Experiment No. IR

Date :

Assignment No. - 01

Q.1 Brief overview of Information Retrieval Process.

→ Ans :-

- Information retrieval refers to the process, methods, & procedures of searching, locating & retrieving recorded data & information from a file or database.
- An Information Retrieval system is supported by basic processes such as -

1. Query Handling
2. Indexing &
3. Matching

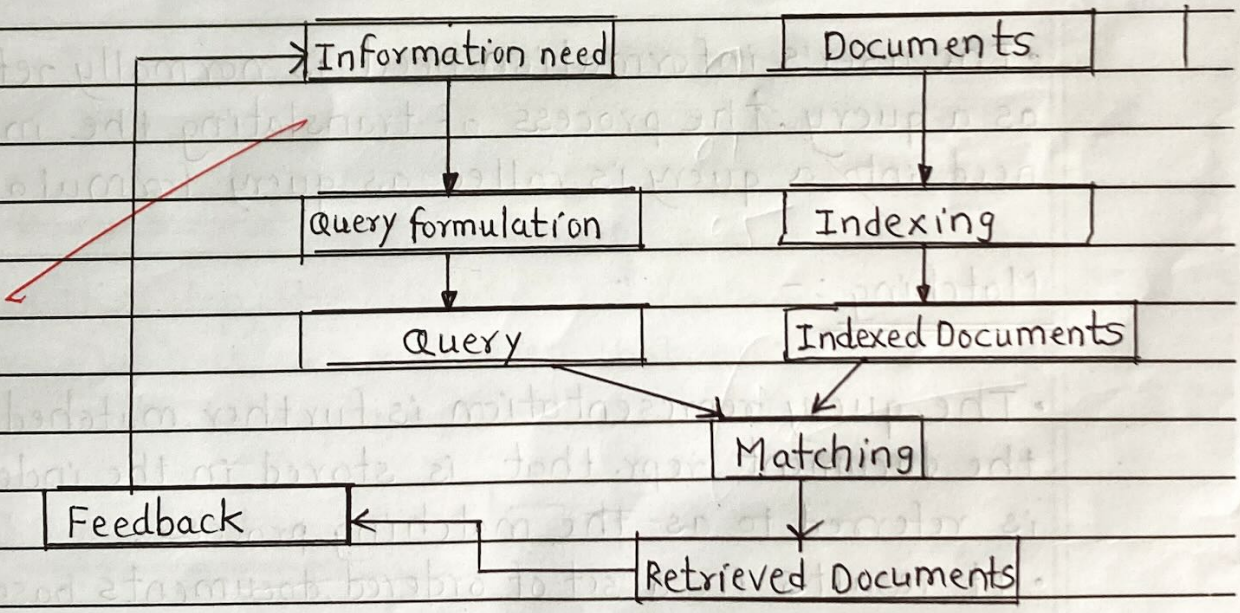


fig: Information Retrieval (IR) Process.

*Indexing :-

- There are three basic processes an information retrieval system has to support, the representation of the documents, the representation of user's information need and the comparison of the two representations.
- Representing the documents is usually called the indexing process.
- The indexing process results in a formal representation of the document.
- The indexing process may include the actual storage of the document in the system but often documents are stored partly.

Query formulation :-

- The user's information need is normally referred to as a query. The process of translating the information need into a query is called as query formulation.

Matching :-

- The query representation is further matched with the document repr. that is stored in the index file & is referred to as the matching process.
- It results in a set of ordered documents based on the relevance & is refered to as the ranked list.

Page No:

Q.2 What are the components of Information Retrieval ? Explain with a diagram.

