MODULE 5

Machine Learning Algorithms for Big Data Analytics

6.1 l INTRODUCTION

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.

Google-owned company Deep Mind developed an Artificial Intelligence (AI) program called AlphaZero, which played 100 chess games in 24 hours, and defeated Stockfish, the highest-rated chess program by 28 games to 0 with 72 games drawn. This was a historical moment. It became a milestone in the history of AI, ML and DL.

The former world champion, Garry Kasparov, noted that achievement of AlphaZero has history-shaping potential. "The ability of a machine to replicate and surpass centuries of human knowledge, is a world-changing tool". (Garry Kasparov, "Deep Thinking — lVhere Artificial Intelligence Ends and Human Creativity Begins", published by the author himself, 2017)

6.2 ESTIMATING THE RELATIONSHIPS, OUTLIERS, VARIANCES

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 dependmetric. feztura a nd category variabl as, Relationships,. outliers.. vari ancas, probability distribution, and the correlations bor eenthe add Tables. Tems,. or entities 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 vorioble 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

VtuCode

his/her studies. A company declares the share of profit to each shareholder in a year after subtracting requirements of money for future growth from the profit.

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.Independent and dependent variables may exhibit a relation or correlation. The relationships may belinear, nonlinear, positive, negative, direct, inverse, scattered or spread. A data point for dependent variable can be an outlier with no relationship.

Data analysis requires studying relationships graphically, mathematically and statistically, studying the outliers, anomalies, variances, correlations, features, categories and probability distributions using a set of variables, and other characteristics.The relationship involves some quantifiable independent variables and the resulting dependent variable or entity. The following subsections explain methods of estimating the relationships, outliers, variances, correlations and probability distributions between a set of variables.

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 [Section 10.5]. Whether two variables are related to each other or not, can be derived from statistical analysis using scatter plots.

A data point is (y, yi) when dependent variable value - y; 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,

VTUCircle

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 Lineor and Non-linenr Befotionships

A linear relationship exists between two variables, say x and y, when a straight line ($y = a_0 + a_1.x$) can fit on a graph, with at least some reasonable degree of accuracy. The a1is the linearity coefficient. For example, a scatter chart can suggest a linear relationship, which means a straight line. Figure 6.1 shows a scatter plot, which fits a linear relationship between the number of students opting for computer courses in years between 2000 and 2017.

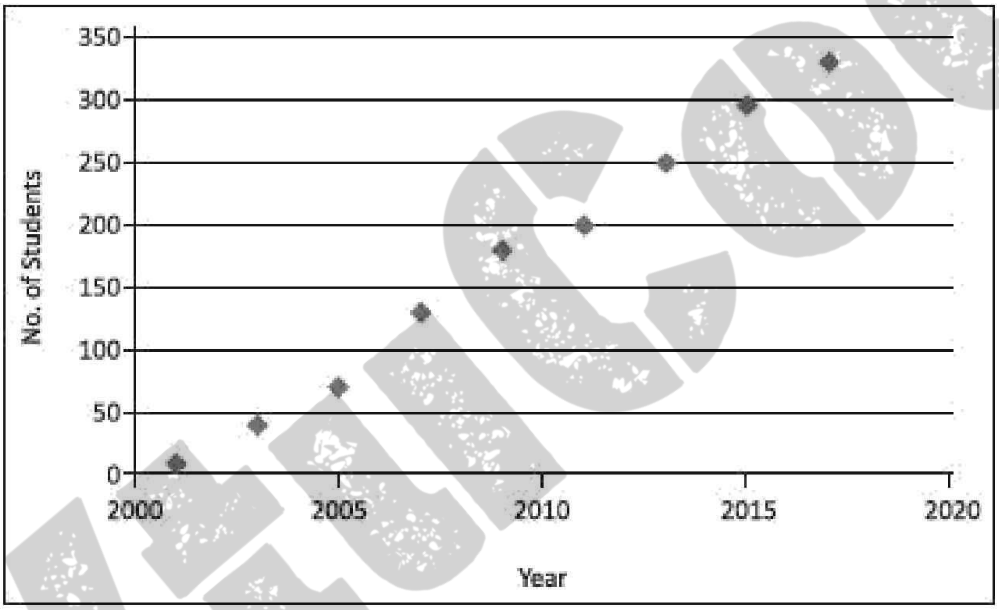


Figure 6.1 Scatter plot for linear relationship between students opting for computer courses in years between 2000 and 2017

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. Perfect, strong or weak linearship categories depend upon the bonding between the two variables.

A non-linear relationship is said to exist between two quantitative variables when a curve ($y = a0 + a1.x + a2.x^2 + ...$) can be used to fit the data points. The fit should be with at least some reasonable degree of accuracy for the fitted parameters, a_0, a_1, a_2 ... Expression for y then

VTUCode

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.

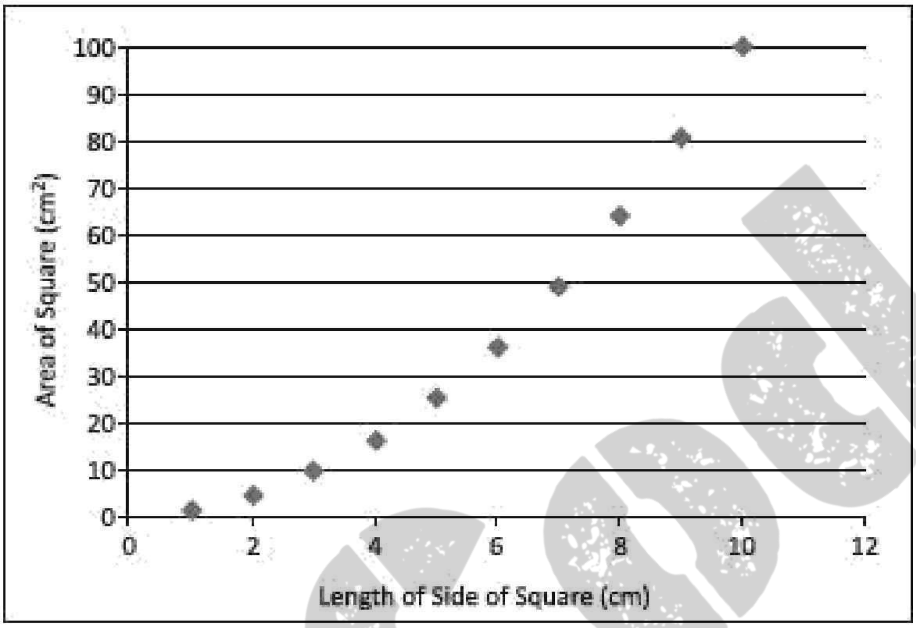


Figure 6.2 shows a scatter plot in case of a non-linear relationship between

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 yq = sales of a car model m in xth year of the start of manufacturing that model. Assume that computations show that the yq relates by a mathematical expression (yq= a0 + a1.xp + a2.xp^2) up to an acceptable degree of accuracy, when a0 - 490, a1 - 10 and a2 - 5.

Estimated first year sales, yp(1) = (490 + 10) - 500, second year yq(2) = (490 + 10 x 2 + 5 • 2^2) = 530, third year yq(3) = (490 + 10 • 3 + 5 • 3^2) = 565, if fit with the desired accuracy, then the results are showing that the expression of yq estimates the relationship between model m sales in next and other years. The yq can also predict the sales in 6^t 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 (Section 5.3.3.1). Outliers are data points that are numerically far distant from the rest of the points in a dataset, are termed as outliers. Outliers show significant variations from the rest of the points (Section 1.5.2.2). 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).

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. The following example illustrates no variance:

vtucode

EXAMPLE 6.1

Consider an examination where everyone gets the same grades. What does it signify?

SOLUTION

Some measurement problem may have taken place in a situation where either the semester examination questions were so easy that everyone got full marks, or it was so hard that everyone got a zero. Now consider the two types of examinations. After each examination, everyone gets the same score on the test, i.e., everyone gets ‘A’ grade in one test and everyone gets ‘B’ in the second test. This is again not telling much about the study or intelligent quotient of the students. Now, these no variance results signify the extreme case and hard to understand or explain. But in general, differences in scores are always found.

6.2.4.1 Standard Devintion and Standard £rror Estimates

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

Standard Deviation With the help of variance, one can find out the standard deviation. Standard deviation, denoted by 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:

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

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

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 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:

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

where sest is the standard error in the estimate, y is an observed value, y4 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 Sest'

1. When sest 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 sest 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 lo0 games, drawn 78 times, and lost none. Then, probability P of winning Pg is 0.22, P of drawn game PD is 0.78 and P of losing, P = 0. The sum of the probabilities is normalized to 1, as only one of the three possibilities exist.

Probnbility 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 x, 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 shows a PDF with normal distribution around , standard deviation and variance

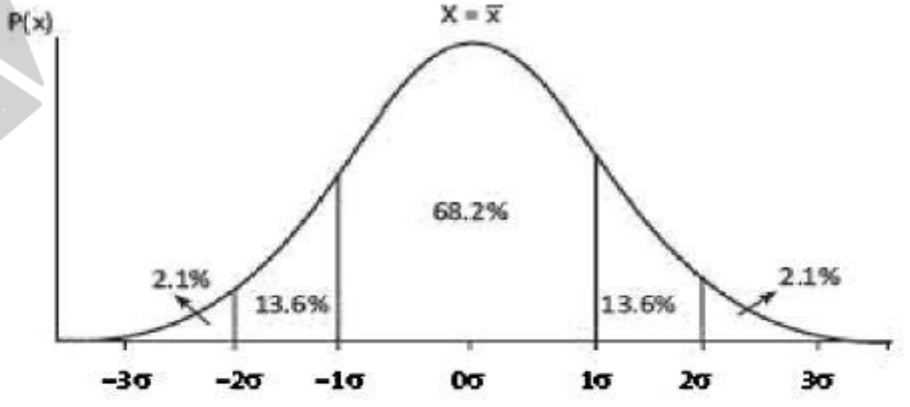


Figure 6.3 Probability distribution function as a function of x assuming normal distribution

vtucircle

The figure also shows the percentages of areas in five regions with respect to the total area under the curve for P(x). The variance for probability distribution represents how individual data points relate to each other within a dataset. The variance is the average of the squared differences between each data value and the mean.

