BDA(18CS72)

MODULE-5

MACHINE LEARNING ALGORITHMS FOR BIG DATA ANALYTICS

ESTIMATING THE RELATIONSHIPS, OUTLIERS, VARIANCES, PROBABILITY DISTRIBUTIONS AND CORRELATIONS

Relationships-Using Graphs, Scatter Plots and Charts

- A relationship between two or more quantitative dependent variables with respect to an independent variable can be well-depicted using graph, scatter plot or chart with data points, shown in distinct shapes.
- Conventionally, independent variables are on the x-axis, whereas the dependent variables on the y-axis in a graph. A line graph uses a line on an x-y axis to plot a continuous function.

- A scatter plot is a plot in which dots or distinct shapes represent values of the dependent variable at the multiple values of the independent variable .
- Whether two variables are related to each other or not, can be derived from statistical analysis using scatter plots.

Linear Relationships

- A linear relationship exists between two variables, say x and y, when a straight line (y = a0 + a1 .x) can fit on a graph, with at least some reasonable degree of accuracy.
- The a1 is the linearity coefficient. For example, a scatter chart can suggest a linear relationship, which means a straight line.
- Figure 6.1 shows a scatter plot, which fits a linear relationship between the number of students opting for computer courses in years between 2000 and 2017

Linear Relationships

- A linear relationship can be positive or negative.
- A positive relationship implies if one variable increases in value, the other also increases in value.
- A negative relationship, on the other hand, implies when one increases in value, the other decreases in value.
- Perfect, strong or weak linearship categories depend upon the bonding between the two variables.

Linear Relationships

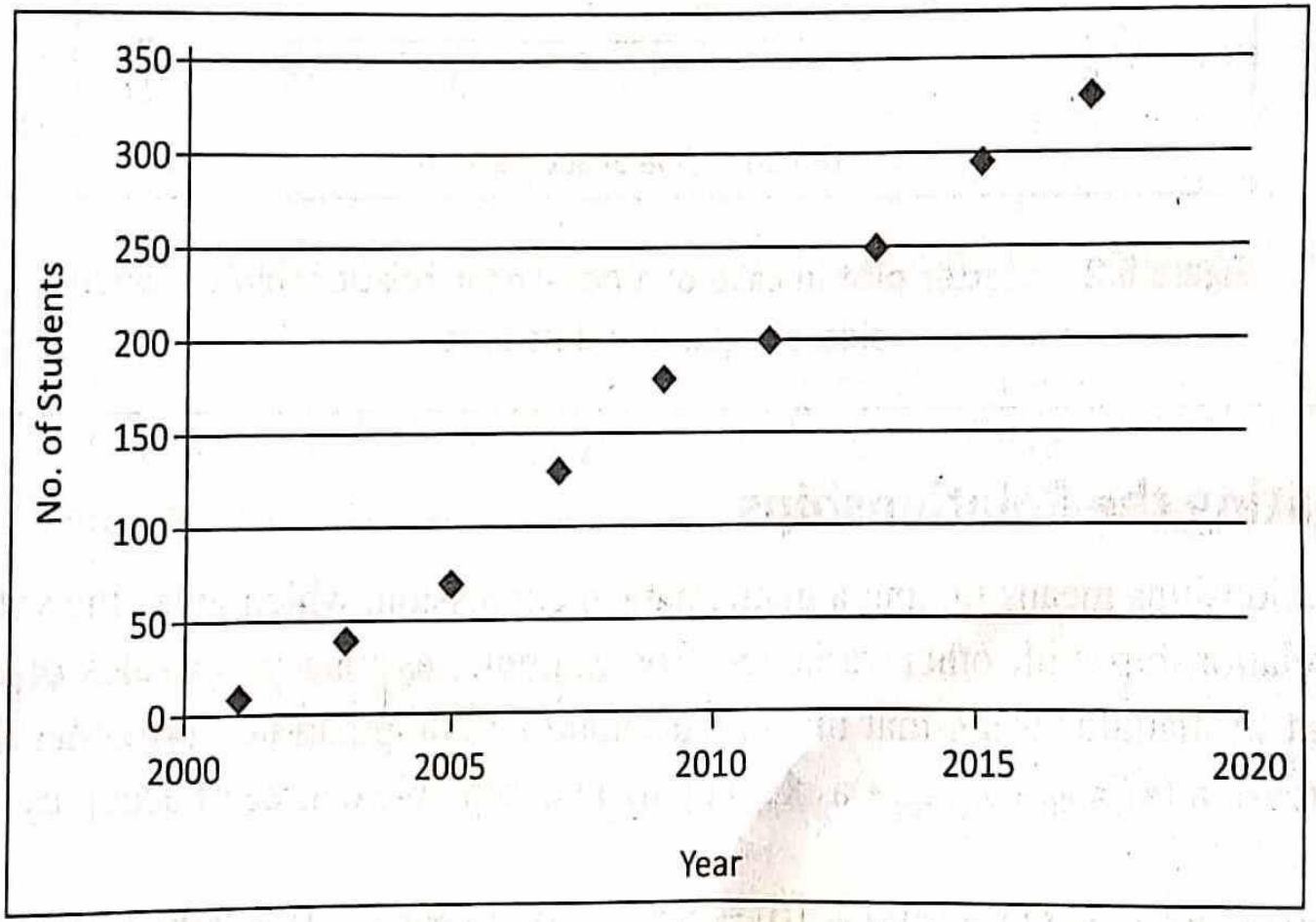

Figure 6.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017

Non-linear Relationships

- A non-linear relationship is said to exist between two quantitative variables when a curve ($y = a0 + a1 .x + a2.x2 + ...$) can be used to fit the data points.
- The fit should be with at least some reasonable degree of accuracy for the fitted parameters, a0, a1, a2 …
- Expression for y then generally predicts the values of one quantitative variable from the values of the other quantitative variable with considerably more accuracy than a straight line.

Non-linear Relationships

- Consider an example of non-linear relationship: The side of a square and its area are not linear.
- In fact, they have quadratic relationship. If the side of a square doubles, then its area increases four times. The relationship predicts the area from the side.

Non-linear Relationships

Figure 6.2 shows a scatter plot in case of a non-linear relationship between side of square and its area

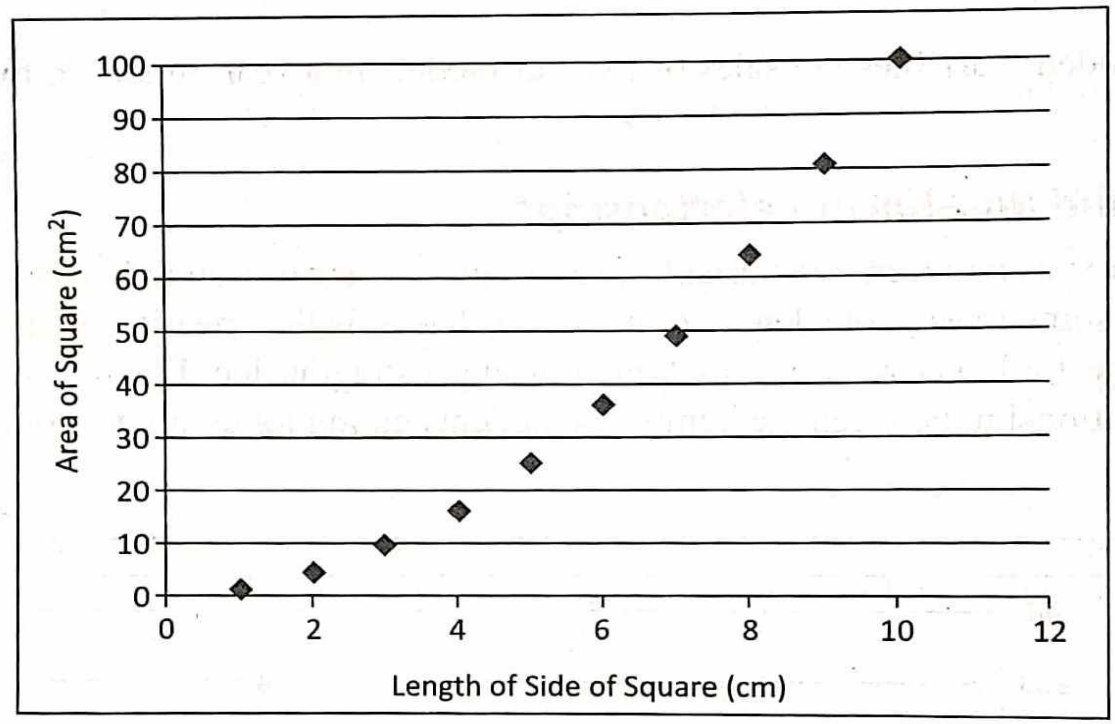

Figure 6.2 Scatter plot in case of a non-linear relationship between side of square and its area

Outliers

- Outliers are data, which appear as they do not belong to the dataset.
- Outliers are data points that are numerically far distant from the rest of the points in a dataset, are termed as outliers.
- Outliers show significant variations from the rest of the points .
- Identification of outliers is important to improve data quality or to detect an anomaly.
- The estimating parameters mathematically, statistically, describing an outcome, predicting a dependent variable value, or taking the decisions based on the datasets given for the analysis are sensitive to the outliers.

Outliers(contd..)

There are several reasons for the presence of outliers in relationships. Some of these are:

- Anomalous situation
- Presence of a previously unknown fact
- Human error (errors due to data entry or data collection)
- Participants intentionally reporting incorrect data (This is common in self-reported measures and measures that involve sensitive data which participant doesn't want to disclose)
- Sampling error (when an unfitted sample is collected from population).

Variance

- A random variable is a variable whose possible values are outcomes of a random phenomenon. A random variable is a function that maps the outcomes of unpredictable processes to numerical quantities.
- A random variable is also called stochastic variable or random quantity.
- Randomness can be around some expected mean value or outcome, and with some normal deviation.

Variance(contd…)

- Variance measures by the sum of squares of the difference in values of a variable with respect to the expected value.
- Variance can alternatively be a sum of squares of the difference with respect to value at an origin. Variance indicates how widely data points in a dataset vary.
- If data points vary greatly from the mean value in a dataset, the variance is large; otherwise, the variance is less.
- The variance is also a measure of dispersion with respect to the expected value.
- A high variance indicates that the data in the dataset is very much spread out over a large area (random dataset), whereas a low variance indicates that the data is very similar in nature.

Standard Deviation and Standard Error Estimates

6.2.4.1 *Standard Deviation and Standard Error Estimates*

The variance is not a standalone statistical parameter. Estimations of other statistical parameters, such as standard deviation and standard error are also used.

Standard Deviation With the help of variance, one can find out the standard deviation. Standard deviation, denoted by σ, is the square root of the variance. The σ says, "On an average how far do the data points fall from the mean or expected outcome?" Though the interpretation is the same as variance but σ is squared rooted, therefore, less susceptible to the presence of outliers. The formulae for the population and the sample standard deviations are as follows:

The Population Standard Deviation: $$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2} \quad (6.1a)$$

The Sample Standard Deviation: $$\sigma = \sqrt{\frac{1}{S-1}\sum_{i=1}^{S}(x_i - \overline{x})^2}\ , \quad (6.1b)$$

where N is number of data points in population, S is number in the sample, μ is expected in the population or average value of x, and $\overline{x}$ is expected x in the sample.

Standard Error The standard error estimate is a measure of the accuracy of predictions from a relationship. Assume the linear relationship in a scatter plot of y (Figure 6.1). The scatter plot line, which fits, is defined as the line that minimizes the sum of squared deviations of prediction (also called the **sum of squares error**). The **standard error of the estimate** is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum(y - y')^2}{N}}, \quad \ldots (6.2)$$

where σ_{est} is the standard error in the estimate, y is an observed value, y′ is a predicted value, and N is the number of values observed. The standard error estimate is a measure of the dispersion (or variability) in the predicted values from the expression for relationship. Following are three interpretations from the σ_{est}:

1. When σ_{est} is small, most of the observed values (y) dots are fairly close to the fitting line in the scatter plot, and better is the estimate based on the equation of the line.
2. When the σ_{est} is large, many of the observed values are far away from the line.
3. When the standard error is zero, then no variation exists corresponding to the computed line for predictions. The correlation between the observed and estimation is perfect.

Analysis of Variance (ANOVA)

- Analysis of Variance (ANOVA) An ANOVA test is a method which finds whether the fitted results are significant or not. This means that the test finds out (infer) whether to reject or accept the null hypothesis.
- Null hypothesis is a statistical test that means the hypothesis that "no significant difference exists between the specified populations". Any observed difference is just due to sampling or experimental error.

