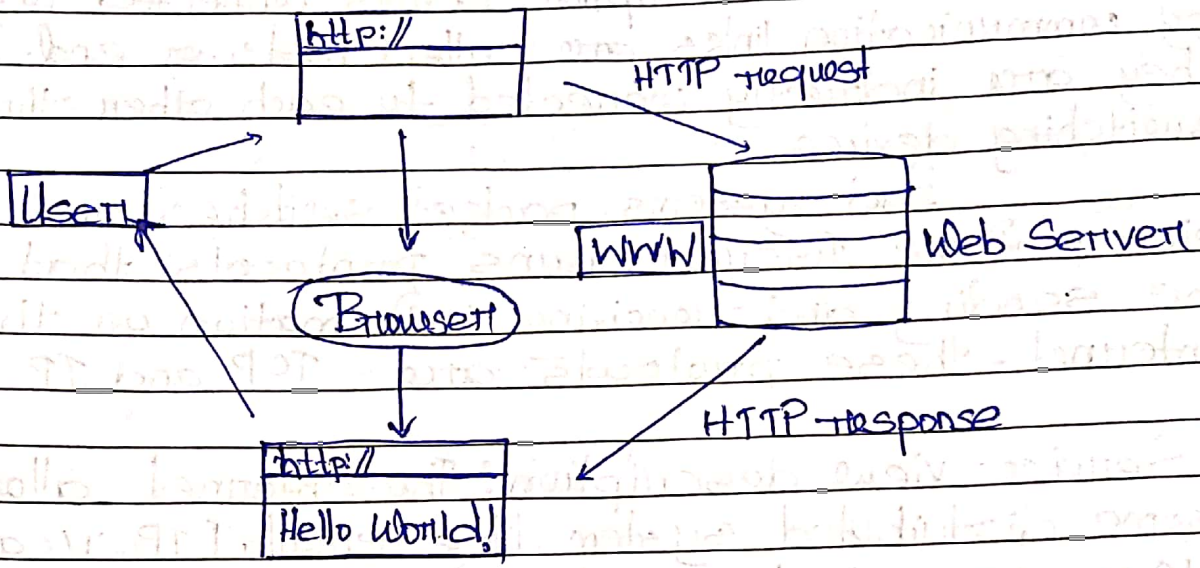
PAGE:

Working Procedure of Web:



Browser generally provides you an interface.

The Web Graph:

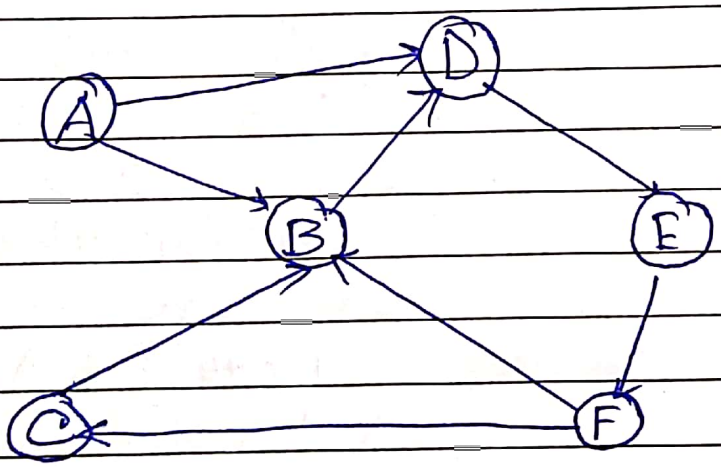
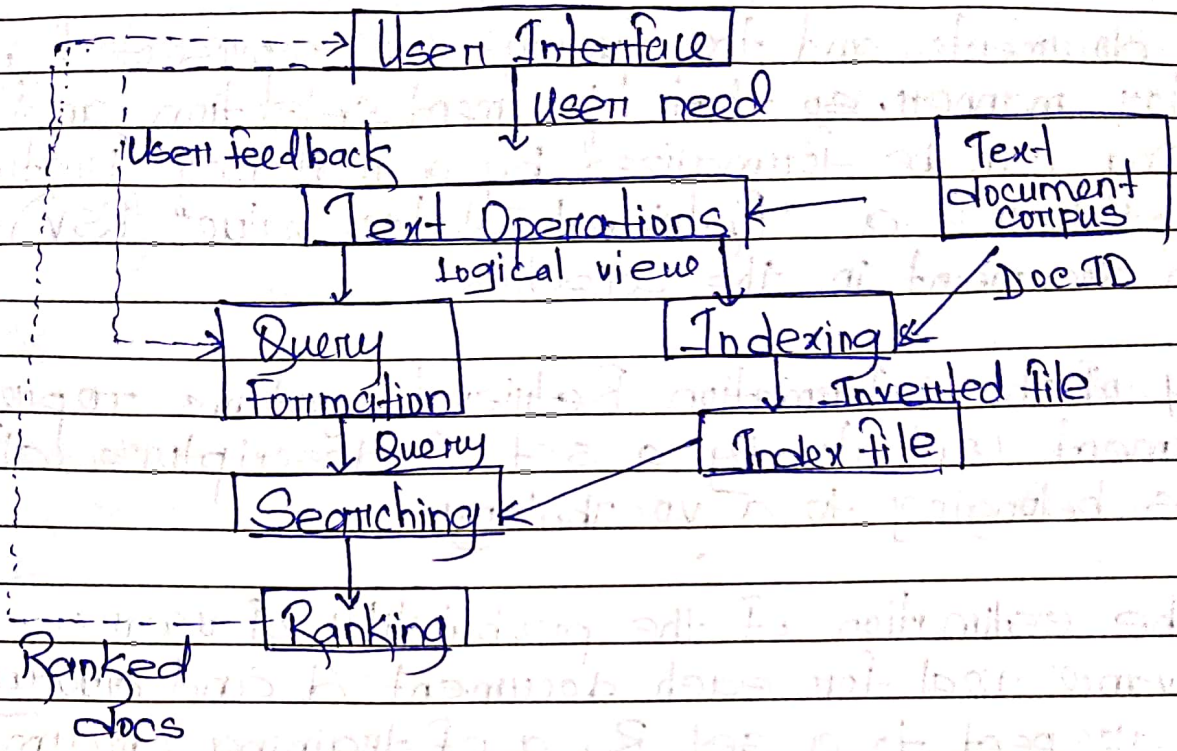


