Module 5
Machine Learning Algorithms for Big Data Analytics, Text, Web Content, Link, and Social Network Analytics

6.1 Introduction to Machine learning:-

Analytics uses the mathematical equations, formulae and models. Analytics also uses the statistics, AI, ML and DL, and predict the behaviour of entities, objects and events. Statistics refers to studying organization, analysis of a collection of data, making interpretations and presentation of analyzed results.

Artificial Intelligence (AI) refers to the science and engineering of making computers perform tasks, which normally require human intelligence. For example, tasks such as predicting future results, visual perception, speech recognition, decision making and natural language processing. Two concepts in AI, 'machine learning' and 'deep learning' provide powerful tools for advanced analytics and predictions.

Machine Learning –Machine Learning (ML) is a field of computer science based on AI which deals with learning from data in three phases, i.e. collect, analyze and predict. It does not rely on explicitly programmed instructions.

6.2 Estimating the Relationships, Outliers, Variances, Probability Distributions and Correlations :-

Methods of studying relationships use variables. Types of variables used are as follows:

Independent variables represent directly measurable characteristics. For example, year of sales figure or semester of study.

Dependent variables represent the characteristics. For example, profit during successive years or grades awarded in successive semesters. Values of a dependent variable depend on the value of the independent variable.

Predictor variable is an independent variable, which computes a dependent variable using some equation, function or graph, and does a prediction. For example, predicts sales growth of a car model after five years from given input datasets for the sales, or predicts sentiments about higher sales of particular category of toys next year.

Outcome variable represents the effect of manipulation(s) using a function, equation or experiment. For example, CGPA (Cumulative Grade Points Average) of the student or share of profit to each shareholder in a year using profit as the dependent variable. CGPA of a student computes from the grades awarded in the semesters for which student completes his/her studies..

Explanatory variable is an independent variable, which explains the behavior of the dependent variable, such as linearity coefficient, non-linear parameters or probabilistic distribution of profit-growth as a function of additional investment in successive years.

Response variable is a dependent variable on which a study, experiment or computation focuses. For example, improvement in profits over the years from the investments made in successive years or improvement in class performance is measured from the extra teaching efforts on individual students of a class.

Feature variable is a variable representing a characteristic. For example, apple feature red, pink, maroon, yellowish, yellowish green and green. Feature variables are generally represented by text characters. Numbers can also represent features. For example, red with 1, orange with 2, yellow with 3, yellowish green 4 and green 5.

Categorical variable is a variable representing a category. For example, car, tractor and truck belong to the same category, i.e., a four-wheeler automobile. Categorical variables are generally represented by text characters.

6.2.1 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.

A data point is (xi, yi) when dependent variable value = yi at the independent variable value = xi. The i = 1, 2… n for number of data points = n. The i varies with the position of projection of the point on X-axis. Scatter plot represents data points by dots. The dot can also be a bubble, triangle, circle, cross or vertical bar. Size or colour of dot distinguishes the dependent variables on the same plot.

Another method is quantifying two or more dependent variables by columns of different widths with filled colours, shades or patterns. The width quantifies the dependent variable. The column-position quantifies the independent variable. Examples of dependent variables are sales of five car models in a year, grades in five courses taken in a semester.

6.2.1.1 Linear and Non-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.

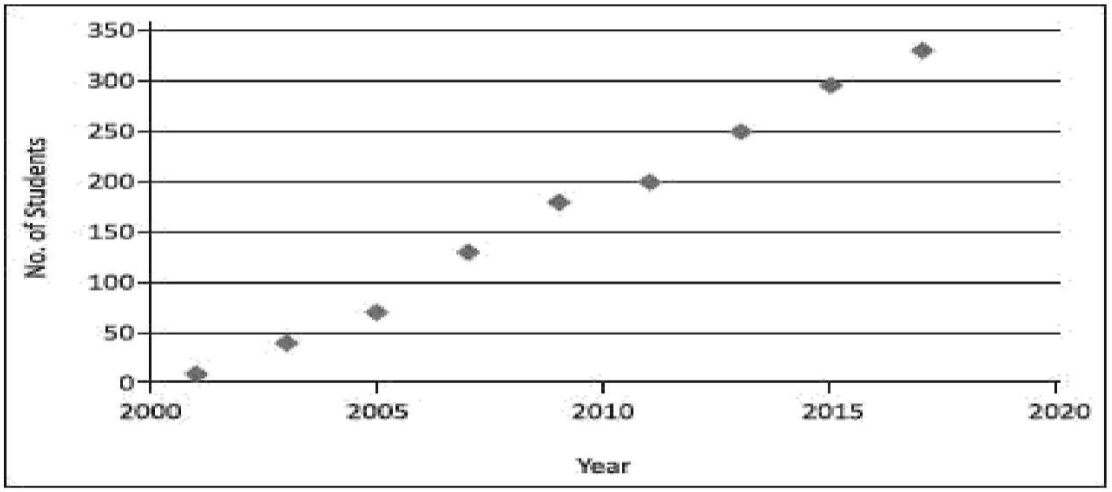


Figure 6.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017

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.

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. 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.

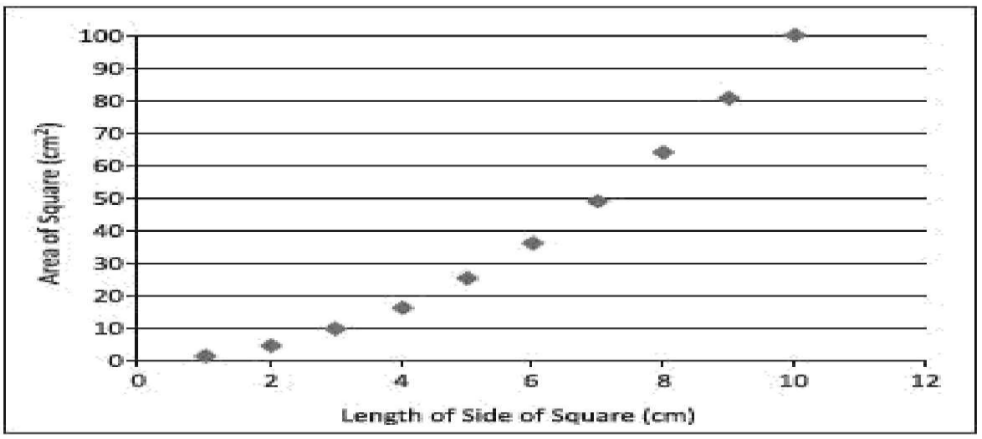


Figure 6.2 shows a scatter plot in case of a non-linear relationship between side of square and its area..

6.2.2 Estimating the Relationships

Estimating the relationships means finding a mathematical expression, which gives the value of the variable according to its relationship with other variables. For example, assume ym = sales of a car model m in xth year of the start of manufacturing that model. Assume that computations show that the ym relates by a mathematical expression (ym= a0 + a1.xm + a2.xm2) up to an acceptable degree of accuracy, when a0 = 490, a1 = 10 and a2 = 5. Estimated first year sales, ym(1) = (490 + 10) = 500, second year ym(2) = (490 + 10 × 2 + 5 × 22) = 530, third year ym(3) = (490 + 10 × 3 + 5 × 32) = 565, if fit with the desired accuracy, then the results are showing that the expression of ym estimates the relationship between model m sales in next and other years. The ym can also predict the sales in 6th or later years. Predictions are up to a certain degree of certainty.

6.2.3 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. 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. 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).

Note: Population means any group of data, which includes all the data of interest. For example, when analysing 1000 students who gave an examination in a computer course, then the population is 1000. 100 games of chess will represent the population in analysis of 100 games of chess of a grandmaster. Sample means a subset of the population. Sample represents the population for uses, such as analysis and consists of randomly selected data.

6.2.4 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 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. **No variance** is sometimes hard to understand in real datasets.

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: Standard deviation, denoted by s, is the square root of the variance. The s 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 s 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: $\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)^2}$ (6.1a)

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

where N is number of data points in population, S is number in the sample, m is expected in the population or average value of x, and $\bar{x}$ 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 s_{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 s_{est}:

1. When s_{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 s_{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.

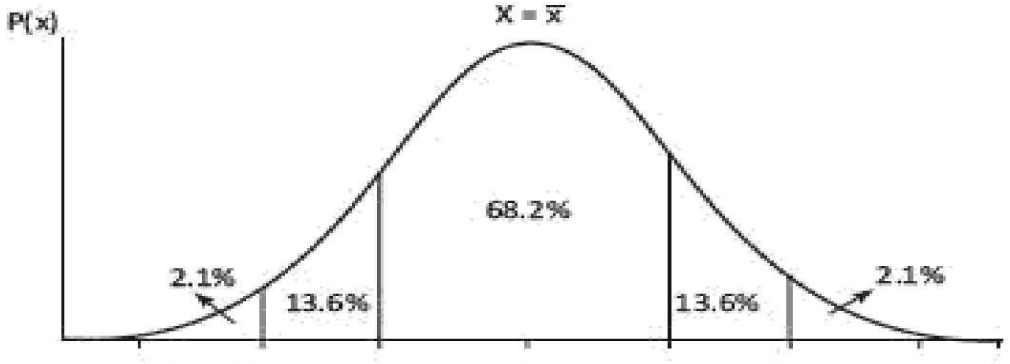
6.2.5 Probabilistic Distribution of Variables, Items or Entities

Probability is the chance of observing a dependent variable value with respect to some independent variable. Suppose a Grandmaster in chess has won 22 out of 100 games, drawn 78 times, and lost none. Then, probability P of winning Pw is 0.22, P of drawn game PD is 0.78 and P of losing, PL = 0. The sum of the probabilities is normalized to 1, as only one of the three possibilities exist.

Probability distribution is the distribution of P values as a function of all possible independent values, variables, situations, distances or variables. For example, if P is given by a function P(x), then P varies as x changes. Variations in P(x) with x can be discrete or continuous. The values of P are normalized such that sum of all P values is 1. Assuming distribution is around the expected value the standard normal distribution formula is:

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}$$

Normal distribution relates to Gaussian function. Figure 6.3 below shows a PDF with normal distribution around x = ¯x standard deviation = s and variance = s2.



