IJCSN International Journal of Computer Science and Network, Volume 5, Issue 2, April 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Patents and Publications Web Scraping

[1] Sushitha S, [2] Vijayalakshmi S Katti, [3] Sowmya H N, [4] Samanvita N

[1, 2, 3] MCA, Nitte Meenakshi Institute of Technology
Bangalore, Karnataka, India

[4] EEE, Nitte Meenakshi Institute of Technology
Bangalore, Karnataka, India

Abstract - Web scraping is a software technique of extracting information from websites. This type of software programs simulates human browsing or exploration of the World Wide Web by the means of applying low-level Hypertext Transfer Protocol (HTTP). Web scraping is a technique of collecting information from WWW using a Web Crawler and it is a common technique used by most Application Programming Interface (API's). This paper is used to fetch the recent publications and patents related to the pharmaceutical industries by developing an efficient web crawler which will fetch all possible information on publications and patents from different pharmaceutical industry websites based on their recent news, meetings and innovations.

Keywords - *Web Crawler, Web Scraping, Hypertext Transfer Protocol.*

1. Introduction

The Pharmaceutical Industries need to have the information on the recent publications and patents of pharmaceuticals. The main problem is user has to maintain the list of topic and specific URLs of Publication and Patents in excel sheet and navigate among them manually through browser and the task of keeping URL's and navigating to multiple sources is quite strenuous. The purpose of this paper is to build a "Web Crawler", which fetches information on Patents and Publications of the Pharmaceutical industries.

A web crawler copies the web pages and indexes, the downloaded pages so that users can search them much more quickly. A Web crawler begins with a list of URLs stored in the database. As it visits these URL's it identifies the links related to the Publications & Patents of Pharmaceutical industries and collects the information and display's the result to user. In this project the crawler also performs the archiving of pharmaceutical websites, it also fetches and stores the information as it goes.

2. Existing System

The existing system is a manual system where web scraping is done by human browsing using computer software. Clients used to search and fetch the Patents & Publications information manually and store the browsed information prepared by the organizations themselves. The clients would then select the categories continuously each time.

2.1 Drawbacks of Existing System

The system is vulnerable to various types of errors. As the system is a manual one, the scraped information is less and not accurate.

3. Proposed System

The main objective of proposed system is to develop a highly efficient web crawler which is used to fetch patents and publications of the pharmaceuticals industry by understanding different API's, collecting the specific information from the API sources and formatting it to get the equal information of the extracted data.

3.1 Advantages

The proposed system has an option to fetch more information and do more work efficiently as Web Crawlers are involved which does the task on behalf of the humans , making it accessible to use anywhere with a click of button.

4. System Specification

4.1 Hardware Specification

Processor Pentium Processor - 2.16 GHZ
RAM - 2 GB
Hard Disk - 250 GB

4.2 Software Specification

Back end - MySql 5.5
Front end - HTML, CSS
Web Server - WAMP Server
Development Tool b - Net Beans 5.3.8

