→ Define Information Retrieval?

Information Retrieval (IR) deals with representation, storage and organization of unstructured data. Information retrieval is the process of searching within a document collection for a particular need (a query).

→ IR can be defined as a Software Program that deals with the organization, Storage, retrieval and evalution of information from document repositories, particularly textual information.

→ Traditional examples of information-retrieval systems are online library catalogs and online document management systems such as those that store newspaper articles.

→ The data in such systems are organized as a collection of documents. A newspaper article and a catalog entry (in a library catalog are examples of documents.

→ It uses a particular keyword (or) set of keywords in document for retrieval.

↳ user search for a keyword

→ Eg The keywords "stock" and "scandals" locate articles about stock-market scandals. The keyword "databa system" may be used to locate books on database systems.

→ Keyword based information retrieval can be used not only for retrieving textual data but also for retrieving other types of data, such as video and audio data.

For eg: Video movie may have associated with its keywords ~~associated with its~~ such as its title, director, actors and genre.

→ While an image or video clip may have tags, which are keywords describing the image (or) video clip associated with it.

Differences b/w Database System and information retrieval systems

→ Database systems deal with the updates and with the associated transactional requirements of concurrency control and durability where as in information retrieval (Scheme) systems it is viewed as less important.

→ Database systems deal with structured information organized with relatively complex data models (such as relational model or object-oriented data models). Where as in information retrieval the information is stored is unstructured documents.

Keyword Search

In fulltext retrieval, all the words in each document are considered to be keywords.

→ we use the word term to refer to the words in a document.

→ Information retrieval systems typically allow query expressions formed using keywords and the logical connectives and, or and not.

→ Ranking of documents on the basis of estimated relevance to a query is critical.

Relevance ranking is based on factors such as

Term frequency (TF):- Frequency of occurance of query keyword in document.

Inverse document frequency :- (IDF)

How many documents the query keyword occurs in. Fewer give importance to keyword.

Hyperlinks to documents

More links to a documents.

Relevance Ranking using Terms

TF-IDF (Term Frequency / Inverse Document frequency) Ranking

Let $n(d)$ = number of terms in the document

$n(d, t)$ = number of occurances of term t in the document d.

Relevance

$$TF(d, t) = \log\left(1 + \frac{n(d, t)}{n(d)}\right)$$

If two terms "database", "silberschatz" are there

The document containing "silberschatz" but not "database" should be ranked higher than document containing term "database" but not "silberschatz"

To fix this Problem, weight are assigned to term using the inverse document frequency (IDF) defined as

$$IDF(t) = \frac{1}{n(t)}$$

$n(t)$ denotes no of documents.

The relevance of a document d to a set of terms Q is then defined as

$$r(d, Q) = \sum_{t \in Q} TF(d, t) * IDF(t)$$

→ Most systems add to the above model

→ words that occur in title, author list, section headings etc, are given greater importance.

→ words whose first occurrence is late in the document are given lower importance.

→ very common words such as "a", "an", "the", "it" etc are eliminated called stopwords.

→ Proximity: If keywords in query occur close together in the document, the document

has higher importance than if they occur apart.

→ Documents are returned in decreasing order of relevance score. usually only few documents are returned, not all.

→ Similarity Based Retrieval

Similarity based retrieval retrieve documents similar to given document.

→ Similarity may be defined on the basis of common words

E.g: find k terms in A with highest $TF(d,t)/n(t)$ and use these terms to find relevance of other documents

→ The model of documents as points and vectors in an n-dimensional space is called the vector space model.

→ The resultant set of documents is likely to be what the user intended to find. This idea is called

relevance feedback.

→ Relevance feedback can also be used to help users find relevant documents from a large set of document matching the given query keywords.

→ The cosine of the angle between the vectors of two documents is used as a measure of their similarity.

Relevance using Hyperlinks :- Popularity Ranking

The basic idea of Popularity ranking also called Prestige ranking is to find Pages that are popular, and to rank them higher than other pages that contain the specified keywords.

Eg: The term "google" may occur in vast numbers of Pages, but the Page gogle.com is most popular among the Pages that contain the term "google". The Page google.com should therefore be ranked as the most relevant answer to a query consisting of the term "google".

use number of hyperlinks to a site as a measure of the Popularity or Prestige of the site.

Hub and authority based ranking

→ A Hub is a Page that stores links to many Pages (on a topic)

→ An authority is a Page that contains actual information on a topic.