Moments (0, 1, 2 ...) refer to the expected values to the power of (0, 1, 2,) of random variable variance (Section 6.2.5.3). The variance is the second central moment of a distribution, which equals to the square of the standard deviation, and the covariance of the random variable with itself, and it is often represented by s2 or var (x). The variance is computed as follows:

$$\sigma^2 = \frac{\sum(x_i - \bar{x})^2}{N}$$

Assume that probability distribution (PDF) is normal, called Gaussian distribution, which is like a bell-shaped curve (Figure 6.3). The PDF of the normal distribution is such that 68% of area under the PDF is within (+ s) and (– s), 95% of area under the PDF is within ($\bar{x}$ + 2s) and (– 2s) and 99.7% is within(+ 3s) and (– 3s).

6.2.5.1 Kernel Functions .

A probability or weight can be represented by a kernel function1 like a Gaussian or tri-cube function. Kernel function is a function which is a central or key part of another function. For example, Gaussian kernel function is the key part of the probability distribution function. Figure 6.3 shows the probability normal distribution, which is a Gaussian function based on the Gaussian kernel function.

A kernel function[1], K* defines as

$$K^*(u) = \lambda . K(\lambda . u), \qquad (6.6a)$$

where $\lambda > 0$. Gaussian kernel function is

$$K^*(x) = \left[\frac{1}{(\sqrt{2\pi})}\right] e^{\left[-\frac{u^2}{2^2}\right]}, \qquad (6.6b)$$

and when $u = \sigma$ the distribution function is proportional to

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}}.$$

$\lambda = \left(\frac{1}{\sigma\sqrt{2}}\right)$ in Equation (6.3).

Tricube kernel function is:

$$K^*(u) = (70/81)\,(1 - |u|^3)^3 \lambda . K(\lambda . u), \qquad (6.6c)$$

where $|u| \leq 1$.

6.2.5.2 Moments

Moments (0, 1, 2, ...) refer to expected values to the powers of (0, 1, 2 ...) of random variable variance. 0^{th} moment is 1, 1^{st} moment = $E(x) = \bar{x}$, (expected value), 2^{nd} moment is squared $V[(x_i - \bar{x})^2]$ = sum of product of $(x_i - \bar{x})^2$, and $P(x = x_i)$.

6.2.5.3 Unequal Variance Welch's t-test

A test in statistics is unequal-variance t test, also called Welch t test. (i) The test assumes that two groups of data are sampled data which consist of Gaussian distributed populations (Equation (6.3)). (ii) The test does not assume those two populations have the same standard deviation.

Unequal variances t-test is a two-sample location test. It tests the hypothesis that two populations have equal means. (Hypothesis means making assumption statements about certain characteristics of the population. For example, an assumption that most students of a specific professor will excel as a programmer. Hypothesis when tested for a decade may pass or fail depending up on whether the statistically significant results show that the students of that professor really excelled as programmers.) Welch's t-test is an adaptation of student's t-test in statistics. The t-test is more reliable when the two samples have unequal variances and unequal sample sizes.

6.2.5.4 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. 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 (ANOVA) 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) 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.

F-test : F-test requires two estimates of population variance— one based on variance between the samples and the other based on variance within the samples. These two estimates are then compared for F-test:

$$F = \frac{E1(V)}{E2(V)}$$

where E1(V) is an estimate of population variance between the two samples and E2(V) is an estimate of population variance within the two samples.

6.2.5.5 No Relationship Case

Relationships between variables need to be studied and analyzed before drawing conclusions based on it. One cannot determine the right conclusion or association when no relationship between the variables exists.

6.2.6 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.
R-Square: R is a measure of correlation between the predicted values y and the observed values of x. R-squared (R2) is a goodness-of-fit measure in linear-regression model. It is also known as the coefficient of determination. R2 is the square of R, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.
Interpretation of R-squared The larger the R2, the better the regression model fits the observations, i.e., the correlation is better. Theoretically, if a model shows 100% variance, then the fitted values are always equal to the observed values, and therefore, all the data points would fall on the fitted regression line. Correlation differs from a regression analysis.
Regression analysis predicts the value of the dependent predictor or response variable based on the known value of the independent variable, assuming a more or less mathematical relationship between two or more variables within the specified variances.

6.2.6.1 Correlation Indicators of Linear Relationships

Correlation is a statistical technique that measures and describes the ‘strength’ and ‘direction’ of the relationship between two variables.
The significant questions are: Does y increase or decrease with x? For example, expenditure increases with income or does the number of patients decrease with proper medication. (Direction)

(i) Suppose y does increase with x; then, how fast? (ii) Is this relationship strong? (iii) Can reliable predictions be made? That is, if one tells the income, can the expenditure be predicted?

Relationships and correlations enable training model on sample data using statistical or ML algorithms. Statistical correlation is measured by the coefficient of correlation. The most common correlation coefficient, called the Pearson product-moment correlation coefficient. It measures the strength of the linear association between variables. The correlation r between the two variables x and y is:

$$r = \left[\frac{1}{(n-1)}\right] \times \sum \left\{ \left[\frac{(x_i - \bar{x})}{\sigma_x}\right] \times \left[\frac{(y_i - \bar{y})}{\sigma_y}\right] \right\},$$

where n is the number of observations in the sample, xi is the x value for observation i, x– is the sample mean of x, yi is the y value for observation i, y– is the sample mean of y, sx is the sample standard deviation of x, and sy is the sample standard deviation of y.
Summation is over all n values of i, i = 1, 2, ..., n.
Use of Statistical Correlation Assume one sample dataset is {u1, ..., un} containing n values of a parameter r. The ru,i is i-th data point in dataset u. (i = 1, 2, …, n). Another sample dataset is {v1, ...,

vn} containing n values of r. rv,i is i-th data point in dataset v. Let the correlation among two samples is being measured. Sample Pearson correlation metric cr measures how well two sample datasets fit on a straight line.

$$c_r(u,v) = \frac{\sum_i (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_i (r_{u,i} - \overline{r_u})^2 \sum_i (r_{v,i} - \overline{r_v})^2}}$$

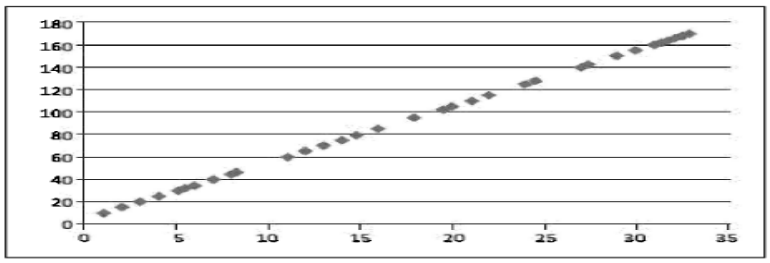
where the summations are over the values of parameter in the datasets.

Three other similarities based on correlation are: (i) Constrained Pearson correlation – It is a variation of Pearson correlation that uses midpoint instead of mean rate. (ii) Spearman rank correlation – It is similar to Pearson correlation, except that the ratings are ranks. (iii) Kendall's G correlation – It is similar to the Spearman rank correlation, but instead of using ranks themselves, only the relative ranks are used to calculate the correlation.

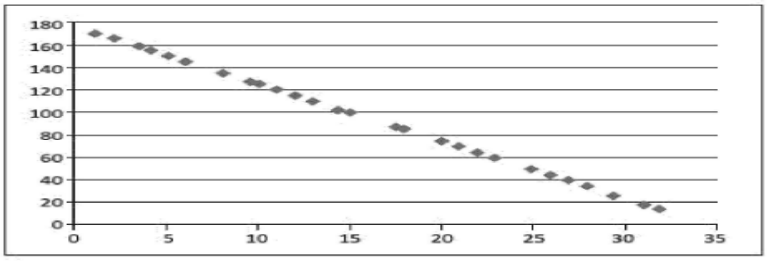
Table 6.1 gives rough guidelines on the strength of the relationship.

Value of r	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

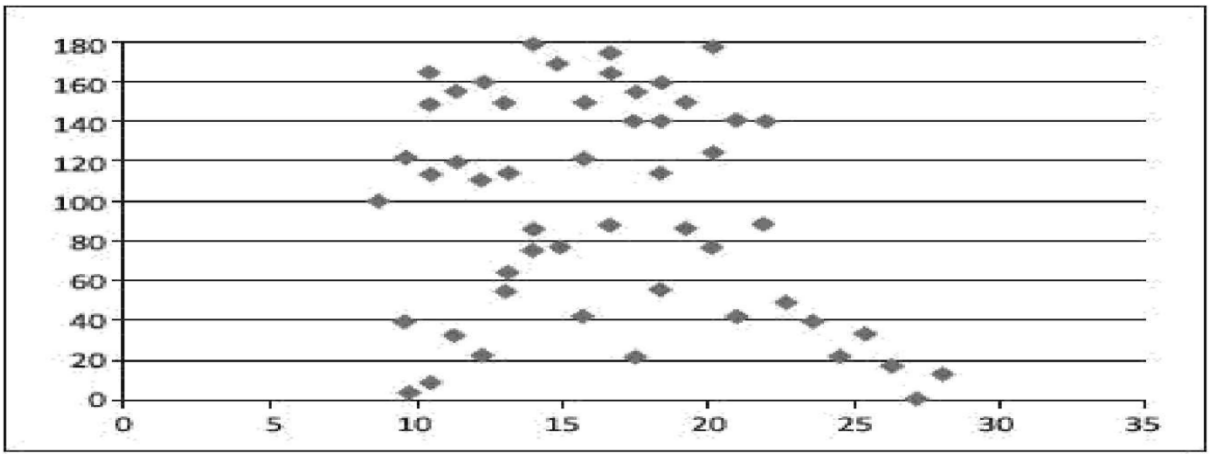
Correlation is only appropriate for examining the relationship between meaningful quantifiable data (such as, temperature, marks, score) rather than categorical data, such as gender, color etc. Figure 6.4 shows perfect and imperfect, linear positive and negative relationships, and the strength and direction of the relationship between variables