→

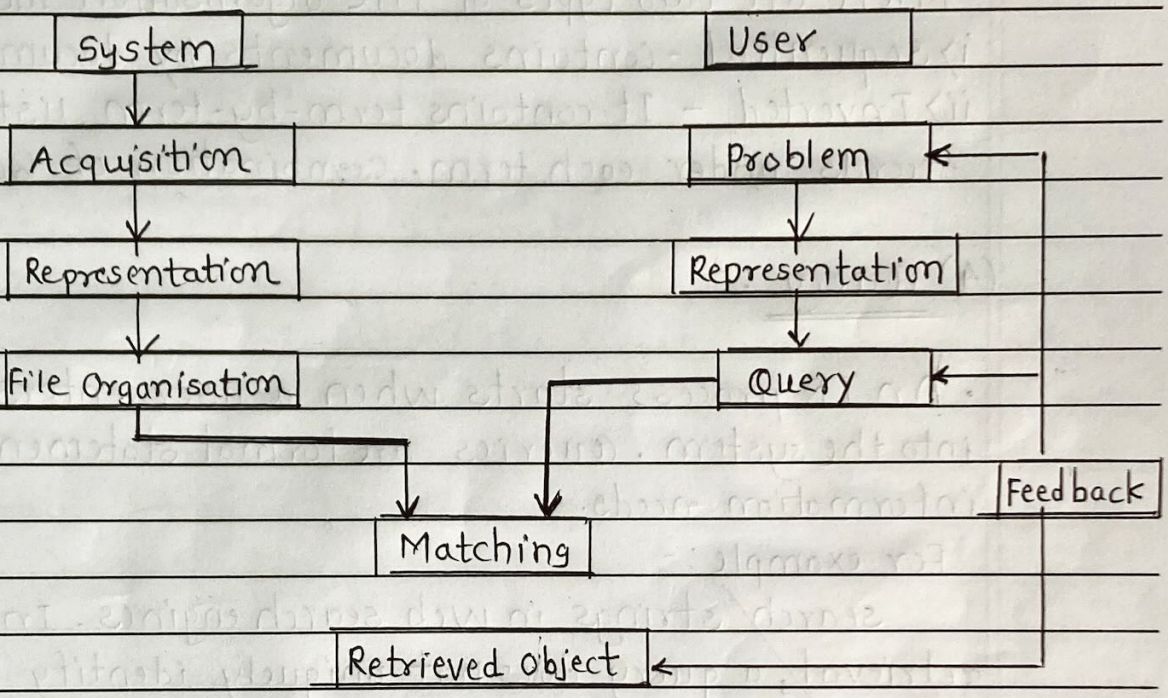


fig : Information Retrieval Components

(1) Acquisition :-

- In this step, the selection of documents & other objects from various web resources that consists of text-based documents takes place.

(2) Representation :-

- It contains indexing that contains free-text terms controlled vocabulary, manual & automatic techniques as well.

(3) File Organisation :-

- There are two types of file organisation methods
  i) sequential - contains documents by document data.
  ii) Inverted - It contains term-by-term, list of records under each term. combination of both.

(4) Query :-

- An IR process starts when a user enters a query into the system. Queries are formal statements of information needs.

For example :-

search strings in web search engines. In information retrieval, a query does not uniquely identify a single object in a collection. Instead several objects may match the query, perhaps with different degrees of relevancy.

Q.3 What is the need of information Retrieval? Discuss.

