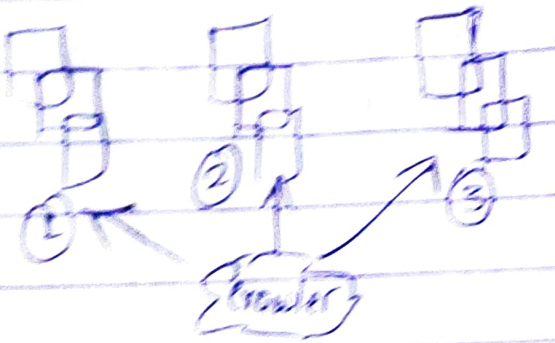
① Web Crawling:

It is a process by which we gather pages from the web, in order to index them and support a search engine

② Objective:

→ to quickly and efficiently gather as many useful web pages as possible, together with the link structure that interconnects them



③ Features:

① Robustness: Crawlers must be designed to be resilient to traps

② Politeness: web servers

→ have both implicit and explicit policies regulating the rate at which a crawler can visit them

③ Distributed:

④ Scalable:

⑤ Performance and Efficiency:

⑥ Quality ⑦ Freshness

⑧ Extensible

Date:
Page:

⓪ Basic Operation:

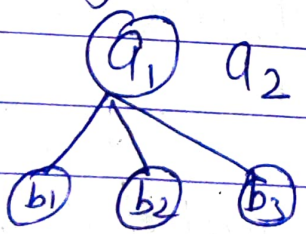
① The crawler begins with one or more URLs that constitute a seed set

② It picks a URL from this seed set, then fetches the web page at the URL

③ The fetched page is then parsed, to extract both text and the links from the page (each of which points to another URL)

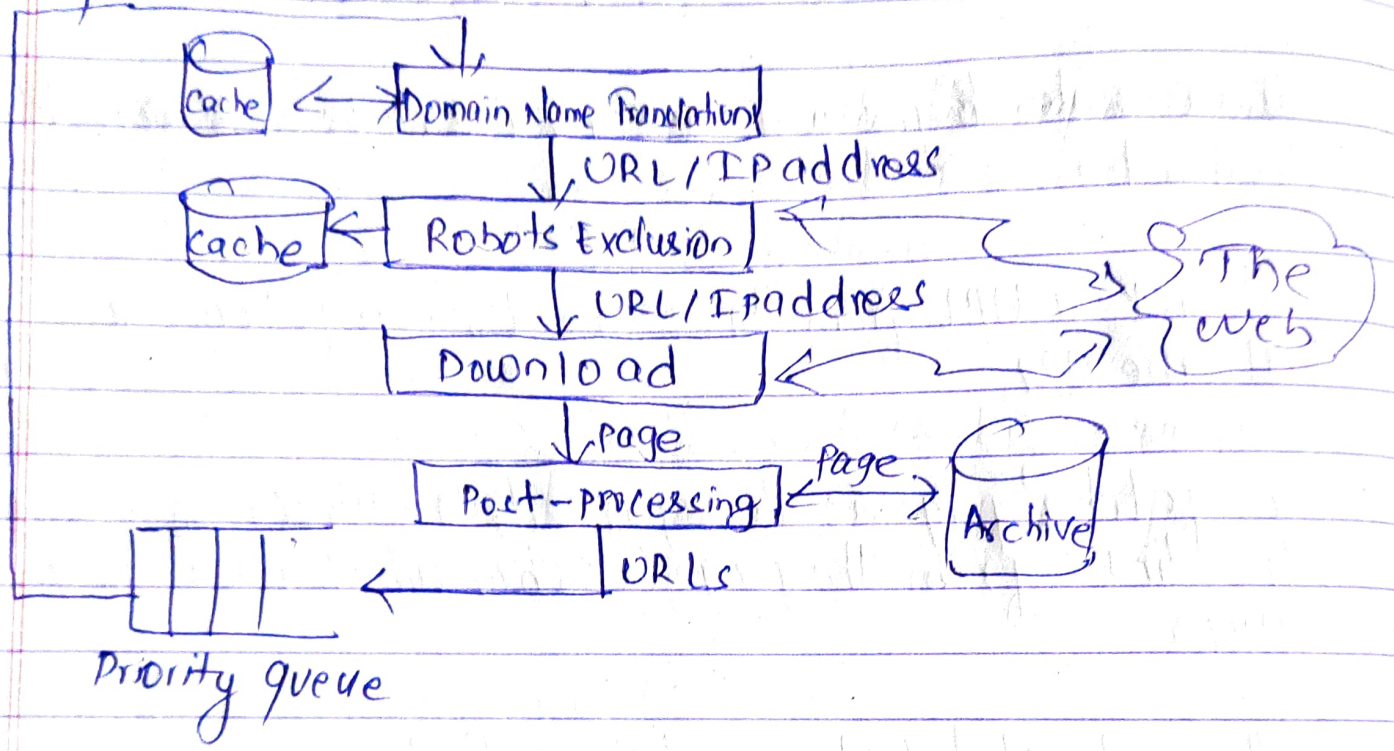
④ The extracted text is fed to a text indexer.

⑤ The extracted links (URLs) are then added to a URL frontier, which at all times consists of URLs whose corresponding pages have yet to be fetched by the crawler.



Date:
Page:

ⓐ Components:



ⓐ Domain Name Translation:

→ Translating host name into 32-bit IP address
→ for speed, maintains its own cache which require thousands of translations per second.
→ Cache maintains mapping between host names and IP addresses

ⓐ Robots Exclusion:

→ With IP address, available crawler must check that access to the page is permitted by the website.
→ robots exclusion protocol
→ access permissions are obtained by appending the path "/robots.txt" to the host name

Date:
Page:

⓪ Download:
→ Access the website via the HTTP protocol and downloads the page
→ Crawlers may be required to resolve directions during download
→ Use of JS cause trouble.

⓪ Post-processing:
→ After download, the crawler stores the page in an archive for indexing by the search engine

→ During ~~indexing~~ post-processing, the crawler must respect conventions requesting that it not index a page or follow certain links:

```
<meta name="robots" content="noindex">
<a rel="nofollow">
```

⓪ Priority queue:
→ URLs extracted during post processing are inserted into pq.