Fig: In this example, we have six pages labeled A-F. Page B has in degree 3 and out-degree 1. This example graph is not strongly connected, there is no path from any of page B, F to A page.

Good Write

DATE: __/__/____
PAGE: _____

Web IR (Information Retrival) Process:



IR model: Selects and ranks the document that is required by the user or user asked for in the form of query.

i) Before any of the retrieval processes are initiated it is necessary to define the text database
   ⇒ It is usually done by the database manager
   ⇒ Includes specifying the documents to be used.

ii) The operations to be performed on the text: it transforms the original documents and generate a logical view of them.

iii) Once the logical view of the documents is defined, the database module builds an index of the text.

Good Write

DATE: __/__/____
PAGE: _____

Index: An index is a critical data structure and it allows fast searching over large volumes of data.

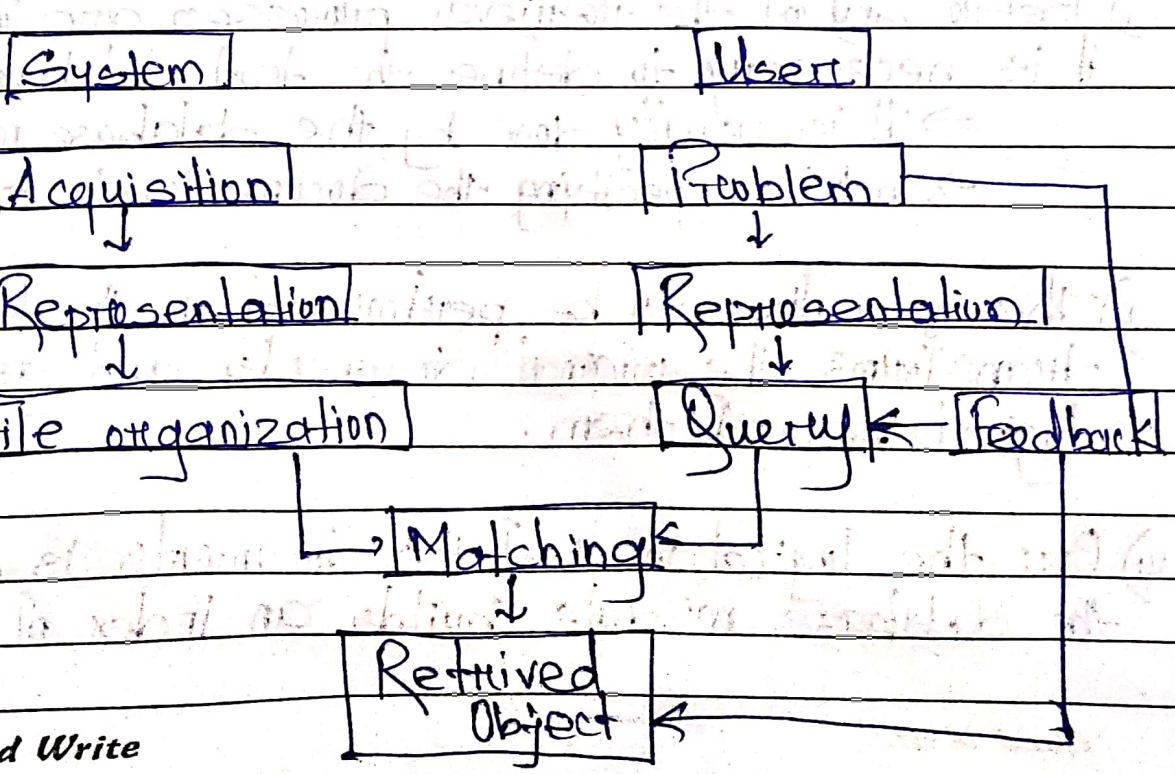
The documents and the queries are represented in a similar manner, so that document selection and ranking can be formalized by a matching function that returns a "Retrieval status value" (RSV) for each document in the collection

Many of the Information Retrieval systems represent document contents by a set of descriptors, called terms, belonging to a vocabulary.

