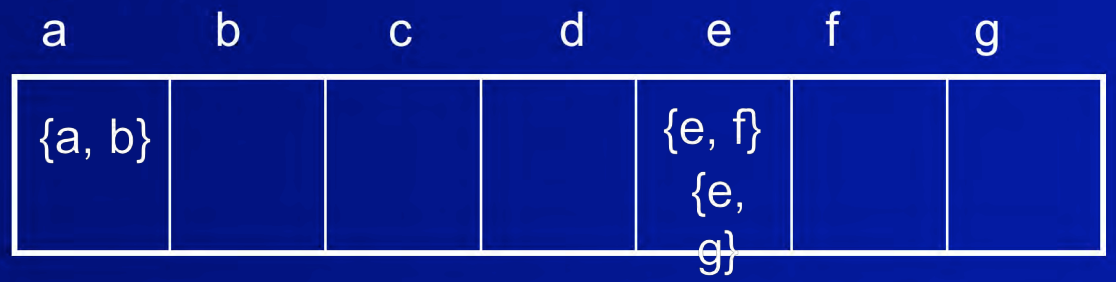
How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { {a, b} {e, f} {e, g} }
  - T : {c, e, f}



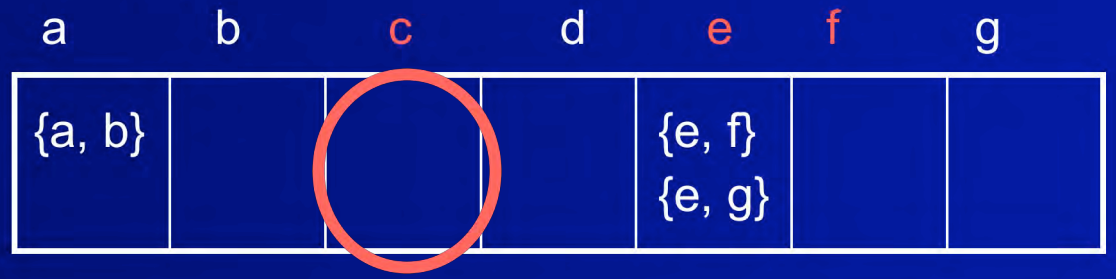
How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { }
  - T : {c, e, f}



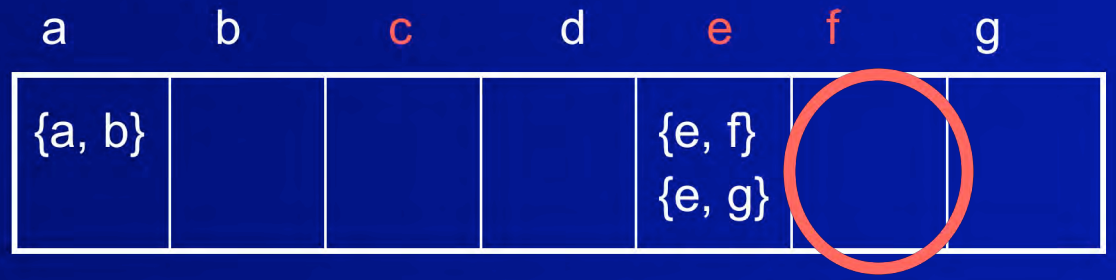
How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { }
  - T : {c, e, f}



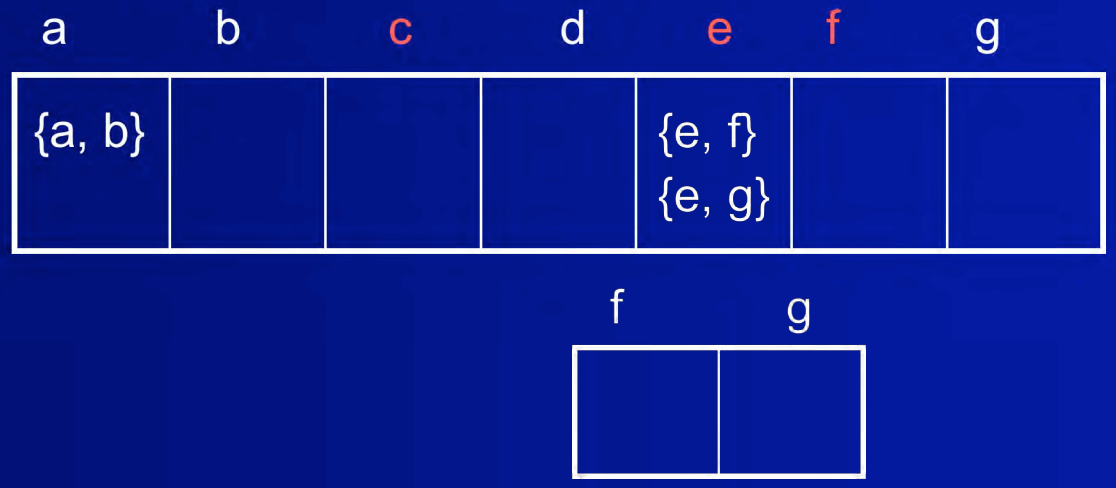
How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { }
  - T : {c, e, f}



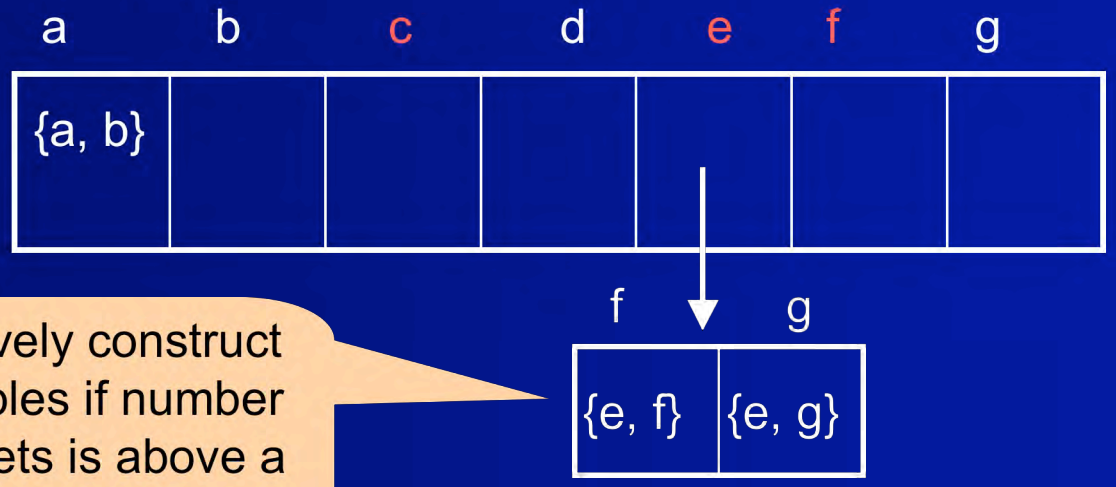
How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { }
  - T : {c, e, f}



How do you count?

- Given a set of candidates C_k, for each transaction T:
  - Find all members of C_k which are contained in T.
- Hash-tree data structure [VLDB '94]
  - C2 : { }
  - T : {c, e, f}



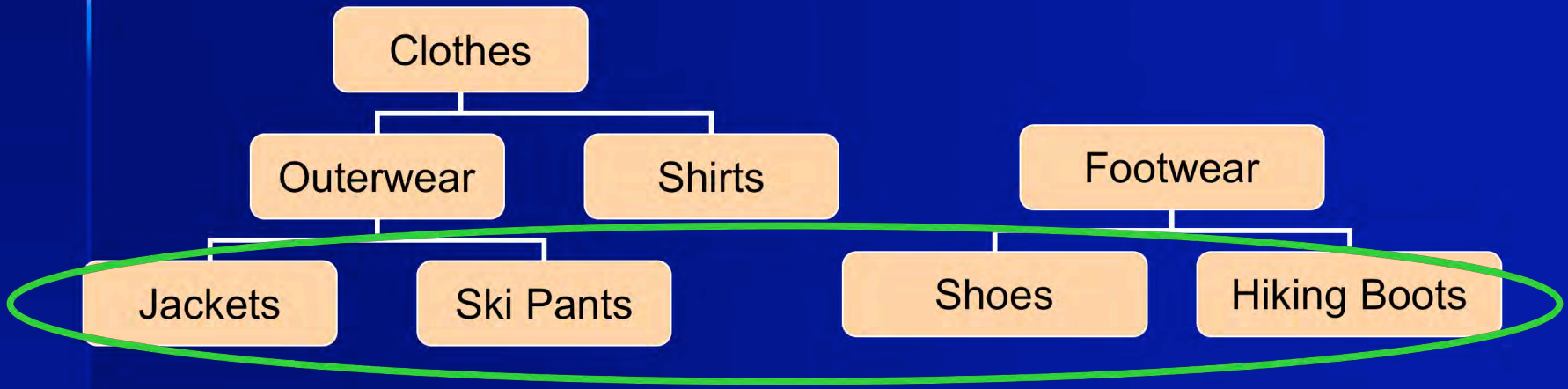
Synthetic Datasets

- Transactions designed to mimic transactions in the retailing environment.
- Model: People tend to buy certain sets of items together.
  - Each such set is potentially a maximal frequent itemset.
  - Some people may only buy some of the items from such a set.
  - Frequent itemsets often have common items.
- Used in many studies for comparing the performance of association rule discovery algorithms.

Outline

- Association Rules & the Apriori Algorithm
  - A quick retrospection of VLDB 94 paper
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
    - Generalizations
    - Restrictions
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
- Whither?

Taxonomies



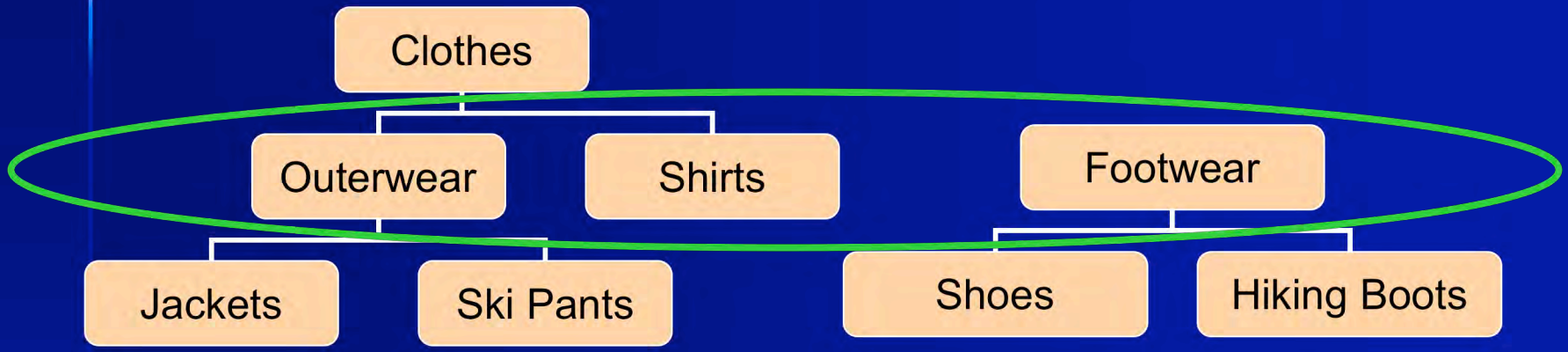
- Can we just replace each item with its generalization?
  - "Navy Squall Jacket Med" τ "Jackets"

R. Srikant and R. Agrawal, "Mining Generalized Association Rules", VLDB '95.

J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases", VLDB '95.

Taxonomies

Can miss rules that run across levels.



