**Chp 2: 5 Attribute types :Qual 1) Nominal(categoricl)** symb/names/nos. No meaningful order. Values called enumerations. Mode imp**. 2)** **Bin/Boolean:** only 2 states. Sym(gender)/asymm(HIV). **3) Ordinal:**ranking of val, magnitude unknown.(drink\_size).Mode-median imp.

**Quant: 4)Numeric:Interval-scaled:**measured on scale of equal size units. +,0,-. Eg temp(C/F). No true zero-point. **Ratio-scaled:** inherent zero-point. Temp(K), wt,ht. 5) **Discr/Cont:** discrete has finite or countably infinite vals. Eg. Cust\_id. Hair\_color.

**Mean** is sensitive to extreme values. **Trimmed mean**-loss of info. **Median** for even no of ordinal values is the 2 middle values and any in between. **Midrange** is avg of max and min. + skew-mode occurs before median. **Data Dispersion:** Range is max-min. Quantiles: points at regular intervals which split the distribution into equal size consecutive sets. There are q-1 q quantiles. **Outliers 1.5\*IQR from Q1 and Q3. IQR=Q3-Q1**. Five no summary-endpoints,Q1,Q2,Q3. Boxplots O(nlogn).

**Data Similarity and Dissimilarity: Data matrix**(two-mode) n objects by p att. **Dissim matrix (**one mode) d(I,j)=dissim betn I j. Non –ve no.

Normalizing data gives equal weight to all att. **Propeties of distance measures: non-negative, identity of discernible (d(I,j)=0, symettric, triangle ineq.** A measure that satisfies these conditions is called a metric. **Ordinal measures:** Replace the att value xif by its rank. Normalize so that all att lie betn 0 to 1. Zif=(rif-1)(Mf-1). Mf is total no of states of att. Use any numeric dissim measure to calc dist. **Cosine sim:** docs represented by term freq vectors. Sparse.

Bar chart for nominal, histogram for numeric. **Bivariate**: Scatter. **Data Visualization: 1) Pixel oriented:** data record in circle-segment, laying out pixels in circle-segments. 2)**Geometric:** 2,3,4-d scatter plot. Scatter plot matrix (n by n), Parallel coordinates(disadv clutter). 3)**Icon-based:** Chernoff faces: display multi-dim data upto 18 dims. Dims mapped to facial charac such as eye size, eye spacing, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, mouth size, and mouth opening. **Disadv**: exact values not shown. Asymm: upto 36 dims.**Stick fig**: A 5-piece stick figure (1 body and 4 limbs w. different angle/length). 2 dims on x-y, remaining limbs. 4) **Hierarchical:** Partition dims into subspaces, visualize hierarchical. **Worlds-within-worlds (n-vision):** Fix non-imp dims as origin in inner world, 3-d plot of 3 imp dims with origin. Outer world also 3-d plot (eg of 6 dims). **Treemap:** hierarchical data as nsted rectangles.**Dimensional stacking, Info cube, 3-d cone trees. Visualize complex data:** Tag cloud-single item-multiple users tag size, multiple items-size tag determines popularity of items.

**Chp 3:** **Data Quality:** accuracy, completeness, consistency, timeliness, believability, interpretability.

**A) Data Cleaning:** fill in missing values, smooth noisy data, identify/remove outliers, resolve inconsistencies.

**1) Methods for missing value:** ignore tuple, manual, fill global constant, enter mean, attribute mean, most probable value.

**2) Noisy data:** Binning methods consider the neighbourhood of values, called **local smoothing.** Smoothing by bin means, boundaries. **Regression:** Linear regression: best line fit to 2 attributes so that one can predict the other. **Outlier analysis.**

**3) Data cleaning as a processs: a) Discrepancy detection: identify** datatype, domain, acceptable values, statistical descriptions.

**Examine for rules: Unique:** each attribute value should be unique.

**Consecutive:** no missing value betn highest and lowest, and unique

**Null:** special characters for null, how to handle. **Data Scrubbing and Data Auditing tools. B) Data Transformation**

**B) Data Integration:** **1) Entity Identification Problem:** Schema integrn and object matching are tricky. E.g customer\_id/cust\_id

Dependencies and constrains in src and tgt should match.

