Journal of
Electrical & Electronic Systems

Kanwal et al., J Electr Electron Syst 2018, 7:4
DOI: 10.4172/2332-0796.1000277

Research Article Open Access

Pubmedinfo Crawler: Data Extraction System from PUBMED for Fast Research

Kanwal A[1]*, Fazal S[1], Bhatti AI[1] and Khalid MA[2]

[1]Department of Biosciences, Capital University of Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, GC University, Faisalabad, Pakistan

Abstract

Background: During last decades, an extraordinary improvement of bioinformatics has been observed that has prompted growth of a huge measure of biological data. The Bioinformatics and computational science aim to manage this huge volume of information. In the way biological data can be extracted, producing scientific knowledge, handling and mining huge information is at present a subject of incredible conspiracy and significance. Automation particularly in the information extraction step might be an essential technique to lessen the time important to finish an efficient research. However, the state of the art of automatically extricating information components from biological databases has not been all around portrayed.

Methods: Systematically PubMedInfo Crawler will identify potentially relevant articles and their details against different keywords. The included methodology met the following criteria: 1) Get keywords from user, send request to eutils.ncbi.nlm.nih.gov, fetch results from this server and then generate database to display output containing title and links of articles found against given keywords. 2) Transform these links into useful and structured form to get detailed information of each article: like PubMed id, title, abstract, journal name, authors' name, publication date and country name to which authors belong. 3) Analyze the obtained details from different aspects that are discussed in detail in the methodology section.

Results: PubMedInfo Crawler (PMIC) has been developed to provide data extraction utilities for commonly used database PubMed. It is a simple web interface that enables input of query in the form of keyword and generates detailed useful information of each article against input keyword. Tool has been experimentally tested on different query keywords and has validated the results from PubMed. The overall accuracy of the crawler was found to be 96% for the number of articles against query terms. The tool is freely available.

Conclusion: This tools with help the public users to extract the data from Pubmed through an automatic way with less consumption of time. Furthermore it will help the researchers to accomplish their research in a better way with less effort.

Keywords: Bioinformatics; Computational biology; Data retrieval; Biological databases

Introduction

The exponentially expanding amount of information being created every year makes getting valuable data from that information more critical. The data is frequently stored in a databases, a repository of information accumulated from different sources, including corporate databases, summarized data from interior frameworks, and information from outer sources [1]. Analysis of the information incorporates straightforward, simple query and reporting, statistical analysis, more complex multidimensional analysis, and data mining. A few evaluations show that the measure of new data doubles every three years [2]. With a central repository to keep the gigantic measures of information, organizations require tools that can help them to extract the most valuable data from the information. A data warehouse can unite information in a solitary arrangement, supplemented by metadata through utilization of an arrangement of info instruments known as extraction, transformation, and loading tools. These tools empower to rapidly settle knowledgeable decisions based on good information analysis from the data [3]. As the intricacy of data analysis develops, so does the amount of data being stored and analyzed; ever more powerful and faster analysis tools and hardware platforms are required to maintain the databases.

Information mining can be characterized as the way towards extricating information, analyzing it from many viewpoints, and then delivering a synopsis of the data in a valuable frame that recognizes relationship within the data. Information mining, the extraction of concealed prescient data from vast databases, is an intense new innovation with extraordinary potential to help researchers and biologists concentrate on the most critical data in their biological databases. Information mining instruments anticipate future patterns and practices, permitting making proactive, knowledge driven choices [4].

The initial step of an ETL process includes extracting the information from the source frameworks. However, this is the most difficult part of ETL, as extracting information accurately will set the phase for how subsequent procedures will go. Extraction is the operation of separating information from a source framework for further utilization. After the extraction, this information can be changed and stacked into the databases. In general, the objective of the extraction step is to change

***Corresponding author:** Kanwal A, Department of Biosciences, Capital University of Science and Technology, Kahuta Road, Sihala, Islamabad Capital Territory, Pakistan, Tel: +92-51-111-555-666; E-mail: attiya_kanwal@yahoo.com

Received June 13, 2018; **Accepted** September 15, 2018; **Published** September 27, 2018

Citation: Kanwal A, Fazal S, Bhatti AI, Khalid MA (2018) Pubmedinfo Crawler: Data Extraction System from PUBMED for Fast Research. J Electr Electron Syst 7: 277. doi: 10.4172/2332-0796.1000277

over the information into a solitary configuration which is suitable for transforming processing [5]. An intrinsic part of the extraction includes the parsing of extracted information.

