Information Retrieval Navigation System For Knowledge Discovery From Biomed Articles

K.Prabavathy, Dr.P.Sumathi
1ResearchScholar in ComputerScience,Manonmanium Sundaranar University,Tirunelveli Town TamilNadu,India, praba_bud@yahoo.co.in
2Doctoral Research Supervisor &Assistant Professor, Post Graduate & Research Department of Computer Science Government Arts College, Coimbatore-18,TamilNadu, India, sumathirajes@hotmail.com

ABSTRACT: The terminology used in Biomedicine shows lexical peculiarities that have required the elaboration of terminological resources and Information Retrieval systems with specific functionalities. Our discovery process enables the scientists to generate novel hypothesis to address the most crucial questions for understanding the knowledge basis from biomed articles or documents. However, formulation of a flexible and general approach for retrieving required information and integrating heterogeneous data and knowledge sources for discovery is elusive and highly dependent upon the specific underlying scientific question. The target of Information Retreival has been taken to mean in our work to interpret the knowledge discovery from Biomed articles. Our work has been framed with base of Keyword search, Information Retrieval and Information Extraction to bound the knowledge of articles in databases like PubMed, Oxford journals etc., Thus, the true impact of text and data mining is only realized if it goes beyond the methods for Information Retrieval into enable and cram of the entities like Protein Protein Interaction, Gene and Human diseases relationship presented in articles and documents. It acts as an underpinnings process to form a network is modelled here.

Keywords: Information Retrieval, NLP, Text Mining, Text Classification, Document indexing, Document Classification, Ranking.

Abbreviations: Natural Language Processing (NLP), Information Retrieval (IR), Information Extraction (IE).

1. INTRODUCTION

Our work bounds with a hybrid discipline of Text Mining, BioNLP and Information Retrieval. Text Mining is the process of finding a useful and interesting patterns, directions, models and rules from semi / un structured data. Biomedical text mining (also known as BioNLP) refers to text mining applied to texts and literature of the biomedical and molecular biology domain. Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. It is a rather recent research field on the edge of natural language processing, bioinformatics, medical informatics and computational linguistics. The Challenges of Text mining in biomed terminology are dynamic nature of the domain inclusive of new terms (genes, proteins, chemical compounds, drugs) which frequently and constantly being created and which biomedical resources need constant updating. In recent database research, studies have been done to model and implement semi-structured data. This is a point why we in need to ground and develop a new application to normalize and find structural information from semi-structured data present in open access database articles. Usually there is a huge gap from the stored data to the knowledge that could be constructed from the semi structured data. Information Retrieval model has been taken to triumph over this target as in our work.

II. A REVIEW OF RELATED WORK

There are a number of Information Retrieval techniques and text mining applications which can be used for discovering knowledge from articles are available. But traditional Information Retrieval techniques become inadequate for the increasingly vast amount of text data. The pitfalls of existing are due to size of the widely used database which has a negative impact on the relevance of users’ query results and also simple free-text queries would return many false positives. Query expansion or reformulation is used to improve retrieval of documents relevant to a free-text query or related to a document of
interest. Although applications are useful in exploring such information in the literature, not many of them provide real-time responses—the users often have to wait for several minutes before they receive the results. Some of the systems provide reasonably quick responses by limiting the number of documents to be analyzed to a very small number, but such limitation leads to a significant deterioration of the coverage[10].

To complement existing applications, we chose a search mechanism to groundwork and retrieve the knowledge and hidden structured data from abstract, articles, discussions etc., in accuracy and efficient based model.

III. KEY COMPONENTS OF TEXT MINING PROCESS

The key components and tasks grounding the structured information used are as

- Information Retrieval (IR)
- Information Extraction (IE)
- Information Indexing
- Text Classification
- Text Clustering

- Information Retrieval (IR)

  It is a process of recovery of documents from a collection of documents, open access database etc., which persuade a given information demand.

- Information Extraction (IE)

  It is the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. It focuses on the collection, organization and application of information to answer questions.

- Information Indexing

  Efficient Indexing is required to reduce vocabulary of terms and query formulation.

- Text Classification

  Common problem in information science is assignment and classification of an electronic document to one or more categories, based on its contents.

- Text Clustering

  Cluster documents into topics according to user query keywords. A clustering program tries to find the groups in the data. Text Clustering programs often choose first the documents that seem representative of the middle of each of the clusters. Then it compares all the documents to these initial representatives.

IV. INFORMATION RETRIEVAL

As the first and main process of knowledge discovery it does the automatic text processing. Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)[1]. Some techniques are used for representing the documents and some techniques are used for representing the information needed. Our idea in Information Retrieval can improve the precision of retrieval systems by filtering relevant documents for the given search query. It has been used to describe the applications of data mining techniques to automated discovery knowledge from text.

The key activities included in our IR Navigation system are Query Evaluation process, Topic tracking and Concept linkage.

A. Query Evaluation Process
Information demand is posed in form of a user flexibility query. It includes of crucial steps as Keyword Selection and Keyword Searching

- **Keyword Selection** - If given query consists of list of words it use Stoplist method and middle frequency word method to select important keywords. The function words do not bear useful information for IR. The removal of stop words with usual standard stoplist usually improves IR effectiveness

- **Keyword Searching** - To what extent does a document correspond to a query is determined by keyword searching step. Here the keyword in the query starts its search effort in the document. It is also referred as Stemming.

