International Journal of Information Movement Vol.3 Issue X (February 2019)
Website: www.ijim.in ISSN: 2456-0553 (online) Pages 1-5

BASIC CONCEPTS IN INFORMATION RETRIEVAL PROCESS

Bibina C. B.
Librarian
Marian College of Architecture and Planning,
Kazhakuttom-695582
Email: bibina.cb92@gmail.com[1]

Abstract

Information retrieval is a continuous process of selection of problem, appropriate keywords, information sources, retrieval techniques and finally evaluates the information retrieved. Due to the advances in research in the field of information technology results in the development of efficient retrieval techniques. By using these retrieval techniques, user can easily retrieve their needed information's effectively and efficiently. Information retrieval techniques can reduce the risk of information overload to a certain context.

Keywords: Information Retrieval Process, Basic Retrieval Techniques, Advanced Retrieval Techniques, Search Strategies, Evaluation of Information Retrieval System.

1.1 Introduction

In the modern world, information is available in different format, print as well as non print. Internet is the greatest source of information in the world. Information overload is the main issue in finding relevant information's from the abundant of information. Developments in the field of information technology results in the emergence of information retrieval techniques, so that relevant documents can be retrieved easily and effectively. The finding and recall of information from a store; earlier methods included comprehensive classification and cataloguing, and searching databases by various mechanical means (Prytherch, 2005).

1.2 Information Retrieval Process

Information retrieval (IR) is the activity of obtaining information system resources relevant to an information need from a collection of information resources (Information retrieval, 2019). It is a continuous process that can be divided into different stages.

1.2.1 Define a problem/ Research Question

It starts with defining a research question/a problem that need more information's required solve this problem, which leads to **analysis of the search topic, choosing appropriate information resources, retrieval techniques and define keywords, performing the planned information retrieval techniques, evaluating the search results by finding the relevant search result, and finally reaching an outcome.** During the information retrieval process, information seeker can consider, reconsider and refine the topic.

1.2.2 Planning

Choosing search terms is one of the most important stages of the information seeking process. An appropriate search term gives positive results. Search terms may be regular words (key words), coordinate terms or subject terms from thesauri. Thesauruses are helpful to find the coordinate terms and synonyms as well as foreign equivalents of terms.

Eg: 1. Merriam-Webster - Dictionary and Thesaurus
2. Eric Thesaurus - The Institute of Education Sciences thesaurus.

Sometimes search terms are divided into separate themes, synonyms, acronyms, equivalent terms in other languages, broader and narrower terms etc. If information on the topic can be found easily from many sources, one can narrowing down the topic to something more specified. For example,

- **Search Terms About Treatment:** Therapeutics, Remedy, Therapy
- **Synonyms:** Heart Health – Cardiovascular
- **Broader Term for Human Resource Management:** Management
- **Narrower Term for Human Resource Management:** Recruitment
- **Abbreviations**: AI- Artificial Intelligence
- **Equivalent Terms in Other Languages:** ΑΡΧΙΤΕΚΤΟΝΙΚΗ - Architecture

1.2.3 Retrieval Techniques

Due to the developments in information technology, a variety of retrieval techniques are available to users. By using retrieval techniques, users can easily retrieve their needed information's efficiently and effectively. Retrieval techniques are mainly classified into two groups:

- Basic retrieval techniques
- Advanced retrieval techniques.

1.2.3.1 Basic Retrieval Techniques:

Basic Retrieval Techniques are supported by most of the IR systems. These are:

a. Boolean Search

Boolean searching was named after the Englishman George Boole who conducted mathematical analysis of logic (Chu, 2003). Boolean logic is an essential search tools which combines multiple search terms by using Boolean operator such as AND, OR, NOT. The **AND** operator is used to retrieve results that contain all the search terms used (i.e., AND operator for narrowing down a search).

E.g.: Chemical Energy AND Thermal Energy

OR operator is commonly used to combine two or more similar terms (synonyms & related terms). It retrieve results contain all or any of the search terms (i.e., OR operator broadens the search).

E.g.: Thermal Power OR Hydel Power

NOT operator excludes the chosen search terms from the search result. It can be used to omit unwanted results. (i.e., NOT operator narrows the search results).

E.g.: Biofuels NOT Biogas

Boolean searches are two types: Simple Boolean search and Compound Boolean search. In a simple Boolean search, only one Boolean operator is used at a time, but compound Boolean search combines several operators by using parentheses to specify the order in which the search terms are interpreted. Terms within the parentheses are executed first. The natural order for processing the three Boolean operators in most IR systems is:

NOT>>AND>>OR (First>>Second>>and Third)

E.g. 1: *(green OR renewable OR sustainable) AND (resources OR energy OR power) AND (opinion OR view point).* Without the parentheses most databases will read terms combined by AND first, which will change the search result.