Perfect Positive Linear Relationship (r = 1)



Perfect Negative Linear Relationship (r = −1)



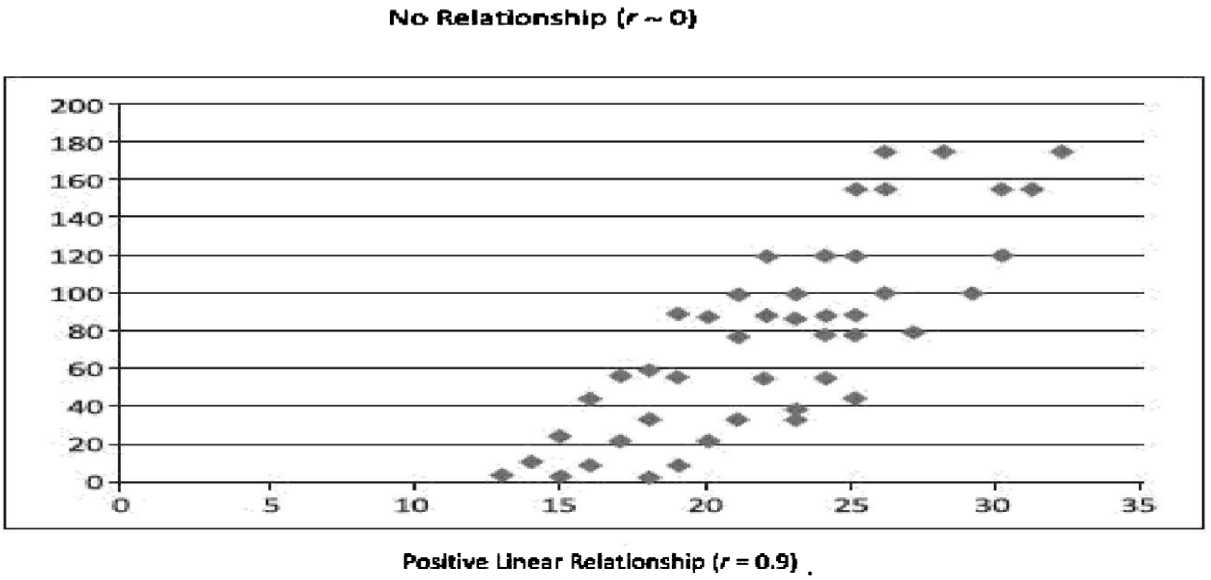


Figure 6.4

6.3 Regression Analysis

Regression analysis is a set of statistical steps, which estimate the relationships among 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. Non-linear regression equation is as follows:

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3,$$

where number of terms on the right-hand side are 3 or 4.

Linear regression means only the first two terms are considered.

6.3.1 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.

$$y = f(a_0, a_1) = a_0 + a_1 x,$$

where a0 is a constant and a1 is the linearity coefficient. Simple linear regression is performed when the requirement is prediction of values of one variable, with given values of another variable.

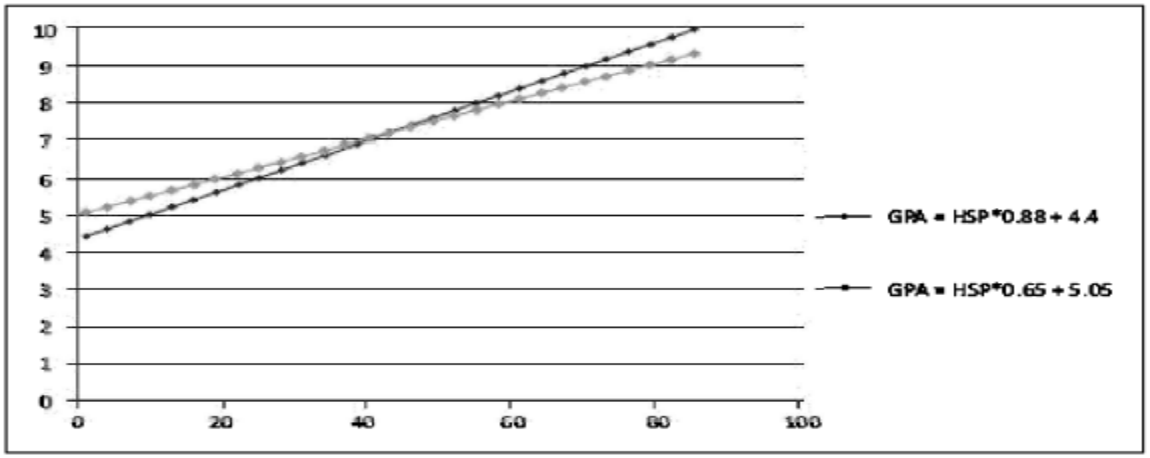


Figure 6.6 Linear regression relationship with two regression lines with different coefficient in regression equation

6.3.2 Least Square Estimation

Assume n data-points, i = 1, 2, …, n. A line out of two lines (Figure 6.6) that fits the data best will be one for which the sum of the squares of the n prediction errors (one for each observed data point) is as small as possible. This is the 'least squares criterion', which says that the best fit is one, which

'minimizes the sum of the squared prediction errors'. This implies that when the equation of the best fitting line is:

$y_i' = b_0 + b_1 x_i$ where b0 and b1 are the coefficients which minimize the errors. The coefficients values make the sum of the squared prediction errors as small as possible. Thus,

$$\text{Minimize } Q = \sum_{i=1}^{n} (y_i - y_i')^2 \quad (6.15)$$

Q is also called chi-square function. To minimize $Q = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$, compute the derivative with respect to b_0 and b_1, set to 0, respectively, and get the 'least squares estimates' for b_0 and b_1 as follows:

$$b_0 = \bar{y} - b_1 \bar{x}, \quad (6.16)$$

and

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \ldots (6.17)$$

6.3.3 Multiple Regressions

A criterion variable can be predicted from one predictor variable in simple linear regression. The criterion can be predicted by two or more variables in multiple regressions.

Multiple regressions are used when two or more independent factors are involved. These regressions are also widely used to make short- to mid-term predictions to assess which factors to include and which to exclude. Multiple regressions can be used to develop alternate models with different factors.

The prediction takes the form: $y = a + c_1 x_1 + c_2 x_2 + \ldots + c_n x_n$

where a is the intercept of line on the y axis (means value of y when all independent variable values = 0). The c1, c2, …, and cn are coefficients, representing the contributions (weights) of the independent variables x1, x2, …, xn in the calculation of y.

An example of a regression study is to examine the effect of education, experience, gender and social background on income.

6.3.4 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:

- Using linear analysis on sales data with monthly sales, a company could forecast sales for future months.
- 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.
- 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.
- An insurance company exploits a linear regression model to obtain a tentative premium table using predicted claims to Insured Declared Value ratio.
- 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.
- To predict the characteristics of child based on the characteristics of their parents.
- A company faces an employment discrimination matter in which a claim that women are being discriminated against in terms of salary is raised.
- Predicting the prices of houses, considering the locality and builder characteristics in a locality of a particular city.
- 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.

6.3.5 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.
Two steps for predicting the dependent variable:
Estimation step: A function is hypothesized and the parameters of the function are estimated from the data collected on the dependent variable.
Prediction step: The independent variable values are then input to the parameterized function to generate predictions for the dependent variable.
Consider an example of data that contain two variables, viz., crop yield and rainfall. Assume that the yield depends on rainfall (in certain critical growth phases). Using past yield data as a function of rainfall, the crop yield can be predicted. The application of linear regression upon these two variables will generate a linear equation, y = a + b.x, where y and x variables denotes crop yield and rainfall, respectively. Constants, a and b are the model's parameters known as the intercept and slope of the equation.

6.3.6 K-Nearest-Neighbour Regression Analysis

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.
Assume that population (all data points of interest) consist of k-data points. A data point independent variable is xi, where i = 1 to k. 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). It predicts the unknown value yp using predictor variable xp using the available values at the neighbours. The training dataset consists of available values of yni at xni with ni = 1 to K, where ni is the K-the neighbour, means just the local examples. 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, yi). K = 3 means the next to next nearest neighbour data points (xi, yi).
First find all available neighbouring target (xi, yi) cases and then predict the numerical value to be predicted based on a similarity measure.
Prediction methods are as follows: (i) Simple interpolation, when predictor variable is outside the training subset (ii) Extrapolation, when predictor variable is outside the training subset (iii) Averaging, local linear regression or local-weighted regression.
KNN analysis assumes that weight is inversely proportional to the square of distance (w a D−2), inverse of the distance (a D−1) or inverse of qth power of the distance (a D−q) called Euclidean DEu, Manhattan DMa and Minkowski DMi distances, respectively.

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

Euclidean distance $D_{Eu} = \left[\sum_{i=1}^{v}(x_i - x_i')^2\right]^{1/2}$ (6

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:

Euclidean distance $D_{Eu} = [(x_j - x_{j+1})^2 + (y_j - y_{j+1})^2]^{1/2}$ (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}:

Manhattan distance $D_{Ma} = D_{Ma} = \sum_{i=1}^{v}[|x_i - x_i'|]$ (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 DMi:

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

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.

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

when $xi = x¢_i$ then $D_H = 0$ and when x_i not equal to $x¢_i$ then $D_H = 1$. For example, Hamming distance $D_H = 1$ between 101001 11100 and 111001 11100 because just one substitution is needed, change second bit from 0 to 1 at 10^{th} place from the right to left positioned bits. Hamming distance $D_H = 4$ between 111001 00000 and 011001 11100 because we need four substitutions, change 3^{rd}, 4^{th}, 5^{th} and from 0 to 1 and 11^{th} bit from 1 to 0

Normalization Concept: Normalization factor in p-norm form in a v-dimensional space is

$$x_i = N^{-1} \cdot x_i \text{ where } N = \left(\sum_{i=1}^{v}|x_i|^p\right)^{1/p}$$

Here, xi is ith component of the vector X. The total number of components are v. Two-dimensional space v = 2, three-dimensional v = 3. The following example explains the meaning of distances, use of Euclidean and Manhattan distances, use distances for predictions, and the KNN regression analysis.

