* Introduction

→ An IRS stores, retrieves and maintainence information.
→ this information can exist in multiple forms such as :- text format
- audio
- video
- images
- multimedia objects.

→ An IRS refers to a system which is capable of representing, organizing, storing & accessing information.
→ It is a proper way of representing & organizing information.
→ And it provides an easy access to it users.
→ the main objective of an IRS is to reduce user's overhead while searching for desired information.
→ Overhead is the difference b/w :
— time required to obtain information
— & time required to read the extracted information

*overhead = (time required to obtain info − Actual reading time required to read data).

→ An IRS is a S/w.
→ It consists of all features & functions needed to manipulate information.
→ Retrieval system is a system used to storing & extracting information.
→ IRS is a S/w program, that helps to users in obtaining desired information.

* Objectives of IRS

→ The main objective of an IRS is to reduce user's overhead while searching for desired information.

→ Overhead means the difference b/w the time required to obtain the information & time required to read the data.

Overhead = { time required to obtain the desired info — time required to read the data }

→ To provide right information to the users at the right time.

→ To retrieve or provide relevant information to users on time.

→ Two measures of IRS are: 1. Precision
2. Recall

1. Precision

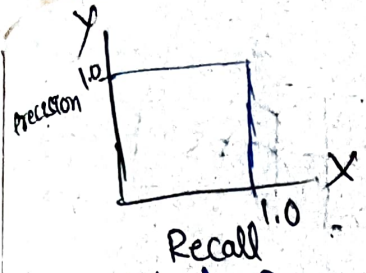
$$\text{Precision} = \frac{\text{No. of relevant items retrieved}}{\text{Total no. of retrieved items.}}$$

2. Recall

$$\text{Recall} = \frac{\text{No. of relevant items retrieved}}{\text{Possible no. of relevant items.}}$$

↓

all relevant items available in DB



Ideal Precision/Recall Graph.

Goal: Maximize recall to ~~ensure~~ get most relevant items

Ex: If a search engine has high recall of 80%
- it means, it retrieves 80% of relevant items
- but misses 20% items.

High Precision: Retrieve only relevant documents/items.

High Recall: Retrieve all relevant documents/items

→ If a search engine has recall of 100%
- It means it retrieves all relevant document, missing none.

* Functional Overview

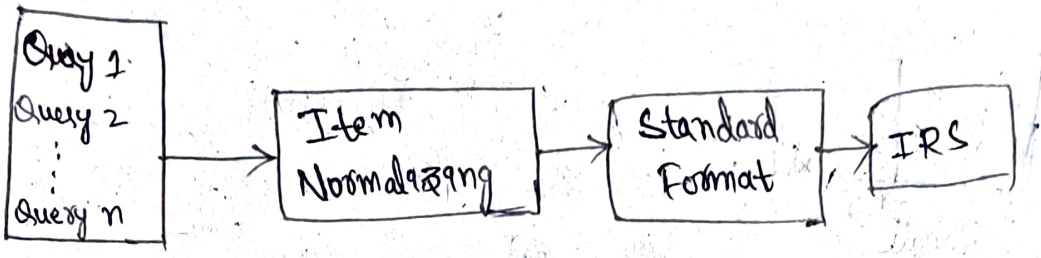
→ IRS consists of following four major functional processes.

1. Item Normalization.
2. SDI (Selective Dissemination of Information)
3. Document database Search
4. Index database Search

1. Item Normalization

→ It is a process of decomposing the incoming item into a standard form.

→ It normalizes the incoming data in to a standandard format.



→ To normalize the data into standard form, it includes :

- Token identification
- Token characterization
- Apply stop lists
- Apply Stemming Algorithms
- ~~Zoning~~ Zoning

• Apply stop lists

→ It removes the common words like 'the', 'is', 'what'

Ex: what is the current cricket match score now

After applying stop lists: Cricket score.

