DATE: __/__/____
PAGE: _____

Unit-5

Syllabus
- [x] Web IR system
- [x] Search Engine
- [x] Web Crawling
- [x] Search Engine Optimization
- [x] Web Analytics
- [x] Web mining taxonomy
- [x] Web mining framework
- [x] Social web mining
- [x] Text mining
- [x] opinion mining
- [x] Recommendation system.

---

* Web IR system

Web can be considered as a large-scale document collection for which classical text retrival techniques can be applied.

Web IR examines the combination of evidence from both the textual content of documents and the structure of the web as well as search behavior of users and issues related to the evaluation of retrieval effectiveness in the web setting.

IR is dominant form of info access.

DATE : __/__/____
PAGE : _____

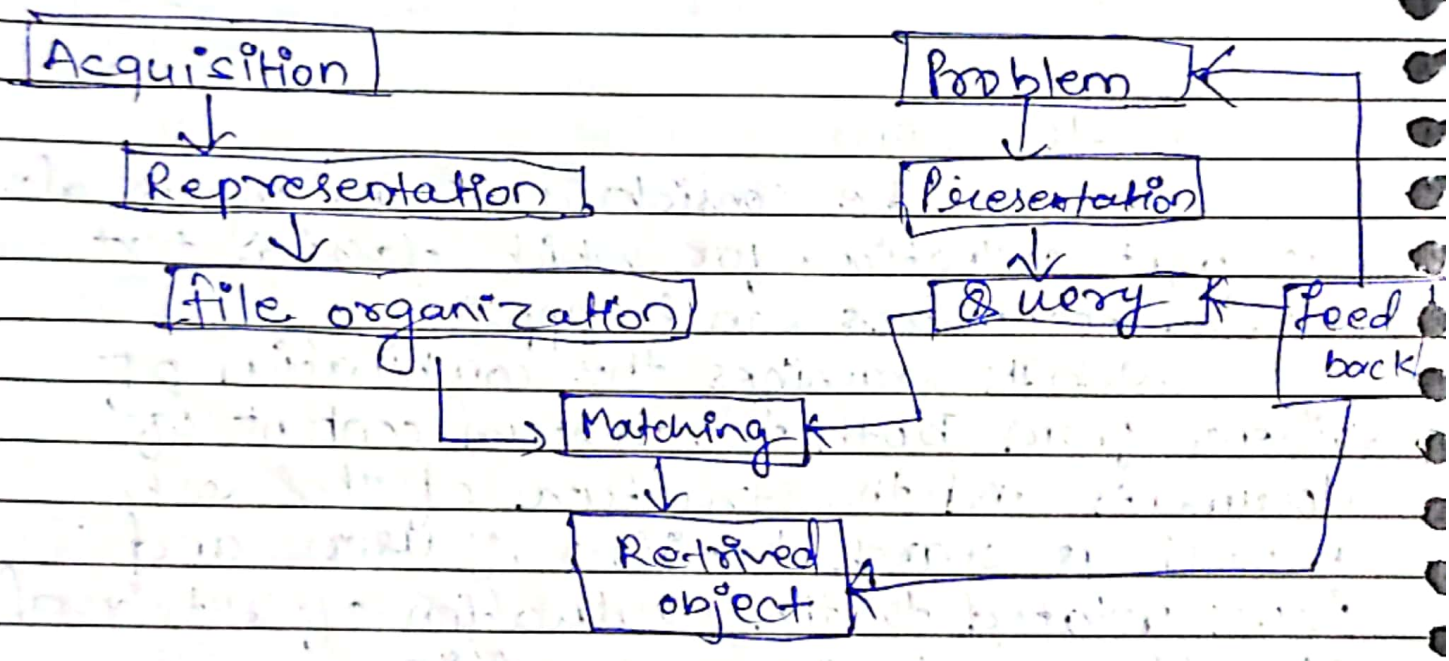
software program that deals with the organization, storage, retrival and evaluation of information from document repositories, particularly textual information.

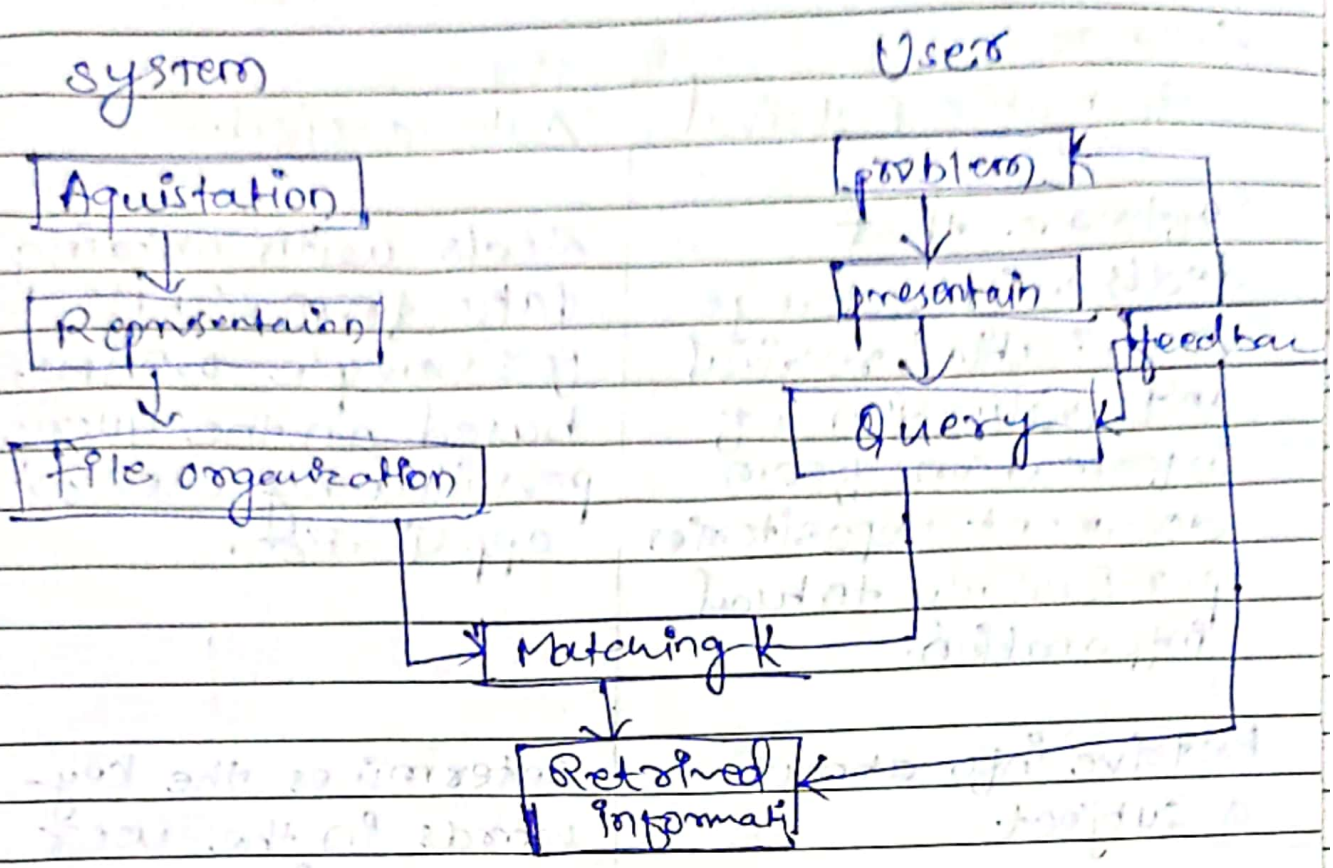
IR Model

selects & rank the document that is required by the user or user asked for in the form of query.

Retrieval status value (RSV)

components of IR model





- Acquisition
  Section of document or other object takes place.

- Representation:
  contains of indexing that contains free-text terms, controlled vocabulary, manual & atomatic techniques as used.