6.4 Finding Similar Items, Similarity of Sets and Collaborative Filtering

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.)

The following subsections describe methods of finding similar items using similarities, application of near-neighbour search, Jaccard similarity of sets, similarity of documents, Collaborative Filtering (CF) as a similar-set problem, and the distance measures for finding similarities.

6.4.1 Finding Similar Items

An analysis requires many times to find similar items. For example, finding similar excellent performance of students in Python programming, similar showrooms of a specific car model which show high sales per month, etc.

6.4.1.1 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. NNS algorithm is as follows: Consider set S having points in a space M. Consider a queried point q ∈M, which means q is member of M. k-NNS algorithm finds the k-closet (1-NN) points to q in S.

Three problems with the Pearson similarities (6.2.6.1):

- 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.)
- If two users overlap on only one item, no correlation can be computed.
- The correlation is undefined if series of preference values are identical.

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:

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

A ∩ B means the number of elements or items that are same in sets A and B. A ∪ 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 Jjava (CA, CS) = 30/(60 + 50) × 100% = 27%. Two sets are sharing 27% of the members for Java course.

6.4.2.1 Similarity of Documents: An application of Jaccard similarity coefficient is in Natural Language Processing (NLP) and text processing. It quantifies the similarity in documents. Computational steps are as follows:

- Find Bag of Words and remove words such as is, are, does, at, in, ….
- Assign weighting factor is the Term frequency and Inverse Document Frequency (TF-IDF). Consider the frequency of words in the document.
- Find k-shingles. A shingle is a word of fixed length.The k-shingles are the number of times the similar shingles extracted from a document or text. Examples of a shingle are Java, GP, 8.0, Python, 80%, Programming.
- Find n-grams. A gram is a contiguous sequence of fixed length item (word or set of characters, letters, words in pairs, triplets, quadruplets, ...) in a document or text. The n-grams are the number of times the similar items (1-grams, 2-grams, ..) extracted from a document or text. The 3-gram examples are Java GP 8.0, Python Programming 7.8, Big Data Analytics, 23A 240C 8LP, the numbers of which are extracted from the text.
- Compute Jaccard similarity coefficient using Equation (6.22) between the documents.

A number of other methods exist for computing similarity of documents. One method is Latent Semantic Indexing method (LSI).

6.4.3 Collaborative Filtering as a Similar-Sets Finding Problem

An analysis requires finding similar sets using collaborative filtering. Collaborative filtering refers to a filtering algorithm, which filters the items sets that have similarities with different items in a dataset. CF finds the sets with items having the same or close similarity coefficients. Following are some examples of applications of CF: Find those sets of students in computer application, and computer science who opt for the Java Programming subject in a semester. Find sets of students in Java Programming subjects to whom same teacher taught and they showed excellent performance. An algorithm finds the similarities between the sets for the CF. Applications of CF are in many ML methods, such as association rule mining, classifiers, and recommenders.

6.4.4 Distance Measures for Finding Similar Items or Users

Distance measures compute the dissimilarities. Complement of dissimilarity gives similarity.

Distance can be defined in a number of ways. Distance is the measure of length of a line between two values in a two-dimensional map or graph. Set of Equations (6.20) measures distances. For example, distance between (2014, 6%) and (2018, 8%) on a scatter plot when year is on the x axis and profit % on the y axis is Distance = √ [(2014 − 2018)2 + (6 − 8) 2] = √ (16 + 4) = 4.47, using Equation (6.20b). Distance can also be similarly defined in v-dimensional space using Equation (6.20a).

Distances between all members in a set of points can be computed in metrics space using a mathematical equation. Metrics space means measurable or quantifiable space. For example, profit and year on a scatter plot are in metric space of two dimensions. Probability distribution function values are in metric space. Consider student-performance measures ‘very good’ and ‘excellent’. These parameters are in non-metric space. How are they made measurable? They become measurable when very good is specified as grade point average 8.5 which implies that a score between 8.0 to 9.0 is very good, and define 9.5 which implies that a score between 9.0 to 10.0 is excellent on a 10-point scale

Distance can be defined as the reciprocal of weight in v-dimensional space. For example, a point at unit distance can be taken as weight w = 1, and a point at distance = 2, w = ½ and so on. Distance can also be defined as dissimilarity coefficient in v-dimensional space.

Greater distance means greater dissimilarity. Subtracting dissimilarity coefficient from 1 gives similarity coefficient. Many different algorithms exist to compute distance and thus similarity between entities, number of users or items. An algorithm computes the distances DEu, DMa, DMi, DHa [Equations (6.20a to e)] or any other distance metric, for example, Jaccard distance DJa, cosine distance DCos, edit distance DEd. Jaccard similarity, Cosine similarity, edit distance or correlation methods are used to find out similarities between users.

6.4.4.2 Euclidean Distance

Euclidean distance , $D_{Eu} = \left[\sum_{i=1}^{v} (x_i - x_j)^2\right]^{1/2}$ refer Equation (6.20a) in Section 6.3.6 for details.)

6.4.4.3 Jaccard Distance

Equation (6.22) gives J (A, B). Jaccard distance, DJa (A, B) measures the dissimilarity between two sets. It is equal to result of subtraction of Jaccard similarity coefficient J (A, B) from 1.

$$D_{Ja}(\mathfrak{A}, \mathfrak{B}) = 1 - J(\mathfrak{A}, \mathfrak{B})$$

(Refer Section 6.4.2 for details.)

6.4.4.4 Cosine Distance

Cosine similarity is a measure of similarity in the inner-product space between two vectors of finite magnitudes. Cosine distance DCos is measure of dissimilarity between vectors. A measure of cosine distance is in terms of the angle between the vectors. Cosine similarity has low complexity. Cosine distance has applications in text mining, finding similarity of documents, and similarities in sparse vectors, column-vectors (fields) and matrices (Section 3.3.3.1).

No Triangular Inequality Property Cosine distances do not exhibit triangular inequality property, while the Euclidean distances exhibit triangular inequality (Section 6.4.1.1).

Vector Cosine-Based Similarity Vector cosine similarity in terms of angle ϕ between two vectors U and V is given by equation:

$$\phi_{UV} = \cos^{-1}(U, V) = \frac{U \cdot V}{[\|U\| - \|V\|]}$$

Differing Similarity Coefficients for SPIs Calculated from Cosine distances and Euclidean Distances: Consider a comparison between the cosine and Euclidean similarities when finding similar items. Several situations exist in which predictions from two computational approaches differ. The reason is that triangular inequality holds true for Euclidean distances, while does not hold true for cosine distances. Certain dimensions have widely different values. For example, let us compare sales JLRS and ZS in column 3 and 5 of Table 6.2. ZS values are nearly ten times the value of JLRS values. A solution is normalizing the values in all dimensions by dividing with the mean values using Equation (6.21). However, that also may give differing and incorrect results using DCos. Cosine singularity is found to exhibit correct results for similarities in text documents. Cosine similarity is very efficient to evaluate situations of sparse vectors and those where one needs to consider non-zero values in the dimensions.

Concept of Sparse and Dense Vectors :Sparse vector uses a hash-map and consists of non-zero values. Hash-map is a collection, which stores data in (key-value) format (Section 3.3.1). Format is also called random access. Hashing means to convert a large value or string into shorter value or string so that indexing for searching is fast.

For example, assume a vector, which consists of array elements, (subject, number of students opting, average GPA).

1. Dense vectors have elements (Hive, 40, 8.0), (Java, 30, 8.5), (FORTRAN, 0, 0), (Pascal, 0, 0). Dense vector consists of all elements, whether the element value is 0 or not 0.
2. Sparse vectors will be two only with elements (4, 40, 8.0) and (3, 30, 8.5). Random access Sparse vector means access to elements (key, value pairs) using key. Sparse vector consists of elements for which key is such that value is not 0.
3. Sparse vector has an associated hash-map in form of a hash-table. First row— Pascal, 1, second row— FORTRAN, 2, third row— Java, 3 and fourth row— Hive. Hashing is a process of assigning a small number or small-sized string indexing, searching and memory saving purposes. Hash

process uses a hash function, which results into not-colliding values. In case of two colliding numbers, the process assigns a new number.

4. Sequential access sparse vectors mean two parallel accessing vectors, i.e., one to access keys and the other for values.

6.4.4.5 Edit Distance

Edit distance DEd is a distance measure for dissimilarity between two set of strings or words. DEd equals the minimum number of inserts and deletes of characters needed to transform one set into another. Applications of edit distances are in text analytics and natural language processing, similarities in DNA sequences etc. DNA sequences are strings of characters.

6.4.4.6 Hamming Distance If both U and V are vectors, Hamming distance DHa is equal to the number of different elements between these two vectors. Recall Example 6.5 (iv) for Hamming distance between Jspi and Zspi. Hamming similarity-coefficient between car models Jaguar Land Rover and Zest is (1− 2/7) = 0.7. [70%]

If M is a matrix, then DHa is equal to the number of different elements between the rows of M ignoring the columns.

DHa between two strings of equal length is the number of positions at which the corresponding characters differ. DHa is also equal to the minimum number of substitutions required to transform one string into the other. DHa is also equal to the minimum number of errors that need correction using transformation or substitution. Hamming distance is therefore another distance measure for measuring the edit distance between two sets of strings, words or sequences.

6.5 Frequent ItemsetSets and Association Rule Mining

The following subsections describes frequent itemset mining, market basket model, association rules mining, and their applications.

6.5.1 Frequent Itemset Mining

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. Frequent itemset refers to a frequent itemset, which is a subset of items that appears frequently in a dataset.

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. FIM is one of the popular techniques to extract knowledge from data.

The extraction is based on frequently occurring events. An algorithm specifies a given minimum frequency threshold for considering an itemset as frequent. The extraction generally depends on the specified threshold. 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.

