Data Mining and Data Warehousing 3

Data Visualization & Data Similarity and DIssimilarity Measurement

Chittaranjan Pradhan
School of Computer Engineering,
KIIT University

Statistical Descriptions Visualization

Boxplot

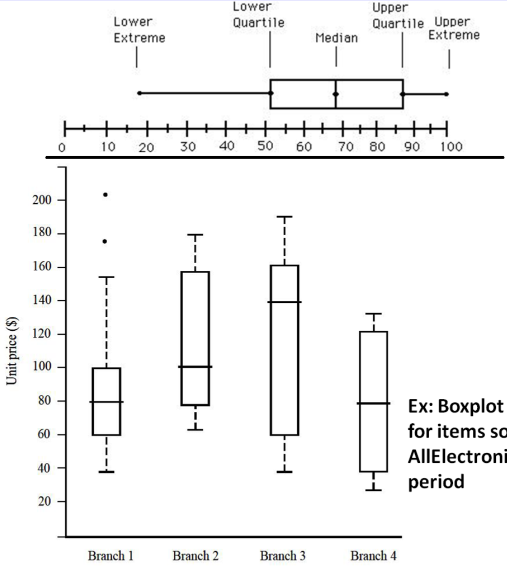
- Boxplots are a popular way of visualizing a distribution
- A boxplot incorporates the five-number summary as:

Minimum, Q1, Median, Q3, Maximum

  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations
  - Points beyond a specified outlier threshold, plotted individually

Statistical Descriptions Visualization...

Boxplot...

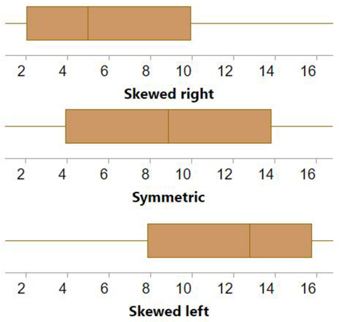


Ex: Boxplot for the unit price data for items sold at four branches of AllElectronics during a given time period

Statistical Descriptions Visualization...

Boxplot...

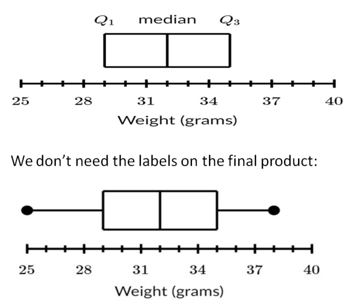
- Boxplots often provide information about the shape of a data set
- If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa
- If a distribution is symmetric, the observations will be evenly split at the median



Statistical Descriptions Visualization...

Boxplot...

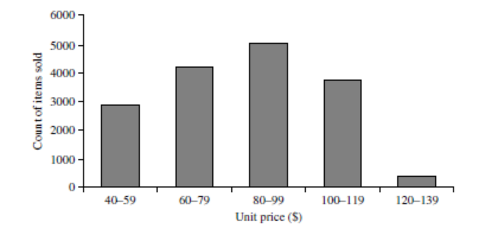
- Ex: 25, 28, 29, 29, 30, 34, 35, 35, 37, 38
- Median -> 32
- First quartile is the median of the data points to the left of the median: 25, 28, 29, 29, 30. So, Q1->29
- Third quartile is the median of the data points to the right of the median: 34, 35, 35, 37, 38. So, Q3->35
- Min-> 25 and Max->38



Statistical Descriptions Visualization...

Histogram Analysis

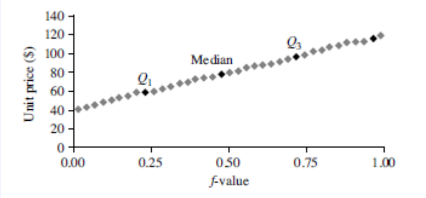
- Plotting histograms is a graphical method for summarizing the distribution of a given attribute X. The height of the bar indicates the frequency (or count) of that X value
- Differs from a bar chart in that it is the area of the bar that denotes the value, not the height as in bar charts, a crucial distinction when the categories are not of uniform width
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Statistical Descriptions Visualization...

Quantile Plot

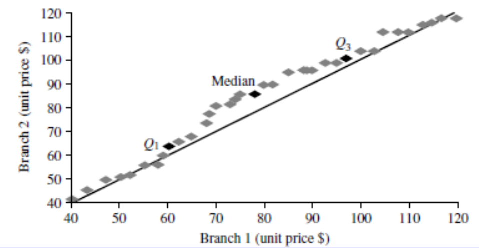
- It displays all of the data for the given attribute (both behavior and occurrences)
- It plots quantile information
- For a data x_i data sorted in increasing order, f_i indicates that approximately $100 * f_i\%$ of the data are below or equal to the value x_i
  where $f_i = \frac{i-0.5}{N}$
- It compares different distributions based on their quantiles
- It allows the user to view whether there is a shift in going from one distribution to another



Statistical Descriptions Visualization...