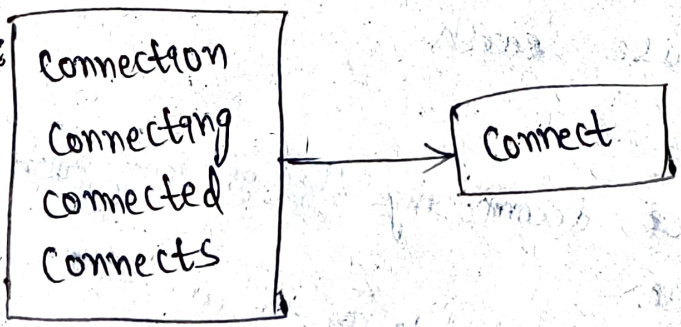
• Applying Stemming Algorithms

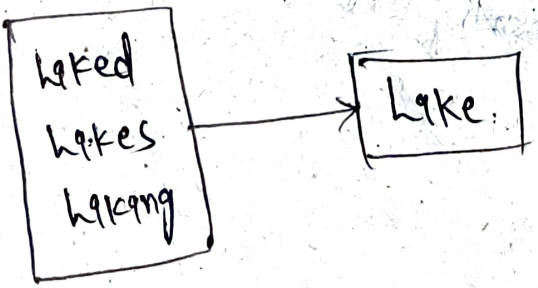
→ It reduces words to their base form.

→ It improve recall & reduce index size

→ Enhance query matching.

Ex:





- Zoning

→ It Parse the data into logical subparts.
→ This Parsing is user-transparent.
→ It is done, to improve the precision.
→ the term zone considered as field.

Example: (Normalization)

Name	age	DOB	address
John	20	02/05/1992	123, Main street, 2nd town, USA
Smith	30	05/02/1990	321, 2nd street, town, USA

⇒ It is converted into:

Name	age	DD/MM/YYYY	(street, city, state, Pin code)
John	20	02/05/1992	123, Main street, town, USA

2. Selective Dissemination of Information

→ It is commonly referred as SDI.
→ SDI is a system that sends relevant information to users based on their interests.
→ It is a process of categorizing data
→ SDI is also known as Mail Process
→ It is a process where users receive new updates based on their interests or profiles.

Ex:1. A farmer subscribes to a weather update service.
- And get notification about rainfall or temperature changes.

2. A business person subscribes to stock market alerts
- So, he gets all updates about their industries

3. A teacher subscribes to an education portal,
- So, they gets all new updates about teaching of their subject.

→ SDI system push information to user based on their profile interests.

Pull System: A System where users pull information by searching.

Push System: A system where information is pushed to users based on their interests.

Difference:
→ Pull System requires users to search actively.
→ Push System that sends information to users based on their interests.

3. Document database Search Process:

User enters a query:

→ A user submits a search query to the system

Search Process:

→ The system processes the query & search in the document database

↓ it contains only documents

Document database: the system searches all items in the document database which contains all received, processed & stored items.

Ex: User enters a query

- A user searches for "AI applications"
- the system process the query and searches the document database for documents containing the known word "AI application"
- the system searches all documents in ~~the~~ DB document DB.

ex: If it contains 10,000 documents including research papers, articles & reports

Search Result

→ the system return a list of 20 documents that match the search query.

4. Index & Database Search

→ When you find something interesting you want to save it for later on and information system
– this is called Indexing

→ there are two classes of index files:
1. Public Index file
2. Private Index file

1. Public Index File:

– Maintained by library professionals.
– Index every item in the Document DB
– Smaller no. of files
– Accessable by any one with proper privilages
↓
Permissions

Ex: A public index file called "All documents" that contains entries for every documents in the DB

2. Private Index File

→ Created by individual users
→ Each user can have multiple private index file.
→ Limited access (only owners & specified users can access)

* Relationship to DBMS

DBMS	IRS
1). It is an organizational structure and handles structured data.	1). It is a system which stores, organize, retrieve info.
2). It provides Precise Semantics.	2). It provides imprecise semantics.
3). It provide Data Modeling Facility.	3). It does not provide DMF.
4). query language is artificial.	4). query language is almost similar to natural language.
5). query specification is complete.	5). Query specification is incomplete.
6). In DBMS, there exists structured data format.	6). In IRS, there exists unstructured data format.
7). Structured data	7). Unstructured data
8). Query matching is exact match	8). Query matching is partial match.
9). Inference is deduction.	9). Inference is Induction.
10). Uses deterministic model.	10). Uses Probabilistic model.
11). We get exact results.	11). Sometimes relevant, often not.
12). Knowledge Representation is Transparent.	12). Knowledge Representation is complex process.