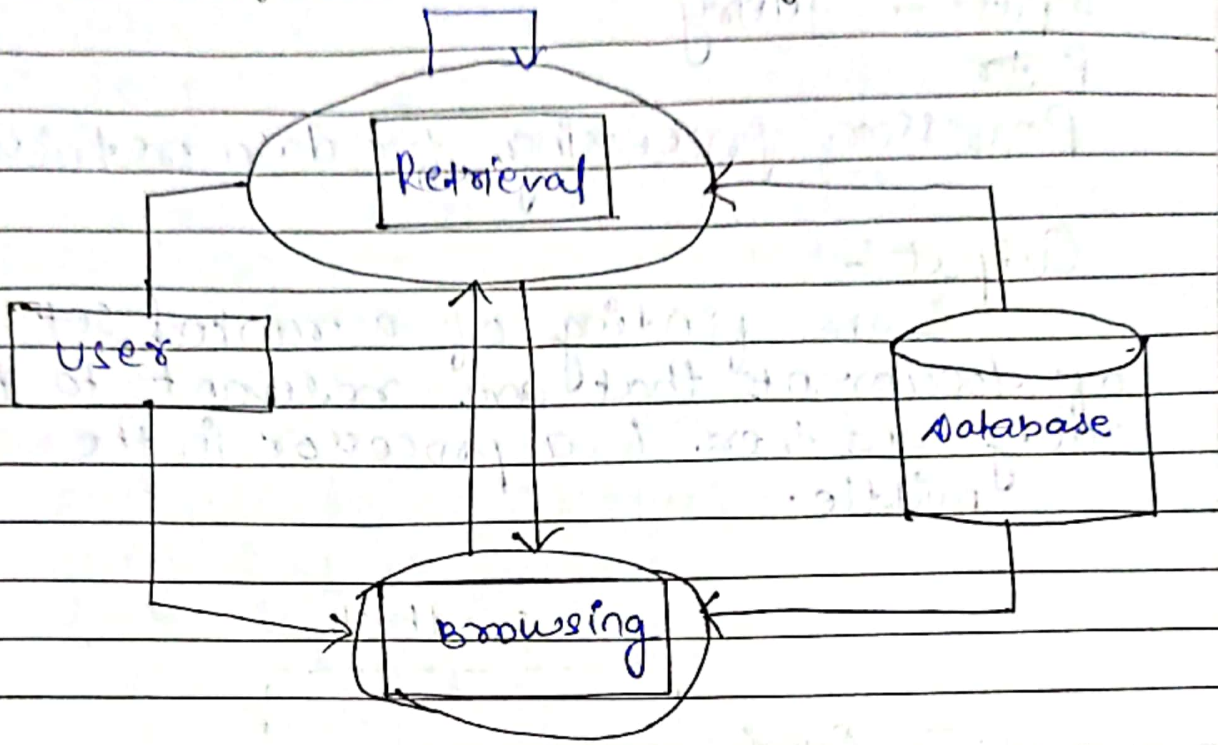
- File Organization:
  two types of file organization
  - Sequential
  - Inverted.

- Query :-
  An IR process starts when a user enters a query into the system.

Information Retrival	Data retrival.
Software that deals with, storage organication, retrival and evaluation of information from document repositories particularly textual information.	deals with obtaing data from DBMS for example DBMS, based on the query provided by user or application.
Retrive info about a subject	determines the key words in the user query and retrive the data.
Small errors are likely to go.	A single error object means total failure.
Doesnot provide a solution to the user of the database system.	Provide
• Results obtained are approximate matches	The results obtained are exact matches
• Result are ordered by relevance.	Results are unordered by relevance.
• probabilistic model	Deterministic model.

DATE: __/__/____
PAGE: _____

User Interaction with Information Retrieval system.

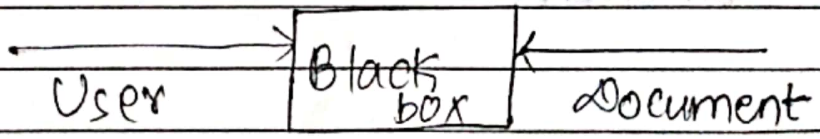


Past present & future of Information Retrieval.

- Early Development
- Information Retrieval in libraries.
- The web & digital libraries.

The logical view of the document by is provided by representative keywords or index terms which are frequently used historically to represent documents in collection.

IR as a bridge.



DATE: __/__/____
PAGE: _____

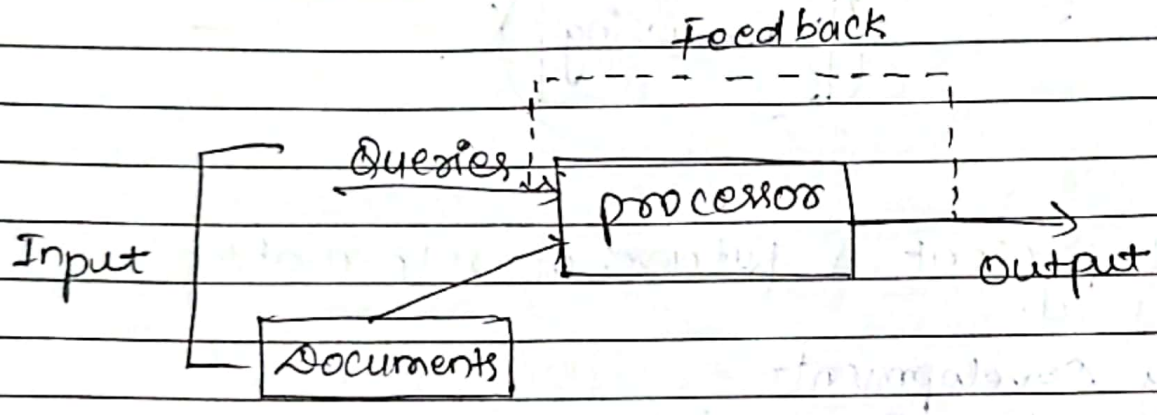
IR System Architechture.

Input :- query.

Processor. processing for data retrieval.

Output :-

is the finding of a ranked set of documents that are relevant to the query and there is a processor in the middle.



Search Engine :-

Tool designed to search for information, when this occurs on the world wide web it is then called a web search engine.

Search Results are usually presented in a list and are commonly called hits.

Common search engine includes google, Yahoo, AOL Search, Ask.com, Bing & Look Smart.

DATE: __/__/____
PAGE: _____

Aspect of IR

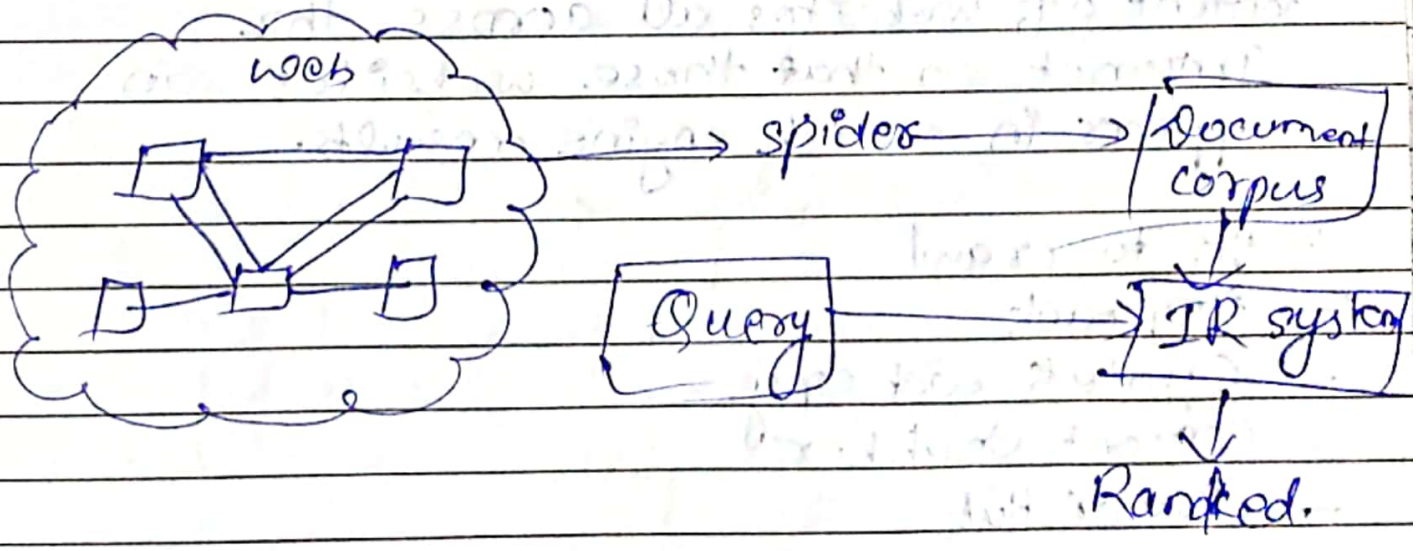
i) processing and presentation of the document collection.

ii) processing queries.