Quantile-Quantile Plot

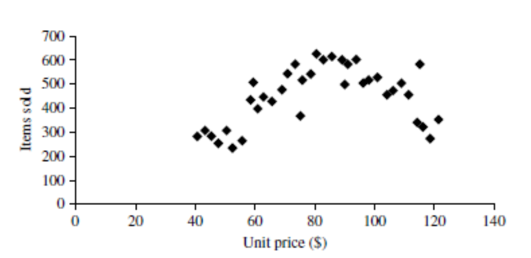
- It graphs the quantiles of one univariate distribution against the corresponding quantiles of another
- It is a powerful visualization tool in that it allows the user to view whether there is a shift in going fromone distribution to another
- Ex: unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2



Statistical Descriptions Visualization...

Scatter Plot

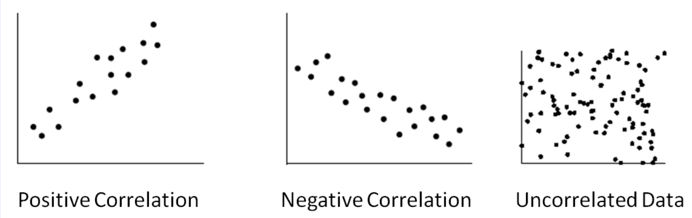
- Scatter plot is a type of data visualization that shows the relationship between different variables
- This data is shown by placing various data points between X-axis and Y-axis
- Provides a first look at bivariate data to see clusters of points, outliers, corelation etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Statistical Descriptions Visualization...

Scatter Plot with Data Correlation

- Two attributes, X, and Y, are correlated if one attribute implies the other
- Correlations can be positive, negative, or null (uncorrelated)
  - If the plotted points pattern slopes from lower left to upper right, this means that the values of X increase as the values of Y increase, suggesting a positive correlation
  - If the pattern of plotted points slopes from upper left to lower right, the values of X increase as the values of Y decrease, suggesting a negative correlation



Measuring Data Similarity and Dissimilarity

Measuring Data Similarity and Dissimilarity

- In data mining applications such as clustering, outlier analysis, classification it is required to find how much similar or dissimilar one data is in comparison to the other
- **Similarity**
  - A Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike and lower when objects are more dissimilar
  - Often falls in the range [0,1]
- **Dissimilarity**
  - A Numerical measure of how different two data objects are
  - Value is higher when objects are more dissimilar and lower when objects are more alike
  - Minimum dissimilarity is often 0 and the upper limit varies
- Proximity refers to a similarity or dissimilarity

Measuring Data Similarity and Dissimilarity...

Data Matrix and Dissimilarity Matrix

- **Data Matrix**
  - object-by-attribute structure
  - It stores the n data objects in the form of a relational table, or n-by-p matrix (n objects x p attributes)
  - Each row corresponds to an object
  - It is a *two-mode* matrix as it is made up of rows and columns
- **Dissimilarity Matrix**
  - object-by-object structure
  - It stores a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table
  - It is a *one-mode* matrix as it is made up of dissimilarity

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}.$$

Data Matrix

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix},$$

Dissimilarity Matrix

Measuring Data Similarity and Dissimilarity...

Similarity/Dissimilarity for Simple Attributes

Similarity and dissimilarity between two objects, x and y w.r.t. single, simple attribute

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n - 1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Proximity Measures for Nominal Attributes

Proximity Measures for Nominal Attributes

- It can take on two or more states
- map_color is nominal with state <red, yellow, green, pink, blue>

- **Dissimilarity between Nominal Attributes**: simple ratio of mismatches
  $d(i,j) = \frac{p-m}{p}$

  where,
  m-> number of matches
  p->total number of variables / attributes

- **Similarity between Nominal Attributes**: simple ratio of matches
  $sim(i,j) = 1 - d(i,j) = \frac{m}{p}$

Proximity Measures for Nominal Attributes

Proximity Measures for Nominal Attributes

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Dissimilarity Matrix: d(i,j)

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

Similarity Matrix: sim(i, j)= 1-d(i,j)

$$\begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0 & 0 & 1 & \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

Proximity Measure for Binary Attributes

Proximity Measure for Binary Attributes

- Binary attribute has only two states: 0 and 1
- Contingency table is used to describe the distance between two objects (let p & q) with only binary attributes
- Compute dissimilarity/similarity using the quantiles:
  - f01 -> the number of attributes where p was 0 and q was 1
  - f10 -> the number of attributes where p was 1 and q was 0
  - f00 -> the number of attributes where p was 0 and q was 0
  - f11 -> the number of attributes where p was 1 and q was 1

	Object q			
		1	0	sum
Object p	1	f11	f10	f11+f10
	0	f01	f00	f01+f00
	Sum	f11+f01	f10+f00	

Proximity Measure for Binary Attributes...