### B. Topic Tracking

After the keyword is selected from given user query, it is fed into the documents or articles present in the database with the forwarding steps of Topic tracking:

- Document Classification
- Document Ranking
- Document Indexing
- Document Clustering
- Ambiguity documents

1) **Document Classification** – Classification and categorization of documents are done according to the keyword given by user. It helps to best represent their contents in the documents and articles from the database. It does its test collections, test topics, relevance assessments and evaluation methods.

2) **Document Ranking** – Only a small fraction of the many available and classified documents will be relevant to a given individual or user. Without knowing what could be in the documents, it is difficult to formulate effective queries for analyzing and extracting useful information from the data. Users need tools to compare different documents, rank the importance and relevance of the documents, or find patterns and trends across multiple documents. It is done by a link analysis algorithm. It assigns a numerical weighting to each element of a hyperlinked set of documents with the purpose of "measuring" its relative importance within the set. The numerical weight that it assigns to any given element is referred as the document ranking

3) **Document indexing** – Designed to make search faster the ranked documents to commence automatic text processing. The steps of transformation, analysis, and retrieval of information are done to index the documents. *MapReduce* is a distributed programming tool designed for indexing and analysis tasks

4) **Document Clustering** – Identifies and infers information from the enormous amount of data from indexed documents[2]. Importance of document clustering is now widely acknowledged by researchers for better management, smart navigation, efficient filtering, and concise summarization of large collection of documents. It represent the document in such a form that inherently captures semantics of the text and defines a similarity measure based on the semantic representation such that it assigns higher numerical values to document pairs which have higher semantic relationship.

5) **Ambiguity documents** – This may also help to reduce dimensionality of the document. From the clustered documents the ambiguity documents are steeld downed here. A similarity measure is proposed based upon the inferred information through topic maps data and structures. The comparative experiment reveals that the proposed approach is effective in improving the cluster quality.

### C. Concept Linkage

The clustered documents with references to the user query and input are stored and visualized in concept linkage.
Query Evaluation Process

- User Input
- Keyword Selection
- Keyword Searching

Documents/Database Search

- Document Classification
- Document Ranking
- Document Indexing
- Clustering
- Ambiguity Documents

Topic Tracking

Concept Linkage

Visualization

Resulted Documents
V. IMPLEMENTATION AND DISCUSSION
The continuous growth in the number of research articles and the corresponding data stored in online repositories require better connections to be established between scientific articles, annotations and data. The main concern underlying retrieval systems are search by content, retrieval model, implementation and indexing [3].

The keyword from user given query is fed into database with articles, documents and scrap which undergoes the implementation step of new designed IR navigation system produces the visualization of result as shown as figure above. For example the user given keyword is of any entity like disease name, Gene name, and protein names and family it hits the database and brings out the articles with specified keyword as highlighted. For effective IR morphological, syntactic, semantic and discourse-level variation and ambiguity need to be taken into account. This BioNLP project addresses linguistic problems typical of biosciences, as well as general domain independent linguistic problems affecting the effectiveness of IR in the field. The problems specific to biosciences involve gene name synonymy and ambiguity, and the identification of phrasal gene and protein names. It also involves orthographic variation of names, named entity recognition and the identification of acronyms’ full names. IR navigation system allows us to generate a new representation based on concepts and it enriches good classification using the concepts and the hypernyms drawn from the domain ontology which has significantly boosted their representation.

VI. CONCLUSION
Information Retrieval is not an ephemeral problem. We addressed some of the technical issues in representing, analysing primitive features and variation in the use of natural language expressions constitutes a major problem in Information Retrieval. These are the building blocks of any such systems. They can be generalized into a much broader range of applications as well. The results of our experiments on the benchmark biomedical collection confirm the importance of the approach by a very significant improvement in the performance of the ontology-based classification compared to the classical representation. Many IR systems can modify keyword sets by estimating the user's particular requirement. Even though such IR systems have better retrieval performance, the complicated estimation process entailed by a large number of keywords makes it difficult for a user to understand how the system behaves. However, the factors considered are accuracy to represent meanings; exhaustiveness and facility for computer to manipulate are fulfilled.

REFERENCES
5. from Medline“, ACM portal, pp1410–1412
7. Mónica Marrero, Sonia Sánchez-Cuadrado, et al ,” Information Retrieval Systems Adapted to the Biomedical Domain”
BIOGRAPHY

1. **Mrs. K.Prabavathy M.Sc,M.Phil.**, Doctoral Research scholar in department of Computer Science, Manonmanium Sundaranar University, Tirunelveli, Tamil Nadu, India. She completed M.Phil in the area of Data Mining at PSG College of Arts and Science, Coimbatore and received MCA degree through Bharathiar University, Coimbatore and M.Sc degree through APA College of Arts and Science, Palani. She has published number of papers in reputed journals and conferences. She has about five years experience of teaching and research experience at PSG College of Arts and Science. Her area of interest includes Data Mining, Bioinformatics and Computer Networks.

2. **Dr. P. Sumathi** is working as an Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore, Tamilnadu, India. She received her Ph.D., in the area of Grid Computing in Bharathiar University. She has done her M.Phil in the area of Software Engineering in Mother Teresa Women’s University and received MCA degree at Kongu Engineering College, Perundurai. She has published a number of papers in reputed journals and conferences. She has about Sixteen years of teaching and research experience. Her research interests include Data Mining, Grid Computing and Software Engineering.