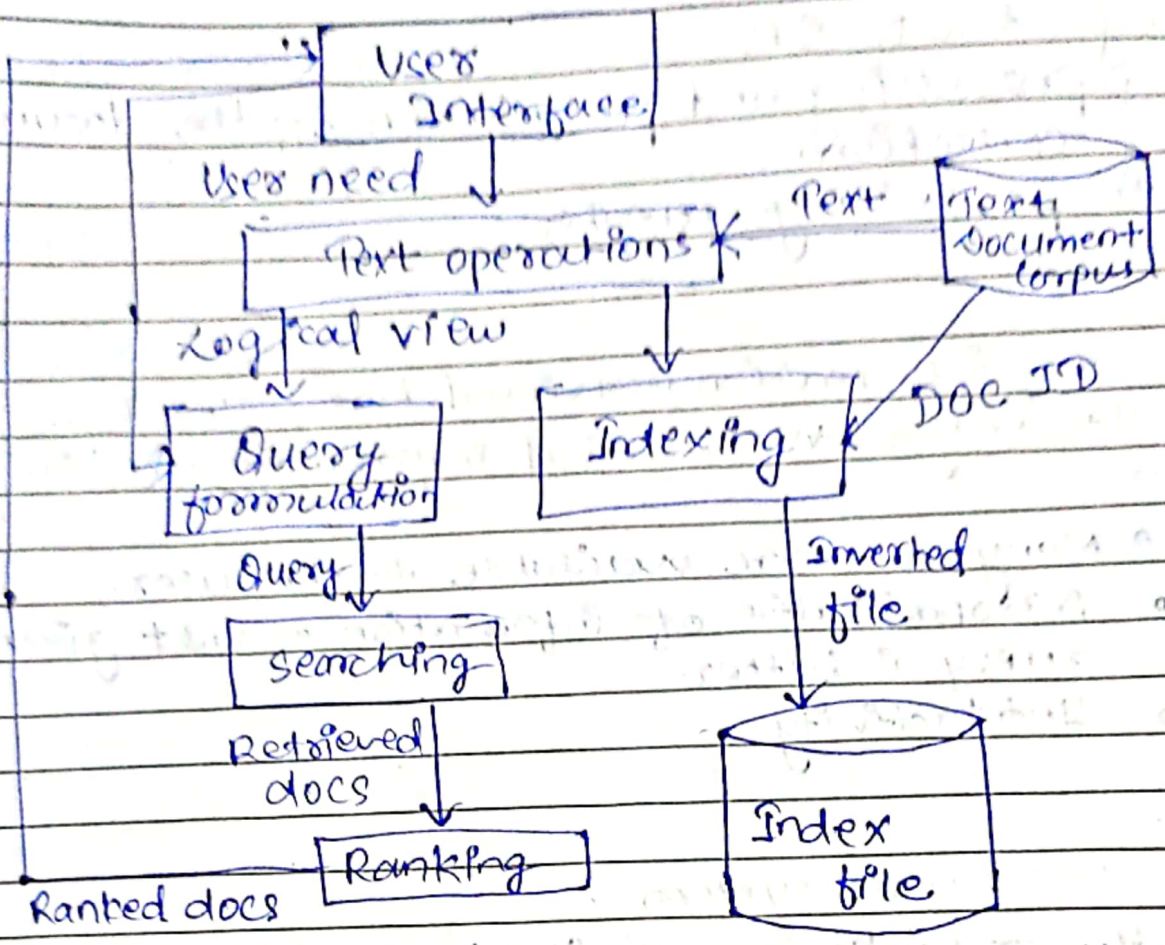
Web information retrieval has following advantages over classical information retrieval.

- Many tools are available to the user.
- personalization of information result given a query is better
- Interactivity.

- collection or system
- Hyper links are available to link one document to the other.
- statistic is easy to gather even in large sample sizes.
- Interactivity.



Good Write



Web crawling :-

- A web crawler or spider is a type of bot that is typically operated by search engine like google and Bing. Their purpose is to index the content of websites all across the Internet so that those websites can appear in search engine results.

ways to crawl

- HTTrack
- Cyotek web copy
- Content Grabber
- Parser Hub
- Outwit Hub

DATE: __/__/____
PAGE: _____

Search engine optimization:-

→ process of improving the ranking or visibility of website in search engines.

→ higher the appearance of website in search engine list, the more visitors it is expected to receive.

→ considers how search engine work, what people search for, and which search term are typed.

promoting a site to increase the number of links, is another SEO tactics.

Importance of SEO:-

- To help gain more visitors
- Important for social promotion of website.
- It play important role in improving the business of a commercial site.
- Improving users experience.

Basic principles in the working of a SE

- Crawling :- search web pages linked to the website
- Indexing
- Processing
- Calculating Relevancy
- Retrieving Result.

DATE: __/__/____
PAGE: _____

Customer Analytics

act of studying customer behavior or intrest and intreprating them to take important business decision.

Web analytics

→ defined as the measurement, collection as well as analysis and reporting of the data for optimization of the web page usage.

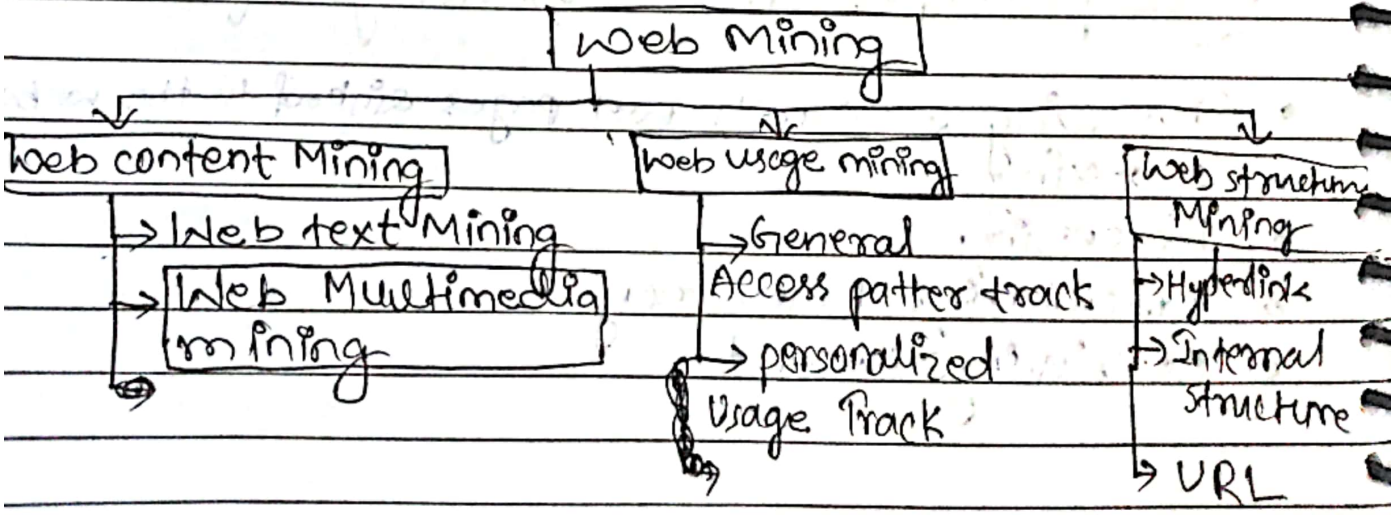
→ provides understanding of user behaviours over the web page or website.

→ tool for marketing research as well.

→ keeps track of :-
- number of users visiting the website.
- Duration of their stay.
- the pages they visit.

Hence used to improve effectiveness and performance of the website.

* Web mining taxonomy :-



Good Write

DATE: __/__/____
PAGE: _____

Web mining is the process of Data mining techniques do automatically discover and extract information from web documents & services.

