Subject: MCA DMBI 1003

Unit 4:
Different Approaches to resolving data mining problems

Made By: Priyanka Karale

Approaches to Data Mining Problems

- Discovery of Sequential Patterns
- Discovery of Patterns in Time Series
- Regression
- Neural Networks
- Genetic Algorithms

Sequential Pattern Mining

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity

Problems:

- string mining
- itemset mining

String Mining

- This is the subset of Sequence Pattern Mining that deals with text data in a sequence. The data can contain only a limited number of characters. For example, a DNA sequence contains only the letters ‘A’, ’T’, ’C’, and ’G’, and therefore analysis of the same falls within String Mining. Similarly, finding patterns in ASCII character sequences falls under String Mining.

Itemset Mining

- This is the broader subset of Sequence Pattern Mining that aims to find patterns in ordered datasets. Itemset Mining generally finds use in Marketing and Sales Applications (increasing co-purchases of items that are frequently brought together, cross-promoting products, managing inventory, setting price levels, and so on).

Methods for Sequential Pattern Mining

- Apriori-based Approaches
  - GSP
  - SPADE
- Pattern-Growth-based Approaches
  - FreeSpan
  - PrefixSpan

Applications of Sequence Pattern Mining

- Sequence Pattern Mining finds applications in multiple fields ranging from science, business, and finance to meteorology and geology. Some of them are listed below:
  - Determination of buying patterns (“If a person bought product A, he is likely to purchase product B”)
  - Stock trading (where else do people make huge bets on patterns than in the stock market?)
  - Analyzing DNA and protein sequences in computational biology
  - Studying website logs to identify a user’s online behavior
  - Predicting natural disasters based on past indicative patterns.
  - Studying telephone calling patterns

Discovery of Patterns in Time Series

- A time series is a set of attribute values over a period of time.
- A **time series** is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals.
- Typical data mining applications for time series include determining the similarity between two different time series and predicting future values for an attribute.
- Time series analysis may be viewed as finding patterns in the data and predicting future values. Detected patterns may include:
  - Trends
  - Cycles
  - Seasonal
  - Outlier

- Trends: A trend can be viewed as systematic nonrepetitive changes (linear or nonlinear) to the attribute values over time. An example would be that the value of a stock may continually rise.
- Cycles: Here the observed behavior is cyclic.
- Seasonal: Here the detected patterns may be based on time of year or month or day. As an example, the sales volumes from department stores always jump around Christmas.
- Outliers: To assist in pattern detection, techniques may be needed to remove or reduce the impact of outliers

Trend Analysis

- Smoothing is an approach that is used to remove the nonsystematic behaviors found in a time series.
- Smoothing usually takes the form of finding moving averages of attribute values.
- Smoothing is used to filter out noise and outliers

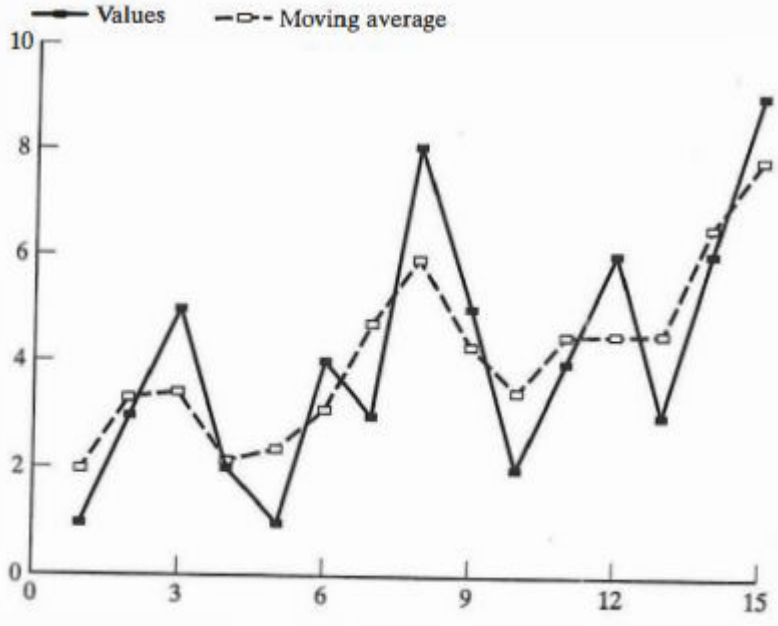


FIGURE 9.5: Smoothing using a moving average.

Example

Month	Sales ($000)	Three-month moving total ($000)	Three-month moving average ($000)
January	125		
February	145	456=(125+145+186)	(456 ÷ 3) = 152
March	186		

Month	Sales ($000)	Three-month moving total ($000)	Three-month moving average ($000)
January	125		
February	145	456=(125+145+186)	(456 ÷ 3) = 152
March	186	462=(145+186+131)	(462 ÷ 3) = 154
April	131	468=(186+131+151)	(468 ÷ 3) = 156
May	151	474	158
June	192	480	160
July	137	486	162
August	157	492	164
September	198	498	166
October	143	504	168
November	163	510	170
December	204		

Subject: MCA DMBI 1003

Time Series Analysis Example

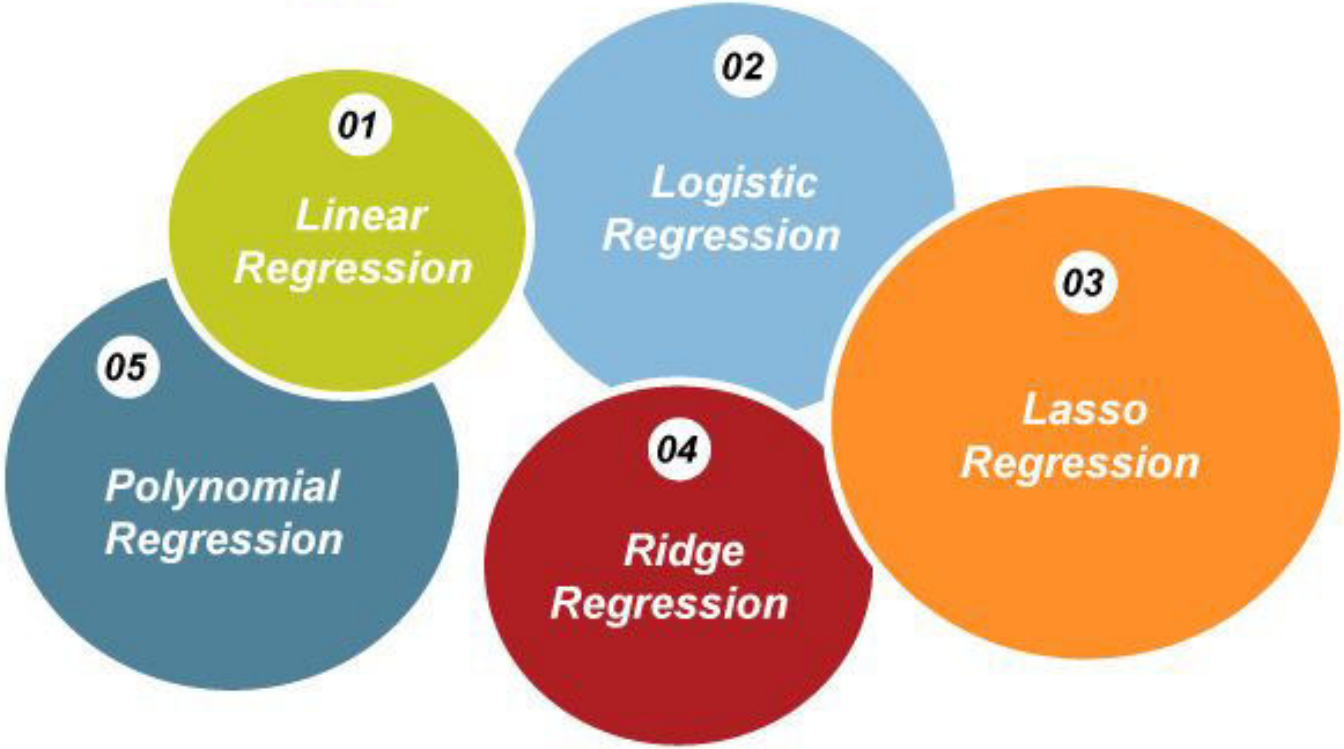
Made By: Priyanka Karale

Regression

- supervised learning method
- Regression in Data Mining involves using statistical methods to examine the connection between a dependent variable and multiple independent variables.
- It is often used to predict numerical outcomes such as a house's price or a product's sales
- In Data Mining, regression models are developed using past data and can be utilized to make predictions for new data
- In addition to prediction, regression analysis helps identify the most significant variables affecting the dependent variable and develop a mathematical equation to describe the relationship between the variables.