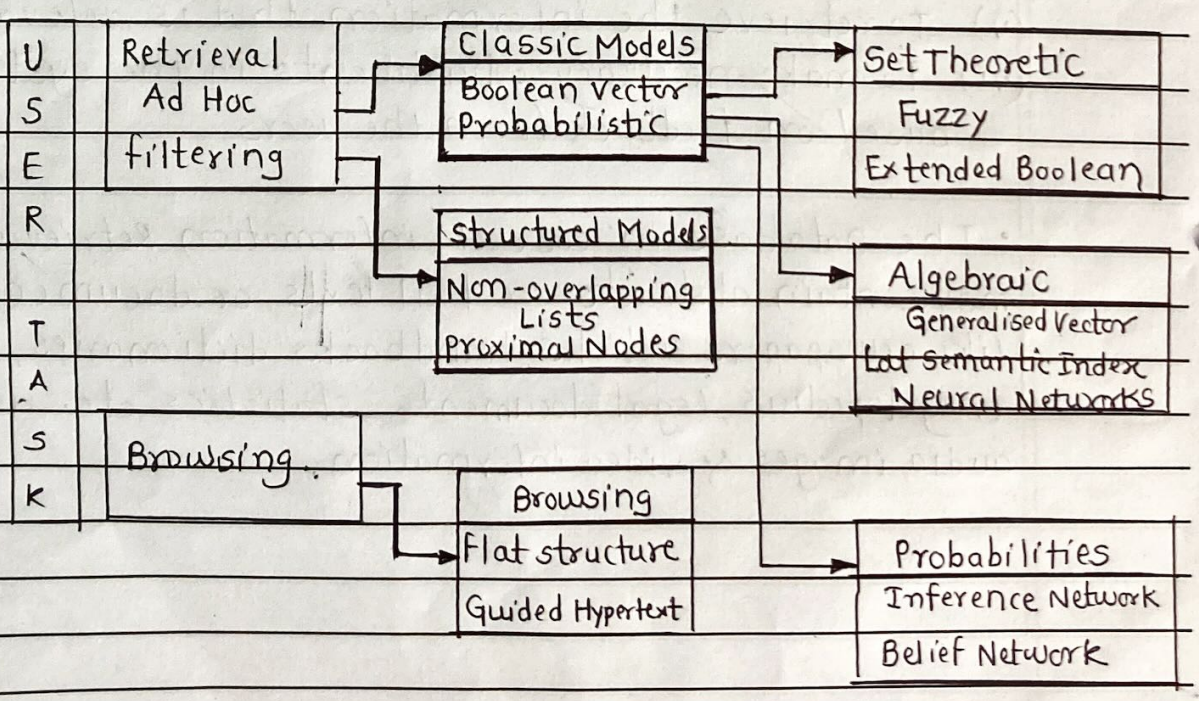
→ • The major objective of an IRS is to retrieve the required documents information whenever needed.
• The search output may contain bibliographic details of the documents that matches the query, or the actual text, image, video etc. that contain the required information.
• The major need of information Retrieval system are :-
(i) To identify the sources of information relevant to the areas of interest of the target users community.
(ii) To analyze the contents of the sources (documents)
(iii) To represent the contents of the analyzed sources for matching with the user's queries.
(iv) To match the search statements with the stored database.
(v) To retrieve the information that is relevant.
(vi) To make necessary adjustments in the system based on feedback from the users.

• The Database in case of information Retrieval system may contain abstracts or full texts or documents like newspapers articles, handbooks dictionaries, encyclopedias, legal documents, statistics etc. as well as audio, images & video information.

Q.4 Explain the taxonomy of Information retrieval with neat diagrams.

→ • An Information Retrieval model is defined as a quadruple $[D, Q, F, Rel(q_i, d_j)]$ where,

D - represents group of documents found in a collection
Q - Represents group of information needs which is referred to as queries.
F - Framework that consists of documents, queries with their relationships with those documents.
$Rel(q_i, d_j)$ - Represents the score which is associated with the query q_i & document d_j.



- The documents and basic set operations are available in the framework for the traditional Boolean model, the algebra operations on vectors are available in the framework for the vector model & the Bayes Theorem and probability operations are available in the framework for the probabilistic model.

- Alternative set-theoretic models include the extended boolean model & the fuzzy model;

- Alternative algebraic model include the neural network model, latent semantic indexing model & generalised vector models.

&

- Alternative probabilistic models include the inference network & the belief network.

Q.5 Explain the logical view of document during text pre-processing.

→ • Due to Historical reasons, documents in a collection are frequently represented through a set of index terms or keywords.

• No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a logical view of the document.

• With very large collections, however even modern computers have to reduce the set of representative keywords. This can be accomplished through the elimination of stopwords, the use of stemming & the identification of noun groups.

• Text operations reduce the complexity of the document representation & allow moving the logical view from that of a full text to that of a set of index terms.