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 s^{2}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 (x,+ s) and (, - s), 950Zo of area under the PDF is within (x‘ + 2s) and (x, - 2s) and 99.7% is within(x, + 3s) and (x, - 3s).

Standard deviation and empirical rule help in computing the population distribution over 68%, 95% and 99.7% of data under normally distributed population. This further helps in forecasting. The following example explains the meaning of population, expected values, normalized probabilities, PDF and interpretation using mean value.

EXAMPLE 6.2

Assume that N students gave the examination. Let N_1 is number of students obtained grade pointer average - 1, N_2 got 2, ..., N_{10} got 10. Highest-grade pointer is 10.0. Grade pointer obtained is not a random variable. Grade pointer variation is a random variable with an expected value and standard deviation.

Expected value among the distributed x; values, where i varies discretely from 0.0 to 10.0 will depend on the expected performance of the student. If teaching in the class is very good andstudents prepare for the examination very well, then expected value of GPA is 8.0 for very good performing students and standard deviation found is 1.0.

(i) What do you mean by population? What do you mean by sample?

(ii) What will be the normalized probabilities?

(iii) How will you define Probability Distribution Function (PDF)?

(iv) How will you interpret the results in terms of normal distribution?

(v) When will you interpret the results as poor and poorer in terms of normal distribution?

vtucode

SOLUTION

(i) Population is GPA of all the students of the university who gave the examination. Population size is N. Sample means datasets used in the analysis. It can be N or less than N students and GPA of each one.Probability that students obtained grade pointer 1 is 2 (on normalization of probability. ($N = N_1 + N_2 + \dots$)

ii) PDF represents a curve for independent variable x between GPA = 0 and GPA = 10, such that the sum of all P values is 1, where P, is the ratio of number of students getting GPA = i with respect to the total population N or the sample.

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

between x - 0 and 10.0, where s - 1.0 and x- - 8.0.

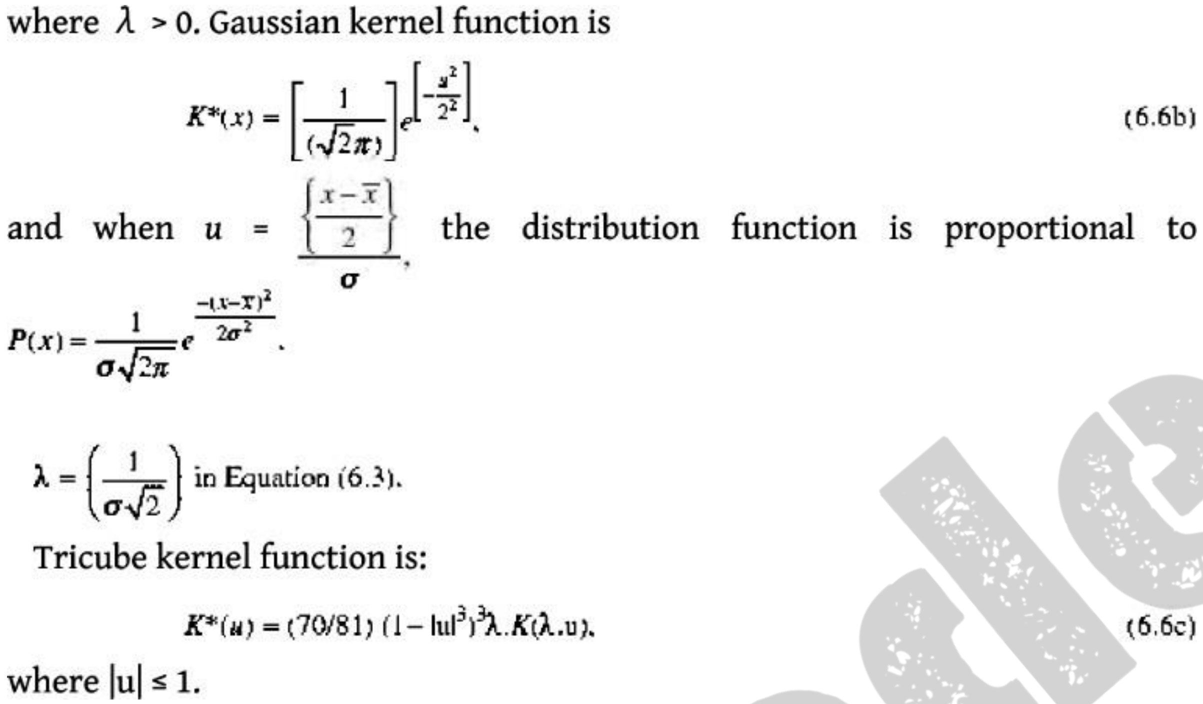
(iv) GPA value is 8.0 and standard deviation is 1.0, which means 68OZoof the students will get GPAs between 7.0 and 9.0, 95% between 6.0 and 10.0, and 99.7% between 5.0 and 10.0.

(v) The expected value of 3.0 (less than 3.0) and standard deviation of 1.0 means poor performance of students because 68% students get between 2.0 and 4.0. The expected value of 3.0- (less than 3.0, say 2.5) and standard deviation of 1.5 means poorer performance of students because 680/»students get between 1.0 and 4.0.

6.2.5.1 Kernel Functions

A probability or weight can be represented by a kernel function[1] like a Gaussian or tri-cubefunction. (Kernel in English means some thing central and key (important) part. Forexample, the kernel inside a walnut's shell is important because it is the edible part. Kernel in an operating system is key or central component.)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 [Equation (6.5)]. Figure 6.3 shows the probability normal distribution, which is a Gaussian function based on the Gaussian kernel function.

A kernel functionl, K* defines as



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 $u = \dfrac{\left\{\dfrac{x-\overline{x}}{2}\right\}}{\sigma}$, the distribution function is proportional to

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

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

Tricube kernel function is:

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

where $|u| \leq 1$.

where 3 » 0. Gaussian kernel function is

6.2.5.2 Moments

Moments (0, 1, 2, ...) refer to expected values to the powers of (0, 1, 2 ...) of random variable variance. O^{t}h moment is 1, 1st moment - £(x) - x. (expected value), 2^{n}d moment is squared V[(xi - x,)^{2}] - sum of product of (x; - x)^{2}, and P(x =

Here, P is the probability at x= x; when i is varying from 1 to n, for n values of random variable x. The rth moment is rth power of variance V[(x; - xJ"]. Moments are evaluated from the results obtained for the randomly distributed probabilistic values of the variable, such as sales. 1st moment assigns equal weight to variances of outliers and inliers, i.e., equal weight for variance of each.2^{n}d moment assigns higher weight to outliers compared to inliers. 3'd moment assigns greater weight to outliers compared to inliers. Moment can be defined with respected to the origin, and in that case, x- is considered 0.

Let P is along y axis and variable x on x axis. Central moment means that moments computetaking x, equals to variable x at x axis point where the probability curve partitionsequally by a vertical axis, parallel to the y axis.

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.

(ii) The test does not assume those two populations have the same standard deviation.

vtucode

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. Variation between samples and also within sample items may exist. For example, compare the effect of three different types of teaching methodologies on students. This may be done by comparing the test scores of the three groups of 20 students each. This technique provides inferences about whether the samples have been drawn from populations having the same mean. It is done by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.fi-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. Several different F-tables exist. Each one has a different level of significance. Thus, look up the numerator degrees of freedom and the denominator degrees of freedom to find the critical value.

vtucircle

The value of F calculated using the above-mentioned formula is to be compared to the critical value of F for the given degrees of freedom. If the F value calculated is equal or exceeds the critical value, then significant differences between the means of samples exist. This reveals that the samples are not drawn from the same population and thus null hypothesis is rejected.

6.2.5.5 No Refosonshfp Case

Statistical relationship is a dependence or association between two random variables or bivariate data. Bivariate means ‘two variables’. In other words, there are two types of data. 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 etween 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 ft is a measure of correlation between the predicted values y and the observed values of x. R-squared (ft^2) is a goodness-of-fit measure in linear- regression model. It is also known as the coefficient of determination. ft^2 is the square of ft, the coefficient of multiple correlations, and includes additional independent (explanatory) variables in regression equation.

Interpretation of R-squared The larger the it^2, 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 o/Linenr itelntionships

Correlation is a statistical technique that measures and describes the ‘strength’ and ‘direction’ of the relationship between two variables. Let us explore the relations between only 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)

VTUCircle

(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 andy is:

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

where n is the number of observations in the sample, x; is the x value for observation i, x- is the sample mean of x,y is the y value for observation i, y- is the sample mean of y, sx is the sample standard deviation of x, and s is the sample standard deviation of y.

Summation is over all n values of i, i - 1, 2, ..., n.[N is square of sample correlation coefficient between the observed outcomes and the observed predictor values, and includes intercept on y-axis in case of linear regression.]

Use of Statistical Correlation Assume one sample dataset is (ul, Hal containing n values of a parameter r. The r,, is i-th data point in dataset u. (i = 1, 2, ..., n). Another sample dataset is (vl,..., v,) containing n values of fi‹ ’v,i is i-th data point in dataset v. Let the correlation among two samples is being measured. Sample Pearson correlation metric c, measures how well two sample datasets fit on a straight line.

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.

