Absolute support sup (count) Relative support s The fraction of transactions that contains X (the probability that a transaction contains X) Association Rules (s,c)Support of X ∪ Y Ex. s{Diaper, Beer} = 3/5 = 0.6 (i.e., 60%) Confidence of X → Y The conditional probability that a transaction containing X also contains Y: c = sup(X, Y) / sup(X) Ex. c = sup{Diaper, Beer}/sup{Diaper} = ¾ = 0.75

Closed patterns: A pattern (itemset) X is closed if X is frequent, and there exists no super-pattern Y X sa troppus emas eht htiw ,X ⊃ lossless compression:Reduces the # of patterns but does not lose the support information Thus more desirable

Max-patterns: A pattern X is a max-pattern if X is frequent and there exists no frequent super-pattern Y ⊃ X lossy compression We only know a subset of the max-pattern is frequent But we do not know the real support any more.

Downward closure (Apriori): Any subset of a frequent itemset must be frequent.Apriori pruning principle: If there is any itemset which is infrequent, its superset should not even be generated! Outline of Apriori (level-wise, candidate generation and test) :[Scan] DB once to get frequent 1-itemset [Repeat] Generate length-(k+1) candidate itemsets from length-k frequent itemsets--Test the candidates against DB to find frequent (k+1)-itemsets--Set k := k +1--[Until] no frequent or candidate set can be generated [Return] all the frequent itemsets derived Partitioning: Scan Database Only Twice.Direct Hashing and Pruning (DHP)Hash TableExploring Vertical Data Format: ECLAT

An element may contain a set of items (also called events) Customer shopping Medical treatments Natural disastersScientific Experiments Stocks Markets Biological sequences, DNA /Protein

depth-first. Apriori: A breadth-first search mining algorithm

GSP (Generalized Sequential Patterns);SPADEVertical format-based mining.
PrefixSpan:Pattern-growth methods

Mining Multiple-Level Associations(Uniform support, Reduced support;Efficient mining: Shared multi-level mining;Use group-based "individualized" min-support) Mining Multi-Dimensional Associations(Inter-dimension,Hybrid-dimension)
Mining Quantitative Associations
Mining Negative Correlations (s(A U B)/s(A) + s(A U B)/s(B))/2 = (0.01 + 0.01)/2 < ϵ
Mining Compressed and Redundancy-Aware Patterns

Measure	Definition	Range	Null-Invariant?
$\chi^2(A,B)$	$\sum_{i,j}\frac{(e(a_i,b_j)-o(a_i,b_j))^2}{e(a_i,b_j)}$	$[0,\infty]$	No
$Lift(A,B)$	$\frac{s(A\cup B)}{s(A)\times s(B)}$	$[0,\infty]$	No
$Allconf(A,B)$	$\frac{s(A\cup B)}{max\{s(A),s(B)\}}$	$[0,1]$	Yes
$Jaccard(A,B)$	$\frac{s(A\cup B)}{s(A)+s(B)-s(A\cup B)}$	$[0,1]$	Yes
$Cosine(A,B)$	$\frac{s(A\cup B)}{\sqrt{s(A)\times s(B)}}$	$[0,1]$	Yes
$Kulczynski(A,B)$	$\frac{1}{2}\left(\frac{s(A\cup B)}{s(A)}+\frac{s(A\cup B)}{s(B)}\right)$	$[0,1]$	Yes
$MaxConf(A,B)$	$max\{\frac{s(A\cup B)}{s(A)},\frac{s(A\cup B)}{s(B)}\}$	$[0,1]$	Yes

Let
$$p=\frac{s(A\cup B)}{s(A)}=P(B|A)$$
$$q=\frac{s(A\cup B)}{s(B)}=P(A|B)$$
p, q are null invariant

- Pattern distance measure $Dist(P_1,P_2)=1-\frac{|T(P_1)\cap T(P_2)|}{|T(P_1)\cup T(P_2)|}$
- δ-clustering: For each pattern P, find all patterns which can be expressed by P and whose distance to P is within δ (δ-cover)

IR (Imbalance Ratio): measure the imbalance of two itemsets A and B in rule implications:

Neutral (Ku=0.5) $IR(A,B)=\frac{|s(A)-s(B)|}{s(A)+s(B)-s(A\cup B)}$

Kulczynski and Imbalance Ratio (IR) together present a clear picture for all the three datasets D_4 through D_6

erate length-2 candidate sequences

singleton * singleton – *Total:* (6 * 6)

	<a>	<b>	<c>	<d>	<e>	<f>
<a>	<aa>	<ab>	<ac>	<ad>	<ae>	<af>
<b>	<ba>	<bb>	<bc>	<bd>	<be>	<bf>
<c>	<ca>	<cb>	<cc>	<cd>	<ce>	<cf>
<d>	<da>	<db>	<dc>	<dd>	<de>	<df>
<e>	<ea>	<eb>	<ec>	<ed>	<ee>	<ef>
<f>	<fa>	<fb>	<fc>	<fd>	<fe>	<ff>