**2) Redundancy and correlation analysis:** Nominal data: **chi-sqr test/Pearson x2 statistic:** Numeric data: **Pearson coeff (-1 to 1).**

Variance is a covariance of an attribute with itself.

**3)Tuple Duplication:** cause may be denormalized tables.

**4) Data value conflict detection and resolution:** different units, currencies, abstraction levels.

**C) Data Reduction;** reduced representation of the data set, which produces almost same analytical results.

**1)Dim reduction:** reduce no of attribs under consideration.

**a) Wavelet transform:** X to X’ with wavelet coeff (same length). Transformed data can be truncated. Store strongest coef(>thresh). Others set to 0. Approximate original dataset by inverse DWT.

**b) PCA:** normalize, k orthogonal vectors (linear combo), sort in decreasing order of significance (variance), eliminate weak ones.

**c)Attribute subset selection:** remove redundant/irrelevant dims.

Stepwise fwd selecn, backward elimn, combination, dec tree induc

**2)Numerosity reducn:** replace original data by alternative represnt

**a) Parametric:** Linear Regression, Log linear models

**b) Non-parametric:** Histograms, Clustering, Sampling (SRSWOR, SRSWR, cluster sample (pages), stratified)

**D) Data Transformation:** Smoothing (Binning, clustering, regression), Attribute construction, Aggregation, Normalization, Discretization, Concept hierarchy generation

**Discretization:** supervises/unsupervised,top-down/bottom-up

(Binning, Histogram, Clustering

**Normalization:** min-ax, z-score, decimal scaling

**Concept Hierarchy generation for nominal data:** (specify partial ordering, specify a portion of hierarchy, specify attributes but not partial ordering, specify only partial set of attribs)

**Chp 4:** A DW is a subject-oriened, integrated, time-variant and non-volatile collection of data in support of management’s decision making process.

**OLTP vs OLAP: (**used for, user type, user number, function, DB design, current data, summary level, type of transacn, num of records, db size, metric(transacn vs query thruput)

**3 models:** EW, Data Mart, Virtual warehouse

**A data cube measure** is a numeric function that can be evaluated at each point in thr data cube space.

**Distributive:** can be computed by partitioning data into subsets, applying the function to them, and then merging the results (max, min, count, sum)

**Algebraic:** can be computed by an algebraic func with M args, where each is obtained by a dist. Agg func. (sd, min\_N, max\_N)

**Holistic:** Must be computed on the entire data set as a whole (median, mode, rank)

**Index OLAP data:** bitmap and join indexing

**Attribute Oriented Induction: (Concept description)** query oriented, online data analysis

**1) Data Focusing:** Get task relevant data (initial working relation)

**2) Data Generalization: a) Attribute removal b)Attrib Generalizn**

InitialRel, PreGen, PrimeGen, Presentation. Associate count

**-Attribute generalizn threshold control:** no of distinct values an attrib can have, If more, remove or generalize (2-8). **Generalized relation threshold control:** No of tuples in relation

**Class discrimn/comparison:** mines descriptions that distinguish a target class from its contrasting classes. Data collen, Dim relevance analysis, synchronous generalizn, presentation of comparison

**Differences betn OLAP and concept descpn: 1)**OLAP has systematic preprocessing, query independent, and can drill down to rather low level. **2)** AOI has automated desired level allocation, and may perform dimension relevance analysis/ranking when too many dimensions **3)** AOI works on non-relational data

**Chp 5**: **Sparse:** product of cardinalities of dims is large relative to number of non-zero valued tuples. **Optimization:** 1) Sorting, hashing, grouping 2) Simultaneous aggregn, agg from child not base 3) Agg from smallest child 4) Apriori pruning (iceberg)

**Multi-Way Array agg:** 1)Partition array into chunks. 2)Compress chunks to remove empty cells. Cell addressed using chunk id+ofset

3) Compute aggregates in “multiway” by visiting cube cells in the order which minimizes the # of times to visit each cell, and reduces memory access and storage cost.. 40\*400 + 40\*1000+100\*1000. Moderate dims, data not sparse, normal cardinality. Full cube. **BUC:** sparse and iceberg cubes. Top-down. Divides dimensions into partitions and facilitates iceberg pruning, If a partition does not satisfy min\_sup, its descendants can be pruned. Usually, entire data set can’t fit in main memory, Sort distinct values, partition into blocks that fit