Designing and making the extraction process is often a standout amongst the most tedious assignments in the ETL procedure and, indeed, in the whole information warehousing process. The source frameworks may be exceptionally intricate and ineffectively recorded, and in this way figuring out which information should be extracted can be troublesome. The information must be separated not only once, but several times in a periodic manner to supply all changed data to the warehouse and keep it up-to-date. In addition, the source framework ordinarily can't be altered, nor can its execution or accessibility be balanced, to suit the requirements of the data warehouse extraction process [6].

Information extraction is the demonstration or procedure of recovering information out of (generally unstructured or inadequately organized) information sources for further information processing or data storage. The import into the intermediate extracting framework is in this manner typically taken after by information transformation and potentially the expansion of metadata before exporting to another phase in the data work process [7].

For the most part, the term information extraction is connected when (trial) data is first imported into a PC from primary source. Today's electronic gadgets will generally introduce an electrical connector (e.g. USB) through which 'crude information' can be streamed into a PC.

Typically, unstructured information sources incorporate web pages, electronic mails, archives, PDFs, filtered content, centralized server reports, spool records, classifieds, and so on which is additionally utilized for various analysis purposes. Extricating information from these unstructured sources has developed into a significant specialized test whereas generally information extraction has needed to manage changes in physical equipment arranges, the dominant part of current information extraction manages extracting information from these unstructured information sources, and from various software formats. This developing procedure of information extraction from the web is alluded to as Web scratching [7].

The act of adding structure to unstructured data takes a number of forms

- Using text pattern matching such as regular expressions to recognize small or large-scale structure e.g. records in a report and their related data from headers and footers;
- Using text analytics to endeavor to know the text and link it to other information.

The evaluated measure of the information to be separated and the phase in the ETL procedure (starting burden or upkeep of information) may likewise affect the choice of how to remove, from an intelligent and a physical point of view. Essentially, you need to choose how to extricate information intelligently and physically.

There are two kinds of logical extraction:

- Full Extraction
- Incremental Extraction

In full extraction the information is extricated totally from the source framework. Since this extraction mirrors every one of the information at present accessible on the source framework, there's no compelling reason to monitor changes to the information source since the last fruitful extraction. The source information will be given as-is and no extra intelligent data (for instance, timestamps) is essential on the source site. A case for a full extraction might be a fare document of a particular table or a remote SQL explanation checking the total source table.

In incremental extraction at a particular point in time, just the information that has changed since an all-around characterized occasion back in history will be extricated. To distinguish this delta change there must be a probability to recognize all the changed data since this particular time occasion. This data can be either given by the source information itself like an application section, mirroring the last-changed timestamp or a change table where a fitting extra system monitors the progressions other than the starting exchanges. By and large, utilizing the last technique implies adding extraction rationale to the source framework.

Contingent upon the picked intelligent extraction technique and the capacities and limitations on the source side, the removed information can be physically separated by two components. The information can either be removed online from the source framework or from a disconnected structure. Such a disconnected structure may as of now exist or it may be produced by an extraction schedule.

There are the following methods of physical extraction:

- Online Extraction
- Offline Extraction

The data is extracted directly from the source system itself by using the online extraction technique. The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner.

Through offline extraction technique the data is not extracted directly from the source system but is staged explicitly outside the original source system. The data already has an existing structure.

An important consideration for extraction is incremental extraction, also called Change Data Capture. If a data warehouse extracts data from an operational system on a nightly basis, then the database enquires only the data that has changed since the last extraction (that is, the data that has been modified in the past 24 hours).

When it is possible to efficiently identify and extract only the most recently changed data, the extraction process (as well as all downstream operations in the ETL process) can be much more efficient, because it must extract a much smaller volume of data. Unfortunately, for many source systems, identifying the recently modified data may be difficult or intrusive to the operation of the system. Change Data Capture is typically the most challenging technical issue in data extraction.

Literature Review

A handful of investigations are accessible in the prose that is related to discover interesting information from the huge scale biological databases. This aspect of data mining has gotten much deliberation among the researchers in recent times. A succinct analysis of some topical revelation associated with data extraction is presented it. But as it is already mention that biomedical natural language processing techniques have not been used to automate the information extraction scheme of research process.