Sets (unordered) – *Total*: (6*5) / 2

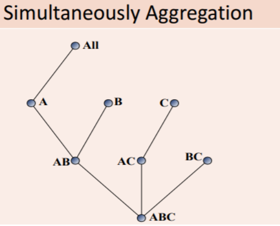
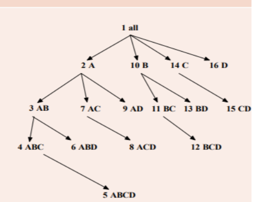
	<a>	<b>	<c>	<d>	<e>	<f>
<a>		<(ab)>	<(ac)>	<(ad)>	<(ae)>	<(af)>
<b>			<(bc)>	<(bd)>	<(be)>	<(bf)>
<c>				<(cd)>	<(ce)>	<(cf)>
<d>					<(de)>	<(df)>

Apriori Pruning

- w/o pruning *(includes g and h*
  8*8 + 8*7/2 = 92 length-2 candidate
- w/ pruning:
  6*6 + 6*5/2 = 51 length-2 candidate

Why Iceberg Cube:1.No need to save nor show those cells whose value is below the threshold (iceberg condition) 2.Efficient methods may even avoid computing the un-needed,intermediate cells; 3.Avoid explosive growth Data Cube: A Lattice of Cuboid

Closed cube: A cell c is closed if there exists no cell d, such that d is a descendant of c, and d has the same measure value as c A closed cube is a cube consisting of only closed cells CubeShell: The cuboids involving only a small # of dimensions, e.g.,2 Idea: Only compute cube shells, other dimension combinations can be computed on the fly

	multiway	BUC
Input format	Multi-dimensional array	Relational database
Good for	Full cube	Iceberg cube
Key idea	Simultaneously Aggregation	Partition and sort
Calculation direction		

Pattern space pruning constraints	Data space pruning constraints
Anti-monotonic: If constraint *c* is violated, its further mining can be terminated	Data succinct: Data space can be pruned at the initial pattern mining process
Monotonic: If *c* is satisfied, no need to check *c* again	Data anti-monotonic: If a transaction *t* does not satisfy *c*, then *t* can be pruned to reduce data processing effort
Convertible: *c* can be converted to monotonic or anti-monotonic if items can be properly ordered in processing	
Succinct: If the constraint *c* can be enforced by directly manipulating the data	

Semi-Online Computational Model Use Frag-Shells for Online OLAP Query Computation
Given a database of T tuples, D dimensions, and F shell fragment size, the fragment cubes' space requirement is: $O\left(T\left\lceil\frac{D}{F}\right\rceil(2^F-1)\right)$

Data Mining in Cube Space Reports generated from a Data Cube can easily by drilled into through query in a drilldown fashion.

0D (Apex) cuboid be pre-calculated (Materialization)

$$T=\prod_{i=1}^{n}(L_i\boxed{+1})$$

Nominal	categories, states, or "names of things"	• Hair_color = {auburn, black, blond, brown, grey, red} • marital status, occupation, ID numbers, zip codes
Binary (0 or 1)	Symmetric: equally important	gender
	Asymmetric: not equally important	Medical test (negative & positive); assign 1 to most important outcome
Ordinal	Need order but no magnitude	Size = {small, medium, large}, grades, army rankings
Numeric	Interval: • equal-sized units; • ordered; • no true zero-point;	temperature in C˚or F˚, calendar dates
	Ratio: inherent zero-point; being an order of magnitude larger than the unit of measurement	temperature in Kelvin, length, counts, monetary quantities

Pixel-oriented:(in Circle Segments)Direct Data Visualization

Inter-quartile range: IQR = Q3 (75th percentile) − Q1 (25th percentile)

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}}\right) width$$

Types of data sets:Record Data, Graphs and Networks, Ordered Data,Spatial,Image and multimedia Data

Geometric projection: Scatterplot Matrices, Landscapes, Parallel Coordinates? Icon-based Chernoff Faces,Stick Figures Hierarchical: Dimensional Stacking,Worlds-within-Worlds,Tree-Map,Cone Trees, InfoCube

Variance: (algebraic, scalable computation)

Q: Can you compute it incrementally and efficiently?