**Shell Fragment approach:** semi-online computation model. Compute aggregates on small number of dims of interest. Inverted index data structure. 1) Partition the set of dimensions into **shell fragments** 2)Compute data cubes for each shell fragment while retaining **inverted indices** or **value-list indices (Top down depth first order-AB, then ABC)** 3) Given the pre-computed **fragment cubes**, dynamically compute cube cells of the high-dimensional data cube ***online.* Point query:** All dims have been instantiated. Answer is vale of measure/ **Subcube query:** atleast 1 relevant dim is inquired ?. \* irrelevant. Returns data cube based on dims.

**Challenges in sampling:** Sparse multidim data, in spite of large. Drill down, few samples. Sampling bias, outliers

**Confidence interval** is algebraic: 1)mean is alg 2) t is based on lookup, depends on l(dist) 3) linear sum and squared sum are alg

**Query Expansion:** Confidence interval is large (i.e., low confidence). Small sample size .High standard deviation with samples.

**1) Intracuboid:** include nearby cells in same cuboid. Weakly correlated/uncorrelated dims. Will not change query semantics or shift query answer. Eg. Avg income of educ=college and birth month-july. Can expand to include more birth months.

2) **Intercuboid:** Expansion by looking to a more general cell.

**Ranking cubes:** Efficient processing of Top-k queries. User specified preference consists of 1)selection condition (where) 2) ranking function (order by)

***Select top 5\* from R where producer=”Ford” and type=”sedan”***

***Order by (price-10K)^2 + (mileage-30k)^2 asc***

Materialize cube on a set of selection dimensions.

**Step 1: Partition Data on Ranking Dimensions**

**Step 2: Group data by** **Selection Dimensions**

**Step 3: Compute Measures for each group**

**Chp 6:** **Antimonotonic**: If an itemset does not satisfy the constraint, none of the supersets can satisfy it. **monotonic**: If an itemset satisfies the constraint, so do the supersets.

**Apriori: 1)Join:** To find Lk, join Lk-1 with itself to get Ck. Sorted in lexicographic order. **2)Prune:** Ck is a superset of Lk. To reduce Ck size, use apriori property: All non-empty subsets of a fre itemset must also be frequent. **Improving Apriori Efficiency: 1)Hash based 2) Transacn reducn:** if a transc does not contain a frequent k itemset, it cannot contain a frequent k+1 itemset. Mark such. **3) Partitioning:** 2 DB scans. **4) Sampling:** Mine patterns from sample instead of DB. Lower threshold. **5) Dynamic itemset counting: O**nce both A & D are determined frequent, counting of AD begins. **Problems with apriori:** BFS, huge num of candidates, scan DB

**FP Growth:** For each frequent item, construct its conditional pattern-base, and then its conditional FP-tree. Repeat the process on each newly created conditional FP-tree Until the resulting FP-tree is empty, or it contains only one path—single path will generate all the combinations of its sub-paths, each of which is a frequent pattern

**Vertical Data Format: item,TID set**

t(X) = {T1, T2, T3}, t(XY) = {T1, T3} Diffset (XY, X) = {T2}

**Pattern Evaluation:** Lift, chi-square

**Null invariant measure**: the value is free from the influence of null transaction (does not contain any of the examined items).

**Kulczynski and Imbalance Ratio** together present a clear picture

**Chp 7: Pattern Mining Classificn (Pattern diversity): 1)Basic 2)Abstraction level 3)number of dims 4) Type of values** (Boolean/Quantitative)

Association rules generated by mining data at multiple abstracn levels is are called **multilevel asso rules**. **Types:** single-dim, multi-dim(interdim (non-repeated), hybrid (repeated))

Techniques can be categorized by how numerical attributes are treated 1)Static discretization based on predefined concept hierarchies 2)Dynamic discretization based on data distribution

An **infrequent or rare** pattern has a support below user defined support. 3 definitions: Negatively correlated if

1) sup(X U Y) < sup (X) \* sup(Y)

2) **write formula 3)If itemsets X and Y are frequent, but (P(X|Y) + P(Y|X))/2 < є, where є is a negative pattern threshold**