# On-AIR based Information Retrieval System for Semi-Structure Data

**Srujan K[1], Radha K[2]**

[1]*III.BTECH-CSE-Rudraram, GITAM UNIVERSITY Hyderabad, Telanagana, India*
[2]*Asst Professor, CSE, Rudraram, GITAM UNIVERSITY, Hyderabad, Telangana, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract** – *Retrieving of information has been the fundamental process to any storage file. With advent of technology, it is possible to store a large amount of data and retrieving them has become tedious task for the incorporators and users. Information retrieval systems emerged from the demand minimize of human resources required in the finding of needed information to accomplish a task. This paper talks about On-AIR architecture for retrieving data from video containing databases using Information Retrieval Systems ,Information Retrieval Systems Capabilities like search and browse capabilities, Functional Overview of IRS and Automatic Indexing that is performed on data items. This paper presents the overview of an Information retrieval system, Information Retrieval Systems Capabilities, Functional Overview of IRS and Automatic Indexing.*

***Key Words***: ***On****-AIR, Automatic Indexing, IRS, Precision, Recall*

## 1. INTRODUCTION

An Information Retrieval System is basically used to store, maintain and retrieve the information. Even though the information is diverse, the text aspect is the only data type that has been undergoing full functional processing [1]. New Techniques are being invented to search the other multimedia data types like ExCaliber's Visual Retrieval Ware which uses natural language processing and semantic networks for processing the information [13].

### 1.1. OBJECTIVES AND PERFORMANCE DMEASURES FOR INFORMATION RETRIEVAL SYSTEM

The main objective of IRS would be to reduce its overhead i.e., the time taken from leading to reading the required information. Distinguishing between a relevant and irrelevant can be the challenging aspect of IRS which might be due to problems like synonyms and polysemy[4], which is a word with multiple meanings. To evaluate the performance of search engines, measures like precision, recall and response time can be taken into consideration. Precision is the ratio between the number of relevant documents retrieved through a query to the total number of documents retrieved with help of a query. Recall is the ratio between the numbers of documents retrieved to the total number of documents that are relevant with respect to a query. Response time is the elapsed time gap after the submission of a query and presentation of documents retrieved by the software [4].

$$\text{Precision} = \frac{\text{Number of Relevant documents retrieved}}{\text{Total number of documents retrieved}}$$

$$\text{Recall} = \frac{\text{Number of documents retrieved}}{\text{Total number of documents that are retrieved}}$$

Precision can be optimized by retrieving a certainly relevant document and recall by retrieving all documents in the database. Therefore, F-square combines both precision and recall and results in the harmonic mean of precision and recall[2].

$$F - Square = 2\,\frac{Recall * Precision}{Recall + Precision}$$

*Example 2.1:*

*Assume the following:*

- *A database contains 100 records regarding a topic*
- *50 were recorded as a result of search.*
- *Of the 50 records retrieved, 45 were relevant.*

*Using the designations above:*

- *X =The number of relevant records retrieved*
- *Y =The number of relevant records not retrieved,*
- *Z =The number of irrelevant records retrieved.*

*In this example X = 45,Y = 65 (100-45) and Z = 5 (50-45).*

*Recall = (45 / (45 + 65)) * 100% => 45/110 * 100% = 40.909%*

*Precision = (45 / (45 + 5)) * 100% => 45/50 * 100% = 90%*

## 2. INFORMATION RETRIEVAL SYSTEMS FUNCTIONALITY

Information retrieval and storage system basically contains of four different functional processes, Item normalization, Selective Dissemination, Document Database data search, Index Database search[1].