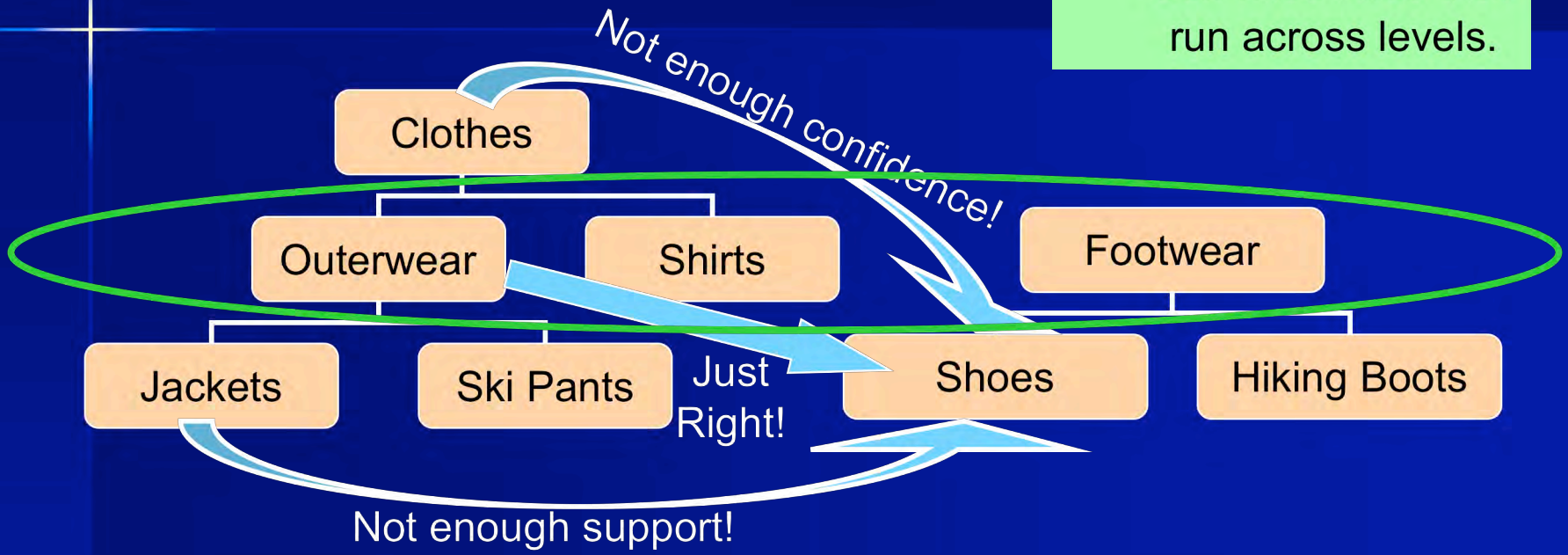
- Can we just replace each item with its generalization?
  - "Navy Squall Jacket Med" τ "Jackets"

R. Srikant and R. Agrawal, "Mining Generalized Association Rules", VLDB '95.

J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases", VLDB '95.

Taxonomies

Can miss rules that run across levels.



- Can we just replace each item with its generalization?
  - "Navy Squall Jacket Med" τ "Jackets"

R. Srikant and R. Agrawal, "Mining Generalized Association Rules", VLDB '95.

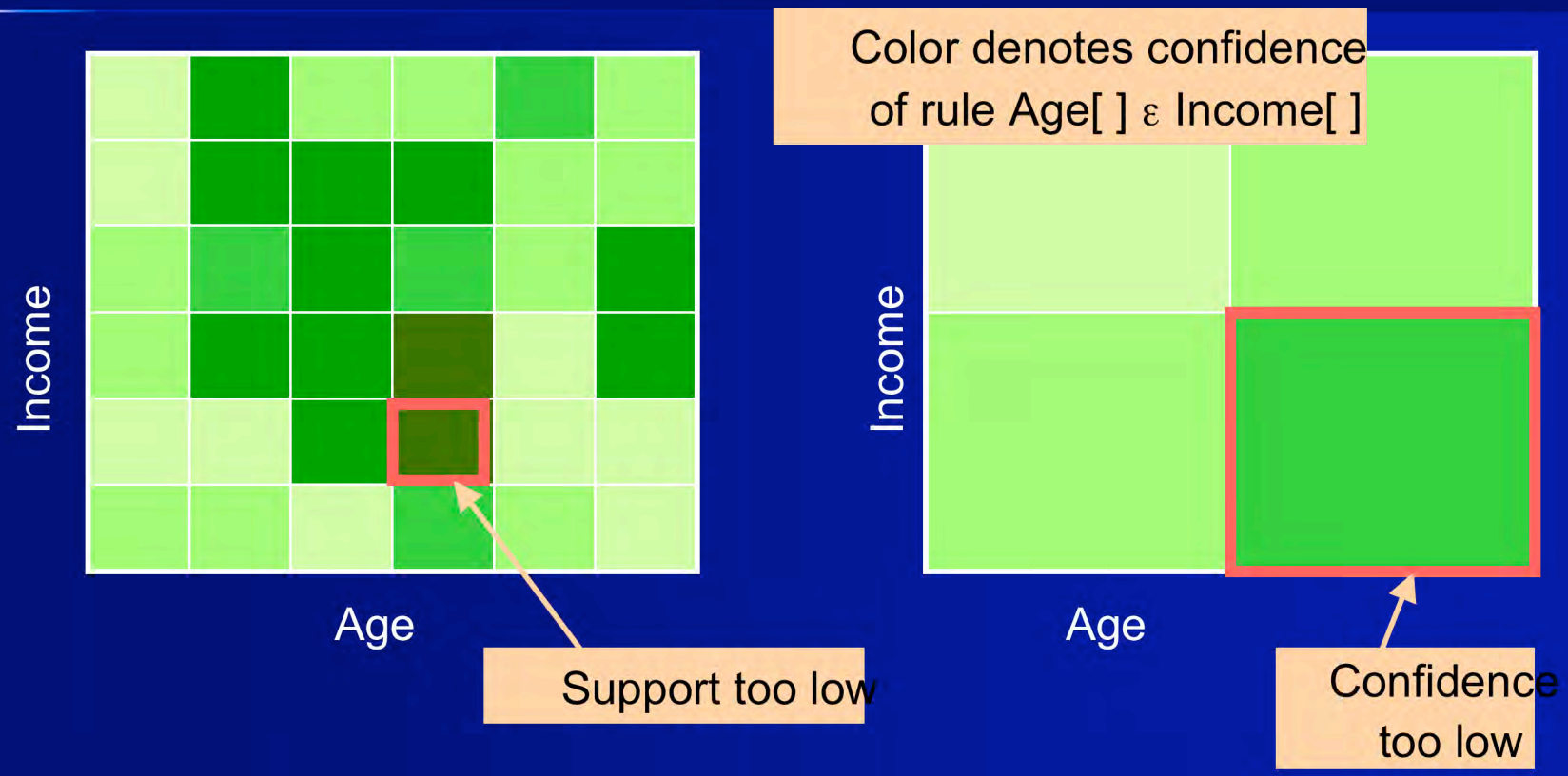
J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases", VLDB '95.

Quantitative Associations

- How do we deal with categorical & quantitative attributes?
  - Example: “30% of married people between age 45 and 60 have at least 2 cars; 5% of records have these properties”
- Can we just map this problem to boolean associations?
- Can discretize intervals & map each attribute into a set of boolean attributes.
- But …

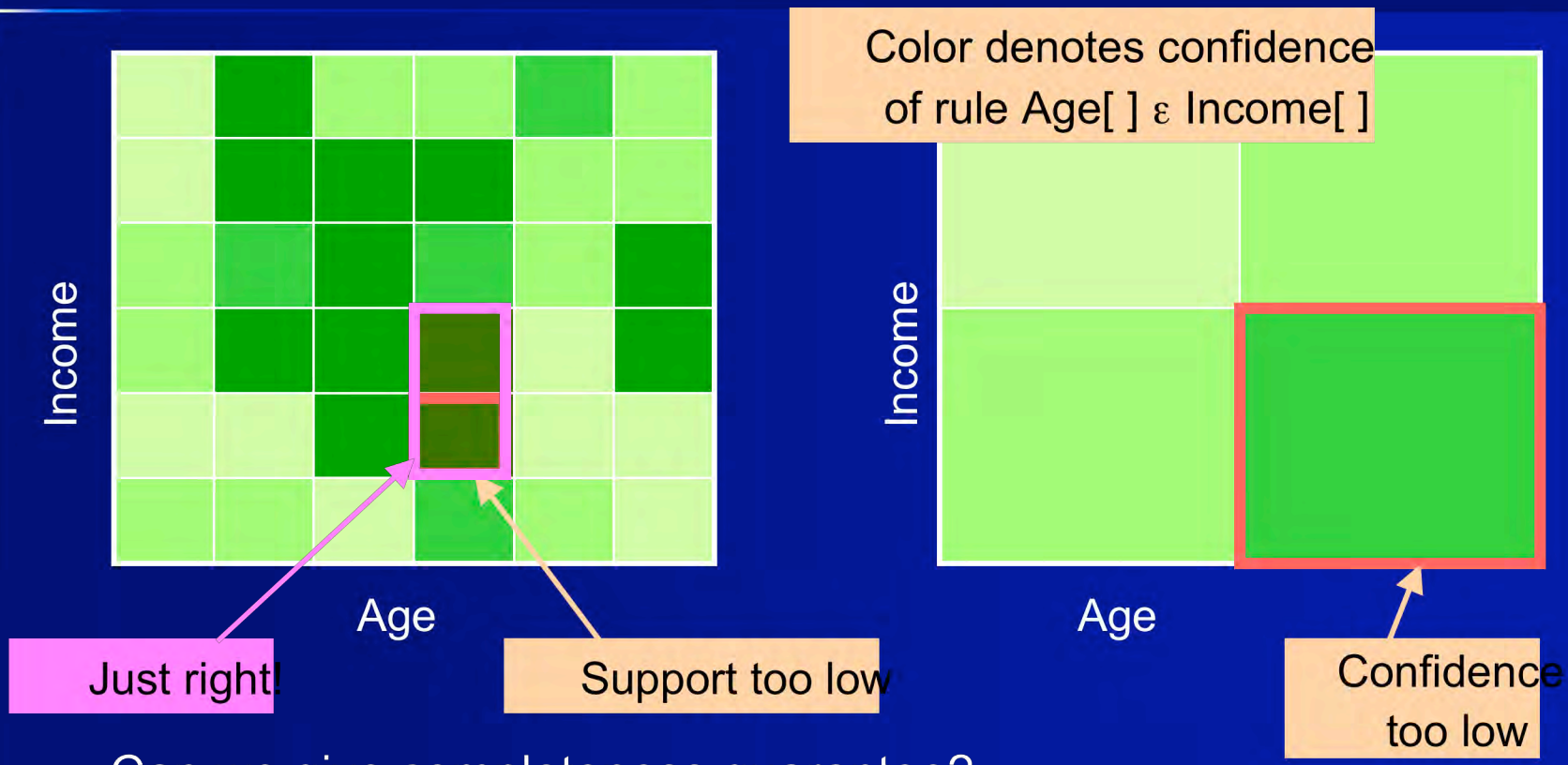
R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", SIGMOD ‘96

Quantitative Association: The Discretization Dilemma



R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", SIGMOD '96

Quantitative Association: The Discretization Dilemma

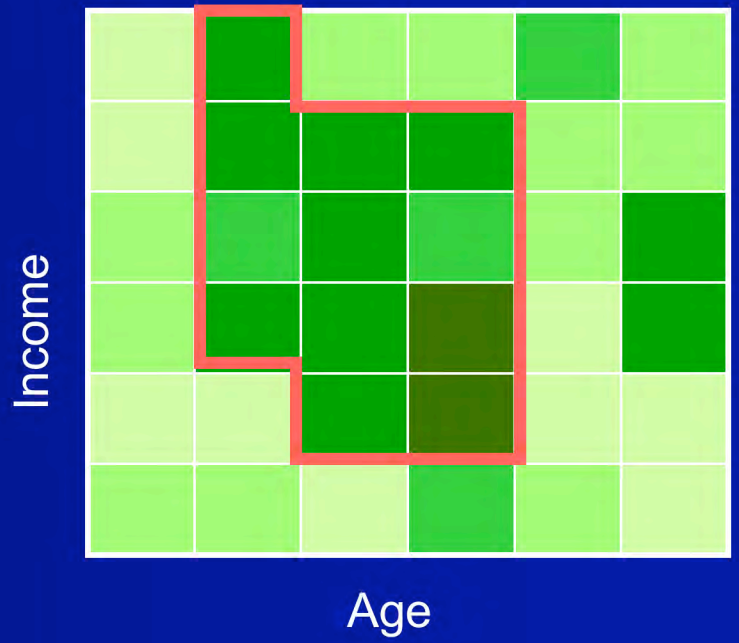


- Can we give completeness guarantee?

