**Subject: IRS**

**Unit 1. Introduction**

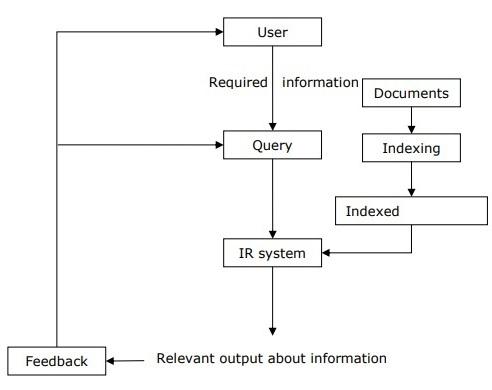
1.Textbook: <https://srikarthiks.files.wordpress.com/2016/07/t2-modern-information-retrieval.pdf>

**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.

### **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. |

**The objective of Information Retrieval System (IRS) is-**

* From a terminological perspective, the term "Information Retrieval System" (IRS) refers to a "system which retrieves information."
* IRS is focused on two fundamental issues: Information storage and retrieval techniques.
* Retrieving the required information is a major goal of an information retrieval system.
* It could be actual information or information surrogates from documents that completely or partially match the user's search.
* An information retrieval system is made to assess, process, and store information sources, then retrieve the ones that best meet the needs of a certain user.
* From a database of documents that has been stored, modern information retrieval systems can either return bibliographic entries or the specific text that meets a user's search criterion.
* Since they dealt with text-based documents, the IRS initially meant text retrieval systems.
* Modern information retrieval systems deal with multimodal information, which includes text, audio, images, and video in addition to text-only information.
* As a result, contemporary information retrieval systems handle the organizing, storage, and access of textual information sources as well as multimedia ones.

## **What are the types of information retrieval?**

Methods/Techniques in which information retrieval techniques are employed include:

* Adversarial information retrieval
* Automatic summarization
* Multi-document summarization
* Compound term processing
* Cross-lingual retrieval
* Document classification
* Spam filtering
* Question answering

## **What are the components and features of Information retrieval systems?**

### **1. 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.

### **2. 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](https://www.engati.com/glossary/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.

### **3. Stemming**

[Stemming](https://www.engati.com/glossary/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.

### **4. Crawling**

Crawling is the process of gathering different web pages to index them to support a search engine. The purpose of crawling is to quickly and efficiently gather as many relevant web pages as possible and together with the link structure that interconnects them.

### **5. Query**

Queries are search statements which describe the information requirements in search engines. A query will never identify one particular result, it will find many results which match the query with different degrees.

### **6. Relevance Feedback**

Relevance feedback helps in taking results that are initially returned from a specific query, to gather user feedback, and determine whether those results are relevant to perform a new query.

### ‍**Search Engine:**

### A search engine is a kind of website through which users can search the content available on the Internet. For this purpose, users enter the desired keywords into the search field. Then the search engine looks through its index for relevant web pages and displays them in the form of a list. The Internet is a huge source of information & resources and to access the resource from the Internet there are some kinds of software, this software is known as a Search Engine. Some of the popular ones are Google, Bing, Yahoo, Duck duck go, Baidu, etc.

**There are three main components of the Search engine:**

* **Crawler:** Crawlers are software programs sometimes referred to the bots. It regularly scans the websites automatically for URLs, keywords, and links to discover the new updates. The crawler can follow the links present on some other webpage.
* **Index:** As we know, the Crawler continuously scans the websites, it develops an index of URLs, links, and keywords to make the search results more effective.
* **Search Algorithm**: The search algorithm is the complete mechanism behind the whole searching process. It is working by searching for the index and finding the most suitable web pages by matching keywords that are searched by the users.

**Web Browser:**

The web browser is an example of application software that is developed to retrieve and view the information from web pages or HTML files present on the web servers. The first web browser was invented by Sir Tim Berners-Lee in 1990 and the very first graphical web browser was developed in 1993 and is named the mosaic. After that, various web browsers were developed. Some of them are navigator which is developed by Netscape communication, Microsoft Edge, Google Chrome, Mozilla Firefox, Opera, and Apple safari.

**The main characteristics of a Web Browser are:**

* It consists of Graphical User Interface.
* It contains the search box where the user can type the address or URL.
* Page style can be static or dynamic. It depends upon the interactivity and the formatting.
* TCP/IP and HTTP protocols are used by web browsers.