5. System Architecture

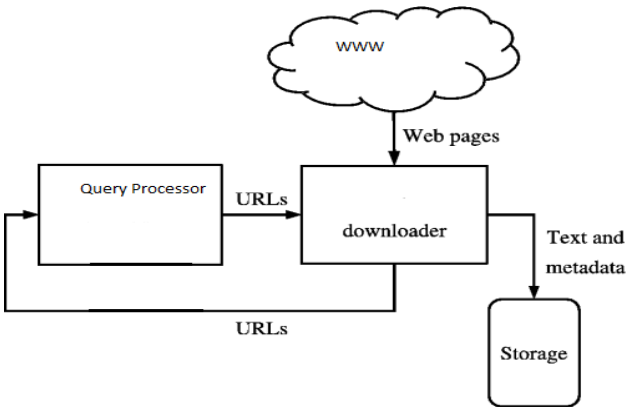

Fig 1System Architecture

5.1 Flow Chart

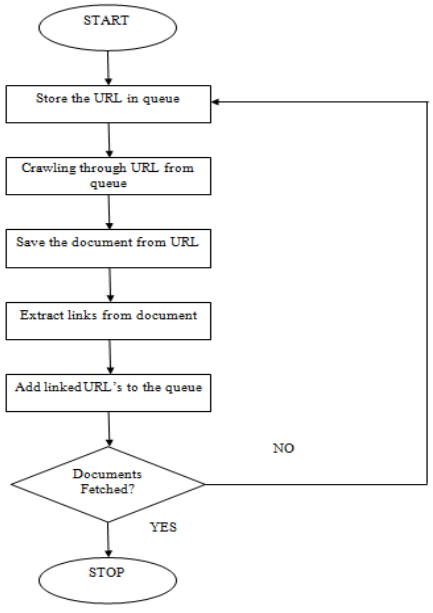

Fig 2: Flow Chart

5.2 Class Diagram

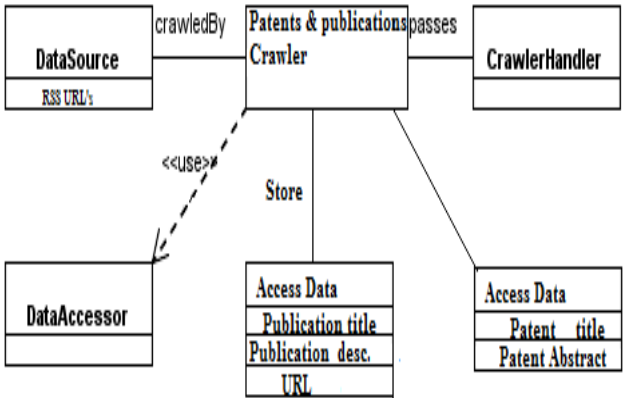

Fig 3: Class Diagram

5.3 Sequence Diagram

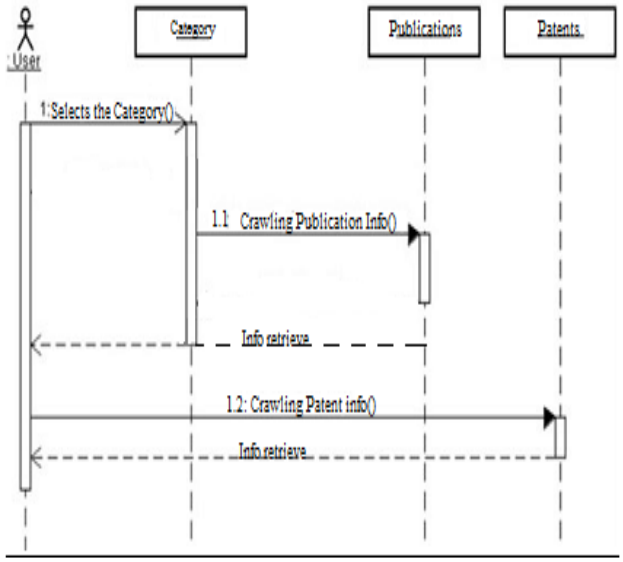

Fig 4: Sequence Diagram

6. Modules Description

- Publications Web Crawler
- Patents Web Crawler

6.1 Publications Crawler

A Publication Web Crawler visits URLs & identifies all the hyperlinks of recent publications of the Pharmaceutical industries. The Publications web crawler visits the stored RSS URL's, as it visits these RSS URL's it identifies the links related to the Publications of Pharmaceutical

industries and collects the Publications information and display's the Crawled result to user.

6.2 Patents Crawler

A Patent Web Crawler does the similar task of publication Web crawler; it visits URLs & identifies all the hyperlinks of Patent documents instead of publications. The Patents crawler visits the stored RSS URLs, it identifies the links related to the Patents of Pharmaceutical industries and collects the patents information and display's the Crawled result to use.