R. Srikant and R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", SIGMOD '96

Quantitative Associations: Optimized Rectilinear Rules

- Find the rectilinear region that maximizes support * confidence (or other criteria) for a fixed RHS.
- Example: Color denotes confidence.
- Optimize tradeoff between including more regions (higher support) vs. focusing only on high confidence regions.



K. Yoda, T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, "Computing Optimized Rectilinear Regions for Association Rules", KDD '97.

Sequential Patterns

- Example: 10% of customers bought “shirts” and “jackets” in one transaction, followed by “shoes" in another transaction.
  - Inter-transaction patterns.
  - Can extend to sequential pattern rules.
- Constraints, e.g., max/min time gap between successive elements of the pattern.
- Not all subsequences of a frequent sequence are frequent.
  - Sequence not supported if it doesn’t meet the constraint.

R. Agrawal and R. Srikant, “Mining Sequential Patterns”, ICDE ’95.

H. Mannila, H. Toivonen, and I. Verkamo, "Discovering frequent episodes in sequences," KDD ’95.

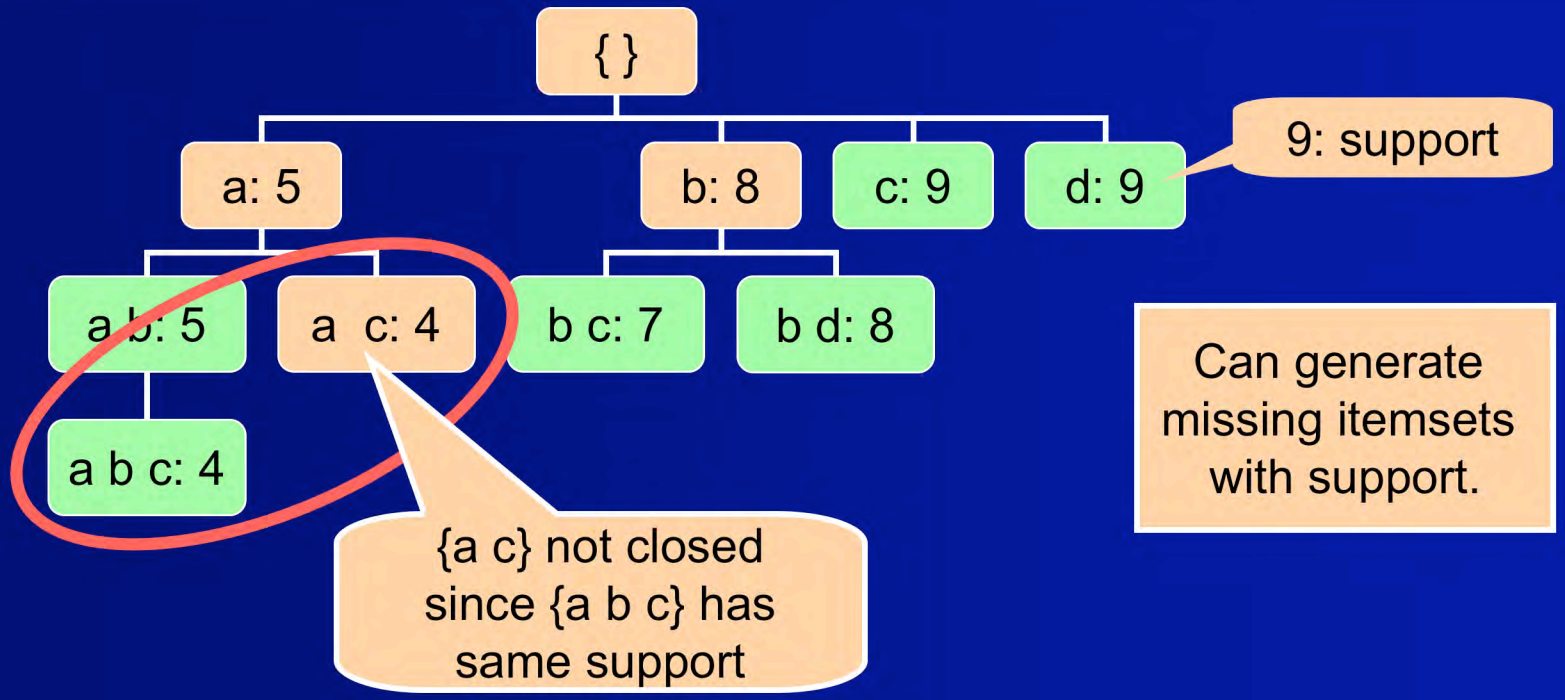
Graphs, Trees, …

- Graphs:
  - M. Kuramochi and G. Karypis, “Discovering frequent geometric subgraphs”, ICDM ‘01
  - X. Yan and J. Han , “gSpan: Graph-Based Substructure Pattern Mining”, ICDM '02.
- Trees:
  - M. Zaki, “Efficiently mining frequent trees in a forest”, KDD ‘02

Outline

- Association Rules & the Apriori Algorithm
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
    - Generalizations
    - Restrictions
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
- Whither?

Closed Itemsets

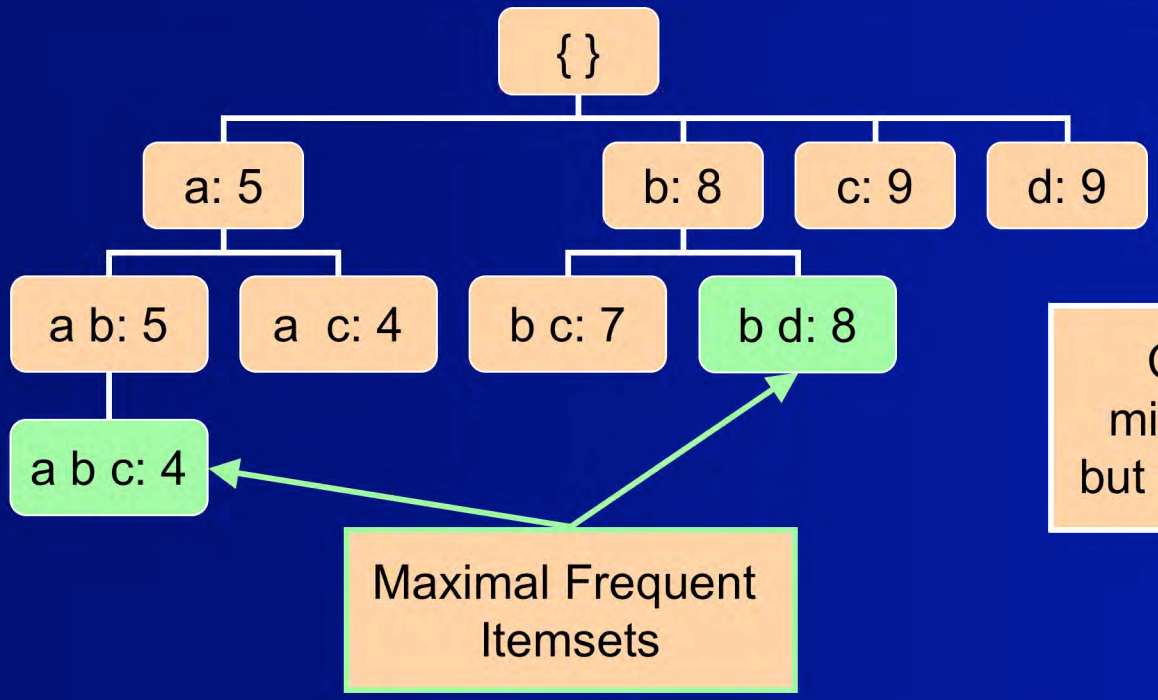


R.J. Bayardo et al., “Brute-force mining of high-confidence classification rules”, KDD ‘97

N. Pasquier et al., “Discovering frequent closed itemsets for association rules”, ICDT ’99.

M.J. Zaki, “Generating non-redundant association rules”, KDD 2000.

Maximal Itemsets



R.J. Bayardo, “Efficiently Mining Long Patterns from Databases", SIGMOD '98.

D-I Lin and Z. Kedem, “Pincer Search Pincer-Search: A New Algorithm for Discovering the Maximum Frequent Set”, EDBT ’98

Constraints over Items

- Boolean expressions over (the presence of) items.
  - (Shirts AND Shoes) OR (Outerwear AND NOT Hiking Boots)
- Anti-monotonicity property no longer holds.
  - {Shirt, Shoes} satisfies the constraint but not {Shirt} or {Shoes}.

R. Srikant, Q. Vu and R. Agrawal, “Mining Association Rules with Item Constraints”, KDD '97

Constraints over Properties of Items

- Example: Avg. price of items in rule should be > $10.
- Can classify constraints into groups:
  - Anti-monotone, e.g., minimum support, min(S) μ v
  - Monotone, e.g., min(S) [v
  - Succint, e.g., min(S) [v, min(S) μ v
  - Convertible, e.g., avg(S) [v, avg(S) μ v
- These four groups can be pushed (to varying degrees).
- Not all important constraints fall into these groups, e.g., min confidence, sum(S) μ v

R. Ng et al., "Exploratory Mining and Pruning Optimizations of Constrained Association Rules", SIGMOD '98.

J. Pei and J. Han, "Can We Push More Constraints into Frequent Pattern Mining?", KDD 2000.

Outline

- Association Rules & the Apriori Algorithm
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
- Whither?

Algorithmic Innovations

- Reducing the cost of checking whether a candidate itemset is contained in a transaction:
  - TID intersection.
  - Database projection.
- Reducing the number of passes over the data:
  - Sampling & Dynamic Counting
- Reducing the number of candidates counted:
  - For maximal patterns & constraints.
- *Many other innovative ideas … unfortunately, not covered in this talk – Apologies in advance!*

Algorithmic Innovations

- Reducing the cost of checking whether a candidate itemset is contained in a transaction:
  - TID intersection.
  - Database projection.

Themes

Focused on CPU cost.
Dense data / longer rules.
Transform the dataset.
Cost of memory & disk.

TID Intersection

- Convert data from “horizontal” to “vertical” format.
- For each frequent item, list of TIDs of transactions that contain the item.

Trans. ID (TID)	Items
10	a, c, e
20	b, d
30	a, e
40	b, e

Item	TID List
a	10, 30
b	20, 40
e	10, 30, 40

TID Intersection

- Convert data from “horizontal” to “vertical” format.
- For each frequent item, list of TIDs of transactions that contain the item.

Item	TID List
a	10, 30
b	20, 40
e	10, 30, 40

TID Intersection

- Convert data from “horizontal” to “vertical” format.
- For each frequent item, list of TIDs of transactions that contain the item.
- Intersect the TID lists of two subsets to get the TID list for a candidate itemset.

Item	TID List
a	10, 30
b	20, 40
e	10, 30, 40

Itemset	TID List
a, b	
a, e	10, 30
b, e	40

TID Intersection

- Convert data from “horizontal” to “vertical” format.
- For each frequent item, list of TIDs of transactions that contain the item.
- Intersect the TID lists of two subsets to get the TID list for a candidate itemset.