Analysis of Variance (ANOVA) (Contd…)

- Consider two specified populations (datasets) consisting of yearly sales data of Tata Zest and Jaguar Land Rover models.
- The statistical test is for proving that yearly sales of both the models, means increments and decrements of sales are related or not.
- Null hypothesis starts with the assumption that no significant relation exists in the two sets of data (population).
- The analysis (ANOV A) is for disproving or accepting the null hypothesis.
- The test also finds whether to accept another alternate hypothesis. The test finds that whether testing groups have any difference between them or not.

Analysis of Variance (ANOVA) (Contd…)

- Analysis of variance (ANOV A) is a useful technique for comparing more than two populations, samples, observations or results of computations.
- It is used when multiple sample cases are involved. Variation between samples and also within sample items may exist.
- For example, compare the effect of three different types of teaching methodologies on students. This may be done by comparing the test scores of the three groups of 20 students each.
- This technique provides inferences about whether the samples have been drawn from populations having the same mean.
- It is done by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.

Correlation

- Correlation means analysis which lets us find the association or the absence of the relationship between two variables, x and y.
- Correlation gives the strength of the relationship between the model and the dependent variable on a convenient 0-100% scale.

Correlation (contd..)

R-Square

- R is a measure of correlation between the predicted values y and the observed values of x.
- R-squared (R pow 2) is a goodness-of-fit measure in linear-regression model. It is also known as the coefficient of determination.
- R square is the square of R, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.

REGRESSION ANALYSIS

- Correlation and regression are two analyses based on multivariate distribution. A multivariate distribution means a distribution in multiple variables.
- **Regressive analysis means estimating relationships between variables.**
- Regression analysis is a set of statistical steps, which estimate the relationships among variables.
- Regression analysis may require many techniques for modeling and performing the analysis using multiple variables.
- The aim of the analysis is to find the relationships between a dependent variable and one or more independent, outcome, predictor or response variables.
- Regression analysis facilitates prediction of future values of dependent variables. It helps to find how a dependent variable changes when variation is in an independent variable among a set of them, while the remaining independent variables in the set are kept fixed.

Simple Linear Regression

- Linear regression is a simple and widely used algorithm. It is a supervised ML algorithm for predictive analysis.
- It models a relationship between the independent predictor or explanatory, and the dependent outcome or variable, y using a linearity equation

EXAMPLE 6.3

How can a university student's GPA be predicted from his/her high school percentage (HSP) of marks?

SOLUTION

Consider a sample of ten students for whom their GPAs and high school scores, HSPs, are known. Assume linear regression. Then,

$$GPA = b_1 . HSP + A \quad \ldots (6.11)$$

Figure 6.5 shows a simple linear regression plot for the relationship between the college GPA and the percentage of high school marks. Plot the values on a graph, with high school scores in percentage on the x axis and GPA on the y axis.

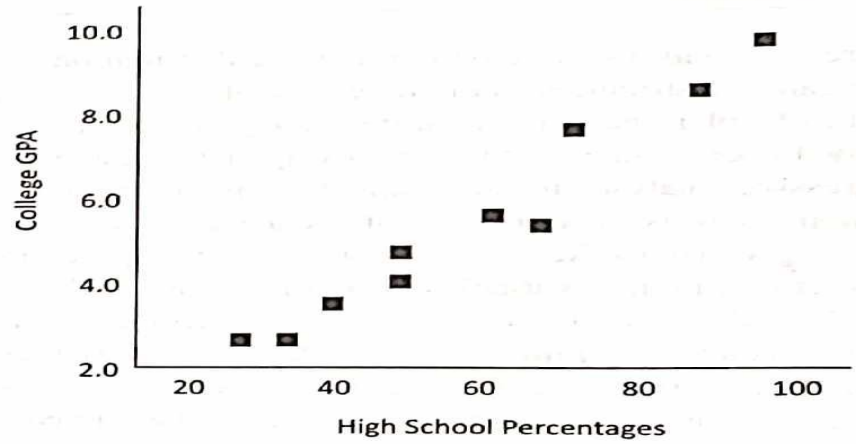

Figure 6.5 Linear regression relationship between college GPA and percentage of high school marks

Whenever a perfect linear relationship between GPA and high school score exists, all 10 points on the graph would fit on a straight line. However, this is never the case. Whenever an imperfect linear relationship exists between these two variables, a cluster of points on the graph, which slope upward, may be obtained. In other words, students who got more marks in high school should get more GPA in college as well.

One variable, denoted by x, is regarded as the predictor, explanatory or independent variable. The other variable, denoted by y, is regarded as the response, outcome or dependent variable.

- The purpose of regression analysis is to come up with an equation of a line that fits through a cluster of points with minimal amount of deviation from the line.
- The best-fitting line is called the ***regression line***. The deviation of the points from the line is called an '***error***'.
- Once this regression equation is obtained, the GPA of a student in college examinations can be predicted provided his/her high school percentage is given.
- Simple linear regression is actually the same as a correlation between independent and dependent variables.

- Figure 6.6 shows a simple linear regression with two regression lines with different regression equations.
- Looking at the scatter plot, two lines can fit best to summarize the relation between GPA and high school percentage.

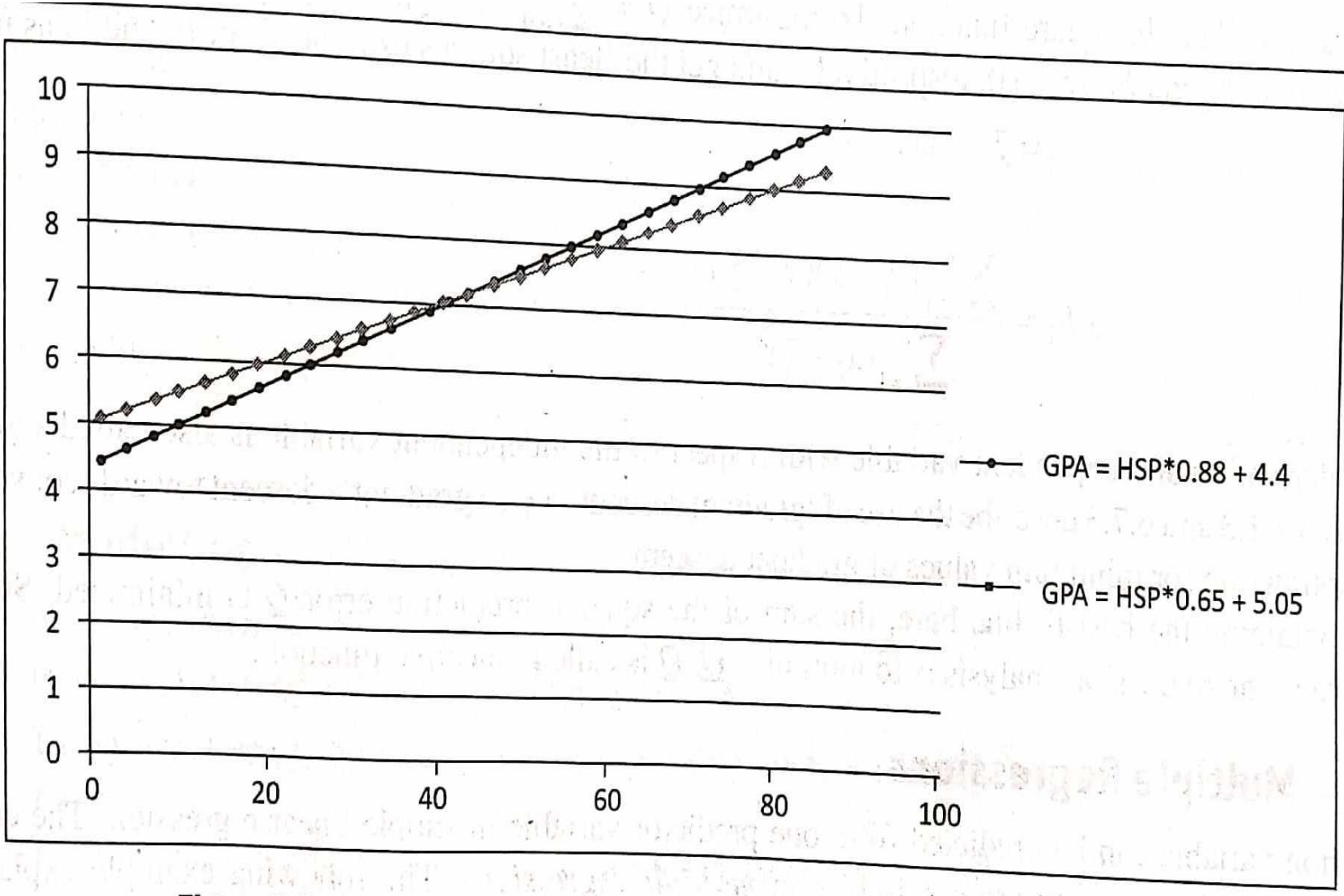

Figure 6.6 Linear regression relationship with two regression lines with different coefficient in regression equation

Modelling Possibilities using Regression

- Regressions range from simple models to highly complex equations. Two primary uses for regression are forecasting and optimization. Consider the following examples:

1. Using linear analysis on sales data with monthly sales, a company could forecast sales for future months.
2. For the funds that a company has invested in marketing a particular brand, an analysis of whether the investment has given substantial returns or not can be made.

3.Suppose two promotion campaigns are running on TV and Radio in parallel. A linear regression can confine the individual as well as the combined impact of running these advertisements together.

4. An insurance company exploits a linear regression model to obtain a tentative premium table using predicted claims to Insured Declared Value ratio.

5. A financial company may be interested in minimizing its risk portfolio and hence want to understand the top five factors or reasons for default by a customer.

6. To predict the characteristics of child based on the characteristics of their parents.

7. A company faces an employment discrimination matter in which a claim that women are being discriminated against in terms of salary is raised.

8. Predicting the prices of houses, considering the locality and builder characteristics in a locality of a particular city.

9. Finding relationships between the structure and the biological activity of compounds through their physical, chemical and physicochemical traits is most commonly performed with regression techniques.

10. To predict compounds with higher bioactivity within groups.

Predictions using Regression Analysis

- Regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of another variable.
- Regression analysis is generally a statistical method to deal with the formulation of a mathematical model depicting the relationship amongst dependent and independent variables.
- The dependent variable is used for the purpose of prediction of the values. One or more variables whose values are hypothesized are called independent variables.
- The prediction for the dependent variable can be made by accurate selection of independent variables to estimate a dependent variable.

Two steps for predicting the dependent variable:

1. *Estimation step: A function is hypothesized and the parameters of the function are estimated from the data collected on the dependent variable.*
2. *Prediction step: The independent variable values are then input to the parameterized function to generate predictions for the dependent variable.*

K-Nearest-Neighbour Regression Analysis

- Consider the saying, 'a person is known by the company he/she keeps.' Can a prediction be made using neighbouring data points?
- K-Nearest Neighbours (KNN) analysis is an ML based technique using the concept, which uses a subset of K = 1, 2 or 3 neighbours in place of a complete dataset. The subset is a training dataset.
- K-Nearest Neighbours (KNN) is an algorithm, which is usually used for classifiers. However, it is useful for regression also. Predictions can use all k examples (global examples) or just K examples (K-neighbours with K = 1, 2 or 3).

- A subset of training dataset restricts k to K-neighbours, where K = 1, 2 or 3. This means using local values near the predictor variable.
- K = 1 means the nearest neighbour data points. K = 2 means the next nearest neighbour data points (xi,YJ K = 3 means the next to next nearest neighbour data points (xi,Yi) .

Assume continuously varying values as a function of independent variables. Assume v denotes the number of variables, independent as well as dependent. The following equations give the KNN distances in *v-dimensional space* for the purpose of using weights.

Euclidean Distance The following equation computes the Euclidean distance D_{Eu}:

Sum of the squared Euclidean distance, $[D_{Eu}]^2 = \left[\sum_{i=1}^{v}(x_i - x_i')^2\right]$, and

$$\text{Euclidean distance } D_{Eu} = \left[\sum_{i=1}^{v}(x_i - x_i')^2\right]^{1/2} \tag{6.20a}$$

Sum is over v dimensions. If one independent and one dependent variable, then $v = 2$. For example, if $v = 2$ and two data points are (x_j, y_j) and (x_{j+1}, y_{j+1}), then Euclidean distance between the points is as follows:

$$\text{Euclidean distance } D_{Eu} = [(x_j - x_{j+1})^2 + (y_j - y_{j+1})^2]^{1/2} \tag{6.20b}$$

Euclidean distance for three variables $v = 3$ (two independent variables and one dependent variable case) consists of three terms on the right-hand side in Equation (6.20b).