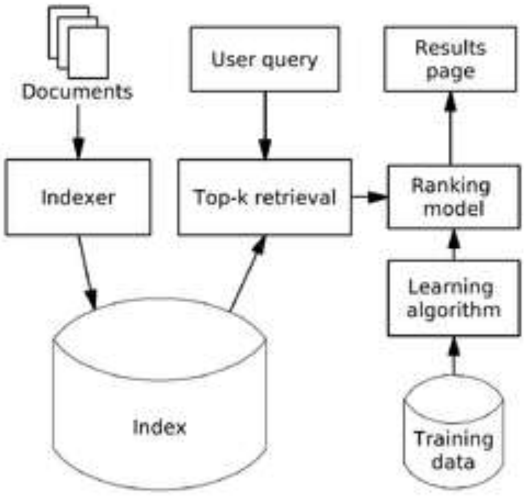


**Fig-1**. Functional Overview of Information Retrieval Systems

### 2.1. Item Normalization

This is the first step before storing in the database. All the information in different format is converted into a single standard format to avoid ambiguity between information and to provide logical restructuring. The process of transforming multiple formats into a single data structure that can be easily processed is termed as item normalization. This process includes identification of process tokens, characterization of tokens, Stemming of tokens, and then transforming into searchable data structure.

### 2.2. Selective Dissemination

Item is processed with help of profile. It gives the capability to compare newly found items with the interested statements and results the documents matching the interest dynamically. If there is a profile match, the item is placed mail file with the profile after processing against every user profile.

### 2.3. Document Database Search

Query is processed along with items which are prior processed. The items are searched in Document Database and can take longer time periods.

### 2.4. Index Database Search

Saving a file for any future references, this is done through indexing process.

## 3. RETRIEVING DIGITAL VIDEO DATABASES USING ON-AIR

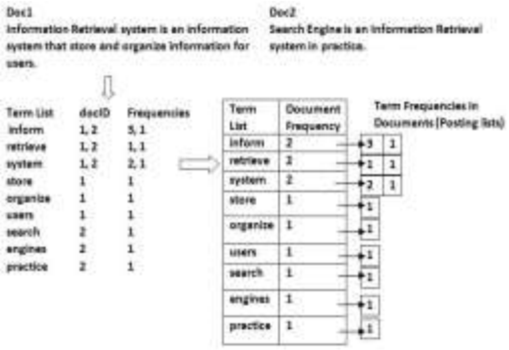
Ontologies are used in Information retrieval system to search for a related term and improve precision and recall. Ontology is developed priory defining the terms and their corresponding relations. Ontology aided information retrieval is used to retrieve data in digital video databases using domain ontology .It consists of mainly two processes indexing and retrieval.

### 3.1. Indexing process

Indexing process is responsible for creating searchable data structure. Initially this process consists of Video collection and Domain ontology. Video collection consists of clips of videos of a long video assigned with keywords, set of images and transcription of speech. Domain Ontology contains the information for query expansion. Indexing process is performed by registering the video clips, the ontology and establishing the configuration values.

This process produces three outputs

- **XML Configuration file:** Contains the information about the associations between clips, transcriptions, keywords, resources etc.
- ***Inverted Index:*** An Inverted Index data structure is used to store the frequencies of word occurrences along the word itself. It also stores weights of terms to contribute to precision and recall .Stemming is performed and stop words are excluded.



Doc1
Information Retrieval system is an information system that store and organize information for users.

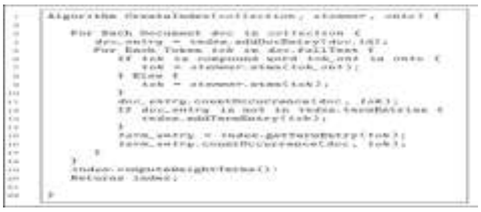
Doc2
Search Engine is an information Retrieval system in practice.

| Term List | docID | Frequencies |
|---|---|---|
| inform | 1, 2 | 3, 1 |
| retrieve | 1, 2 | 1, 1 |
| system | 1, 2 | 2, 1 |
| store | 1 | 1 |
| organize | 1 | 1 |
| users | 1 | 1 |
| search | 2 | 1 |
| engines | 2 | 1 |
| practice | 2 | 1 |

| Term List | Document Frequency | Term Frequencies in Documents (Posting lists) | |
|---|---|---|---|
| inform | 2 | 3 | 1 |
| retrieve | 2 | 1 | 1 |
| system | 2 | 2 | 1 |
| store | 1 | 1 | |
| organize | 1 | 1 | |
| users | 1 | 1 | |
| search | 1 | 1 | |
| engines | 1 | 1 | |
| practice | 1 | 1 | |