The estimation of the probability of user's relevance Real for each document d and query q with respect to a set R, q of training documents:

$$Prob(rel\ d, q, R_q)$$

Components of IR model:



Good Write

DATE: __/__/____
PAGE: _____

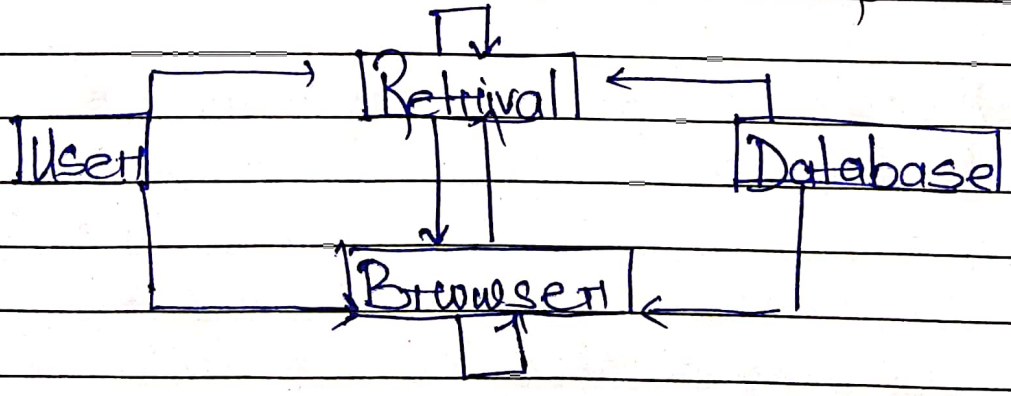
Acquisition: In this step, the selection of documents and other various objects from web resources that consists of text based documentation takes place. The required data is collected by the web crawlers and stored in the database.

Representation: It consists of indexing that contains text-free terms, controlled vocabulary, manual and automatic techniques as well.

File organization: There are two types of file organization
Sequential: It contains documents by document data
Inverted: It contains term by term, list of records under each term.

Query: An IR process starts when an user enters a query into the system. Queries are formal statements of information need.

User Interaction with Information Retrival System:



DATE: __/__/____
PAGE: _____

Past, Present, and Future of Information Retrival.

i) Early Development: As there was an increase in the need for a lot of information, it became necessary to build data structures to get faster access. The index is the data structure for faster retrieval of information.

ii) Information Retrieval in Libraries: Libraries were the first to adopt IR systems for information retrieval. In first generation, it consisted, automation of previous technologies, and the search was based on author name and title. In second generation, it included searching by subject heading. In third generation, it consisted of gene graphical interfaces, electronic forms, hypertext features etc.

iii) Web and Digital Libraries: It is cheaper than various sources of information, it provides greater access to networks due to digital communication.

DATE: __/__/____
PAGE: _____

Information Retrieval	Data Retrieval
(i) The software that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual info.	(i) Data retrieval deals with obtaining data from database management system such as DDBMS.
(ii) Retrieves information about a subject.	(ii) Determines the keywords in the user query and retrieves the data.
(iii) Small errors are likely to go unnoticed.	(iii) A single error object means total failure.
(iv) It is a probabilistic model.	(iv) It is a deterministic model.
(v) Results are ordered by relevance.	(v) Results are unordered by relevance.
(vi) Results obtained are approximate matches.	(vi) The results obtained are exact matches.

Good Write

DATE: __/__/____
PAGE: _____

Search Engine: A search engine is a software program that provides information according to the user query. It finds various websites or web pages that are available on the Internet and gives related results according to the search.

How do Search Engine works:

1. Crawling: Search engines have a number of computers programs that are responsible for finding information that is publicly available on internet. These program scan the web and create a list of all available websites. Then they visit each website and by reading HTML code, they try to understand the structure, type, meaning of the page's content.

@ Crawling is important because our first concern when optimizing the website for search engines is to make sure they access it correctly. If it cannot find the content, we won't get any ranking or search engine traffic.

2. Indexing: Information identified by the crawlers needs to be organized, sorted and stored so that it can be processed later by the ranking algorithm. Search engine don't store all the information in the index, but they keep things like title and description of the page.

@ Indexing is important because if the website is not in the index, it will not appear in any searches. So, if we index the pages, it has more

changes of appearing in the search results for a related query.

3. Ranking: Ranking is the position in which the websites are listed in the search engine.

Procedure of ranking:

Step 01: Analyze user query ⇒ This step is to understand what kind of information the user is looking for. To do that, analyze the user's query by breaking it down into keywords.

Step 02: Finding matching pages ⇒ This step is to look into their index and find the best matching page.

Step 03: Present the result to the user ⇒ A typical search result page includes ten organic result in most cases

Components of Search Engine: There are three components of search engine. They are:

i) Web Crawler: A search engine uses multiple web crawlers to crawl through www. and gather information. This software is also known as bot or spider.

ii) Data base: The information which is gathered by the web crawler is stored on the database

DATE: __/__/____
PAGE: _____

iii) Search Interface: Search interface is just an interface to database which is employed by an user to search through database.

