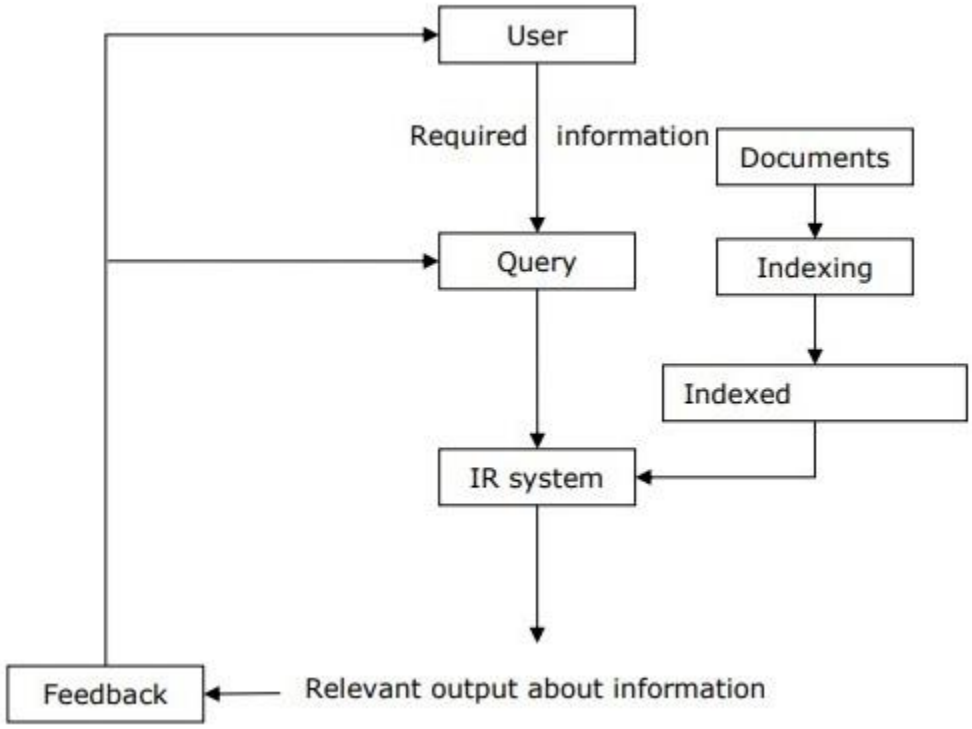
INFORMATION RETRIEVAL

Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information. The system assists users in finding the information they require but it does not explicitly return the answers of the questions. It informs the existence and location of documents that might consist of the required information. The documents that satisfy user's requirement are called relevant documents. A perfect IR system will retrieve only relevant documents.

With the help of the following diagram, we can understand the process of information retrieval (IR) −



It is clear from the above diagram that a user who needs information will have to formulate a request in the form of query in natural language. Then the IR system will respond by retrieving the relevant output, in the form of documents, about the required information.

Classical Problem in Information Retrieval (IR) System

The main goal of IR research is to develop a model for retrieving information from the repositories of documents. Here, we are going to discuss a classical problem, named **ad-hoc retrieval problem**, related to the IR system.

In ad-hoc retrieval, the user must enter a query in natural language that describes the required information. Then the IR system will return the required documents related to the desired information. For example, suppose we are searching something on the Internet and it gives some exact pages that are relevant as per our requirement but there can be some non-relevant pages too. This is due to the ad-hoc retrieval problem.

Aspects of Ad-hoc Retrieval

Followings are some aspects of ad-hoc retrieval that are addressed in IR research −

- How users with the help of relevance feedback can improve original formulation of a query?
- How to implement database merging, i.e., how results from different text databases can be merged into one result set?
- How to handle partly corrupted data? Which models are appropriate for the same?

Information Retrieval (IR) Model

Mathematically, models are used in many scientific areas having objective to understand some phenomenon in the real world. A model of information retrieval predicts and explains what a user will find in relevance to the given query. IR model is basically a pattern that defines the above-mentioned aspects of retrieval procedure and consists of the following −

- A model for documents.
- A model for queries.
- A matching function that compares queries to documents.