Manhattan Distance The following equation computes the Manhattan distance D_{Ma}:

$$\text{Manhattan distance } D_{Ma} = \sum_{i=1}^{v}\left[|x_i - x_i'|\right] \tag{6.20c}$$

Manhattan distance for three variables $v = 3$ (two independent variables and one dependent variable case) consists of three terms on the right-hand side in Equation (6.20c).

Comparison between Euclidean and Manhattan Distances Basically, Euclidean distance is the direct path distance between two data points in v-dimensional metric spaces. Manhattan distance is the staircase path distance between them. Staircase distance means to move to the next point, first move along one metric dimension (say, x axis) from the first point, and then move to the next along another dimension (say, y axis).

When $v = 2$, Euclidean distance is the diagonal distance between the points on an x-y graph. Manhattan distances are faster to calculate as compared to Euclidean distances. Manhattan distances are proportional to Euclidean distances in case of linear regression.

Minkowski Distance The following equation computes the Minkowski distance D_{Mi}:

$$\text{Minkowski distance } D_{Mi} = \left\{\sum_{i=1}^{v}\left[(x_i - x_i')^q\right]\right\}^{1/q} \tag{6.20d}$$

Hamming Distance When predictions are on the basis of categorical variables, then use the Hamming distance. It is a measure of the number of instances in which corresponding values are found.

$$\text{Hamming Distance, } D_H = \sum_{i=1}^{v}|x_i - x_i'|, \tag{6.20e}$$

FINDING SIMILAR ITEMS, SIMILARITY OF SETS AND COLLABORATIVE FILTERING

Finding Similar Items

- Similar item search refers to a data mining method which helps in discovering items which have similarities in datasets. (Data mining means discovering previously unknown interesting patterns and knowledge from apparently unstructured data. The process of data mining uses the ML algorithms. Data mining enables analysis, categorization and summarization of data and relationships among data.)

Application of Near Neighbour Search

- Similar items can be found using Nearest Neighbour Search (NNS). The search finds that a point in a given set is most similar (closest) to a given point.
- A dissimilarity function having larger value means less similar. The dissimilarity function is used to find similar items.

Three problems with the Pearson similarities

1. Do not consider the number of items in which two users' preferences overlap. (e.g., 2 overlap items==> 1, more items may not be better.)
2. If two users overlap on only one item, no correlation can be computed.
3. The correlation is undefined if series of preference values are identical.

Greater distance means greater dissimilarity. Dissimilarity coefficient relates to a distance metric in metrics space in v-dimensional space. An algorithm computes Euclidean, Manhattan and Minkowski distances using Equations.

6.4.2 Jaccard Similarity of Sets

Let A and B be two sets. Jaccard similarity coefficient of two sets measures using notations in set theory as shown below:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6.22}$$

$A \cap B$ means the number of elements or items that are same in sets A and B. $A \cup B$ means the number of elements or items present in union of both the sets. Assume two set of students in two computer courses, Computer Applications CA, and Computer Science CS in a semester. Set CA 40 students opted for Java out of 60 students. Set CS 30 students opted for Java out of 50 students. Jaccard similarity coefficient

- Jaccard similarity coefficient J java (CA, CS) = 30/(60 + 50) x 100% = 27%.
- Two sets are sharing 27% of the members forJava course. (n is symbol for intersection in set theory. U is symbol for union in set theory.)

FREQUENT ITEMSETS
AND
ASSOCIATION RULE MINING

Frequent Itemset Mining (FIM)

- Extracting knowledge from a dataset is the main goal of data analytics and data mining.
- Data mining mainly deals with the type of patterns that can be mined.
- A method of mining is Frequent Patterns (FPs) mining method. Frequent patterns occur frequently in transactional data.
- ***Frequent itemset*** refers to a set of items that frequently appear together, for example, Python and Big Data Analytics. Students of computer science frequently choose these subjects for in-depth studies.
- Frequent itemset refers to a frequent iternset, which is a subset of items that appears frequently in a dataset .

Frequent Itemset Mining(Contd..)

- Frequent Itemset Mining (FIM) refers to a data mining method which helps in discovering the itemsets that appear frequently in a dataset. For example, finding a set of students who frequently show poor performance in semester examinations.
- Frequent subsequence is a sequence of patterns that occurs frequently. For example, purchasing a football follows purchasing of sports kit.
- Frequent substructure refers to different structural forms, such as graphs, trees or lattices, which may be combined with itemsets or subsequences .

Frequent Itemset Mining(Contd..)

- FIM finds the regularities in data. Frequent itemset mining is the preceding step to the association rule learning algorithm. Most often the algorithm is used for analyzing a business.
- For example, customers of supermarkets, mail order companies and online shops use FIM to find a set of products that are frequently bought together. This provides the knowledge of important pairs of items that occur much more frequently than the items bought independently. A sales person can learn the pattern of what should be bought together for sales.

The analysis results in:

- Improvement of arrangement of products in shelves and on catalog pages
- Marketing and sales promotion
- Planning of products that a store should stock up.
- Support cross-selling (suggestion of other products) and product bundling.

Association Rule- Overview

- An important method of data mining is association rule mining or association analysis. The method has been widely used in many application areas for discovering interesting relationships which are present in large datasets. The objective is to find uncovered relationships using some strong rules. The rules are termed as association rules for frequent itemsets.
- Mahout includes a 'parallel frequent pattern growth' algorithm. The method analyzes the items in a group and then identifies which items typically appear together (association)

Apriori Algorithm

- Apriori algorithm is used for frequent itemset mining and association rule mining.
- Apriori algorithm is considered as one of the most well-known association rule algorithms. The algorithm simply follows a basis that any subset of a large itemset must be a large itemset. This basis can be formally given as the Apriori principle.
- The Apriori principle can reduce the number of itemsets needed to be examined.
- Apriori principle suggests if an itemset is frequent, then all of its subsets must also be frequent.
- For example, if itemset {A, B, C} is a frequent itemset, then all of its subsets {A}, {B}, {C}, {A, B}, {B, C} and {A, C} must be frequent.
- On the contrary, if an itemset is not frequent, then none of its supersets can be frequent. This results into a smaller list of potential frequent itemsets as the mining progresses.

Apriori Algorithm(Contd…)

- Steps of the algorithm can be stated in the following manner:

1. Candidate itemsets are generated using only large itemsets of the previous iteration. The transactions in the database are not considered while generating candidate itemsets.
2. The large itemset of the previous iteration isjoined with itself to generate all itemsets having size higher by 1.
3. Each generated itemset that does not have a large subset is discarded. The remaining itemsets are candidate itemsets.

Figure 6.8 shows Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset.

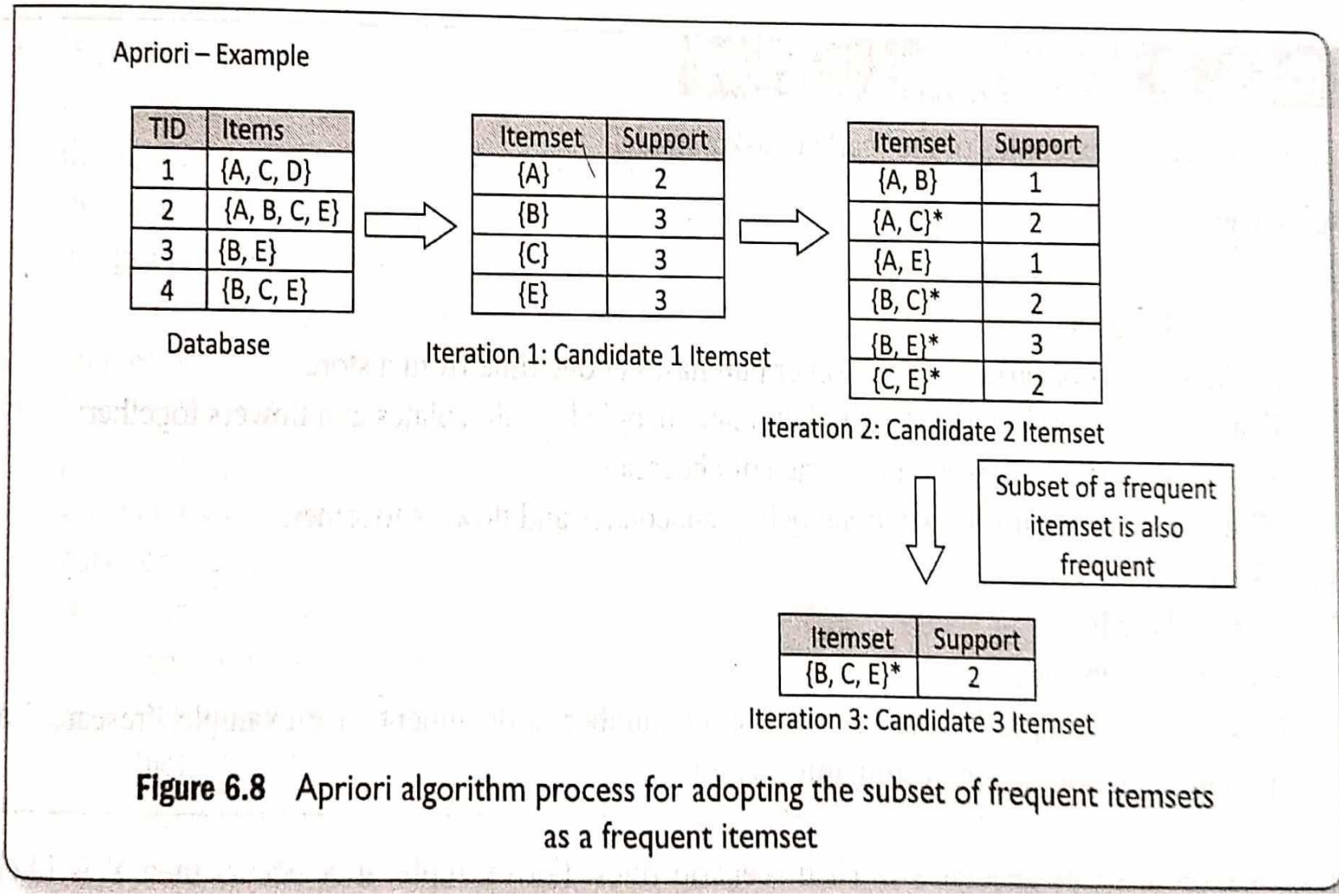

Figure 6.8 Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset

Apriori Algorithm(Contd…)

- It is observed in the Apriori example that every subset of a frequent itemset is also frequent. Thus, a candidate itemset in Ck+1 can be pruned even if one of its subsets is not contained in Fk·
- The Apriori algorithm adopts the fact that the subset of a frequent itemset is also a frequent itemset.
- The algorithm thus reduces the number of candidates being considered by only considering the itemsets whose support count is greater than the minimum support count.
- All infrequent itemsets are pruned if they have an infrequent subset.

Applications of Association Rules

- FIM is a popular technique for market basket analysis

Market Basket Model

- Market basket analysis is a tool for knowledge discovery about co-occurrence of items. A co-occurrence means two or more things occur together.
- It can also be defined as a data mining technique to derive the strength of association between pairs of product items.
- If people tend to buy two products (say A and B) together, then the buyer of product A is a potential customer for an advertisement of product B.

Market Basket Model (Contd…)

- The concept is similar to the real market basket where we select an item (product) and put it in a basket (itemset).
- The basket symbolizes the transactions. The number of baskets is very high as compared to the items in a basket.
- A set of items that is present in many baskets is termed as a frequent itemset. Frequency is the proportion of baskets that contain the items of interest.

EXAMPLE 6.8

Suggest application examples of the market basket model.

SOLUTION

Application 1:

1. Items = Products

   Baskets = Sets of products a customer purchases at one time from a store.

   Example of an application: Given that, many people buy chocolates and flowers together:

   - Run sales on flowers; raise price of chocolates.

   The knowledge is useful when many buy chocolates and flowers together.

Application 2:

2. Items = Words

   Baskets = Web pages

   Unusual words appearing together in a large number of documents, for example, 'research' and 'plastic' may provide interesting information.

- Market basket analysis generates If-Then scenario rules.
- For example, if X occurs then Y is likely to occur too. If item A is purchased, then item B is likely to be purchased too.
- The rules are derived from the experience. This may be the result of frequencies of co-occurrence of items in past transactions.

The applications of market basket analysis in various domains other than retail are:

1. Medical analytics: Market basket analysis can be used for conditions and symptom analysis. This helps in identifying a profile of illness in a better way.