E.g. 2: *("transferable skills" NOT ("study skills" OR "presentation skills")) AND "research students"*

The retrieved documents will be transferable skills of research students excluding material that mentioned study skills or presentation skills. Multiple pairs of parentheses can be used to specify a particular order of processing in a compound Boolean search statement. So, compound Boolean searching is also called nested Boolean.

b. Case Sensitive Search

For languages such as English, French, Spanish upper and lower cases makes a difference. Case sensitive searching allows pinpointing exactly how a term is represented in a query and the system. For example...

Web - World Wide Web.
web - web woven by spiders (Bachchhav, 2016).

c. Truncation or Wild Card

Truncation is a retrieval technique that broadens the search result by including various word endings and spellings. The symbols used for truncation varies as databases varies. Most commonly used symbols/characters are *, ?, #. Commonly used truncation is right truncation (i.e., truncating the term suffix). In right truncation, enter the root of a word and put truncation symbol at the end.

E.g.: Use of the truncated term of *process** as a query results in retrieving documents on *processor, processing, processed, process, processes* and *processors.* Truncation can also be done by taking away the prefix is called left truncation.

E.g.: Use of the truncated term **rary* as a query results in retrieving documents on library, literary etc.

Taking away the infix of the term is called middle or internal truncation or wildcard. In internal truncation, truncation character substitutes one letter of a word.

Eg: Using the search term analy?e will include both analyze and analyse in the search result.

Simultaneously put truncated character on the left and right side of the truncated term broadens the search result.

Eg: *chemical* (Displays chemical analysis, chemical engineering, Sustainable chemical engineering etc)

Too much truncation retrieves a lot of unwanted information (e.g., truncate *Library* to *Lib**), whereas too little truncation will not achieve the goal for truncation (e.g., truncate *Library* to *Librar*).

d. Phrase Search

Phrase search is an another retrieval techniques for searching an exact phrase (two or more words in a specific order), by using quotation mark (" ") around the word. For Example, "Marketing Management". Without using quotation mark around the word, many databases search words individually, and the result will contain lot of irrelevant documents.

e. Field Searching

Field search is a commonly used search tool for narrowing the search by using fields such as author, title, publishing year, document type (journal article, book article, book etc), file type, full text, language, keywords, peer reviewed/full text etc.

f. Proximity Search

This search technique allows user to search two or more words or a phrase in relation to one another. It not only searches for the occurrence of two or more search terms but also specifies the distance between the search terms. Proximity searches use operators to designate how closely, and in what order, you want the search terms to appear (Northcentral University Library, 2019). Commonly used proximity operators are:

1. **Near (N):** Does not matter the order in which word appears

E.g.: *Construction N3 Management*, Finds construction within three words of management, in any order. I.e., it retrieve documents mentioning *Construction company project management*, *Construction cost management*, *Management jobs in construction*, *Management of construction* or other similar phrases.

2. **With (W):** Terms are in the exact order in which they are entered.

E.g.: *Construction W3 Management,* Finds construction within 3 words of Management. I.e., It retrieves *Construction of material management/Construction of technology management*, but not *management of materials in construction/Management of technology in construction* etc.

1.2.3.2 Advanced Retrieval Techniques

a. Fuzzy Search

Fuzzy searching and truncation are related to each other. But it occur some difference that, fuzzy searching finds terms that are spelled incorrectly at the time of entering a query, but truncation results different forms of a term. For example, Architect could be misspelled as architecht, arkitect, architekt or arkitecht. But it does not affect the outcome. Fuzzy searching can endure the mistake made at the time of data entry.

b. Weighted Search

Weighted search are very useful techniques for assigning weights (stress/importance) on each terms in a query. Weights can be given in the form of symbols (*) or numbers. Weighing scale is designed by the IR systems.

For example, in the search query *leadership2* AND *administration5, if* the user is more interested in the *administration* aspect rather than *leadership* .

1.2.3.3 Retrieval by Searching- Search Strategies

Search strategy is any method used to identify an object of interest (Search strategy, 2009). For conducting searches, following search strategies are used:

a. The Building Block Approach

This is the most widely used online search strategy. The building block approach starts with single concept searches by entering each concept individually.