To date, there is limited knowledge and methods on how to automate the data extraction phase of the systematic research, despite being one of the most time consuming steps. The tools that facilitate extraction of the data and integration of diverse data types are therefore the need of the hour. To address this gap in knowledge, we sought to perform a review of methods to automate the data extraction component along with their positive and negative aspects.

The free PubCrawler [8], web service (http://www. pubcrawler. ie) has been working for quite a long while thus far has brought literature and sequence updates to enormous level. It gives data on a customized page whenever new articles show up in PubMed or when new successions are found in GenBank that are specific to customized queries. The server likewise goes about as an automatic alerting system by conveying short warnings or messages with the most recent updates when they wind up plainly accessible. PubCrawler keeps on being a critical apparatus for biomedical analysts. . But it adds the information on the particular web page and further we have to manually extract information of the newly added article from that page.

Various other specific scattering of data (SDI) administrations exists, both business and allowed to general society. A few cases incorporate PubMed Cubby, BioMail, JADE, OVID and Science Direct yet have an indistinguishable constraint from Pubcrawler has. BioDB Extractor (BDE) Rajiv k et al., gives a typical information extraction stage for numerous databases like ENA, UniprotKB, PDB, and KEGG. But, this is not a substitute to fundamental keywords based database searches. It empowers contribution to the type of promotion numbers/ID codes, selection of utilities and choice of fields/subfields of information by the clients. NCBI E-utilities [9], give tweaked information extraction utilities to different databases accessible at NCBI. Be that as it may, these utilities require era of URL by the client I the configuration particular for separate databases physically. There are restricts on the quantity of URL asked for every second and every day, which can be submitted through an IP address. PDB Goodies (2002), BioDownloader (2007), BioMart (2015) [10], are likewise a portion of the cases of information extraction from natural databases yet each of them is particular to specific database not for PubMed.

Methods

The methodology and the materials that were adopted to carry out the current work have been discussed in detail.

Data collection from biological database

The typical process flow of utilities in PMIE is given below.

- Get keywords from the user. Keywords may be a disease or gene/protein name.
- Connects the relevant keyword to the respective database.
- Extract records against the particular keyword from the specified database
- Generates and returns the output in the database containing titles and links of the records from the database
- Using links of articles to extract more detailed information about each article such as abstract of the paper, authors' name, authors' country, publication date and journal name.

System requirements

Minimum 32 bit Operating system

Operating system: Windows xp, 7, 8, 8.1, 10, Macintosh, Linux, Unix etc,

Need an Apache server for local system run: In windows (xampp, Wamp, local-server) for Macintosh (Mamp, local-server) for all other operating system (Lamp).

Language used In This project: HTML, PHP, CSS

Framework Used: Bootstrap

Output: Output was stored in the local disk Drive->Xampp->htdocs->crawler->output

Getting keywords

The PMIE provides the opportunity to extract articles from PubMed by using disease or gene/protein name and through year wise. To efficiently identify and extract only the most recently added data from the PubMed makes the extraction process much more efficient, because it must extract a much smaller volume of data of the given year.

Connecting to PubMed

After getting disease or gene name from the user the request was sent to eutils.ncbi.nlm.nih.gov, to fetch results from this server and then generate html file to display output. The output contains the links to articles and the URLs' against entered disease or gene name.

Extracting articles details

By using the keywords as query terms in PUBMED, we have saved all the papers' links that are returned to us as a result of these queries. From each paper, the following information was extracted; Pubmed id of each document, title, abstract, author's names and year of publication. All this work was done by screen scrapping methodology, a code was written in PhP platform to extract all the required information from the PUBMED database against each query term. The retrieved information was saved in different formats like SQL, CSV or excels, PDF etc. to carry out further processing [11].

Results and Evaluation

As the common diseases with high incidence, Type 2 diabetes mellitus gains much attention among researchers and has a rather large literature accumulation. We used Type 2 diabetes as testing disease for system evaluation. We experimentally tested our application on Type II Diabetes Mellitus data for the last five years 2013 to 2017. The tables below show the different possible query keywords for Type II diabetes mellitus that were used for testing the proposed application. In order to be more accurate in our results we used all possible terms for Type II Diabetes for the collection of data. These queries include "Type II Diabetes", "T2D", "T2DM", "type 2 diabetes", "diabetes type 2", "type 2 diabetes mellitus", "diabetes mellitus type 2" Figure 1.