→ Each Page gets a hub Prestige based on Prestige of authories that it Points to.

→ Each Page gets an authority Prestige based on Prestige of hubs that Points to.

→ Page Rank :- The web search engine Google introduced PageRank, which is a measure of Popularity of a Page based on Popularity of Pages that link to the Page.

One drawback of PageRank algorithm is that it assigns a measure of Popularity that does not take query keywords into account.

→ Search Engine Spamming refers to the practice of creating webpages or sets of web pages designed to get a high relevance rank for some queries, even though the sites are not actually popular site.

Eg: A Travel site may want to be ranked high for queries with the keyword "travel". It can get high TF-IDF scores by repeating the word "travel" many times in its page.

→ Combining TF-IDF and Popularity Ranking Measures. → By using Machine learning Techniques.

Synonyms, Homonyms and ontologies

"Motorcyle Maintenance" replaced by "Motorcycle refair" → words with same Meaning

Homonyms :- object with different meanings

Eg:- Table → single word with multiple meaning as a noun and as a verb

dinnertable (or) table in a database

Concept based querying :- What concept each word in a document represents & similarly to understand what concepts a user looking for, and to return documents that address the concepts that user is interested in.

Advantages :- A query in one language can retrieve documents in other languages, so long as they relate to the same concept.

→ Automated translated mechanisms can be used. if user does not understand the language in which the document is written.

ontologies :- ontologies are hierarchical structures that reflects relationships between concepts.

→ The most common relation is, the is-a relationship.

Eg :- A leopard is-a mammal and a mammal is-a animal.

→ other relationships such as Part-of are also possible.

Eg : An airplane wing is part-of an airplane

→ There are two basic measures for assessing the quality of text retrieval

2 Performance Metrics

Precision – This is the percentage of retrieved data that are actually relevant to the query (i.e, "correct" responses). It is formally represented as

$$\text{Precision} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|}$$

Recall :- This is the percentage of records that are relevant to the query and were actually retrieved. It is formally represented as

$$\text{recall} = \frac{|\{\text{Relevant}\} \cap \{\text{Retrieved}\}|}{|\{\text{Relevant}\}|}$$

→ $$\text{F\_Score} = \frac{\text{recall} \times \text{Precision}}{(\text{recall} + \text{Precision})^2}$$

Indexing of documents

→ An efficient index structure is important for efficient Processing of queries in an information retrieval system.

→ · Documents that contain a specified key can be located efficiently by using an inverted index that maps each keyword K_i to a list S_i of (identifiers of) the documents that contain K_i.

Inverted Index:-

An Inverted Index is a data structure used in Information retrieval systems to efficiently retrieve documents or web pages containing a specific term or set of terms.

Document1 → The, quick, brown, fox, jumped, over, lazy, dog.

Document2 → The, lazy, dog, slept, in, the, sun

create an index of terms

The → Document1, document2 (1, 1) (1, 7) [2

quick → Document1 (1, 2)

brown → Doc1 (1, 3)

fox → Doc1 (1, 4)

jumped → Doc_1 (1, 5)

over → Doc_1 (1, 6)

lazy → Doc_1, Doc_2 (1, 8) (2, 2)

dog → Doc_1, Doc_2 (1, 9) (2, 3)

slept → Doc_2 (2, 4)

in → Doc_2 (2, 5)

sun → Doc_2 (2, 7)

Example

→ The documents d_1, d_9, d_{21} contain the term "silberschatz", the inverted list for the keyword Silberschatz would be "d_1; d_9; d_{21}"

→ It also provide the list of locations within the documents where the keyword appears.

Eg:- If "silberschatz" appeared at Position 21 in d_1, Positions 1 and 19 in d_9, Positions 4, 29 and 46 in d_3, inverted list with positions would be "d_1/21"; "d_9/1,19"; "d_{21}/4,29,46".

→ and operation: Finds documents that contain all of $k_1, k_2 \dots k_n$.

Intersection $S_1 \cap S_2 \cap \dots \cap S_n$

or operation: documents that contain one of $k_1, k_2 \ldots k_n$

union, $S_1 \cup S_2 \cup \ldots S_n$.

→ not operation: documents that do not contain a specified keyword k_i.

Merging of intersection / union.

Measuring Retrieval Effectiveness

→ Information-retrieval systems save space by using index structures that supports only appropriate retrieval. May result in

→ false negative (false drop) – some relevant documents may not be retrieved

→ false positive → some irrelevant documents may be retrieved.

