Information Retrieval System

Priyanka Debnath
Dept. Library and Information Science
Student of M.li.s
Jadavpur University

Information Retrieval System

Information retrieve is a software program that deals with organization, storage retrieval and evaluation of information from document repositories, particularly textual information.

Information Retrieval System helps to locate and retrieve relevant information from a large collection of data, such a database or a document collection.

Examples of IRS :

Google search Yahoo search Bing search

How does it works?

When a user puts a query into a system the IR system creates an index in a content collection or information database of documents. Including text documents , images , audio, and videos are processed to extract relevant terms. And also gets ready to surrogate data, and data structures are used to efficiently store and retrieve those entities.

When a user enters a query , the system immediately identifies the most relevant terms and determines the importance. And then the system ranks documents based on their relevance to the query.

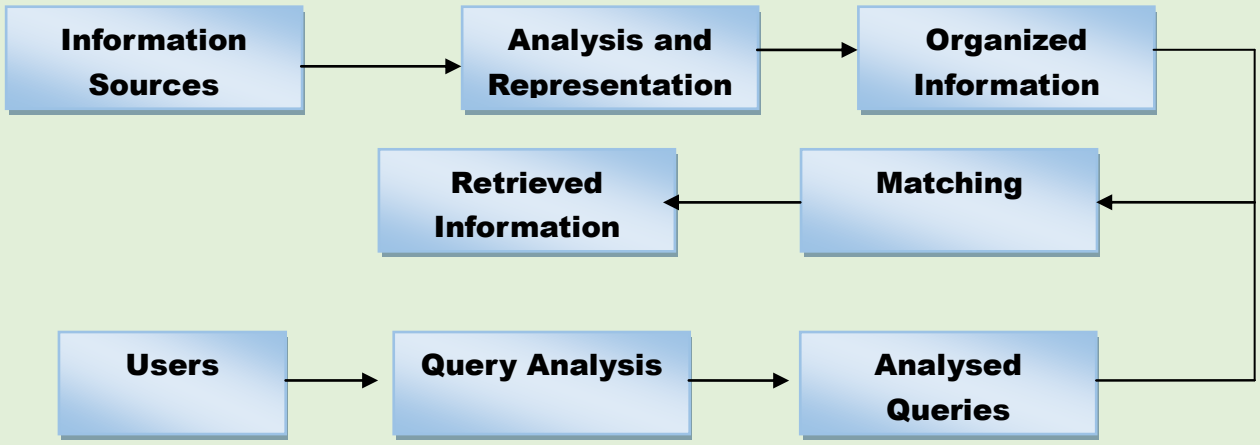

Fig.Taken from introduction to modern information retrieval

Objectives:

- The main objective of an IRS is to reduce the overhead of a user locating required information.

- Overhead is expressed as the time a user spends in all of the steps leading to reading an item containing the needed information.

- Overhead for a user to find the needed information.

Key features of Information Retrieval System:

- Data collection and Indexing: any kind of text data, image can be searched by IR system.

Fast searching, analysing documents and creating optimized data structures to map terms to documents containing them.

- Automatic Indexing: automatic indexing refers to the process of generating indexes for large collections of documents.

Information Organization

It is a systematic arrangement of data. This process is important in various contexts, like personal,academic, and professional environments. In this process we make it easier to retrieve and utilize when needed.

Types of Information organizations:

Libraries
Museums
Documentation centers
Publishing houses
Archives

Relation Between Organization of information and IRS

- Foundation for retrieval: if entered data is logically categorized, indexed and stored, an IR system can- 1. retrieve relevant documents faster. 2. Deliver more accurate details.
- Quick retrieval of indexing data. A digital library that categorises books by author, genre, and keywords allows the IR system to support advanced searches.
- Improves search Accuracy:
  Well organized information in a library allows IR system to:
  1. Understand user queries better.
  2. Match queries to the right documents
  3. Reduce irrelevant results

Organization tools like taxonomies, ontologies and metadata enhance the precision and recall of searches.