Date:
Page:

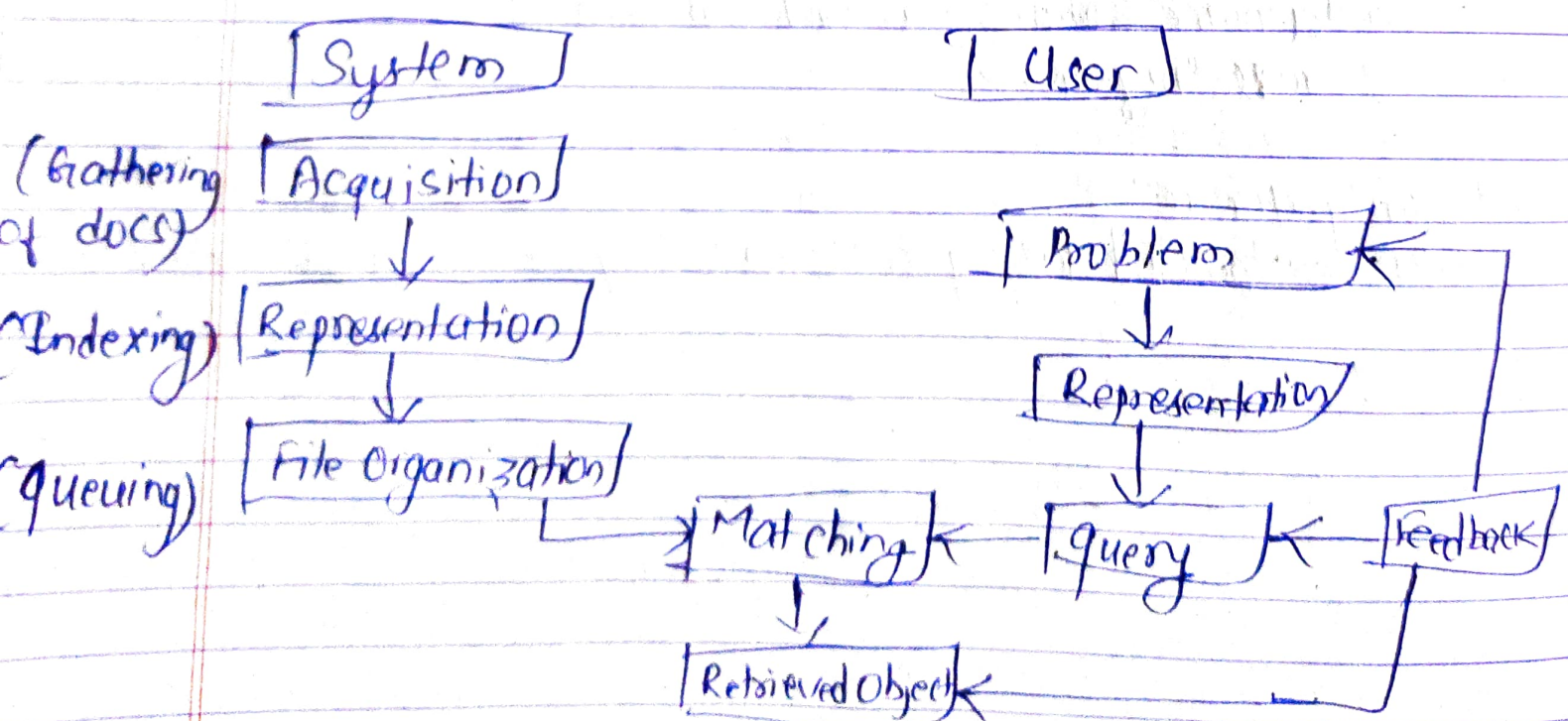
ⓐ Web Information Retrieval:

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

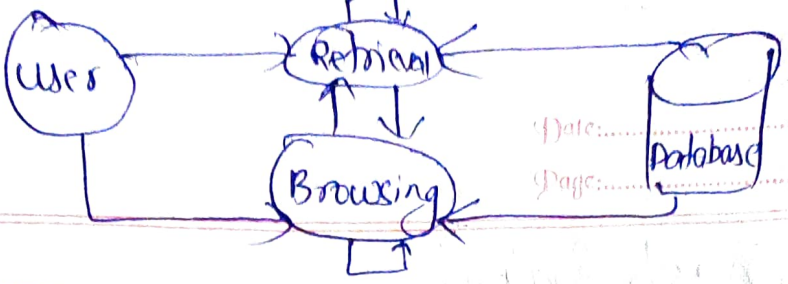
ⓑ IR Mode:

→ Selects and ranks the document that is required by the user or the user has asked for in the form of query.

ⓒ Components of IR model:



User Interaction:



Q Difference :

	IR retrieval	Data Retrieval
(I)	from document repositories	from a database management system
(II)	Retrieves information about a subject	Determines the keywords in user query and retrieves the data
(III)	Small error are likely to go unnoticed.	Single Error object means total failure
(IV)	Results are ordered by relevance	Not ordered by relevance
(V)	probablistic model	deterministic model
(VI)	Results are approximate matches	exact matches
(VII)	does not provide solution to user of database system.	provides solution
(VIII)	Not well defined structure and semantics	Well defined

Date:
Page:

① Search Engine:

→ the practical application of information retrieval techniques to large scale text collections