2. Web usage analytics: FIM approaches can be used with viewing data on websites. The information contained in association rules can be exploited to learn about website browsing of visitor's behavior, developing website structure by making it more effective for visitors, or improving web marketing promotions. The results of this type of analysis can be used to inform website design (how items are grouped together) and to power recommendation engines.

3. Fraud detection and technical dependence analysis: Extract knowledge so that normal behavior patterns may be obtained in illegal transactions from a credit card database in order to detect and prevent fraud.

4 .**Click stream analysis or web link analysis:** Click stream refers to a sequence of web pages viewed by a user. Analysis of clicks is the process of extracting knowledge from web logs. This helps to discover the unknown and potentially interesting patterns useful in the future.

5. Telecommunication services analysis: Market basket analysis can be used to determine the type of services being utilized and the packages customers are purchasing. For example, telecommunication companies can offer TV Internet, and web-services by creating combined offers.

6. Plagiarism detection: It is the process of locating instances of similar content or idea within a work or a document. Plagiarism detection can find similarities among statements that may lead to similar paragraphs if all statements are similar and that possibly lead to similar documents.

Finding Association

- Association rules intend to tell how items of a dataset are associated with each other.
- The concept of association rules was introduced in 1993 for discovering relations between items in sales data of a large retailing company.
- The following examples give rules between items found associated in the sales data of a retailer.

Finding Association(Contd…)

EXAMPLE 6.9

Suggest association rules between items found in the sales data of a retailer, and rules for course choice for a computer science student in college.

SOLUTION

1. {Bread} →{Butter}
   The rule suggests a relationship between the sales of bread and butter. A customer who buys bread also buys butter.
2. {Chocolates} → {a Gift Box}
   The rule suggests a that relationship between the sales of chocolates and empty gift boxes exists. A customer who buys chocolates also buys a gift box.
3. {Java programming} → {advanced web technology} and
   {Python programming} → {Big Data Analytics}
   The rules suggest relationships between Java and advanced web technology, and Python programming and data analytics. Students who opt for Java programming also want to learn advanced web technology, and those who opt for Python programming also opt for Big Data Analytics.
4. {Data Mining} → {Data Visualization}
   The rule may be that 90% of students who select data mining as a major subject will opt for the data visualization course as well.
5. {Computer Graphics, Modeling Techniques} → {Animation}
   The rule may be that students who study computer graphics and modeling techniques courses are likely to choose the course on animation in higher semesters.

Text, Web Content ,Link , and Social Network Analytics

TEXT MINING

- ***Four definitions are:***

1. "Text mining refers to the process of deriving high-quality information from text." (Wikipedia)
2. "Text mining is the process of discovering and extracting knowledge from unstructured data." (National Center of Text Mining The University of Manchester+)
3. "Text mining is the process of analyzing collections of textual contents in order to capture key concepts themes, uncover hidden relationships, and discover the trends without requiring that you know the precise words or terms that authors have used to express those concepts." (IBM2)
4. "Text mining is a technique which helps in revealing the patterns and relationships in large volumes of textual content that are not visible to the naked eye, leading to new business opportunities and improvements in processes." (Amazon BigData Official Blog3)

- Applications of text mining in business domains are predicting stock movements from analysis of company results, decision making for product and innovations developed at the company and contextual advertising.
- Some other applications are (i) mail filtering (spam), (ii) drug action reports (iii) fraud detection (iv) knowledge management, and (iv) social media data analysis.

Areas and Applications of Text Mining

- **Natural Language Processing (NLP)** is a technique for analyzing, understanding and deriving meaning from human language.
- NLP involves the computer's understanding and manipulation of human language.
- NLP algorithms are typically based on ML algorithms. They automatically learn the rules.
- NLP contributes to the field of human computer interaction by enabling several real-world applications such as automatic text summarization, sentiment analysis, topic extraction, named entity recognition, parts-of-speech tagging, relationship extraction and stemming. The common uses of NLP include text mining, machine translation and automated question answering.

Figure 9.1 shows process-pipeline in text-analytics.

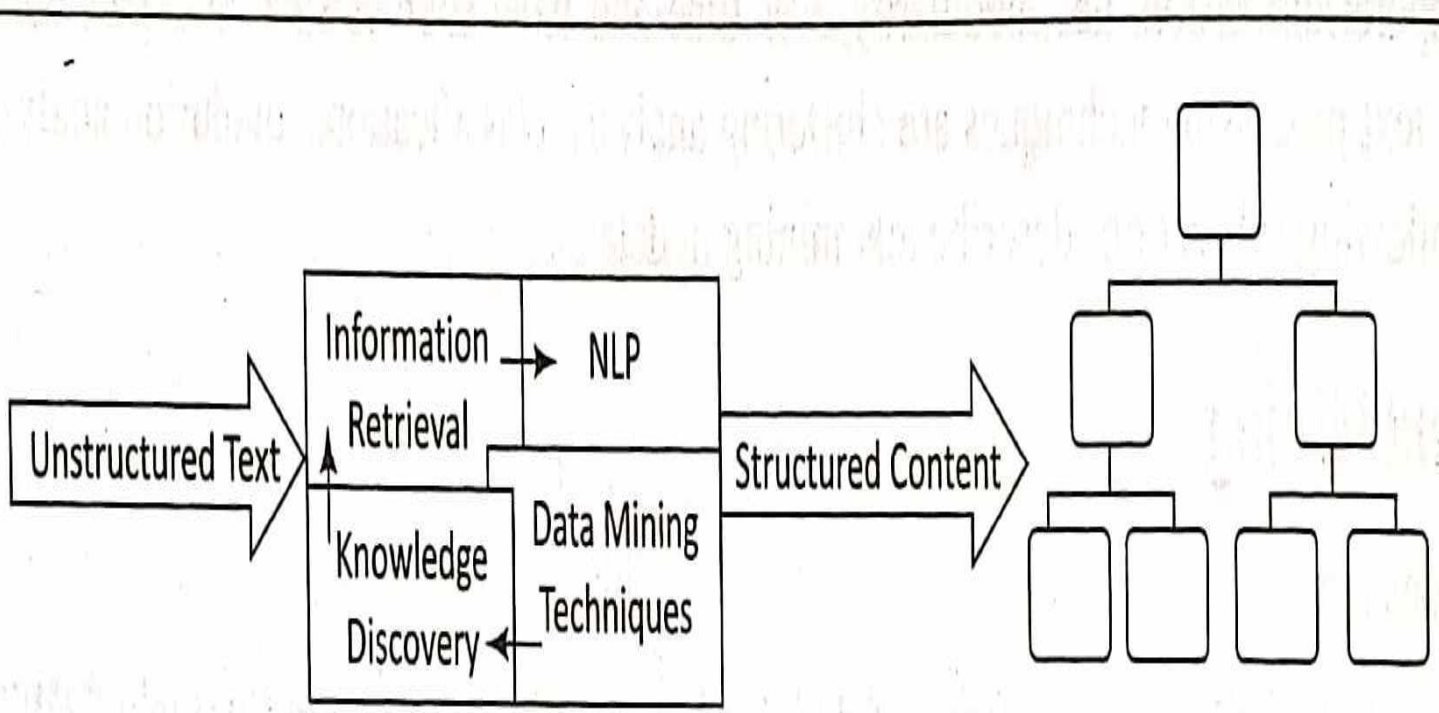

Figure 9.1 Text analytics process pipeline

- **Information Retrieval (IR)** is a process of searching and retrieving a subset of documents from the abundant collection of documents. IR can also be defined as extraction of information required by a user.
- IR is an area derived fundamentally from database technology. One of the most popular applications of IR is searching the information on the web. Search engines provide IR using various advance techniques.
- For example, the crawler program is capable of retrieving information from a wide variety of data sources. Search methods use metadata or full-text indexing.
- **Information Extraction (IE)** is a process in which the software extracts structured information from unstructured and/or semi-structured documents.
- IE finds the relationship within text or desired contents from text. IE ideally derives from machine learning, more specifically from the NLP domain. Content extraction from the images, audio or video is an example of information extraction .

- **Document Clustering** is an application which groups text documents into clusters. Automating document organization, topic extraction and fast information retrieval or filtering use the document clustering method. For example, web document clustering facilitates easy search by users.
- **Document Classification** is an application to classify text documents into classes or categories. The application is useful for publishers, news sites, biogs or areas where lot of contents are present.
- **Web Mining** is an application of data mining techniques. They discover patterns from the web Data Store. The patterns facilitate understanding. They improve the services of web-based applications. Data mining of web usage provides the browsing behavior of a website.
- **Concept Extraction** is an application that deals with the extraction of concept from textual data. Concept extraction is an area of text classification in which words and phrases are classified into a semantically similar group.

Text Mining Process

- Text is most commonly used for information exchange. Unlike data stored in databases, text is unstructured, ambiguous and difficult to process. Text mining is the process that analyzes a text to extract information useful for a specific purpose.
- Syntactically, a text document comprises characters that form words, which can be further combined to generate phrases or sentences. Text mining steps are (i) recognizing, extracting and using the information present in words. Along with searching of words, mining involves search for semantic patterns as well.
- Text mining process consists of a process-pipeline. The pipeline processes execute in several phases. Mining uses the iterative and interactive processes. The processing in pipeline does text mining efficiently and mines the new information.
- Figure 9.2 shows five phases of the process pipeline.

Text Mining Process Phases

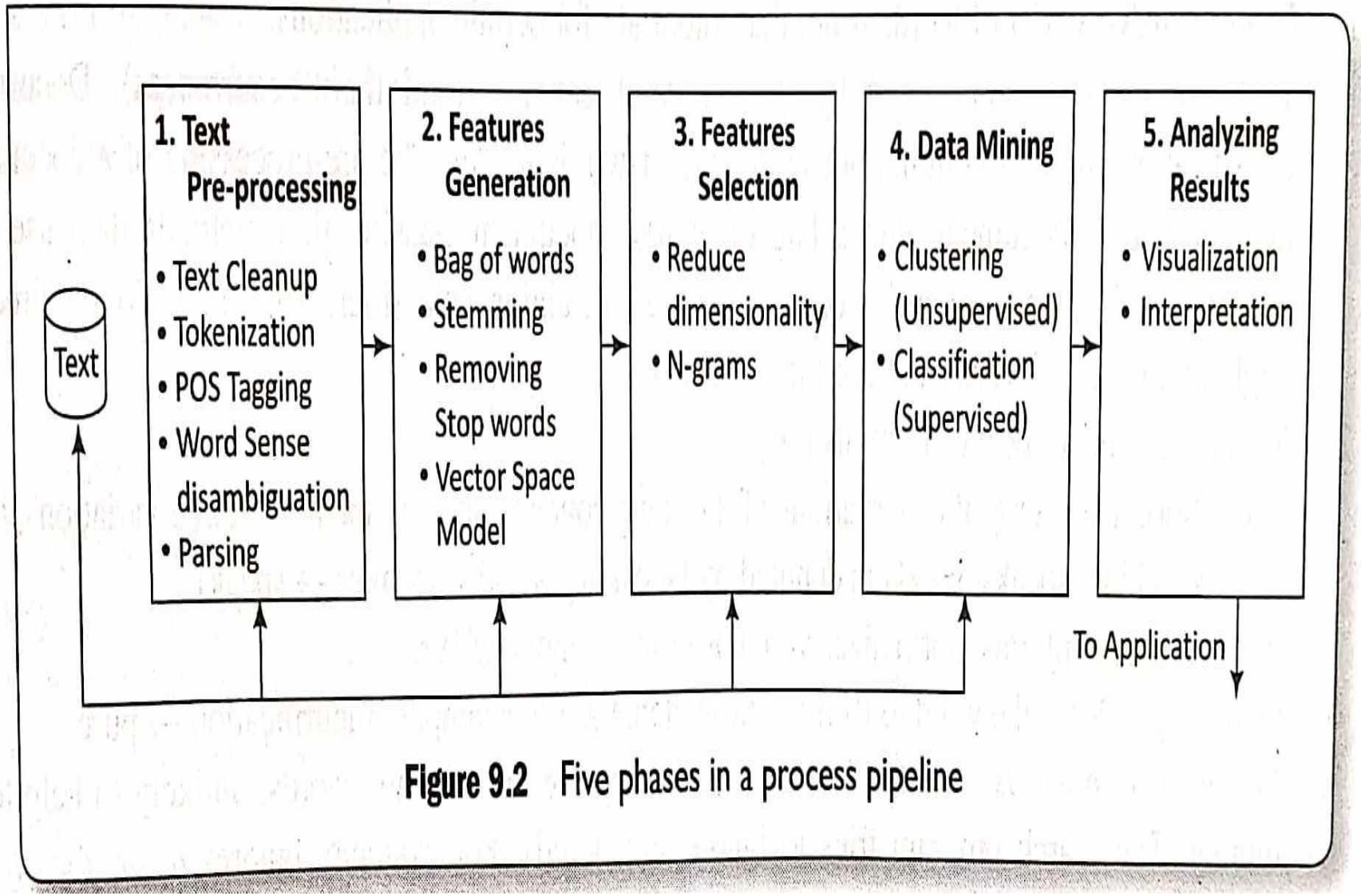

Figure 9.2 Five phases in a process pipeline

Text Mining Process Phases(Contd..)