Performance of Search Engine:
Mainly 2 requirements ⇒
(i) effectiveness (quality of result)
(ii) efficiency (response time & throughput)

Basic building block of search engine:

i) Indexing: Indexing performs mainly 3 activities.

⇒ (i) Text acquisition: Text acquisition basically defines and, identifies and store documents into a database for indexing.
⇒ (ii) Text transformation: It transforms document into indexed terms
⇒ (iii) Index creation: It collects the features like position and count of words, calculates weight on the indexed terms and as the format of inverted files is fast for query it converts document term information to term document information.

ii) Querying: Querying performs 4 tasks

⇒ User Interaction: Provides a query input which gives an interface and parsers for query language.

Good Write

DATE: __/__/____
PAGE: _____

⇒ (ii) Ranking: It firsts calculates the score of the document by using ranking algorithm.
⇒ (iii) Score: $q_i * d_i$; where q_i and d_i are terms of weights for term i query & document
⇒ (iv) Evalution: In this step, it logs user queries and interaction for improving search engines efficiency and effectiveness.

Search Engine Optimization:

SEO is the process of improving the ranking (visibility) of a website in search engines. The higher (or more frequently) a website is displayed in a search engines list, the more visitors it is expected to receive

SEO considers;
⇒ How engines work.
⇒ What people search for
⇒ which search terms are typed.

Optimizing a website involve:
(i) editing the content to increase it's relevance to specific keywords.
(ii) promoting a site to increase the number of links, is another SEO tactic

Effective search engine optimization may require changes to the HTML source code of a site and to the site

DATE: __/__/____
PAGE: _____

Importance of SEO:

i) To help gain more visitors: majority of users click on only the top 4-5 web pages appearing in search results, so it is very important for a website to appear in the top results of a search engine.

ii) Important for social promotion of a website: If a website appears in top results of a search engine such as Google, Bing etc. then it gains instant popularity and to some extent trust of a user.

iii) It plays an important role in having improved the business of a commercial site: If two websites are selling the same product, then the site having a better position in the search result of a search engine has chances of getting more users as compared to the other.

iv) Improving user experience: SEO doesn't focus only on improving search results but also improving the user experience and usability of a website so that a website is more appealing to a user.

DATE: __/__/____
PAGE: _____

Basic principles in working of a search engine:

i) Crawling: Process of fetching all the web pages linked to a website. This task is performed by a software, called a crawler or a spider.

ii) Indexing: Process of creating index for all the fetched web pages and keeping them into a giant database from where it can later be retrieved.

iii) Processing: When a search request comes, the search engine processes it, i.e. it compares the search string in the search request with the indexed pages in the database.

iv) Calculating relevancy: It is likely that more that one page contains the search string, so the search engine starts calculating the relevancy of the each of the pages in it's index to the search string.

v) Retrieving results: The last step in search engine activities is retrieving the best matched results.

Web Analytics: Web analytics is the collection, reporting, and analysis of data generated by users' visiting and interacting with a website.

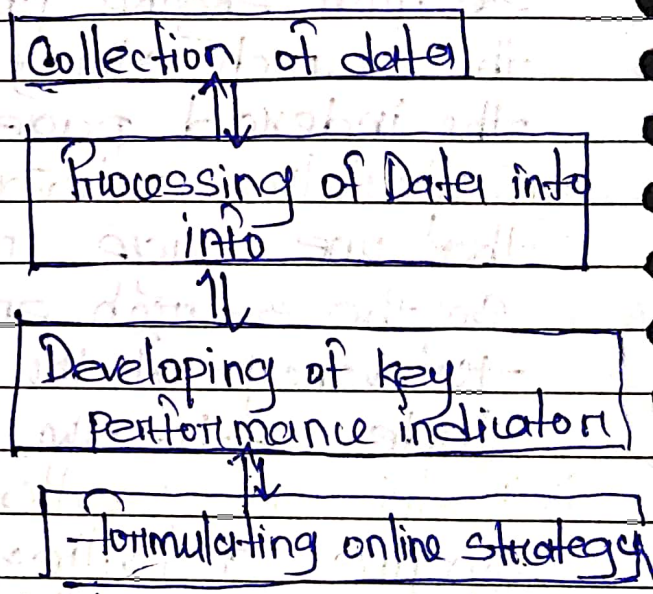
The purpose of web analytics is to measure user behaviour, optimize the website's user experience and flow, and gain insights that help meet business objectives like increasing conversion.

DATE: __/__/____
PAGE: _____

⇒ provides understanding of user behavior over the web page or website.
⇒ tool for marketing research as well.
⇒ Keeps track of -
  i) no. of users visiting the website.
  ii) Duration of their stay.
  iii) the pages they visit.

Hence, used to improve effectiveness and performance of the website.

⇒ Web Analytics example:
  i) Pageviews
  ii) Unique pageviews
  iii) Sessions
  iv) New visitors
  v) Returning Visitors
  vi) Traffic sources
  vii) Bounce Rate.



Web Mining: The process of Data Mining technique to automatically discover and extract information from web documents and services.

The main purpose of web mining is discovering useful information from the www and its usage patterns.

DATE: __/__/____
PAGE: _____

Applications of web mining:

i) Web mining helps to improve the power of web search engine by classifying the web documents and identifying the web pages.
ii) It is used for Web Searching.
iii) Web mining is used to predict user behaviour.
iv) Web mining is very useful of a particular website and e-service.