→ Eg: Google, Yahoo

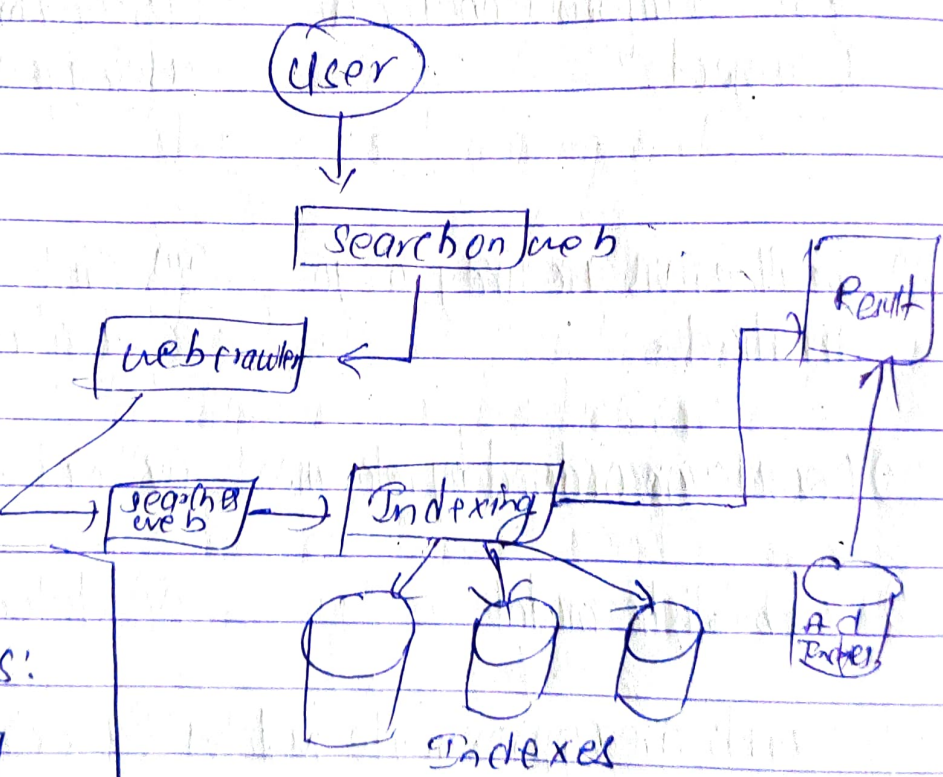
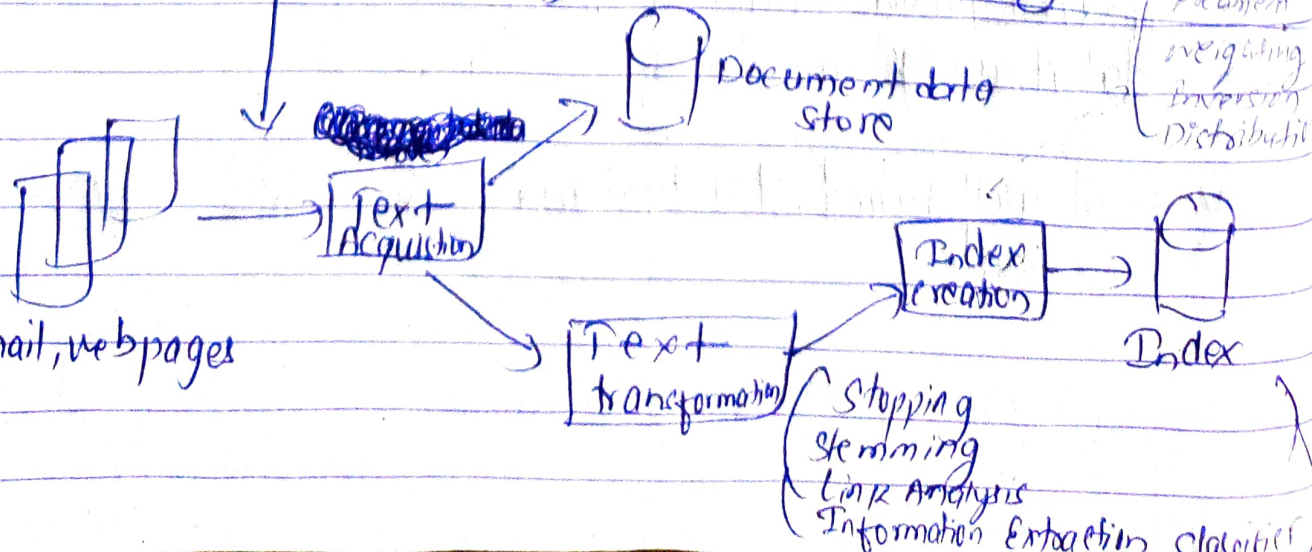


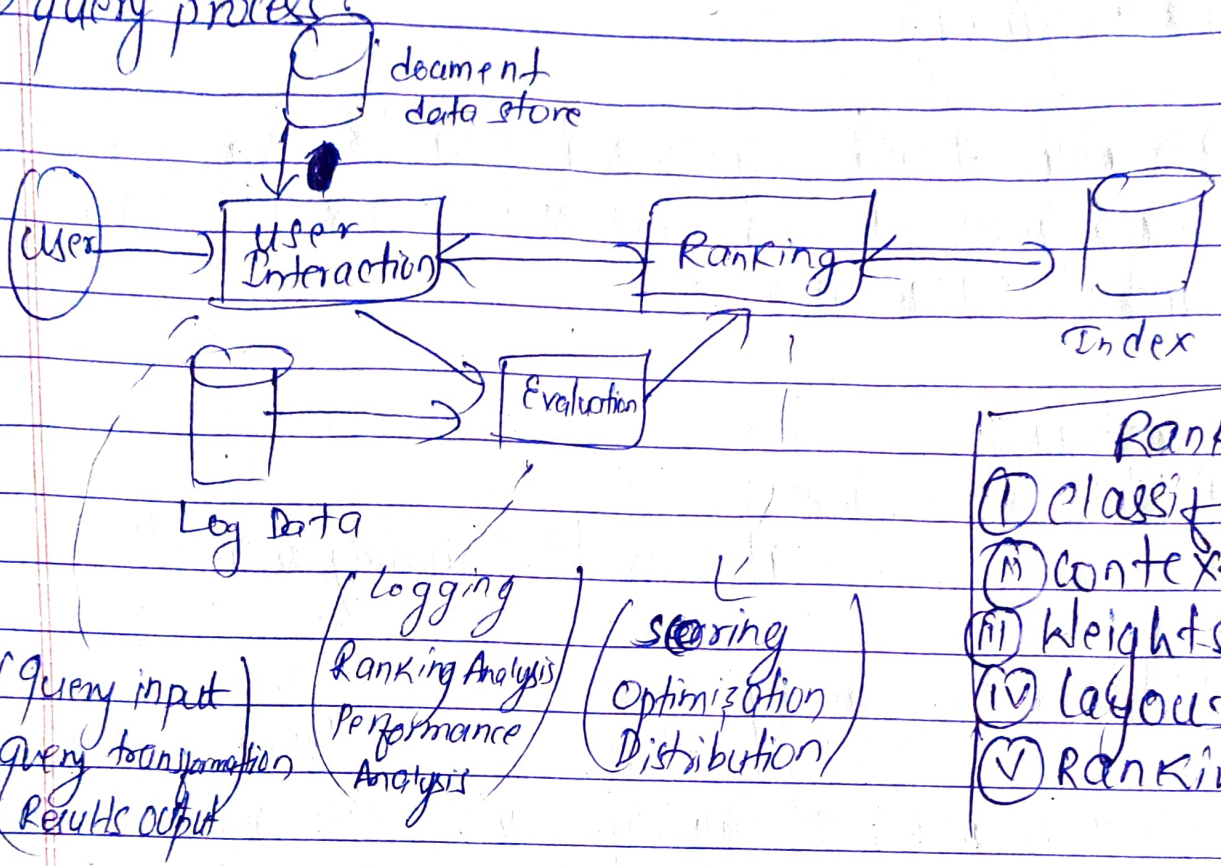
fig: Search Engine working

① Two Major functions:
(i) Indexing process
(ii) Query process



Date:
Page:

⑥ query process:



Ranking:
(I) classify
(II) context
(III) weights
(IV) layouts
(V) Ranking

⑦ Search Engines Features:

① Performance:
→ Efficient search and Indexing

② Incorporate new data:
→ Coverage and freshness

(Statistics)

③ Scalability:
→ Growing with data and users

④ Adaptability:
→ Tuning for applications

⑤ Specific Problems:
→ Eg: Spam

Date:
Page:

⑥ Web Mining:

→ To automatically discover and extract information from web documents and services. i.e. to extract from WWW and its usage patterns

→ Techniques:

① Web Content Mining:
It relates to text mining and NLP.