The five phases for processing text are as follows:

Phase 1: Text pre-processing enables Syntactic/Semantic text-analysis and does the followings:

1. Text cleanup is a process of removing unnecessary or unwanted information. Text cleanup converts the raw data by filling up the missing values, identifies and removes outliers, and resolves the inconsistencies.
2. Tokenization is a process of splitting the cleanup text into tokens (words) using white spaces and punctuation marks as delimiters.
3. Part of Speech (POS) tagging is a method that attempts labeling of each token (word) with an appropriate POS. Tagging helps in recognizing names of people, places, organizations and titles.
4. Word sense disambiguation is a method, which identifies the sense of a word used in a sentence; that gives meaning in case the word has multiple meanings. The methods, which resolve the ambiguity of words can be context or proximity based. Some examples of such words are bear, bank, cell and bass.
5. Parsing is a method, which generates a parse-tree for each sentence. Parsing attempts and infers the precise grammatical relationships between different words in a given sentence.

Text Mining Process Phases(Contd..)

- **Phase 2: Features Generation** is a process which first defines features (variables, predictors). Some of the ways of feature generations are:

1. *Bag of words-Order of words is not that important for certain applications.*

   Text document is represented by the words it contains (and their occurrences). Document classification methods commonly use the bag-of-words model. The pre-processing of a document first provides a document with a bag of words. Document classification methods then use the occurrence (frequency) of each word as a feature for training a classifier. Algorithms do not directly apply on the bag of words, but use the frequencies.

2. Stemming-identifies a word by its root.

   (i) Normalizes or unifies variations of the same concept, such as *speak for three variations, i.e., speaking, speaks, speakers denoted by [speaking, speaks, speaker → speak]*

   (ii) Removes plurals, normalizes verb tenses and remove affixes.

   Stemming reduces the word to its most basic element. For example, impuriflcation → pure.

Text Mining Process Phases(Contd..)

3. Removing stop words from the feature space-they are the common words, unlikely to help text mining. The search program tries to ignore stop words. For example, ignores a, at, for, it, in and are.

4. Vector Space Model (VSM)-is an algebraic model for representing text documents as vector of identifiers, word frequencies or terms in the document index. VSM uses the method of term frequency-inverse document frequency (TF-IDF) and evaluates how important is a word in a document.

- When used in document classification, VSM also refers to the bag-of-words model. This bag of words is required to be converted into a term-vector in VSM. The term vector provides the numeric values corresponding to each term appearing in a document. The term vector is very helpful in feature generation and selection.

Text Mining Process Phases(Contd..)

- **Phase 3: Features Selection** is the process that selects a subset of features by rejecting irrelevant and/or redundant features (variables, predictors or dimension) according to defined criteria. Feature selection process does the following:

1. Dimensionality reduction-Feature selection is one of the methods of division and therefore, dimension reduction. The basic objective is to eliminate irrelevant and redundant data. Redundant features are those, which provide no extra information. Irrelevant features provide no useful or relevant information in any context.
- Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) are dimension reduction methods. Discrimination ability of a feature measures relevancy of features. Correlation helps in finding the redundancy of the feature. Two features are redundant to each other if their values correlate with each other.

2. N-gram evaluation-finding the number of consecutive words of interest and extract them. For example, 2-gram is a two words sequence, ["tasty food", "Good one"]. 3-gram is a three words sequence, ["Crime Investigation Department"].
3. Noise detection and evaluation of outliers methods do the identification of unusual or suspicious items, events or observations from the data set. This step helps in cleaning the data.
- The feature selection algorithm reduces dimensionality that not only improves the performance of learning algorithm but also reduces the storage requirement for a dataset. The process enhances data understanding and its visualization.

Text Mining Process Phases(Contd..)

Phase 4: Data mining techniques enable insights about the structured database that resulted from the previous phases. Examples of techniques are:

1. Unsupervised learning (for example, clustering)

(i) The class labels (categories) of training data are unknown

(ii) Establish the existence of groups or clusters in the data

- Good clustering methods use high intra-cluster similarity and low inter-cluster similarity. Examples of uses - blogs, patterns and trends.

2. Supervised learning (for example, classification)

(i) The training data is labeled indicating the class

(ii) New data is classified based on the training set

Classification is correct when the known label of test sample is identical with the resulting class computed from the classification model.

- Examples of uses are news filtering application, where it is required to automatically assign incoming documents to pre-defined categories; email spam filtering, where it is identified whether incoming email messages are spam or not.
- Example of text classification methods are Naive Bayes Classifier and SVMs.

3. Identifying evolutionary patterns in temporal text streams-the method is useful in a wide range of applications, such as summarizing of events in news articles and extracting the research trends in the scientific literature.

Text Mining Process Phases(Contd..)

Phase 5: Analysing results

(i) Evaluate the outcome of the complete process.

(ii) Interpretation of Result- If acceptable then results obtained can be used as an input for next set of sequences. Else, the result can be discarded, and try to understand what and why the process failed.

(iii) Visualization - Prepare visuals from data, and build a prototype.

(iv) Use the results for further improvement in activities at the enterprise, industry or institution.

Support Vector Machines

- Support vector machines (SVM) is a set of related supervised learning methods (the presence of training data) that analyze data, recognize patterns, classify text, recognize hand-written characters, classify images, as well as bioinformatics and bio sequence analysis.
- A vector has in general n components, x2, x3, ... , xn. A datapoint represents by (Xl, X2, ... , Xn) in n-dimensional space.
- Assume for the sake of simplicity, that a vector has two components, Xl and X2 (Two sets of words in text analysis).

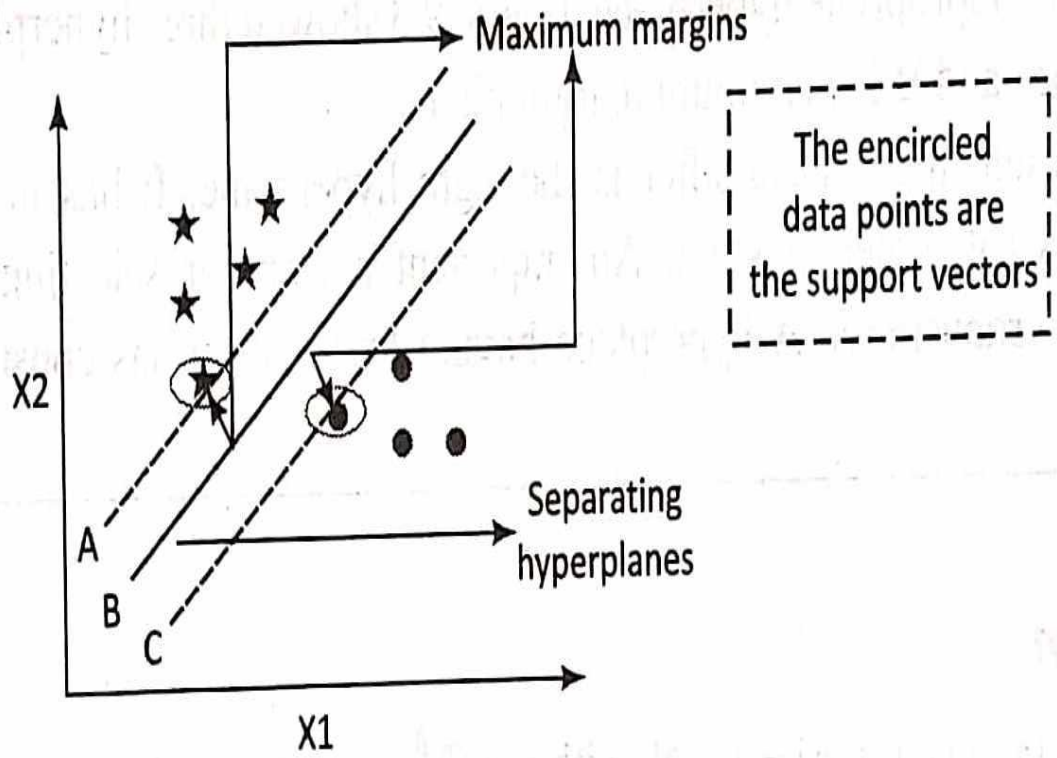

Figure 9. 3 Support vectors, separating hyperplane (B) and margins

WEB MINING, WEB CONTENT AND WEB USAGE ANALYTICS

WEB MINING, WEB CONTENT AND WEB USAGE ANALYTICS

- Web is a collection of interrelated files at web servers. Web data refers to

(i) web content-text, image and records, (ii) web structure-hyperlinks and tags, and (iii) web usage-http logs and application server logs.

Features of web data are:

1. Volume of information and its ready availability
2. Heterogeneity
3. Variety and diversity (Information on almost every topic is available using different forms, such as text, structured tables and lists, images, audio and video.)
4. Mostly semi-structured due to the nested structure of HTML code
5. Hyperlinks among pages within a website, and across different websites
6. Redundant or similar information may be present in several pages
7. Mostly, the web page has multiple sections (divisions), such as main contents of the page, advertisements, navigation panels, common menu for all the pages of a website and copyright notices
8. A web form or HTML form on a web page enables a user to enter data that is sent to a server for processing
9. Website contents are dynamic in nature where information on the web pages constantly changes, and fast information growth takes place such as conversations between users, social media, etc.

Web Mining

- Data Mining is a process of discovering patterns in large datasets to gain knowledge. The process can be shown as [Raw Data - Patterns - Knowledge].
- Web data mining is the mining of web data.
- Web mining methods are in multidisciplinary domains: (i) data mining, ML, natural language, (ii) processing, statistics, databases, information retrieval, and (iii) multimedia and visualization.
- Web consists of rich features and patterns. A challenging task is retrieving interesting content and discovering knowledge from web data.
- Web offers several opportunities and challenges to data mining.

Definition of Web Mining

- Web mining refers to the use of techniques and algorithms that extract knowledge from the web data available in the form of web documents and services. Web mining applications are as follows:

(i) Extracting the fragment from a web document that represents the full web document

(ii) Identifying interesting graph patterns or pre-processing the whole web graph to come up with metrics, such as PageRank

(iii) User identification, session creation, malicious activity detection and filtering, and extracting usage path patterns

Web Mining Taxonomy

- Web mining can broadly be classified into three categories, based on the types of web data to be mined. Three ways are web content mining, web structure mining and web usage mining.
- Figure 9.6 shows the taxonomy of web mining.