Subject: MCA DMBI 1003

Types of Regression



Made By: Priyanka Karale

Linear Regression

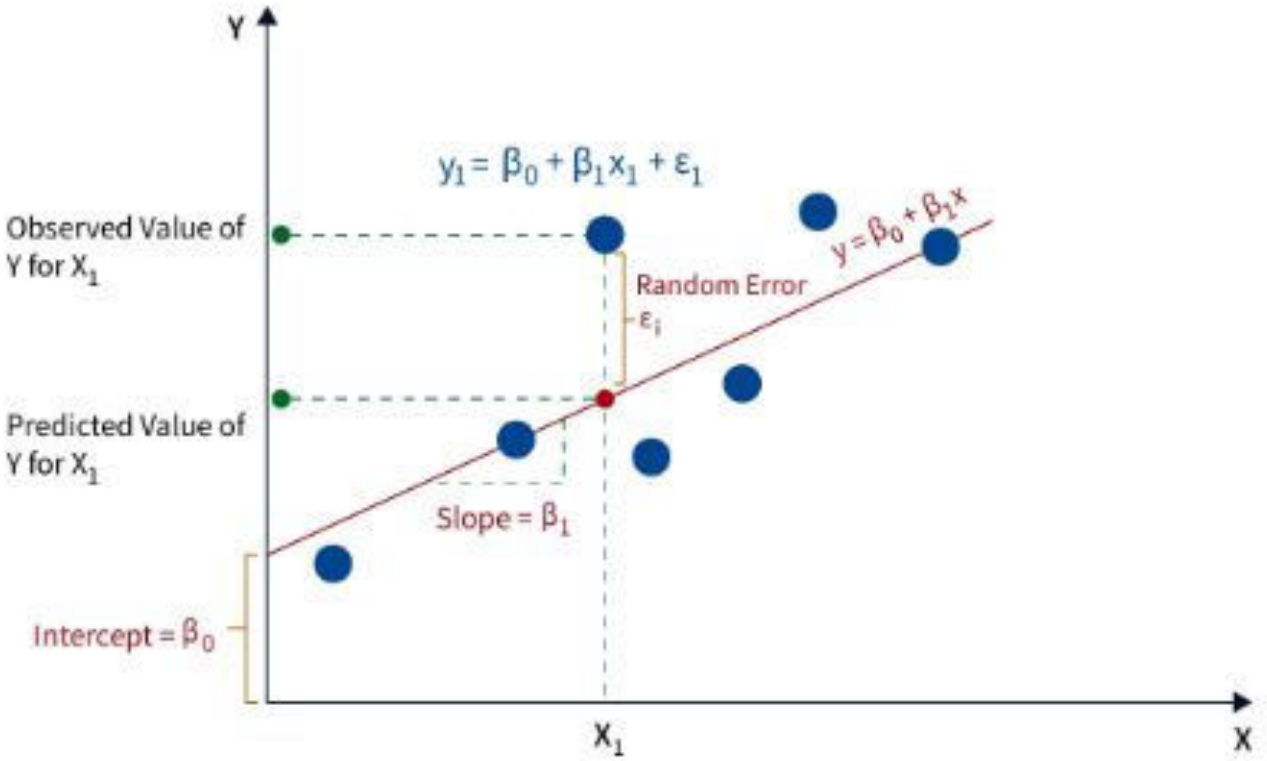
- Linear regression in data mining is a statistical technique used to model the relationship between a dependent variable and one or more independent variables, assuming a linear relationship between them. The goal is to find the best-fit line that minimizes the distance between the observed and predicted values.
- Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

- A simple linear regression is represented as an equation

$$Y = \beta 0 + \beta 1X$$

- Where Y the dependent variable and X being the independent variable, β0 the intercept term, and β1 is the slope term.
- The intercept term represents the value of Y when X = 0, and the slope term represents the change in Y for a unit change in X. (In statistics, the symbol β represents the parameters or coefficients in a regression model.)

Subject: MCA DMBI 1003



Made By: Priyanka Karale

Problem Statement:

Suppose we have a dataset that represents the relationship between the number of hours students study and the scores they receive on a test. We want to build a linear regression model to predict the test score based on the number of hours studied.

Hours Studied (x)	Test Score (y)
2	60
3	70
4	80
5	85
6	90

Step 1: Calculating the Mean

First, calculate the mean of the hours studied and test scores:

Mean of x: $\bar{x} = \frac{2+3+4+5+6}{5} = 4$

Mean of y: $\bar{y} = \frac{60+70+80+85+90}{5} = 77$

Step 2: Calculate the Slope (m)

Using the formula for the slope of a linear regression line:

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

Substituting the values:

$$m = \frac{(2-4)(60-77)+(3-4)(70-77)+(4-4)(80-77)+(5-4)(85-77)+(6-4)(90-77)}{(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2}$$

$$m = \frac{-34-7+9+40+26}{2+1+0+1+4}$$

$$m = \frac{34}{8} = 4.25$$

Step 3: Calculate the Y-Intercept (b)

Using the formula for the y-intercept of the linear regression line:

$$b = \bar{y} - m\bar{x}$$

$$b = 77 - 4.25 \times 4 = 59.5$$

Step 4: Write the Regression Equation

The linear regression equation for this problem is:

$$y = 4.25x + 59.5$$

This equation can be used to predict the test score (y) based on the number of hours studied (x). For example, if a student studies 7 hours, the predicted test score would be:

$$y = 4.25 \times 7 + 59.5 = 89.25$$

Solve This Example using linear regression

Given Data Set Consist of Chips in gms and their cost in rs.

Build the linear regression model to predict cost of chips based on their weight in gms.