6.5.2 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). A formal statement of the association rule problem is: Let I = {I1, I2, …, Id} be a set of d distinct attributes, also called literals. Let T = {t1, t2, …, tn} be set of n transactions and contain a set of items such that T ⊆ I. An association rule is an implication of the form, X → Y, where X, Y belong to sets of items called itemsets (X, Y ⊂ I), and X and Y are disjoint itemsets (X ∩ Y = ⦰). Here, X is called antecedent, and Y consequent.

Explanation: ⊆ means 'subset of', ⊂ means 'proper (strict) subset of', ∩ means intersection and ⍉ means disjoint, no commonality in members. Consider an If () then () form of a rule. The If part of the rule (A) is known as antecedent and the THEN part of the rule (B) is known as consequent. The condition is antecedent. Result is consequent.

6.5.3 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. Support is an indication of how popular an itemset is. That is the frequency of the itemset for appearing in a database. Assume X and Y are two itemsets.

Apriori principle holds due to the following property of support measure:

$\forall X, Y: (X \subseteq Y) \rightarrow s(X) \geq s(Y)$

Explanation: ∀ means for all, and ⊆ means 'subset of' and can be 'equal to or included in'. Support of an itemset never exceeds the support of its subsets. This is known as the anti-monotone property of support. The algorithm uses k-itemsets (An itemset which contains k items is known as a k-itemset) to explore (k+1)-itemsets in order to mine frequent itemsets from transactional database for the Boolean association rules (If Then rule is a Boolean association rule, as it checks if true or false). The frequent itemset algorithm uses candidate generation process. The groups of candidates are then tested against the dataset.

6.5.4 Evaluation of Candidate Rules Apriori algorithm evaluates candidates for association as follows:

Ck: Set of candidate-itemsets of size k
Fl: Set of frequent itemsets of size k
F1 = {large items}
for (k=1; Fk != 0; k++) do {
Ck+1 = New candidates generated from Fk
for each transaction t in the database do
Increment the count of all candidates in Ck+1 that are contained in t
Fk+1 = Candidates in Ck+1 with minimum support
}

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 is joined 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.

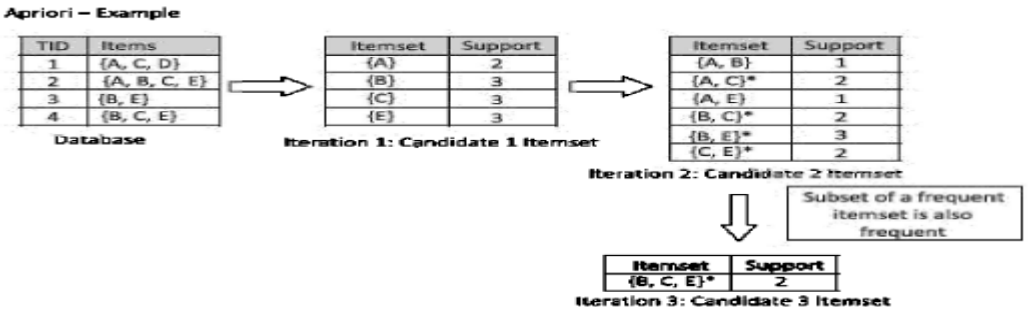


Figure 6.8 Apriori algorithm process for adopting the subset of frequent itemsets as a frequent itemset

6.5.5 Applications of Association Rules

FIM is a popular technique for market basket analysis.

6.5.5.1 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. 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.

Market basket analysis generates If-Then scenario rules. The rules are derived from the experience. This may be the result of frequencies of co-occurrence of items in past transactions. The rules can be used in several analytical strategies. The rules can be written in format If {A} Then {B}.

The If part of the rule (A) is known as antecedent and the THEN part of the rule (B) is known as consequent. The condition is antecedent and the result is consequent. If-then rules about the contents of baskets: {p1, p2…, pk} → q means, "If a basket contains all of p1, p2…, pk then it is likely to contain q."

Scale of analysis:

- Amazon sells more than 12 million products and can store hundreds of millions of baskets.
- www has 1000 million words and several billion pages.
- 75 million credit card transactions in a month in India (RBI statistics of June, July 2016) at Point of Sales (POS) terminals.

Market basket analysis signifies shopping carts and supermarket shoppers at once. The approach behind Amazon's users who bought a particular product also reviewed or bought other list of items is a well-known example of market basket analysis.

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

- 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. The analysis is also useful in genome analysis, molecular fragment mining, drug design and studying the role of biomarkers in medicine.
- Web usage analytics: FIM approaches can be used with viewing data on websites.. The results of this type of analysis can be used to inform website design (how items are grouped together) and to power recommendation engines (Section 6.8). Results are helpful in targeted marketing. For example, advertising content that people are probably interested in, based on past behavior of users.
- 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. Another example can be to find frequently occurring relationships or FIM rules between the various parties involved in the handling of the financial claim.

Some examples are: ♦ Financial institutions to analyze credit card purchases of customers to build profiles for fraud detection purposes and cross-selling opportunities. ♦ Insurance institution builds the profiles to detect insurance claim fraud. The profiles of claims help to determine if more than one claim belongs to a particular victim within a specified period of time.

- 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
- 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. The analysis might also be useful to determine capacity requirements.
- Plagiarism detection: It is the process of locating instances of similar content or idea within a work or a document. Formation of relevant word and sentence sequences for detection of plagiarism using association rule mining technique is also very popular technique.

6.5.5.2 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.

6.5.5.3 Finding Similarity

Section 6.4 describes finding similarity of an item attribute, such as sales percentage increase using Euclidean or cosine similarity coefficients. Section 6.4.2 describes Jaccard similarity of sets. The similarity of sets applies to recommenders and collaborative filtering.

Text, Web Content, Link, and Social Network Analytics

Text Mining:-

Four definitions are:

- "Text mining refers to the process of deriving high-quality information from text." (Wikipedia)
- "Text mining is the process of discovering and extracting knowledge from unstructured data." (National Center of Text Mining—The University of Manchester1)
- "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)
- "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.

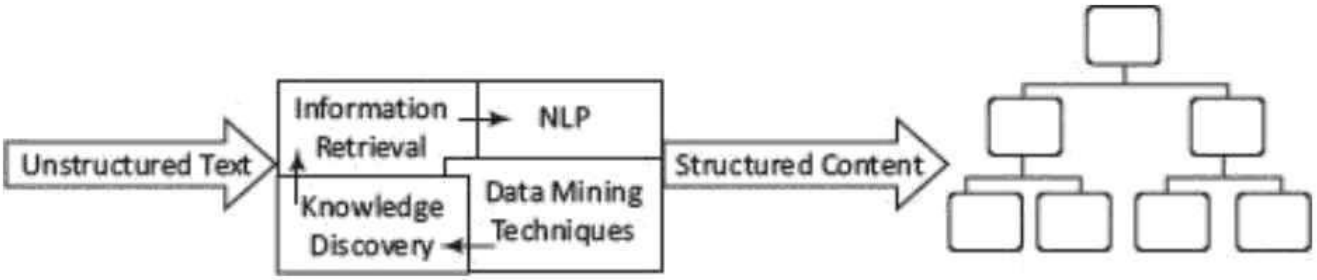
The applications provide innovative and insightful results. The results when combined with other data sources, find the answers to the following:

(i) Two terms which occur together (ii) Information linkage with another information (iii) Different categories that can be created from extracted information (iv) Prediction of information or categories.

Text Mining Overview :

Text mining includes extraction of high-quality information, discovering and extracting knowledge, and revealing patterns and relationships from unstructured data available in the form of text. The term text analytics evolves from provisioning of strong integration with the already existing database technology, artificial intelligence, machine learning, data mining and text Data Store techniques.

Information retrieval, natural language processing (NLP), classification, clustering and knowledge management are some of such useful techniques. Figure 9.1 shows process-pipeline in text-analytics.



Areas and Applications of Text Mining

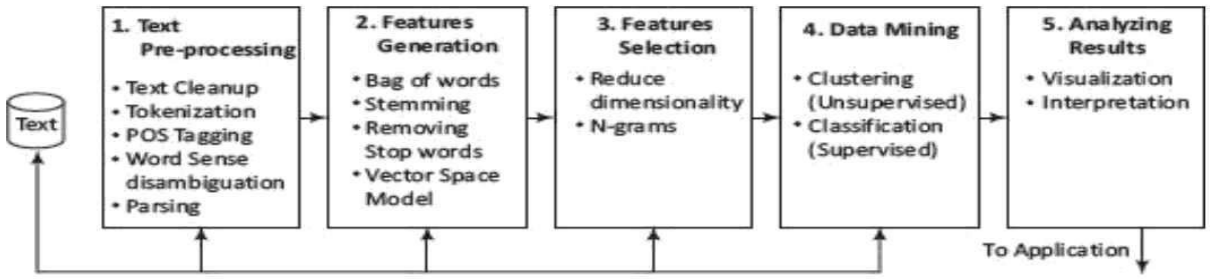
- Natural Language Processing (NLP) is a technique for analyzing, understanding and deriving meaning from human language. NLP involves the computer's understanding and manipulation of human language. NLP algorithms are typically based on ML algorithms. They automatically learn the rules. First, they analyze set of examples from a large collection of sentences in a book. Then, they make the statistical inferences. Figure 9.1 Text analytics process pipeline 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.

- 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. IE requires a dictionary of extraction patterns (For example, "Citizen of <x>, or "Located in <x>") and a semantic lexicon (dictionary of words with semantic category labels).
- 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, blogs 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 phases

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.



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. For example, removing comments, removing or escaping "%20" from URL for the web pages or cleanup the typing error, such as teh (the), do n't (do not) [%20 specifies space in a URL].
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. English language set includes the noun, verb, adverb, adjective, prepositions and conjunctions. Part of Speech encoded in the annotation system of the Penn Treebank Project has 36 POS tags.[4]
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.

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, impurification → pure.

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.