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

sample

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Population

Note: The formulae f • n : the • N : the

Standard deviation *s (or σ)* is the square root of variance s^2 (or σ^2)

Minkowski	$\left(\sum_{l=1}^{n} \lvert x_{il} - x_{jl}\rvert^p\right)^{1/p}$	$0 \to \infty$	• Pros: • Most commonly used distance for numerical data • Positivity/Symmetry/Triangle Inequality
Manhattan	$Minkowski, p = 1$ $\sum_{l=1}^{n} \lvert x_{il} - y_{jl}\rvert$	$0 \to \infty$	• Pros: • Not sensitive to outliers. • Cons: • Non differentiable
Euclidean	$Minkowski, p = 2$ $\left(\sum_{l=1}^{n} \lvert x_{il} - x_{jl}\rvert^2\right)^{1/2}$	$0 \to \infty$	• Pros: • differentiable • Cons • Sensitive to outliers
Supremum	$Minkowski, p \to \infty$ $\max_{f=1}^{l} \lvert x_{if} - x_{jf}\rvert$	$0 \to \infty$	
Symmetric Binary Variable	$\frac{r+s}{q+r+s+t}$	[0, 1]	• Null variant • if 0 and 1 are equally in
Asymmetric Binary Variable	$\frac{r+s}{q+r+s}$	[0, 1]	• Null invariant • If 0 is not important (such as meaning did not appear, too common in data, …)
Jaccard Coefficient / Coherence	$\frac{q}{(q+r)+(q+s)-q}$	[0, 1]	• This is a similarity measure • The higher the value, the more similar the two vector

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Proximity Measure

Ordinal Variables

Cosine Similarity	$cos(d_1, d_2) = \frac{d_1 \bullet d_2}{\lVert d_1 \rVert \times \lVert d_2 \rVert}$	$[-1, 1]$ In many applications, d_i are all positive, then [0, 1]	Commonly used in text mining 1-> similar 0-> irrelevant -1-> opposite
Chi-Squared Test	$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$	$[0, +\infty]$	Correlation measure for categorical da Higher value->strong correlation
Variance / Covariance	$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)] = E[X_1X_2] - \mu_1\mu_2 = E[X_1X_2] - E[X_1]E[X_2]$	$[-\infty, +\infty]$	Correlation measure for continuous da High positive value->strong positive correlation Very negative value->strong negative correlation
Correlation coefficient	$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2\sigma_2^2}}$	$[-1, 1]$	Correlation measure for continuous da High positive value->strong positive correlation Very negative value->strong negative correlation

$$\sigma^2 = \mathrm{var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Covariance Matrix

Object *j*

i	1	0	sum
1	q	r	$q+r$
0	s	t	$s+t$
sum	$q+s$	$r+t$	p

Data Compression : Lossless vs Loss

Dimensionality reduction:Feature selection and feature extraction;PCA; attribute subset selection (heuristic search); attribute creation

Z-Score −∞, +∞ But scores outside −3, 3 are likely to be outliers • Pros: • Easy to calculate • Good for outlier detection • Cons: • Small data sets skew the results $z = \frac{x - \mu}{\sigma}$ Mean Absolute Deviance $[0, +\infty]$ $\frac{\sum_{i=1}^{n} \lvert x_i - \bar{x}\rvert}{n}$

Numerosity Reduction: Parametric (Regression) Non-Parametric: Histogram, Clustering , Sampling

Min/Max Normalization • Pros: • Allows for custom range of data

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

$nw\_min_A \to nw\_max_A$

Unsupervised / Top-down Split:Binning,Histogram analysis,Clustering analy
Unsupervised / bottom-up merge:Clustering analysis
Supervised / top-down split:Decision-tree analysis
Data Discretization

Data Warehouse: Long time horizon (e.g., past 5-10 years) ,Contains an element of time, explicitly or implicitly. Operational Database: current value data; data may or may not contain "time element"

Three Data Warehouse Models:[Enterprise warehouse] - Specially designed for the entire organization; [Data Mart]: Specific, selected groups,Independent vs. dependent (directly from warehouse) data mart; [Virtual warehouse]A set of views over operational databases Only some of the possible summary views may be materialized

Conceptual Modeling:[Star Schema]A fact table in the middle connected to a set of dimension tables[Snowflake Schema] A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables [Fact Constellation], Multiple fact tables share dimension tables

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Lossy:smart photo is lossy compression Lossless: Zip is lossy compression

Common Warehouse Index: Bitmap Index

OLAP Operation: Roll up, Drill down, Dice, Slice, Pivot (rotate)(reorient the cube, visualization, 3D to series of 2D planes), Drill across(involving (across) more than one fact table), Drill through(through the bottom level of the cube to its back-end relational tables (using SQL))

Data Cube Measures: [Distributive]: count(), sum(), min(), max [Algebraic]: avg(), standard_deviation() .[Holistic] median(), mode(), rank(

Server Architectures: 【Relational OLAP (ROLAP) 】Data is stored in a relational database. Greater scalability 【Multidimensional OLAP (MOLAP)】Everything is in multi-dimensional storage (see page 26 for an example)Fast indexing to pre-computed summarized data 【Hybrid OLAP (HOLAP)】Used by : Microsoft SQLServer Combines both ROLAP & MOLAP. Theoretically provides best performance

Extraction, Transformation, and Loading (ETL):Data extraction, Data cleaning, Data transformation, Load, Refresh