② Web Structure mining:
Used to extract the structure of web pages e.g. pages as nodes and hyperlinks and edges connecting related pages, → describes the structure summary of a website.
→ to determine the connection between two commerical websites, web structure is useful.

③ Web usage Mining:
→ Identifying interesting usage pattern from large datasets
→ It determines user behaviour
→ users can collect data from logs, so it is called log mining

⑥ Applications:

① Help improve power of search engines by classifying web documents.

(2) Used for scheduling e.g. google, yahoo ~~and vertical~~

(3) Predict User behaviours

(4) Landing Page Optimization in particular website and e-service

(5) Tools include Apache, Scrapy logs.

⑥ Text Mining:

→ Essentially a sub field of data mining as it focuses on brining structured to unstructured data and analyzing it to generate novel insights.

→ Types of data: (i) Structured data → (SVM, Naive Bayes)
(ii) Unstructured data
(iii) Semi-Structured data

→ Steps involved:
(i) Gathering Unstructured data (Info Extraction)
(ii) Remove Anomalies (Information Retrieval)
(iv) Convert Unstructured to Structured data(↓)
(v) Analyze the patterns within the data (categorization and clustering)
(vi) store all the valuable information (summarization) in a database

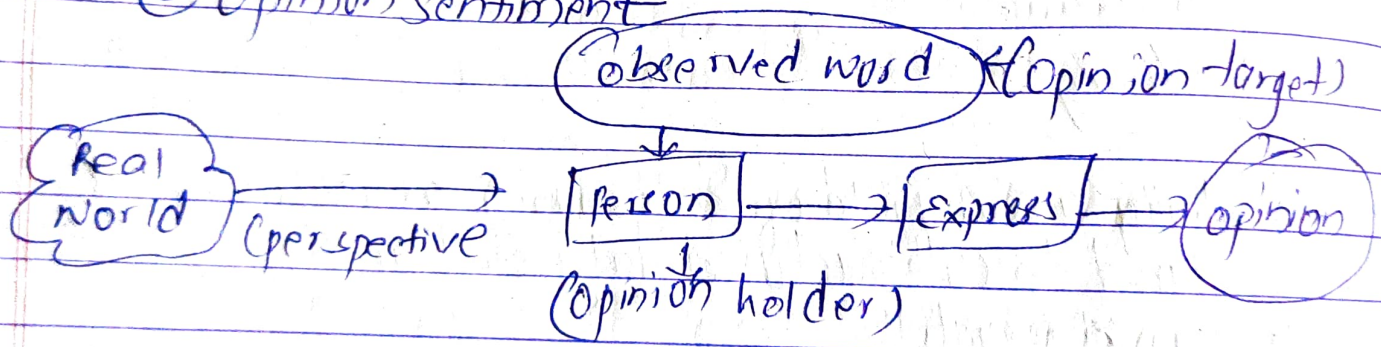
Date:
Page:

⓪ Opinion Mining:

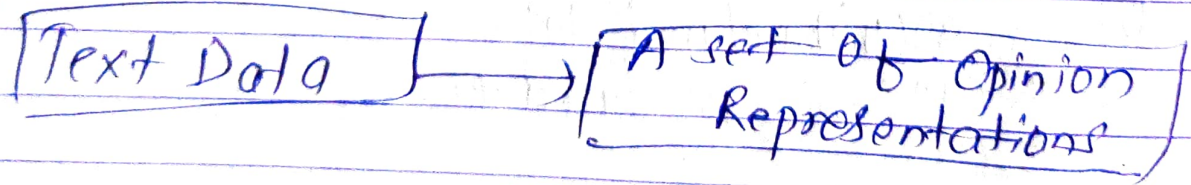
↳ a subjective statement describing what a person believes or thinks about something

↳ opinion representation:

1. Opinion holder
2. Opinion target
3. Opinion Content
4. Opinion Context
5. Opinion Sentiment



→ Task:



→ Why Opinion Mining:

(i) Decision Support

(ii) Understand People

(iii) "Voluntary Survey"

Date:
Page:

⊙ Social Web Mining:

→ the process of mining social data

→ involves collection, processing and analysis of raw data obtained from social media platforms such as Fb, Instagram, twitter etc to uncover meaningful patterns and trends, draw conclusions, and provide insightful and actionable information

⊙ Data collected:

→ Age → clicks
→ Gender → No. of followers
→ location → likes

⊙ How it works:

→ involves:

- Combination of (I) Statistical techniques
  - (II) Mathematics
  - (III) Machine learning

⊙ Used by (I) Government Agencies

(II) Companies
(III) Hotels
(IV) Retailers
(V) Airlines

⊙ Used for:

(I) Trend Analysis
(II) event Detection
(III) Social spam detection
(IV) Ecommerce
(V) Brands
(VI) Research purposes

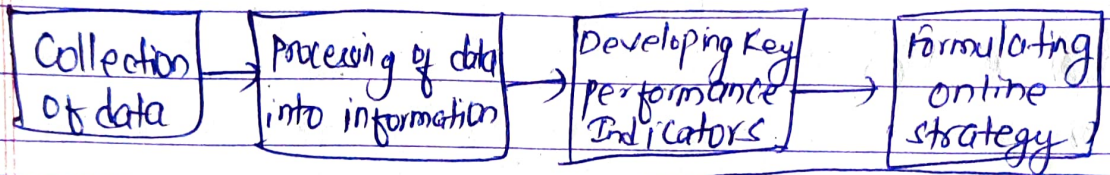
Date:
Page:

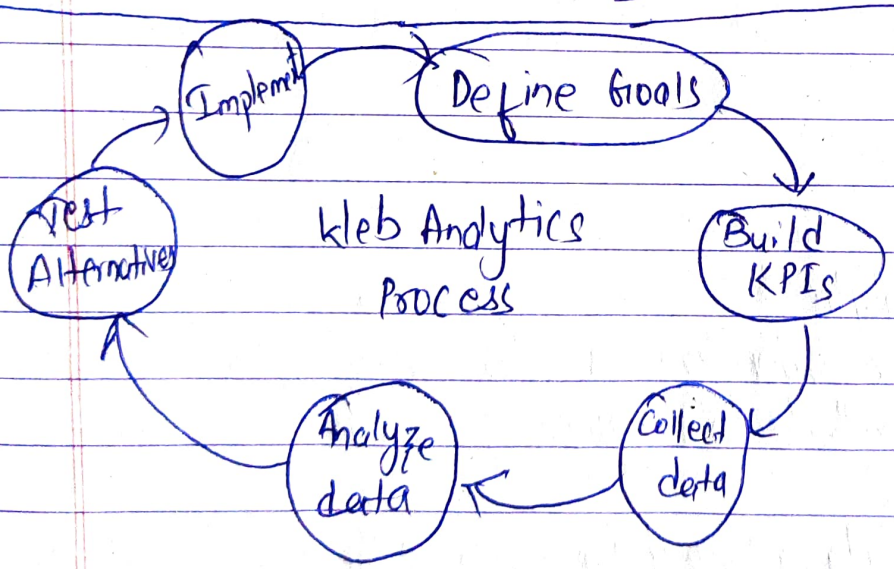
	Web Mining	Data Mining
①	Extract info from web documents	Extract info from large datasets
②	Very useful for particular web-site.	Very useful for webpage analysis
③	access data publicly	access data privately
④	information from structured, unstructured and semi-structured web pages	information from explicit structure
⑤	Special tools like Scrapy, PageRank	ML algorithms
⑥	Web content Mining, Web Structure Mining	Clustering, Regression etc

