An Approach to Mine Textual Information From Pubmed Database

G.CHARLES BABU [1], Dr. A.GOVARDHAN [2]

[1] Professor & Head in CSE Department
Holy Mary Institute of Technology & Science,
Bogaram, Hyderabad.

[2] Professor in CSE Department
Director of Evaluation, JNTU Hyderabad.

Abstract

The web has greatly improved access to scientific literature. A wide spectrum of research data has been created and collected by researchers. However, textual information on the web are largely disorganized, with research articles being spread across archive sites, institution sites, journal sites and researcher homepages. Data was widely available over internet and many kinds of data pose the current challenge in storage and retrieval. Datasets can be made more accessible and user-friendly through annotation, aggregation, cross-linking to other datasets. Biomedical datasets are growing exponentially and new curative information appears regularly in research publications such as MedLine, PubMed, Science Direct etc. Therefore, a context based text mining was developed using python language to search huge database such as PubMed based on a given keyword which retrieves data between specified years.

Keywords: *Text mining, data, database, PubMed, python.*

I.INTRODUCTION

Computer-aided discovery of information based on text data is an exciting challenge in the field of research in biology and medicine [1]. It has been reported in literature that research data, in particular, is a valuable resource and making the data publicly available would aid towards progressing research. Moreover, the importance of data in a particular contextual reference and infrastructure that necessitates in managing data should not be ignored [2].

A wide spectrum of research data has been created and collected by researchers on all subjects and disciplines during the course of their research. In this event, data segregation and generation of datasets through different processes and methodologies are preserved and and shared with others through online resources [3]. Moreover, datasets can be made more accessible and user-friendly through annotation, aggregation, cross-linking to other datasets as well as developing various tools for data analysis and curation [4].

It has been emphasized that researchers working in scientific disciplines should be properly trained on datasets, usage, accumulation, storage and retrieval. Moreover, metadata provides information about data resources enabling efficient curation, management and re-use of the data [5]. However, when there is a lack of informative metadata or file format inconsistencies, datasets would lose the community and consigned to obscurity [6].

Data curation is the major task in storage and retrieval of textual information as they are prone to become unused if they are not curated efficiently. The two main ways of storing and curating data are: using large, centralized national or international data centres; or using a distributed array of local data stores (journals etc) [7]. Such publicly-available data are a valuable long-term essential resource. On the contrary, there are several obstacles to data sharing, such as the confidentiality of those from whom primary data are gathered, or the expense of creating datasets [8-9].

Biomedical datasets are growing exponentially and new curative information appears regularly in research publications such as MedLine,

PubMed, Science Direct etc [2]. Generation of datasets requires massive collection of data which can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia and/or hypertext documents [10-11]. With the enormous amount of data stored in databases and other repositories, powerful tools for analysis and interpretation that could help in decision-making has been developed [12].

Text mining is the extraction of useful information from large volumes of text. Text mining involves the application of techniques from areas such as information retrieval, natural language processing, information extraction and data mining. In this paper, we present text mining procedure written in python object-oriented language to extract or mine specific keywords from biomedical literature database like PubMed [13-14].

II. MATERIALS AND METHODS

Using the **urllib** module in the Python standard library, we *could* send a request directly to the Entrez server that handles normal interaction. Instead, NCBI use **eUtils**. In addition, providing an email address and other information helps us to process the request.

```
>>> import urllib
>>> Entrez.email = "email id"
>>>u="http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?term=%s&mindate=%d/01/01&maxdate=%d/12/31"
```

Dynamically providing the search terms, startYear, endYear and recordSize can be done by using the "**raw_input**" statement for the strings and "**input**" statement for the variables.

```
>>> Term = raw_input("Enter Search Term to find No. of Records: ")
>>> startYear = input("Enter the start Year: ")
>>> endYear = input("Enter the end Year: ")
>>> recordSize = input("No. of Records to be Retrieved: ")
```

Entrez esearch handles search terms and database to connect and the option to mention the start year and end year and these terms are predefined.

```
>>> handle = Entrez.esearch(db="pubmed", term= Term , mindate= startYear, maxdate=endYear, retmax=recordSize)
```