E.g.: *Socialization or Social development or Social skill.*

After completing all single concept searches, try different concept combination by using AND connector (Boolean Operator) to produce a document set relevant to the problem as a whole. This strategy is more popular because it simplifies a complex search task into manageable parts. It also allows the user to modify the searches if needed.

b. *The Successive Fraction Approach*

In this approach, searches starts with a broad concept, then narrows the search by applying various' limiting techniques such as Boolean Operators (AND, NOT), Proximity searches (WITH), Field Searching (by using subject attributes).

c. *Pearl Growing Approach*

This method begins with a specific document or document set that is known to be relevant (a pearl) and uses the characteristics of that document to successively grow a set of related (and presumably similarly relevant) documents (Marchionini, 1997). This approach is also called snowball approach. This search strategy starts with a well defined and solid entry. Compared with building blocks and successive fraction approaches, it is less algorithmic and more interactive with the system.

d. *Interactive Scanning Search*

In this approach, searcher conducts a broad search including thesaurus terms and/or codes with the intention of obtaining a broad picture of the literature in the area. (Kent, 1994). This strategy requires continuous cognitive attention, changing criteria for judging relevance as the problem and its associated literature is better understood, and plausible reasoning about when to terminate search (Marchionini, 1997).

e. *The Least Specific Facet First Approach*

The least specific aspect of the subject is searched first. Searches continue by entering broader the search terms, if insufficient documents are retrieved.

1.2.4 Evaluation of an information retrieval system

Evaluation of information retrieval system assesses the level of performance of an existing information system and **identifies the most relevant sources of information from the abundant information's. Basic parameters for accessing quality of information retrieval are: Recall, Precision, Fallout, and Generality.**

Recall is defined as the ratio of number of relevant items retrieved to the total number of relevant items in the collection. It is the ability of a systems to retrieve all useful items. Recall calculation formula is

$$\text{Recall (R)}=\frac{\text{Number of relevant item retrieved}}{\text{Total number of relevant items in the collection}} \text{ x } 100,$$

$$\text{Recall}=\frac{a}{(a+c)} \text{ x } 100$$

Precision is defined as the proportion of relevant documents retrieved. It is the ability of a system to omit unwanted documents. The formula for calculation of precision is:

$$\text{Precision (P)}=\frac{\text{Number of relevant item retrieved}}{\text{Total number of items retrieved}} \text{ x } 100,$$

$$\text{Precision}=\frac{a}{(a+b)} \text{ x } 100$$

	Relevant	**Not-relevant**
Retrieved	a (suggested)	b (noise)
Not-retrieved	c (lose)	d (not accepted)

Table 1: Recall-Precision Matrix

Fallout ratio is the proportion of number of non-relevant items retrieved to the total number of non-relevant items in the collection.

$$\text{Fallout ratio (F)}=\frac{\text{Number of non relevant items retrieved}}{\text{total number of non}-\text{relevant items in the collection}} \text{ x } 100,$$

$$\text{Fallout ratio}=\frac{b}{(b+d)} \text{ x } 100$$

Generality is the ratio of number of relevant items retrieved to the total number of items in the collection. Generality number is directly proportional to the density of relevant items in total collection.

$$\text{Generality ratio (G)}=\frac{\text{Number of relevant items retrieved}}{\text{Total number of items in the collection}} \text{ x } 100$$

$$\text{Generality ratio}=\frac{(a+c)}{(a+b+c+d)} \text{ x } 100$$

1.3 Conclusion

In a nutshell, information retrieval is a continuous process of searching and retrieving information's from large collection of documents. It start when user input a query into the systems. User must need **some basic skills of information retrieval techniques while seeking information from structured databases.** Information's retrieved from the internet sources does not guarantee the reliability, rationality or quality. So, user must critically evaluate the information retrieved from the internet.

References

1. Bachchhav, K. P.(2016). Information retrieval: Search process, techniques and strategies. *International Journal of Next Generation Library and Technologies, 2*(1), 7. Retrieved from www.ijnglt.com
2. Chu, H. (2003). Retrieval techniques and query representation. In *Information Representation and Retrieval in the Digital Age (p.59).* United Kingdom: Information Today.
3. *Information Retrieval.* (2019). Retrieved from https://en.wikipedia.org
4. Kent, A. (1994). Educating and users of remote online system. In *Encyclopedia of Library and Information Science* (Vol.54, p.164). United State: CRC Press
5. Lancaster, F W. (1979). Information retrieval systems: characteristics, testing,
6. *and evaluation* (2nd ed.). New York: John Wiley
7. Marchionini, G. (1997). *Information Seeking in Electronic Environments*. New York: Cambridge University. Retrieved from https://ils.unc.edu/~march/isee_book
8. Northcentral University Library. (2019). *Proximity searching.* Retrieved from https://ncu.libguides.com/researchprocess
9. Prytherch, R. (2005). *Harrod's librarians' glossary and reference book: A directory of*
10. *over 10,200 terms, organizations, projects and acronyms in the areas of information management, library science, publishing and archive management* (10th ed.). England: Ashgatc Publishing, 352.
11. Search strategy. (2009). In *Medical Dictionary*. Retrieved from https://medical-
12. dictionary.thefreedictionary.com