Connecting to PubMed

After getting above mentioned query terms from the user the request will be sent to eutils.ncbi.nlm.nih.gov, to fetch results from this server and then generate html file to display output. The output against each query was returned in the form of .html files that contains titles of articles and the URLs' against the specific query term. The Figures 1 and 2 below shows the screenshot of the results against the query Type 2 Diabetes for the year 2017 from PubMed.

These .html files contain the articles that are in PubMed related to Type 2 Diabetes. On clicking each html file, we get the titles and URL's of articles against Type 2 Diabetes. Each html file contains maximum 20 articles for the specific query. Type-2-Diabetes-2017-0 contains the first 20 results for Type 2 diabetes. Type-2-Diabetes-2017-1 contains next 20 articles found for the query Type 2 Diabetes for the year 2017. The Figure 2 shows the first 20 results of html file Type-2-Diabetes-2017-0.

In Figure 3, the address bar shows first 20 articles against the query Type 2 Diabetes for the year 2017. File name "Type+2+Diabetes-2017-0.html" means this is the first page of the PubMed results against the specific query. Similarly in the address bar "%20diabetes%20Mellitus" represents that each .html page of Figure 2 contains 20 articles per page.

Extracting articles details

By using the keywords as query terms in PUBMED, we have saved all the papers' links that are returned to us as a result of these queries as shown in Figure 3. From each paper, the following information was extracted; Pubmed id of each document, title, abstract, author's names and year of publication. All this work was done by screen scrapping methodology, a code was written in PhP platform to extract all the required information from the PUBMED database against each query term. The retrieved information was saved in different formats like SQL, CSV or excels, PDF etc to carry out further processing. Figure 4 shows the screenshot of what type of information we get as a result of further processing on the above mentioned steps.

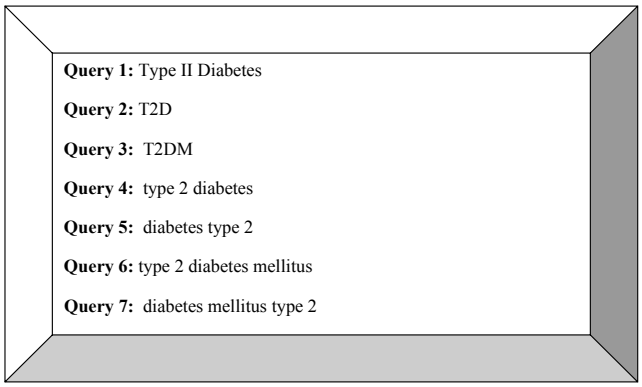

Figure 1: Queries.

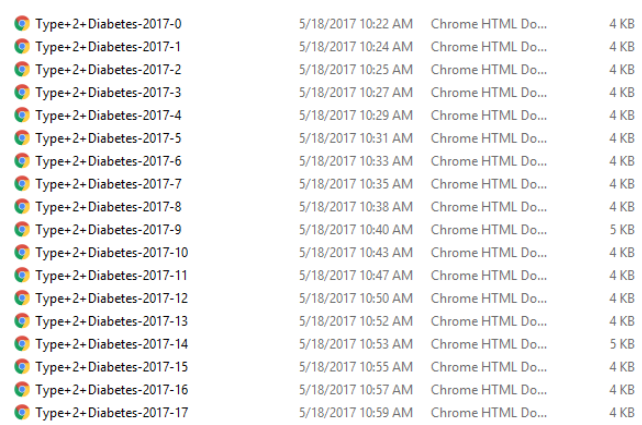

Figure 2: Output of the PubMed Info Extractor (PMIE) against Type 2 Diabetes for the year 2017.

The highlighted portion shows the author's affiliation (Department of Molecular and human Genetics, Journal Name: Carcinogenesis and PubMed ID: 28535186). Using the screen scrapping code, for each article link shown in Figure 3, the abstract, Authors' names of an article, authors' affiliation, Journal name and PubMed ID has been extracted.

We evaluated our tool by using Heuristic expert based evaluation methods. The articles and their detailed information related to different possible query terms against Type II diabetes mellitus were extracted. The results of each phase of preprocessing methodology were evaluated and validated. The Figure 5 shows the graphical representation of the data articles obtained from PubMed from 2013 to 2017 against different query terms of Type II Diabetes. This figure shows the rise and fall of research work in the area of Diabetes Mellitus.