Web mining can be broadly divided into 3 different types of techniques of mining:

i) Web content mining: Web content mining is the application of extracting useful information from the content of the web documents. Web content consists of several types of data - text, image, audio, video etc.

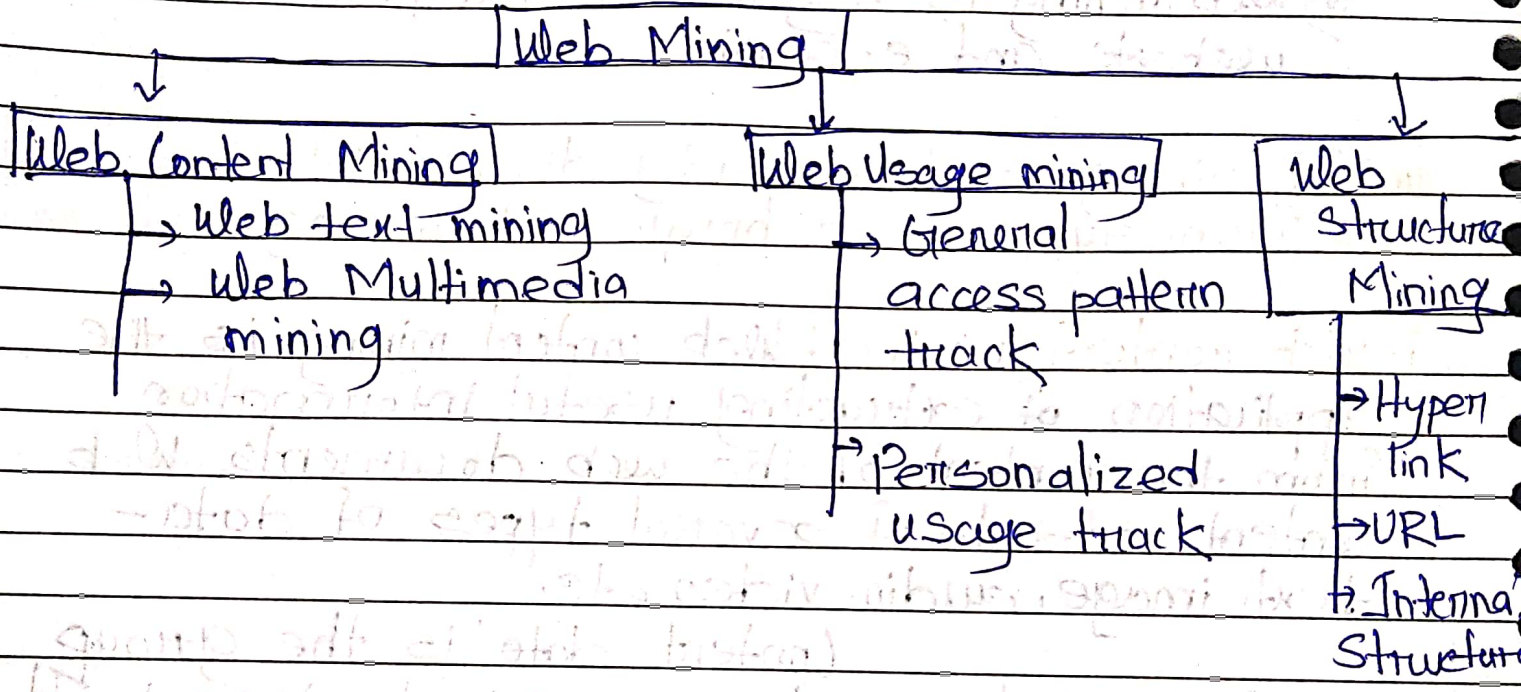
Content data is the group of facts that a web page is designed. It can provide the effective and interesting patterns about user needs. Text documents are ~~data~~ related to text mining, machine learning and NLP. This mining is known as text mining.

ii) Web structure mining: Web structure mining is the application of discovering structure information from the web.