Term frequency and inverse document frequency (IDF) are important metrics in text analysis. TF-IDF weighting is most common—Instead of the simple TF, IDF is used to weight the importance of word in the document.

TF-IDF stands for the 'term frequency-inverse document frequency'. It is a numeric measure used to score the importance of a word in a document based on how often the word appears in that document and in a given collection of documents. It suggests that if a word appears frequently in a document, then it is an important word, and should therefore be high in score. But if a word appears in many more other documents, it is probably not a unique identifier, and therefore should be assigned a lower score. The TF-IDF is measured as:

$$\text{TF} - \text{IDF}(t) = \frac{\text{No. of times t appears in a document}}{\text{Total No. of terms in the document}} \times \log \frac{\text{No. of documents in the collection}}{\text{No. of documents that contain t}}, \quad (9.1)$$

where **t** denotes the term vector.

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.

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 *Naïve 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.

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.

Text mining Challenges:

The challenges in the area of text mining can be classified on the basis of documents area-characteristics. Some of the classifications are as follows:

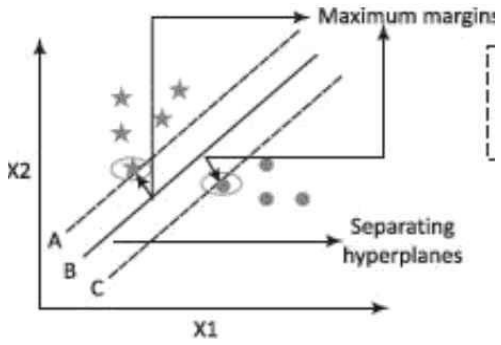
1. NLP issues:
   (i) POS Tagging
   (ii) Ambiguity
   (iii) Tokenization
   (iv) Parsing
   (v) Stemming
   (vi) Synonymy and polysemy
2. Mining techniques:
   (i) Identification of the suitable algorithm(s)
   (ii) Massive amount of data and annotated corpora
   (iii) Concepts and semantic relations extraction
   (iv) When no training data is available
3. Variety of data:
   (i) Different data sources require different approaches and different areas of expertise
   (ii) Unstructured and language independency
4. Information visualization
5. Efficiency when processing real-time text stream
6. Scalability

Supervised Text Classification

The supervised text classification requires labeled documents and additional knowledge from experts. The algorithms exploit the training data (where zero or more categories) to learn a classifier, which classifies new text documents and labels each document. A document is considered as a positive example for all categories with which it is labeled, and as a negative example to all others. The task of a training algorithm for a text classifier is to find a weight vector which best classifies new text documents.

The different approaches for supervised text classification are: (i) K-Nearest Neighbour Method (ii) Support Vector Machine (iii) Naïve Bayes Method (iv) Decision Tree (v) Decision Rule

- **K-Nearest Neighbours (KNN)** method makes use of training text document. The training documents are the previously categorized set of documents. They train the system to understand each category. The classifier uses the training 'model' to classify new incoming documents. KNN assumes that close-by objects are more probable in the same category. KNN finds k objects in the large number of text documents, which have most similar query responses. Thus, in KNN, predictions are based on a method that is used to predict new (not observed earlier) text data. The predictions are by (i) majority vote method (for classification tasks) and (ii) averaging (for regression) method over a set of K-nearest examples.
- **Naïve Bayes Analysis** Naïve Bayes classifier is a simple, probabilistic and statistical classifier. It is one of the most basic text classification techniques, also known as multivariate Bernoulli method. Naïve Bayes classifies using Bayes theorem along with the Naïve independence assumptions (conditional independence). The classifier computes condition probabilities for the conditional independence .
- **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 (X1, X2, ..., Xn) in n-dimensional space. Assume for the sake of simplicity, that a vector has two components, X1 and X2 (Two sets of words in text analysis).



- **Binary classification:**

For a given training data $[x_i, y(x_i)]$ for i = 1 ...N, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, learn a classifier f(x) such that:

$$f(x_i)\begin{cases}\geq 0 & y_i = +1 \\ <0 & y_i = -1\end{cases} \quad (9.19)$$

The above equation implies that $y_i f(x_i) > 0$ for correct classification.

Figure 9.4 shows a two-class classification. The method is using one hyperplane B for separating two-class classification of data points.

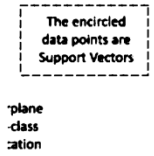


Figure 9. 4 Concept of training set data using support vectors

Let us take the simplest case of two-class classification. Suppose there are two features X1 and X2 and it is required to classify objects as shown in the Figure 9.4. Stars and dots represent the objects (itemsets, sets of words, entities) of two classes. The goal is to design a hyperplane (B) that classify all training vectors in two classes for linearly separable binary set.

Web mining, web content and web usage analytics

Web is a collection of interrelated files at web servers. Web data refers to web content—text, image and records, (ii) web structure—hyperlinks and tags, and (iii) web usage—http logs and application server logs.

Web Mining:

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

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.

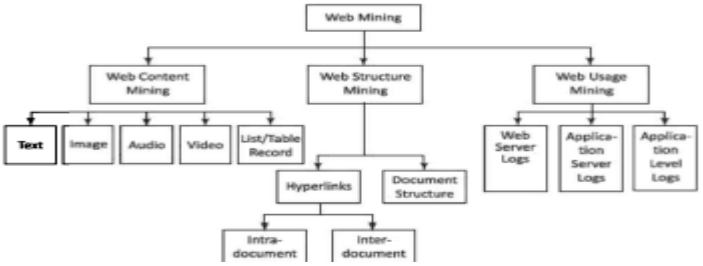


Figure 9.6 Web mining taxonomy

Web 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: 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) Document Structure: The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting the related pages.

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. Web content mining relates to both, data mining as well as text mining.

Following are the reasons: (i) The content from web is similar to the contents obtained from database, file system or through any other mean. Thus, available data mining techniques can be applied to the web. (ii) Content mining relates to text mining because much of the web content comprises texts. (iii) Web data are mainly semi-structured and/or unstructured, while data mining is structured and the text is unstructured.

Applications Following are the applications of content mining from web documents:

Classifying the web documents into categories

Identifying topics of web documents

Finding similar web pages across the different web servers

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.

Common Web Content Mining Techniques

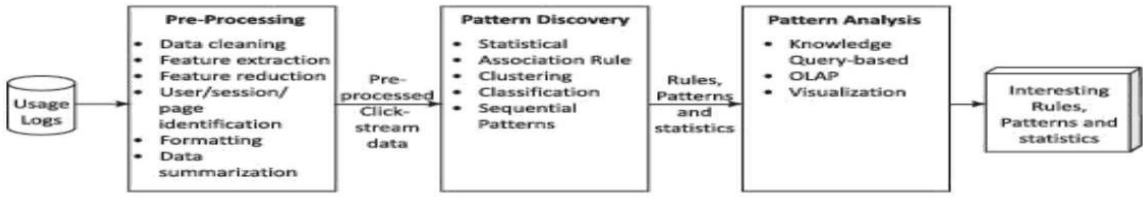
Pre-processing of contents The pre-processing steps are quite similar to the pre-processing for text mining. The content preparation involves:

1. Extraction of text from HTML
2. Data cleaning by filling up the missing values and smoothing the noisy data
3. *Tokenizing*: Generates the tokens of words from the cleaned up text
4. *Stemming*: Reduce the words to their roots. The different grammatical forms or declinations of verbs identify and index (count) as the same word. For example, stemming will ensure that both "closed" and "closing" are derived from the same word "close". Stemming algorithm, *Porter*, can be used here. The java code for *Porter* stemming algorithm can be obtained from *https://tartarus.org/martin/PorterStemmer/java.txt,*
5. *Removing the stop words*: The common words unlikely to help in the mining process such as articles (a, an, the), or prepositions (such as, to, in, for) are removed.
6. *Calculate collection wide-word frequencies*: The distinct-word stem obtained after stemming process and removing the stop words results into a list of significant words (or terms). Calculating the occurrence of a significant term (t) in a collection is called collection frequency (CF_t). CF counts the multiple occurrences.)

   Now, find the number of documents in the collection that contains the specific term (t). This numeric measure is the document frequency (DF_t).
7. *Calculate per Document Term Frequencies* (TF). TF is a numeric measure that is used to score the importance of a word in a document based on how often it appeared in that document (Refer Example 9.1).
8. *Bag of words*: Web document is represented by the words it contains (and their occurrences.

Web Mining Usage:-

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.



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.

Usage data are collected at server, client and proxy levels. The usage data collected at the different sources represent the navigation patterns of the overall web traffic. This includes single-user, multi-user, single-site access and multi-site access patterns.

1. Preprocessing

The common data mining techniques apply on the results of pre-processing using vector space model (Refer Example 9.2).

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.

2. 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.

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.

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.

Classification The method classifies data items into predefined classes. Classification is useful for:

(i) Developing a profile of users belonging to a particular class or category

(ii) Discovery of interesting rules from server logs. For example, 3750 users watched a certain movie, out of which 2000 are between age 18 to 23 and 1500 out of these lives in metro cities.

Classification can be done by using supervised inductive learning algorithms, such as decision tree classifiers, Naïve 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. Transition probability between two states can represent the probability that a user will navigate from one state to the other. This representation allows for the computation of a number of significant user or site metrics that can lead to useful rules, pattern, or statistics.

3. 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.

Data cube enables visualizing data from different angles. For example, *toys* data visualization using category, colour and children preferences. Another example, news from category, such as sports, success stories, films or targeted readers (children, college students, etc).

PAGE RANK, STRUCTURE OF WEB AND ANALYZING A WEB GRAPH

Hyperlinks links exist between the web contents. Link analysis finds the answers to the following:

- Can a linked (web) page rank them higher or lower?
- Can the links be modeled as edges of graphs, structure of web as graph network, and applied the tools same as for graph analytics?
- Can web graph mining method analyze and find a link sending spams?
- Does a set of links correspond to a hub? Do the links correspond to an authority?
- Does a linked page has higher authority compared to others?

Links analysis applies to domains of social networks and e-mail.