The **ElementTree** wrapper adds code to load XML files as trees of Element objects, and save them back again. We can use the **parse** function to quickly load an entire XML document into an ElementTree instance

```
>>> import xml.etree.ElementTree as ET
>>> url = u % (Term.replace(" ", "+"), year, year)
>>> page = urllib.urlopen(url).read()
>>> doc = ET.XML(page)
```

Entrez efetch fetches the data from the database based on the search terms defined in the Research.

```
>>> handle = Entrez.efetch(db="pubmed", id=idlist, rettype="medline",
                    retmode="text")
```

Medline has a parser to parse the records retrieved using Entrez.efetch. Medline.parse parses the records.

```
>>> record = Entrez.read(handle)
>>> idlist = record["IdList"]
>>> handle = Entrez.efetch(db="pubmed", id=idlist, rettype="medline",
                    retmode="text")
>>> records = Medline.parse(handle)
>>> records = list(records)
```

Output of these records retrieved by Entrez.efetch statement can be done by calling

predefined PubMed Search Field Descriptions and Tags.

PubMed ID [PMID]:

>>> print "PUBMED ID:" , record.get("PMID")

Title [TI]:

>>> print "TITLE:", record.get("TI", "Error (or) No Title")

Authors [AU]:

>>> print "AUTHORS:", record.get("AU", "Error (or) No Authors")

Abstract [AB]:

>>> print "ABSTRACT:", record.get("AB", "Error (or) No Abstract")

Source [SO]:

>>> print "SOURCE:", record.get("SO", "Error (or) No Source")

III. RESULTS AND DISCUSSION

Entering Search Term, start year, end year and number of records will retrieve the relevant data from PubMed database. As given in Figure-1, finding out the number of records on "cancer" from year 2008 to 2009 retrieved 5 records.