Proximity Measure for Binary Attributes...

- **DIssimilarity measure of Binary attributes**
  - **Symmetric binary variables**

    $$\frac{f10+f01}{f11+f10+f01+f00}$$

  - **Asymmetric binary variables**

    $$\frac{f10+f01}{f11+f10+f01}$$

- **Similarity measure of Binary attributes**
  - **Symmetric binary variables [Simple matching coefficients]**

    $$\frac{f11+f00}{f11+f10+f01+f00}$$

  - **Asymmetric binary variables [Jaccard index similarity]**

    $$\frac{f11}{f11+f10+f01}$$

Proximity Measure for Binary Attributes...

Proximity Measure for Binary Attributes...

Dissimilarity between binary variables

- Gender is a symmetric attribute and the remaining attributes are asymmetric binary
- Let the values Y and P be 1, and the value N be 0
- DIssimilarity in terms of only asymmetric attribute= $\frac{f10+f01}{f11+f10+f01}$

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$
$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$
$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$$

- Among three patients, Jack and Mary have similar diseases [as their distance is less]

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Proximity Measure for Binary Attributes...

Proximity Measure for Binary Attributes...

- Similary in terms of only asymmetric attribute [**Jaccard index similarity**] = $\frac{f11}{f11+f10+f01}$

  $d(Jack, Mary) = \frac{2}{2+0+1} = 0.66$
  $d(Jack, Jim) = \frac{1}{1+1+1} = 0.33$
  $d(Jim, Mary) = \frac{1}{1+1+2} = 0.25$
- Among patients, Jack and Mary have similar diseases [as their similarity is more]

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

Dissimilarity of Numeric Attributes

Dissimilarity of Numeric Attributes

Commonly used distance measures for computing dissimilarity are:

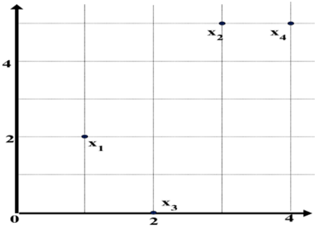
- **Euclidean Distance**
  $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
  where,
  n->number of dimensions (attributes)
  x_i and y_i are the i^{th} attributes of data objects x and y

- **Manhattan Distance**
  $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

- **Minkowski Distance**: is a generalization of the Euclidean and Manhattan distances. It is defined as
  $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
  When p=1 -> it is Manhattan
  When p=2 -> it is Euclidean

Dissimilarity of Numeric Attributes...

Dissimilarity of Numeric Attributes...



Point	Attribute1	Attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix (Euclidean Distance)

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

Dissimilarity Matrix (Manhattan Distance)

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Dissimilarity of Numeric Attributes...

Dissimilarity of Numeric Attributes...

Point	Attribute1	Attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Minkowski Distance $$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

Where p-> real number >=1

Minkowski Distance with p value 1

Manhattan Distance $$d(x, y) = \sum_{i=1}^{n}|x_i - y_i|$$

	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Minkowski Distance with p value 2

Euclidean Distance $$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	5.1	5.1	0	
x4	4.24	1	5.39	0

Proximity Measure for Ordinal Attributes

Proximity Measure for Ordinal Attributes

- Ordinal variables are discrete or continuous with meaningful order or ranking about them but the magnitude between successive ranks is unknown
- Ex: Attribute Size: <small, medium, large>

  Attribute temperature: <-30 to -10: cold, -10 to 10: Moderate, 10 to 30: warm>
- For an attribute f, let M represent the total number of possible states. Then the order/rank states is $< 1...M_f >$
- For an object i, let r_{if} is the rank of the i^{th} object: $r_{if} \in 1, ..., M_f$
- Since any ordinal attribute can have a different number of states, it is necessary to normalize the range onto [0.0 to 1], where each of the rank r_{if} will be normalized to z_{if}: $z_{if} = \frac{r_{if}-1}{M_f-1}$ and then find it's dissimilarity matrix

Proximity Measure for Ordinal Attributes...