The evaluation was done by the real user that belongs to department of Computer science, department of Bioinformatics and Biosciences of Capital University of Science and technology. The Table 1 shows the evaluation results along with tool's accuracy against different query keywords and their results of the crawler. The accuracy was obtained through "the ratio of results in PubMed to the results through crawler". Percentage accuracy is obtained in the same way.

The number of articles in PubMed against the query "Type II Diabetes" for the year "2013" was found to be "10353". The number of articles we got through PubMedInfo Crawler was "10,965" for the year 2013 against the same query term. The %age accuracy of the results calculated through (Accuracy = Results in PubMed/Results through crawler * 100) was approximately 94%. Similarly the number of articles in PubMed against the query "Type II Diabetes" for the year "2014" was found to be "11088". The number of articles we got through PubMedInfo Crawler was "10,562" for the year 2014 against the same query term. The %age accuracy of the results calculated through (Accuracy = Results in PubMed/Results through crawler * 100) was approximately 95%. In the same way we have checked the accuracy of our results for the years 2015, 2016 and 2017 against the query term "Type II Diabetes". The average accuracy calculated against this query term for years 2013 to 2017 was found to be approximately 94%. Likewise the average accuracy calculated against the queries "T2D", "T2DM", "Type 2 diabetes", "Diabetes type 2", "Type 2 diabetes mellitus" and "Diabetes mellitus type 2" for years 2013 to 2017 was found to be approximately 97%, 96%, 95%, 96%, 95%, 96%. The overall accuracy of the crawler was found to be 96% for the number of articles against query terms.

The graphical representation of how much accuracy obtained in the extraction of papers' details like abstract, authors' names, authors country, journal name, publication date is shown in Figures 6-10. From those URLs, we got through PubMedInfo Crawler, the abstracts, Authors' names, Authors' affiliation, and Journal names got the accuracy 97% for the year 2013 against the query term "Type II Diabetes". The publication dates were extracted with the accuracy of 86% for the year 2013 against the same query term. The accuracy obtained for the abstracts, Authors' names, Authors' affiliation, and Journal names against the query term "T2D" was 92% and the publication dates were extracted with the accuracy of approximately 88% against the query term "T2D" for the year 2013. In the same way the accuracy of extracted detailed information was evaluated manually against all other query terms for the years 2013, 2014, 2015, 2016 and

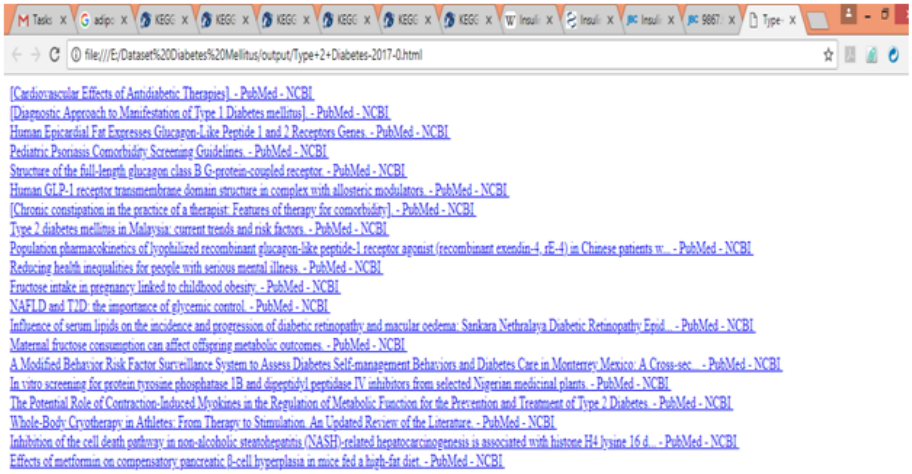

Figure 3: Titles of the output with URL's.