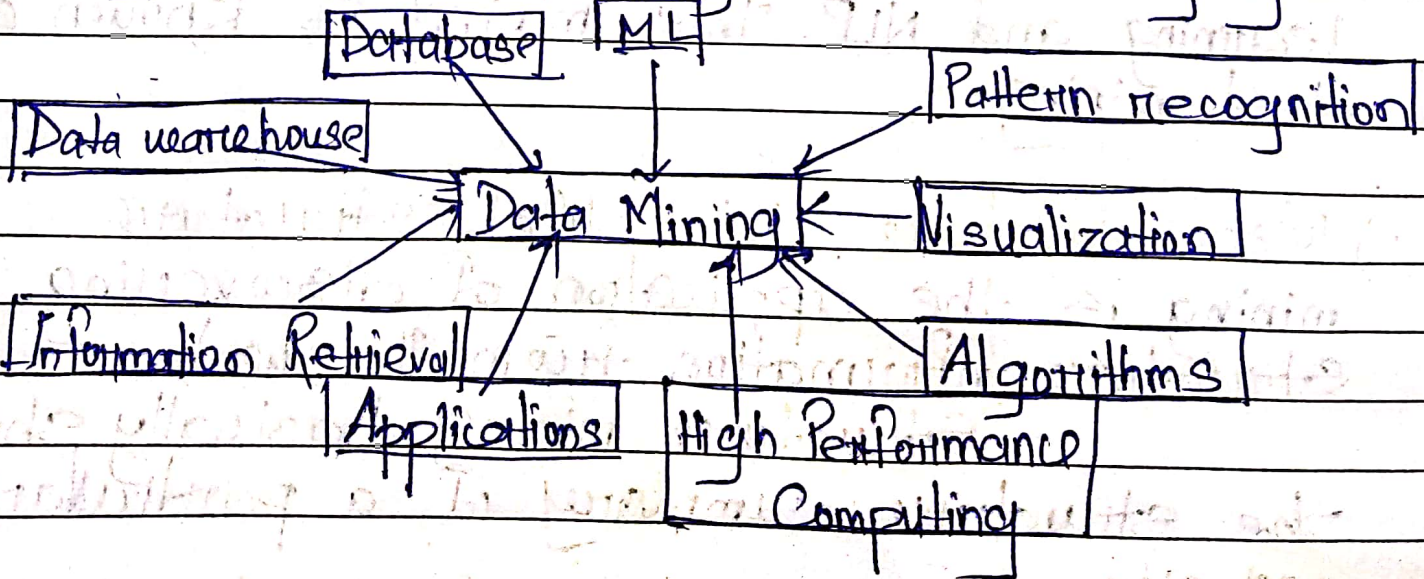
Structure mining basically shows the structure summary of a particular website.

iii) Web Usage Mining: Web usage mining is the application of identifying or discovering interesting usage patterns from large datasets. And these patterns enable to understand the user behavior.

Web mining taxonomy:



Data Mining: Data mining can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology and data dredging.



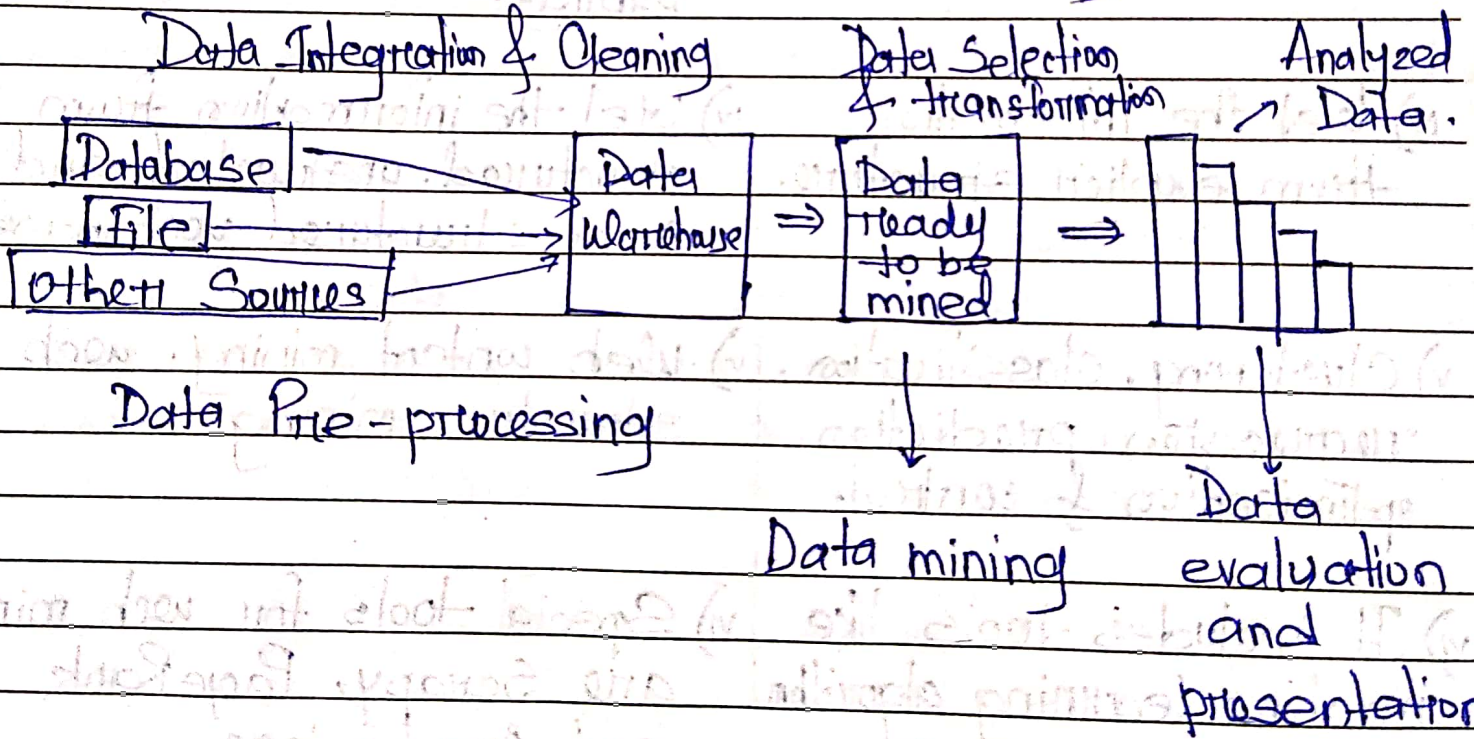
DATE: __/__/____
PAGE: _____

Data mining can be applied to any type of data:

i) Data Warehouse
ii) Transactional database
iii) Relational database.
iv) Multimedia database.
v) Spatial database.
vi) Time series database.