Date:
Page:

① Web Analytics:
→ process of analysing the behaviour of visitors to a website
→ quantifiable measure
→ (i) bounce rates,
(ii) Unique users,
(iii) User sessions,
(iv) On-site search queries

② Process:





③ Terminology:
(i) Site References: Means from which website or source your readers come from

(ii) Keywords and Key Phrases:

(iii) Building Block Terms:
(Page, Page Views, Visits/ Sessions, Return, Unique Visitors, New, Repeat)

Date:
Page:

(a) Visit Characterisation Terms:
(I) Entry page
(II) Landing Page
(III) Exit Page
(IV) Visit duration
(V) Referrer
(VI) Click through / click-through rate
(VII) Page View per Visit

(b) Content Characterisation Terms:
(I) Page Exit ratio
(II) Single Page Visit
(III) Bounce Rate

(c) Conversion Metrics:
(I) Events
(II) Conversion

→ Types:
(I) Off site Web Analytics:
→ practice of monitoring visitor activity outside of an organization's website to measure potential audience

→ focuses on data collected from across the web, such as social media, search engines and forums.

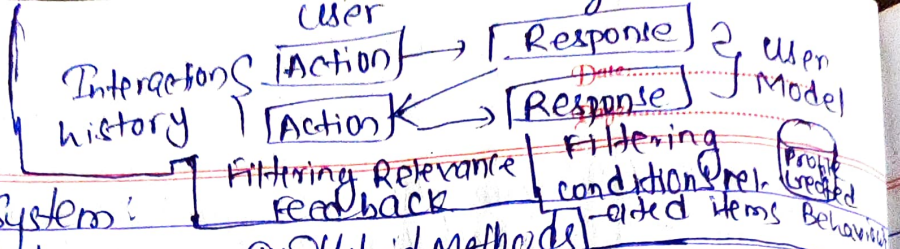
→ provides industry wide analysis

(2) Onsite Web Analytics:
→ Analytics to track the activity of visitors to a specific site to see how the site is performing
→ One's own site

⓪ Web Analytics Platforms:
(I) Google Analytics
(II) Adobe Analytics
(III) Stat Counter
(IV) Mixpanel
(V) Baidu Analytics
(VI) Crazy Egg
(VII) HubSpot

⓪ Need/Advantages:
(I) To understand the traffic
(II) To create better user experience
(III) To understand customer behaviour
(IV) To track effectiveness of online marketing campaign
(V) Help to understand ROI of social media campaign
(VI) Improve SEO
(VII) Identify pain points

⓪ Disadvantages:
(I) Time Consuming
(II) Costly
(III) Not Accurate (100%)



⓪ Recommender System:
① Content Based
② Collaborative filtering:
(II) Hybrid Methods

↳ methods that are based solely on the past interactions recorded between users and items in order to produce new recommendations

↓

→ Two classes:

① Memory based → User-User, Item-Item (Less Scalable)

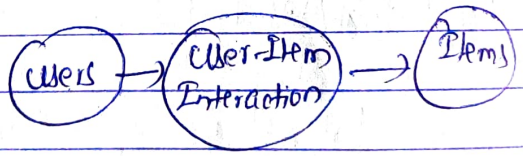
② Model based → Matrix factorisation

→ Advantage:

① no pre information about users or items.

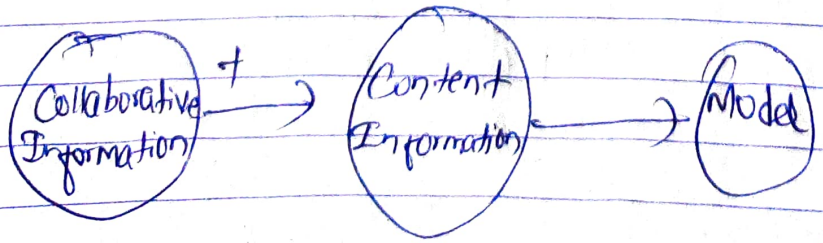
→ Disadvantage:

① "Cold start Problem"



① Content Based Method:

→ build a model based on the available features, that explain the observed user-item interactions



→ Advantage

① Suffer less from cold start problem

→ Disadvantage:

→ totally new users or hard to adapt

→ User prof
interest

⓪ JDBC: C

→ A technolo
with data

JAVA Application

→ is an API
by which

→ JDBC A

① Jav

Driver
connecti
statem
Result

⇒ JDBC
-systems

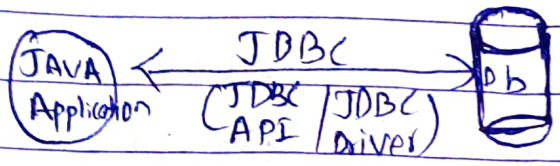
⓪ Features
① Stand
② Datab
③ stor

→ User profiling building refers to identifying users interests and domains of relevance ()

Date
Page

ⓐ JdBC : (Java Database Connectivity)

→ A technology which is used to connect Java Application with database



→ part of JDK software

→ is an API which provides various classes and interfaces by which we can connect Java Application with DB.

