What is Information Retrieval?

Information Retrieval (IR) can be defined as a software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories, particularly textual information. Information Retrieval is the activity of obtaining material that can usually be documented on an unstructured nature **i.e.** usually text which satisfies an information need from within large collections which is stored on computers. For example, Information Retrieval can be when a user enters a query into the system.
Not only librarians, professional searchers, etc engage themselves in the activity of information retrieval but nowadays hundreds of millions of people engage in IR every day when they use web search engines. Information Retrieval is believed to be the dominant form of Information access.

The **IR system** assists the users in finding the information they require but it does not explicitly return the answers to the question. It notifies regarding the existence and location of documents that might consist of the required information. Information retrieval also extends support to users in browsing or filtering document collection or processing a set of retrieved documents. **The system searches over billions of documents stored on millions of computers**. A spam filter, manual or automatic means are provided by Email program for classifying the mails so that it can be placed directly into particular folders.
An IR system has the ability to represent, store, organize, and access information items. A set of **keywords** are required to search. Keywords are what people are searching for in search engines. These keywords **summarize the description of the information**.

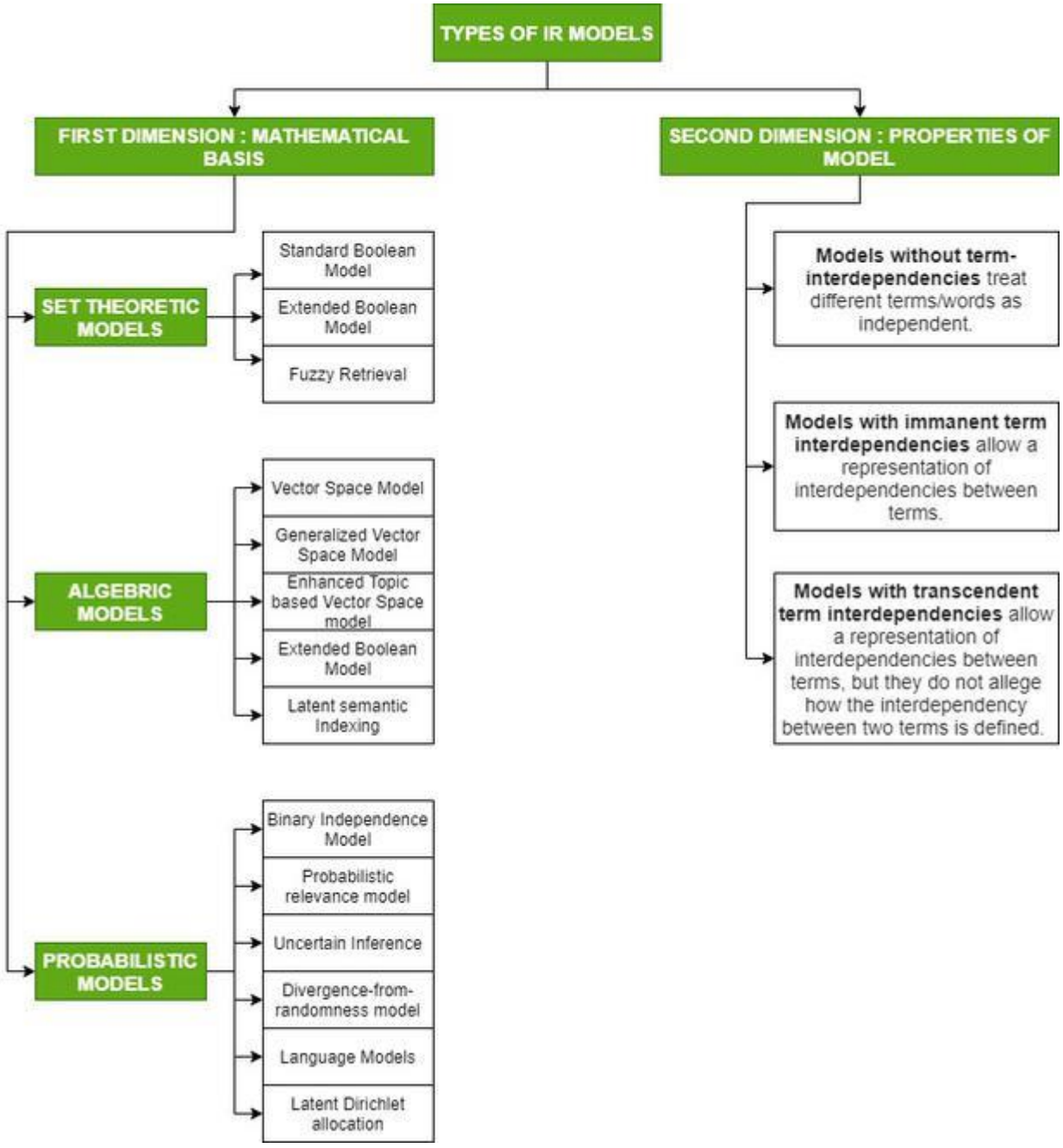
What is an IR Model?