Procedure of data mining:

i) Data preprocessing = data cleaning, integration, selection and transformation.
ii) Data extraction/mining = occurance of exact data mining
iii) Data evaluation and presentation: Analyzing and presenting results.



Applications of Data Mining:

i) Financial Analysis
ii) Biological Analysis
iii) Scientific analysis
iv) Intrusion detection
v) Fraud detection
vi) Research Analysis

Good Write

DATE: __/__/____
PAGE: _____

Difference of web Mining & data mining:

Data Mining	Web Mining
i) The process that attempts to discover pattern and hidden knowledge in large data sets in any system.	i) The process of data mining techniques to automatically discover and extract info from web documents.
ii) Data Mining is very useful for web page analysis.	ii) Web mining is very useful for a particular website & e-service.
iii) It access data privately.	iii) Web mining access data publicly.
iv) Get the information from explicit structure.	iv) Get the information from structured, unstructed and semi-structured web pages.
v) Clustering, classification, regression, prediction, optimization & control.	v) Web content mining, web structure mining.
vi) It includes tools like machine learning algorithm.	vi) Special tools for web mining are Scrapy, PageRank and Apache logs.

DATE: __/__/____
PAGE: _____

Social Web Mining: Social media mining includes social media platforms, social networks analysis, and data mining to provide a convenient and consistent platform for learners, professionals, scientists, and project managers to understand the fundamentals and potentials of social media mining.

It suggests various problems arising from social media data and presents fundamentals and potential concepts, emerging issues, and effective algorithms for data mining and networks analysis. It includes multiple degrees of difficulity that enhance knowledge and help in applying ideas, principles and techniques in distinct social media mining situations.

Data mining techniques can assit effectively in dealing with the three primary challenges with social media:

i) Social media data sets are large.
ii) Social media site's data sets can be noisy.
iii) Data from online-social media platforms are dynamic regular modifications and updates over a short period are not common but also a significant aspect to consider in dealing with social media

DATE: __/__/____
PAGE: _____

Text Mining:

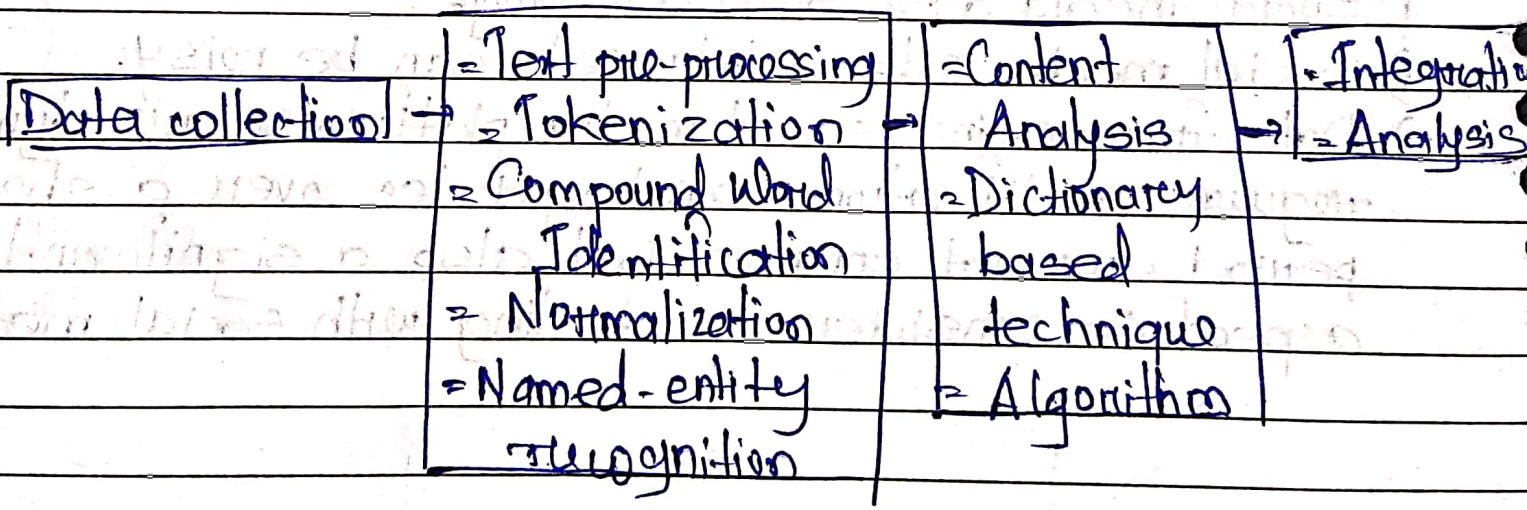
Text mining is the procedure of synthesizing information, by analyzing relations, patterns, and rules among textual data.

Text mining is a process of extracting useful information and non-trivial patterns from a large volume of text databases.

⇒ application of extracting useful information from the content of the web documents.
⇒ contains several types of data-text, image, audio, video etc.
⇒ provides interesting patterns about users needs.

creation → storage → management → Delievery

The conventional process of text mining as follows:



Procedures of analyzing text mining:

i) Text summarization
ii) Text Categorization
iii) Text Clustering

DATE: __/__/____
PAGE: _____

Text Mining techniques:

i) Information Extraction.
ii) Information Retrival
iii) Natural language processing
iv) Clustering
v) Text summarization.

Opinion Mining: An approach to natural ~~learning~~ language processing (NLP) that defines the emotion tone behind a body of text.

This is a popular way for organizations to determine and categorize opinions about a product, service, or idea. It involves the use of data mining, machine learning (ML) and artificial intelligence (AI) to mine text for sentiment and subjective information.

Types of sentiment analysis:

i) Fine-grained sentiment analysis provides a more precise level of polarity by breaking down into further categories.
ii) Emotion detection identifies specific emotions rather than positivity and negativity.
iii) Intent-based analysis recognizes actions behind a text in addition to opinion.
iv) Aspect based analysis gathers the specific component being positively or negatively mentioned.

DATE: ___/___/____
PAGE: _____

Applications of sentiment analysis:

i) Identifying brand awareness, reputation and popularity at a specific moment or over time.
ii) Tracking consumer reception of new products.
iii) Evaluating the success of a marketing campaign.
iv) Pinpointing the target audience or demographics.
v) Conducting market research.

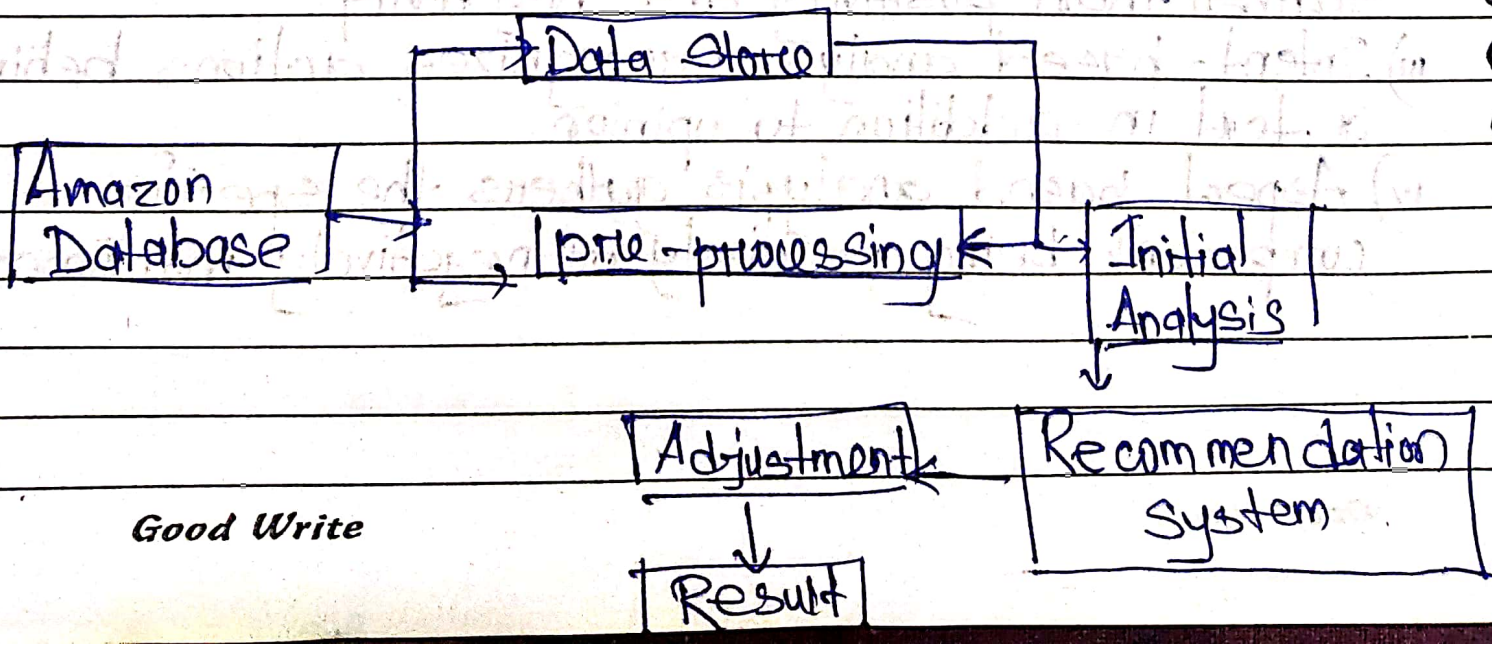
Recommendation system:

Mainly deals with the likes and dislikes of the users, by learning from the large set of repositories generated day by day.

Types:

i) User based recommendation: Here, we calculate persons' similarity measure, which in needed to determine the closely related users, i.e. whose likes and dislikes follows the same pattern.

ii) Item based recommendation: Initial aim is to obtain the mean adjusted matrix.



Good Write