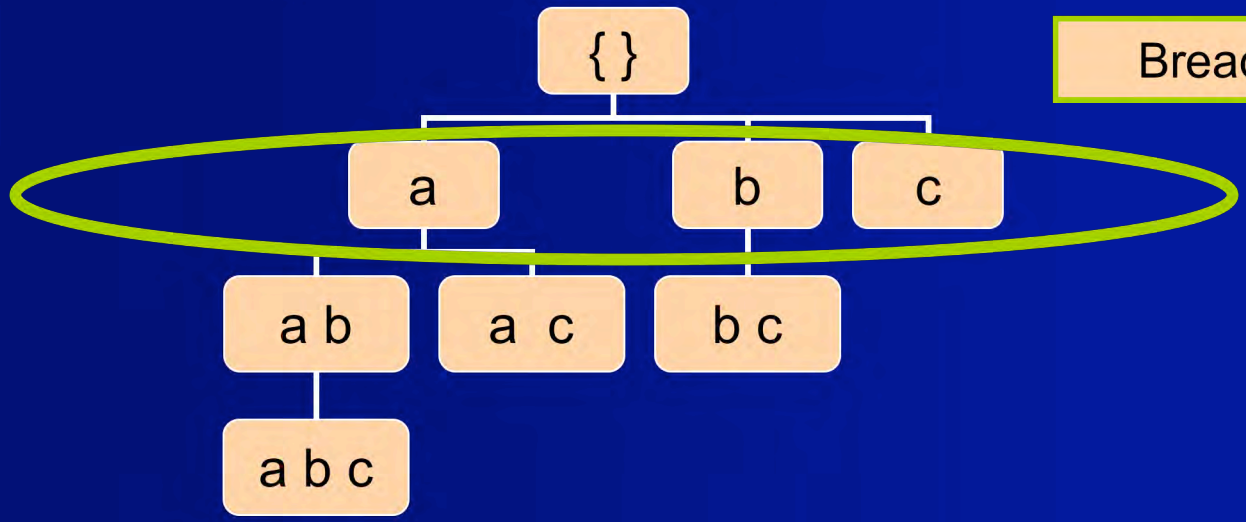
Cost of intersection is low if lists are in memory.

Item	TID List
a	10, 30
b	20, 40
e	10, 30, 40

Itemset	TID List
a, b	
a, e	10, 30
b, e	40

TID Intersection: Depth First Search

- Use depth-first search to keep TID lists in memory.

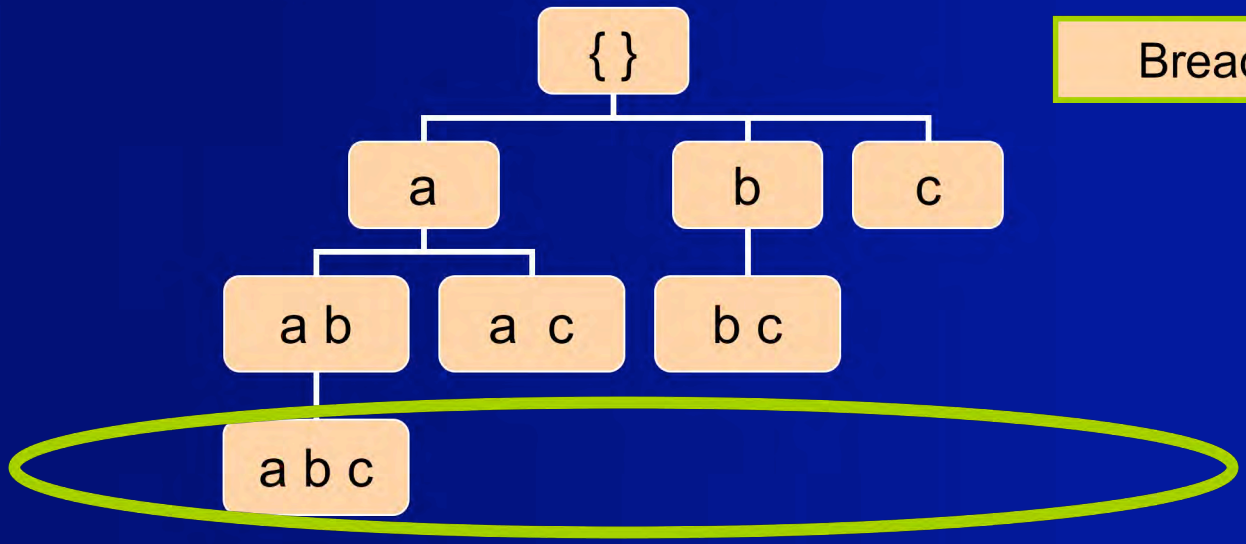


Breadth-first Search

M. Zaki et al., “New Algorithms for Fast Discovery of Association Rules", KDD ‘97.

TID Intersection: Depth First Search

- Use depth-first search to keep TID lists in memory.

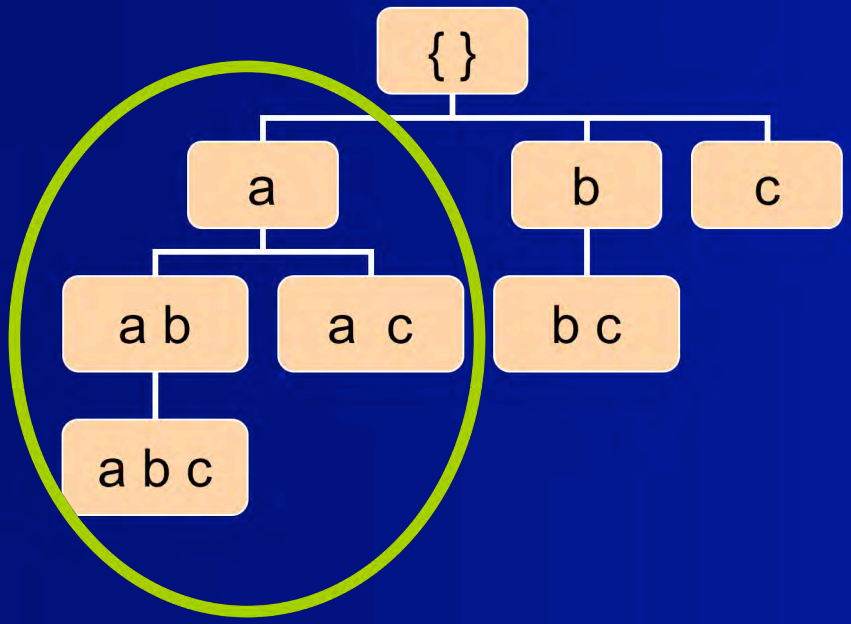


Breadth-first Search

M. Zaki et al., “New Algorithms for Fast Discovery of Association Rules", KDD ‘97.

TID Intersection: Depth First Search

- Use depth-first search to keep TID lists in memory.

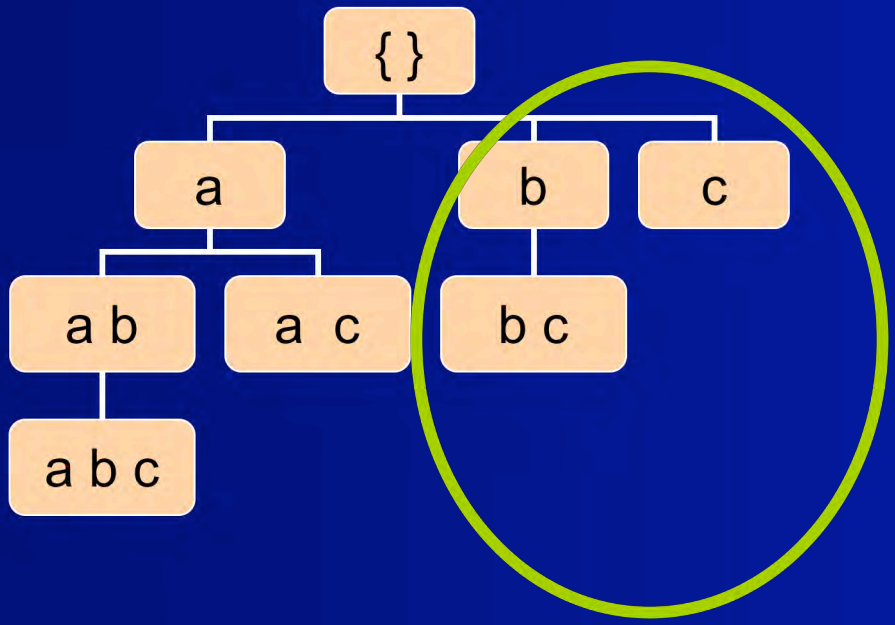


Depth-first Search

M. Zaki et al., “New Algorithms for Fast Discovery of Association Rules", KDD ‘97.

TID Intersection: Depth First Search

- Use depth-first search to keep TID lists in memory.



Depth-first Search

M. Zaki et al., “New Algorithms for Fast Discovery of Association Rules", KDD ‘97.

Database Projection

- Example: To count all itemsets starting with {a}, with possible extensions {b, e, f}:

Trans. ID (TID)	Items
10	a, c, e, f
20	b, d, f
30	a, e, f
40	b, e

R.C. Agarwal et al., "A Tree Projection Algorithm for Generation of Frequent Itemsets", JPDC.

Database Projection

- Example: To count all itemsets starting with {a}, with possible extensions {b, e, f}:

Trans. ID (TID)	Items
10	a, c, e, f
20	b, d, f
30	a, e, f
40	b, e

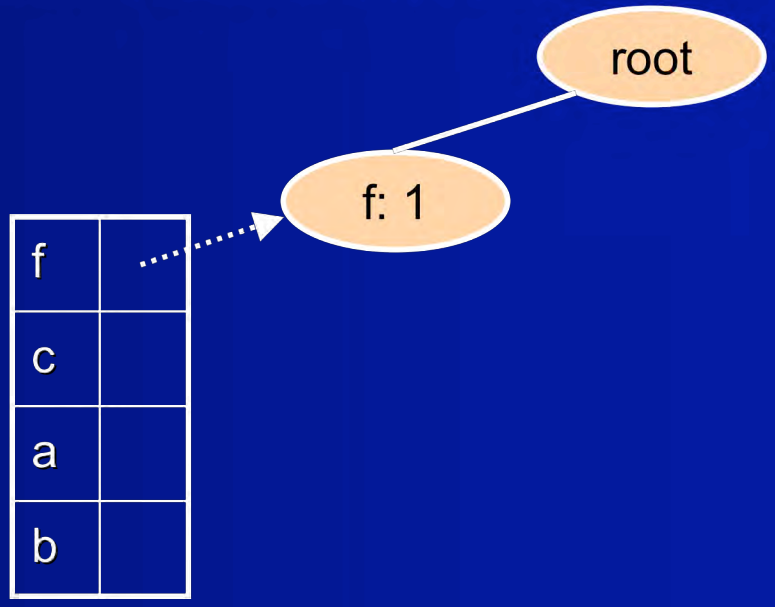
Trans. ID (TID)	Items
10	e, f
30	e, f

R.C. Agarwal et al., "A Tree Projection Algorithm for Generation of Frequent Itemsets", JPDC.