* Digital Libraries

→ Digital Library acts as a repository.
→ It stores data.
→ It is a system, used to extract relevant information
→ At first, it is known as Electronic Libraries
→ BCZ, they utilize electronic devices to perform functions.
→ Later US government renamed electronic libraries as Digital Libraries.
→ As Internet technology is growing exponentially, the digital libraries conception is also emerging.
→ It is responsible for extracting relevant information.

* Data Warehouses

→ A Data warehouse is a special type of database.
→ And it is used by the organizations for storing large amount of business data.
→ It is a centralized repository.
→ It stores data from various sources.
→ It is designed to handle large volume of data.
→ It makes easier to access, analyze data.
→ It support decision-making & analytics.

Example:

1. A school can stores all data about student.
- Such as grades, attendance, test results
- from different years on one place.
- We can see student's performance has improved or not.
- So, based on this information, we can make decisions.

2. A hospital stores Patient records, treatment history, appointment data.

3. Banks stores transaction data.

4. Library.

5. Airlines: Stores flight data, schedules, & customer experience.

* Browse Capabilities

→ After the completion of the search, browse capabilities allows user to identify and display the required items.

→ It allows users to explore data from categories or topics.

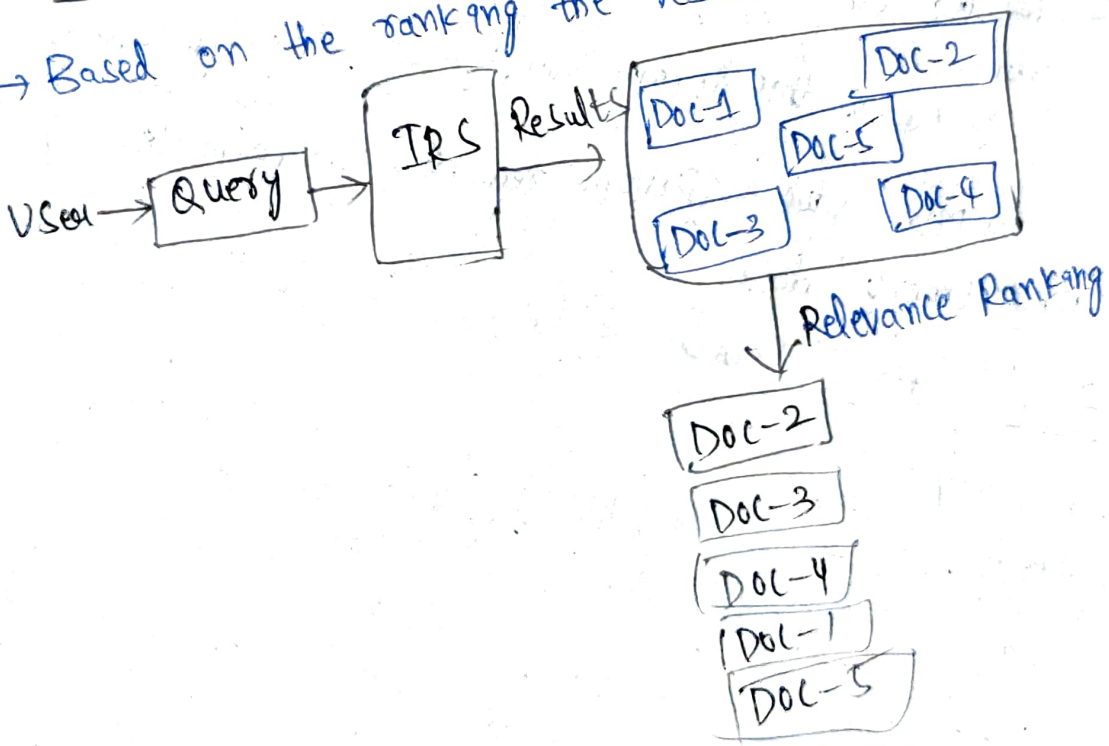
→ Data Summary can be displayed ~~into~~ in two ways:

1. Line the status of items
2. Data Visualization.

→ Browse capabilities includes: - Relevance Ranking
- Zooming
- Highlighting.