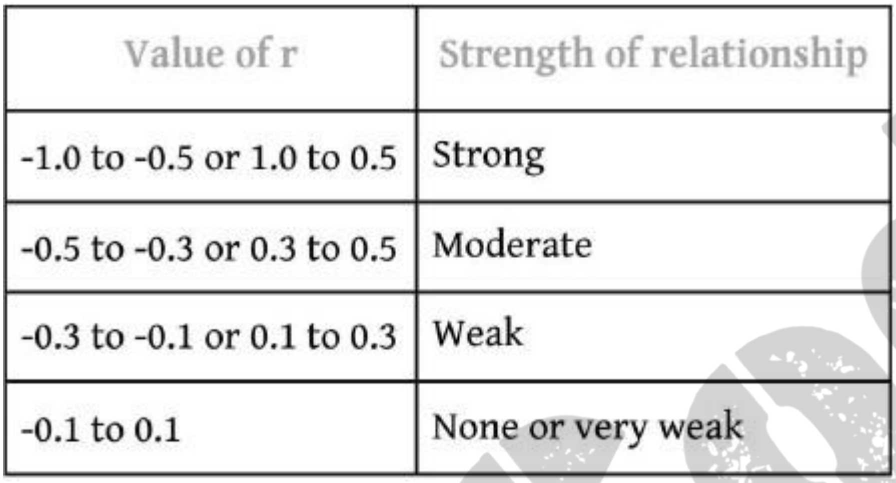
(iii) Kendall's G correlation - It is similar to the Spearman rank correlation, but instead of using ranks themselves, only the relative ranks are used to calculate the correlation.Numerical value of correlation coefficient ranges from +1.0 to -1.0. It gives an indication of both the strength and direction of the relationship between variables.

VTUCode

In general, a correlation coefficient r» 0 indicates a positive relationship; r • 0 indicates a negative relationship; r - 0 indicates no relationship (or that the variables are independent of each other and not related). Here r - ml.0 describes a perfect positive correlation and r = -1.0 describes a perfect negative correlation.

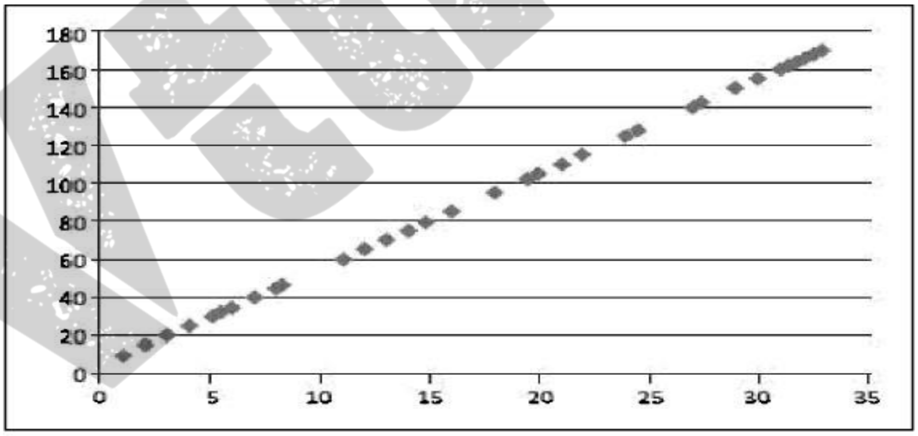
The closer the coefficients are to +1.0 and -1.0, the greater is the strength of the relationship between the variables.

Table 6.1 The strength of the relationship as a function of r



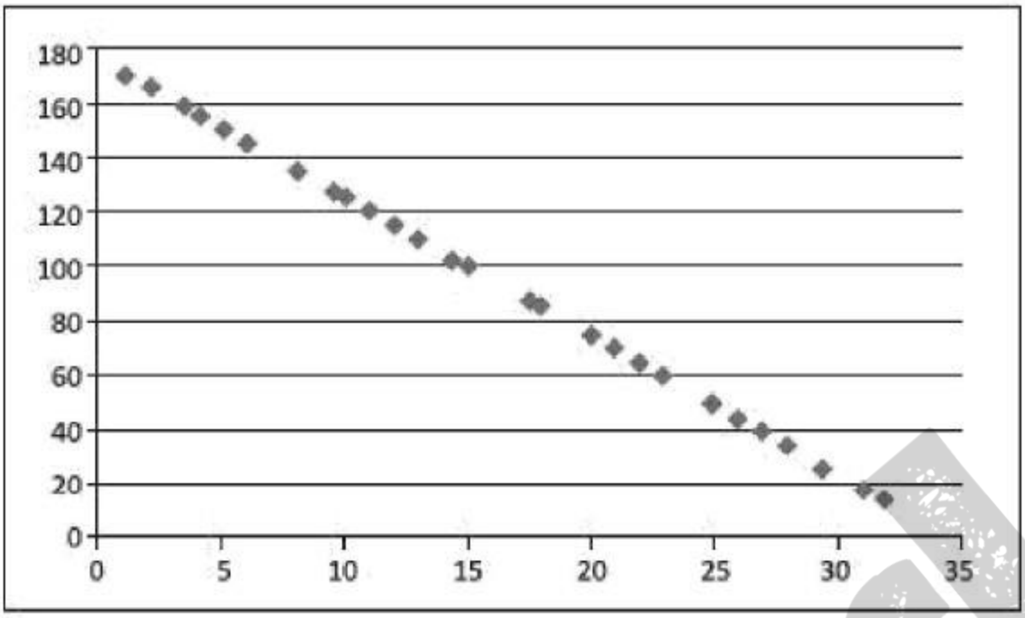
Value of r	Strength of relationship
-1.0 to -0.5 or 1.0 to 0.5	Strong
-0.5 to -0.3 or 0.3 to 0.5	Moderate
-0.3 to -0.1 or 0.1 to 0.3	Weak
-0.1 to 0.1	None or very weak

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

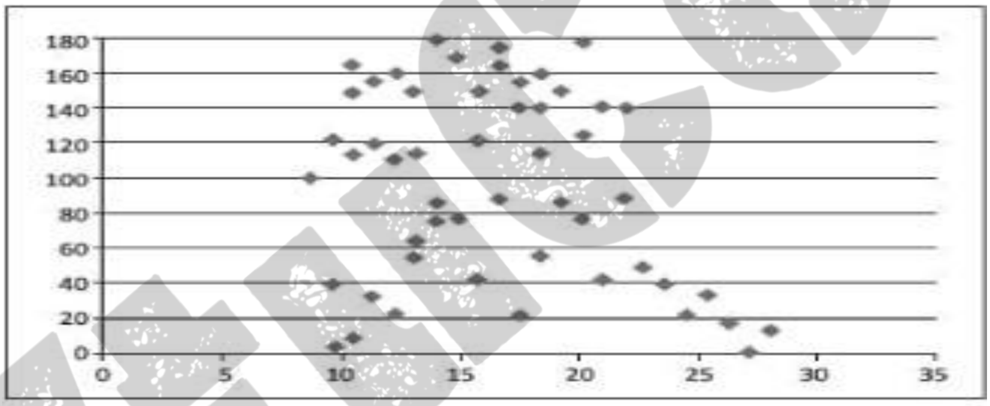


Perfect Positive Linear Relationship (r = 1)

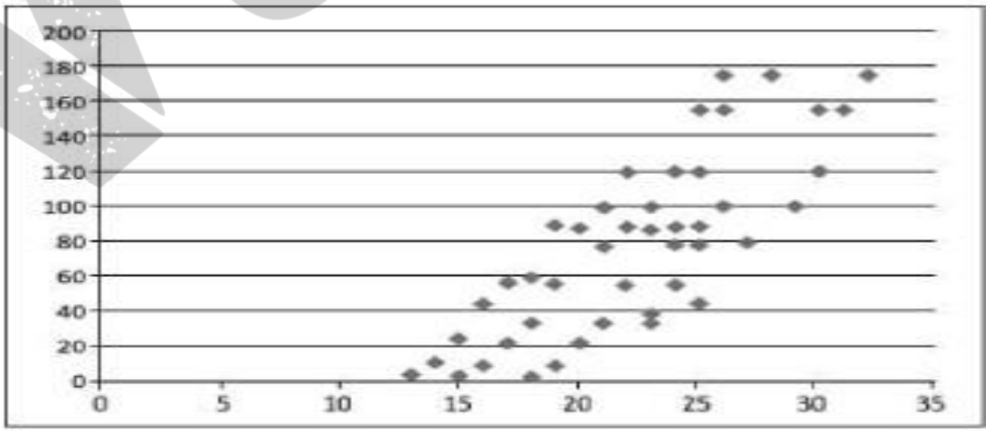
VTUCode



Perfect Negative Linear Relationship ($r = -1$)



No Relationship ($r \sim 0$)



Positive Linear Relationship ($r = 0.9$)

VtuCode

Self-Assessment Exercise linked to LO 6.1

1. Define non-linear relation. Plot on the same graph, a company car sales,for its two models every year between 2012 to 2017, using the formula -— n0 + n1.xp + n2.xq^2). How will you predict the sales in 2010? Assume for first model n0 — 490, a1 - 10 and n2 — 5. Assume for second model n0 4900, n_1 = 100 and n_2 = 50. Assume, xp - 0 for year 2011, xp = 1 for 2012and xq - 6 for 2017.

2. How does the P(x) vary in normal distribution when expected mean is at x = 6.0 and standard deviation s is 1.0? Show a plot of P(x) and x and points at deviations of 1.0, 2.0 and 3.0 (means at cr, 2P and 3 cr).

3. Define mean, variance and standard deviation. How do the O^{t}h moment, 1st moment, 2nd moment and 3’d moment compute from the values and their probabilities?

4. When will you perform t-test and F-test?

5. What does variable R-squared mean? How is the correlation parameter between predicted valued and observed value evaluated? When do you use R, r, R^2 and when N?

6. Consider correlation r between two variables. How do you interpret r » 0, r • 0 and r = 0?

7. How is the inference made that two variables do not correlate?

6.3 Regression Analysis

Correlation and regression are two analyses based on multivariate distribution. A multivariate distribution means a distribution in multiple variables.Suppose a company wishes to plan the manufacturing of Jaguar cars for coming years. The company looks at sales data regressively, i.e., data of degrussian analysis using Iinear and n on- linaar regression models,K-Nearest-Noi gh bours,. an d using distance measures for prediction sprevious years' sales. Regressive analysis means estimating relationships between variables. Regression analysis is a set of statistical steps, which estimate the relationships among variables. Regression analysis may require many techniques for modeling and performing the analysis using multiple variables. The aim of the analysis is to find the relationships between a dependent variable and one or more independent, outcome, predictor or response variables. Regression analysis facilitates prediction of future values of dependent variables.It helps to find how a dependent variable changes when variation is in an independent variable among a set of them, while the remaining independent variables in the set are kept fixed.

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

VTUCode

the first two terms are considered. The following subsections describe regression analysis in detail.