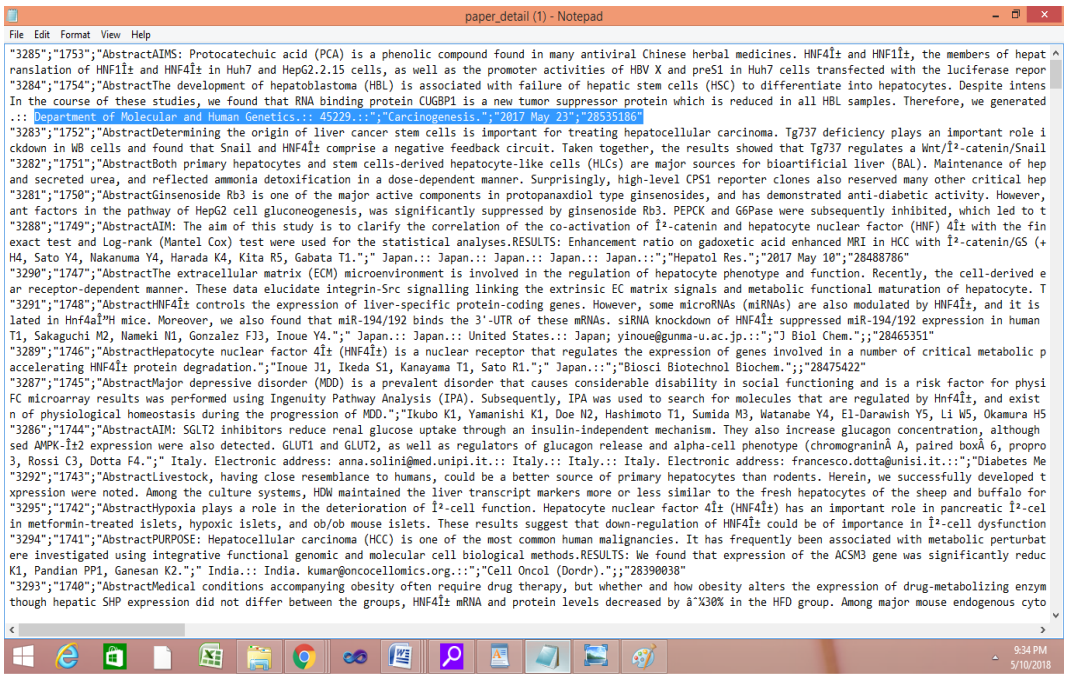

Figure 4: Screenshot of the paper details against specific query.

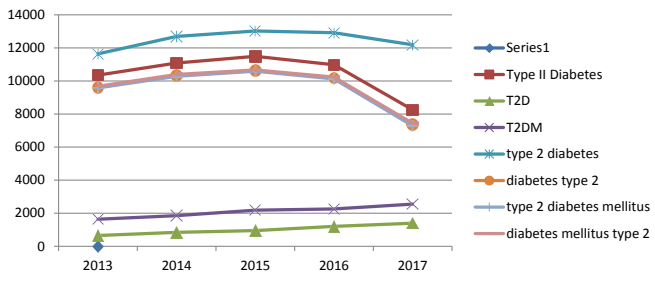

Figure 5: Graphical representations of the data against PubMed queries.

2017. The overall accuracy for abstracts, authors, author's country and journal name was found to be approximately 98% whereas the publication dates got relatively less accuracy as compared to other information of about 83%. The reason for this was that many articles do not have exact date including day, month and year, they just have year of publication.

PMIC has been developed with the target of giving a platform to information extraction and mining helpful data from the extracted information from the biological database PubMed. It adds to the abilities of the current apparatuses and servers for information extraction and analysis.

	Type II Diabetes				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	10353	11088	11492	10982	8257
Results through crawler	10,965	10,562	10,789	10,167	7863
Accuracy	94%	95%	94%	93%	95%
	T2D				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	651	842	951	1206	1395
Results through crawler	651	856	912	1149	1337
Accuracy	100%	98%	96%	95%	96%
	T2DM				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	1636	1854	2183	2257	2549
Results through crawler	1596	1791	2147	2196	2456
Accuracy	97%	96%	98%	97%	96%
	Type 2 diabetes				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	11641	12694	13022	12917	12186
Results through crawler	10967	12145	12323	12175	11586
Accuracy	94%	96%	95%	94%	95%
	Diabetes type 2				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	9612	10323	10622	10172	7359
Results through crawler	9087	9865	10289	9693	7159
Accuracy	94%	96%	97%	95%	97%
	Type 2 diabetes mellitus				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	9581	10281	10592	10124	7284
Results through crawler	8869	10112	10163	9599	6839
Accuracy	93%	98%	96%	95%	94%
	Diabetes mellitus type 2				
	2013	**2014**	**2015**	**2016**	**2017**
Results in PubMed	9674	10394	10666	10234	7438
Results through crawler	9359	9973	10259	9993	6934
Accuracy	97%	96%	95%	98%	93%

Table 1: Last five years results evaluated by different users for type 2 Diabetes Mellitus against possible queries.