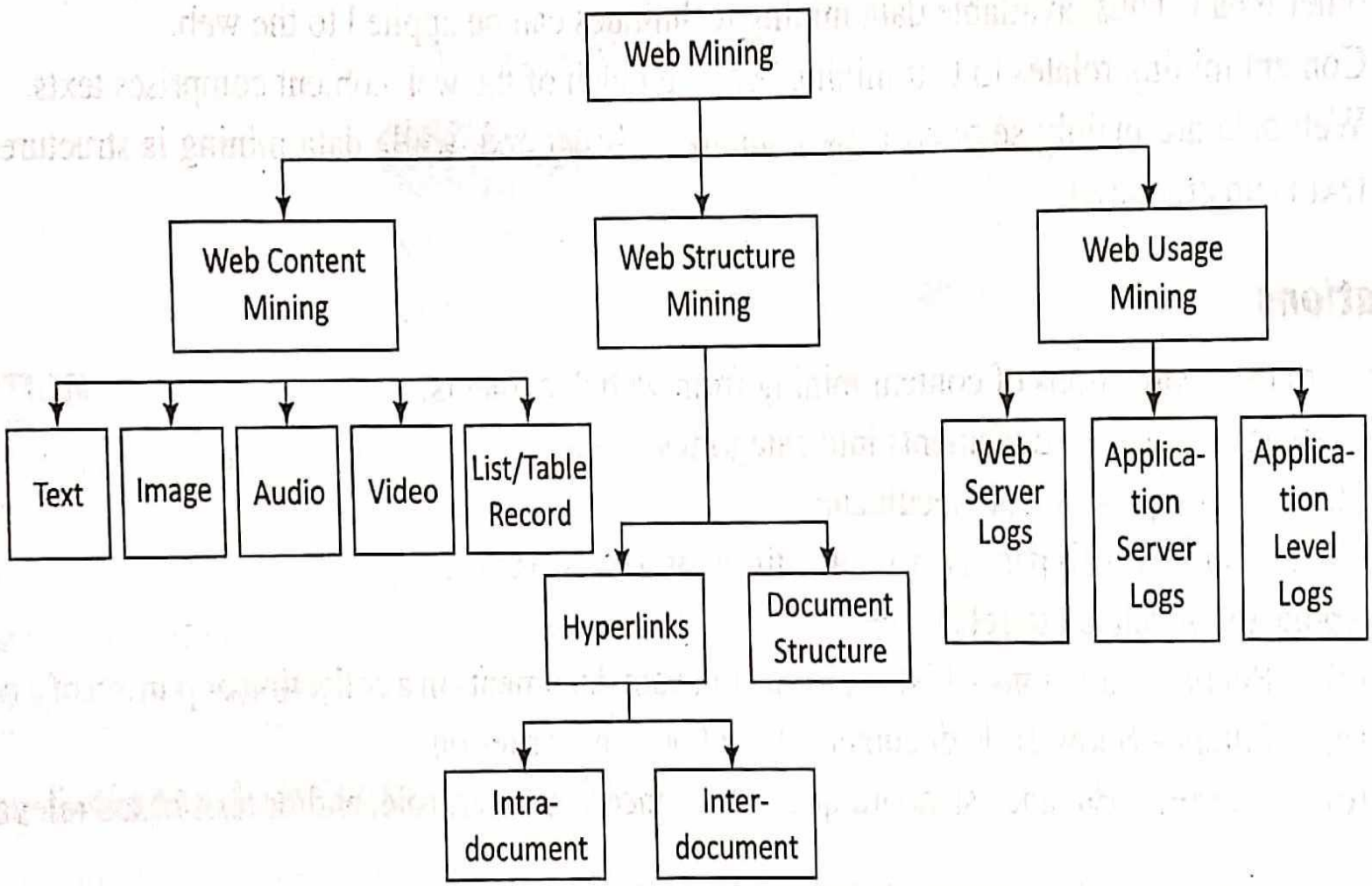

Figure 9.6 Web mining taxonomy

Web Mining Taxonomy(Contd..)

- ***Web content mining*** is the process of extracting useful information from the contents of web documents. The content may consist of text, images, audio, video or structured records, such as lists and tables.
- **Web structure mining** is the process of discovering structure information from the web.
- Based on the kind of structure-information present in the web resources, web structure mining can be divided into:

1. **Hyperlinks:** the structure that connects a location at a web page to a different location, either within the same web page (intra-document hyperlink) or on a different web page (inter-document hyperlink)

2. **Document Structure:** The structure of a typical web graph consists of web pages as nodes, and hyper links as edges connecting the related pages.

Web Mining Taxonomy(Contd..)

- ***Web usage mining*** is the application of data mining techniques which discover interesting usage patterns from web usage data.
- The data contains the identity or origin of web users along with their browsing behavior at a web site. Web usage mining can be classified as:

(i) **Web Server logs:** Collected by the web server and typically include IP address, page reference and access time.

(ii) **Application Server Logs:** Application servers typically maintain their own logging and these logs can be helpful in troubleshooting problems with services.

(iii) **Application Level Logs:** Recording events usually by application software in a certain scope in order to provide an audit trail that can be used to understand the activity of the system and to diagnose problems.

Web Content Mining

- Web Content Mining is the process of information or resource discovery from the content of web documents across the World Wide Web. Web content mining can be

(i) direct mining of the contents of documents or

(ii) mining through search engines. They search fast compared to direct method.

Applications of Web Content Mining

Following are the applications of content mining from web documents:

1. Classifying the web documents into categories
2. Identifying topics of web documents
3. Finding similar web pages across the different web servers
4. Applications related to relevance:

(a) Recommendations - List of top "n" relevant documents in a collection or portion of a collection

(b) Filters - Show/Hide documents based on some criterion

(c) Queries - Enhance standard query relevance with user, role, and/or task-based relevance.

Web Usage Mining

- Web usage mining discovers and analyses the patterns in click streams.
- Web usage mining also includes associated data generated and collected as a consequence of user interactions with web resources.

Phases of Web Usage Mining

Figure 9.7 shows three phases for web usage mining.

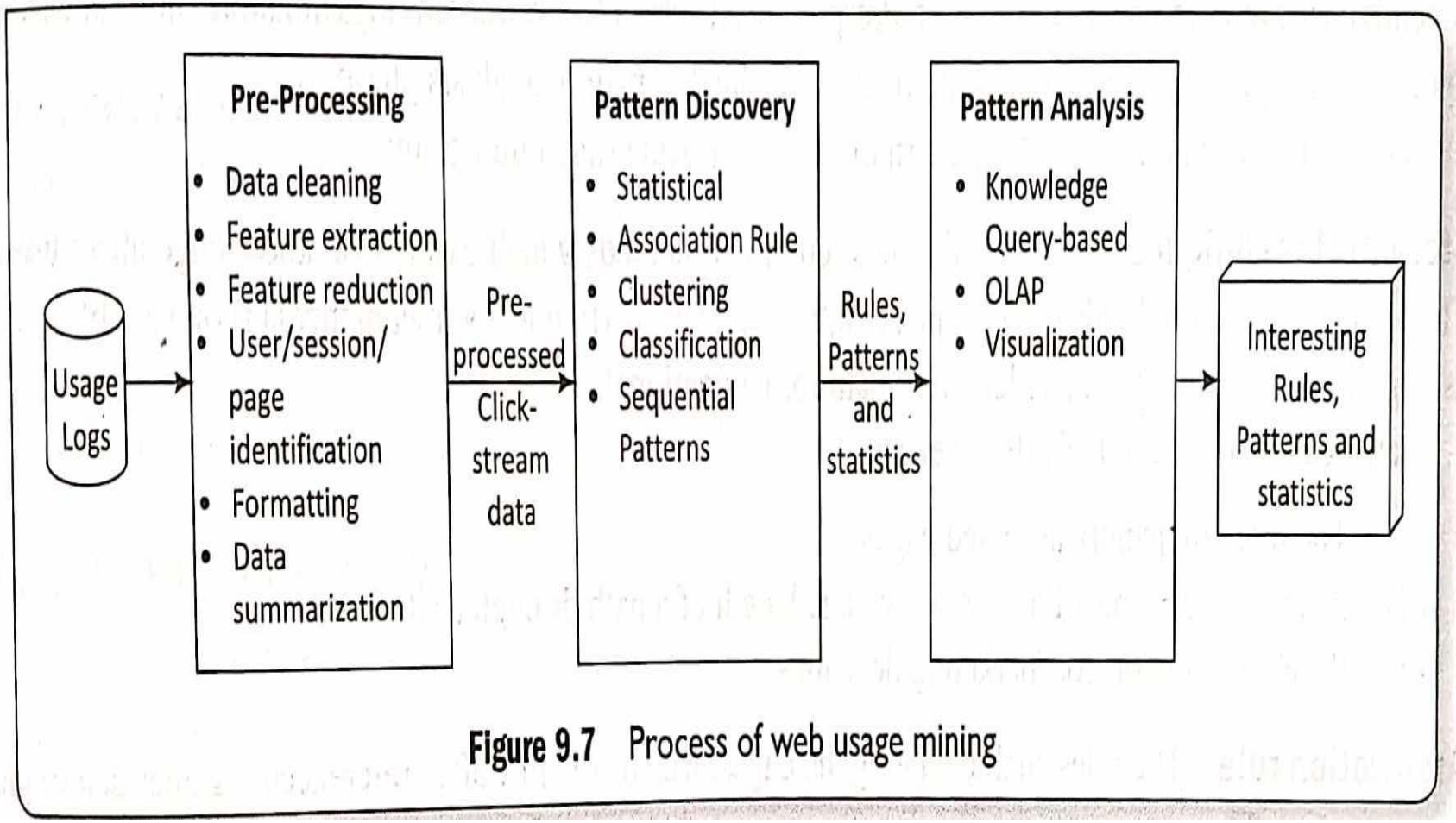

Figure 9.7 Process of web usage mining

Phases of Web Usage Mining (Contd..)

The phases are:

1. **Pre-processing** - Converts the usage information collected from the various data sources into the data abstractions necessary for pattern discovery.
2. **Pattern discovery** - Exploits methods and algorithms developed from fields, such as statistics, data mining, ML and pattern recognition.
3. **Pattern analysis** - Filter outs uninteresting rules or patterns from the set found during the pattern discovery phase.

Phases of Web Usage Mining (Contd..)

- ***Pre-processing***
- The common data mining techniques apply on the results of pre-processing using vector space model .Pre-processing is the data preparation task, which is required to identify:

(i) User through cookies, logins or URL information

(ii) Session of a single user using all the web pages of an application

(iii) Content from server logs to obtain state variables for each active session

(iv) Page references.

- The subsequent phases of web usage mining are closely related to the smooth execution of data preparation task in pre-processing phase. The process deals with (i) extracting of the data, (ii) finding the accuracy of data, (iii) putting the data together from different sources, (iv) transforming the data into the required format and (iv) structure the data as per the input requirements of pattern discovery algorithm.
- Pre-processing involves several steps, such as data cleaning, feature extraction, feature reduction, user identification, session identification, page identification, formatting and finally data summarization

Phases of Web Usage Mining (Contd..)

- **Pattern Discovery**
- The pre-processed data enable the application of knowledge extraction algorithms based on statistics, ML and data mining algorithms.
- Mining algorithms, such as path analysis, association rules, sequential patterns, clustering and classification enable effective processing of web usages. The choice of mining techniques depends on the requirement of the analyst.
- Pre-processed data of the web access logs transform into knowledge to uncover the potential patterns and are further provided to pattern analysis phase.

Phases of Web Usage Mining (Contd..)

- ***Some of the techniques used for pattern discovery of web usage mining are:***
- **Statistical techniques:** They are the most common methods which extract the knowledge about users. They perform different kinds of descriptive statistical analysis (frequency, mean, median) on variables such as page views, viewing time and length of path for navigational.
- Statistical techniques enable discovering:

(i) The most frequently accessed pages

(ii) Average view time of a page or average length of a path through a site

(iii) Providing support for marketing decisions

- **Association rule :** The rules enable relating the pages, which are most often referenced together in a single server session. These pages may not be directly connected to one another using the hyperlinks.
- Other uses of association rule mining are:

(i) Reveal a correlation between users who visited a page containing similar information. For example, a user visited a web page related to admission in an undergraduate course to those who search an eBook related to any subject.

(ii) Provide recommendations to purchase other products. For example, recommend to user who visited a web page related to a book on data analytics, the books on ML and Big Data analytics also.

(iii) Provide help to web designers to restructure their websites.

(iv) Retrieve the documents in prior in order to reduce the access time when loading a page from a remote site.

Phases of Web Usage Mining (Contd..)

- **Clustering** is the technique that groups together a set of items having similar features. Clustering can be used to:

(i)Establish groups of users showing similar browsing behaviors

(ii)Acquire customer sub-groups in e-commerce applications

(iii) Provide personalized web content to users

(iv) Discover groups of pages having related content. This information is valuable for search engines and web assistance providers.

- Thus, user clusters and web-page clusters are two cases in the context of web usage mining. Web page clustering is obtained by grouping pages having similar content. User clustering is obtained by grouping users by their similarity in browsing behavior.

Phases of Web Usage Mining (Contd..)

- **Classification** can be done by using supervised inductive learning algorithms, such as decision tree classifiers, Naive Bayesian classifiers, k-nearest neighbour classifiers, support vector machines.
- **Sequential pattern** discovery User navigation patterns in web usage data gather web page trails that are often visited by users in the order in which pages are visited. Markov Model can be used to model navigational activities in the website. Every page view in this model can be represented as a state.

Phases of Web Usage Mining (Contd..)

- **Pattern Analysis**
- The objective of pattern analysis is to filter out uninteresting rules or patterns from the rules, patterns or statistics obtained in the pattern discovery phase.
- The most common form of pattern analysis consists of:

(i) A knowledge query mechanism such as SQL

(ii) Another method is to load usage data into a data cube in order to perform Online Analytical Processing (OLAP) operations

(iii) Visualization techniques, such as graphing patterns or assigning the colors to different values, can often highlight overall patterns or trends in the data

(iv) Content and structure information can filter out patterns containing pages of a certain usage type, content type or pages that match a certain hyperlink structure.