An Information Retrieval (IR) model selects and ranks the document that is required by the user or the user has asked for in the form of a query. The documents and the queries are represented in a similar manner, so that document selection and ranking can be formalized by a matching function that returns a **retrieval status value (RSV)** for each document in the collection. Many of the Information Retrieval systems represent document contents by a set of descriptors, called **terms**, belonging to a **vocabulary V**. An IR model determines the query-document matching function according to four main approaches:

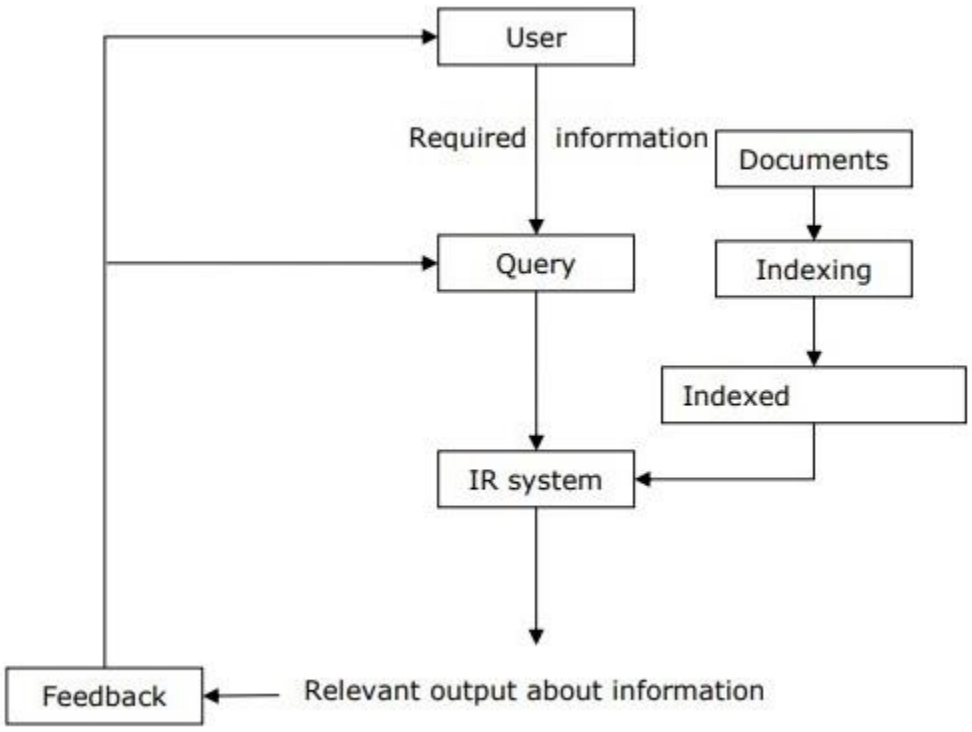
The estimation of the probability of user's relevance ***rel*** *for each document* ***d*** *and query* ***q*** *with respect to a set* ***R****.*

q *of training documents:* $\boldsymbol{Prob\ (rel|d, q, R_q)}$

Types of IR Models


TYPES OF IR MODELS
FIRST DIMENSION : MATHEMATICAL BASIS
SECOND DIMENSION : PROPERTIES OF MODEL
SET THEORETIC MODELS
Standard Boolean Model
Extended Boolean Model
Fuzzy Retrieval
ALGEBRIC MODELS
Vector Space Model
Generalized Vector Space Model
Enhanced Topic based Vector Space model
Extended Boolean Model
Latent semantic Indexing
PROBABILISTIC MODELS
Binary Independence Model
Probabilistic relevance model
Uncertain Inference
Divergence-from-randomness model
Language Models
Latent Dirichlet allocation
Models without term-interdependencies treat different terms/words as independent.
Models with immanent term interdependencies allow a representation of interdependencies between terms.
Models with transcendent term interdependencies allow a representation of interdependencies between terms, but they do not allege how the interdependency between two terms is defined.

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.

Boolean models are discussed.

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 −

$((text \vee information) \wedge rerieval \wedge \tilde{} \; theory)$

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.