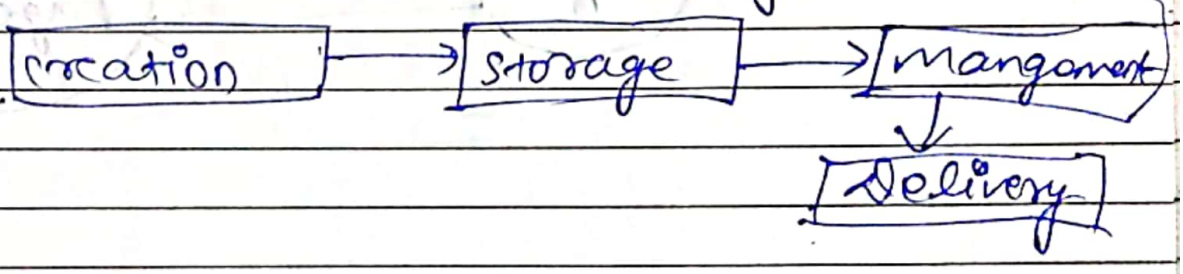
Application of Web Mining
- Used for searching
- predict user behavior.
- Improve power of web search engine by classifying the web documents and identifying the webpage.
- ~~Improve the po~~
- useful of a particular website and e-service eg: landing page optimization.

Web content Mining :-
→ application of extracting useful info from the content of the web documents.

→ contains several types of data - text, image, audio, video etc.

→ provides intresting patterns about user's needs

→ Also known as text mining.



DATE: __/__/____
PAGE: _____

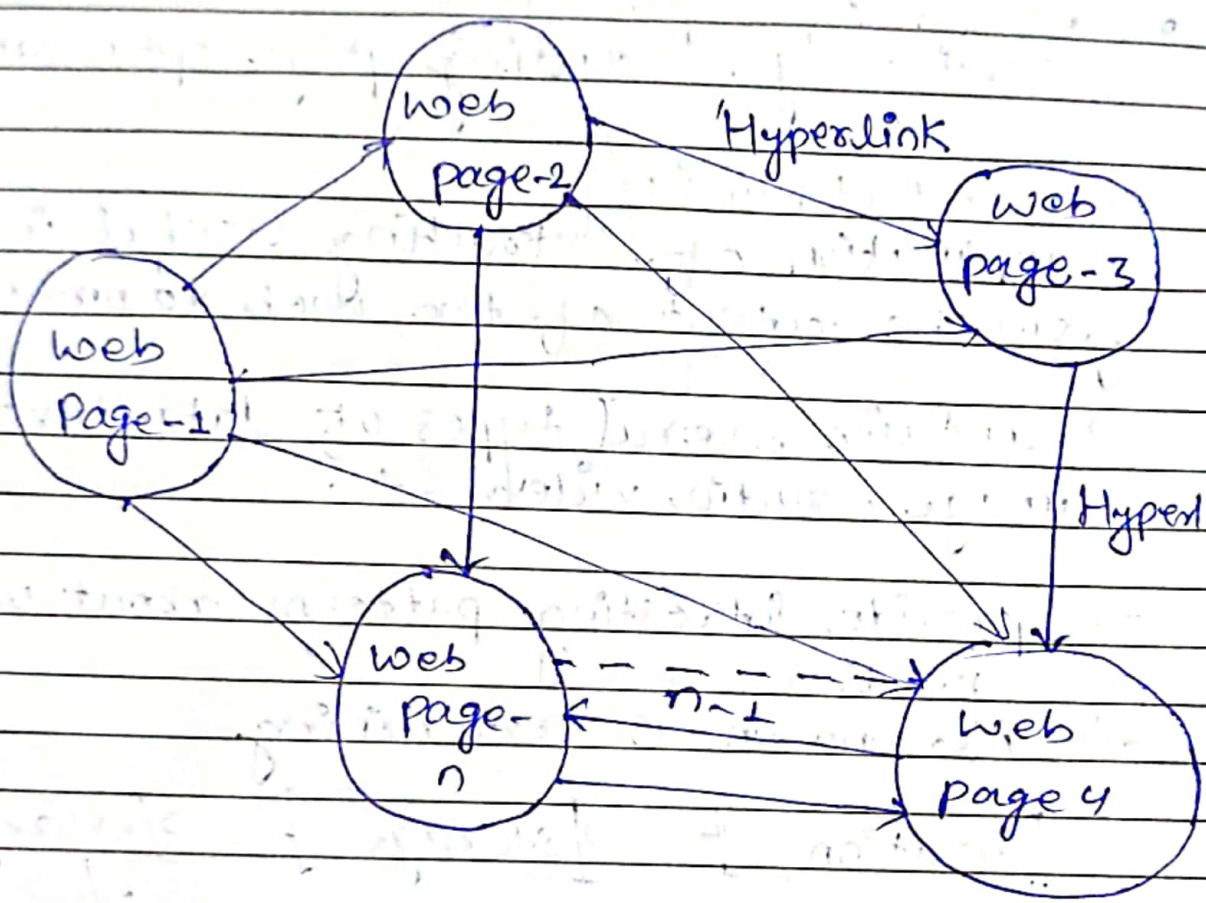
- Web structure Mining :-

→ Application of discovering structure information from the web.

→ consists of web pages as nodes, and hyperlinks as edges connecting related pages.

→ Shows the structed summary of a particular website.

→ To determine the connection between two commercial websites.

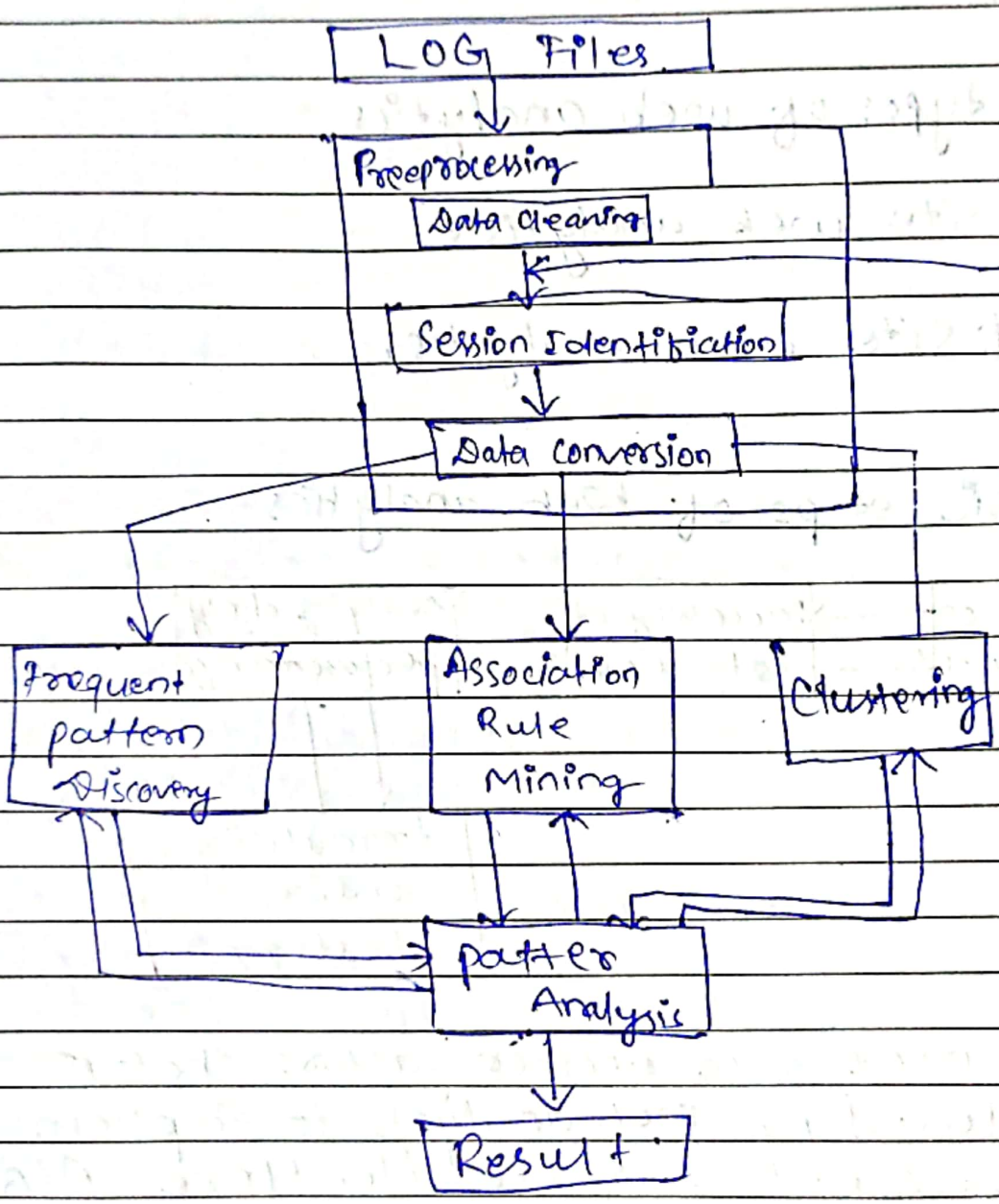


Good Write

DATE: __/__/____
PAGE: _____

Web usage mining :-

Application of identifying or discovering interesting patterns from large data sets



Good Write

DATE: __/__/____
PAGE: _____

⇒ process of gathering, processing and evaluating data from website.

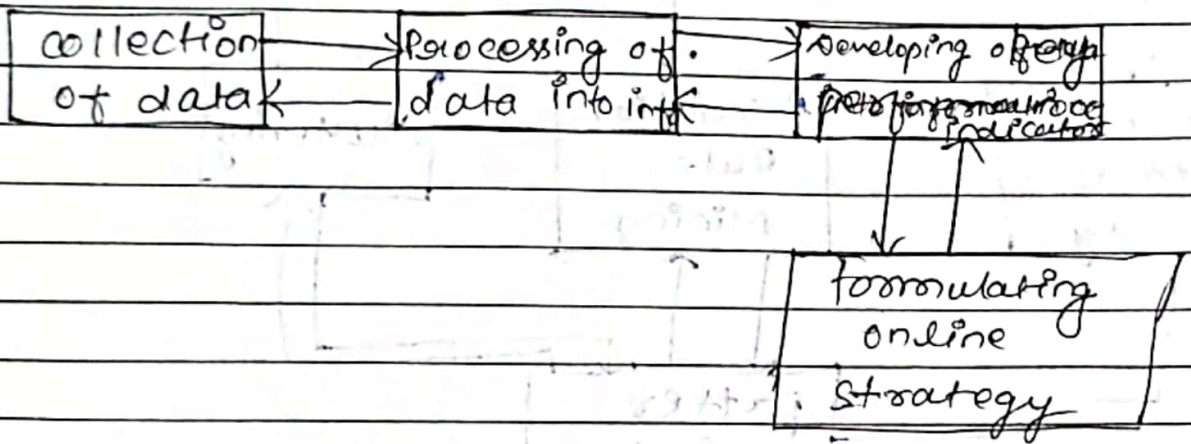
Web analytics has many applications like it shows web navigation, time-specific website traffic, search engine data, visitor data.

Two types of web analytics.

On-site web analytics

Off site web analysis.

Basic steps of web analytics



Experiments are another factor that is introduced by web analysts to improve the performance of website. Here A/B testing is done.

DATE: __/__/____
PAGE: _____

* Tools used in web analytics

- Google analytics
- Yahoo web analytics
- Google website optimizer
- facebook insights
- 4Q by iPerceptions
- Click table
- Optimizely.

Advantage of web analytics:-

- Measure online traffic
- Tracking bounce rate
- Optimizing and tracking of Marketing campaign.
- finding right target audience and its capitalization.
- Improves and optimize websites and web services.
- Conversion rate optimization
- Tracking business goal online.

Good Write

DATE: __/__/____
PAGE: _____

Sentiment analysis

→ an approach to NLP that identifies the emotional tone behind a body of text.

→ It involves the use of data mining, machine learning and artificial intelligence (AI) to mine text for sentiments and subjective information.

i) fine grained :- further categorization to provide a more precise level of polarity.

ii) Emotion detection :-
specifies emotion rather than polarity.

iii) Intent based
action behind a text in addition to opinion.

iv) Aspect based analysis.
specific component being positively or negatively mentioned rather than whole product.

Challenges of sentiment analysis.
i) Inaccuracies in training model.
ii) Sentiments written in neutral manner.
iii) When system cannot understand the tone or context.

Good Write

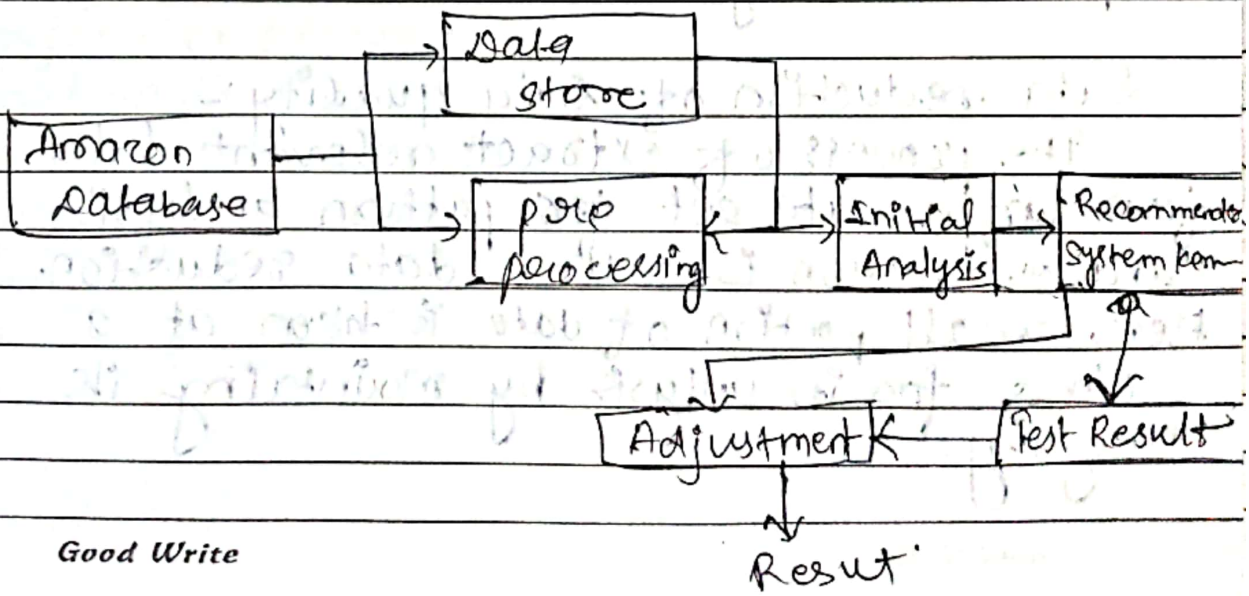
- Use of emoji's and irrelevant information
- contradictory sentiments.

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:-

User based recommendation:-
Here we calculate person's similarity measure, which in needed to determine the closely related users. i;e whose like and dislikes follows the same pattern.

Item based Recommendation:-
Initial aim is to obtain the mean adjusted matrix. Here we use cosine similarity matrix.



Good Write

DATE: __/__/____
PAGE: _____

Working steps of Data mining

1. Data cleaning (removing noisy data & filling gaps)
2. Data Integration (combing differ data sets)
3. Data reduction for data quality (extract relevant data)
4. Data transformation
5. Data mining
6. Patter Evaluation
7. Representing knowledge in Data mining

Data cleaning :-

After collecting data from different sources, it need to be cleaned using methods like binning. This methods cleans up the noisy data and mate it eligible for data integration. Dirty data leads to poor insights and increase the cost & time.

Data Integration :-

miner

Combining different data sets for proper data analysis is called data integration. Some inconsistency of the data also cleaned up in this stage.

Data reduction of Data quality :-

The process of extract relevant data from large data set for pattern analysis and evaluation is called data reduction. Here small portion of data is taken at a time for its anlysis by maintaing its integrity.

DATE: __/__/____
PAGE: _____

4 unit

Data transformation

In this standard process, engineers transform data into an acceptable form to align with mining goals.
Technique to do so are smoothing or eliminating noise from data, aggregation normalization or discretization.

Data mining:-

to extract useful trends and optimiz. knowledge discovery to generate business intelligence. Use methods like clustering classification or other modelling techniques to ensure accuracy.

Pattern evaluation

From these data specialist can generate patterns & business knowledge. Visualization data mining is use for better understa

Representing knowledge in datamining:-

Use combination of data visualization, reports and others mining tools to share information with others. These ensights presented in reports help the higher authority to take decisions, generate new business, eliminate waste, and create better advertising campaigns.

DATE: __/__/____
PAGE: _____

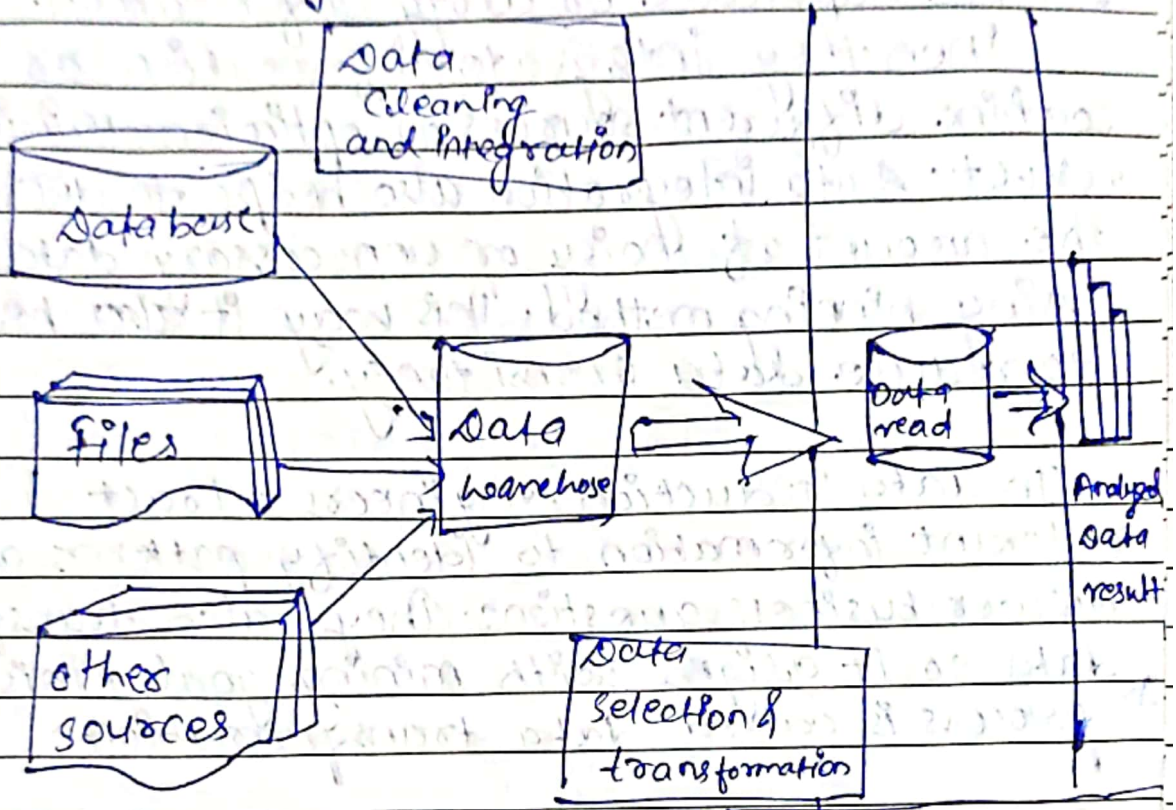
Difference between Data Mining & web Mining..

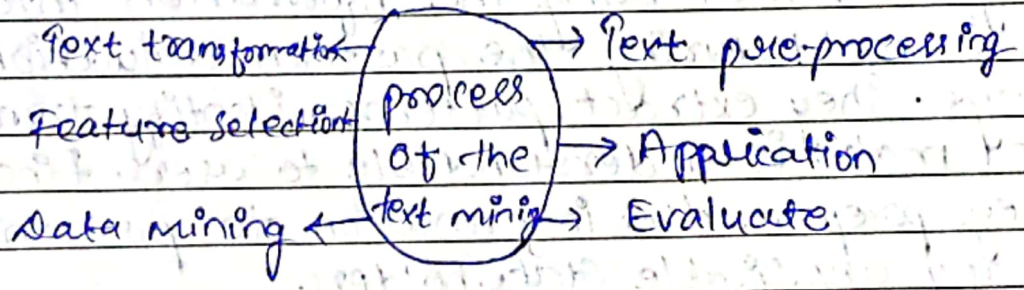
Data Mining	Text mining
• Statical technique of processing raw data in a structured form.	part of data mining which involves processing of text from documents.
• Pre processing database and spreadsheets are used to gather information.	The text is used to gather high quality information.
• Processing of data is done directly.	Processing of data is done linguistically.
• Statical techniques are used to evaluate data.	Computational linguistic principles are used to evaluate text.
• Data stored in structured format.	Data is stored in unstructured format.
• Data is homogenous and easy to retrieve.	Data is heterogenous and is not easy to retrieve.
• Mining of mixed data.	Mining of text data is only done.
• AI, ML and statistics.	Pattern recognition & NLP
• Marketing medicine healthcare.	Business customer profile analysis

Good Write

DATE: __/__/____
PAGE: _____

Data mining





Good Write

DATE: __/__/____
PAGE: ____

Data mining process

First, specialist need to cleans data to remove duplicate or dirty information.

Then they integrate information or combine different sources to optimize mining result. Data integration also helps to decrease the amount of noisy or unnecessary data, using binning method. This way it also helps to reduce data inconsistency.

In data reduction, engineers extract relevant information to identify patterns and answer business questions. They also transform data so it aligns with mining goals. This process is called data transformation.

In data mining, engineers assign relevant patterns to each data set before they extract it. They then generate models ~~which~~ with clustering or classification techniques.

Engineers then bring the information into the real world during the pattern evaluation stage. They extract patterns, identify trends and make it understandable to users. Finally they prepare the information to present to any applicable stakeholders.

Business owner use data mining insights to optimize decision making, increase sales and learn more about customers.

DATE: __/__/____
PAGE: _____

Unit 4

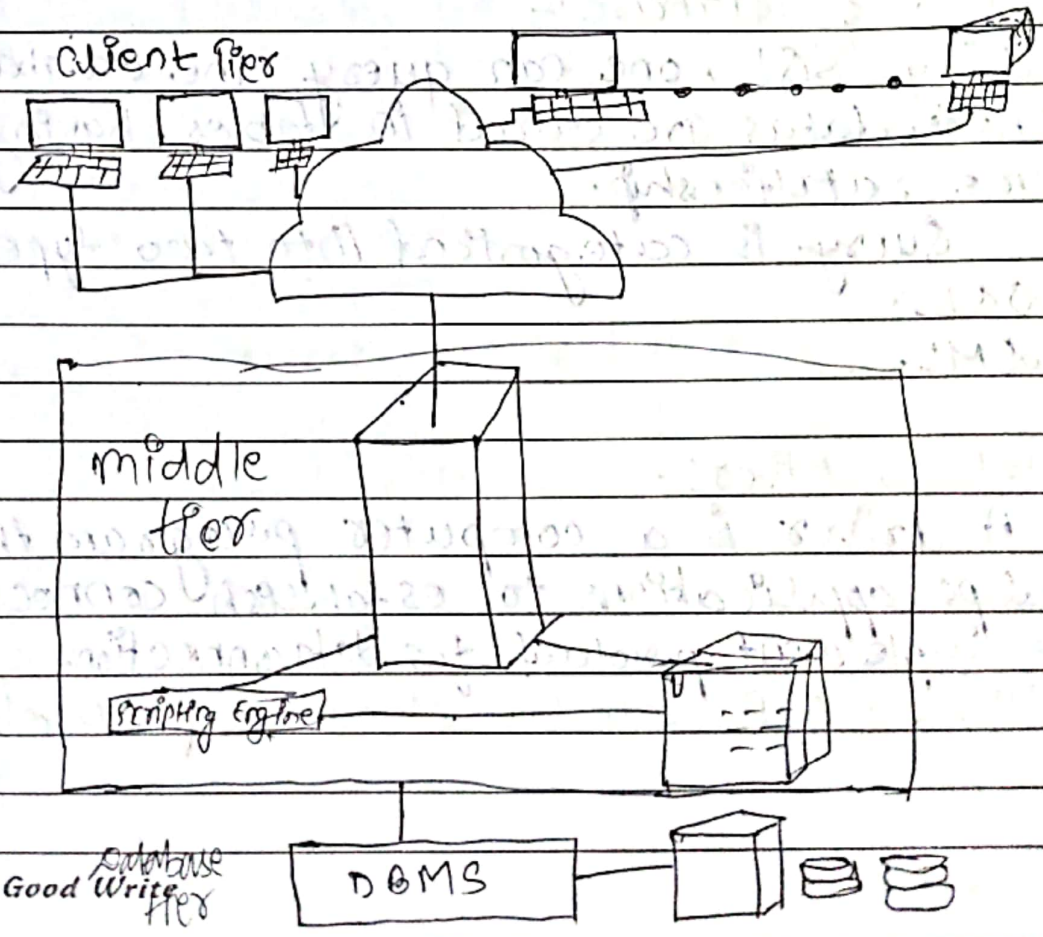
- [x] web database
- [ ] database connectivity
- [ ] JDBC
- [ ] ODBC
- [ ] Database to web connectivity.

web database

A web database is a database application designed to be managed and accessed through the internet.

Website operators can mange this collection of data and present analytical result based on the database application.

Three tier web database architecture.



Good Write

DATE: / /
PAGE:

Database connectivity

Database connectivity allows the client software to communicate with the database server software.

Elements of frontend application/ websites like buttons, fonts or menus need to be connected to the database to deliver relevant information to the end-user.

Database connectivity can be done using different programming languages, Java, C++ and HTML. By using these language one can connect a program to a particular database and can query data to access & manipulate them.

Querying database:-

Using SQL, one can query the database where datas are stored in tables having some relationship.

Query is categorised into two types.

- DQL.
- DML.

Database driver:-

A driver is a computer program that helps applications to establish connection. It implement protocol for data connection.

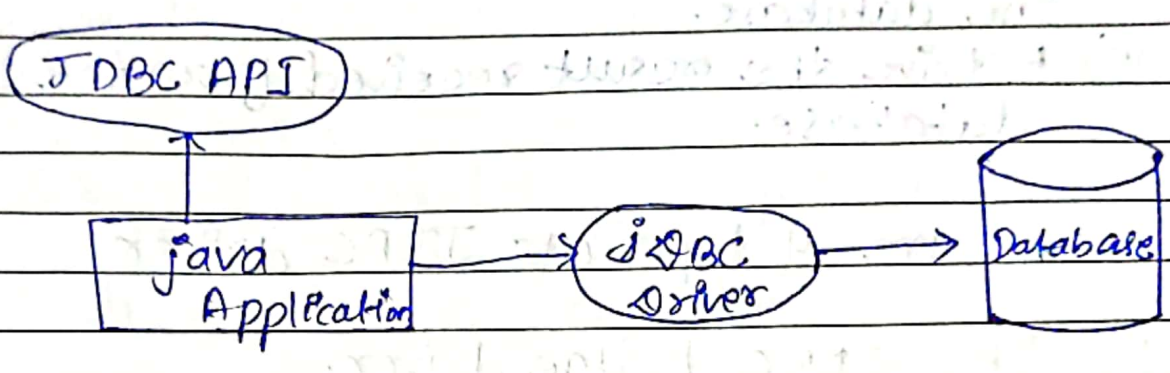
JDBC & ODBC are two standard protocols.

DATE: __/__/____
PAGE: _____

The driver work like an adaptor which connects a generic interface to a specific database vendor implementation.

Java Database Connectivity (JDBC)

Its an API for java programming language to connect and execute query with the database.



It access data from the relational database table. JDBC use drivers to connect with data base. Some of the popular JDBC API Interfaces include ResultSet interface, callableStatement interface, Driver Interface Connection Interface, statement Interface and Rowset Interface.

popular classes of JDBC API

- DriverManger API
- Blob class
- Clob class
- Types class

Good Write

DATE: __/__/____
PAGE: _____

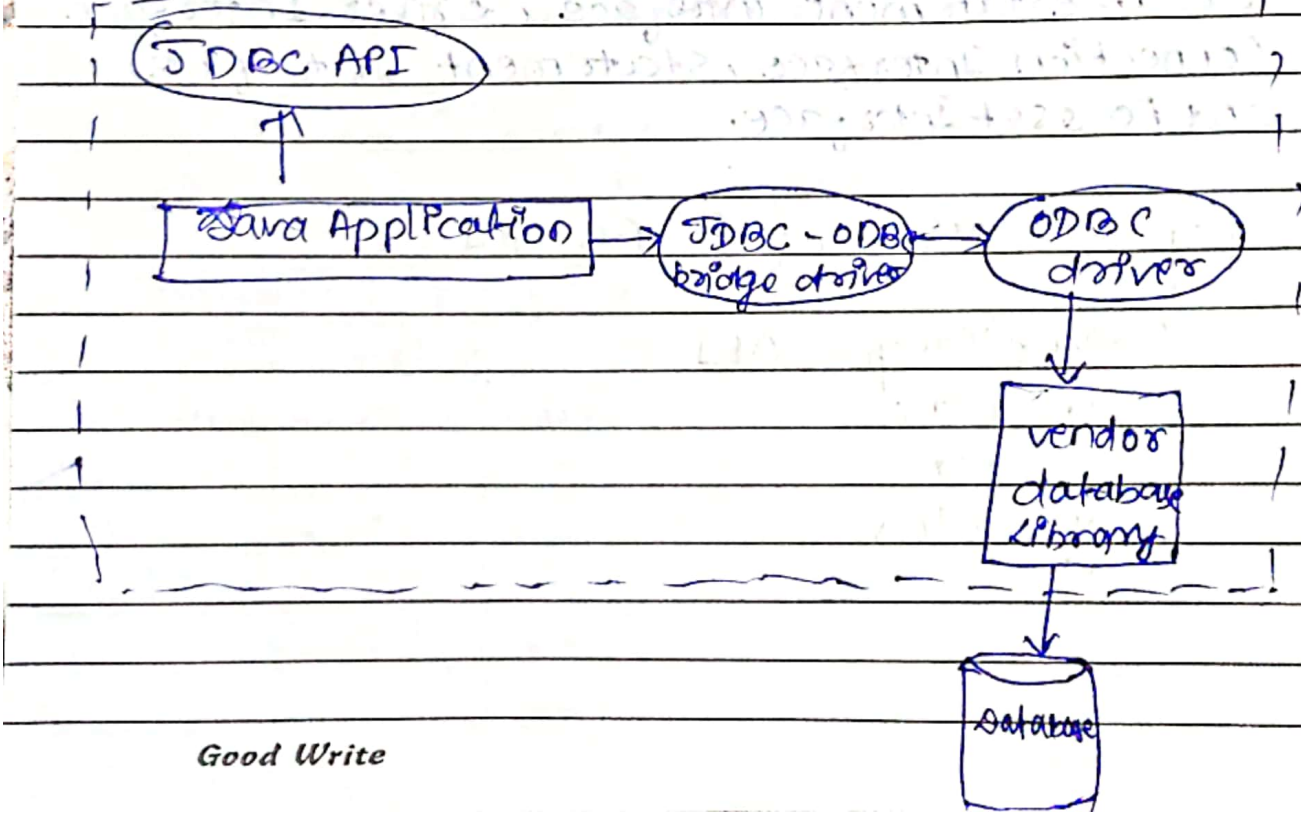
ODBC API uses ODBC driver which is written in C language (i.e platform dependent and unsecured)

JDBC API can do:-

i) Connect to the database
ii) Execute queries and update statements to the database.
iii) Retrive the result received from the database.

There are 4 types of JDBC driver

1) JDBC-ODBC bridge driver.
Uses ODBC driver to connect with database. The JDBC-ODBC bridge converts the JDBC method calls into the ODBC function calls



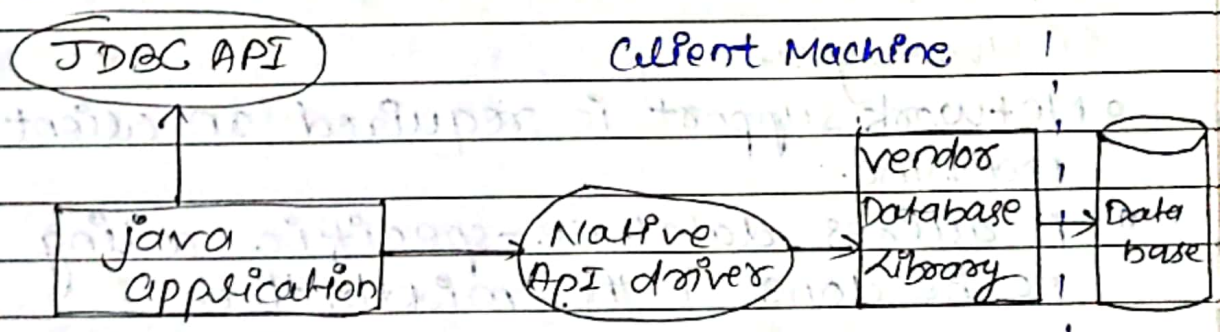
DATE: __/__/____
PAGE: _____

Advantage:-
- easy to use
- can be easily connected to any database.

Disadvantage:-
- Performance degraded because JDBC method calls is converted to ODBC function calls.
- Since its platform based, so it should be installed on the client machine.