The following sub-sections describe the applications of link analysis:

PAGE RANK:

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.

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 u1, u2, … un. 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 v1, v2, … vm. 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:

- Text at the hyperlink represents the property of a vertex u that describes the destination V of the out-going edge.
- 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 below shows a web structure consisting of pages and hyperlinks.

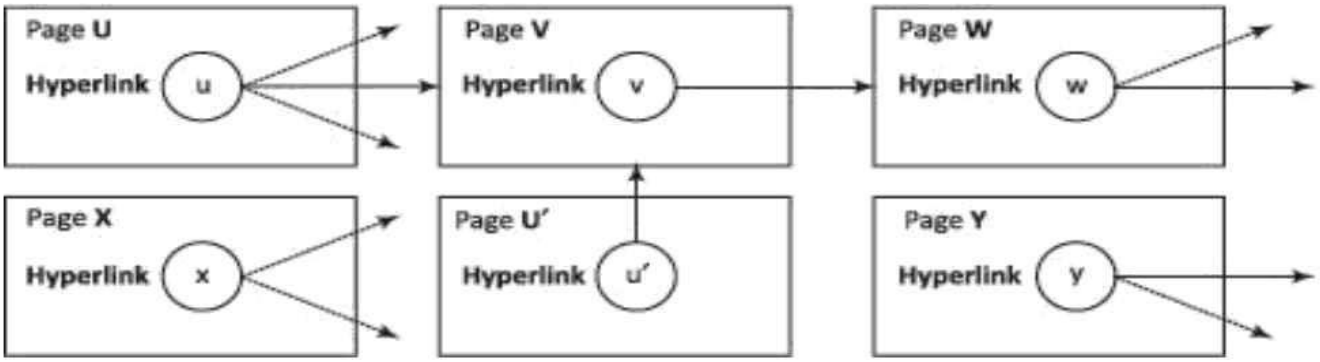


Figure 9.8 Web structure with hyperlinks from a parent to one or more pages

Dead Ends: Dead-end web pages refer to pages with no out-links. When a web page links to such pages, its page rank gets reduced. Dead ends are on a website having poor linking structure. The web structure of service pages may have pages with a dead end. The end causes no further flows for further action and no internal links.

Analyzing and Implementing a System with Web Graph :

Mining Number of metrics analyze a system using web graph mining. Following are the examples:

- In-degrees and out-degrees
- Closeness is centrality metric. Closeness, $Cc(v) = 1 / \sum_{u \neq v} gdist(v,u)$, where gdist is the geodesic distance between vertex v with u and sum is over all u linked with V. Geodesic distance means the number of edges in a shortest path connecting two vertices. Assume v has an edge with w, and w has an edge with u. Assume u does not have direct edge from v. Then, geodesic distance = 2 (two edges between v and u in shortest path).
- Betweenness
- PageRank and LineRank
- Hubs and authorities
- Communities parameters, triangle count, clustering coefficient, K-neighbourhood
- Top K-shortest paths

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 N_{out} [(Pg (u)] pages. Figure 9.9 shows Pg (v) in-links from Pg (u) and other pages.

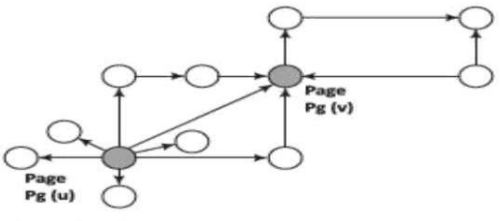


Figure 9.9 Page Pg (v) in-links from Pg (u) and other pages

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:

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

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 ∈ 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:

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

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:

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

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.

Significance of *a* PageRank can be seen as modeling a "random surfer" that starts on a random page and then at each point: E(v) models the probability that a random link jumps (surfs) and connect with out-link to Pgv. R(v) models the probability that the random link connects (surf) to Pgv at any given time. The addition of E(v) solves the problem of Pgv by chance out-linking to a link with dead end (no outgoing links).

PageRank Iteration using MapReduce functions in Spark Graph

The computation of PageRank using SparkGraph method (Section 8.5),

```
graph.pageRank(0.0001).vertices
ranksByUsername = users.join(ranks).map{case id, (username, rank)) => (username, rank).
```

The method includes conversions to MapReduce functions and using HDFS compatible files. Functions PageRank (), ranksByUsername () do the computations using the PageRankObject. GraphX consists of these functions (GraphOps). GraphX Operators includes the functions (Section 8.5).

Static PageRank algorithm runs for a fixed number of iterations, while dynamic PageRank runs until the computed rank converges. Convergence means that after certain iterations, the rank does not change significantly and any change remains within a pre-specified tolerance. Thereafter the iterations stop.

Assume specified tolerance at the start of iterations is 0.0001 (1 in 10000). When the rank does not change beyond that tolerance, it means rank value will converge and then the iterative process will stop.

Topic Sensitive PageRank and Link Spam

Number of methods have been suggested for computations of topic-sensitive page ranking, R_{TS}. The R_{TS} (v) of a page P (v) may be higher for a specific topic compared to other topics. A topic associates with a distinct bag of words for which the page has higher probability of surfing than other bags for that topic.

Topic-sensitive PageRank method uses surfing weights (probabilities) for the pages containing the topic or bag of words corresponding to a topic. Method for creating topic-sensitive PageRank is to compute the bias to rank R(v) and thus increase the effect of certain pages containing that topic or bag of words.

Recapitulate equation (9.26). Probability of random jump to page v is E(v). An alternative equation for topic sensitive PageRank, R(v) computation for page P (v) is as follows:

$$R(v) = nc \cdot \left\{ (1-\alpha_t) \cdot P(v) \sum_{u \in PA(v)} \left[\frac{R(u).}{|ch(u)|} \right] + \alpha_t \cdot E(v) \right\}. \quad (9.29)$$

Probability of random jump to page v is *E* (*v*). $\alpha_t = 0$ for page unrelated to a topic α is not 0 for page related to a topic. α_t = surfing probability for in-links for a topic t. Further, coefficient (1- α) is considered as biasing factor depending on the web page P (*v*) selected for a queried topic *t*.

Link Spam

Effects of a *link spam* can be nullified using the topic-sensitive PageRank algorithm. Link Spam tries to mislead the PageRank algorithm. A link spam attempts to make PageRank algorithm ineffective. The spam assisting pages connects to the page repeatedly and increases the in-degree of a page, thereby enhancing the rank to a large value.

A link spam creator website *ws* also has a page l_s for whom w_s attempts to enhance the PageRank. The w_s has a large number of assisting pages al_s which out-links to *ls* only. The al_s pages also prevent the PageRank of l_s from being lost. A spam mass consists of w_s, l_s and its al_s pages.

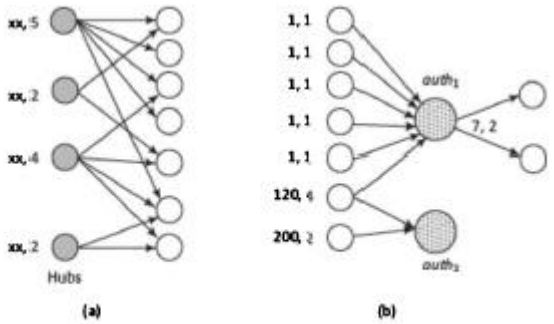
Following are the steps for finding spam mass:

1. A distant topic sensitive page has unusually high in-degrees compared to the other pages of the same topic. A plot known as power-law plot is drawn between the log of number of web pages on the y-axis out-linking to the page v and logs of them in-degrees of v on the x-axis.
2. Plot is nearly linear as the number exponential decays is within degrees. N is proportional to exp (-d), where d is decay constant.
3. An unusual pattern with marked deviation from near linearity identifies the distant link spam mass.

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.

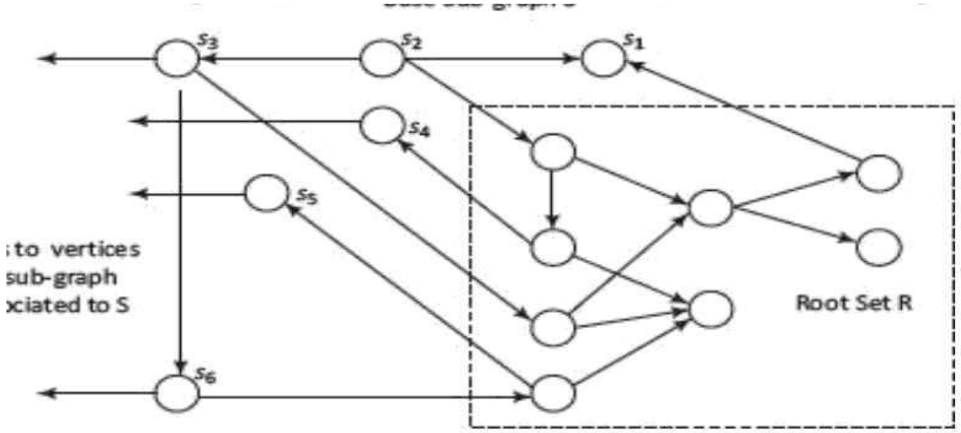


igure 9.10 (a) Hubs (shaded circles) and (b) Authorities (dotted circles)

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, $auth_1$ in Figure 9.10(b) has in-links from 6 vertices (in-degrees = 6) and $auth_2$ has in-links to just 2 (in-degree = 2). However, $auth_1$ has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125). Authority, $auth_2$ has links with two vertices with in-degrees = 120 and 200 (total = 220). $Auth_2$ has association with greater authorities. Therefore, in-degrees may not be a good measure as compared to authority.

Sub-graph for HITS consisting of root set **R** of pages and children of parents in the sub-graph **S**. Figure 9.11 shows subgraph **S** for HITS consisting of root set R of pages and all the pages pointed to by any page of R.



