CHAPTER -4

INFORMATION RETRIEVAL

4.1 INTRODUCTION

The Information Retrieval is one of the tasks of Text Mining. Information Retrieval is identifying unstructured documents from the huge collection of documents which satisfies the given stipulation and is an interdisciplinary task based on Computer Science, Mathematics, Information Science etc..

Information Retrieval is about recognizing a subset of document from the corpus. The content of subset document is most relevant to the users’ need. The users’ requirement will be specified in a query. The Information Retrieval system retrieves the relevant document from the huge collection of document based on the users’ query.

An Information Retrieval process starts when a user inputs the need in terms of query in to the Information Retrieval System. The queries are formal statement which describes the need of the end users. The Information Retrieval system retrieve the documents in different perspectives of relevance.

The Information Retrieval (IR) system consists of three modules:

- Document Acquisition Module.
- Document Structuring Module
- Validation of Relevance
4.2 ARCHITECTURE OF INFORMATION RETRIEVAL SYSTEM

The task of the Information Retrieval system is to retrieve the subset of documents from the collection of documents based on the criteria specified by the end users. Information Retrieval is the Initial process of Information Extraction. The study focus on biomedical documents for Information Extraction. The biomedical documents from the biomedical databases such as Pubmed, Medline etc., which are under the (repository of) National Center for Biotechnology Information (NCBI). The major task of Information Retrieval is to generate documents based on the criteria specified by the end users in the user requirement specification from the corpus of documents.

![Information Retrieval System Diagram]

Fig :4.2 Information Retrieval System
4.2 Describes the overall structure of the Information Retrieval system.

4.3 DATA SET

To extract the biomedical named entities and relations, the biomedical documents are needed as a data set. The biomedical documents will be available in various biomedical databases like NCBI[56]. Based on the criteria given by end users, the information retrieval system will retrieve the documents from the database and produces the corpus.

The National Center for Biotechnology Information (NCBI) comes under the United States National Library of Medicine (NLM). It is a branch of the National Institutes of Health. The NCBI accommodates a series of databases relevant to biotechnology and biomedicine and is a vital resource for bioinformatics tools and services. Key databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. Other databases consist of the NCBI Epigenomics database.

PubMed is a free search engine mainly accessing the MEDLINE database of references and abstracts on life sciences and biomedical topics. The United States National Library of Medicine (NLM) at the National Institutes of Health maintains the database as part of the Entrez system of information retrieval.
MEDLINE (Medical Literature Analysis and Retrieval System Online, or MEDLARS Online) is a bibliographic database of life sciences and biomedical information. It consists of bibliographic information for articles from academic journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, and health care. MEDLINE also covers much of the literature in biology and biochemistry and also molecular evolutions.

More than 5,500 biomedical journals are indexed in MEDLINE. Selection is based on the recommendations of a panel, the Literature Selection Technical Review Committee, based on scientific scope and quality of a journal. The Journals Database holds information, such as its name abbreviation and publisher, about all journals.

4.4 USER REQUIREMENTS

The initial process of Information Retrieval is getting the requirement from the end users. The end users specifies their requirements in conditions of terms in the user interface design created for getting the user requirements. The Terms includes main information's for query such as keywords, Organisms. Query properties such as Authors details, Journal details, journal type, Publication year and date. The repository information is about the databases weather the documents are from Pubmed Central or MedLine or from both databases. And the article detail includes information's about the
availability of articles whether the articles are only available abstracts, Open Access articles or Available Full Text. The user requirements are specified based on the terms specified in the user interface design. The Fig 4.4 illustrates the user interface design for the information retrieval. IR System converts the requirements as Boolean Query Using Jquery.

**4.4.1 JQuery**

JQuery is a rich featured java script library. It is a DOM (Document Object Model) handling library. JQuery can be used for finding an element in the document with particular properties.

![Fig : 4.4 User interface design](image)
4.5 DOCUMENT ACQUISITION SYSTEM

The requirements gathered using the user interface design are given to the JQuery in terms of query to the Document Acquisition System for the further process. The aim of the Document Acquisition System is to generate a query based on the requirements identified and using that query to retrieve the desired documents from the biomedical databases to generate the corpus. The Document Acquisition System is having various sub system to complete the task. The sub systems are (i) IR System, (ii) eUtils System and (iii) corpus generation System. The Fig: 4.5 illustrates the Document Acquisition System.