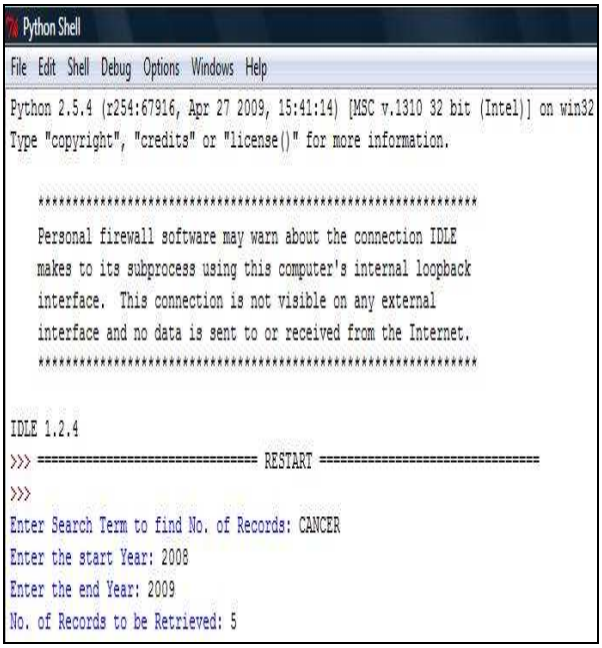

Figure 1. Entry terms to search PubMed Database

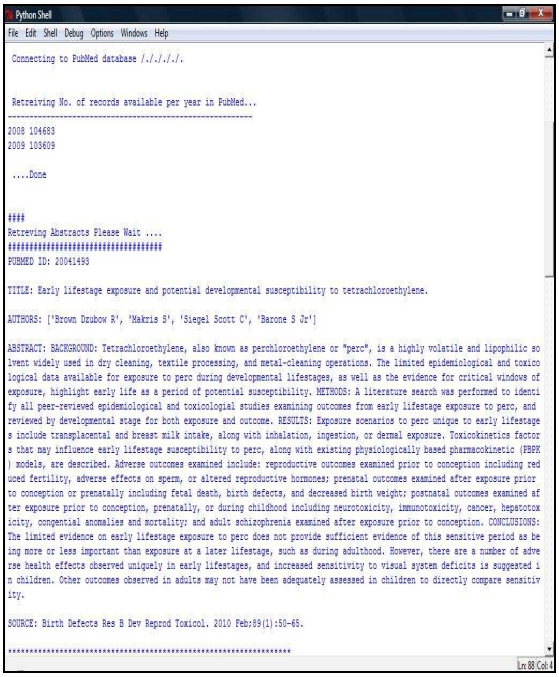

Figure 2. Retrieved abstracts based on the search terms provided

IV. CONCLUSION

Given the complexity and sheer magnitude of the task of searching the vast amount of literature and other databases for a certain piece of information, it is necessary to develop improved computer-based tools to aid the human expert. Also, this information is often scattered throughout the published literature and it first must be translated into computer-readable form and associated with the data records to which they are referring. The emergence of scripting languages such as Python, Tcl, and Perl as major tools in software development represents a potentially revolutionary change in computer programming. Therefore, an attempt to extract keyword based data from PubMed database was made as a much helpful resource to researchers using Python language.

REFERENCES

[1] Ingrid Petrič "Text Mining for Discovering Implicit Relationships" in Biomedical Literature Informatica 344 (2010) 261-262

[2] Latha .K, Kalimuthu.S, Dr.Rajaram.R "Information Extraction from Biomedical Literature using Text Mining Framework" International Journal Of Imaging Science And Engineering (Ijise) Vol.1,No.1, January 2007.

[3] Aaron Griffiths, The Publication of Research Data: Researcher Attitudes and Behaviour, The International Journal of Digital Curation 4(1): 2009 46-56

[4] http://www.rin.ac.uk/files/Data%20publication%20report,%20main%20-%20final.pdf

[5] van Bemmel J.H. et al (Eds). Data, Information and Knowledge in Medicine. Methods of Information in Medicine, Special issue, 27(3), 1988

[6] Geographic information metadata for spatial data infrastructures: resources, By Javier Nogueras-Iso, F. Javier Zarazaga-Soria, Pedro R. Muro-Medrano, Springer Publishers, Chapter 1, pp1-2

[7] Ian H Witten, Eibe Frank, Mark A Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition, Morgan kaufmann publishers, Chapter 1, pp 3-9

[8] Jan-Marco Bremer and Michael Gertz, Integrating document and data retrieval based on XML, The VLDB Journal Volume 15, Number 1, 53-83, DOI: 10.1007/s00778-004-0150-4

[9] Donna J. Peuquet & Niu Duan, An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data, International journal of geographical information systems Volume 9, Issue 1, 1995 pp7-24

[10] JENSEN , C. S. , CLIFFORD , J. , GADIA , S. K. , SEGEV , A. , and SNODGRASS , R. T. , 1993 , A glossary of temporal database concepts . In Proceedings of the International Workshop on an Infrastructure for Temporal Databases , (Arlington , TX : Association for Computing Machinery) pp. A25 – A29 .

[11] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, U.Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., MIT Press, Cambridge, Mass., 1-36

[12] Yeh, A. S., Hirschman, L. and Morgan, A. A. (2003), 'Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup', Bioinformatics, Vol. 19 Suppl. 1, pp. i331–339

[13] Schwartz, A. S. and Hearst, M. A. (2003), 'A simple algorithm for identifying abbreviation definitions in biomedical text', in 'Proceedings of the 8th Pacific Symposium on Biocomputing', 3rd–7th January, Hawaii, pp. 451–462.

[14] Yu, H. and Agichtein, E. (2003), 'Extracting synonymous gene and protein terms from biological literature', Bioinformatics, Vol. 19 Suppl. 1, pp. i340–349

G.Charles Babu received the degree in computer science & engineering from KLCE in 1997, Masters in Software Engineering from JNTU in 1999.Presently working as a Professor & Head in CSE department in Holy Mary Institute of Technology & Science.

Dr.A.Govardhan received Ph.D in Computer Science & Engineering from JNTU in 2003.M.Tech from JNU in 1994.B.E from Osmania University in 1992. He is Presently working as a Director of Evaluation in JNTU, Hyderabad.

His research interest includes Data Mining, Information Retrieval & Search Engines. He has Published more than 120 papers in Various international Journals.