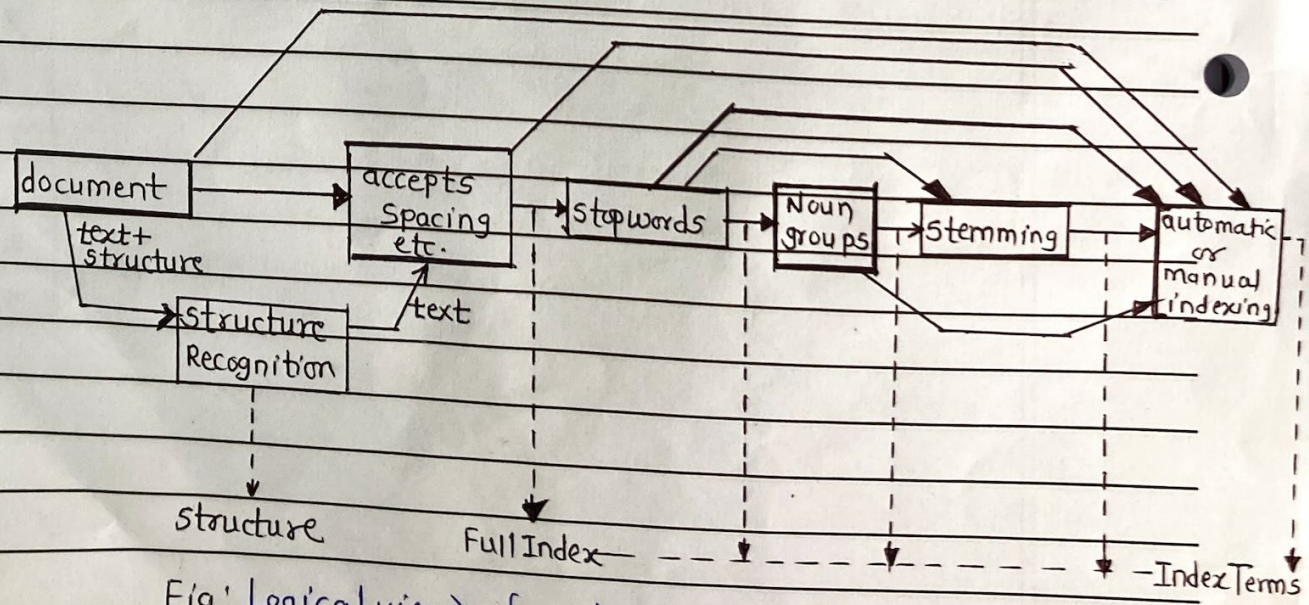


Fig: Logical view of a document: from full text to a set of Index terms.

- The full text is clearly the most complete logical view of a document but its usage usually implies higher computational costs.
- small set of catagories provides the most concise logical view of a document but its usage might lead to retrieval of poor quality.

Q.8 What is the tf-idf weight of term 't' in document d?

→ Term frequency Weights (tf)

Definition :- Luhn Assumption.

The value, or weight of a term K_i, that occurs in a document d_j is simply proportional to the term frequency $f_{i,j}$. Which means, the more often the term K_i occurs in the text of the document d_j, the higher its term frequency weight $TF_{i,j}$ is.

$$tf_{i,j} = \begin{cases} 1 + \log f_{i,j} & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Inverse Document frequency Weights :-

- The inverse document frequency is a measure of how much information the word provides. i.e. if it is common or rare across all documents.
- It is the logarithmically scaled inverse fraction

of the documents that contain the word -

- It is obtained by dividing the total Number of documents by the number of documents containing the term & then taking the logarithm of that quotient.

$$idf(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with

- N : total number of documents in the corpus $N = |D|$
- $|\{d \in D : t \in d\}|$ - Number of documents where the term t appears (i.e. $tf(t,d) \neq 0$)

tf - idf weighting Schems.

$$TF(t,d) = \log(1 + f_{t,d})$$

$$IDF(t,d) = (1 + \log f_{t,d}) \cdot \log \frac{N}{n_t}$$

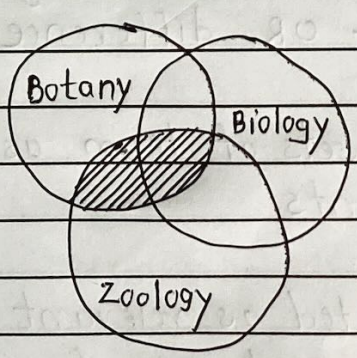
$$IDF_t = \log \frac{N}{n_t}$$

Page No:

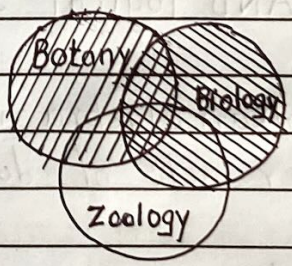
Q.6 How is a document & Query represented using the Boolean model? How is the relevance of a document to a user query defined? Outline with example.