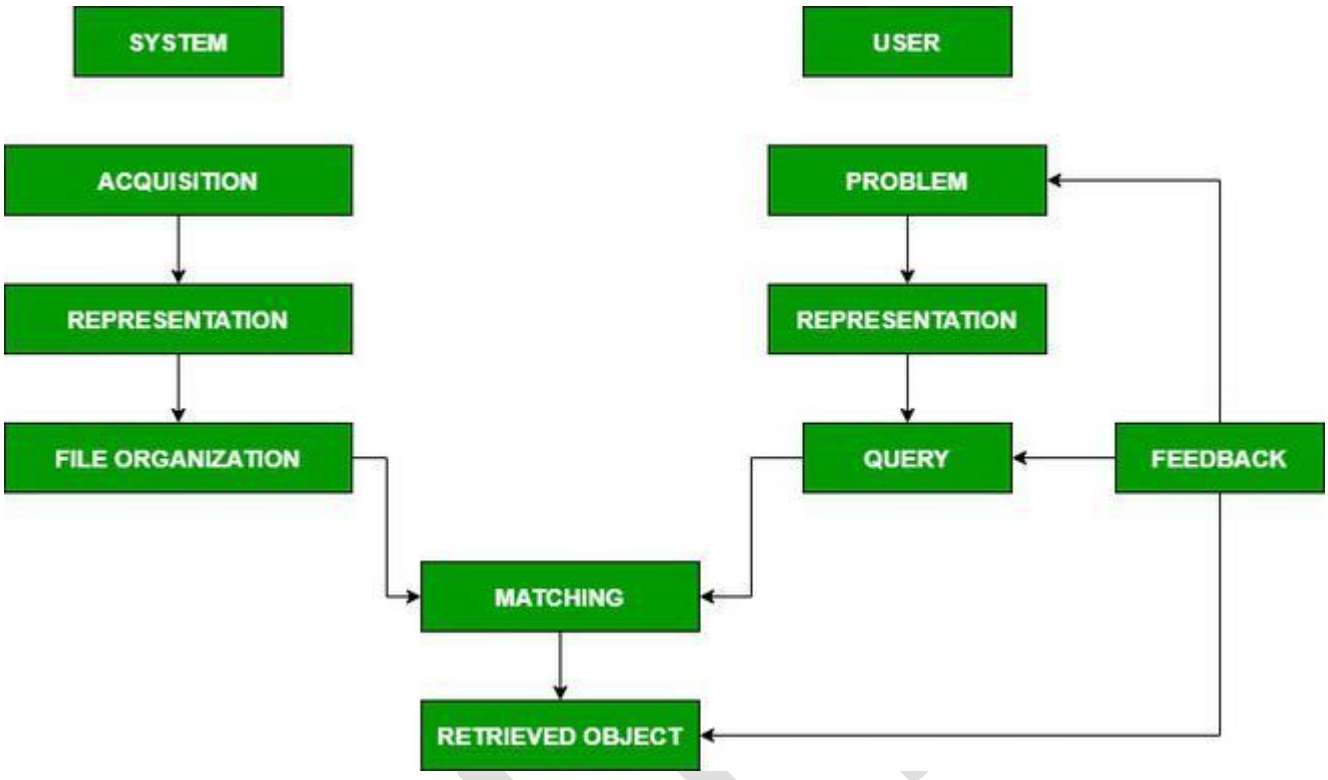
Now, what would be the result after combining terms with Boolean OR operator? It will define a document set that is bigger than or equal to the document sets of any of the single terms. For example, the query with terms ***"social"*** or ***"economic"*** will produce the documents set of documents that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