FP Growth

- Transaction database converted to prefix-tree structure.
  - Items ordered in frequency descending order.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

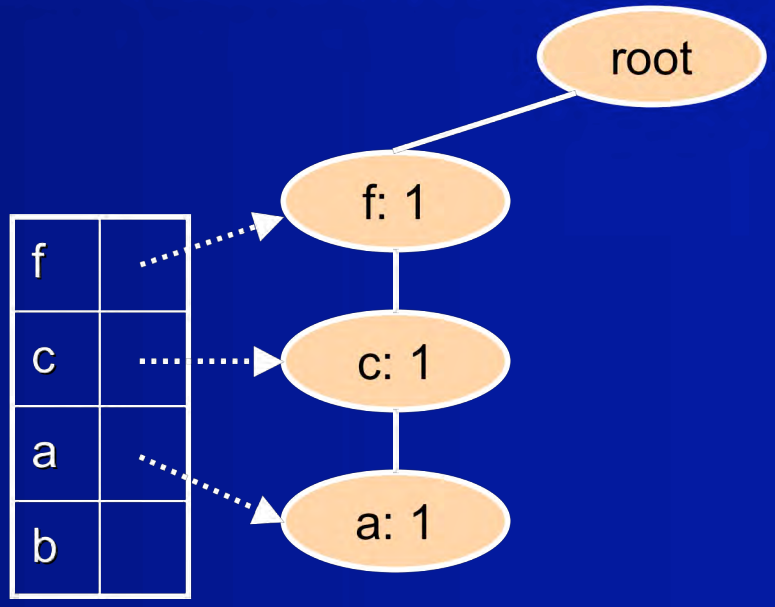


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.
  - Items ordered in frequency descending order.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

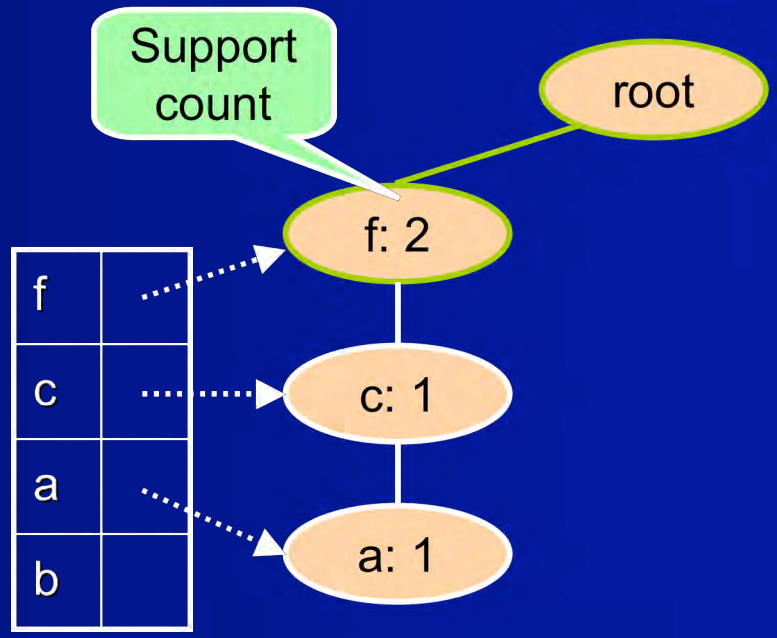


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

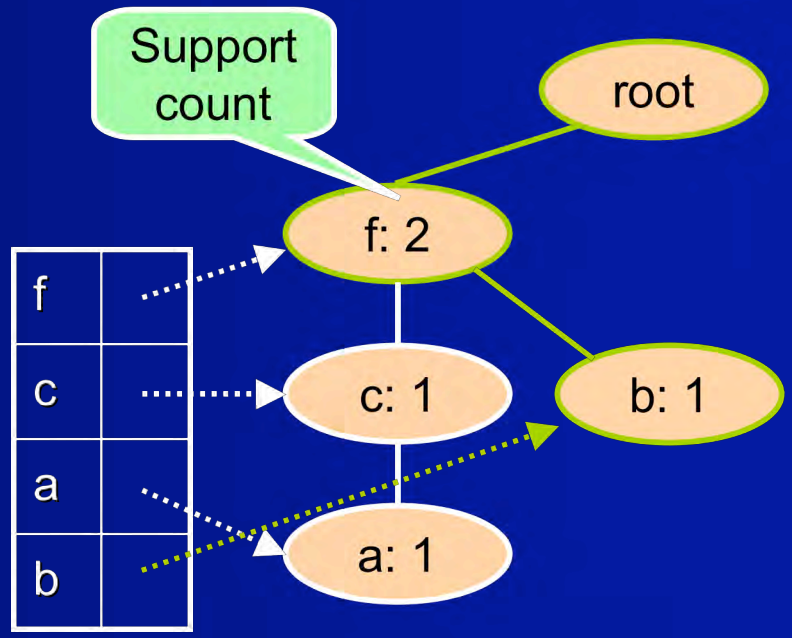


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

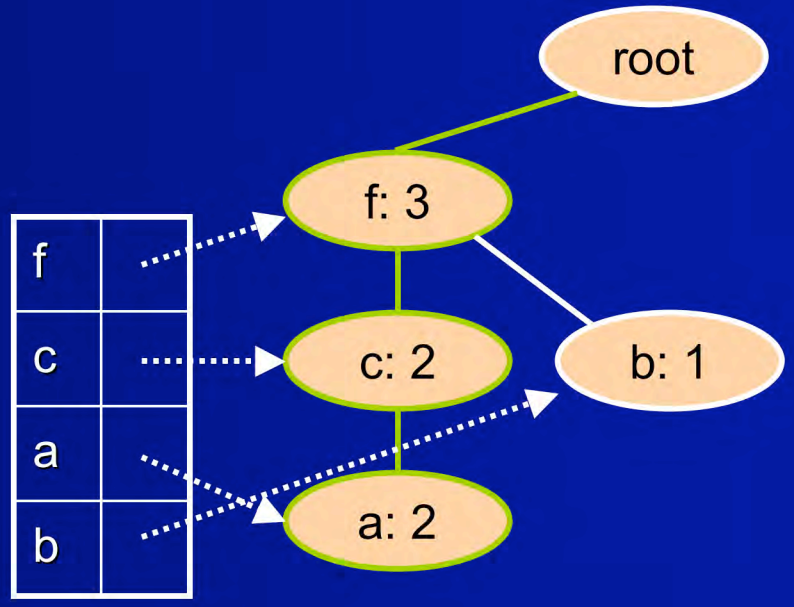


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

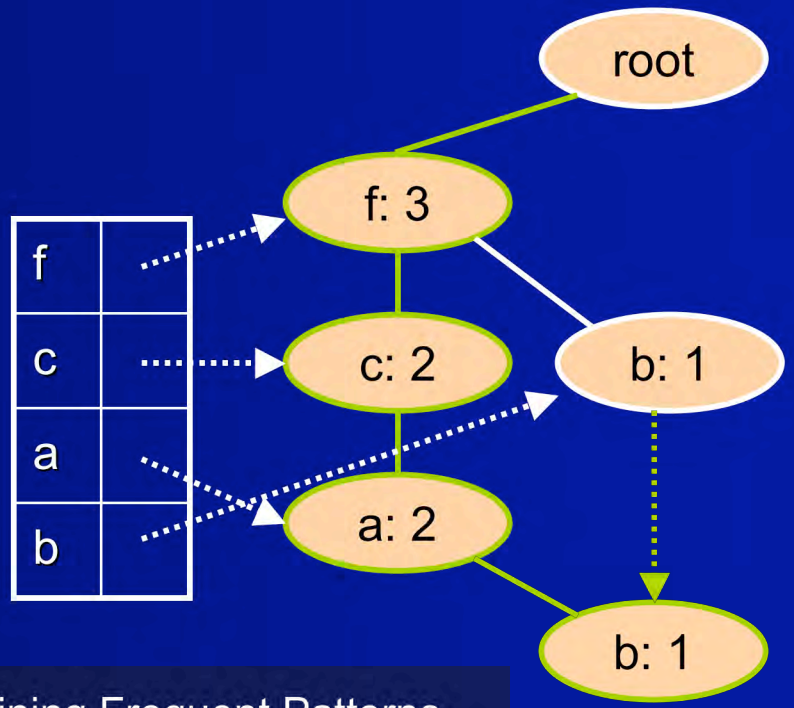


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b

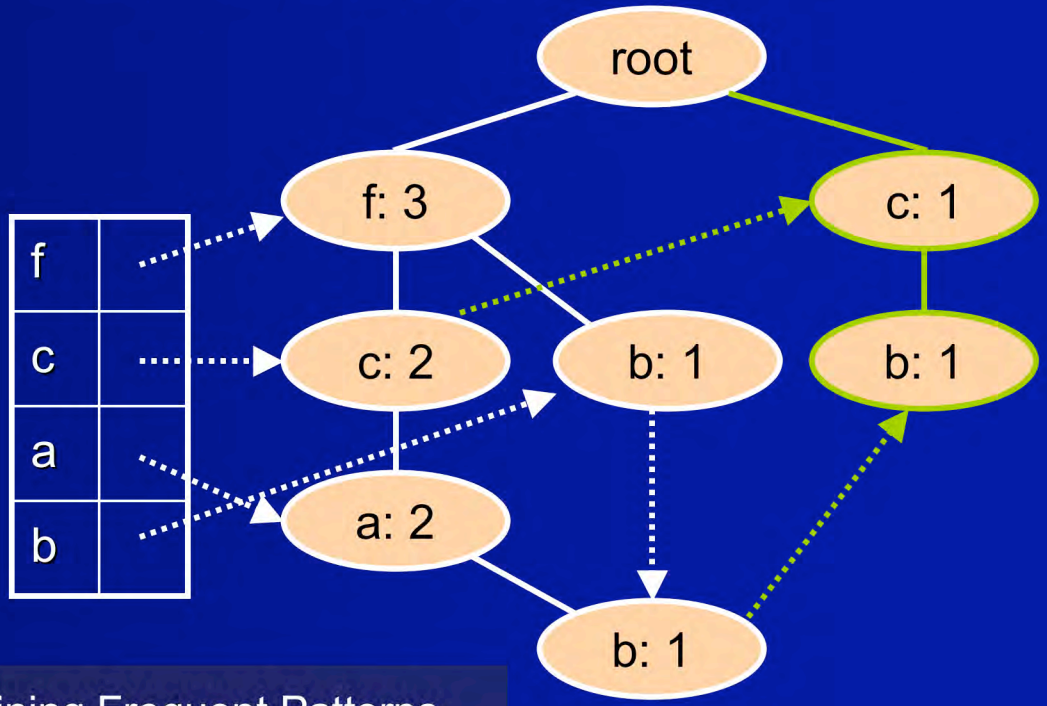


J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.
  - Tree typically much smaller than database.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b



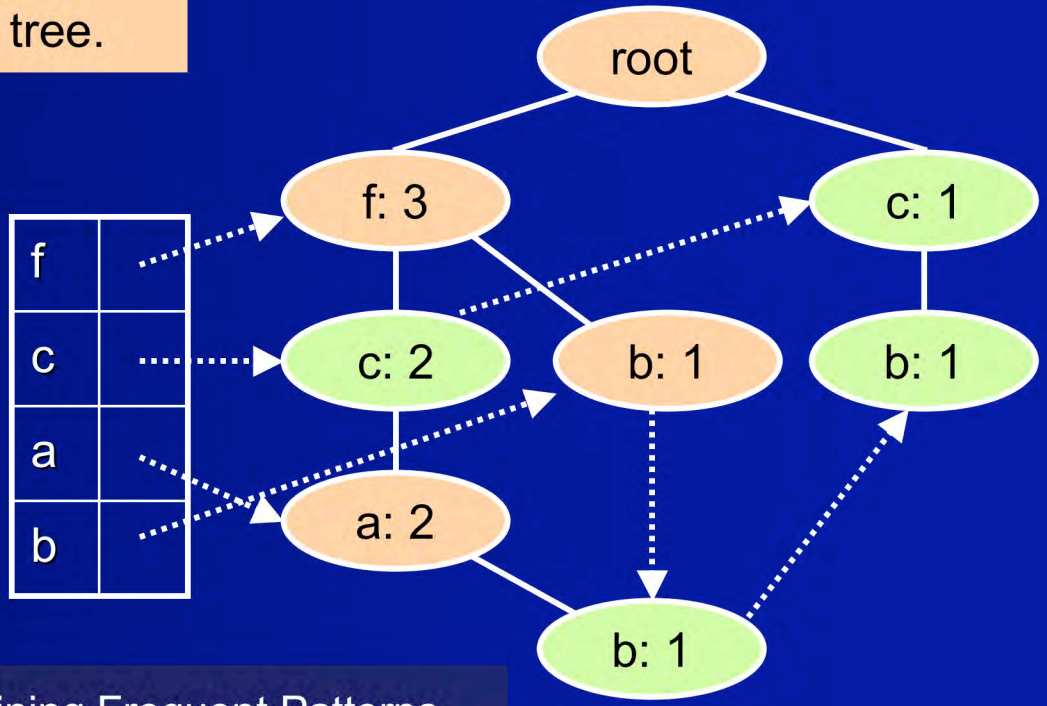
J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

FP Growth

- Transaction database converted to prefix-tree structure.