→ JDBC APIs has two packages:

① Java.sql

- (I) Driver, connection, statement, ResultSet
- (C) Date, Time, Timestamp, Types

⑪ javax.sql

- (I) DataSource, RowSet, RowSet Listener
- (C) RowSetEvent, ConnectionEvent

⇒ JDBC is an abstraction which is provided by Sun Micro-systems and implemented by Database Vendors

ⓑ Features:

① Standard API
⑪ Database Independent API
(M) stored procedure, triggers, views
⑪ platform independent technology
(N) CRUD or SCRUD

Date:

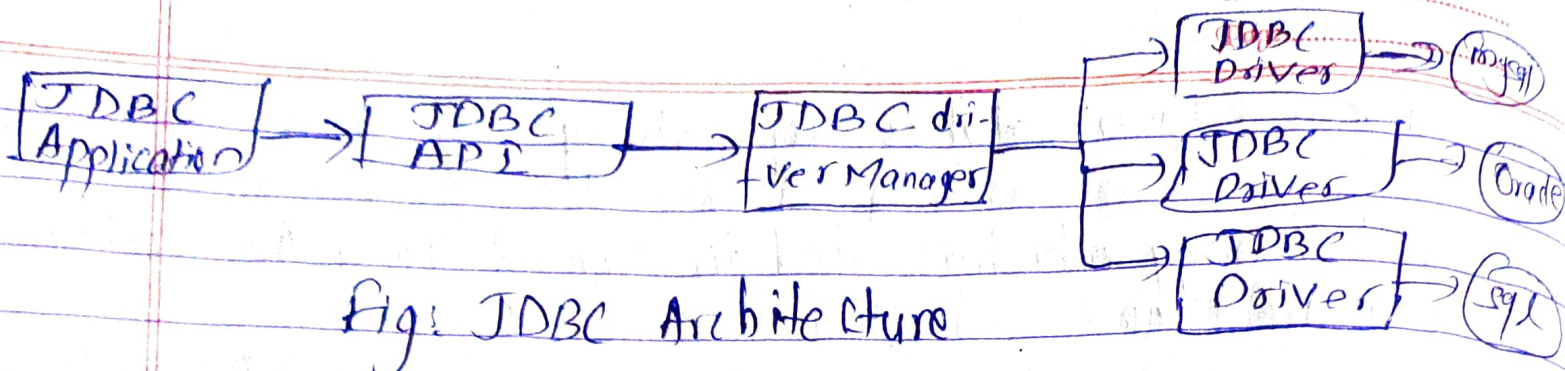


fig: JDBC Architecture

Q) Why use JDBC?

⟹ before functions were appended in front-end files provided by database vendors to connect to the database

	JDBC	ODBC
(i)	both platform and database independent	platform dependent and database independent
(ii)	Implemented in Java without pointer	Implemented in C with pointer
(iii)	Introduce d by SUN Microsystems in 1997	Introduced by M/s in 1992
(iv)	Performs comparatively better with java applications	performs comparatively slower
(v)	object oriented	procedural
(vi)	Stands for Java DB Connectivity	Stands for Open DB Connectivity

Steps to write JDBC program:

(I) load the JDBC driver class or register the JDBC driver

(II) Establish the connection

(III) Create a statement

(IV) Execute the sql commands on database and get the result

(V) print the result

(VI) Close the connection

Date:
Page:

⑥ ODBC:

→ Open Database Connectivity

→ standard database access method developed by Ms Corporati

→ makes it possible to access data from any application, regar
of which database management system is handling the

→ based on CLI (Call-Level-Interface) specifications
uses SQL as its database access language

→ functionality of ODBC involves insertion of middle laye
called a database driver, between application and D

→ purpose of database driver is to translate application's
queries into commands that DBMS understands

→ To access an ODBC db, you must have the appropriate
driver for the db you wish to access

→ To access a data source, connection to the data source (provided by data) must be first be established.

⑦ ODBC Architecture:

→ Four components:

① Application:

performs processing and calls ODBC functions to
SQL statements and retrieve results

(ii) Driver Manager :→ loads and unloads driver on behalf of an application.
→ processes ODBC function calls or passes them to a driver

(iii) Driver : → processes ODBC function calls, submits SQL requests to a specific data source and returns results to the applications.

(iv) Data source : → Consists of the data user wants to access and its associated OS, DBMS, and network platform (if any) used to access the DBMS.

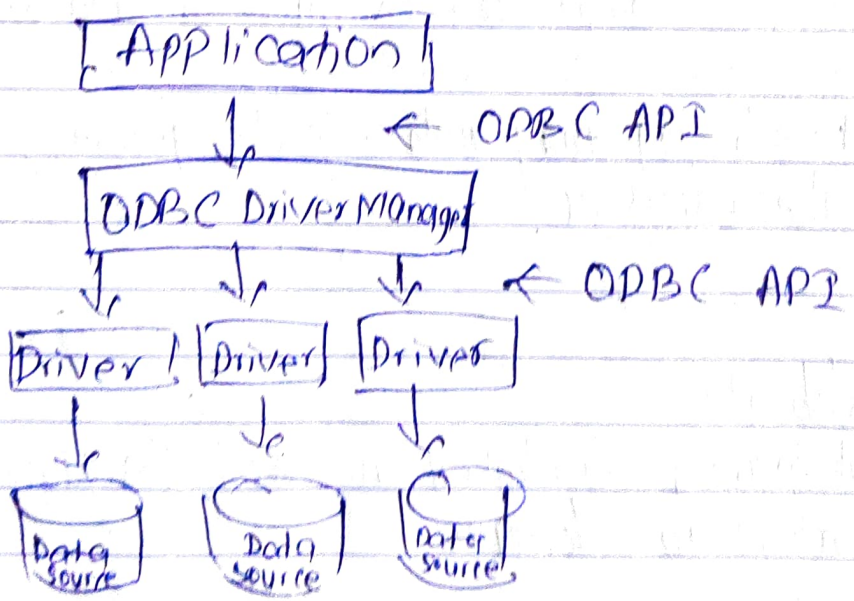


Fig: ODBC Architecture

Date:
Page:

① Web Pages:
→ Single document written using HTML
↓
(a text that contains hyperlinks that lead to other documents, usually texts)

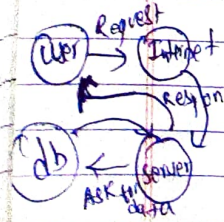
(is a text, coord, button, icon, which when clicked or hover on it, you will be directed to another web page)

→ We access a web page by entering its URL address using a web browser.
A typical URL looks like:
http://ww.example.com/index.html

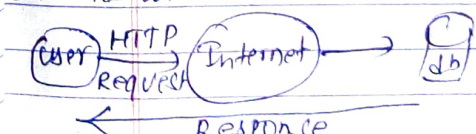
→ A web page may contain texts, graphics, hyperlinks and many other resources

② Web sites:
→ A group of interlinked and well-structured web pages that exist on the same domain

→ Two types: [→ Static websites → Static web pages (A document that, every time is requested, is shown on the browser exactly as it is stored on server)
→ Dynamic websites → Dynamic webpages (Not changed until changed manually)
(URL ust end with a document extension like .html, .php)



(Generated using scripting language and programs interacting with db on the server side)



③ Web Applications:
→ Software programs that exist on the server and runs using a web browser, through a web page

→ Created using a combination of web programming languages and web application frameworks