7. Results

7.1 Publications Screenshot

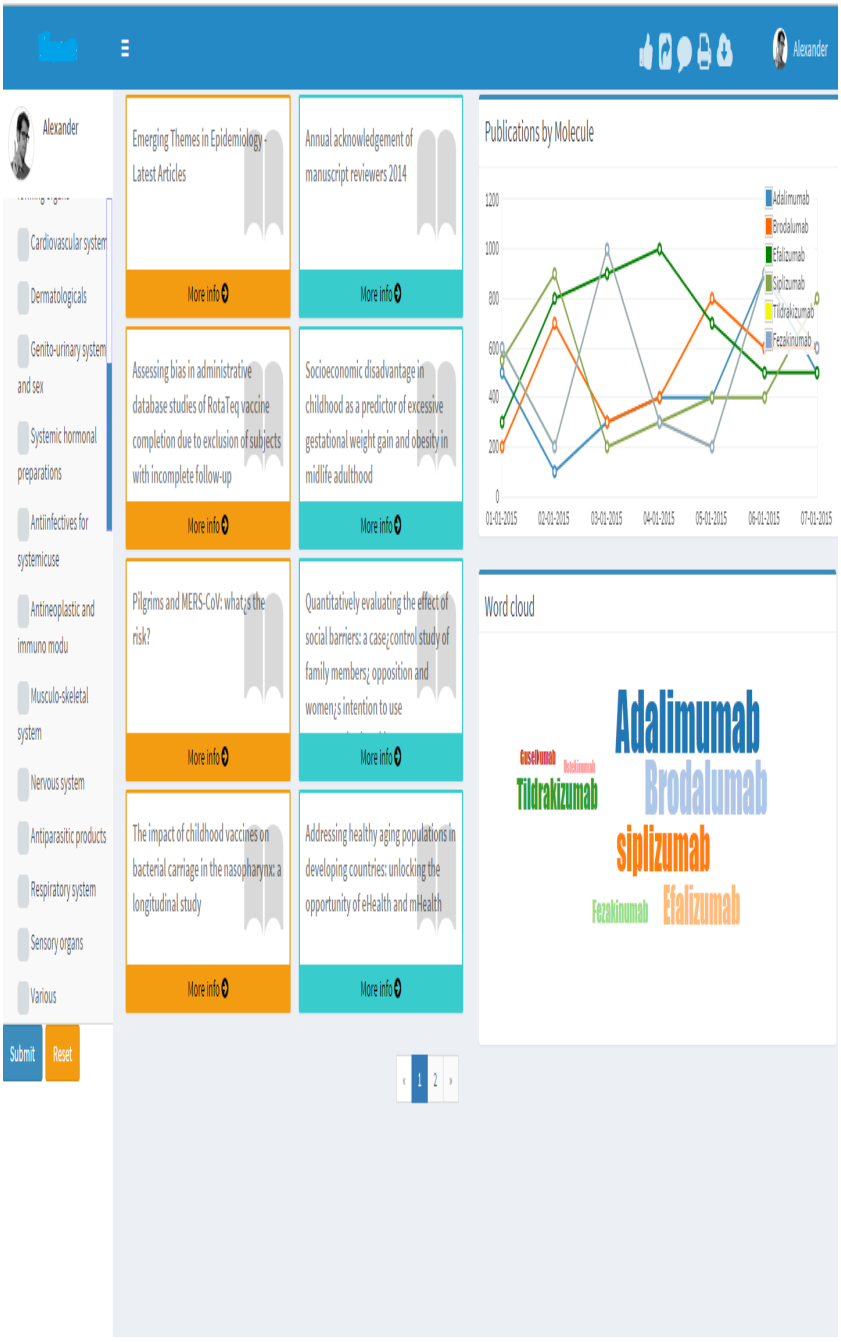

Fig 5: Publications Screenshot

The above diagram (Fig 5) shows the result of Web Crawler visited URLs & identified hyperlinks of recent publications of the Pharmaceutical industries.