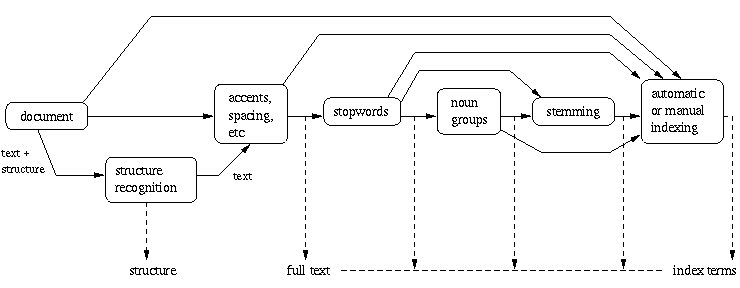
**Difference between Search Engine and Web Browser:**

| **S. No.** | **Parameters** | **Search Engine** | **Web Browser** |
| --- | --- | --- | --- |
| 1. | Definition | A search engine is used to find the information in the World Wide Web and displays the results at one place by returning web pages available on internet. | Web Browser uses the search engine to retrieve and view the information from web pages present on the web servers. |
| 2. | Usage | Search engine is intended to gather Information regarding several URL’s and to maintain it. | Web Browsers are intended to Display the web page of the current URL available at the server. |
| 3. | Installation | Search Engine need not to be installed on our system (i.e. comes as default). | Many Web Browsers can be installed on our system. |
| 4. | Accessibility | The search engine is accessed through a web browser. | Typically, all devices are supported. |
| 5. | Components | The Search Indexer, Crawler, and Database are the three essential components of a search engine. | A web browser uses a graphical user interface to help users have an interactive online session on the Internet. |
| 6. | Database | A search engine contains its own database. | No database is required in Web browser. It contains only cache memory to store cookies as well as browsing history until we remove it from our system. |
| 7. | Dependency | A search engine is not required to open the browser. This means that the search engine is reliant on the browser. | A browser is required to open a search engine. This means that the browser is not reliant on the search engine. |
| 8. | History | Unless you actively clear this data or use a private browsing mode, browsers will retain your browsing history, cookies, and cache in memory. | Typically, search engines acquire information on their users and their search queries. Some search engines, such as DuckDuckGo, do not gather user information. |
| 9. | Advantages | The major Advantages of using search engines are to Get the Consumer Trust, Trackable Results, Generates Targeted Traffic, Sustainable Clicks and Grow Your Small Business. | The major Advantages of using web browser are open standards, security sandbox, Robust GUI and Simple networking. |
| 10. | Disadvantages | The disadvantages of using search engines are difficult Of Competitive Keywords, Changing Algorithms and Results are Not Guarantee. | The disadvantages of using web browsers are slow down with the new version and no-add on support. |
| 11. | Examples | Example of famous search engines are: Google, Yahoo, Bing, DuckDuckgo, Baidu, Internet Explorer. | Some of the widely used web browsers are: Mozilla Firefox, Netscape Navigator, and Google Chrome. |

## **Logical View of the Documents**

Due to historical reasons, documents in a collection are frequently represented through a set of index terms or keywords. Such keywords might be extracted directly from the text of the document or might be specified by a human subject (as frequently done in the information sciences arena). No matter whether these representative keywords are derived automatically or generated by a specialist, they provide a [*logical view of the document*](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#L). For a precise definition of the concept of a document and its characteristics.

Modern computers are making it possible to represent a document by its full set of words. In this case, we say that the retrieval system adopts a [full text](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#F) logical view (or representation) of the documents. With very large collections, however, even modern computers might have to reduce the set of representative keywords. This can be accomplished through the elimination of [stopwords](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#S) (such as articles and connectives), the use of [stemming](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#S)  (which reduces distinct words to their common grammatical root), and the identification of noun groups (which eliminates adjectives, adverbs, and verbs). Further, compression might be employed. These operations are called text operations  (or transformations) and are covered in detail in Chapter 7. Text operations reduce the complexity of the document representation and allow moving the logical view from that of a full text to that of a set of [index terms](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#I) .

**Figure:** 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. A small set of categories (generated by a human specialist) provides the most concise logical view of a document but its usage might lead to retrieval of poor quality. Several intermediate logical views (of a document) might be adopted by an information retrieval system as illustrated in Figure . Besides adopting any of the intermediate representations, the retrieval system might also recognize the internal structure normally present in a document (e.g., chapters, sections, subsections, etc.). This information on the structure of the document might be quite useful and is required by structured text retrieval models such as those discussed in Chapter 2.