6.3.1 Simple Linear Regression

Linear regression is a simple and widely used algorithm. It is a supervised ML algorithmfor 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 n_0 is a constant and n_1 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. The following example explains the meaning of linear regression.

The purpose of regression analysis is to come up with an equation of a line that fitsthrough a cluster of points with minimal amount of deviation from the line. The best-fitting line is called the regression line. The deviation of the points from the line is called an 'error'. Once this regression equation is obtained, the GPA of a student in college examinations can be predicted provided his/her high school percentage is given. Simple linear regression is actually the same as a correlation between independent and dependent variables. Figure 6.6 shows a simple linear regression with two regression lines with different regression equations. Looking at the scatter plot, two lines can fit best to summarize the relation between GPA and high school percentage.

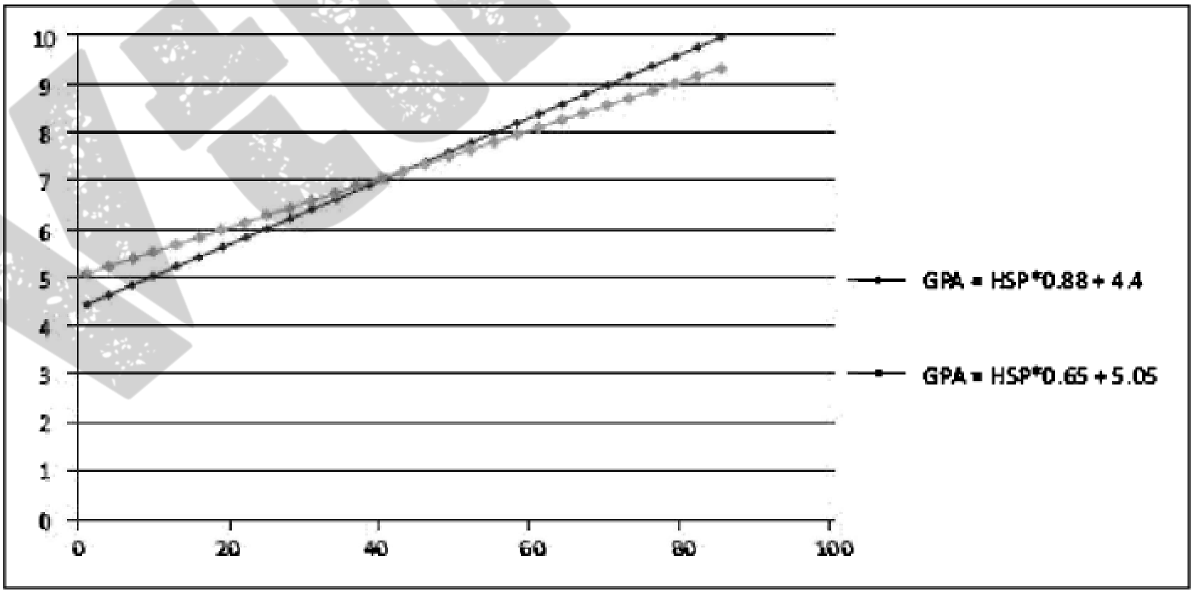


Figure 6.6 Linear regression relationship with two regression lines with different coefficient in regression equation

VTUCode

Following notations can be used for examining which of the two lines is a better fit:

1. y denotes the observed response for experimental unit i
2. x, denotes the predictor value for experimental unit i
3. y is the predicted response (or fitted value) for experimental unit i

Then, the equation for the best fitting line using a sum of the error estimating function is:

$$y'_i = a'_0 + a_1 x_i,$$

where n'$_0$ and n'l £tre the coefficients in Equation (6.10). Use of the above equation to predict the actual response 7i› leads to a prediction error (or residual error) of size:

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 b$_0$ and bl are the coefficients which minimize the errors. The coefficients values make the sum of the squared prediction errors as small as possible. Thus,

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

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 bl as follows:

$$b_0 = \overline{y} - b_1 \overline{x},$$

and

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

vtucircle

The derivative of a dependent variable with respect to the independent variable is also called a gradient. For obtaining the best-fit line here, the sum of the squared prediction error Q is minimized. Since the objective in the regression analysis is to minimize Q, Q is called objective function.

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. (Datn 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.)

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 carmodel which show high sales per month, recommending books on similar topic such as in Internet of Things by Raj Kamal from McGraw-Hill Higher Education, etc.

6.4.1.1 Appficotiox o/¥eex JVefghbouz 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 I 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):

1. Do not consider the number of items in which two users' preferences overlap. (e.g.,

2 overlap items --» 1, more items may not be better.)

2. If two users overlap on only one item, no correlation can be computed.

3. The correlation is undefined if series of preference values are identical.

Greater distance means greater dissimilarity. Dissimilarity coefficient relates to a distance metric in metrics space in v-dimensional space. An algorithm computes Euclidean, Manhattan and Minkowski distances using Equations (6.20a) to (6.20d).

Distance metric is symmetric and follows triangular inequality. Meaning of triangular inequality can be understood by an example. Consider three vectors of lengths x, y, and z.

VTUCircle

Then, triangular inequality means z • x + y. It is similar to the theorem of inequality that the third side of a triangle is less than the sum of two other sides, and never equal. The theorem applies to v-dimensional space also. Dissimilarity can be asymmetric, i.e., triangular inequality is not true (Bergman divergence).

Consider a linear search (also referred as Naive search) algorithm, Naive, one of meaning is simple in English. Search requires computations of distances to every other point. The algorithm running time is large. The time function, 0 (v.c) which measures the efficiency of the search algorithm in terms of means v.c. The v is dimensionality of M and c is cardinality of S. Cardinality refers to the number of relationships. For example, one independent variable and two dependent variables in a relationship, then cardinality is 3. Cardinality in the context of databases means the uniqueness of values contained in a column fields.

Note: Space partitioning followed by the search algorithm is an efficient method using ak-d tree or R-tree data structure. Search is made after arranging the tree-like data structure.Space partitioning problems become complex in case of high dimensionality.Naive search algorithm outperforms space partitioning approaches when using high dimensionalspaces M and high cardinality.2

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 intersection B means the number of elements or items that are same in sets A and B. A U 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 of60 students. Set CS 30 students opted for Java out of 50 students. Jaccard similarity coefficient JJava (CA, CS) = 30/(60 + 50) x 100% = 27%. Two sets are sharing 27% of the members for Java course.(H is symbol for intersection in set theory. U is symbol for union in set theory.)

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:

VTUCode

- 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. The following subsections describe the distance measures.

6.4.4.1 Definition o/a Distance

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 0/ on the y axis is Distance=4.47, 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 amathematical 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.

Consider a chart between number of students passing in examination with best grades vslanguages C••, Java, Node.js and Python. Languages are in non- metric space. They become measurable when numbers, say 0, 1, 2 and 3 are assigned for a language for the purpose of using distance measure for similarity analysis.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=1/2 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

VTUCode

similarities between users.

6.4.4.2 Euclidean Distance

Euclidean distance

$$D_{\mathrm{Eu}} = \left[\sum_{i=1}^{\nu} (x_i - x_i')^2\right]^{1/2}$$

6.4.4.3 Jaccard Distance

Equation (6.22) gives (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.

$$\mathbf{D}_{\mathrm{Ja}}(\mathfrak{A}, \mathfrak{B}) = 1 - \mathbf{J}(\mathfrak{A}, \mathfrak{B})$$

6.5 FREQUENT ITEMSETS 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 datamining. Data mining mainly deals with the type of patterns that can be mined. A method of mining is Frequent Patterns (FPs) mining method. Frequent patterns occur frequently in transactional data.

Frequent itemset refers to a set of items that frequently appear together, for example, Python and Big Data Analytics. Students of computer science frequently choose these subjects for in-depth studies. frequent itemset refers to a frequent itemset, which is a subset of items that appears frequently in a dataset.

Frequent Itemset Mininy (FIM) refers to a data mining method which helps in discoveringthe itemsets that appear frequently in a dataset. For example, finding a set of students who frequently show poor performance in semester examinations. Frequentsubsequence 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 technique has been an essential part of data analysis and data mining. 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.

VtuCode

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

The analysis results in:

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

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) (Section 6.8). A formal statement of the association rule problem is:

Let I - (f_1, f_2, ..., Id) be a set of d distinct attributes, also called literals. Let T -{t_1, t2,..., UJbe set of n transactions and contain a set of items such that T U 1. An association rule is an implication of the form, X Y, where X, Y belong to sets of items called itemsets (X, Y subset of I), and X and Y are disjoint itemsets (X intersectionY = e). Here, X is called antecedent, and Y consequent.

Explanation:

1. U means 'subset of, O means 'proper (strict) subset of, H means intersection and m means disjoint, no commonality in members.

2. Consider an If () then () form of a rule. The If part of the rule iA) 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 ñ 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. Apriori uses breadth-first search method and a hash tree structure to count candidate itemsets. Also, it is assumed that items within an itemset are kept in lexicographic order. The algorithm identifies the frequent individual items in the database and extends them to larger and larger itemsets as long as those itemsets are found in the database. The frequent itemsets provide the general trends in the database as well.