Web Structure

9.4.2 Web Structure

Web structure models as directed-graphs network-organization. Vertex of the directed graph models an anchor. Let n = number of hyperlinks at the page U. Assume **u** is a vector with elements u_1, u_2, ... u_n. Each page Pg (**u**) has anchors, called hyperlinks. Page Pg (**v**) consists of text document with m number of hyperlinks. v is a vector with elements v_1, v_2, ... v_m. The m is number of hyperlinks at Pg (**v**). A vertex u directs to another Page **V**. A page Pg (**v**) may have number of hyperlinks directed by out-edges to other page Pg (**w**). Consider the following hypotheses:

1. Text at the hyperlink represents the property of a vertex u that describes the destination **V** of the out-going edge.
2. A hyperlink in-between the pages represents the conferring of the authority.

Pages **U** and **U′** hyperlinks u and u′ out-linking to Page **V**. Let Page U has three hyperlinks parenting three Pages, V one, W two, X two, U' one, and Y two, respectively. Figure 9.8 shows a web structure consisting of pages and hyperlinks.

Web Structure(Contd..)

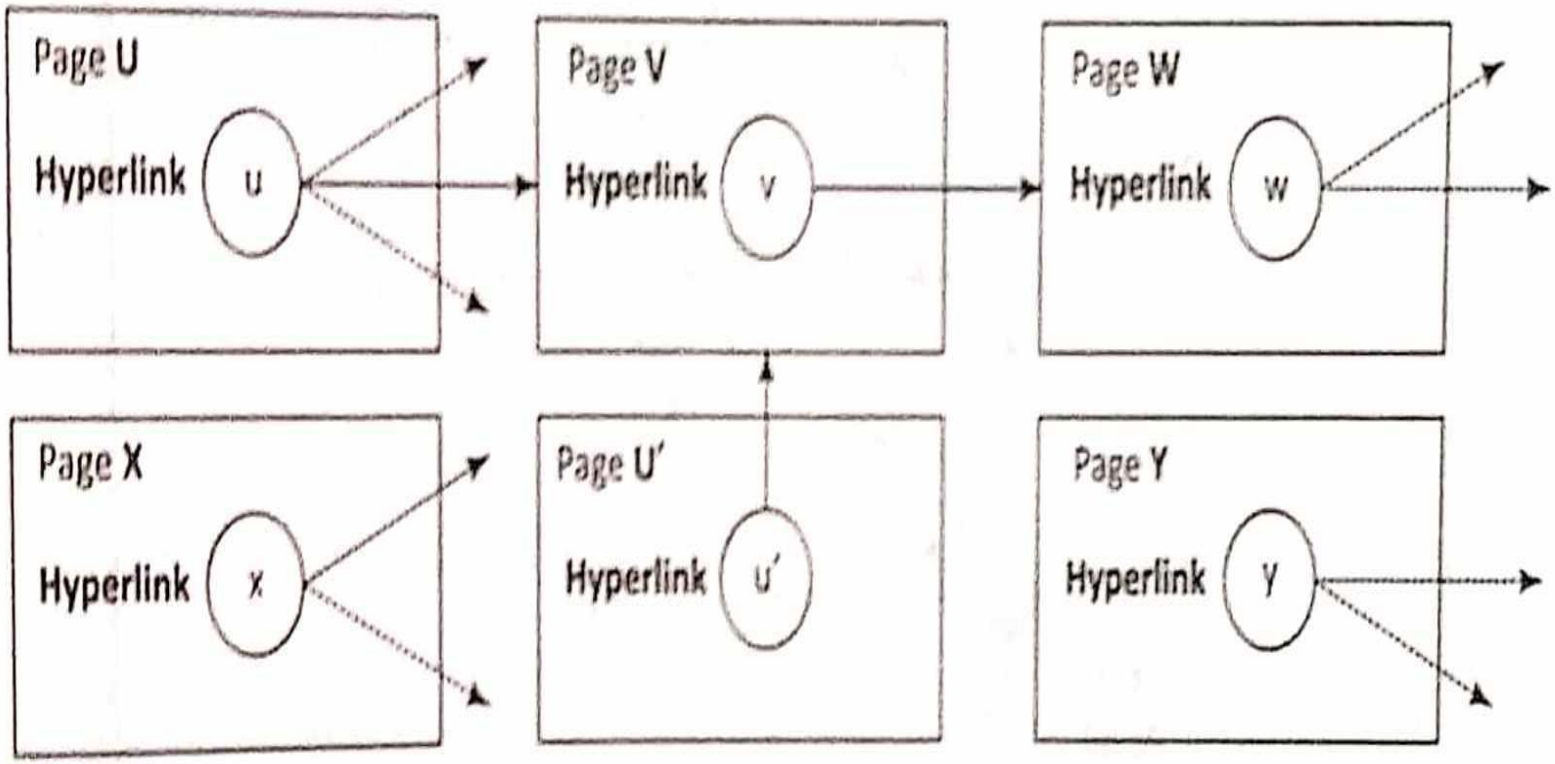

Figure 9.8 Web structure with hyperlinks from a parent to one or more pages

Page Rank Definition

- The in-degree (visibility) of a link is the measure of number of in-links from other links. The out-degree (luminosity) of a link is number of other links to which that link points.

Computation of PageRank and PageRank Iteration

- Assume that a web graph models the web pages. Page hyperlinks are the property of the graph node (vertex).
- Assume a Page, Pg (v) in-links from Pg (u), and Pg (u) out-linking similar to Pg (v), to total Nout [(Pg (u)] pages.
- Figure 9.9 shows Pg (v) in-links from Pg (u) and other pages.

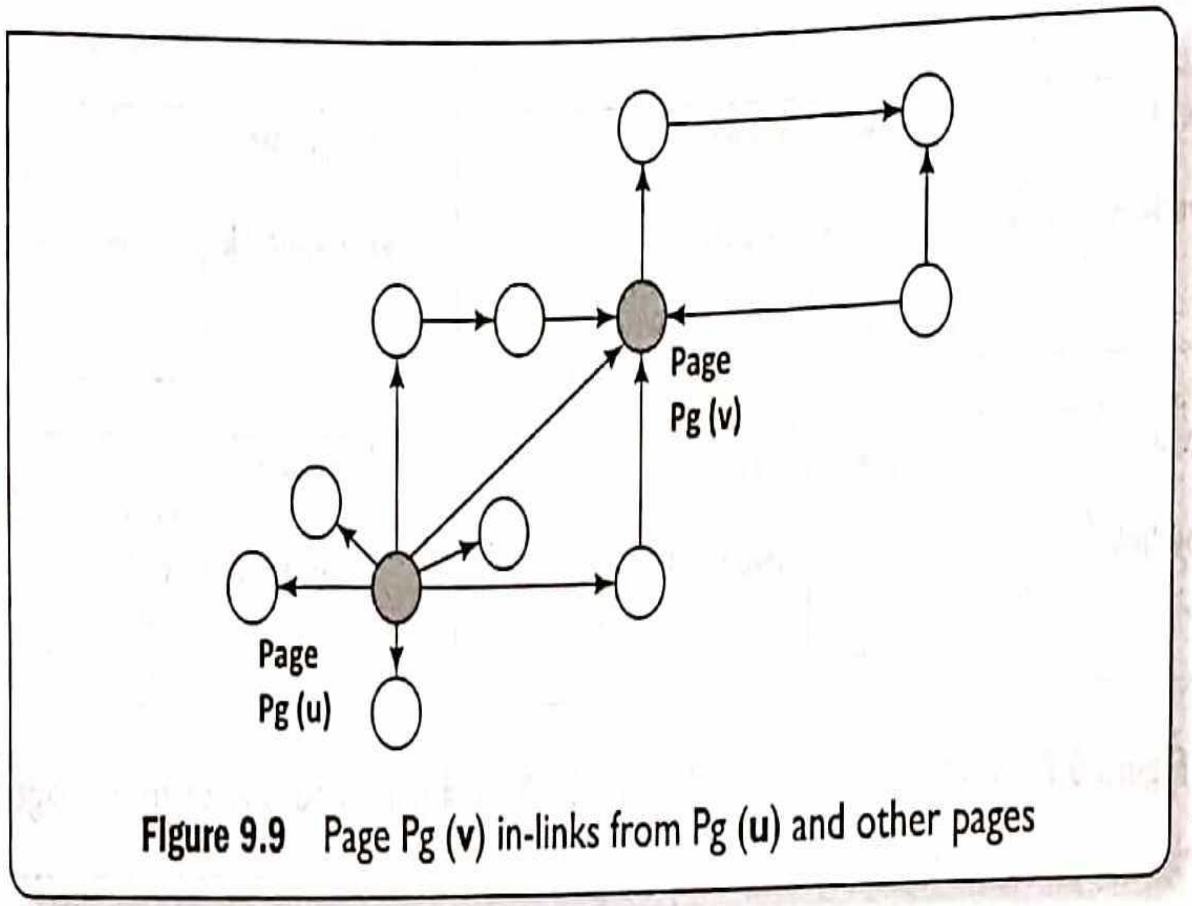

Figure 9.9 Page Pg (**v**) in-links from Pg (**u**) and other pages

N_{out} for page U is 7 and for V is 1 in the figure. Number of in-linking N_{in} for page V is 4. Two algorithms to compute page rank are as follows:

1. PageRank algorithm using the in-degrees as conferring authority

Assume that the page **U**, when out-linking to Page **V** "considers" an equal fraction of its authority to all the pages it points to, such as Pgv. The following equation gives the initially suggested page rank, PR (based on in-degrees) of a page Pgv:

$$PR(Pgv) = nc \cdot \sum_{Pgu:Pgu \to Pgv} \left[PR(Pgu)/N(Pgu)\right] \tag{9.21}$$

where N(Pgu) is the total number of out-links from **U**. Sum is over all Pgv in-links. Normalization constant denotes by nc, such that PR of all pages sums equal to 1.

However, just measuring the in-degree does not account for the authority of the source of a link. Rank is flowing among the multiple sets of the links. When Pgv in-links to a page Pgu, its rank increases and when page Pgu out-links to other new links, it means that N (Pgu) increases, then rank PR(Pgv) sinks (decreases). Eventually, the PR (Pgv) converges to a value.

2. *PageRank algorithm using the relative authority of the parents over linked children*

A method of PageRank considers the entire web in place of local neighbourhood of the pages and considers the relative authority of the parents (children). The algorithm uses the relative authority of the parents (children) and adds a rank for each page from a rank source.

The PageRank method considers assigning weight according to the rank of the parents. Page rank is proportional to the weight of the parent and inversely proportional to the out-links of the parent.

Assume that (i) Page v (Pgv) has in-links with parent Page u (Pgu) and other pages in set PA (v) of parent pages to v that means $\in$ PA(v), (ii) R(v) is PageRank of Pgv, (iii) R (u) is weight (importance/rank) of Pgu, and (iv) ch (u) is weight of child (out-links) of Pgu. Then the following equation gives PageRank R (v) of link v:

$$R(v) = \sum_{u \in PA(v)} \left[R(u) / |ch(u)| \right] \tag{9.25}$$

where PA(v) is a set of links who are parents (in-links) of link v. Sum is over all parents of v. nc is normalization constant whose sum of weights is 1.

Assume that a rank source E exists that is addition to the rank of each page R (v) by a fixed rank value E(v) for Pgv. E(v) is fraction α of [1/|PA(v)|].

An alternative equation is as follows:

$$R(v) = nc \cdot \left\{ (1-\alpha) \sum_{u \in PA(v)} \left[\frac{R(u)}{|ch(u)|} \right] + \alpha \cdot E(v) \right\}, \tag{9.26}$$

where nc = [1/R(v)]. R(v) is iterated and computed for each parent in the set PA(v) till new value of R(v) does not change within the defined margin, say 0.001 in the succeeding iterations.

Hubs and Authorities

- A hub is an index page that out-links to a number of content pages. A content page is topic authority. An authority is a page that has recognition due to its useful, reliable and significant information.
- Figure 9.10(a) shows hubs (shaded circles) with the number of out-links associated with each hub.
- Figure 9.10(b) shows authorities (dotted circles) with the number of in-links and out-links associated with each link.

Hubs and Authorities(Contd..)