- Enables Advanced search features: Searching by file type, category, or language depends on the presence of structured metadata.

Tools of Organizing Information Resources

1. Classification Schemes:

A classification scheme is an organized system. It arranges knowledge into a structured order. It allows libraries and information organizations to manage vast collections of documents efficiently. The primary purpose of classification scheme is to group resources based on their subject content. For that the users can easily locate, retrieve, and browse related materials. It provides each subject areas with a unique notation or class number, letters, or a combination of both, which serves as an identifier on catalogues, indexes, and

shelves. Some classification schemes such as Dewey Decimal Classification (DDC), Library of Congress Classification (LCC), Universal Decimal Classification (UDC), and Colon Classification (CC) have become the foundation of knowledge organization worldwide. These systems are hierarchical in nature. They move from broader classes to narrower subdivisions, thus reflecting the logical relationships between subjects. For example, in DDC, knowledge is divided into ten main classes, and each classes further broken into hundred specific topics, again each of the topic broken into thousand specific topic ensuring both simplicity and flexibility. A good classification scheme ensures consistency, standardization, and accessibility. Classification not only saves time but also reduces duplication and confusion in large collections. In addition modern classification schemes also play a significant role in digital libraries and online retrieval systems to supporting physical arrangement, where classification notations and subject hierarchies are linked with metadata and search engines to enhance discoverability. Thus, for information organizations the classification schemes are essential tools, as they enable systematic arrangement, facilitate interdisciplinary research, and promote universal access to knowledge.

2. Thesaurus:

Thesaurus is a structured vocabulary tool in the context of information science that establishes relationships between terms to facilitate accurate and efficient information retrieval. A thesaurus not only provides synonyms unlike a simple dictionary or glossary, but also organizes terms hierarchically, associatively, and equivalently, thereby ensuring consistency and precision in subject representation. It standardizes the language of indexing by controlling synonyms, homonyms, and variant spellings so that users retrieve all relevant documents irrespective of the terminology used in their queries. For example, if a user searches for “automobile” a thesaurus can guide the system to also retrieve documents indexed under “car” or “vehicle”. In practice, a thesaurus contains preferred terms (descriptors), non-preferred terms (use for), broader terms (BT), narrower terms (NT), and related terms (RT), each linked to indicate conceptual

relationships. This structure of semantic not only improves retrieval accuracy but also assists users in exploring broader or narrower aspects of a subject, thereby enhancing discovery and learning. Standard thesauri follows international guidelines. They are ANSI/NISO Z39.19 and ISO 25964, which define rules for the construction, display, and maintenance of controlled vocabularies. It is widely used in bibliographic databases, digital libraries, and subject-specific indexing services such as ERIC Thesaurus for education and MeSH (Medical Subject Heading) for biomedical sciences. Sometime Thesauri are integrated with metadata and ontologies to support advanced search functions and semantic web applications. Thus, a thesaurus serves as an essential tool for controlled vocabulary. It bridges the gap between user queries and indexed content.

3. Facet:

In library and information science, particularly in the organization and classification of knowledge A facet is a fundamental concept. Facet refers to a distinct and independent characteristic or dimension of a subject. So that it can describe complex topics more precisely. The concept Facet was introduced by S.R. Ranganathan in his Colon Classification system. This CC system revolutionized traditional linear classification by allowing subjects to be broken down into multiple aspects. According to him knowledge can be analyzed into five fundamental categories, popularly known as PMEST.