6.5.4 Evaluation of Candidate Rules

Apriori algorithm evaluates candidates for association as follows:

```
Cl: Set of candidate-itemsets of size k
F1: Set of frequent itemsets of size k
F1: (large itemsJ
for (k=1; Fk !- o; k++) do I
Ck+1=New candidates generated from Fk
for each transaction t in the database do
```

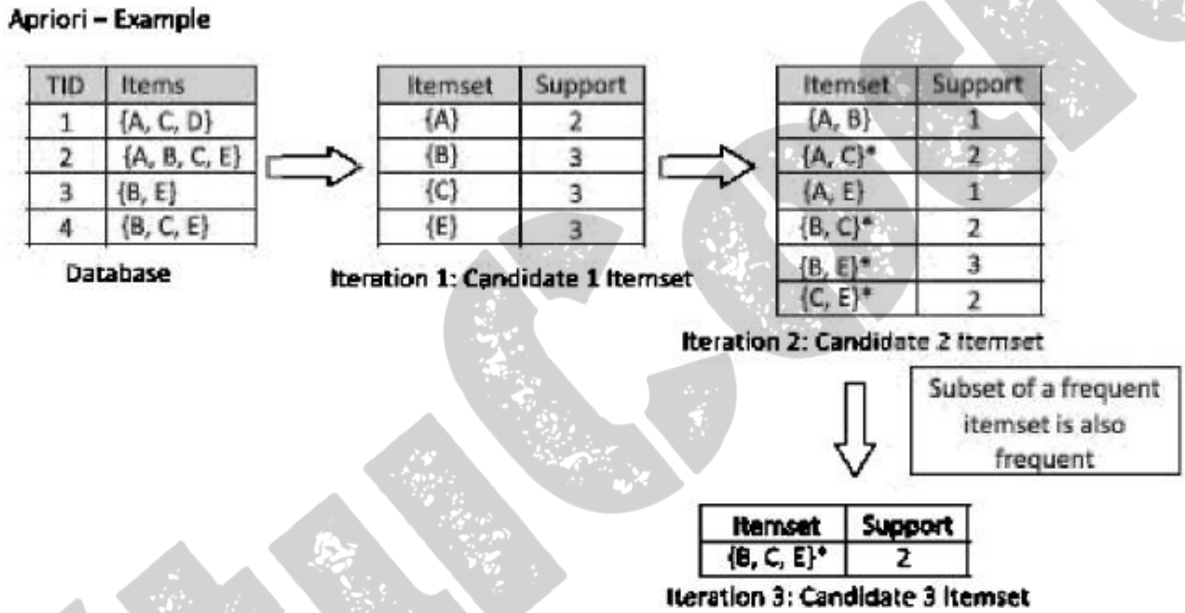
Increment the count of all candidates in Ck+l that are contained in t Fk+1= Candidates in Ck+1 with minimum support

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.



Apriori – Example

TID	Items
1	{A, C, D}
2	{A, B, C, E}
3	{B, E}
4	{B, C, E}

Database

Itemset	Support
{A}	2
{B}	3
{C}	3
{E}	3

Iteration 1: Candidate 1 Itemset

Itemset	Support
{A, B}	1
{A, C}*	2
{A, E}	1
{B, C}*	2
{B, E}*	3
{C, E}*	2

Iteration 2: Candidate 2 Itemset

Subset of a frequent itemset is also frequent

Itemset	Support
{B, C, E}*	2

Iteration 3: Candidate 3 Itemset

Figure 6.8 shows 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. Aco-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.

VtuCode

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 can be applied to many areas. The following example explains the market basket model using application examples.

EXAMPLE 6.8

Suggest application examples of the market basket model.

SOLUTION

Application 1:

1. Items - Products

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

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

- Run sales on flowers; raise price of chocolates.

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

Application 2:

2. Items - Words Baskets - Web pages

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

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 analysis is the mining of transaction data to identify relations between different products. This is normally performed to Applications of FIA1in markei analytics, medical analytics..web usage analytics, fraud direction.. clickstream analytics identify products that a customer is likely to buy, given the products that they have already bought (or added to basket).

VTUCode

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. The analysis can also help to reveal biologically relevant associations between different genes. Further, it can also help to find the effect of environment on gene expressions.

- Web usage analytics: FIM approaches can be used with viewing data on websites. The information contained in association rules can be exploited to learn about website browsing of visitor's behavior, developing website structure by making it more effective for visitors,or improving web marketing promotions. The results of this type of analysis can be used to inform website design (how items are grouped together) and to power recommendation engines(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. 5ome 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. It facilitates an understanding of the behavior of website visitors. This knowledge can be used to enhance the way that web pages are interconnected or for increasing the sales of the commercial websites.

- Telecommunication services analysis: Market basket analysis can be used to determine the type of services being utilized and the packages customers are purchasing. This knowledge can be used to plan marketing strategies for customers who are interested in similar services. 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. Plagiarism detection can find similarities among statements that may lead to similar paragraphs if all statements are similar and that possibly lead to similar documents. 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 fiindinp Association

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

The following examples give rules between items found associated in the sales data of a retailer.

EXAMPLE 6.9

Suggest association rules between items found in the sales data of a retailer,

and rules for course choice for a computer science student in college.

SOLUTION

1. {Bread)   {Butter)

The rule suggests a relationship between the sales of bread and butter. A customer who buys bread also buys butter.

2. (Chocolates)   (a Gift Box)

The rule suggests a that relationship between the sales of chocolates and empty gift boxes exists. A customer who buys chocolates also buys a gift box.

3. {Java programming)   {advanced web technology) and {Python programming)   {Big Data Analytics)

The rules suggest relationships between Java and advanced web technology, and Python programming and data analytics. 5tudents who opt for Java programming also want to learn advanced web technology, and those who opt for Python programming also opt for Big Data Analytics.

4. (Data Mining)   (Data Visualization)

The rule may be that 90% of students who select data mining as a major subject will opt for the data visualization course as well.

5. {Computer Graphics, Modeling Techniques)   {Animation)

The rule may be that students who study computer graphics and modeling techniques courses are likely to choose the course on animation in higher semesters.

VTUCode

Association analysis is applicable to several domains. Some of them are marketing, bioinformatics, web mining, scientific data analysis, and intrusion detection systems. The applications might be to find: products that are often purchased together, types of DNA sensitive to a new drug, the possibility of classifying web documents automatically, geophysical trends or patterns in seismicity to predict earthquakes and automate the malicious detecting characteristics.

In medical diagnosis, for example, considering the co-morbid (co-occur) conditions can help in treating the patient in better way. This helps in improving patient care and medicine prescription.

9.1 INTRODUCTION

Text Analytics often termed as ‘text mining’ refers to analysing and extracting the meanings, patterns, correlations and structure hidden in unstructured and semi-structured textual data. Text data stores consist of strong temporal dimensions, have modularity over time and sources, such as topics and sentiments.

Methods of machine-learning are prevalent in text analytics also. For example, when a user books an air-flight ticket using a tablet or desktop, the user receives an SMS on the mobile about details of the booking and flight timings. An ML algorithm, such as Windows Crotona at the mobile reads

and learns by itself from the SMSs received at the phone. Crotona uses the ML for the SMS text analysis. learning results in SMS alerts to the user. An alert is reminder a day before the flight. Another alert is two hours before the flight, about the need to reach the airport. Those alerts are system-generated without prior request from the user.

The reader is required to know the meaning of the following select key terms:

Vector refers to an entity with number of interrelated elements. for example, a data point consists of n-elements in an n- dimensional space, and represents a vector to that point from the origin in the space. A word is a vector of characters as the elements. Consider a vector representation of word ‘McGraw-Hill’, then vector V_{MH} = [M, c, G, r, a, w, -, H, i, l, l]. V_{MH} is vector of 11 elements (characters) that refers to word McGraw-Hill.

Feature refers to a set of properties associated with an entity, object or category. for example, feature of properties, such as description of data analysis, data cleaning, data visualization and other topics in a book on data analytics.

Category refers to a classification on the basis of set of distinct features (for example, a category of text, document, cars, toys,students, news or fruits).

Label refers to a name assigned to a category, for example to sports-news, latest data analytics books.

VTUCode

Dimension refers to a number of associated values, features or states, along the distinct spaces (dimensions). I-or example, a sentence has a number of words, each word has a number of characters, each word may have a feature, and so the sentences are in a three-dimensional space.

Two dimensions are in metric spaces, which mean values in quantifiable spaces, such as the number of words, probability of occurrences in sentences, etc. Third dimension is in feature space, measured by a feature such as noun, verb, adverb, preposition, punctuation marks and stop word.

Graph data model refers to the data modelled by a set of entities. The entities identify by vertices V. A set of relations or associations identifies by edges E. An edge e represents a relation or association between two entities. Nodes represent the entities in the graph. The model also represents a hierarchy between the parent and children nodes.

Graph data network organization refers to a structure created by organizing entities or objects in a network, such as social network, business network and student network. A network organization means where persons or entities interconnect with each other, and have areas of common interest, business or study. A graph enables ease in traversing from one entity, person or web page link to another in the network by following a path. Web graph and social network graph enable such analysis. A graph network organization models the web and social networks. Examples of social networks are SlideShare, Linkedln, Facebook and Twitter. The analytics of social networks finds the link ranks, clusters and correlations. The analytics discovers hubs and communities.

Web content mininq refers to the discovery of useful information from web documents and services. Search engines use web content mining. A search provides the links of the required information to the user.hyperlinks refer to links mentioned in the contents that enable the retrieval of contents at web, file, object or resources repository.

Link analytics means web structure mining of hyperlinks between web documents. The analytics of links and analyzing them for metrics such as page ranks, clusters, correlations, hubs and communities. Count triangles Alqorithm is an algorithm that finds a number of triangular relationships among the nodes. Triangular relationships mean interrelations between each other.

Graph node centrality metric means the centrality of a node in reference to other nodes using certain metrics. Metrics used for centrality of a node are degree, closeness, betweenness or other characteristics of the node, such as rank, belief, potential, expectation, evidence, reputation or status.

Degree centrality of a node refers to the number of direct connections. Having more number of direct connections is not always a better metric. Better measure is the fact that the connection directs to significant results and tell how the nodes connect to the isolated node.

Betweenness centrality is a measure that provides the extent to which a node lies on paths between other nodes. A node with high betweenness signifies high influence over what flows in the network indicating importance of link and single point of failure.

VtuCode

Closeness centrality is the degree to which a node is near all other nodes in a network (directly or indirectly). It reflects the ability to access information through the network.The present Chapter focuses on text, web, contents, structure and social network graph analytics.

9.2.1 Text Mining

Four definitions are:

1. "Text mining refers to the process of deriving high-quality information from text." (Wikipedia)

2. "Text mining is the process of discovering and extracting knowledge from unstructured data." (National Center of Text Mining The University of Manchesterl)

3. "Text mining is the process of analyzing collections of textual contents in order to capture key concepts themes, uncover hidden relationships, and discover the trends without

requiring that you know the precise words or terms that authors have used to express those concepts." (IBM2)

4. "Text mining is a technique which helps in revealing the patterns and relationships in large volumes of textual content that are not visible to the naked eye, leading to new business opportunities and improvements in processes." (Amazon BigData Official Blog3)

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

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.