Mathematically, a retrieval model consists of −

D − Representation for documents.

R − Representation for queries.

F − The modeling framework for D, Q along with relationship between them.

R (q,di) − A similarity function which orders the documents with respect to the query. It is also called ranking.

Types of Information Retrieval (IR) Model

An information model (IR) model can be classified into the following three models −

Classical IR Model

It is the simplest and easy to implement IR model. This model is based on mathematical knowledge that was easily recognized and understood as well. Boolean, Vector and Probabilistic are the three classical IR models.

Non-Classical IR Model

It is completely opposite to classical IR model. Such kind of IR models are based on principles other than similarity, probability, Boolean operations. Information logic model, situation theory model and interaction models are the examples of non-classical IR model.

Alternative IR Model

It is the enhancement of classical IR model making use of some specific techniques from some other fields. Cluster model, fuzzy model and latent semantic indexing (LSI) models are the example of alternative IR model.

Design features of Information retrieval (IR) systems

Let us now learn about the design features of IR systems −

Inverted Index

The primary data structure of most of the IR systems is in the form of inverted index. We can define an inverted index as a data structure that list, for every word, all documents that contain it and frequency of the occurrences in document. It makes it easy to search for ‘hits’ of a query word.

Stop Word Elimination

Stop words are those high frequency words that are deemed unlikely to be useful for searching. They have less semantic weights. All such kind of words are in a list called stop list. For example, articles “a”, “an”, “the” and prepositions like “in”, “of”, “for”, “at” etc. are the examples of stop words. The size of the inverted index can be significantly reduced by stop list. As per Zipf’s law, a stop list covering a few dozen words reduces the size of inverted index by almost half. On the other hand, sometimes the elimination of stop word may cause elimination of the term that is useful for searching. For example, if we eliminate the alphabet “A” from “Vitamin A” then it would have no significance.

Stemming

Stemming, the simplified form of morphological analysis, is the heuristic process of extracting the base form of words by chopping off the ends of words. For example, the words laughing, laughs, laughed would be stemmed to the root word laugh.

In our subsequent sections, we will discuss about some important and useful IR models.

Components of Information Retrieval Model

Here are the prerequisites for an IR model:

1. An automated or manually-operated indexing system used to index and search techniques and procedures.
2. A collection of documents in any one of the following formats: text, image or multimedia.
3. A set of queries that serve as the input to a system, via a human or machine.
4. An evaluation metric to measure or evaluate a system’s effectiveness (for instance, precision and recall). For instance, to ensure how useful the information displayed to the user is.

Acquisition

Documents and other things are being chosen from various websites.

1. Documents that are mostly text-based o entire texts, titles, abstracts

2. Other research-based objects like Data, statistics, photos, maps, copyrights, soundscapes, and so on...
3. Web crawlers take data and store it in a database.

Representation

The representation of information retrieval system mainly involves indexing the following:

- Indexing may be done in a variety of methods, including free text keywords (even in entire texts) o regulated vocabulary - thesaurus o manual and automatic procedures.
- Summarizing and abstracting
- Bibliographic information: author, title, sources, date, etc.
- Information about metadata
- Classification and clustering
- Field and limit organization
- Basic Index, Supplemental Index Limits

File Organisation

There are mainly 2 categories of file organization which are: sequential and inverted. The mixture of these two is a combination.

- Sequential

It organizes documents based on document data.

- Reversed

It provides a list of records under each phrase, term by term.

- Combination

Synthesis of inverted indexes as well as sequential documents

When just citations are retrieved, there is no requirement for document files. It leads to approaches for large files and for computer retrieval efficiency.

Query

When a user inputs a query into the system, an IR process begins. Queries, such as search strings in web search engines, are explicit representations of information requests. A query in information retrieval system does not uniquely identify a particular object in a collection. Instead, numerous things may match the query, maybe with varying degrees of significance.