Can efficiently count support of candidates using prefix tree.

TID	Frequent Items
10	f, c, a
20	f, b
30	f , c, a, b
40	c, b



J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, SIGMOD 2000.

Algorithmic Innovations

- Reducing the cost of checking whether a candidate itemset is contained in a transaction:
  - TID intersection.
  - Database projection.
- Reducing the number of passes over the data:
  - Sampling & Dynamic Counting

Themes:

Focused on IO cost.

Sparse data / shorter rules.

Reducing the number of passes

- Use independence assumption to estimate which itemsets will be frequent, and count all the expected frequent itemsets in a single pass (SIGMOD ’93).
  - Eagerly combine multiple passes to reduce cost of scanning the database (VLDB ’94).
- Use sampling instead of independence assumption to estimate which itemsets are frequent, and count all frequent itemsets + negative border in one pass (Toivonen, VLDB ’96)
  - Sample yields more accurate estimates than independence assumption.
  - Need mop-up pass(es) for those itemsets in negative border that turn out to be frequent.

Reducing the number of passes: Dynamic Counting

- Can combine “sampling” with pass over the database.
  - Assuming transactions are independent!

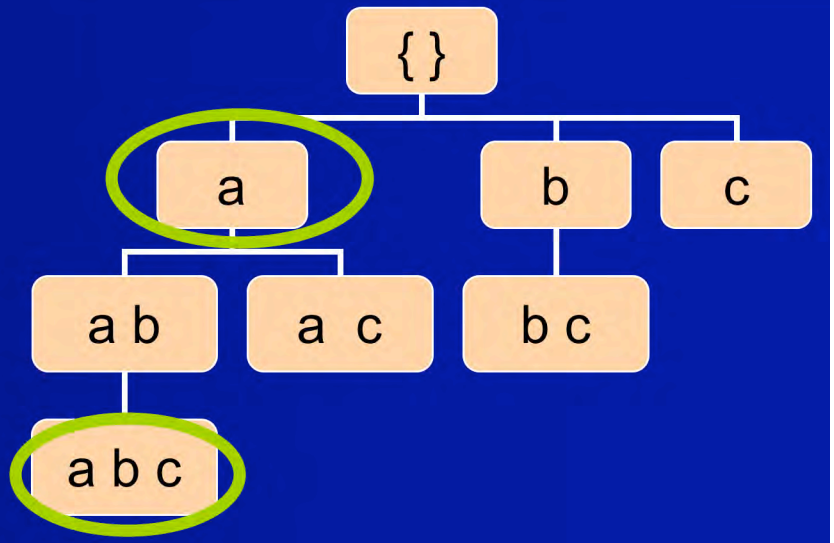
S. Brin et al, “Dynamic Itemset Counting and Implication Rules for Market Basket Data”, SIGMOD ‘97

Algorithmic Innovations

- Reducing the cost of checking whether a candidate itemset is contained in a transaction:
  - TID intersection.
  - Database projection.
- Reducing the number of passes over the data:
  - Sampling & Dynamic Counting
- Reducing the number of candidates counted:
  - For maximal patterns & constraints.

Maximal Itemsets

- MaxMiner: Uses pruning based on superset frequency in addition to pruning on subset infrequency.
- Set enumeration search.
- Lookahead: If {a b c} is frequent, don't need to count {a b} or {a c}.
- Scales roughly linearly in number of maximal sets.



R.J. Bayardo, "Efficiently Mining Long Patterns from Databases", SIGMOD '98.

Outline

- Association Rules & the Apriori Algorithm
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
- Whither?

Database Integration: Tight Coupling

- Motivation: Tighter integration for performance reasons.
  - Avoid copying of records from database address space to the application address space.
  - Avoid process context switching for each record retrieved.
- Push part of the computation into UDFs.
  - Two-fold performance advantage.

R. Agrawal and K. Shim, "Developing Tightly-Coupled Data Mining Applications on a Relational Database System", KDD '96

Database Integration: Using SQL

- Use SQL to perform computation.
- Motivation:
  - Composability: Combine selections and projections.
  - Use full power of database query optimization techniques.
- Need extensions to SQL ’92 and/or query optimization to get good performance.
  - Object-relational extension such as UDFs, Blobs, Table Functions allow for substantial performance improvement.

S. Sarawagi et al., “Integrating Association Rule Mining with Relational Database Systems: Alternatives and Implications”, SIGMOD ’98.

Database Integration: Pushing Apriori into Optimizers

- Flock: Parameterized query (e.g., pairs of items) and a filter condition (e.g., support > 20).
- Can generalize Apriori to apply to broader class of query flocks.
- Incorporating the generalized Apriori optimization into query optimizers can yield benefits for other data mining queries.

D. Tsur et al., "Query Flocks: A Generalization of Association Rule Mining", SIGMOD '98.

Standards: PMML

- Predictive Model Markup Language
  - Markup language for statistical and data mining models.
  - Includes association rules.
- Share models between PMML compliant applications.
- Allows users to mine association rules with one application, and use a different applications to visualize, analyze, evaluate or otherwise use the discovered rules.

PMML: Example

```
<AssociationModel functionName="associationRules"
      numberOfTransactions="4" numberOfItems="3“ minimumSupport="0.6"
      minimumConfidence="0.5" numberOfItemsets="3" numberOfRules="2">
…
<Item id="1" value="Cracker" />
…
<Itemset id="3" support="1.0" numberOfItems="2">
      <ItemRef itemRef="1" />
      <ItemRef itemRef="3" />
</Itemset>
…
<AssociationRule support="1.0" confidence="1.0"
                        antecedent="1" consequent="2" />
…
```

Outline

- Association Rules & the Apriori Algorithm
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
  - Algorithmic Innovations
  - System Issues
  - Usability
    - Interesting Rules
    - Visualization
- Interesting Applications
- Whither?

Statistical Significance

- Statistical Significance, e.g., p-value of independence test between antecedent and consequent.
- Need to adjust the threshold to account for number of hypotheses tested.
- Number of hypotheses tested can be >> number of rules.
  - Upper bound: number of candidates.
  - Upper bound too conservative since the hypotheses are not independent.
- Use resampling to determine appropriate threshold.
  - Generate synthetic dataset using item frequency information (assuming independence) and measure p-values of discovered associations.

N. Megiddo and R. Srikant, “Discovering Predictive Association Rules”, KDD ’98

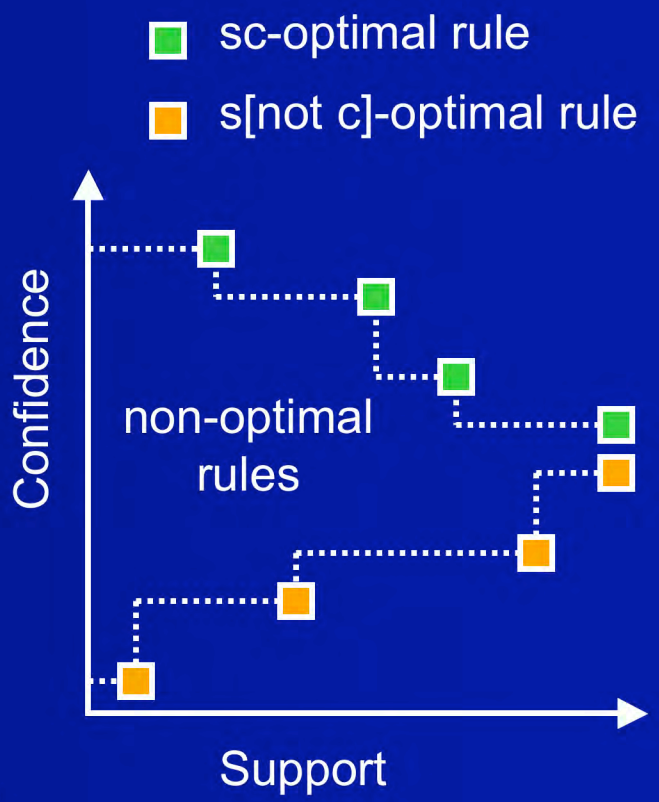
Objective Interest Measures

- Objective Measures: Depend only on the structure of pattern and the underlying data.
- Lift: Ratio of support to expected support assuming independence.
  - For rule A εB, Lift = P(A and B) / (P(A) * P(B))
- Other examples: chi-squared value, gini, conviction, etc.

G. Piatetsky-Shapiro, “Discovery, analysis and presentation of strong rules”, Knowledge Discovery in Databases, 1991.

Mining the Most Interesting Rules

- For fixed consequent, the best rule according to a broad class of interest metrics resides on a support/confidence border.
  - Example Metrics: support, confidence, gain, chi-squared, gini, entropy gain, laplace, lift, conviction.
- Can mine all rules that are best according to any of these criteria.



R.J. Bayardo and R. Agrawal, "Mining the Most Interesting Rules", SIGKDD '99.

Subjective Interest Measures

- Subjective Measures: Depend on the class of users who examine the pattern.
- Actionability: User can use the pattern to take some action.
- Unexpectedness: Surprising to the user.
  - Can be defined via belief systems.

Silberschatz and A. Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery", KDD '95.

B. Padmanabhan and A. Tuzhilin, "A Belief-Driven Method for Discovering Unexpected Patterns", KDD '98.

Interest Measures using Other Rules

- Jackets ε Shoes [70% confidence]
- Jackets and Shirts ε Shoes [71% confidence]
  - This rule not very surprising given the previous rule.
- Can take one step further & consider only the direction of the correlation (positive, negative or neutral).
  - Extension of a rule is interesting only if its direction is different.

B. Liu et al., “Pruning and Summarizing the Discovered Associations", KDD '99.

S. Sahar, “Interestingness via What Is Not Interesting”, KDD ’99.

Pruning based on confidence

- Fixed consequent
- Goal: Find all rules whose confidence is significantly higher than any of their simplifications.
- Can prune candidate itemsets based on confidence.
  - Although confidence does not map into any of the constraint classes, its constituent parts do.
  - Can be used prune search.