Also write what will be the cost of 150 gm chips

X	Y
Chips in gm	Cost in rs.
120	5
23	10
40	20
73	30
90	50

Made By: Priyanka Karale

X mean = 47.6 = 47

Y mean = 23

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

$$m = \frac{(12-47)(5-23)+(23-47)(10-23)+(40-47)(20-23)+(73-47)(30-23)+(90-47)(50-23)}{(12-47)^2+(23-47)^2+(40-47)^2+(73-47)^2+(90-47)^2}$$

$$= \frac{630+312+21+182+1161}{1225+576+49+676+1849}$$

$$= \frac{2306}{4875} = 0.52$$

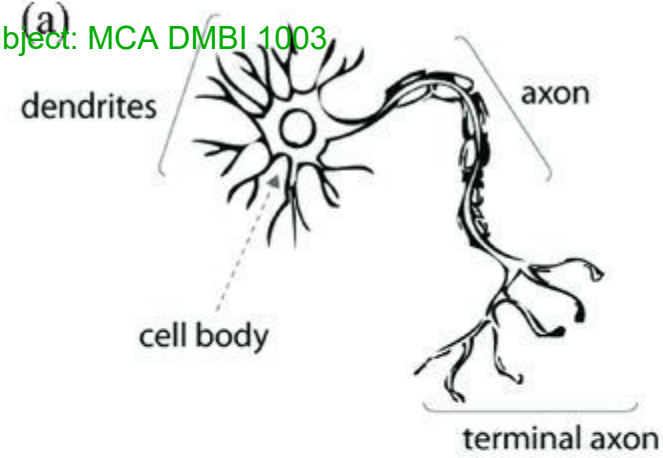
$$b = \bar{y} - m\bar{x}$$
$$= 23 - 0.52 \times 47$$
$$= -1.44$$

$$y = mx + b$$
$$y = 0.52x + (-1.44)$$

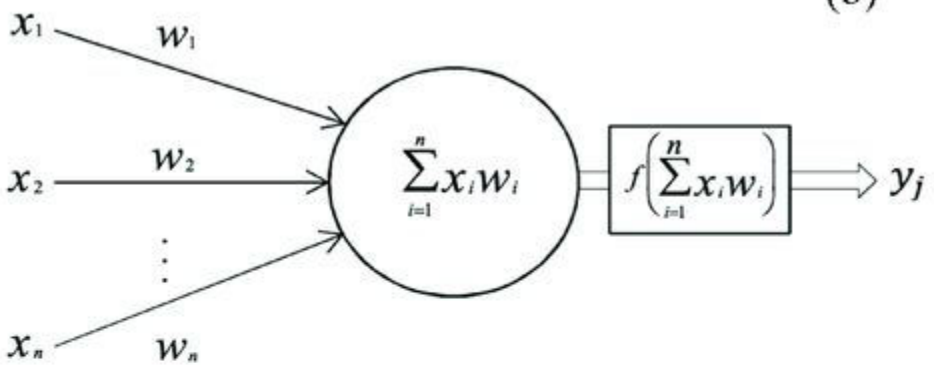
Neural Network

- A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain
- Neural networks, also known as artificial neural networks (ANNs) or simulated neural networks (SNNs)
- As in the Human Nervous system, we have Biological neurons in the same way in Neural networks we have Artificial Neurons which is a Mathematical Function that originates from biological neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. Each neuron receives signals through synapses that control the effects of the signal on the neuron.

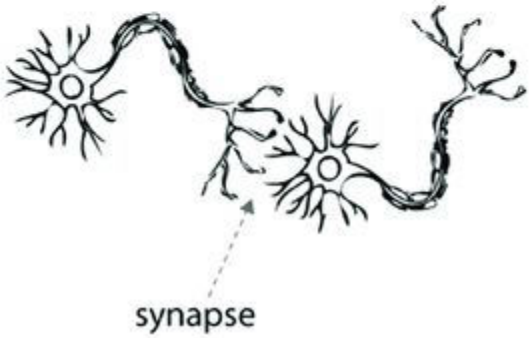
(a)



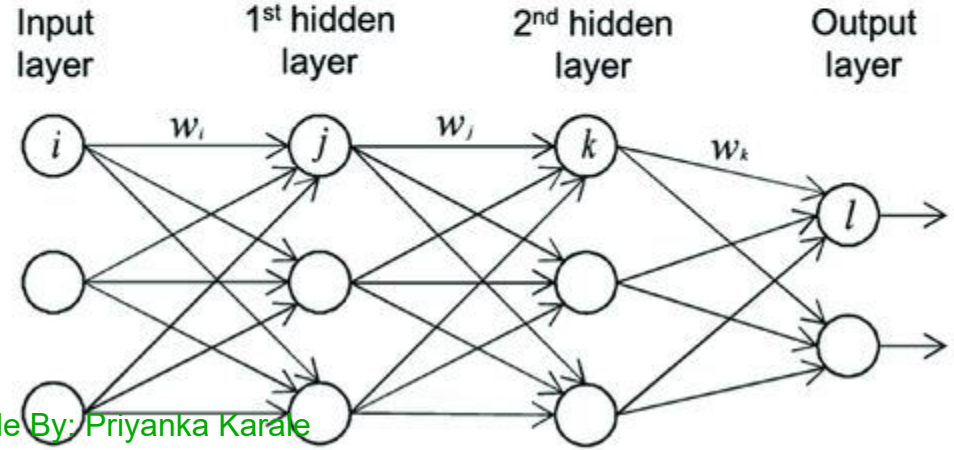
(b)



(c)

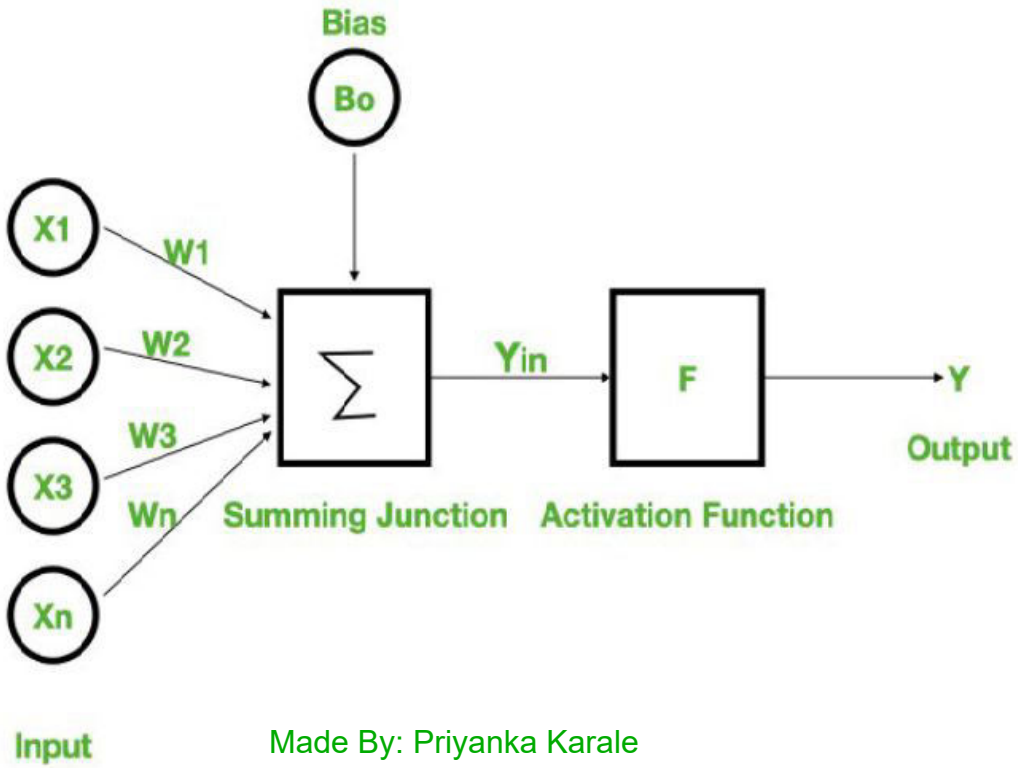


(d)



Subject: MCA DMBI 1003

Neural Network Architecture



Made By: Priyanka Karale

How Artificial Neural Network Work?

Let us Suppose that there are n input like X1,X2,…,Xn to a neuron.

=> The weight connecting n number of inputs to a neuron are represented by [W]=[W1,W2,..,Wn].

=> The Function of summing junction of an artificial neuron is to collect the weighted inputs and sum them up.

$$Yin=[X1*W1+X2*W2+....+Xn*Wn]$$

=> The output of summing junction may sometimes become equal to zero and to prevent such a situation, a bias of fixed value Bo is added to it.

$$Yin = [X1*W1 + X2*W2 + ... + Xn*Wn] + Bo$$

// Yin then move toward the Activation Function.