P=Personality

M=Matter

E=Energy

S=Space

T=Time

Each of these represents a facet. When combined they can describe any subject comprehensively. For example, the subject

“Treatment of Malaria in India during the 20th Century”

By analyzing into facets: Personality= Malaria, Energy= Treatment, Space= India, and Time= 20th Century. This faceted approach ensures flexibility, as new subjects. We can also get interdisciplinary topics can be easily represented without disturbing the existing system. Rigid hierarchical classification schemes, faceted classification provides a synthetic and dynamic structure. It makes particularly suitable for modern digital and online information retrieval systems. We can get facets widely used in online search interfaces like e-commerce sites, library OPACs, and digital repositories. Users can refine their searches by filtering results through categories like author, subject, date, format, language. Facets not only bring order and clarity to knowledge organization but also enhance the user experience.

In short, facets are not just theoretical classification tool but a practical framework for modern information retrieval systems, enhancing both organization and accessibility.

4. Classaurus:

It is a specialized tool in knowledge organization. It combines the principles of a classification scheme with the relational features of a thesaurus. It is developed by G. Bhattacharyya. It represents a hybrid structure where the hierarchical notations of a classification system are integrated. A classaurus provides the benefits of both by allowing subjects to be located systematically while also linking them to related concepts through associative and equivalence indexing, cataloguing, and search refinement by guiding users from broader terms. In a library database, a Classaurus entry on “Education” may not only show its position within the broader classification of “social sciences” but also connect it to narrower terms like “Higher Education”, related terms such as “Pedagogy”, and equivalent terms like “Teaching”.

It is a system of terms having separate hierarchic schedules. They are-
1.Elementary Categories: Discipline, entity, property and action, together with their respective types, and their special modifiers.

2.Schedules for the common modifiers: form, time, environment and place. Each of the terms in this hierarchic schedules constituting the

systematic part which is supplemented by an alphabetical index of chain entries.

Classaurus used in the formulation of subject headings in general, and in particular language. For the construction of classaurus the POPSI language provides guidelines itself. A set of programs has been developed to construct a classaurus using as input.

5. Subject heading list as tools of organizing information resources:

Various information tools like catalogue, indexes and list of subject headings. It will discuss the concept of subject cataloguing. It will illustrate types of controlled indexing languages like list of subject headings and thesaurus. Information Retrieval is the activity of locating or obtaining or getting relevant information from a collection of information resources. A library is an information retrieval system in which we prepare and use different tools and techniques. Also it is a software to describe or represent the contents of the documents. These tools and techniques also help in retrieving the required information in libraries.

In libraries we acquire documents and then we analyse the contents of the documents. We use the notation or code to represent the contents of the documents. For example we use different kinds of classification schemes to represent the contents of the documents. We create a file or database of notation of contents of the documents. We organize and maintain the collection according to the notation. That means we use a certain scheme of classification for representation or for translating the contents of documents into a notation. After that we create an interface which connects the user to the collection of documents. This very interface helps the users to browse and know the contents of books and collections.

In order to organize knowledge librarians and information professionals have to create a variety of tools. They have to use a variety of tools. Traditionally the tools of information retrieval have been catalogues, bibliographies and printed indexes with the arrival, advent of information technologies. We have online databases and

these databases have indexes which help users and learners in retrieving the required information.

In a traditional library when a control vocabulary is set up in the form of an alphabetical listing of index terms the individual terms are known as subject headings. And the control vocabulary as a list of subject headings. The subject headings list is useful to understand the relationship among concepts to a certain degree besides their application in indexing.

The subject headings list is highly valuable for indexing. Subject headings are given in catalog entries to provide subject access to information. Catalogue depends on the list of subject headings from where they can assign subject headings to the catalogue documents. That means there are standard subject headings available which the cataloguers use to assign headings to the documents. The conceptual relationships are indicated in the list and choice of terms and the preferences are given. Recently these lists have also introduced many features of thesaurus.

For example:

- The Library of Congress Subject Headings List is one of the best tools for indexing and retrieval.
- The Sears list of Subject Headings.

References:

Chowdhury, G. G. (2010). *Introduction to modern information retrieval* (2nd ed.). Facet Publishing.

Lancaster, F.W., Information retrieval systems, New York, John Wiley, 1968.