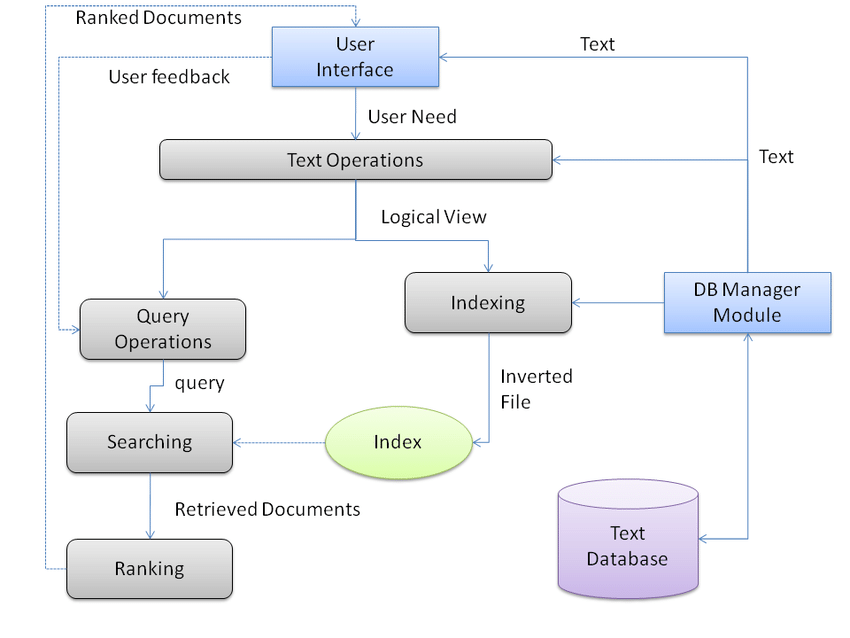
As illustrated in Figure , we view the issue of logically representing a document as a continuum in which the logical view of a document might shift (smoothly) from a full text representation to a higher level representation specified by a human subject.

**The Retrieval Process**

At this point, we are ready to detail our view of the retrieval process. Such a process is interpreted in terms of component subprocesses whose study yields many of the chapters in this book.

To describe the retrieval process, we use a simple and generic software architecture as shown in Figure . First of all, before the retrieval process can even be initiated, it is necessary to define the text database. This is usually done by the manager of the database, which specifies the following: (a) the documents to be used, (b) the operations to be performed on the text, and (c) the text model (i.e., the text structure and what elements can be retrieved). The text operations transform the original documents and generate a logical view of them.

Once the logical view of the documents is defined, the database manager (using the DB Manager Module) builds an index of the text. An index is a critical data structure because it allows fast searching over large volumes of data. Different index structures might be used, but the most popular one is the [inverted file](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#I) as indicated in Figure . The resources (time and storage space) spent on defining the text database and building the index are amortized by querying the retrieval system many times.



**Figure:**

The process of retrieving information (the numbers beside each box indicate the chapters that cover the corresponding topic).

Given that the document database is indexed, the retrieval process can be initiated. The user first specifies a [user need](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#U) which is then parsed and transformed by the same text operations applied to the text. Then, query operations might be applied before the actual query, which provides a system representation for the user need, is generated. The query is then processed to obtain the retrieved documents. Fast query processing is made possible by the index structure previously built.

Before been sent to the user, the retrieved documents are ranked according to a likelihood of relevance. The user then examines the set of ranked documents in the search for useful information. At this point, he might pinpoint a subset of the documents seen as definitely of interest and initiate a [user feedback](https://people.ischool.berkeley.edu/~hearst/irbook/glossary.html#U) cycle. In such a cycle, the system uses the documents selected by the user to change the query formulation. Hopefully, this modified query is a better representation of the real user need.

Consider now the user interfaces available with current information retrieval systems (including Web search engines and Web browsers). We first notice that the user almost never declares his information need. Instead, he is required to provide a direct representation for the query that the system will execute. Since most users have no knowledge of text and query operations, the query they provide is frequently inadequate. Therefore, it is not surprising to observe that poorly formulated queries lead to poor retrieval (as happens so often on the Web).

### **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.

**Questions:**

Q1. Differentiate between data retrieval and information retrieval.

Q2. How does search engine retrieves the information?

Q3. Explain logical view of a document with diagram