Components of Information Retrieval/ IR Model



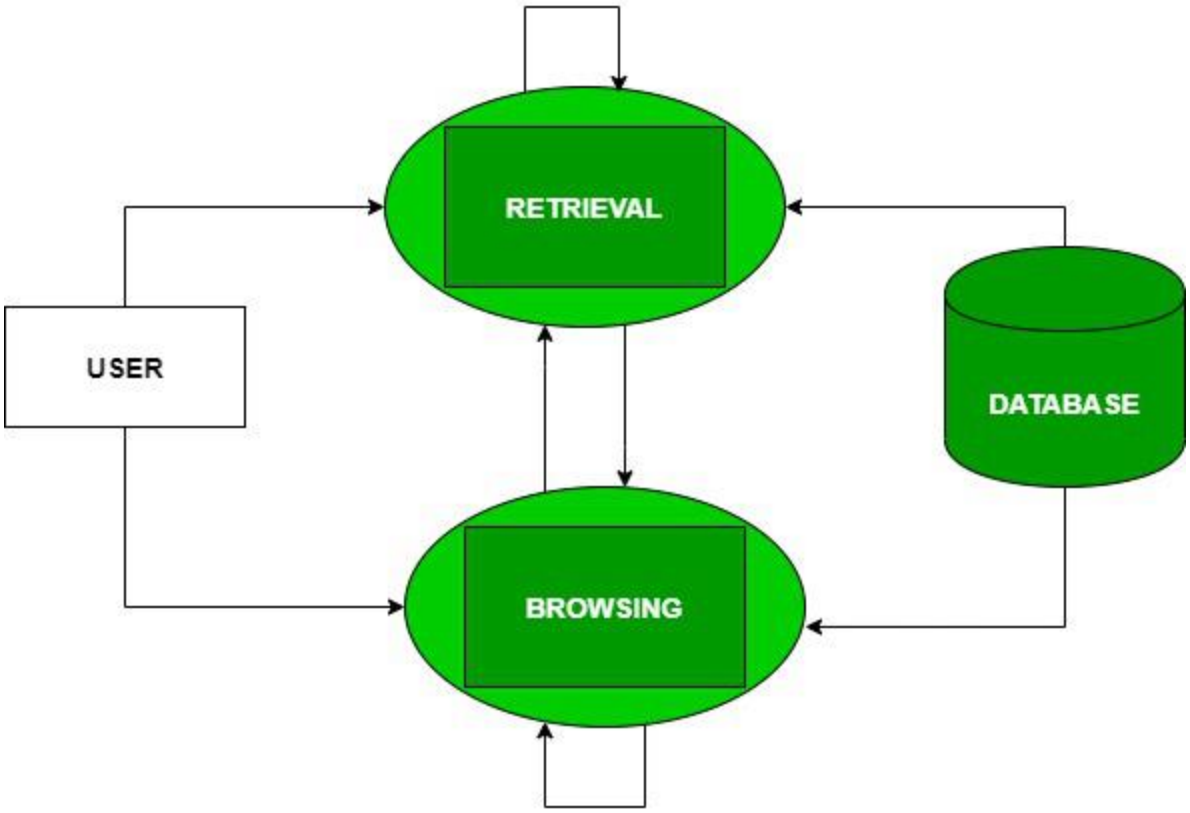
- **Acquisition:** In this step, the selection of documents and other objects from various web resources that consist of text-based documents takes place. The required data is collected by web crawlers and stored in the database.
- **Representation:** It consists of indexing that contains free-text terms, controlled vocabulary, manual & automatic techniques as well. example: Abstracting contains summarizing and Bibliographic description that contains author, title, sources, data, and metadata.
- **File Organization:** There are two types of file organization methods. i.e. *Sequential*: It contains documents by document data. *Inverted*: It contains term by term, list of records under each term. *Combination* of both.
- **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 the collection. Instead, several

objects may match the query, perhaps with different degrees of relevancy.

Difference Between Information Retrieval and Data Retrieval

Information Retrieval	Data Retrieval
The software program that deals with the organization, storage, retrieval, and evaluation of information from document repositories particularly textual information.	Data retrieval deals with obtaining data from a database management system such as ODBMS. It is A process of identifying and retrieving the data from the database, based on the query provided by user or application.
Retrieves information about a subject.	Determines the keywords in the user query and retrieves the data.
Small errors are likely to go unnoticed.	A single error object means total failure.
Not always well structured and is semantically ambiguous.	Has a well-defined structure and semantics.
Does not provide a solution to the user of the database system.	Provides solutions to the user of the database system.
The results obtained are approximate matches.	The results obtained are exact matches.
Results are ordered by relevance.	Results are unordered by relevance.
It is a probabilistic model.	It is a deterministic model.

User Interaction With Information Retrieval System



The User Task: The information first is supposed to be translated into a query by the user. In the information retrieval system, there is a set of words that convey the semantics of the information that is required whereas, in a data retrieval system, a query expression is used to convey the constraints which are satisfied by the objects. Example: A user wants to search for something but ends up searching with another thing. This means that the user is browsing and not searching. The above figure shows the interaction of the user through different tasks.

- **Logical View of the Documents:** A long time ago, documents were represented through a set of index terms or keywords. Nowadays, modern computers represent documents by a full set of words which

reduces the set of representative keywords. This can be done by eliminating stopwords i.e. articles and connectives. These operations are text operations. These text operations reduce the complexity of the document representation from full text to set of index terms.

Past, Present, and Future of Information Retrieval

1. Early Developments: As there was an increase in the need for a lot of information, it became necessary to build data structures to get faster access. The index is the data structure for faster retrieval of information. Over centuries manual categorization of hierarchies was done for indexes.

2. Information Retrieval In Libraries: Libraries were the first to adopt IR systems for information retrieval. In first-generation, it consisted, automation of previous technologies, and the search was based on author name and title. In the second generation, it included searching by subject heading, keywords, etc. In the third generation, it consisted of graphical interfaces, electronic forms, hypertext features, etc.

3. The Web and Digital Libraries: It is cheaper than various sources of information, it provides greater access to networks due to digital communication and it gives free access to publish on a larger medium.