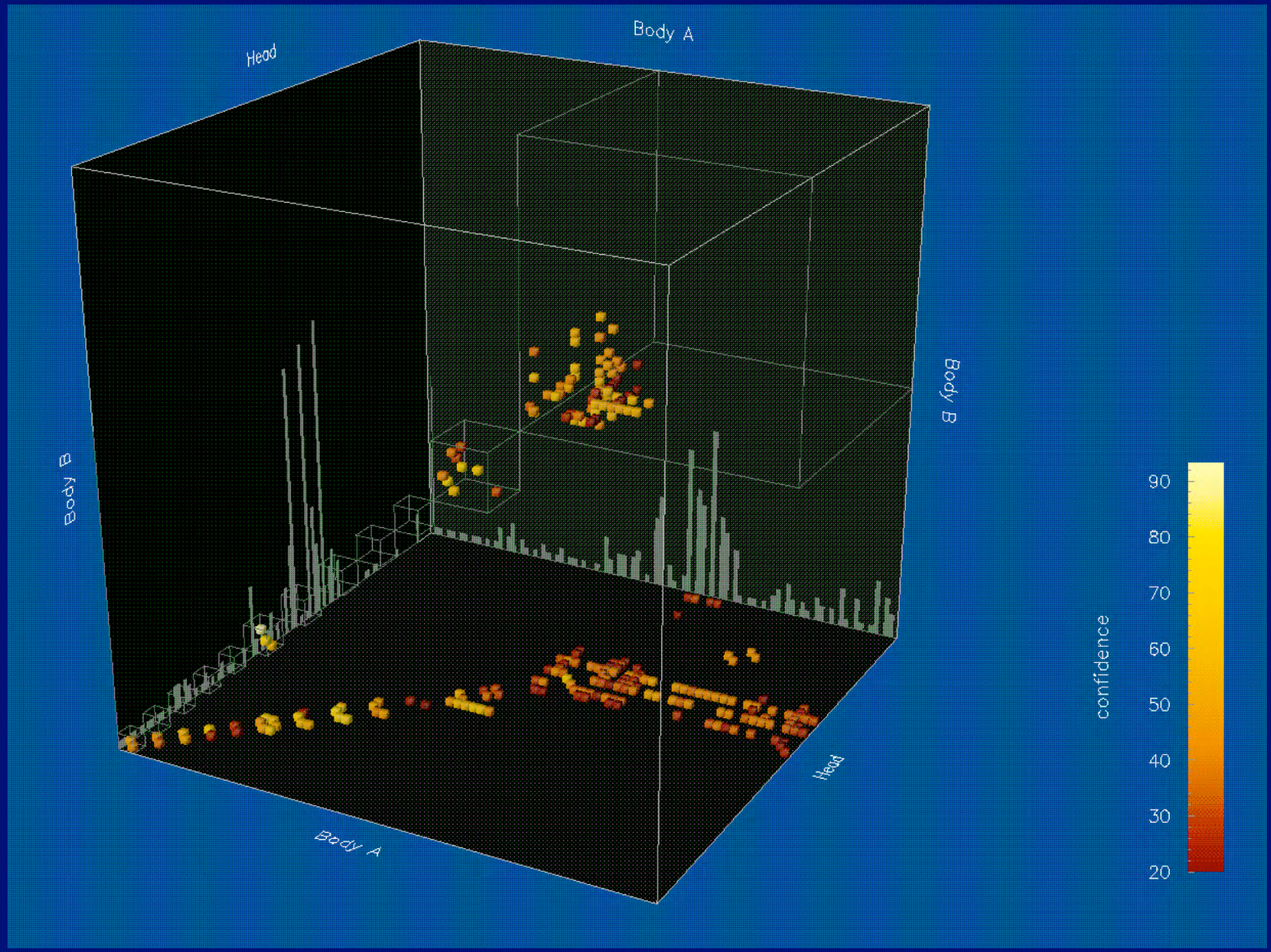
R. Bayardo, R. Agrawal and D. Gunopulos, “Constraint-Based Rule Mining in Large, Dense Series Databases”, ICDE '99.

Interest Measures using Other Rules: Taxonomies

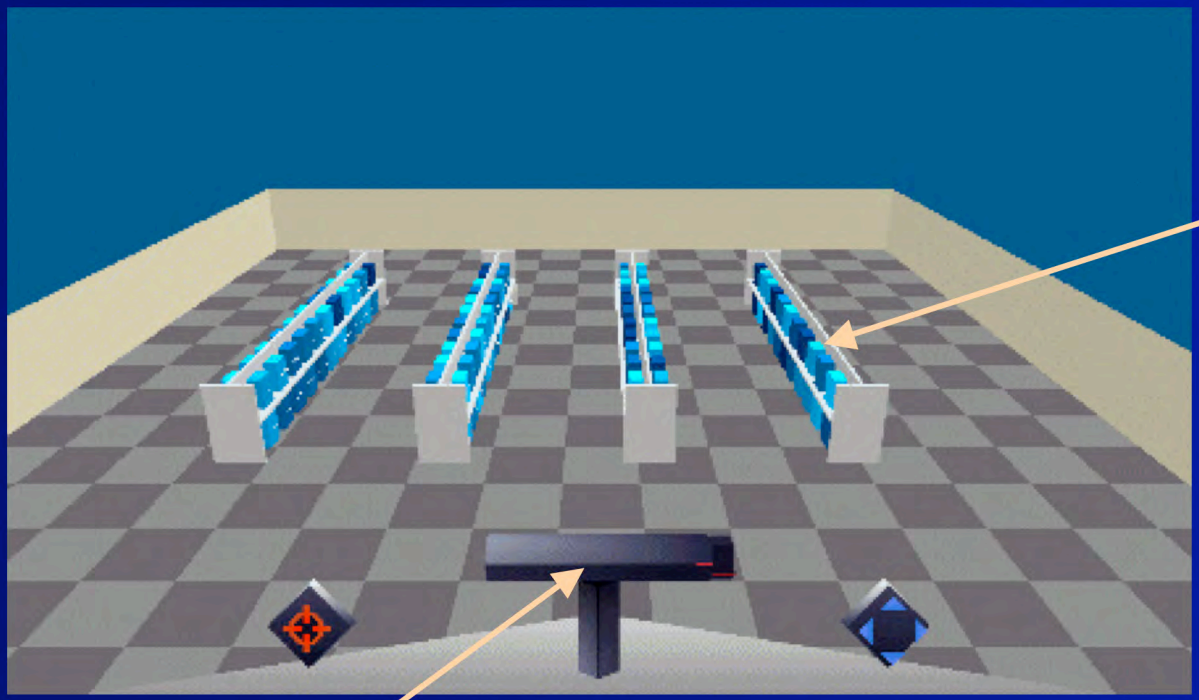
- Clothes ε Shoes
  - 8% support, 70% confidence
- Quarter of sales of Clothes are Jackets
- Jackets ε Shoes
  - expect 2% support, 70% confidence
- Interesting rule if support/confidence is significantly different than “expected” value.
- User-specified “interest level”
- Similar approach for quantitative associations.

R. Srikant and R. Agrawal, “Mining Generalized Association Rules”, VLDB '95.

Visualization Example



Application-Driven Visualization

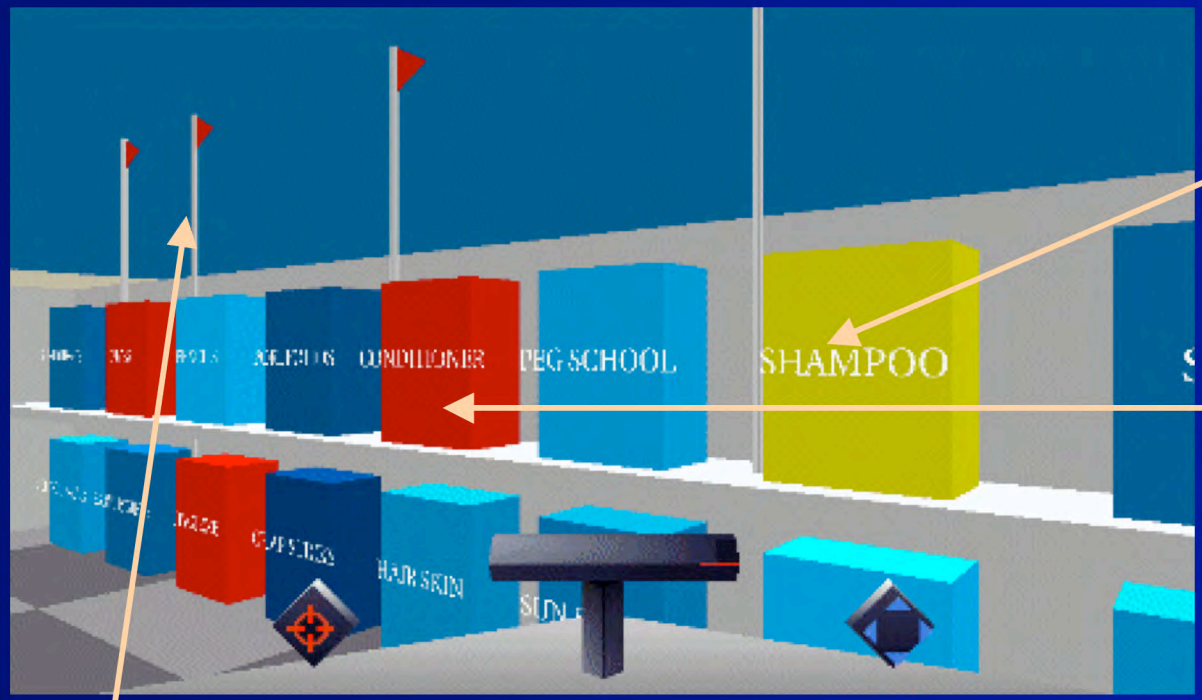


Color shows support.

Can “walk” around and examine placement of frequent items.

Can have multiple colors to show profit & frequency..

Application-Driven Visualization



Select an item.

Associated items in red.

Flags to see associated items in other aisles.

Outline

- Association Rules & the Apriori Algorithm
- Developments since VLDB 94
- Interesting Applications (Beyond Market Basket Analysis)
  - Early Applications
  - Building Block for Other Data Mining Operations
  - Other Innovative Applications
- Whither?

Product Assortment Decisions in Automated Convenience Stores

- Context: Automated Convenience Stores that can hold about 200 products.
- Placement decisions traditionally based on profit per product.
  - Ignored cross-selling effects.
- Use frequent itemsets to identify cross-selling effects.
  - Feed into a microeconomic optimization model.
- Example: Cigarette paper is not justified by its profit margins, but drives sales of tobacco (which has high profit margins).

T. Brijs et al., "Using Association Rules for Product Assortment Decisions: A Case Study", KDD '99

Health Insurance Claim Analysis

- Health Insurance Commission of Australia
  - Claim file over 550 GB
- Associations between medical payment codes
- Found that doctors were frequently claiming MCS2 with OCP, rather than claiming FCS with OCP.
  - $13.55 benefit for MCS2 versus OCP.
- However, this combination should be a rare event, not a common event.
- Potential savings of over $500,000 per year (in just one state).

M. Viveros, J.P. Nearhos and M.J. Rothman, “Applying Data Mining Techniques for Effective Health”, VLDB '96

Analyzing Automobile Warranty Claims

- Analyzed claims on DaimlerChrysler automobiles.
- Found associations "considered as very useful by domain experts":
  - Dependency between problems with justification of headlights and the distance of the axes.
  - 20% of vehicles visiting a specific garage had problems with the cables.
    - Problems with cables not uniformly distributed over garages.
    - Further investigation may identify the root problem.

J. Hipp and G. Lindner, "Analysing Warranty Claims of Automobiles", ICSC '99.

Reducing Telecommunications Failures

- Telecommunications service orders:
  - Around 3.5 sub-parts (USOCs).
    - Example: USOC for “call-waiting” in an order for new line.
  - Failure code (RMA) if automated processing failed.
    - RMA: Request for Manual Assistance.
- Classification problem: what combination USOCs predict an RMA?
  - But only 2.5% of orders fail, 25 different failure codes.
- Found frequent itemsets for each RMA,
  - Checked for statistically significant correlation between RMA & itemset.
  - Ranked itemsets by Pr(RMA | itemset) / Pr(RMA | no itemset).
- Results pointed to overhaul of certain mechanisms for ISDN orders.

K. Ali, S. Manganaris and R. Srikant, “Partial Classification using Association Rules”, KDD ’97.

Outline

- Association Rules & the Apriori Algorithm
  - A quick retrospection of VLDB 94 paper
- Developments since VLDB 94
- Interesting Applications
  - Early Applications
  - Building Block for Other Data Mining Operations
  - Other Innovative Applications
- Whither?

Classification

- Find all association rules where the consequent is a class label.
- Select a subset of rules and build a rule-based classifier.
- More accurate than C4.5

B. Liu, W. Hsu and Y. Ma, “Integrating Classification and Association Rule Mining”, KDD ’98.

W. Li, J. Han, and J. Pei, “Cmar: Accurate and efficient classification based on multiple class-association rules, ICDM ’01.