Applications of the Tool

Text mining -- the automated extraction of information from (electronically) distributed sources – could fulfill a significant task potentially, but only if we know how to control its strengths and overcome its weaknesses. In genomics, this is mostly pressing as more and more rare disease-causing variants are found and there is a need to understand those variations to fight with life threatening diseases. Moreover, finding associations among different biological objects that include genes, proteins and diseases etc. will help scientists and researchers to cope with today's serious health issues. This technology may put scientists and biomedical regulators at a severe advantage. The PubMed Crawler is a freely available application through which one can handle such problems. We hope that this application will serve in many different ways. Some of which are mentioned in this section. This app will provide a timely and useful overview of the current status of any field, including a survey of present challenges; like which countries and their institutes are mostly involve in the research on any particular disease. (ii) to enable researchers to choose how and when to pertain text mining tools in their own research; and (iii) to highlight how the research communities (authors) in genomics and systems biology can help to make text mining from biomedical abstracts and texts more easy [12].

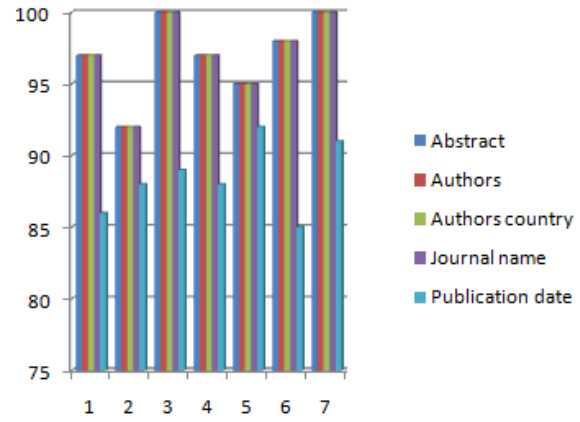

Figure 6: Paper details accuracy results against the query terms for year 2013.

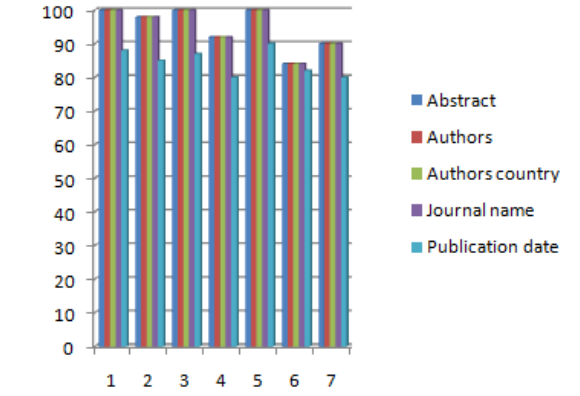

Figure 7: Paper details accuracy results against the query terms for year 2014.

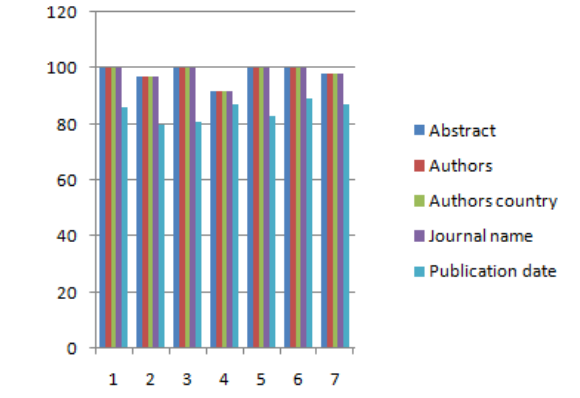

Figure 8: Paper details accuracy results against the query terms for year 2015.