→ For many applications a good index should not permit any false drops, but may permit a few false positives.

→ Recall Vs Precision

1. Can increase recall by retrieving many documents (down to a low level of relevance ranking), but may irrelevant documents would be fetched, reducing Precision.

→ Measures of retrieval effectiveness

Recall as a function of number of documents fetched or Precision as a function of recall.

Eg Precision of 75% at recall of 50% and 60% at a recall of 75%

Problem is which documents are actually relevant and which are not.

→ Crawling and Indexing the web

→ web crawlers are Programs that locate and gather information on web.

→ The recursively follow the hyperlinks Present in the known documents to find the other documents.

→ Starting from an initial set of

→ The fetched documents

- Handed over to an indexing system
- Can be discarded after indexing or as a cached copy.

disadvantage:- Crawling the entire web would take a very large amount of time

- Search engines typically cover only a part of the web, not all of it.
- Take months to perform a single cr

To overcome this.

→ Crawling is done by multiple Pro on multiple machines, running in Parallel

- Set of links to be Crawled stored a database.
- New links found in Crawled Pages a to this set, to be Cavawled later.

→ Indexing process also runs on multip machines.

- Creates a new copy of index instead modifying old index.

. old index is used to answer queries.
. After a crawl is "completed" new index becomes "old" index

→ Multiple machines used to answer queries.
. indices may be kept in memory.
. Queries may be routed to different machines for load balancing.

→ Information Retrieval: Beyond Ranking of Pages

1. Diversity of Query Results.
2. Information extraction :- Systems convert information from textual form to a more structured form.
3. Question Answering :- Information retrieval systems focus on finding documents relevant to a given query. However, the answer to a query may lie in just one part of a document, or in small parts of several documents.

→ Question answering systems attempt to Provide direct answers to question by users.

4. Querying structured data.

Directories and Categories

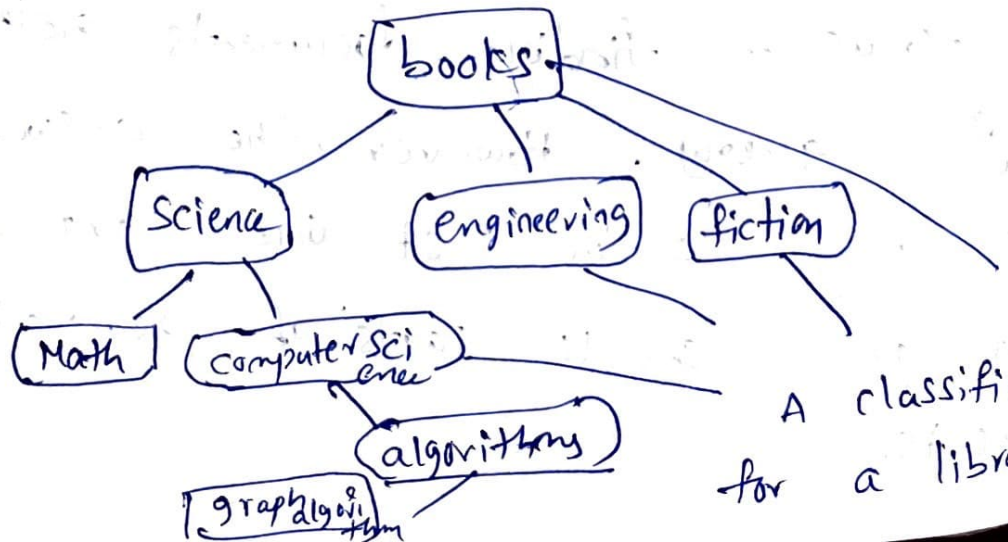
1. Storing related documents together

a. library facilitates browsing.

↳ users can see not only requested document but also related ones.

→ Browsing is facilitated by classification system that organizes logically related documents together.

→ To keep related books close together libraries use a classification hierarchy



A classification for a library sy

→ A directory is simply a classification DAG structure.

→ Each leaf of the directory stores links to documents on the topic represented by the leaf.

→ Internal nodes may also contain links.

→ Documents can reside in multiple places in a hierarchy in an information retrieval system. Since physical location is not important.

→ classification hierarchy is thus Directed Acyclic Graph (DAG)

A classification DAG for a Library information Retrieval System