2. Native API driver:-

The native API driver uses the client-side libraries of the database. Its not entirely written in java. converts JDBC method calls into native calls of the database API



Advantage:-
- performance upgraded than JDBC-ODBC bridge driver.

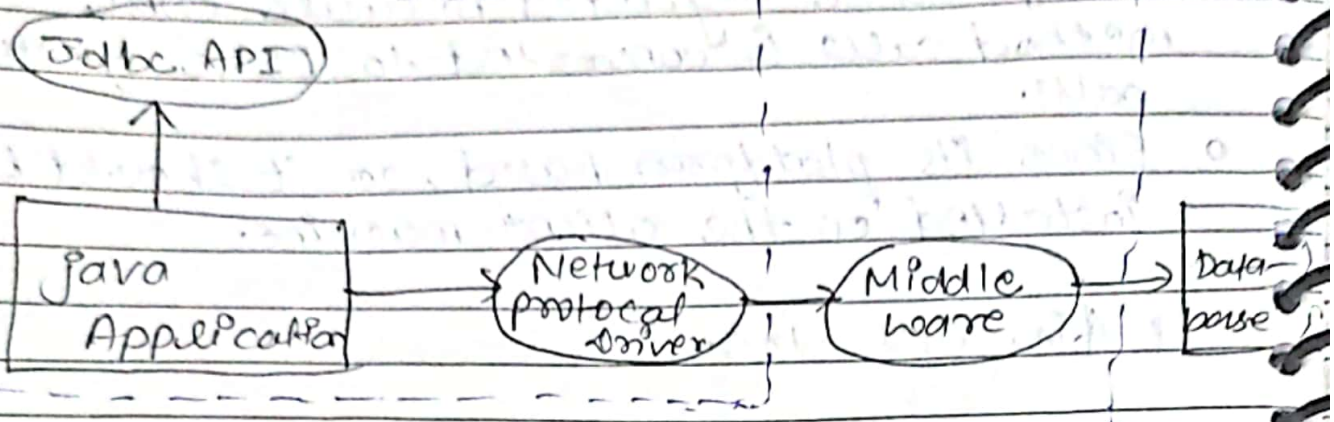
Disadvantage:-
Native driver and vendor client library needs to be installed on client machine.

Good Write

DATE: __/__/____
PAGE: _____

3) Network Protocol driver:

The Network protocol driver uses middle ware that converts JDBC calls directly or indirectly into the vendor-specific database protocol. fully written in Java.



Advantage:-

- No client library is required, because of application server.

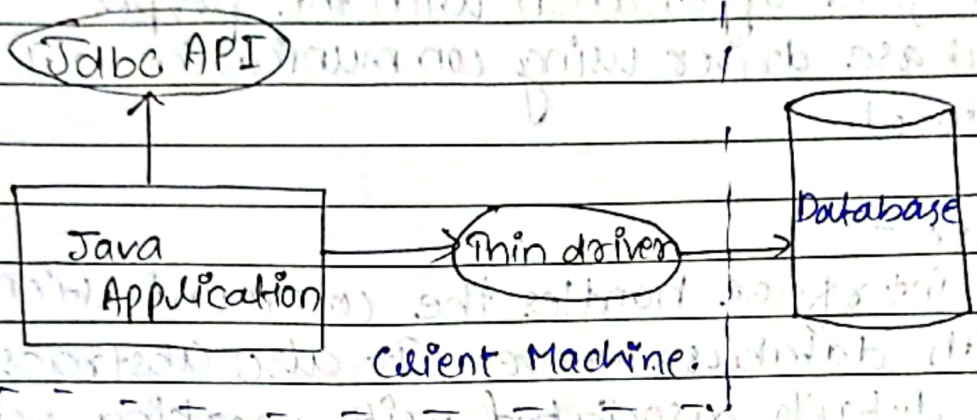
Disadvantage:-

- Network support is required on client machine.
- Requires database-specific coding to be done in the middle tier.
- Cost is high because it requires database specific coding to be done in the middle tier.

DATE: __/__/____
PAGE: _____

4) Thin driver

The thin driver converts JDBC calls directly into the vendor-specific database protocol. That's why its known as thin driver. It is fully written in Java language.



Advantages

- Better performance than all drivers.
- No software is required at client side or server side.

Disadvantage

- Drivers depend on the Database.

JDBC connectivity with 5 steps.

1. Register the Driver class
2. Create connection. (Driver Manager)
3. Create statement (connection interface)
4. Execute queries (statement interface)
5. Close connections. (connection interface)

Good Write

DATE: / /
PAGE:

JDBC components

- **DriverManager:-**
  This class manages a list of database drivers. Matches connection request from the java application with the proper database driver using communication sub protocol.

- **Driver:-**
  This interface handles the communication with database server. It also abstracts the details associated with working with Driver objects.

- **Connection:**
  This interface with all methods for contacting a database. All the communication with database is through connection obj. only.

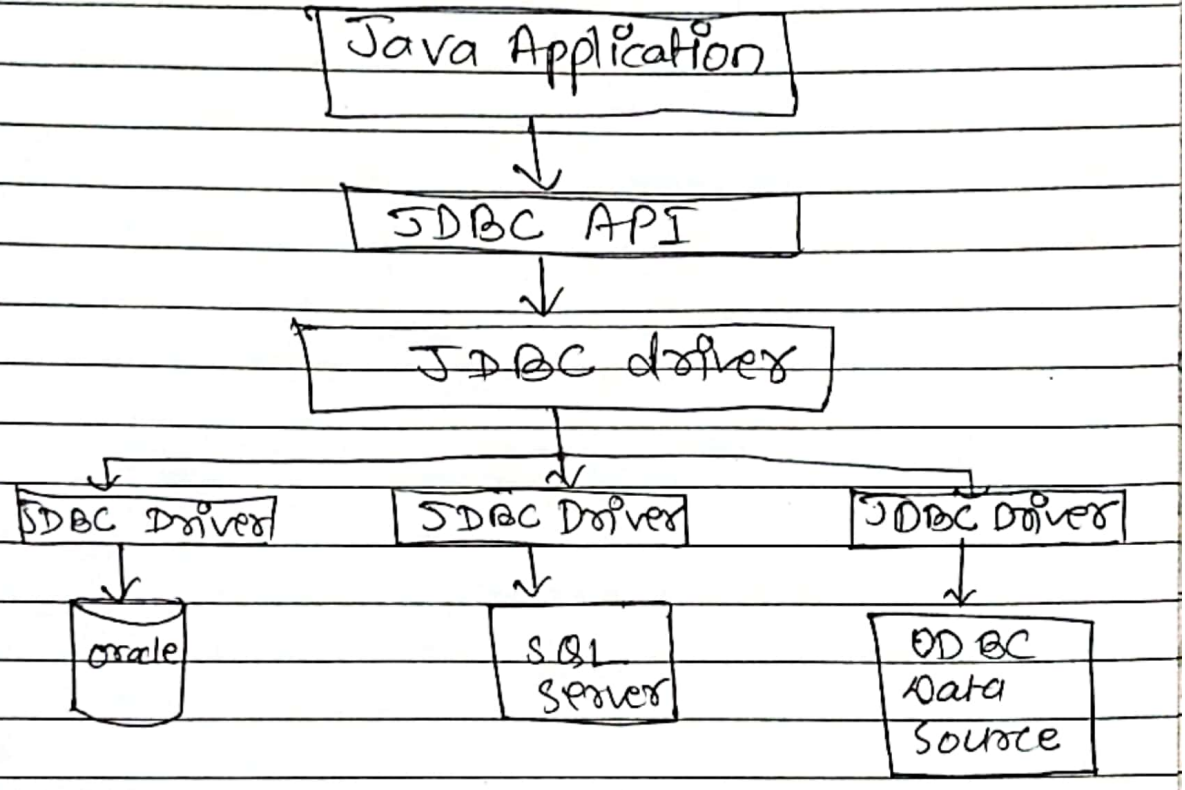
- **Statement:-**
  Here we use objects created from this interface to submit the SQL statements to the database.

- **ResultSet:-**
  These objects hold data retrieved from a database after we execute an SQL query using Statement object.

Good Write

DATE: __/__/____
PAGE: _____

- SQL exception.
  This class handles errors that occurs in a database application.



~~Database~~

ODBC

Open Database Connectivity (ODBC) is a standard (API) for accessing DBMS. The designer of ODBC make it