Discussion and Future Direction

PMIC gives a simple and straightforward UI for information extraction and mining data from the extricated information. It is an effective and dispenses with the requirement for manually downloading articles from PubMed against given question. The objective client of

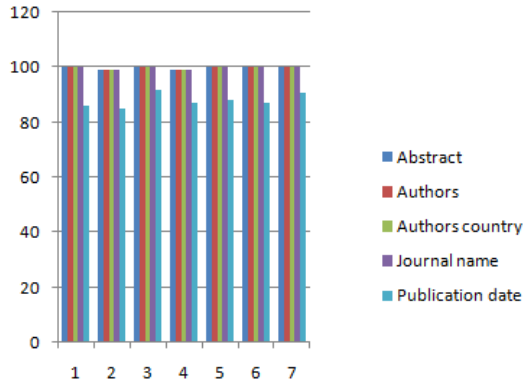

Figure 9: Paper details accuracy results against the query terms for year 2016.

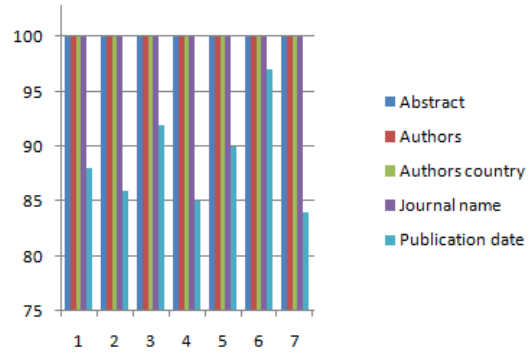

Figure 10: Paper details accuracy results against the query terms for year 2017.

PMIC may incorporate life science analysts/area specialists and also trained computational researcher and bioinformaticians. With its basic and enlightening UI, PMIC likewise can possibly be helpful instrument for tenderfoots and learner clients' keen on creating curated data from the information put away in PubMed. Future version of PMIC would cover the wider range of primary and secondary biological databases.

The next step in this research is to use this tool for above mentioned applications and to present whole work picture on a particular disease. Secondly, an effort will be made to extend this tool to create multithreaded implementation. With this we can extract papers of different years at the same time. Right now, user has to wait to extract results for next year until he/she obtained results of one mentioned year as a query.

Conclusion

PMIC has been developed with the target of giving a platform to information extraction and mining helpful data from the extricated information from the biological database PubMed. It adds to the abilities of the current apparatuses and servers for information extraction and analysis.

Authors' Contributions

Ms. Attiya Kanwal, Dr. Sahar Fazal and Dr. Aamir Iqbal Bhatti contributed in the design and methodology of the tool. Ms. Attiya Kanwal and Muhammad Arslan Khalid contributed in the development of tool. Ms. Attiya Kanwal and Dr. Sahar Fazal contributed in the evaluation of the tool.

Acknowledgement

This work is dedicated to Capital University of Science and Technology which provided us with all necessary facilities to complete this work. Then to our parents, teachers and all others who helped and supported us throughout the work.

References

1. Aniket A, Pawar PM (2015) Mining of complex data using combined mining approach. AVCOE, Sangamner.
2. Han J, Kamber M (2006) Data Mining Concepts and Techniques. 3rd Ed, US.
3. Vassiliadis P, Simitsis A, Georgantas P, Terrovitis M, Skiadopoulos S (2005) A generic and customizable framework for the design of ETL scenarios. Elsevier Science.
4. Larose DT (2012) Discovery knowledge in data, an introduction to data mining. John Wiley & Sons.
5. Ali El-Sappagh SH, Ahmed Hendawi AM, El Bastawissy AH (2011) A proposed model for data warehouse ETL processes. J King Saud Univ – Comput Info Sci 23: 91-104.
6. Rajiv P (2008) Principles and implementation of Data warehousing.
7. Mohanjit K, Kang HS, Maan KS (2014) Iterative algorithm for extraction and data visualization of HL7 data. Int J Comput Sci Eng.
8. Karsten H, Wolfe KH (2004) PubCrawler: keeping up comfortably with PubMed and GenBank. Nucleic Acids Res 32: W16-W19.
9. Sayers E (2010) A general introduction to the E-Utilities. NCBI, US.
10. Rajiv K, Sawant S, Kale UK (2015) BioDB extractor: Customized data extraction system commonly used bioinformatics databases. BioData Min.
11. Nielsen J, Mack RL (1994) Usability Inspection Methods. John Wiley & Sons, New York, pp: 25-62.
12. Wharton C, Rieman J, Lewis C, Polson P (1994) The Cognitive Walkthrough Method: A Practitioner's Guide: Usability Inspection Methods, Nielsen J and Mack R (eds.), Wiley, New York, pp: 105-140.