→ Uses RAM, allows user interactivity and is designed for many uses

→ Google Mail, fb, Youtube

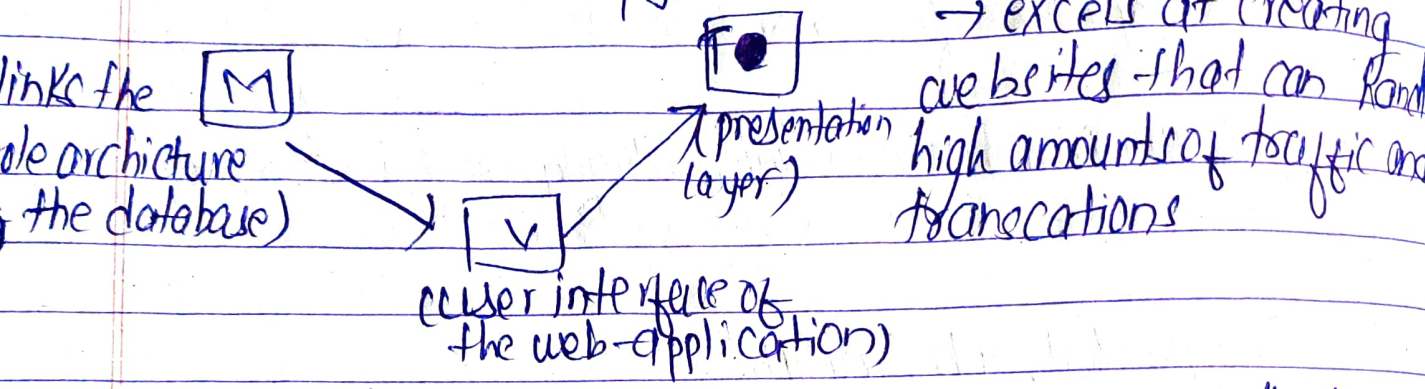
→ Web Applications	Dynamic Websites
(I) based on user Engagement	(I) Allow the users to interact and add content
(II) Almost all content from users	(II) Creator content is dominant
(III) similar to desktop apps	(III) less complicated and [redacted] skills needed are less
(IV) more complicated and more skills than websites do	

Date:
Page:

⓪ Web Application development frameworks:

① Django: (MVT)

→ high speed MVC, framework for building web apps based on python



→ excels at creating websites that can handle high amounts of traffic and transactions

→ A web framework is a software framework that is created to support the development of dynamic sites, web services and web applications.

- Easy to Maintain
- Scalability
- High Compatibility
- Easy to learn

⓪ features:

(I) Vast Community
(II) Open Source
(III) Enhance Security
(IV) Rapid Development
(V) Versatile

⓪ Companies:

(I) Instagram
(II) pinterest
(III) firefox
(IV) Netflix

Date:
Page:

⑥ Ruby on Rails:

→ Ruby:
- → Popular programming languages
- → Acts as the fundamentals for working with framework like Ruby on Rails.
- → Emphasizes on ensuring the language is expressive and easy to understand
- → Free format language
- → Case-sensitive
- → dynamic programming language

→ Ruby on Rails is an MVC-based framework used for web development and app programming at the server-side of the applications

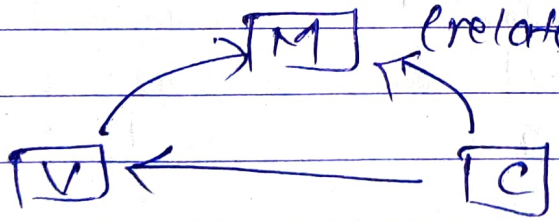


Fig: MVC architecture (separates application layer into M, V, C)

→ Ruby on Rails developers prefer it for its time-saving qualities while writing code.

→ offers ready-made code libraries for tasks such as building

Date:
Page:

forms, menus, and tables

⊙ Benefits :
(i) Cost effective
(ii) Speed
(iii) Easy change management
(iv) Secure
(v) Efficiency
(vi) Larger Developer Community

⊙ Responsibilities :
(i) Web App Development
(ii) Data Integration
(iii) Server side functionalit
(iv) Creation on Backend o
(v) App Connection with Service
(vi) Offering solutions

→ Client side Technology:

Date:
Page:

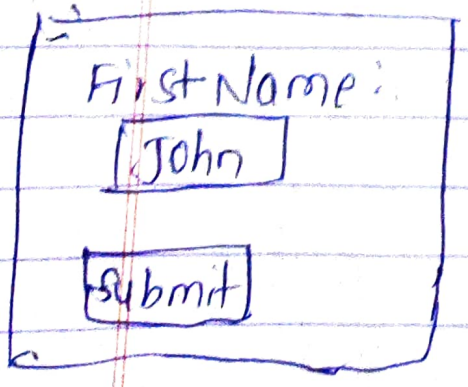
① HTML (5): (A standardized system for tagging text files to achieve font, colour, graphic, and hyperlink effects on World Wide Web pages)

```
<html>
<head> <title> ..... </title>
<script>
— — — — — —
</script>
</head>
<body>
— — — — — —
</body>
</html>
```

(because it is the set of markup symbols or codes inserted into a file intended for display on the internet)

② HTML Forms:

```
<form action="/action_page.php">
  <label for="fname">First Name:</label><br>
  <input type="text" id="fname" name="fname" value="John"><br>
  <input type="submit" value="Submit">
</form>
```



→ Forms attribute:

(i) action

(ii) target = "_blank", "_self", "_parent", "_top", "framedname"

(iii) method = "get", "post", "put", "delete"

(iv) autocomplete = "on" (v) novalidate

⑤ <select id="cars" name="cars">
<option value="volvo">Volvo</option>
</select>

→ (can use size attribute to control no. of dropdown displayed)

→ (multiple: for selecting multiple options)

⑥ <textarea name="message" rows="10" cols="30">
</textarea>

* ① button:

<button type="button" onclick="alert('Hello World!')">
Click Me!</button>

② <input type="radio" id="html" name="fav_language" value="HTML">

↳ "checkbox"

○ HTML

③ <input type="date" id="birthday" name="birthday">

④ <input type="image" src="img_submit.gif" alt="Submit" width="48" height="48">

⑤ <input type="file" id="myfile" name="myfile">

Date:
Page:

⑧ <input type="range" id="vol" name="vol" min="0" max="5">

⑨ About links:

① Absolute URLs: W3C

⑪ Relative URLs: CSS Tutorials

⑪ Img: <img src="Smiley.gif" alt="HTML tutorial" style="width:42px; height:42px;">

⑫ button: <button onclick="document.location='default.asp'">HTML tutorial</button>

⑩ Tables:

```
<table>
  <tr>
    <th>...</th>
    <th>...</th>
  </tr>
  <tr>
    <td>...</td>
    <td>...</td>
  </tr>
</table>
```