=> The output Y of a neuron largely depends on its Activation Function (also known as transfer function).

Neural Network Method in Data Mining

- **Feed-Forward Neural Networks:** In Feed-Forward Network, if the output values cannot be traced back to the input values and if for every input node, an output node is calculated, then there is a forward flow of information and no feedback between the layers. In simple words, the information moves in only one direction (forward) from the input nodes, through the hidden nodes (if any), and to the output nodes. Such a type of network is known as a feedforward network.

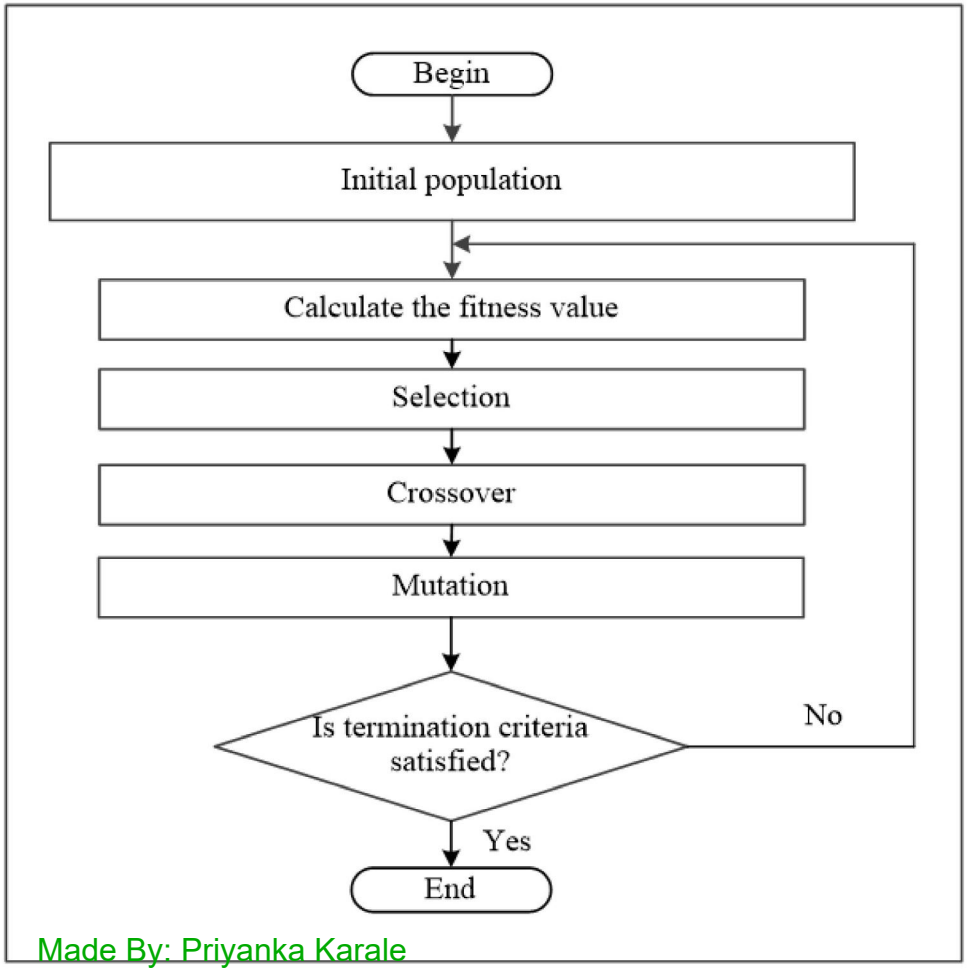
Genetic algorithm

- A **genetic algorithm** (or **GA**) is a search technique used in computing to find true or approximate solutions to optimization and search problems.
- Genetic algorithms are categorized as global search heuristics.
- Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as **inheritance, mutation, selection, and crossover** (also called recombination).
- Genetic algorithms are implemented as a computer simulation in which population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions.

- Genetic algorithms are implemented as a computer simulation in which a population of abstract representations (called chromosomes or the genotype or the genome) of candidate solutions (called individuals, creatures, or phenotypes) to an optimization problem evolves toward better solutions.
- Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible.
- Genetic algorithm is search-based optimization technique based on the principle of genetic and natural selection.

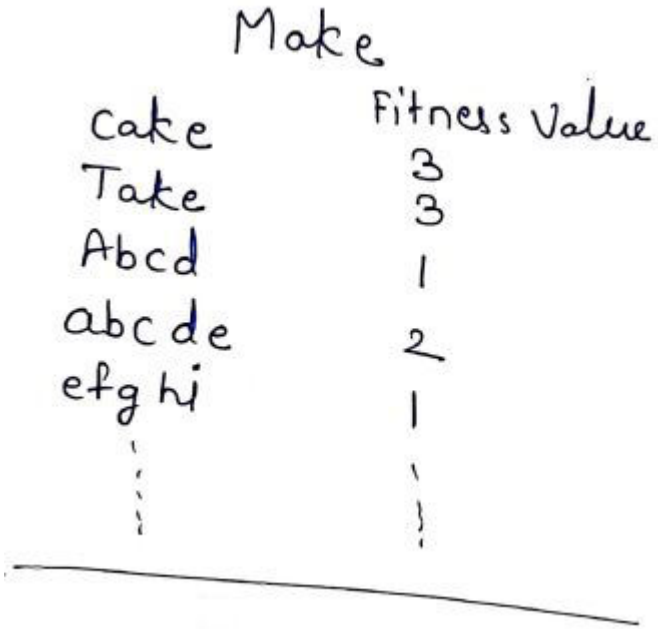
Subject: MCA DMBI 1003

Genetic Algorithm Flowchart



Made By: Priyanka Karale

Example


Make
Cake
Take
Abcd
abcde
efghi
Fitness Value
3
3
1
2
1


Crossover
Fitness
abcde
efghi
abchi
efgde
1
1
akme
kmae
4
4

Text Mining

- **What is Text Mining?**

Text mining (also known as text analysis), is the process of transforming unstructured text into structured data for easy analysis. Text mining uses natural language processing (NLP), allowing machines to understand the human language and process it automatically.

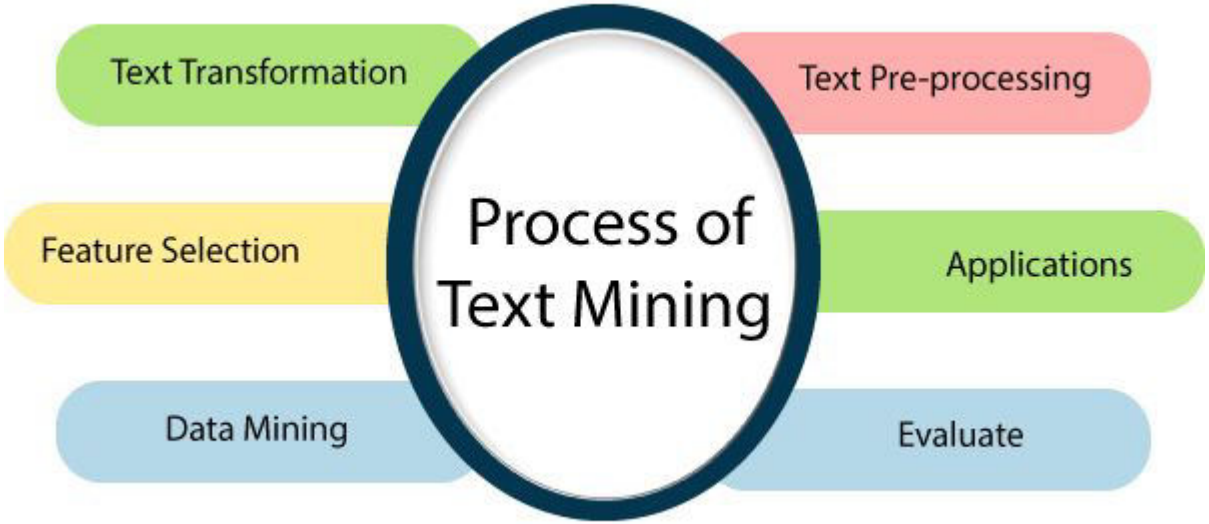
Text mining is a component of data mining that deals specifically with unstructured text data. It involves the use of natural language processing (NLP) techniques to extract useful information and insights from large amounts of unstructured text data. Text mining can be used as a preprocessing step for data mining or as a standalone process for specific tasks.