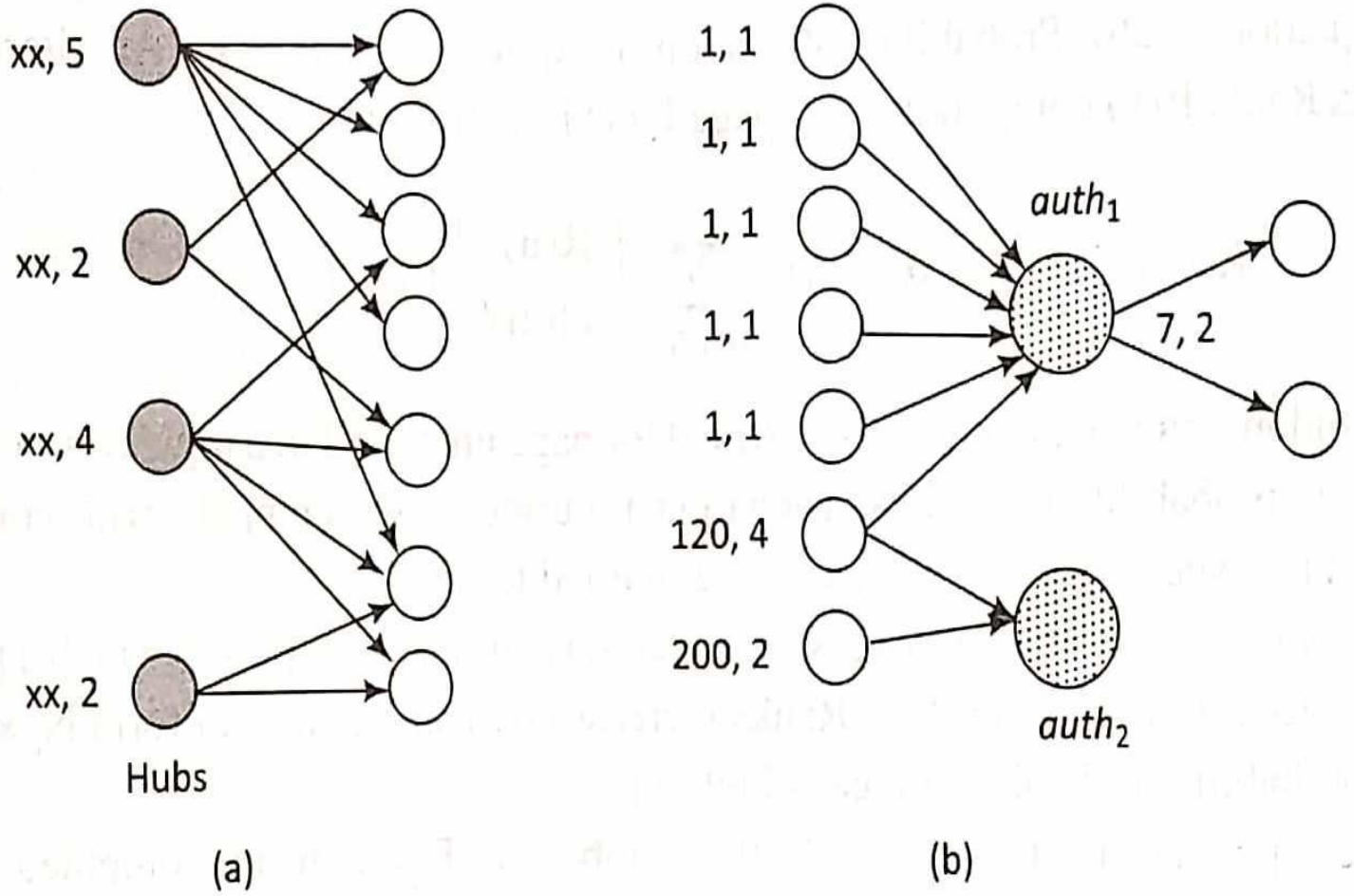

Figure 9.10 (a) Hubs (shaded circles) and (b) Authorities (dotted circles)

Hubs and Authorities(Contd..)

- In-degrees (number of in-edges from other vertices) can be one of the measures for the authority. However, in-degrees do not distinguish between an in-link from a greater authority or lesser authority.
- Authority, auth1 in Figure 9.10(b) has in-links from 6 vertices (in-degrees = 6) and auth2 has in-links to just 2 (in-degree = 2). However, auth1 has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125). Authority, auth2 has links with two vertices with in-degrees= 120 and 200 (total= 220). Auth2has association with greater authorities. Therefore, in-degrees may not be a good measure as compared to authority.

SOCIAL NETWORKS AS GRAPHS AND SOCIAL NETWORK ANALYTICS

SOCIAL NETWORKS

- A social network is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of inter-dependency, such as friendship, kinship, financial exchange, dislike or relationships of beliefs, knowledge or prestige. *(Wikipedia)*
- Social networking is the grouping of individuals into specific groups, like small rural communities or some other neighbourhoods based on a requirement. The following subsections describe social networks as graph, uses, characteristics and metrics.

Counting Triangles and Graph Matches

- One of the methods of detecting communities is counting of triangles. A triangle means three vertices forming a triangle with edges interconnecting them.
- Triangle count refers to the number of triangles passing through each vertex. The count is a measure of clustering. A vertex is part of a triangle when it has two adjacent vertices with an edge between them.
- Graph matches are computed using filtering or search algorithm, which uses the properties, labels of vertices, edges or the geographic locations.
- Figure 9.14 shows triangles and triangles between similar graph properties found from graph matches. Edge labels show the GPAs of students socially connected.

Counting Triangles and Graph Matches(Contd..)

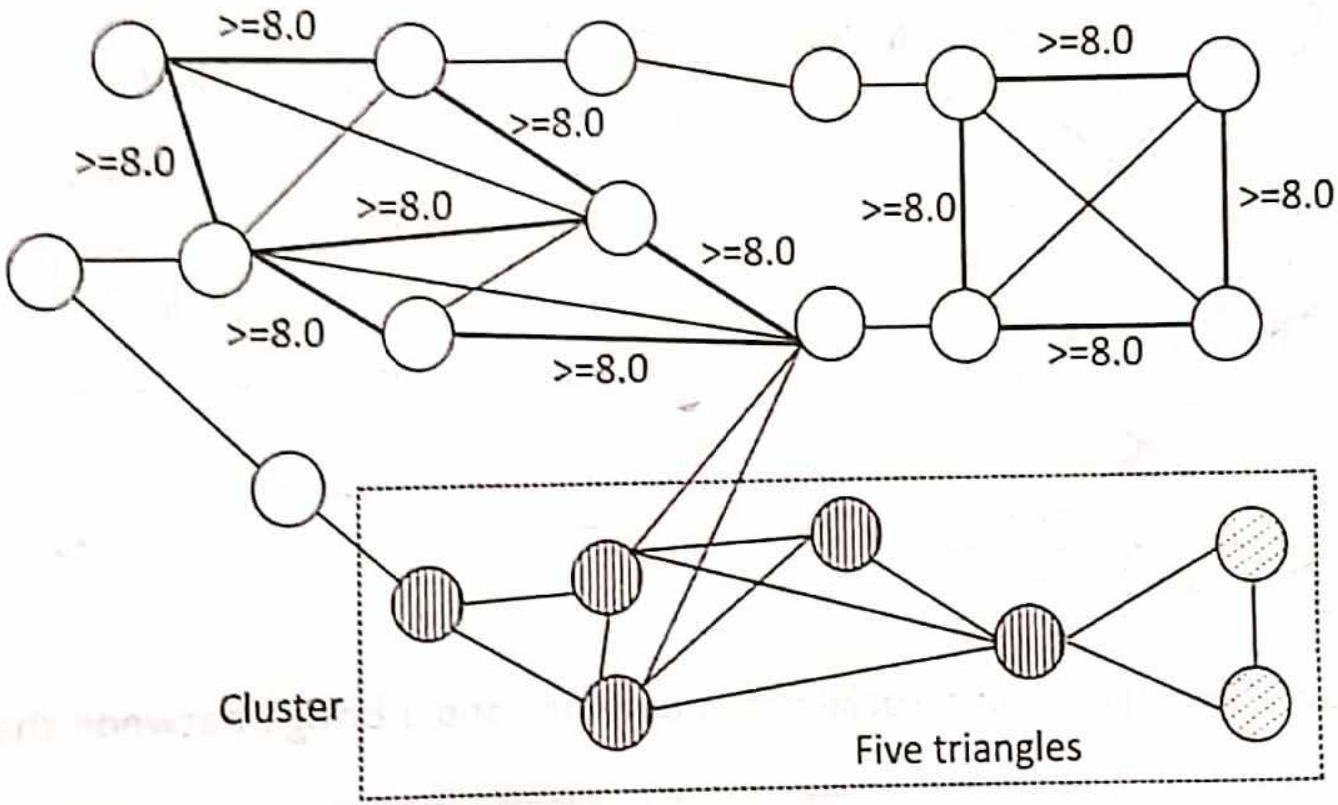

Figure 9.14 Clustering of five triangles and three matches of graphs

Direct Discovery of Communities

- Three metrics identify groups and communities from a social graph:

1. Cliques - A clique forms by a set of vertices when each of the vertices directly connects to every other individual vertex through the edges. Detecting the cliques leads to direct discovery of communities.
2. Structurally cohesive blocks.
3. Social circles from connections and neighbourhoods

- A bridge enables the link between two groups. Application of analyzing communities, Sim Ranks and bridges are finding a set of experts, specific areas of expertise, and ranking the expertise in an organization.
- Experience in social science fields shows that the social network of a person is the key indicator of the stature of the person and his/her success potential. Social graph analysis enables finding key bridges and persons with most connections.

- Figure 9.15 shows a social graph with two cliques and a bridge.

Direct Discovery of Communities(Contd..)

Figure 9.15 shows a social graph with two cliques and a bridge.

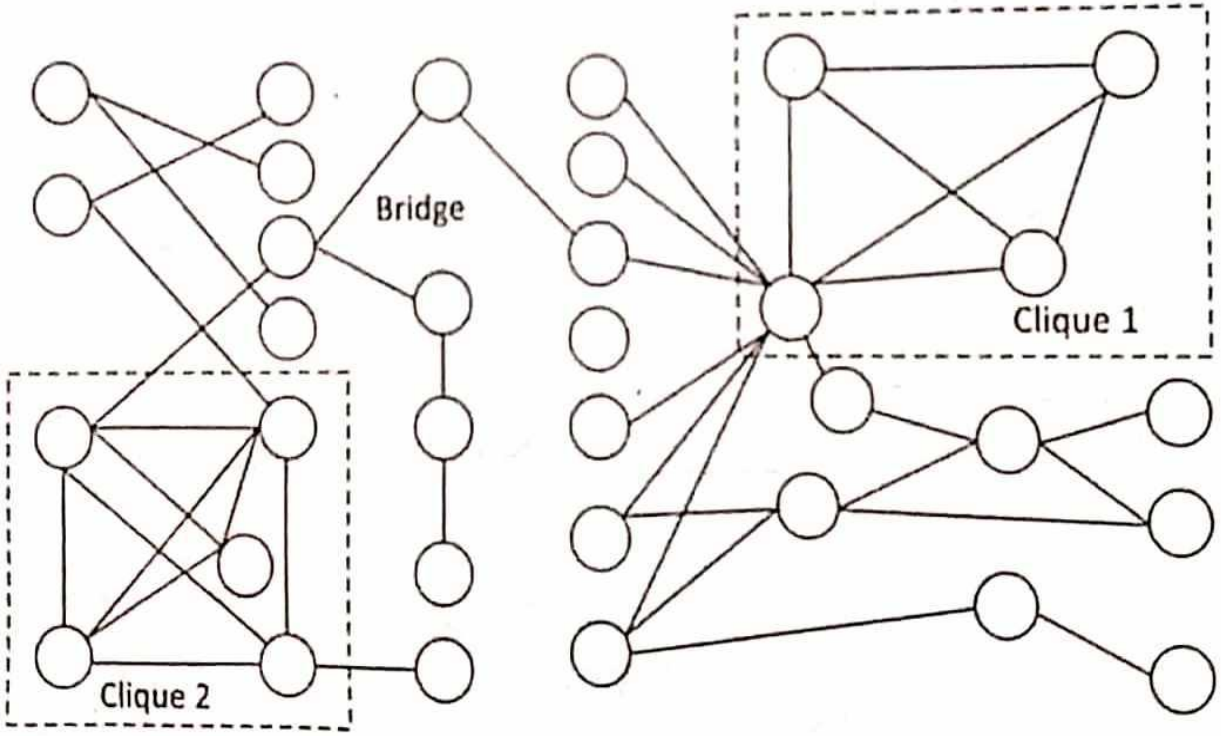

Figure 9.15 Two cliques in a social graph network and a bridge between the cliques