![Document Acquisition System Diagram]

Fig: 4.5 Document Acquisition System
The Document Acquisition System initiates the process with the user requirements. The identified user requirements are given to the IR System. The IR system converts the given requirements into Boolean queries.

The *Boolean retrieval model* is for information retrieval in which anyone can create any query which is in the structure of a Boolean expression of terms. The terms can be combined with the operators AND, OR, and NOT. The Boolean model views every document as a set of words.

A document collection consists of numerous documents that include information on a range of areas or subjects of interests[14]. Queries convert the user's information into a structure that properly represents the user's fundamental information requirement and is appropriate for the corresponding procedure. Query structure depends on the primary model of retrieval.

Boolean query generated by the IR System is given to the Enterz programming Utilities[26]. Based on the preference given by the user in the user requirement, the “Enterz Utility System” will retrieve the biomedical documents.

A Set of nine server side programs which is known as Entrez Programming Utilities (E-utilities) provides a constant interface to
Entrez query and database system. E-utilities and its services are available in the National Center for Biotechnology Information (NCBI).

The E-utilities employ a standard URL syntax that interprets a standard set of input parameter into the measures necessary for various NCBI software modules to search and retrieve the needed data.

The E-utilities are the ordered interface to the Entrez system, which is at present includes 38 databases covering a range of biomedical data, including nucleotide and protein sequences, gene records, three-dimensional molecular structures, and the biomedical literature[17]. Table 4.6 exhibits in detail about the nine E-utilities.

Server side programs.

4.5.1 E-UTILITIES AND ENTREZ DATABASES

The core search engine of the Entrez system is well been accessed by E-utilities. The E-utilities can able to access all the databases available in Entrez. The majority of the databases of NCBI is available in Entrez.

4.5.1.1 ENTREZ CORE ENGINE

The Entrez core engine consist of three major components viz: EGQuery, ESearch, and ESummary. The two fundamental functionalities of Entrez databases are: (i) collect a set of UIDs that match a text query and (ii) Retrieve the document summary for each UID. These two fundamental tasks are taken care by ESearch and ESummary. ESearch returns the list of UIds and ESummary returns
the documents summary. EGQuery is similar to ESearch. It searches all
Entrez databases concurrently. These three E-utilities performs the core
Entrez functions.

<table>
<thead>
<tr>
<th>E-utilities Server side programs</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>EInfo (database statistics)</td>
<td>Provides statistical information like number of records in the database, the last updated date of the database.</td>
</tr>
<tr>
<td>ESearch (text searches)</td>
<td>Reacts to a text query with the set of matching UIDs in a given database with the term translations of the query</td>
</tr>
<tr>
<td>EPost (UID uploads)</td>
<td>Recognizes a set of UIDs from a given database and stores the list on the History Server. Responds with a query key and web location for the uploaded dataset.</td>
</tr>
<tr>
<td>ESummary (document summary downloads)</td>
<td>Responds to a set of UIDs from a given database with the exact document summaries.</td>
</tr>
<tr>
<td>EFetch (data record downloads)</td>
<td>Responds to a set of UIDs in a given database with the equivalent data records in a specified format.</td>
</tr>
<tr>
<td>ELink (Entrez links)</td>
<td>Responds to a set of UIDs in a given database with the list of related UIDs in the same database or a list of linked UIDs in a different Entrez database.</td>
</tr>
</tbody>
</table>
Tab 4.6 E-utilities Server side programs

<table>
<thead>
<tr>
<th><strong>EGQuery (global query)</strong></th>
<th>Responds to a text query with the number of records corresponding the query in every Entrez database.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ESpell (spelling suggestions)</strong></td>
<td>Retrieves spelling suggestions for a text query in a given database.</td>
</tr>
<tr>
<td><strong>ECitMatch (batch citation searching in PubMed)</strong></td>
<td>Retrieves PubMed IDs (PMIDs) corresponding to a set of input citation strings.</td>
</tr>
</tbody>
</table>

4.6 CORPUS OF RETRIEVED DOCUMENTS

A Corpus is a set of documents that could be annotated with entities/events in IE Processes

4.7 DOCUMENT STRUCTURING

Most of the documents will be available in PDF. PDF files need to be converted to a format that can be utilized by NLP modules and the Document Structuring Module is responsible for converting the document from PDF to Text.