- By using text mining, the unstructured text data can be transformed into structured data that can be used for data mining tasks such as classification, clustering, and association rule mining. This allows organizations to gain insights from a wide range of data sources, such as customer feedback, social media posts, and news articles.

Subject: MCA DMBI 1003

Text Mining Process

1. Text Preprocessing
2. Text Transformation
3. Feature Selection
4. Data mining
5. Evaluate
6. Application



Made By: Priyanka Karale

Text Pre-processing

It involves a series of steps as shown in below:

- Text Cleanup

Text Cleanup means removing any unnecessary or unwanted information. Such as remove ads from web pages, normalize text converted from binary formats.

- Tokenization

Tokenizing is simply achieved by splitting the text into white spaces.

- Part of Speech Tagging

Part-of-Speech (POS) tagging means word class assignment to each token. Its input is given by the tokenized text. Taggers have to cope with unknown words and ambiguous word-tag mappings.

Text Transformation (Attribute Generation)

A text document is represented by the words it contains and their occurrences. Two main approaches to document representation are:

i. Bag of words

ii. Vector Space

Feature Selection (Attribute Selection)

Feature selection also is known as variable selection. It is the process of selecting a subset of important features for use in model creation. Redundant features are the one which provides no extra information. Irrelevant features provide no useful or relevant information in any context.

Data Mining

At this point, the Text mining process merges with the traditional process. Classic Data Mining techniques are used in the structured database. Also, it resulted from the previous stages.

Applications

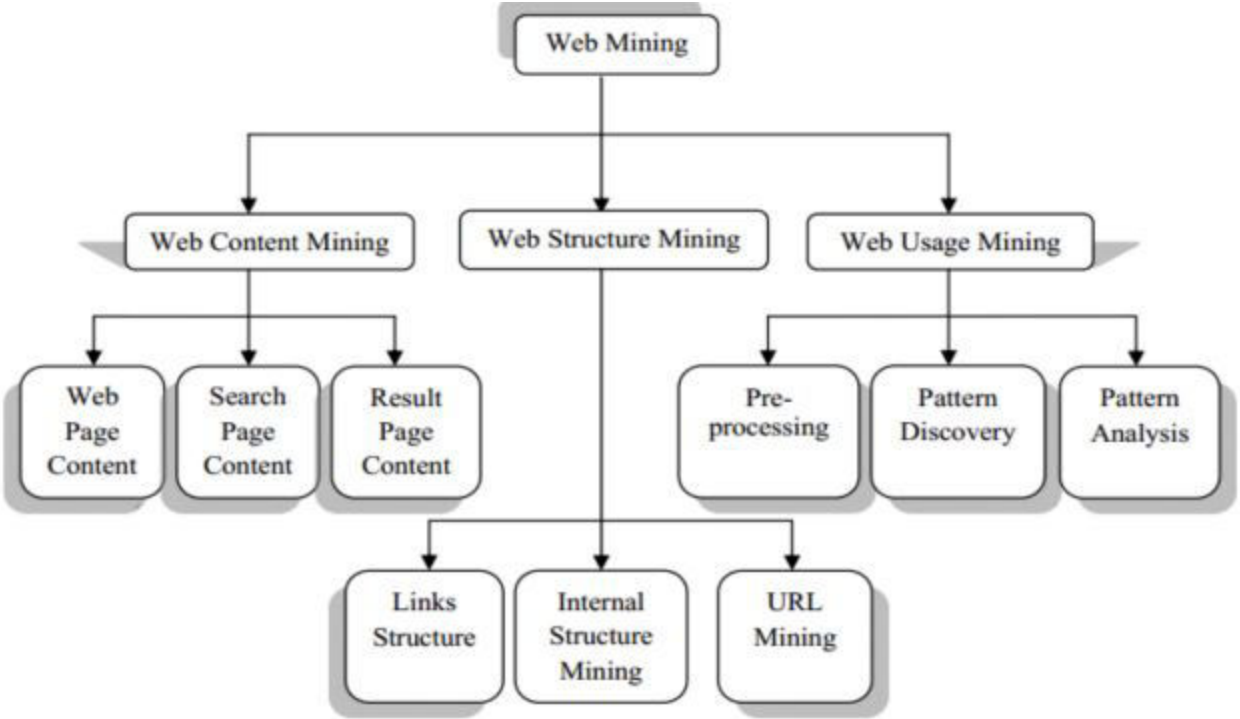
- Web Mining
- Medical
- Resume Filtering

Web Mining

What is Web Mining?

- Web mining is the use of data mining techniques to extract knowledge from web data.
- Web data includes :
    - web documents
    - hyperlinks between documents
    - usage logs of web sites
- The WWW is huge, widely distributed, global information service centre and, therefore, constitutes a rich source for data mining.