7.2 Patents Screenshot

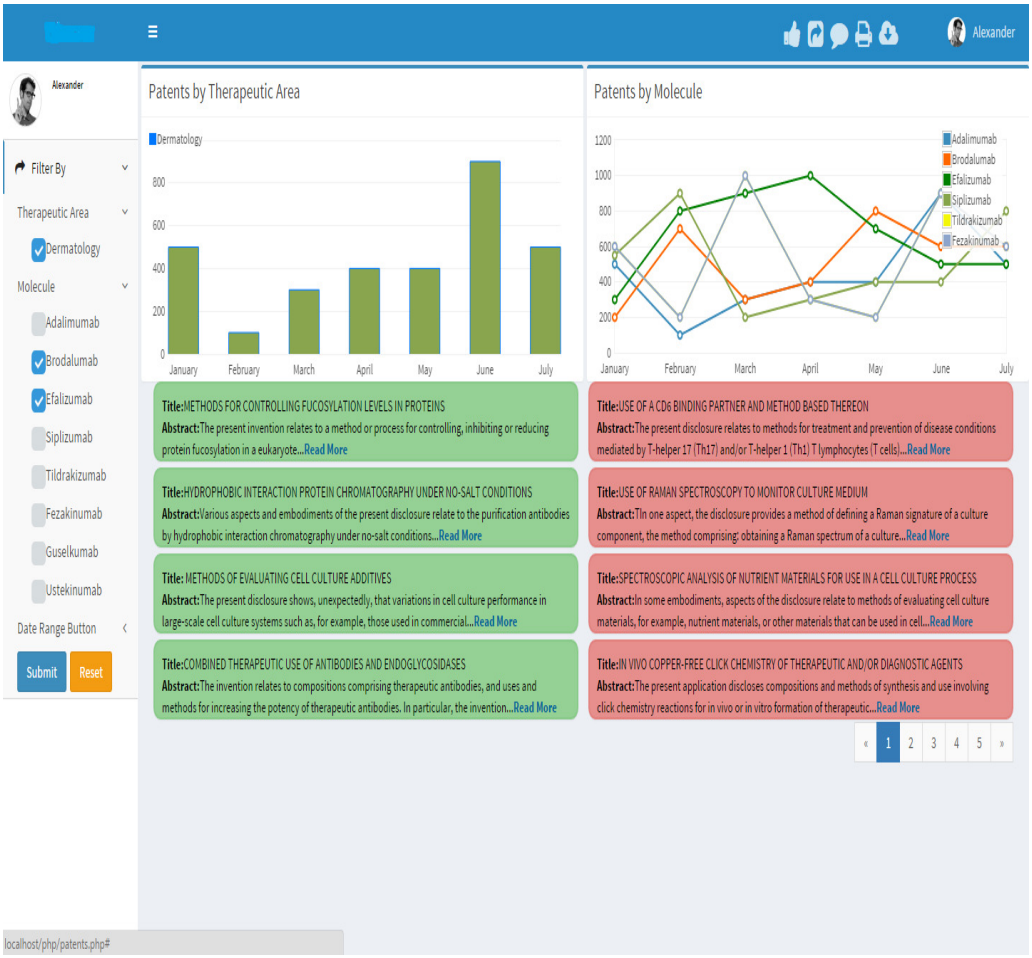

Fig 6: Patents Screenshot

The above diagram (Fig 6) shows the result of Web Crawler visited URLs and identified hyperlinks of recent patents of the Pharmaceutical industries.

8. Conclusion

Overall this is very informative application which gives a clear vision to the user to fetch the information on recent publications & patents of pharmaceutical industry. It's very efficient in this materialistic world, as it saves your time by choosing the right category. The Patents & Publications Web Scraping is a web-based application for fetching. Publications and patents in pharmaceutical industry which provide customized solutions to meet industry needs. This software is developed successfully and is also tested successfully by taking "test cases". It has required options, and it is user friendly by which the user can perform the required operations. This software is developed using PHP for scripting and query processing and MySQL for back end in Windows operating system. Some of the goals achieved by the software are instant access, user friendly, portable and flexible.

9. Future Enhancements

Developing a system is complicated which meets all the requirements of the user. As the system is used by the User, the requirements keep on changing. Some of the

future enhancements that can be done to the present project usually are:

A . As the technology changes, it is possible to update the system and can be adjustable to preferred environment.
B. Based on the long term security problems, security can be increased using prominent technologies like single sign-in.

References

[1] Internet & World Wide Web How to program 3rd edition by Deitel & Deitel
[2] Kevin Tatroe, Peter MacIntyre (2013), **"**Programming PHP**"**, O'Reilly Media.Inc. CA 95472.
[3] www.w3schools.com
[4] www.getbootstrap.com
[5] www.github.com
[6] www.stackoverflow.com
[7] http://www.ncbi.nlm.nih.gov/pubmed

Author Profile:

Sushitha S: Working as Asst Prof in Department Of MCA, NMIT Publications-ERCICA

Vijayalakshmi S Katti,: Working as Asst Prof in Department Of MCA, NMIT

Sowmya H N: Working as Asst Prof in Department Of MCA, NMIT

Samanvita N: Working as Asst Prof in EEE Dept NMIT, Pursuing PhD in VTU in the field of Neuro- Fuzzy control. Publications-IRED, IJEEE, IEEE