**Fig-2**. Diagram representing a Inverted file structure.

Figure-2 represents how tokens are stored in inverted data structure for the document 1 containing "*I did enact Julius Caesar: I was killed i' the capitol; Brutus killed me*" and document 2 containing *"So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious"*[5]. Initially the frequency of each term is tabulated and converted into inverted data structure containing posting lists .The following algorithm is a prototype of generating index file [6]. In the following algorithm, initially the frequencies of each word in the document is calculated and weights are calculated using Term Frequency-Inverse Document Frequency before stemming is performed.



```
Algorithm CreateIndex(collection, stemmer, onto) {

    For Each Document doc in collection {
        doc_entry = index.addDocEntry(doc.id);
        For Each Token tok in doc.fullText {
            If tok is compound word tok_ont in onto {
                tok = stemmer.stem(tok_ont);
            } Else {
                tok = stemmer.stem(tok);
            }
            doc_entry.countOccurrence(doc, tok);
            If doc_entry is not in index.termEntries {
                index.addTermEntry(tok);
            }
            term_entry = index.getTermEntry(tok);
            term_entry.countOccurrence(doc, tok);
        }
    }
    index.computeWeightTerms();
    Return index;

}
```

**Fig-3**. Algorithm representing Creation of index in Index File.

- ***Ontology Data Structure:*** This helps in enhancing processing time. All the information is represented in a graph with the help of classes, sub-classes, instances, relationships etc.

- ***Retrieval process:***

Using the information obtained in inverted file and ontology data structure, Visualizer is used to display the relevant information in the form of list. The process performed by a Visualizer is shown in Figure-4.

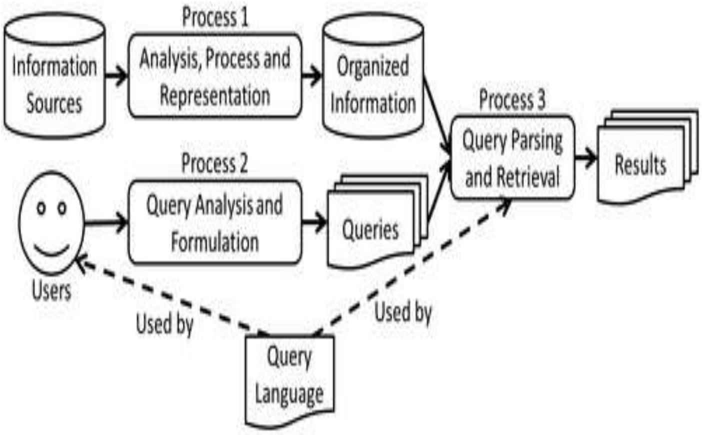


**Fig-4.**Retrieval Process

Pre-Process: In this stage, the misspelling detection is applied and also stop words are removed on encounter. Stemming process is applied and weightsare calculated.

*Query-Expansion:* The degree of similarity is computed in Query Expansion process[7].

*Retrieval:* In this stage, the query is compared to video clip contents and information of maximum importance is returned. Vector space model is used for retrieval[9].

Video player: This software returns all the relevant videos and the user can choose upon his need. Searching for the information in long-duration videos is time consuming process. Currently research is going on in Ontologies which is used to retrieve the information from the web. Through the digital videos we can retrieve the information efficiently [14]. Most of the organizations are producing the massive volumes of data in a semi-structured format only that is in PDF formats, Reports and Tabular extraction of data. To enable the SQL-Like query and analysis of financial tables from annual reports in pdf format, deep learning based technique is used. For the nearest row match table type classification can be done. To query and analyse the financial tables in pdf documents from distinct sources we will use text-match based approach. Similarity measure for query framework returns the similar rows and associated columns via different documents for given table row as a query. Queries can be given in two ways such as Querying table and querying by row/column. Query table permits the users to select a table then similar tables are returned and finally these queries can be refined using the range operators. Deep learning architecture for table classification has been divided into Data collection and pre-processing. Tokenization and Classification. A technique and prototype implementation is presented for deep-learning based table processing pipeline architecture. It provides access to similar table data across various PDF files [15].

***Example of IR Process***

Consider a query specified by the user.

"How do you make money online??"

Initially, stop words "how", 'do", "make" are separated and "money" and "online" are given weights according to the ontology defined. Depending upon the ontology the relevant data from the inverted index is retrieved and displayed.