9.2.1.1 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

vtucode

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.

9.2.1.2 Areas and Applications of Text Mining

Natural Language Processing (NLP) is a technique for analysing, 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.

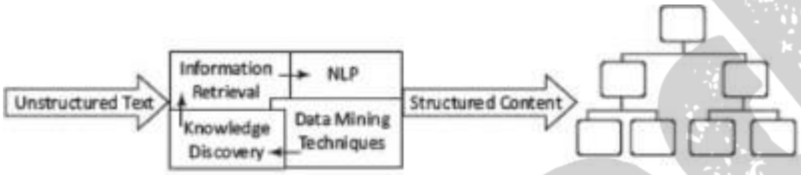


Figure 9.1 Text analytics process pipeline

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. lE finds the relationship within text or desired contents from text. lE 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.

9.2.1.3 Text Mininp Process

Text is most commonly used for information exchange. Unlike data stored in databases, text is unstructured, ambiguous and difficult to process. Text mining is the process that analyzes a text to extract information useful for a specific purpose. Syntactically, a text document comprises characters that form words, which can be further combined to generate phrases or sentences. Text mining steps are (i) recognizing, extracting and using the information present in words. Along with searching of words, mining involves search for semantic patterns as well.

Text mining process consists of a process-pipeline. The pipeline processes execute in several phases. Mining uses the iterative and interactive processes. The processing in pipeline does text mining efficiently and mines the new information. Figure 9.2 shows five phases of the process pipeline.

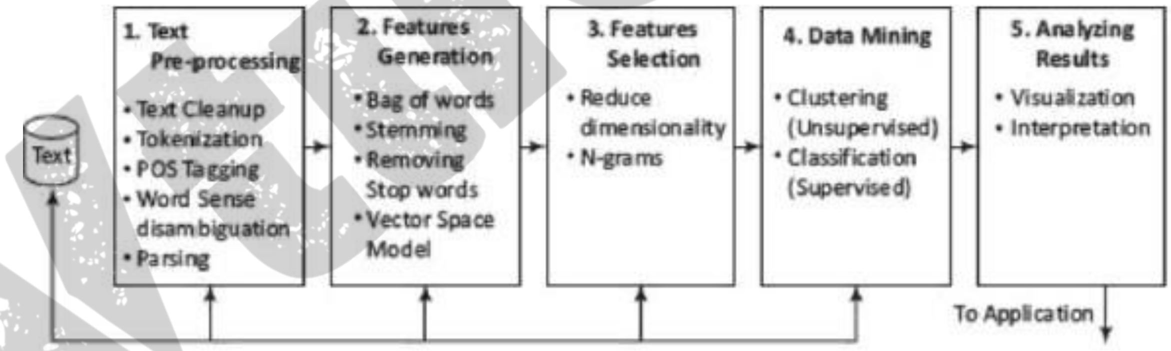


Figure 9.2 Five phases in a process pipeline The following subsection describes these phases:

9.2.1.4 Text Mining Process Phases

The five phases for processing text are as follows:

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

1. Text cleanup is a process of removing unnecessary or unwanted information. Text cleanup converts the raw data by filling up the missing values, identifies and removes outliers, and resolves the inconsistencies. For example, removing comments, removing or escaping “m20” from URL for the web pages or cleanup the typing error, such as teh (the), do n't (do not) [9t20 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. Port of Speech (POS) t«gqinq 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. Stemminq—identifies a word by its root.

(i) Normalizes or unifies variations of the same concept, such as speed 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 ‹{or, it, in and are.

4. Vector 6poce 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:

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

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-grom ev‹iluotion—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 end evuluotion o{ 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, patternsand trends.

2. Supervised feorninq (for ex‹imple, classipcation)

(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 npplicotion, where it is required to automatically assign incoming documents to pre-defined categories; emoil spom altering, where it is identified whether incoming email messages are spam or not.

Example of text classification methods are Naïve Bayes Classifier and SVMs.

3. IdentiJinq 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.

Open source tools, such as rifts are available for text analytics. Online contents accompanying book describe how text analytics tasks can be performed using Python library rifts.

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

VtuCode

(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

9.2.2 Naive Bayes Analysis

Waive 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. Waive Bayes classifies using Bayes theorem along with the Naive independence assumptions (conditional independence). The classified computes condition probabilities for the conditional independence (Refer Section 8.3.2).

When compared with other techniques, such as Random Forest, Max Entropy and SVM, the Waive Bayes classified performs efficiently in terms of less CPU and memory consumption. Naive Bayes classified requires a small amount of training data to estimate the parameters. The classifier is not sensitive to irrelevant features as well. Furthermore, the training time is significantly smaller with Naive Bayes as opposed to other techniques.The classifier is popularly used in a variety of applications, such as email spam detection, personal email sorting, document categorization, language detection, authorship identification, age/gender identification and sentiment detection.

9.2.3 Support Vector Machines

Support vector machines (SVM) is a set of related supervised learning methods (the presence of training data) that analyze data, recognize patterns, classify text, recognize hand-written characters, classify images, as well as bioinformatics and bio sequence analysis.

A vector has in general n components, x2, x3, ..., xn. A datapoint represents by (Xl, X2, ..., Xn) in n-dimensional space. Assume for the sake of simplicity, that a vector has two components, Xl and X2 (Two sets of words in text analysis).

A hyperplane is a subspace of one dimension less than its ambient space in geometry (Figure 6.18). lf a space is 3-dimensional then its hyperplanes are 2-dimensional planes, while if the space is 2-dimensional, its hyperplanes are l-dimensional, which means lines.

The hyperplane which separates the two classes most appropriately has maximum distance from closest data points of the distinct classes. This distance is termed as margin. Figure 9.3 shows the concept of support vectors, separating hyperplane and margins when using B as a classifier. The margin for hyperplane B in Figure 9.3 is more as compared to two hyperplanes, A and C shown by dotted lines. The margin of the data points kom B is maximum. Therefore, the hyperplane B is the maximum margin classifier.

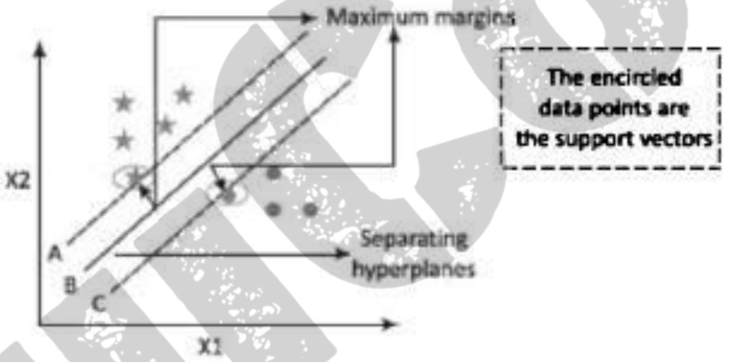


Figure 9. 3 Support vectors, separating hyperplane (B) and margins A and C are closest (least margins) to the data points. These are called the support vectors. They support the classifications of the star and dotted data points. [Remember that with n-dimensional datapoints space, a hyperplane has the vectors along (n - 1) axes.]

The support vectors are such that a set of data points lies closest to the decision (classification) surface (or hyperplane). Those points are most difficult to classify. They have direct bearing on the optimum location of the classification surface. Support vectors along maximum margin classification surface are thus gives the best results.

Thus, a SVM classifier is a discriminative classifiers formally defined by a separating hyperplane.

The concept applies extensively in number of application areas of ML. Applications of SVMs are as follows:

1. Classification based on the outputs taking discrete values in a set of possible categories, SVM can be used to separate or predict if something belongs to a particular class or category. SVM helps in finding a decision boundary between two categories.

2. Regression analysis, if learning problem has continuous real-valued output (continuous

values of x, in place discrete n values,(Xj X2› X3› ., Xp)

3. Pattern recognition

4. Outliers detection.

9.3 WEB MINING, WEB CONTENT AND WEB USAGE ANALYTICS

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

Features of web data are:

1. Volume of information and its ready availability

2. Heterogeneity

3. Variety and diversity (Information on almost every topic is available using different forms,

such as text, structured tables and lists, images, audio and video.)

4. Mostly semi-structured due to the nested structure of HTML code

5. Hyperlinks among pages within a website, and across different websites

6. Redundant or similar information may be present in several pages

7. Mostly, the web page has multiple sections (divisions), such as main contents of the page, advertisements, navigation panels, common menu for all the pages of a website and copyright notices

8. A web form or HTML form on a web page enables a user to enter data that is sent to a server for processing

9. Website contents are dynamic in nature where information on the web pages constantly changes, and fast information growth takes place such as conversations between users, social media, etc.

The following subsections describe web data mining and analysis methods:

9.3.1 Web Mining

Data Mining is a process of discovering patterns in large datasets to gain knowledge. The process can be shown as [Raw Data —• Patterns —• Knowledge]. Web data mining is the mining of web data. Web mining methods are in multidisciplinary domains: (i) data mining,

VtuCode

ML, natural language, (ii) processing, statistics, databases, information retrieval, and (iii) multimedia and visualization.

Web consists of rich features and patterns. A challenging task is retrieving interesting content and discovering knowledge from web data. Web offers several opportunities and challenges to data mining.

Definition of Web Mining

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

(i) Extracting the kagment 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.