Limitations of Link, Rank and Web Graph Analysis:

Following are the limitations of link and web graph analysis:

- Search engines rely on metatags or metadata of the documents. That enhances the rank if metadata has biased information.
- Search engines themselves may introduce bias while ranking the pages of clients higher as the pages of advertising companies may provide higher searches and hence lead to biased ranks.
- A top authority may be a hub of pages on a different topic resulting in increased rank of the authority page.
- Topic drift and content evolution can affect the rank. Off-topic pages may return the authorities.
- Mutually reinforcing affiliates or affiliated pages/sites can enhance each other's rank and authorities.

- The ranks may be unstable as adding additional nodes may have greater influence in rank changes.

Social Networks as Graphs and Social Network Analytics :-

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.

Social Network as Graphs

Social network as graphs provide a number of metrics for analysis. Network topological analysis tools compute the degree, closeness, betweenness, egonet, K-neighbourhood, top-K shortest paths, PageRank, clustering, SimRank, connected components, K-cores, triangle count, graph matches and clustering coefficient. Bipartite weighted graph matching does collaborative filtering.

Apache Spark GraphX and IBM System G Graph Analytics tools are the tools for social network analysis.

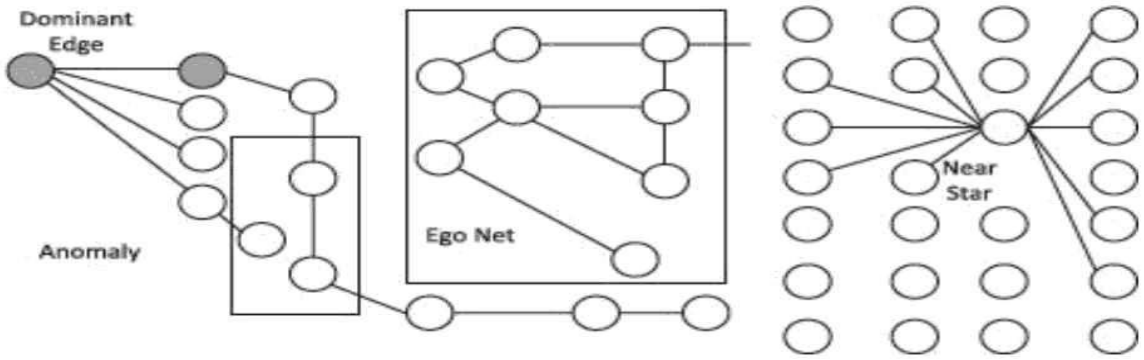
Centralities, Ranking and Anomaly Detection

Important metrics are degree (centrality), closeness (centrality), betweenness (centrality) and eigen vector (centrality). Eigen vector consists of elements such as status, rank and other properties. Social graph-network analytics discovers the degree of interactions, closeness, betweenness, ranks, probabilities, beliefs and potentials. Social network analysis of closeness and sparseness enables detection of abnormality in persons. Abnormality is found from properties of vertices and edges in network graph. Analysis enables summarization and find attributes for anomaly.

Social network characteristics from observations in the organizations are as follows:

1. Three-step neighbourhoods show positive correlation between a person and high performance. Betweenness between vertices and bridges between numbers of structures are not helpful to the organization. Too many strong links of a person may have a negative correlation with the performance.
2. Social network of a person shows high performance outcome when the network exhibits structural diversity. Person with a social network with an abundant number of structural holes exhibits higher performance. This is because having diverse relations help an organization.

Social network analysis enables detection of an anomaly. An example is detection of one dominant edge which other sub-graphs are follow (succeed). Ego network is another example. The network structure is such that a given vertex corresponds to a sub-graph where only its adjacent neighbours and their mutual links are included. The analysis enables spam detection. Spam is discovered by observation of a near star structure.



. Figure 9.12 Discovering anomaly, ego-net and spam (using near star) from the analysis

Social network has concerns of privacy, security and falsehood dissimentation. Security issues are phsishing attacks and malwares.

Social Graph Network Topological Analysis using Centralities and PageRank

Social graph network can be topologically analyzed. The centralities (degree, closeness, effective closeness and betweenness) and PageRank (vertexRank similar to PageRank in web graph network) are the parameters analyzed.

Degree

Degree of a graph vertex means the total number of edges linked to that. *In-degree of a vertex* means the number of in-edges from the other vertices. *Out-degree of a vertex* means the number of out-edges to other vertices to which that vertex directs. *Degree distribution function* means the distribution function for the degrees of vertices (Section 6.2.5 described the common distribution functions).

Closeness

Graph vertex closeness c_C (v) is a way of defining the centrality of a vertex in reference to other vertices. Sum is the overall vertices connected to other vertices **u**. The **u** is a subset of vertices in set **V**.

The centrality (closeness index), c is function of distances of vertices.

$$c_C(v) = [\sum_{u \in V} d(u, v)]^{-1}.$$

where d (u, v) is distance between u and v for path traversal.

Effective Closeness

Effective closeness C_{ec}(v)can also be analyzed. Use approximate average distance from v to all other vertices in place of the shortest paths. C_{ec} reduces run time for cases with a large number of edges and near linear scalability in computations.

Betweenness

Graph vertices betweenness means the number of times a vertex exists between the shortest path and the extent to which a vertex is located 'between' other pairs of vertices. Betweenness c_B (v) of a vertex v requires calculating the lengths of shortest paths among all pairs of vertices and computations of the summation for each pairing vertex in V.

PageRank

PageRank is a metric for the importance of each vertex in a graph, assuming an edge from *v1* to *v2* represents endorsement of importance of *v2* by *v1* by connecting, following, interacting, opting for relationship, sharing belief or some other means.

Contacts Size

Contacts size means a vertex connection to many vertices. The size of each vertex does not convey any meaningful information. A big social graph network will also require high maintenance cost.

Indirect Contacts

Indirect contacts metric means betweenness, which is the sum of the shortest paths within geodesic distances from all other pairing vertices. Three-step contact metric means a number of edges to other vertices plus the number of edges from other vertices within geodesic distances = < 3.

Both metrics convey meaningful information. The indirect contacts metric has meaning in terms of magnitude of betweenness centrality.

Structure Diversity

Structure diversity metric means that social graph has access to diverse sub-graphs (knowledge).

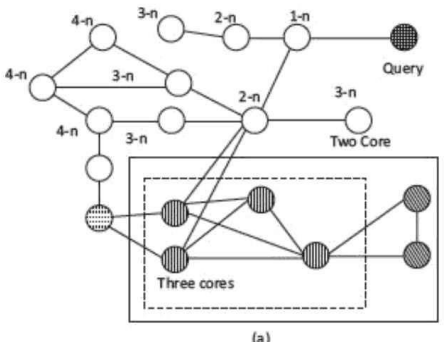
Social Graph Network Analysis using K-core and Neighbourhood Metrics

K-core is a sub-graph in a graph network structure. *Graph Vertex* K^{th} neighbourhood is number of 1^{st} neighbour vertices, 2^{nd} neighbour vertices and so on to a querying vertex that are correlated, linked, and have weighted correlations or the associations.

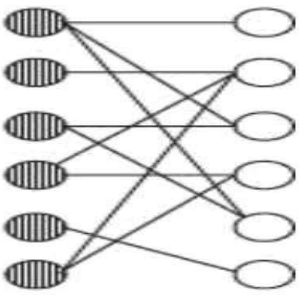
K-nearest neighbourhood (KNN) finds K-similar objects, items, or entities, which are nearest neighbours after computing the similarities. For example, KNN is K-documents (or books) in the large number of text documents (books) that are most similar to the queried document.

Collaborative filtering for frequent itemsets uses weighted bipartite graph matching.

Figure 9.13 shows the K-cores and K-neighbourhood metrics for a social network graph. The figure also shows frequent itemsets obtained from collaborative filtering algorithm (Sections 6.4 and 6.8.1).



(a)



Clustering in Social Network Graphs

One of the methods of detecting communities from social graph analysis is finding clustering and cluster coefficients. A clustering coefficient is a metric for the likelihood that two associated vertices of a vertex are also associated with other vertices. A higher clustering coefficient indicates a greater association and cohesiveness.

Connected components mean components of the datasets (represented by properties of vertices) connected together. For example, finding student–teacher datasets, social network datasets, etc.

SimRank

Similarity can be defined by properties of graph vertices. For example course, subject, student, scientist, Java programmer, status, values, or any other salient characteristic. Social network analysis of graphs computes *SimRank*.

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

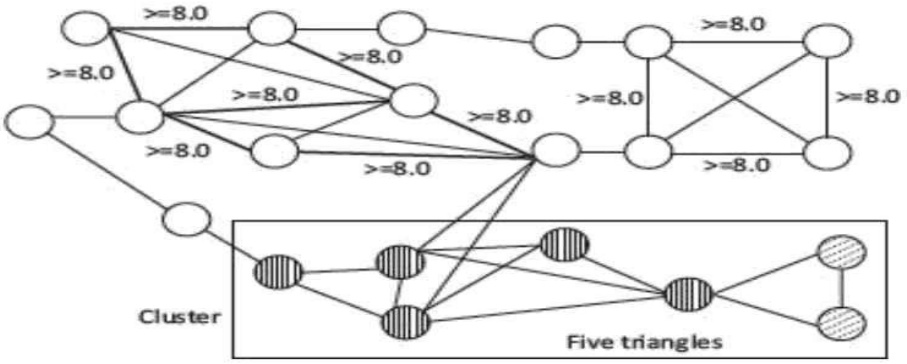


Figure 9.14 Clustering of five triangles and three matches of graphs

Using SparkGraph (Map-Reduce) for Network Graphs

Connected components compute by graph.connectedComponents().vertices method in SparkGraph. Connected Components Algorithm labels each connected component of the graph with an ID. Each connected component ID is ID of the lowest-numbered vertex. For example, in a social network, connected component objects can approximate clusters. GraphX contains an implementation of the algorithm in the ConnectedComponentsObject. The clusters are found by discovering close-by connected components using closeness centrality metric. SparkGraphX triangle-count algorithm computes the number of triangles passing through each vertex.

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, SimRanks and bridges are finding a set o experts, specific areas of expertise, and ranking the expertise in an organization.