Web mining taxonomy

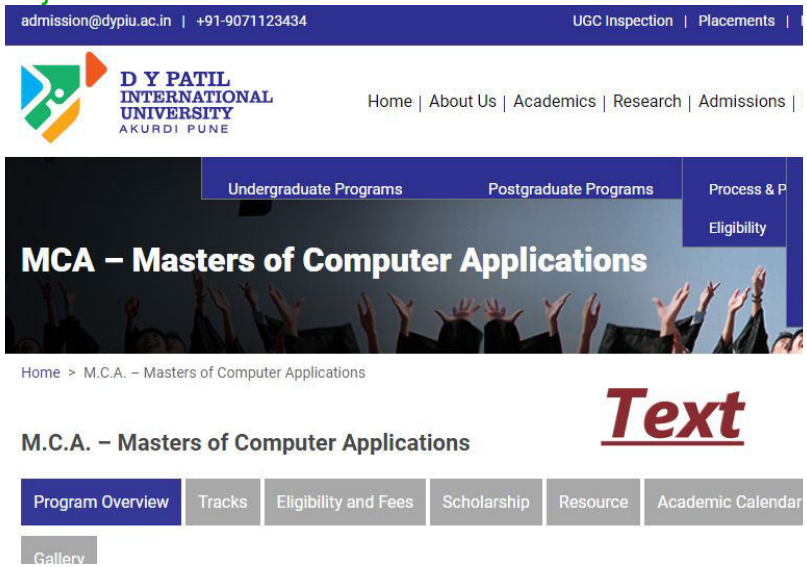


Made By: Priyanka Karale

Web Content Mining

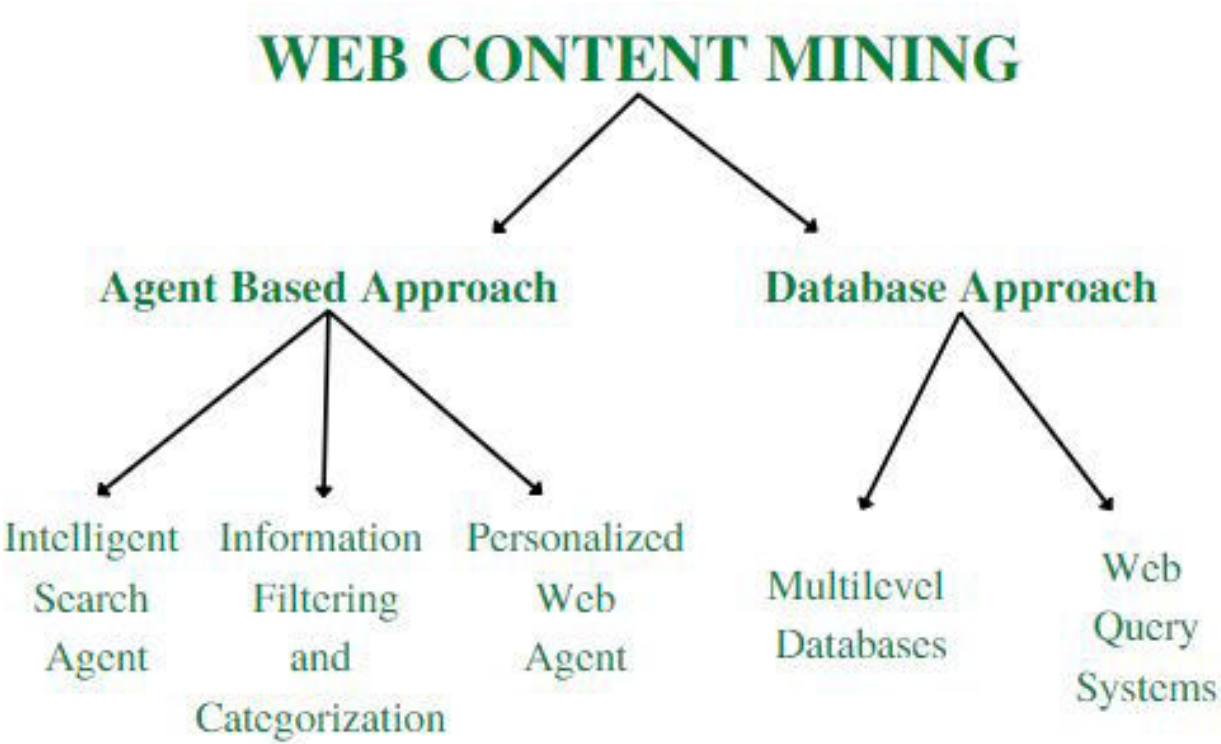
- Mining, extraction and integration of useful data, information and knowledge from Web page content.
- Web content mining is related but different from data mining and text mining.
- Web data are mainly semi-structured and/or unstructured, while data mining deals primarily with structured data

Subject: MCA DMBI 1003



- Text
- Audio
- Video
- Image

Made By: Priyanka Karale


WEB CONTENT MINING
Agent Based Approach
Database Approach
Intelligent Search Agent
Information Filtering and Categorization
Personalized Web Agent
Multilevel Databases
Web Query Systems

Agent-Based Approaches:

- **Intelligent- Search-** This type of search basically refers to a particular goal of the user and will return the results based on the conclusion of that goal.
- **Information-Filtering/ Categorization –** This type of search basically deals with the filtering of data, i.e., removal of unwanted information or redundant information using certain ai based methods. Like, HyPursuit, BO (Bookmark Organizer).
- Growth of **Sophisticated AI systems** replacing users in an automated or unautomated manner. One of these is Deep Learning, wherein the system is trained by feeding it with certain kinds of data.

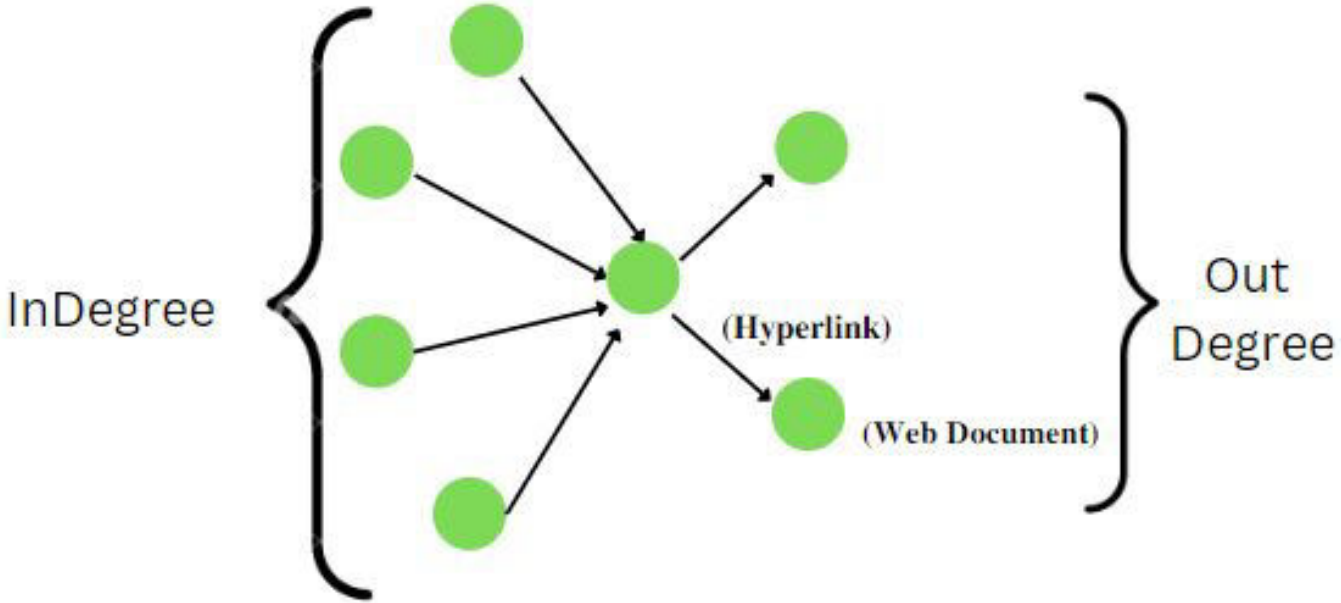
Database Approaches:

Used for transforming unstructured data into a more structured and high-level collection of resources, such as in relational databases, and using standard database querying mechanisms and data mining techniques to access and analyze this information.

- **Multilevel Databases:**
  - Lowest Level – semi-structured information is kept.
  - High Level- generalization from lower levels organized into relations and objects.
- **Web Query Systems:**
  - Web-query systems are developed such as SQL, and Natural Language Processing for extracting data.

Web Structure Mining

- Web structure mining is the process of discovering structure information from the web.
- The structure of typical web graph consists of Web pages as nodes, and hyperlinks as edges connecting between two related pages.
- This type of mining can be performed either at the document level(intra-page) or at the hyperlink level(inter-page).
- The research at the hyperlink level is called Hyperlink analysis.
- Hyperlink structure can be used to retrieve useful information on the web. There are two main approaches:
  - PageRank
  - Hubs and Authorities - HITS