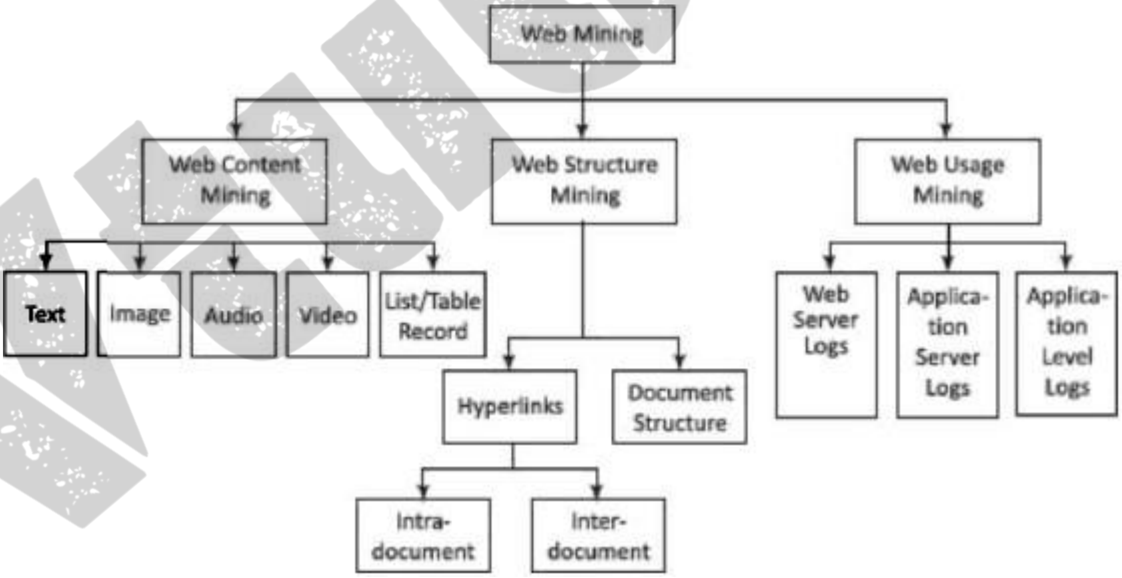


Figure 9.6 Web mining taxonomy

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:

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

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

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

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

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

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

9.3.2 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.

App1ications

Following are the applications of content mining from web documents:

1. Classifying the web documents into categories

2. Identifying topics of web documents

3. Finding similar web pages across the different web servers

VtuCode

4. Applications related to relevance:

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

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

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

9.3.3 Web Usage Mining

Web usage mining discovers and analyses the patterns in click streams. Web usage mining also includes associated data generated and collected as a consequence of user interactions with web resources. Figure s.7 shows three phases for web usage mining.

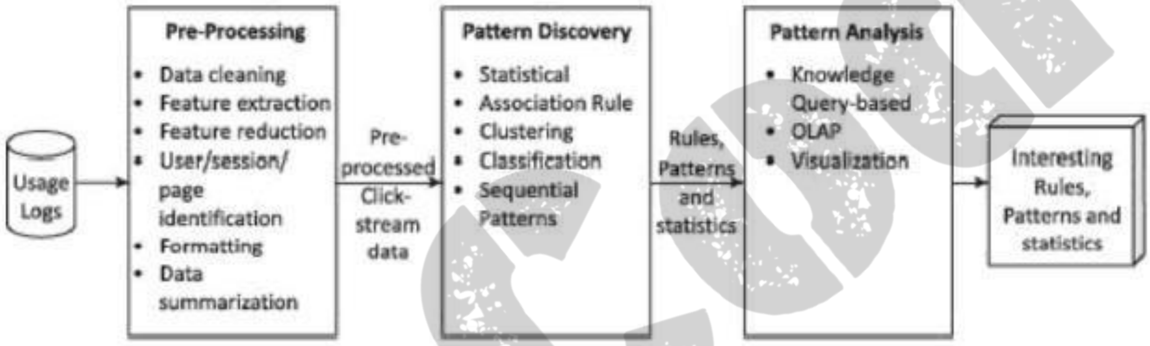


Figure 9.7 Process of web usage mining The phases are:

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.

9.3.3.1 Pre-processing

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.

Pre-processing involves several steps, such as data cleaning, feature extraction, feature reduction, user identification, session identification, page identification, formatting and finally data summarization.

9.3.3.2 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.

VTUCircle

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

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

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-commence 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.

Model-based or distance-based clustering can be applied on web usage logs. The model type is often specified theoretically with model-based clustering. The model selection techniques and parameters estimate using maximum likelihood algorithms, such as Expectation Maximization (EM) determines the structure of model. Distance-based clustering measures the distance between pairs of web pages or users, and then groups the similar ones together into clusters. The most popular distance-based clustering techniques include partitional clustering and hierarchical clustering 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, 375o users watched a certain movie, out of which 20o0 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, Naive Bayesian classifiers, k-nearest neighbour classifiers, support vector machines.

Sequential pattern discovery User navigation patterns in web usage data gather web page trails that are often visited by users in the order in which pages are visited. Markov Model can be used to model navigational activities in the website. Every page view in this model can be represented as a state. 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.

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. I-or 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).

9.4 Page Rank Definition

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

Assume a web structure of hyperlinks. Each hyperlink in-links to a number of hyperlinks and out-links to a number of pages. A page commanding higher authority (rank) has greater number of in-degrees than out-degrees. Therefore, one measure of a page authority can be in-degrees with respect to out-degrees.

PageRank refers to the authority of the page measured in terms of number of times a link is sought after.

9.4.1 Page Rank Definition

According to the new approach Earlier approach of page ranking based on in-links and out-links does not capture the relative authority (importance) of the parents. Page and co-authors (1998) defined a page ranking method,5 which considers the entire web in place of local neighbourhood of the pages and considers the relative authority of the parent links (over children).

9.4.2 Web Structure

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

1. Text at the hyperlink represents the property of a vertex u that describes the destination V of the out-going edge.

2. A hyperlink in-between the pages represents the conferring of the authority.

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

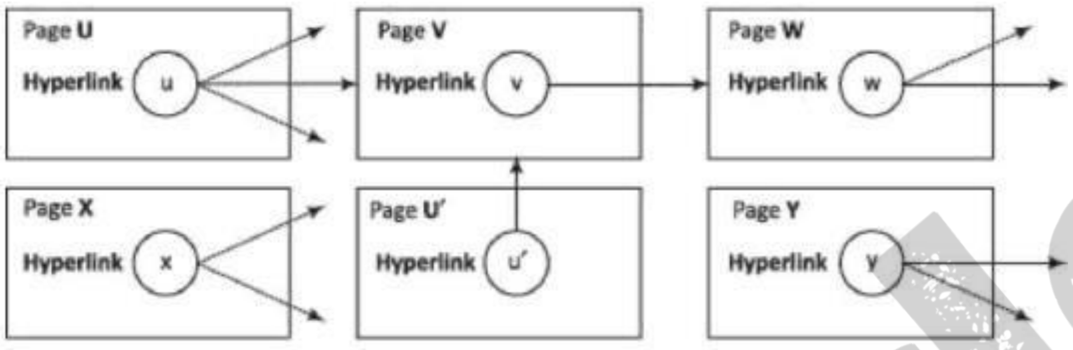


Figure 9.8 Web structure with hyperlinks from a parent to one or more pages

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. Good website structures have the pages designed such that they specifically gently guide the visitors toward actions and towards next step. For example, if one searches for a book title on Amazon, then visitor gets links of other books also on a similar topic.

9.4.2.2 Analyzing and Implementing a System Frith Web Graph Mining

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

1. In-degrees and out-degrees

2. Closeness is centrality metric. Closeness,

$$\mathrm{Cc}(v) = 1 / \sum_{u \in V} \mathrm{gdist}(v, u)$$

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).

3. Betweenness

4. PageRank and LineRank

5. Hubs and authorities

VTUCode

6. Communities parameters, triangle count, clustering coefficient, K-neighbourhood

7. Top K-shortest paths

9.4.3 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„t[(Pg (u)] pages. Figure 9.9 shows Pg (v) in-links from Pg (u) and other pages.

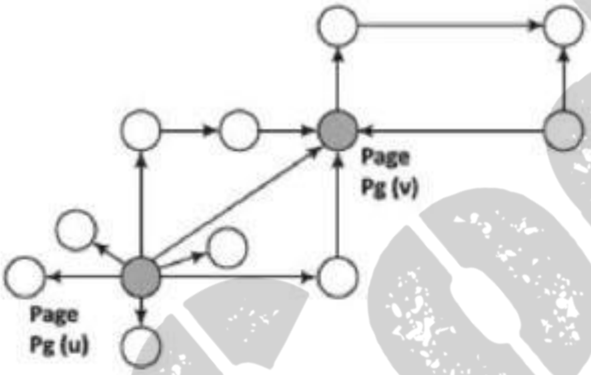


Figure 9.9 Page Pg (v) in-links from Pg (u) and other pages

Two algorithms to compute page rank are as follows:

1. PageRank algorithm using the in-degrees os 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:

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

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

VtuCode

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 I 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

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

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 of [1/PA(V)]

An alternative equation is as follows:

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

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 n 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).

9.4.4 Topic Sensitive PageRank and Link Spam

Number of methods have been suggested for computations of topic-sensitive page ranking, RTS. The RSS (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.

9.4.5 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 9.1o(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.

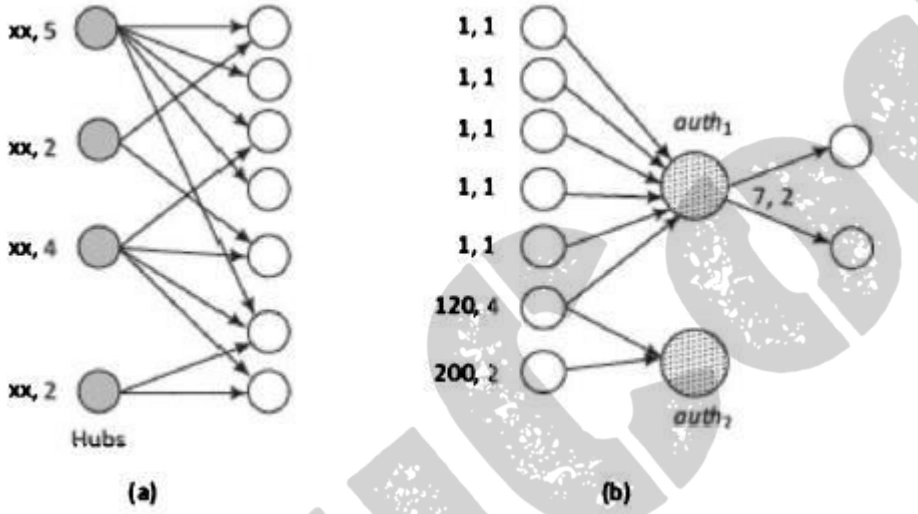


Figure 9.10(a) Hubs (shaded circles) and (b) Authorities (dotted circles)

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, in Figure 9.10(b) has in-links from 6 vertices (in-degrees = 6) and nuth₂ hasin-links to just 2 (in-degree = 2). However, nuth, has link with six vertices with in-degrees = 1, 1, 1, 1, 1 and 120 (total = 125). Authority, nuth, has links with two vertices with in-degrees - 120 and 200 (total - 220). Auth has association with greater authorities. Therefore, in-degrees may not be a good measure as compared to authority.