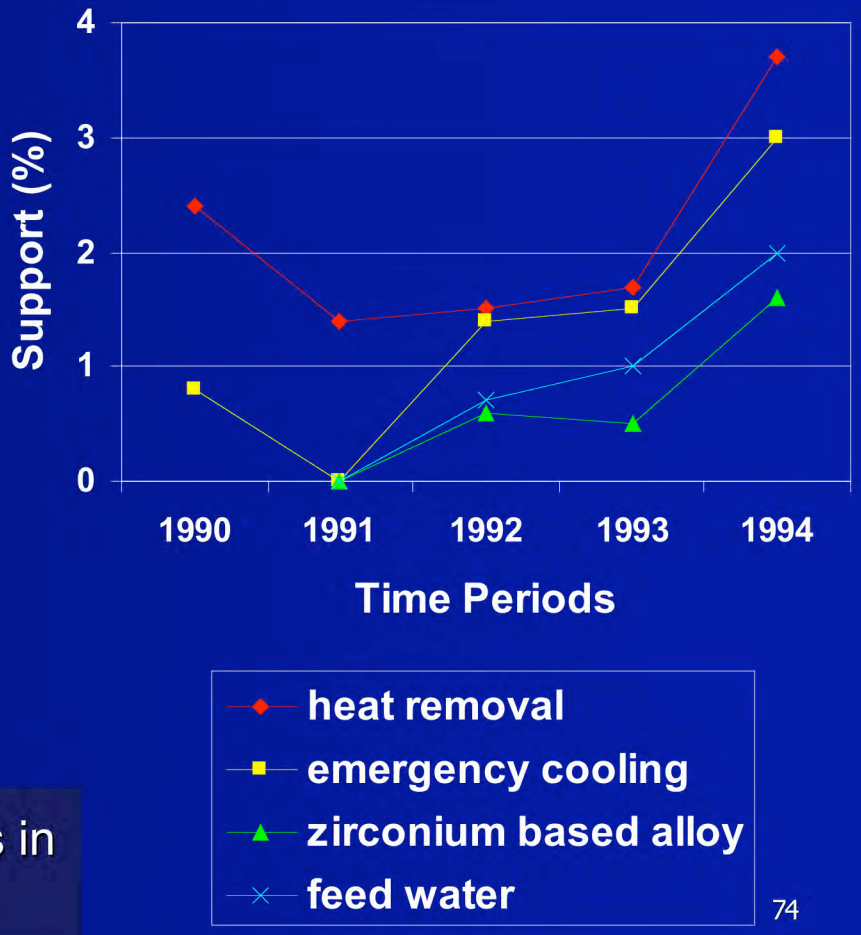
Subspace Clustering

- Problem: Find clusters embedded in subspaces of high dimensional data.
- Cluster: Maximal set of connected dense units in k-dimensions.
- Clusters are analogous to frequent itemsets.
- Anti-monotonicity: If a collection of points S is a cluster in a k-dimensional space, then S is also part of a cluster in any (k-1)-dimensional projection of this space.
- MDL used to identify interesting subspaces (and clusters).

R. Agrawal, J. Gehrke, D. Gunopulos and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", SIGMOD '98.

Discovering Technology Trends from Patent Databases

- Input: i) patent database ii) shape of interest.
- Find frequent patterns (e.g., sequence of phrases) in each time period.
- Use shape-based query language to identify trends over the change in support.



B. Lent et al, "Discovering Trends in Text Databases", KDD '97.

Outline

- Association Rules & the Apriori Algorithm
  - A quick retrospection of VLDB 94 paper
- Developments since VLDB 94
- Interesting Applications
  - Early Applications
  - Building Block for Other Data Mining Operations
  - Other Innovative Applications
- Whither?

Intrusion Detection

- Mine frequent patterns from user command data.
  - Merge similar patterns to form the normal usage profile.
- Given a new session, mine its patterns.
  - Compare similarity of patterns against normal profile.
- Successful at detecting anomalies in DARPA dataset.

W. Lee, S.J. Stolfo and K.W. Mok, “A Data Mining Framework for Building Intrusion Detection Models”, IEEE Symp. On Security and Privacy, 1999.

Identifying Social Links on the Web

- Input: Crawl of about 1 million pages.
- Find associations of what names occur together.

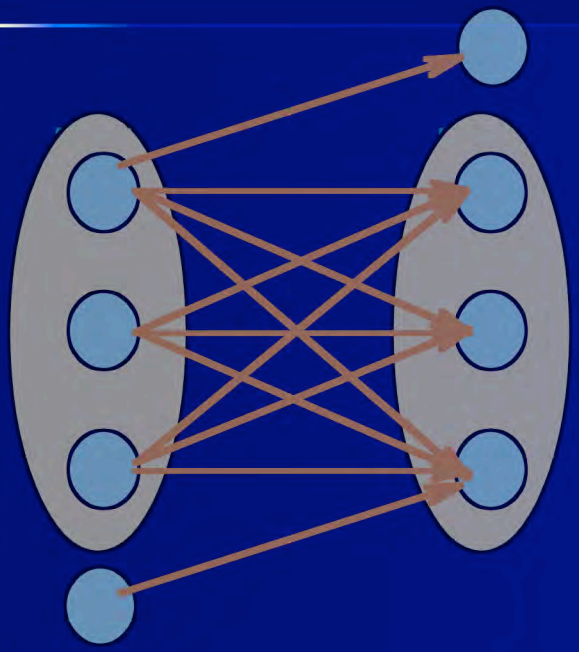
WebFountain Project,
IBM Almaden.

Template Detection for the Web

- Three important Hypertext IR principles:
  - Relevant Linkage: Google, Citeseer
  - Topical Unity: Co-cited documents are related.
  - Lexical Affinity: Proximity is correlated with relevance.
- Templates: Collection of pages that share the same look and feel, and are controlled by a single authority.
  - Templates violate these principles.
  - Detection instance of frequent itemset counting problem.
- Eliminating templates results in “surprising increases in precision at all levels of recall” in (pure) search algorithms.

Z. Bar-Yossef and S. Rajagopalan, “Template Detection via Data Mining and its Applications”, WWW 2002.

Discovering Micro-communities



complete 3-3 bipartite graph

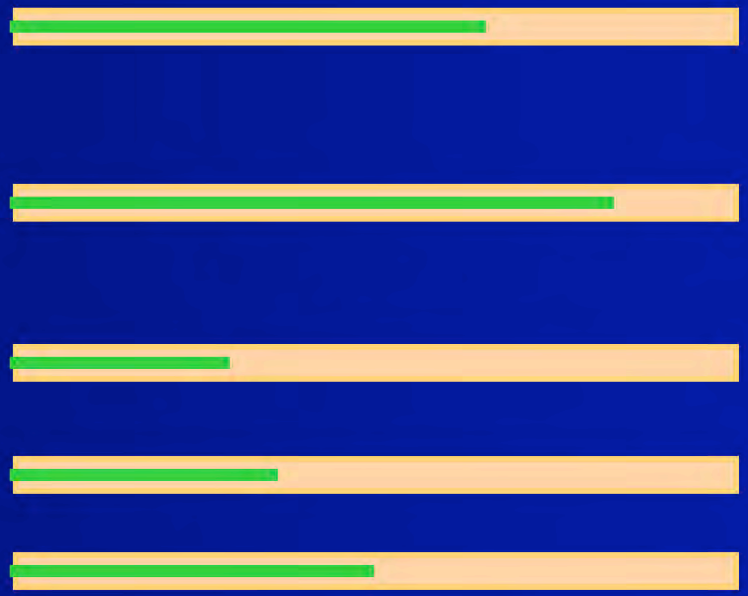
- Japanese elementary schools
- Turkish student associations
- Oil spills off the coast of Japan
- Australian fire brigades
- Aviation/aircraft vendors
- Guitar manufacturers

Frequently co-cited pages are related.
Pages with large bibliographic overlap are related.

R. Kumar et al., "Trawling the web for emerging cyber-communities", WWW '99.

Developments since VLDB '94: A Snapshot

- Extensions to Association Rules
- Algorithmic Innovations
- Database Integration
- Usability
- Published Applications



Speculation On Why Association Rules Got Attention Beyond Our Wildest Dreams

- Essence of formalism was small.
- Grounded in real, unfulfilled need.
  - Being used in the field and getting feedback, allowed us to identify the key generalizations.
- Emphasis on finding all rules rather than strong rules.
  - Could run against whole dataset.
  - Discovery vs. Hypothesis Testing (Completeness).

Speculation On Why Association Rules Got Attention Beyond Our Wildest Dreams

- Essence of formalism was small.
- Grounded in real, unfulfilled need.
  - Being used in the field and getting feedback, allowed us to identify the key generalizations.
- Emphasis on finding all rules rather than strong rules.
  - Could run against whole dataset.
  - Discovery vs. Hypothesis Testing (Completeness).
- Since the abstraction was simple, there was a lot of flexibility in extending the problem.
- While the formalism was simple, efficient solutions were not!
  - Scope for innovation.

Outline

- Association Rules & the Apriori Algorithm
  - A quick retrospection of VLDB 94 paper
- Developments since VLDB 94
  - Extensions to the Association Rules Formalism
  - Algorithmic Innovations
  - System Issues
  - Usability
- Interesting Applications
- Whither?

Whither Data Mining: The Dream (circa 1994)

- Data mining should aspire to become the technology that allows entities to derive maximal economic value from data.

Whither Data Mining: The Dream (circa 2004).

- Data mining should aspire to become the technology that allows entities to derive maximal economic value from data while respecting the sovereignty of data.

Sovereignty of Individual

- Goal: Build accurate data mining models while preserving the privacy of individual records.
- Initial Approach: Data Perturbation.
- Research Challenges:
  - How do you define privacy?
  - What is the tradeoff between privacy and accuracy?
  - Other approaches?

R. Agrawal and R. Srikant, SIGMOD 2000

A. Evfimievski et al., PODS 2003

Sovereignty of Individual

- Goal: Build accurate data mining models while preserving the privacy of individual records.
- Initial Approach: Data Perturbation.
- Research Challenges:
  - How do you define privacy?
  - What is the tradeoff between privacy and accuracy?
  - Other approaches?
- Business Imperative:
  - Competitive Advantage
  - Compliance with legislation.
- Moral Imperative: “Privacy is … the most comprehensive of rights, and the right most valued by civilized man.” -- Louis D. Brandeis

R. Agrawal and R. Srikant, SIGMOD 2000

A. Evfimievski et al., PODS 2003

Organizational Sovereignty

- The network era is forcing organizations to share information and make decisions across organizational boundaries.
- Goal: Data mining that transcends organizational boundaries while preserving sovereignty.
- Initial Approach: Cryptographic Protocols
- Research Challenges:
  - New framework for thinking about ownership, privacy and security.
  - What is the tradeoff between Generality, Performance, Accuracy, and Potential disclosure?

Y. Lindell & B. Pinkas, "Privacy Preserving Data Mining", Crypto 2000

Purdue Toolkit [Clifton et al 2003]

R. Agrawal, A. Evfimievski and R. Srikant, SIGMOD 2003

Organizational Sovereignty

Closing Thought
Tremendous opportunity for innovation and making fundamental contributions.

- Initial Approach: Cryptographic Protocols
- Research Challenges:
  - New framework for thinking about ownership, privacy and security.
  - What is the tradeoff between Generality, Performance, Accuracy, and Potential disclosure?

Acknowledgements

To all the researchers who worked on association rules and thus helped create the field:

Thank You!

Ack: The Quest Group (circa 1995)

Left to Right:
H. Miranda
K. Shim
R. Srikant
D. Lin
J. Shafer
R. Agrawal
M. Mehta
M. Zait

Acks: Our Colleagues in Germany

Left to Right:
L. Knoth-Weber
P. Vigants
U. Baumbach
C. Lingenfelder
N.P. Friedrich
B. Hartel
R. Keuler
W. Staub
A. Arning
T. Bollinger
H. Meiwes
S. Bayerl
P. Greissl

Acknowledgments

Arning	Arnold	Bayardo	Baur	Bollinger	Brodbeck
Baune	Carey	Chandra	Cody	Faloutsos	Gardner
Gehrke	Ghosh	Greissl	Gruhl	Grove	Gunopulos
Gupta	Haas	Ho	Imielinski	Iyer	Lent
Leyman	Lin	Lingenfelder	Mason	McPherson	Megiddo
Mehta	Miranda	Psaila	Raghavan	Rissanen	Sawhney
Sarawagi	Schwenkries	Schkolnick	Shafer	Shim	Somani
Srikant	Staub	Swami	Traiger	Vu	Zait