⑬ Lists:

```
<ul>
<li>...</li>
</ul>
```

```
<ol>
<li>...</li>
</ol>
```

Advantages
① Reduces file transfer size
② easy to customize online page
③ Need few lines to improve site speed

Disadvantages
→ multiple levels which creates confusion
→ cross-browsers issues
→ scarcity of security
→ confusion for web browsers

④ device friendly ⑤ less complex

⑥ CSS: → used to format the layout of a webpage
→ can be added three ways:
(i) Inline: style
(ii) Internal: <style>
(iii) External: using <link>

⑦ JavaScript:
→ a light-weight, interpreted, or just-in-time compiled programming language with first-class functions. Both server side and client side language

→ for External JS:
<script src="myscript.js"></script>

Eg:
```
<script>
    function myFunction()
    {
      document.getElementById("demo").innerHTML="Paragraph Changed";
    }
</script>

<body>
 <p id="demo">A Paragraph</p>
 <button type="button" onClick="myFunction()">Try it</button>
</body>
```

→ let does not allow multiple initializations

Date:
Page:

- window.alert( )
- document.Write( )
- console.log( )
- window.print( )
- (x == y)
- Const cars = ["saab", "Volvo", "Bmw"]
- Const object = { firstName: "John" age: 50 };
- let x = myfunction(4, 3);

  function myfunction(a, b) { return a*b };
- for(; ;) { Do sth; }
- Objectname.firstName or Objectname["firstName"]
- if...else

  e.g.

  if(hour < 18) { greeting = "Good Day" };
- switch (value) { case x: ... break; case Y: ... break; default:
- Events:
  - (I) onChange
  - (II) onclick
  - (III) onmouseover
  - (IV) onmouseout
  - (V) onkeydown
  - (VI) onload

Date:
Page:

⑥ Class:

```
class Car
{
  constructor(name, year)
  {
    this.name = name;
    this.year = year;
  }
};
let mycar = new Car("Ford", 2014);
```

⑦ JSON:

→ JavaScript Object Notation

→ lightweight data-Interchange format

→ Inside curly brackets,

⑧ XML schemas:

① describes the structure of XML document

② XML schema definition (XSD)

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="site link">
<xs:element name="note">
  ---
</xs:element>
</xs:schema>
```

→ XML based and more powerful than DTD

→ hundreds of XML formats are in daily use

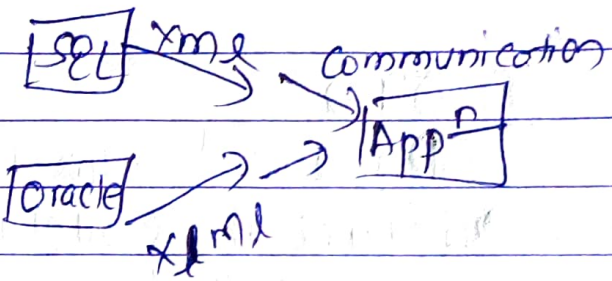
Date:
Page:

⑥ XML:

→ eXtensible Markup Language
→ used to store and transport data
→ Self-descriptive
→ used to carry data
→ self-defined tags
→ platform and language independent
→ helps in easy communication b/w two platforms

→ Features:

(i) Separates data from HTML
(ii) Simplifies data sharing
(iii) " " transport
(iv) Increases data availability
(v) Simplifies platform change



→ XML example:
↳ Hierarchial structure

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<college> ← Root element is must
    <class1>
        <Name>Amit</Name>
    </class1>
</college>
```

Child element

XML schemas
→ Specify restrictions on data
→ Keeps much better control over types of data
→ —

DTD
→ —
→ —
→ doesn't allow creating customized data

HTML
(i) Display data
(ii) Markup language
(iii) Not case sensitive
(iv) predefined tags
(v) static

XML
(i) Transport and store data
(ii) provide framework to define markup language
(iii) Case sensitive
(iv) Can create own tags
(v) Dynamic

(6) DTD:
⇒ Document Type Definition
⇒ defines the structure and legal elements and attributes of an XML document

⇒ With a DTD, independent groups of people can agree on a standard DTD for interchanging data

⇒ An application can use a DTD to verify that XML is valid.

⇒ Can be declared inside XML file or Outside

```
<?xml version "1.0"?>
<!DOCTYPE note]
<!Element note(to,from,heading,body)>
]>
    ↓
<!Element to (#PCDATA)>
```

```
<note>
<to>Tove</to>
</note>
```

Date:
Page:

⓪ TF-IDF:

→ stands for Term Frequency-Inverse Document frequency

→ two matrices that are closely related and search and figure out the relevancy of a given word to a document given a larger body of document.

→ All TF means is how often a given word occurs in a given document so within one web page one wikipedia article, how common is a given word within that document, what is the ratio of that word occurence rate throughout all the words in that document that's it.

→ DF means how often a word occurs in an entire set of documents i.e all of wikipedia or every web page.

① IDF:

→ Inverse Document frequency

→ We use log of IDF since words frequencies are distributed exponentially.

TF-IDF = TF * DF

→ how often the word appears in a document over how often it just appears everywhere.

→ measure of how important and unique this word is for this document.

Date:
Page:

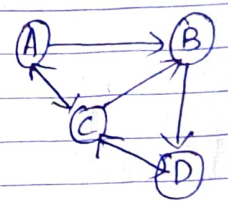
→Sp
reduc
on

⑥ Builtin functions in JS:

① eval:
② parseInt:
③ parseFloat
④ Array
⑤ isNan
⑥ String(object)
⑦ isFinite
⑧ N...

r( )

⑥ PageRank Algorithm:



	Iteration 0	Iteration 1	Iteration 2	PR
A	1/4	1/12	1.5/12	1
B	1/4	2.5/12	2/12	2
C	1/4	4.5/12	4.5/12	4
D	1/4	4/12	4/12	3

→ Formula:

$$PR_{t+1}(P_i) = \sum_{P_j} \frac{PR_t(P_j)}{C(P_j)}$$

$$PR(A) = \frac{1/4}{3} = \frac{1}{12}$$

Go on similarly

$$PR(B) = \frac{\frac{1}{4}}{3} + \frac{1/4}{2}$$

$$PR(C) = \frac{1/4}{2} + \frac{1/4}{1}$$

Hub: A node is high-quality if it links to many high-quality nodes

Authority: A node is high-quality if many high-quality nodes link to it

④ New hub is the sum of authority of its children

⑤ Normalize new authority and hub

⑥ HITS:

→ An algorithm used in link analysis.

→ It could discover and rank web pages relevant for a particular search.

→ Algorithm:

① Initialize the hub and authority of each node with a value 1

② For each iteration, update the hub and authority of every node in the graph

③ New authority is the sum of hub of its parents.