→

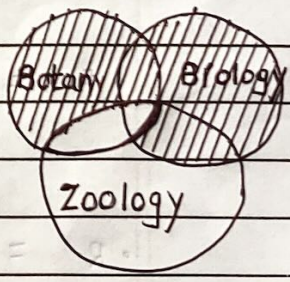
1) It is an exact matching model, which means it either retrieves or does not retrieve documents without ranking them.
2) The model allows for the use of structured queries which include not only query terms but also relationships between the terms defined by the query operator AND, OR & NOT.
3) For example
The information need "Botany AND Zoology" will return a set of documents containing both words "whereas query with the keyword "Botany AND Zoology will return a set of documents containing either the words "Botany" or Zoology.



Botany AND Zoology



Botany ~~AND~~ OR Zoology



(Botany OR Biology) AND NOT (Botany AND Zoology)

4) In case of a query "Botany AND Zoology AND Biology" It will not return a document that contains the terms "family" friends or parents but it will also return a document that contains "Botany" & "Zoology" but not "Biology".

5) However this model has significant limitations such as the inability to provide a ranking based on relevance when retrieving multiple document.

6) The model is based on set theory & the Boolean algebra, where documents are sets of terms & queries are boolean expression on terms.

The Boolean model can be defined as -

- D - A set of word, i.e. the indexing terms present in a document. Here, each term is present (1) or absent (0).
- Q = A boolean expression where terms are the index terms and operator and logical product AND, logical sum - OR, difference NOT
- F = Boolean algebra over sets of term as well as over sets of documents
- R = A document is predicted as relevant to the query expression if and only if it satisfied the query expression as :-

$$((\text{text} \vee_{\text{information}}) \wedge \text{retrieval} \wedge \vee \text{theory})$$

Page No:

Q.7 Explain the concept of cosine similarity with example.

→ Ans :-

1) Every document in the document space & every information need expressed as a query are represented by a vector in the term space.

2) The similarity score can be calculated by measuring the distance between the document vector & query vector, which represents how closely or distantly related the document is to the query.

3) The similarity score is normally the cosine of the angle that separates the two vectors $\vec{q}$ & $\vec{d}$. The cosine of the angle is 0 if the vectors are orthogonal in the space.

$$\text{Score}(\vec{d}, \vec{q}) = \sum_{k=1}^{m} n(d_k) \cdot n(q_k)$$

where $n(v_k) = v_k \sum_{k=1}^{m} v_k^2$

4) The representation of the angle between vector in dimensional space simplifies explanation.

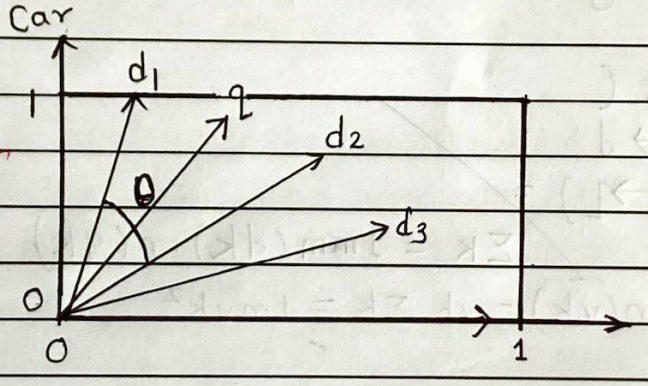
5) The similarity measure of a document vector to a query vector is usually the cosine of the angle between them.

Example :-

The query & documents are represented by two dimensional vector space. The terms are car & insurance. There is one query and three documents in the vector space.

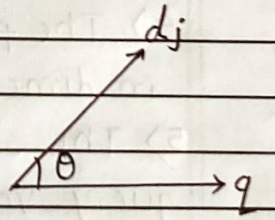
The top ranked document in response to the terms car & insurance will be the document d_2 because the angle between q & d_2 is the smallest The reason behind this is that both the concepts car & insurance are silent in d2. & hence have the high weights.

On the other side, d1 & d3 also mentioned both the terms but in each case, one of them is not a centrally important term in the document.



Degree of Similarity

$$Sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \times |\vec{q}|}$$



$$= \sum_{i=1}^{t} w_{i,j} \times w_{i,q}$$

$$= \sum_{i=1}^{t} w_{i,j}^2 \times \sum_{i=1}^{t} w_{i,q}^2$$

Page No: