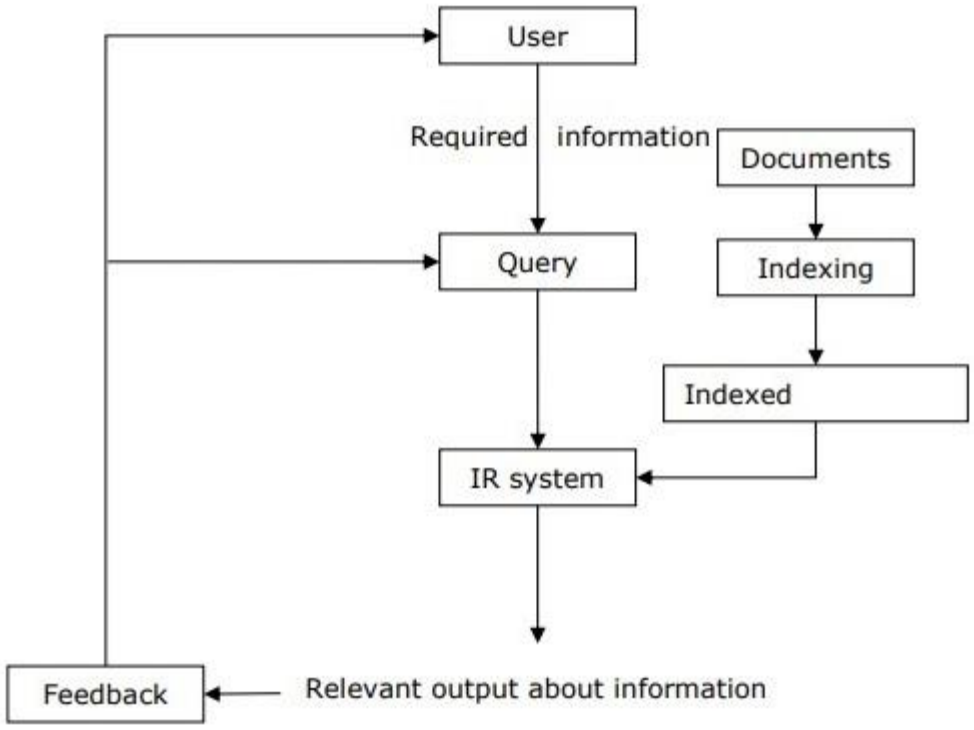
Information retrieval (IR) may be defined as a software program that deals with the organization, storage, retrieval and evaluation of information from document repositories particularly textual information. The system assists users in finding the information they require but it does not explicitly return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

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

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 Zipfs 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.

The Boolean Model

It is the oldest information retrieval (IR) model. The model is based on set theory and the Boolean algebra, where documents are sets of terms and queries are Boolean expressions on terms. The Boolean model can be defined as −

- **D** − A set of words, i.e., the indexing terms present in a document. Here, each term is either present (1) or absent (0).
- **Q** − A Boolean expression, where terms are the index terms and operators are logical products − AND, logical sum − OR and logical difference − NOT
- **F** − Boolean algebra over sets of terms as well as over sets of documents

If we talk about the relevance feedback, then in Boolean IR model the Relevance prediction can be defined as follows −

- **R** − A document is predicted as relevant to the query expression if and only if it satisfies the query expression as −

(( ))

We can explain this model by a query term as an unambiguous definition of a set of documents.

For example, the query term ***economic*** defines the set of documents that are indexed with the term ***economic***.

Now, what would be the result after combining terms with Boolean AND Operator? It will define a document set that is smaller than or equal to the document sets of any of the single terms. For example, the query with terms ***social*** and ***economic*** will produce the documents set of documents that are indexed with both the terms. In other words, document set with the intersection of both the sets.

Now, what would be the result after combining terms with Boolean OR operator? It will define a document set that is bigger than or equal to the document sets of any of the single terms. For example, the query with terms ***social*** or ***economic*** will produce the documents set of documents that are indexed with either the term ***social*** or ***economic***. In other words, document set with the union of both the sets.

Advantages of the Boolean Mode

The advantages of the Boolean model are as follows −

- The simplest model, which is based on sets.
- Easy to understand and implement.
- It only retrieves exact matches
- It gives the user, a sense of control over the system.

Disadvantages of the Boolean Model

The disadvantages of the Boolean model are as follows −

- The models similarity function is Boolean. Hence, there would be no partial matches. This can be annoying for the users.
- In this model, the Boolean operator usage has much more influence than a critical word.
- The query language is expressive, but it is complicated too.
- No ranking for retrieved documents.

Information Extraction (IE)

Definition:

Information Extraction is the task of **automatically identifying and extracting structured information** from unstructured or semi-structured data (like text, web pages, or emails).

Core Tasks in IE:

Task	Purpose
Named Entity Recognition (NER)	Identifies entities like names, dates, places, and organizations
Relation Extraction	Extracts relationships between entities (e.g., "Steve Jobs founded Apple")
Event Extraction	Identifies events and participants (e.g., "Flood in Mumbai in July 2021")
Coreference Resolution	Identifies when different expressions refer to the same entity
Template Filling	Extracts values to fill in pre-defined forms or templates

Techniques in IE:

a. Rule-Based Systems

- Use regular expressions and manually defined patterns
- High precision, low flexibility

b. Statistical Models

- CRFs (Conditional Random Fields), HMMs (Hidden Markov Models)
- Require labeled training data

c. Deep Learning Models

- Use models like **BiLSTM**, **BERT**, **RoBERTa**
- Capture context and semantics for accurate extraction

Example (NER):

Sentence: “Elon Musk was born in South Africa.”
Extraction:

- Person: Elon Musk
- Location: South Africa

Applications of IE:

- Resume parsing and HR automation
- News summarization
- Financial market analysis
- Biomedical information mining
- Social media monitoring