Plain ASCII text is considered the most appropriate system, but this translation involves various information losses. Since present PDF-to-text converters are not aware of the typesetting of each journal, two-column text, footnotes, headers, footers, figures and tables’ captions tend to be isolated and mixed up throughout the conversion. Through the system uses the biomedical documents there will be terminology-
related issues like conversion of Greek letters, superscripts and subscripts, hyphenation and italics. After analysing all existing PDF conversion tools like Optical Character Recognition (OCR), the system includes PDFBox which is part of the Xpdf software.

The Apache PDFBox library is an open source Java tool for working with PDF documents. The Apache PDFBox is involved in converting the PDF to the documents to a text format which can be understood by natural language Processing Module.

![Fig 4.7 Document Structuring](image)

4.8 VALIDATION OF RELEVANCE

The documents have been retrieved based on the requirements specified by the end users. Among the retrieved documents there may be relevant and irrelevant documents towards the requirement specifications. The retrieved documents need to be validated to identify
the relevance. The relevance of the retrieved documents are validated using the **Document – Weight Matrix**.

![Diagram of Corpus with relevant and irrelevant documents](image)

**Fig: 4.8 Corpus with relevant and irrelevant documents**

### 4.8.1 DOCUMENT WEIGHT MATRIX

A matrix that illustrates the frequency of terms that occur in a set of documents is known as **document-term matrix** or **term-document matrix**. In that matrix, rows correspond to documents in the collection and columns correspond to terms. There are various methods for examine the value that each entry in the matrix should take. One such scheme is tf-idf. It is valuable in the field of natural language processing.

A document collection (corpus) composed of \( n \) documents that are indexed by \( t \) terms can be used to represent an \( n \times t \) **document-term**
matrix. An entry in matrix corresponds to the “Weight of the term in the document”. Each document and term assigned a weight $W_{n,t}$.

$$
\begin{pmatrix}
T_1 & T_2 & T_3 & \ldots & T_t \\
D_1 & W_{11} & W_{21} & W_{31} & \ldots & W_{tl} \\
D_2 & W_{12} & W_{22} & W_{32} & \ldots & W_{t2} \\
D_3 & W_{13} & W_{23} & W_{33} & \ldots & W_{t3} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
D_n & W_{1n} & W_{2n} & W_{3n} & \ldots & W_{tn}
\end{pmatrix}
$$

**Fig: 4.9.1 Document-term matrix**

The weight is a statistical measure used to evaluate how important a term in to a document in a corpus. The importance increases proportionally to the number of times a word appears in the document. The Weight is calculated using the **Tf-idf** (*Term frequency-inverse document frequency***).

### 4.8.2 TF-idf

$$Tf-idf = tf \times idf$$

The tf-idf weight is composed by two terms:

- **Term Frequency (TF)**
- **Inverse Document Frequency (IDF)**
4.8.2.1 TERM FREQUENCY (TF)

The number of times a word appears in a document, divided by the total number of words in that document.

\[
TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}
\]

4.8.2.2 INVERSE DOCUMENT FREQUENCY (IDF)

The logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears.

\[
IDF(t) = \log_e\left(\frac{\text{Total number of documents}}{\text{Number of documents with term } t \text{ in it}}\right)
\]

- Consider a document containing 100 words wherein the word 'cancer' appears 3 times.
  
  term frequency (tf) = \( \frac{3}{100} = 0.03 \)

- Now, assume 10 million documents and the word 'cancer' appears in one thousand times.
  
  inverse document frequency (idf) = \( \log(10,000,000 / 1,000) = 4 \)

  tf-idf weight = 0.03 * 4 = 0.12.

To assess the effectiveness of an IR system, need to know two key statistics about the system’s returned results for a query and they are known to be PRECISION and RECALL.
**PRECISION**

What fractions of the returned results are relevant to the information needed?

**RECALL**

What fractions of the relevant documents in the collection were returned by the system?

---

![Precision-Recall Matrix](image)

**Fig: 4.9.2 Precision - Recall**

Precision: The ability to retrieve *top-ranked documents that are mostly relevant*. Recall: The ability of the search to find *all of the relevant items in the corpus*.