Kleinberg (1998) developed the Hypertext-Induced Topic Selection (HITS) algorithm.6 The algorithm computes the hubs and authorities on a specific topic t. The HITS analyses a sub-graph of web, which is relevant to t. Basis of computation is (i) hubs are the ones, which out-link to number of authorities, and

(ii) authorities are the ones, which in-link to number of hubs. A bipartite graph exists for the hubs and authorities. Consider a specifically queried topic t. Following are the steps:

VTUCode

1. Let a set of pages discover a root set R using standard search engine. Root pages may limit to top 200 for t.

2. Find a sub-graph of pages S, using a query that provides relevant pages for t and pointed by pages at R. Sub-graph S pages form Set for computations as it includes the children of parent R and limit to a random set of maximum so pages returned by a "reverse link" query.

3. Eliminate purely navigational links and links between two pages on the same host.

4. Consider only u (i«l! - 4-8) pages from a given hyperlink as pointer to any individual page.

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.

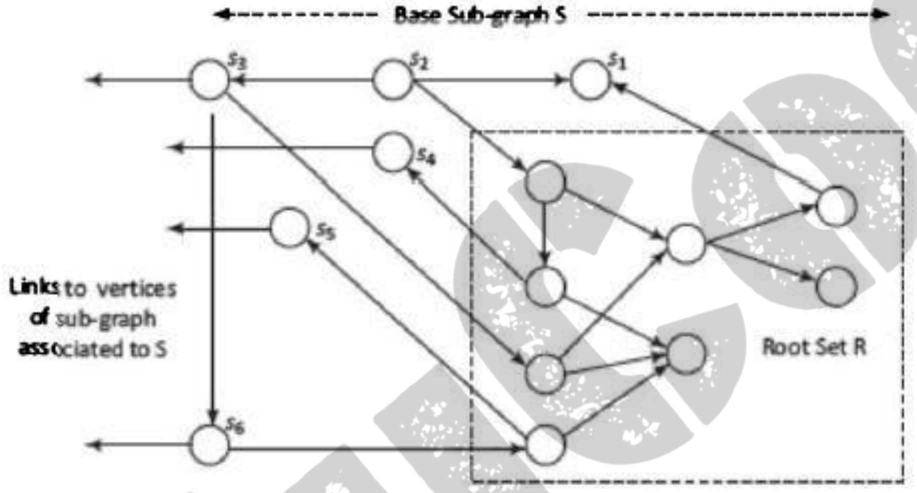


Figure 9.11 Sub-graph for HITS consisting of root set R of pages and base sub-graph S including all the pages pointed to by any page of ft.

The left directed leftmost arrows from s3, s4 , s5 and s6 are pointing to nodes in sub-graph(s) associated to S. The following example explains the algorithm steps to compute hub score and authority score.

9.5 SOCIAL NETWORKS AS GRAPHS AND SOCIAL ANALYTICS 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. The following subsections describe social networks as graph, uses, characteristics and metrics.

9.5.1Social Network as Graphs

Social network as graphs provide a number of metrics for analysis. The metrics enable the application of the graphs in a number of fields. Network topological analysis tools computethe 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 net:work 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.

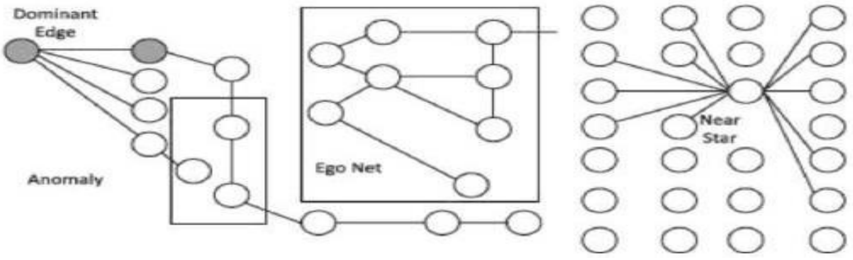


Figure 9.12 shows discovering anomaly, ego-net and spam from the analysis.

Figure 9.12 Discovering anomaly, ego-net and spam (using near star) from the analysis Social network has concerns of privacy, security and falsehood dissemination. Security issues are phishing attacks and malwares.

9.5.2 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

Deqree 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 suction means the distribution function for the degrees of vertices (Section 6.2.5 described the common distribution functions).

Closeness

Grnph vertex closeness cC (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.where d (u, v) is distance between u and v for path traversal.

Effective Closeness

Effective closeness Ce (v)can also be analyzed. Use approximate average distance from v to all other vertices in place of the shortest paths. Ces 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 cB(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 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 vl 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).

9.5.3 Social Graph Network Analysis using K-core and Neighbourhood Metrics

K-core is a sub-graph in a graph network structure. Graph Vertex Kth neighbourhood is number of $1^{s‘}$ neighbour vertices, 2^{n} 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 Jilteringq for frequent itemsets uses weighted bipartite graph matching.

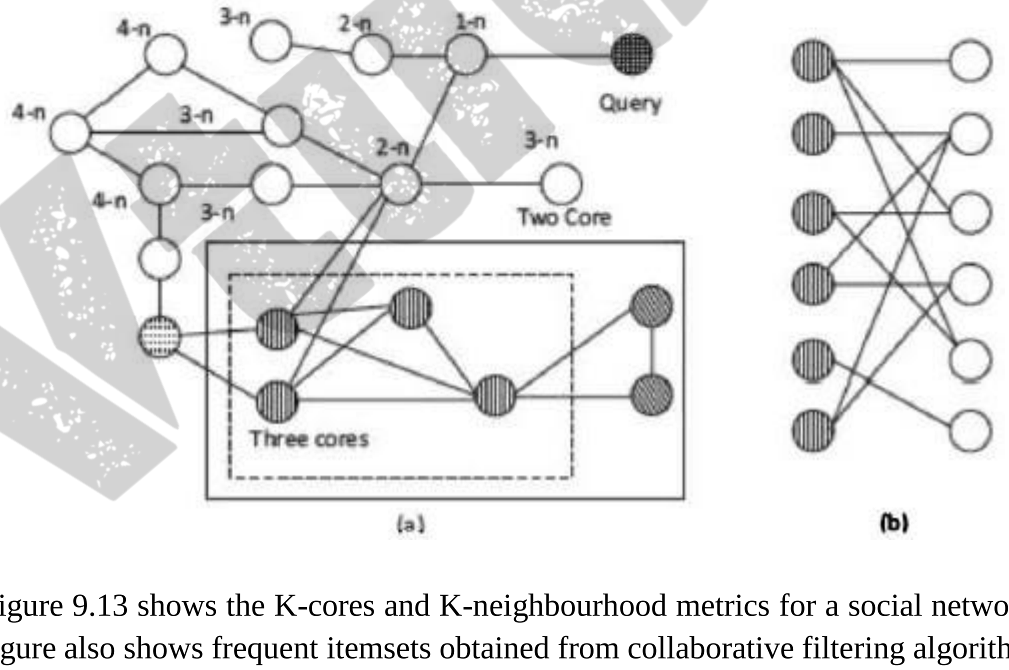


Figure 9.13 shows the K-cores and K-neighbourhood metrics for a social network graph. The figure also shows frequent itemsets obtained from collaborative filtering algorithm

Figure 9.13 (a) K-cores and K-neighbourhoods with K = 1, 2, 3 and 4 and (b) Frequent itemsets from collaborative filtering algorithm (weighted bipartite graph matching)

VTUCode

Figure 9.13(a) shows three cores of two triangles, one quadrilateral, two cores of one pentagon and one traingle) in K- neighbourhoods. K = 1, 2, 3 and 4. Figure 9.13(b) shows frequent itemsets from collaborative filtering algorithm (weighted bipartite graph matching).

9.5.4 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.

9.5.5 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.

SimRank is the metric for measuring similarity between vertices of the same type. The computation starts from a vertex possessing specific property and path traversals through the edges search the similarities. The vertices having similar properties are counted to the SimRank. The counting continues till counts per unit traversals converge within a prefixed margin, say .001. SimRank converges to a value which is applicable for path traversals within, say geodesic distance, say up to 200. The computations are analogous to ones for PageRank as in Example 9.7

9.5.6 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 s.14 shows triangles and triangles between similar graph properties found from graph matches. Edge labels show the GPAs of students socially connected.

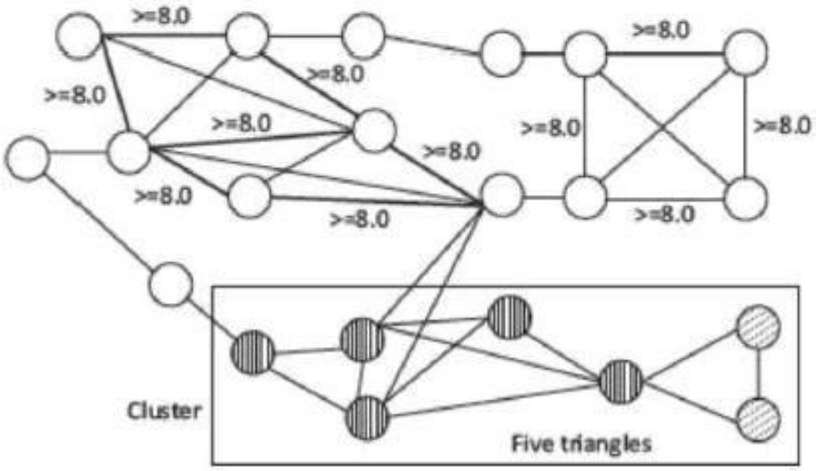


Figure 9.14 Clustering of five triangles and three matches of graphs

VtuCode