InDegree
(Hyperlink)
(Web Document)
Out
Degree

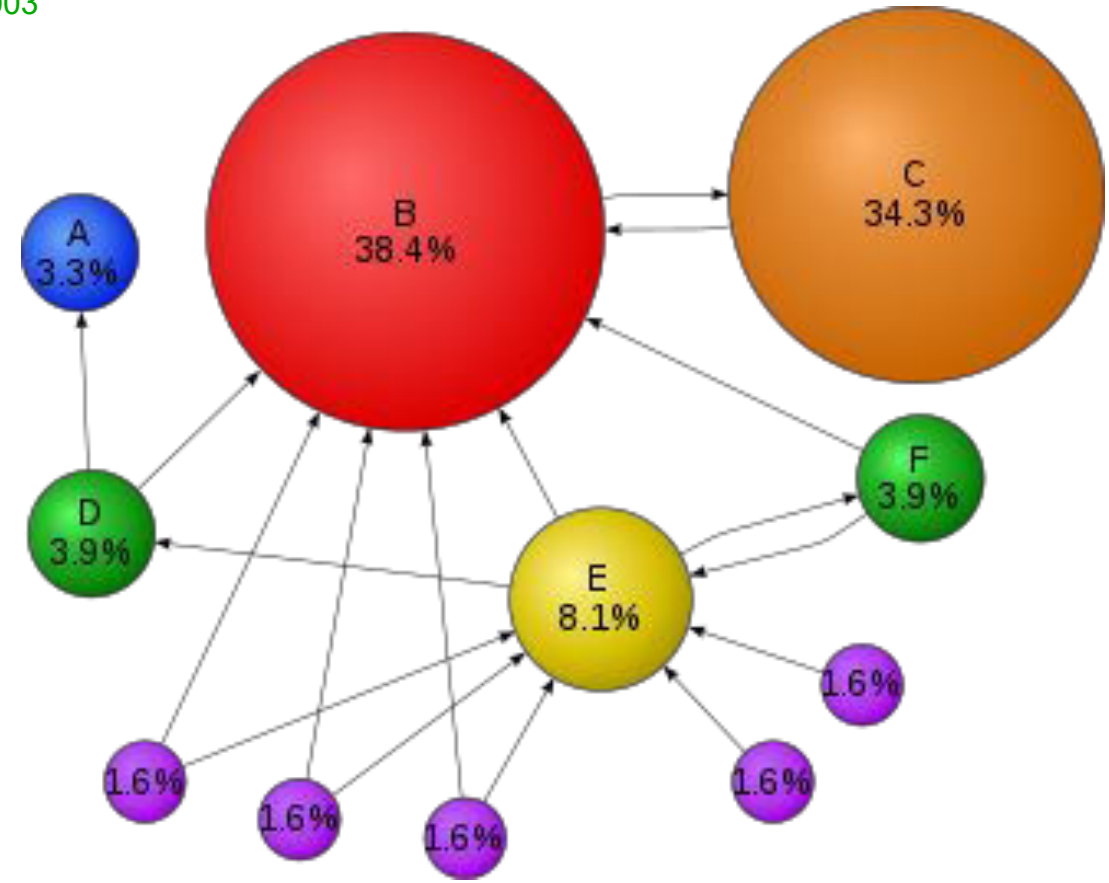
Page Rank

- Used to discover the most important pages on the web.
- Prioritize pages returned from search by looking at web structure.
- Importance of pages is calculated based on the number of pages which point to it (backlinks).
- Weighting is used to provide more importance to backlinks coming from important pages.

$$PR(p) = (1-d) + d\ (PR(1)/N1 + \ldots\ldots + PR(n)/Nn)$$

- PR(i): PageRank for a page i which points to target page p.
- Ni: Number of links coming out of page i.
- d: constant value between 0 and 1 used for normalization.
- (1-d): Bit of probability math magic so that sum of all webpages pageranks should be one.

Subject: MCA DMBI 1003

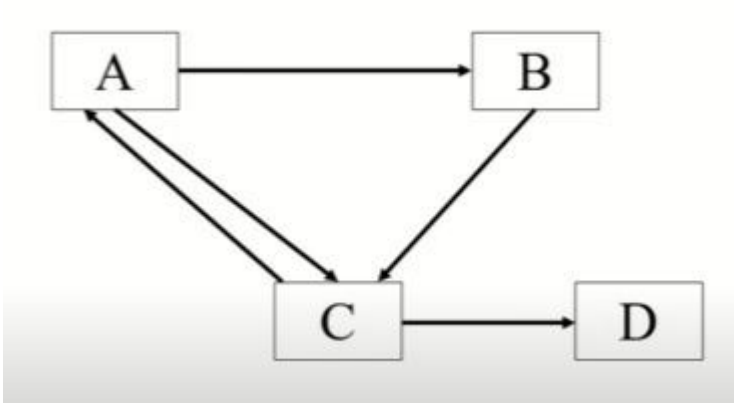


Made By: Priyanka Karale

Example

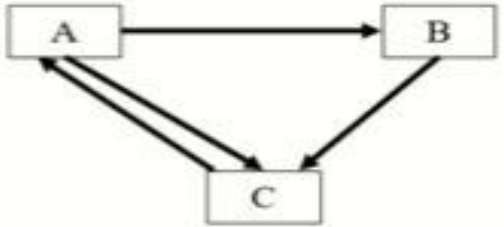
By Default Damping Factor is = 0.85

Initially Page Rank (PR) for all the Web Pages=1



$$PR(A) = (1-d) + d(PR(Ti)/C(Ti) + \ldots + PR(Tn)/C(Tn))$$

$$PR(A) = (1-d) + d\,[\,PR(C) / C(C)\,]$$
$$= (1-0.85) + 0.85\,[\,1/1\,]$$
$$= 0.15 + 0.85[\,1\,]$$
$$= 0.15 + 0.85$$
$$= 1$$

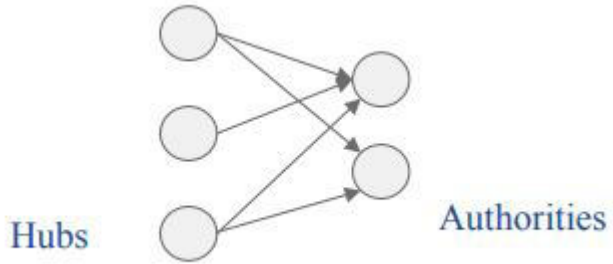


$$PR(B) = (1-d) + d\,[\,PR(A) / C(A)\,]$$
$$= (1-0.85) + 0.85\,[\,(1) / 2\,]$$
$$= 0.15 + 0.85\,[0.5]$$
$$= 0.15 + 0.425$$
$$= 0.575$$

$$PR(C) = (1-d) + d\,[\,PR(A) / C(A) + PR(B) / C(B)\,]$$
$$= (1-0.85) + 0.85\,[\,(1/2) + (0.575 / 1)\,]$$
$$= 0.15 + 0.85[\,0.5 + 0.575\,]$$
$$= 0.15 + 0.85\,[\,1.075\,]$$
$$= 0.15 + 0.91375$$
$$= 1.06375$$

Hubs and Authorities

- Authoritative pages
  - Authors defines an authority as the best source for the request.
  - Highly important pages.
  - Best source for requested information.
- Hub pages
  - Contains links to highly important pages



HITS (Hyperlink Induced Topic Search)