1. Relevance Ranking:

→ Based on the ranking the results will displayed.

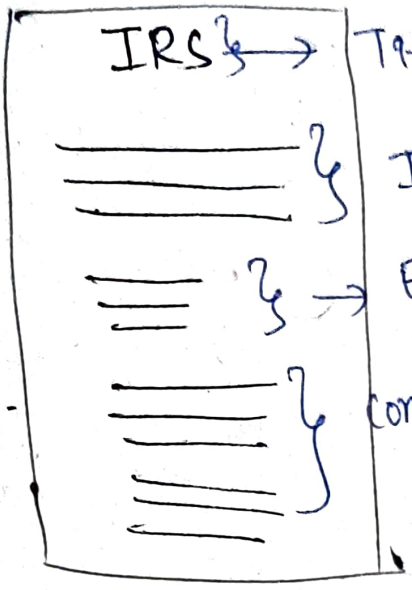


→ Users can find relevant information faster.

2. Zoning

→ It plays a significant role in browsing.

Ex:



→ Zoning enriches browsing capabilities..

→ Ex: when user looking for main concept in book.

- user can browse through conclusion.

3. Highlighting

→ It makes easier for users to identify relevant information quickly.

Ex: If a user search for IRS topic on google.

→ In results, main points are highlighted.

Give a brief note on [redacted] capabilities.

[redacted] Capabilities

search capabilities

→ It creates mapping between user's need and items in the database.

User's need (Source) —mapping→ items in Database (result) (Destination)

→ User's need (or) Search item consists of natural language and boolean logic.

Ex: AI and IRS

→ Every computer contains some algorithms.

→ Different algorithms contains different functions.

→ Algorithms are used for processing the Search Statement.

→ The functions define the relationship between the terms in the Search Statement.

Functions

→ Functions read Search items and creates mapping.

- Boolean Logic
- Proximity
- Contiguous word Phrases
- Fuzzy Searches
- Term Masking
- Numeric and date ranges
- Natural Language Queries

1. Boolean Logic

→ It is used in between query terms.

→ It is a way to combine multiple Search terms using logical operators (AND, OR, NOT)

Boolean Operators:

AND (Intersection)

→ It returns the document containing both terms.

Ex: AI and IRS

OR (Union)

→ It returns document either term

Ex: Datascience OR Machine learning

NOT (Difference)

→ excludes document containing term

Ex: AI NOT IRS

2. Natural language Query

→ It allows to specify the importance of each search term. using a value b/w 0.0 and 1.0

Ex: AI (0.8) and IRS (0.6) or DS(0.4)

→ It means AI is most important

- IRS is moderate in important
- DS is less important

3. Proximity

→ It specifies distance b/w two items in search terms.

→ It is used to increase the precision of the search

→ The typical form of Proximity is TERM 1 within n units of TERM 2

Ex: Data within 5 words of machine

8.4. Fuzzy Search

→ It search for similar words.

Ex: Fuzzy search on "Computer"

- computer (exact match)
- compter (minor spelling error)
- computor (common misspelling)
- computre (another common misspelling)

5. Contiguous word Phrases:

→ It is a Sequence of two or more words that are treated as a single unit.

Ex: United States of America

→ It represents a single concept and can be used in Search queries.

→ And combine a Contiguous word Phrase with Search Operators like "AND", "OR" etc....

6. Term Masking:

→ In the context of term masking, a wild card character is a Special Symbol.

→ It is used to represent one or more characters in a Search query.

→ the most common wild card characters are:

* - for multiple characters

? - for single character

$ - for a Special character

Ex:- "*Puter" → computer

- C?mputer → Computer

- multi$national → multi-national