Proximity Measure for Ordinal Attributes...

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Obj Id	Test-2 (Ordinal)	Rank (r_{ij})	Normalized Rank (z_{ij})
1	Excellent	3	1
2	Fair	1	0
3	Good	2	0.5
4	Excellent	3	1

Dissimilarity Matrix: d(i, j)

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

Similarity Matrix: sim(i, j)= 1-d(i,j)

$$\begin{bmatrix} 1-0 & & & \\ 1-1 & 1-0 & & \\ 1-.5 & 1-.5 & 1-0 & \\ 1-0 & 1-1 & 1-.5 & 1-0 \end{bmatrix} = \begin{bmatrix} 1 & & & \\ 0 & 1 & & \\ 0.5 & 0.5 & 1 & \\ 1 & 0 & 0.5 & 1 \end{bmatrix}$$

Dissimilarity for Attributes of Mixed Types

Dissimilarity for Attributes of Mixed Types

- The real database mostly contains mixed types of attributes. The simple approach is to process all types of attributes together by performing a single analysis to obtain a single dissimilarity matrix with a common scale of an interval [0.0, 1.0] (normalized form)
- Let the data set contains p attributes of mixed type. The dissimilarity d(i, j) between object i and object j is:

  $$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}(f) d_{ij}(f)}{\sum_{f=1}^{p} \delta_{ij}(f)}$$

  where, d_{ij} -> is the dissimilarity between object i and j
  $\delta_{ij}(f)$ -> is the indicator
  $\delta_{ij}(f)$ =0, if x_{if}/x_{jf} is missing or $x_{if} = x_{jf} = 0$ for asymmetric binary attribute
  $\delta_{ij}(f)$ =1, otherwise

Dissimilarity for Attributes of Mixed Types...

Dissimilarity for Attributes of Mixed Types...

$$
d(i,j) = \begin{cases}
\textbf{if f is numeric: } & d_{ij}^{(f)} = \dfrac{|x_{if} - x_{jf}|}{max(h) - min(h)} \text{ where h is the attribute set} \\
\textbf{if f is nominal or binary: } & d_{ij}^{(f)} = 0 \text{ if } x_{if} = x_{jf} \text{ otherwise } d_{ij}^{(f)} = 1 \\
\textbf{if f is ordinal: } & \text{Compute rank } r_{if} \text{ and } z_{if} \text{ where } z_{if} = \dfrac{r_{if} - 1}{M_f - 1}
\end{cases}
$$

Dissimilarity for Attributes of Mixed Types...

Dissimilarity for Attributes of Mixed Types...

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Dissimilarity Matrix of Numeric Attribute (test-3)

max(h)=64 and min(h)=22

max(h) - min(h)= 42

$$\frac{|x_{if}-x_{jf}|}{max(h)-min(h)} = \frac{x_{i3}-x_{j3}}{42}$$

Dissimilarity Matrix of Nominal Attribute (test-1)

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity Matrix of Ordinal Attribute (test-2)

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1 & 0 & \\ 0.4 & 0.14 & 0.8 & 0 \end{bmatrix}$$

Dissimilarity for Attributes of Mixed Types...

Dissimilarity for Attributes of Mixed Types...

Using the generalized formula of mixed Attribute Disimmilarity:

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$$\begin{bmatrix} 0 & & & \\ \frac{1+1+.55}{3} & 0 & & \\ \frac{1+.5+.45}{3} & \frac{1+.5+1}{3} & 0 & \\ \frac{0+0+.4}{3} & \frac{1+1+.14}{3} & \frac{1+.5+.86}{3} & 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

Cosine Similarity

Cosine Similarity

- It is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words
- Let x and y be two vectors for comparison. Cosine similarity = $sim(x, y) = \frac{x.y}{||x||||y||}$

  where, . -> vector dot product
  ||x|| -> Euclidean norm of vector x
  ||y|| -> Euclidean norm of vector y
- The measure computes the cosine of the angle between vectors x and y
  - Cosine value of 0 means the two vectors are at 90 degrees to each other and have no match
  - The closer the cosine value to 1, the smaller the angle and the greater the match between vectors

Cosine Similarity...

Cosine Similarity...

- d1 = 3, 2, 0, 5, 0, 0, 0, 2, 0, 0
  d2 = 1, 0, 0, 0, 0, 0, 0, 1, 0, 2
- $d1.d2 = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$
- $||d1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
- $||d2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$
- $cos(d1, d2) = .3150$

Ref: J. Han, J. Pei and H. Tong, "Data Mining: Concepts and Techniques", Morgan Kaufmann, 4th edition