## 4. INFORMATION RETRIEVAL SYSTEM CAPABILITIES

### *4.1. Search Capabilities:*

The Searching capabilities include mapping the user required query to the information present in database. "Weighting" has been used to rank and procure information in commercial systems[7].

Many functions are been defined to determine the relationships between search statements

- Boolean Logic: Multiple concepts can be logically related by using Boolean logic. The Boolean operators like AND, OR, NOT are used to logically compute intersection, union and negation respectively.

- Proximity: Proximity is used to restrict the distance between two search terms within a query. When a certain proximity parameter is crossed between two search items, they are ignored.

- Contiguous word phrase: When two or more units are considered as single unit then they are called as contiguous word phrase. "United Arab Emirates" is a contiguous word phrase.

- Term Masking :

  Term Masking is the process of masking a certain portion of term and matching it with the unmasked portion of the term. Stemming can be used as the size is huge .It can be either fixed length masking or variable length masking.

- Concept/Thesaurus expansion:

  Thesaurus function is used to group a term along with its term which has a similar meaning. Concept tree is used to group similar items in the form of tree structure [1].Figure-5 depicts concept expansion of Orchestra, In Orchestra different kind of instruments are used like woodwinds, brass, percussion, strings and keyboard. Further woodwinds, brass are subdivided into their respective instruments [6].

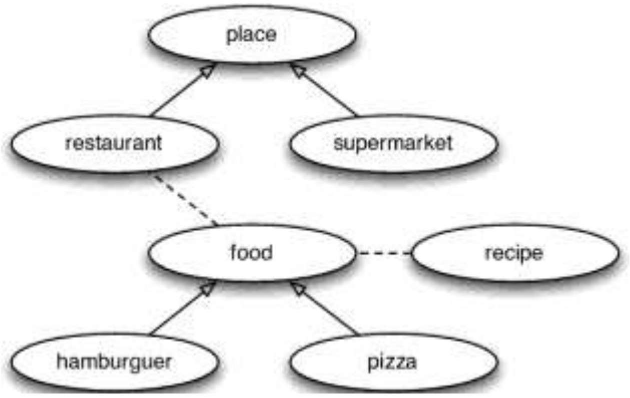


**Figure-**5:Concept Expansion

### 4.2. Browsing Capabilities:

The jargon associated with browsing are ranking, zoning, highlighting. **Ranking** is the process where different tokens of the query are provided with values on their relevance. **Zoning** is the process of dividing the standardized input into meaningful divisions. **Highlighting** is used to grab the user's attention to particular parts.

## 5. INDEXING PROCESS IN INFORMATION RETRIEVAL SYSTEMS

The indexing process consists of three basic steps: defining the data source, transforming document content to generate a logical view, and building an index of the text on the logical view[10] .In order, to retrieve information, the received item must be converted into a data structure that can be searched. Indexing process can be automatic or manual. Search can be either direct search in which information in present in a simple document or it can be indirect i.e., with the help of index files. Some systems instead of creating data structure, the item is completely transformed into a different representation which is used as searchable data structure.

### *5.1. Automatic Indexing*

In automatic indexing process, system determines the index terms and assigns them automatically. Automatic indexing can be complicated when the determining the limited number of index terms from the topic compared to manual indexing [1]. Automatic indexing would result in two classes of indexes. They are weighted and un-weighted. In weighted systems, a value is placed on index term and its related concept in the document. The weight is dependent on the function which is responsible for counting the frequency of occurring in each item. In case of un-weighted systems, the queries are dependent on Boolean logic and occurrences in hit file which is considered as value. Techniques of Automatic Indexing:

### *5.2. Indexing by term:*

When the basis of indexing is the term that belongs to the original item, the index can be created by two techniques.

### 5.3. *Statistical Techniques*:

Statistical techniques can be models **like vector models**, **probabilistic model** or **Bayesian model**[1].

All these models are based on statistical information like frequency and their distribution. Weights are calculated considering frequency and distributions in the database. The weighted systems are also called as Vectorized information systems. In vectorized information system, the weights are stored in the form of vectors. All the queries can be converted into vector forms and the distance between query vector and information present in data base vector is calculated. In probabilistic model, Bayesian product is calculated with the help of Bayesian network. Bayesian network can be considered as an acyclic graph that is directed where the random variables are defined in the form of nodes and the probabilistic dependency among the node and its parents are denoted by the arcs between the nodes.