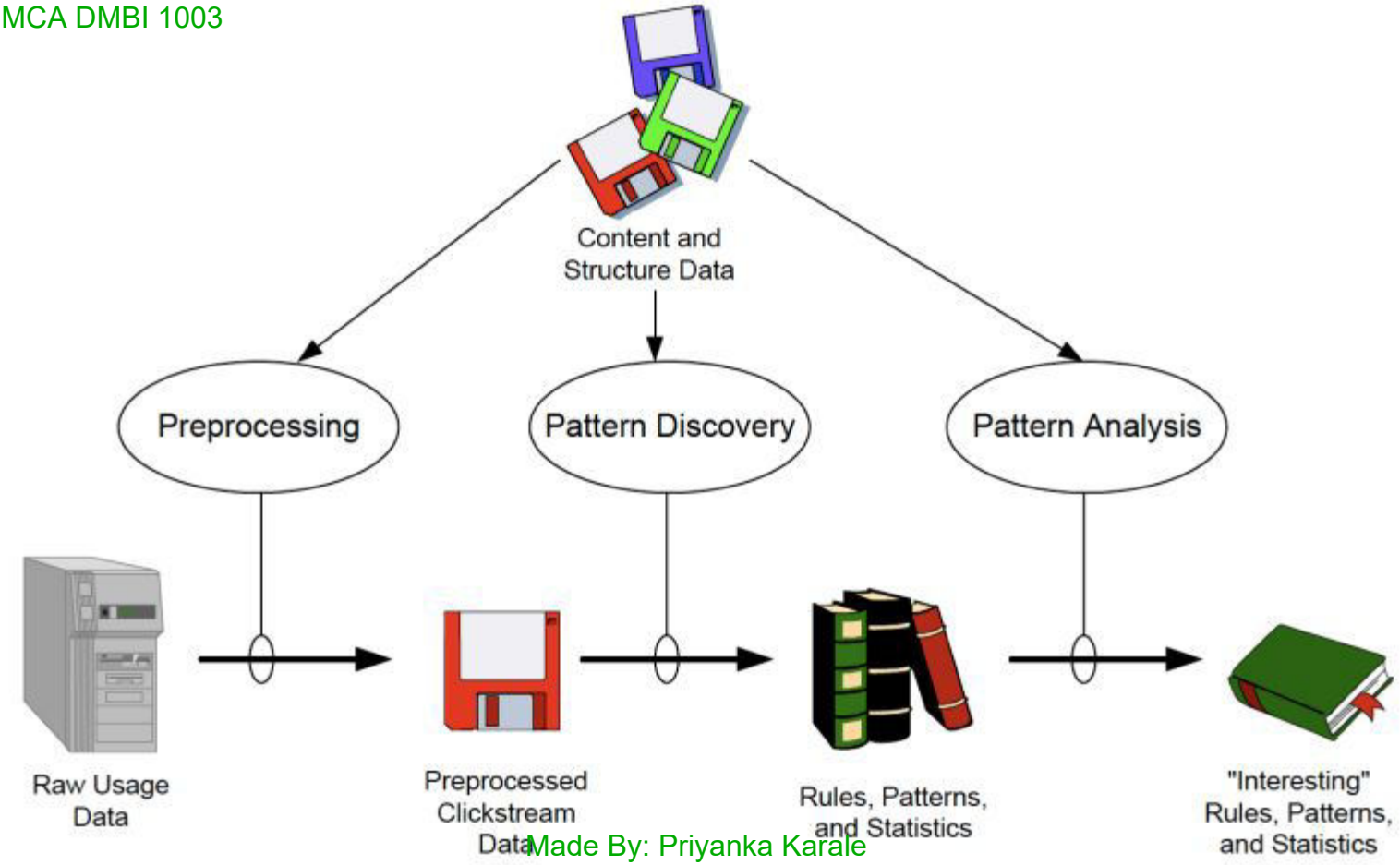
Iterative algorithm for mining the Web graph to identify the topic hubs and authorities.

Web Structure Mining applications

- Information retrieval in social networks.
- To find out the relevance of each web page.
- Measuring the completeness of Web sites.
- Used in search engines to find out the relevant information.

Web Usage Mining

- Web usage mining is used for mining the web log records (access information of web pages) and helps to discover the user access patterns of web pages.
- Web server registers a web log entry for every web page.
- Analysis of similarities in web log records can be useful to identify the potential customers for e-commerce companies.


Content and
Structure Data
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Usage
Data
Preprocessed
Clickstream
Data
Rules, Patterns,
and Statistics
"Interesting"
Rules, Patterns,
and Statistics

Data Visualization

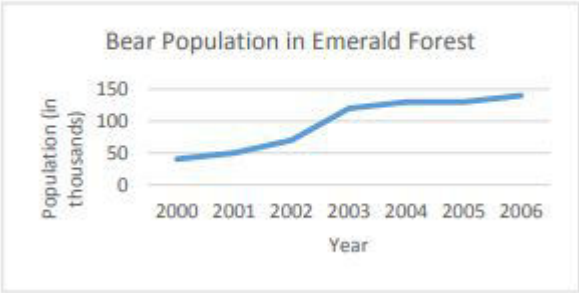
- Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps.
- Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans.
- Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.
- it provides an excellent way for employees or business owners to present data to non-technical audiences without confusion

Tools for visualizing data

- Tableau
- Google Charts
- Dundas BI
- Power BI
- JupyteR
- Infogram
- ChartBlocks
- D3.js
- FusionCharts
- Grafana**et**

Type of Graph	Uses
Line graph	demonstrating change over time or comparing change over time
Bar chart	comparing differences / similarities between groups
Table	providing exact values
Scatter plot	showing correlation (positive, negative, none) between two variables
Pie chart	illustrating large differences in proportions for simple data

Line graph: changes over time



Since the unit for the x-axis is Years, we can easily see the growth of the bear population (the y-axis unit) over time by following the line connecting the data points.

Bar graph: comparing groups

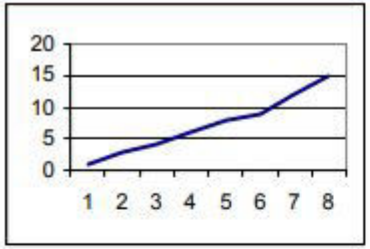
A quick glance at this graph shows a large difference between the groups wearing no chemical protection and the groups wearing either the current gear or the future warrior protective gear. Notice how the gridlines have been removes to eliminate “non data ink” and create a clean look.

Table: presenting exact values

Table 1: Average time to perform task (in seconds)

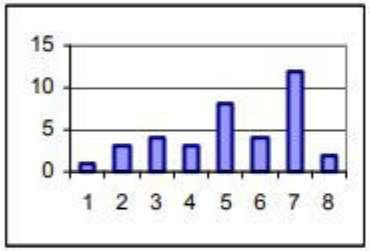
Task	*Firefox 2.0*	*IE 7.0*
Startup speed		
Time to start up from cold	11.6	7.8
Download speeds		
Download benchmark HTML page w/ multiple images	2.0	2.5
Download benchmark Javascript	22.0	36.4

We can easily compare the startup speed between Firefox and IE by glancing at the actual values and knowing that 11.6 is higher than 7.8. It is also useful that we can see not only startup speed in this table, but also download speeds. Again, notice how an absence of gridlines simplifies this graph.



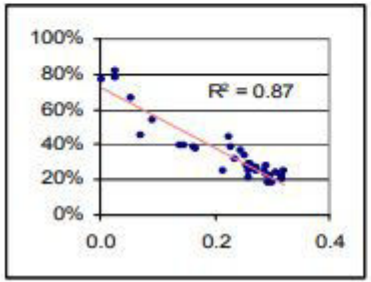
Line Graph

- x-axis requires quantitative variable
- Variables have contiguous values
- Familiar/conventional ordering among ordinals



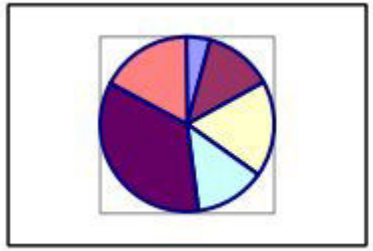
Bar Graph

- Comparison of relative point values



Scatter Plot

- Convey overall impression of relationship between two variables



Pie Chart

- Emphasizing differences in proportion among a few numbers

Fraud detection

- Fraud detection is a process that detects and prevents fraudsters from obtaining money or property through false means.
- It is a set of activities undertaken to detect and block the attempt of fraudsters from obtaining money or property fraudulently.
- Fraud detection is predominant across banking, insurance, medical, government, and public sectors, as well as in law enforcement agencies.

- Fraudulent activities include money laundering, cyberattacks, fraudulent banking claims, forged bank checks, identity theft, and many such illegal practices. As a result, organizations implement modern fraud detection and prevention technologies and risk management strategies to combat growing fraudulent transactions across diverse platforms.

- The adaptive and predictive analytics techniques are used to create a fraud risk score along with real-time monitoring of fraudulent events. This involves continuous monitoring of transactions and crimes in real-time.

Subject: MCA DMBI 1003

Made By: Priyanka Karale