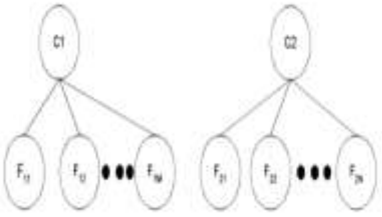


**Figure-6:** Two level Bayesian Network

Natural language processing is another method of defining indexes to terms. The DR LINK systems process the items at different levels like morphological, semantic, lexical, syntactic and discourse. DR LINK stands for document retrieval through linguistic knowledge .Each level in DR LINK uses the information of previous level to perform analysis of the present level.

### *5.4. Indexing by Concept*

There are numerous ways of expressing an idea out of which only one can give the enhanced retrieval performance; this is the basis of Indexing by Concept. The Concept indexing will determine a collection of concepts. The concepts are determined based on the test set terms and these tests are used as basis for indexing the items.

## 6. CONCLUSION

Information retrieval system has the high potential to revolutionize the present and conventional storing and retrieving techniques. Ten years ago, the algorithms developed were restricted in scope allowing theoreticians to limit their focus to very specific areas. But the insertion of technology in systems like Internet has changed the way problems were bounded. Hence, efficient algorithms must be constructed in order to handle enormous data and minimal user search statement information must be considered along with optimal usage of functional aspects of Information Retrieval system.

## REFERENCES

1. Gerald J Kowalski, Mark T Maybury: Information Storage and Retrieval Systems, Theory and Implementation.

2. Paz-Trillo, C., Wassermann, R., and Braga, P: An Information Retrieval Application using Ontologies (2005).

3. R. Baeza-Yates and B. Ribeiro-Neto.Modern Information Retrieval. Addison Wesley Longman, 1999.

4. M. Mauldin. Conceptual Information Retrieval: A case study in adaptive partial parsing. Kluwer Academic Publishers, 1991.

5. Stanford Home Page: https://nlp.stanford.edu/IR-book/html/htmledition/a-first-take-at-building-an-inverted-index-1.html

6.ResearcherGate:https://www.researchgate.net/figure/CYC-assisted-concept-enpansion-tree-An-example-of-demonstrating-concept-expansion-is_fig1_257728125

7. Information Retrieval: search process, techniques and strategies KiranPrakashBachchhav Librarian, SarvajanikShikshanSanstha'sAdv.V.B.Deshpande College of Commerce, Mulund (West), Mumbai- 400080

8.Berkeley.edu:http://people.ischool.berkeley.edu/~buckland/papers/analysis/node1.html

9. M. Smith, C. Welthy, and D. McGuiness.OWL Web Ontology Language Guide. Technical report, World Wide Web Consortium, 2004. http://www.w3.org/TR/2004/ REC-owl-guide-20040210/

10. Indexing:https://www.springer.com

11. J.F. Sequeda, D.P. Miranker, A pay-as-you-go methodology for ontology-based data access, IEEE Internet Comput. 21 (2) (2017) 92–96.

12. M. Rodriguez-Muro, R. Kontchakov, M. Zakharyaschev, Ontology-based data access: ontop of databases, in: International Semantic Web Conference, Springer, 2013, pp. 558–573.

13. Vise, David A. (2004-12-03). "Agencies Find What They're Looking For". The Washington Post. Retrieved 2010-05-22.

14. Christian Paz-Trillo, et al,"Using ontologies to retrieve video information".

15. Rahul Anand, et al, "Integrating and querying similar tables from PDF documents using deep learning", 15 jan 2019.

## BIOGRAPHIES

**First Author**: K. Srujan Kumar, currently pursuing III B.Tech, CSE at GITAM UNIVERSITY, HYDERABAD. My Research areas are Machine Learning, Predictive Analytics.

**Second Author:** K Radha working as an Asst Professor at GITAM University, Hyderabad. She has Completed MTech (CSE) at JNTUH, Pursuing PhD at KL University, Vijayawada. She has 12 years of Teaching Experience and 1 Year Industrial Experience. She has published numerous research papers and presented at various conferences. She is a Member of IAENG.