IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Implementation of Web Content Extraction of Structured Data Using DotNet Framework

M.Florence Dayana[1*], Dr.M.Chidambaram[3]

[1*] Department of Computer Science, A.V.V.M Sri Pushpam College (Autonomous), Poondi, Thanjavur, India
[2] Department of Computer Science, Rajah Serfoji Government College (Autonomous), Thanjavur, India

*Corresponding Author: florencedayana@gmail.com

Available online at: www.ijcseonline.org

Accepted: 14/May/2018, Published: 31/May/2018

***Abstract*—** This paper deals in Web Content Mining for extraction of structured data. While perusing the web, the client needs to experience numerous pages of the Internet, channel the information and download related records and documents. This errand of seeking and downloading is tedious. Now and again the look inquiries call for particular choice, say, restricting inquiry to few connections. To lessen the time spent by clients, a web extraction and capacity apparatus has been composed and executed in .Net framework, that robotizes the downloading task from a given client question. The Test Scenario has been given different catchphrases. The present work can be a valuable contribution to Web Manipulators, Staff, Students and Web Administrators in an Academic Environment.

***Keywords*—** Web Content Mining, Structured Data, Web Data Extraction, HTML, Data mining, Web Mining.

I. INTRODUCTION

The measure of data accessible in the Web increments persistently and in this manner requires an exact printed portrayal of the examined Web pages. Notices, navigational components muddle the assignment of separating the Web page's content. Web content extraction is worried about extricating the significant content from Web pages by expelling inconsequential literary clamor like commercials, navigational components, contact and copyright notes. Utilization of information mining methods to the World Wide Web is alluded to as Web Mining. Web Mining can be extensively characterized as the robotized revelation and examination of valuable data from the web archives and administrations utilizing information mining methods. It finds conceivably helpful and beforehand obscure data or information from web information. Web mining utilizes the system of information mining into the records on the World Wide Web.

Web Mining assignments can be arranged into three kinds in light of which part of the Web to mine. They are Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining manages finding helpful data or information from web page contents. Web content comprises of various kinds of information, for example, printed, picture, video, sound, metadata and hyperlinks. Web Structure Mining centers on the hyperlink structure of the web. Web Usage Mining (otherwise called Web Log Mining), is the way toward removing fascinating examples in Web get to logs. Use data can be utilized to restructure a web webpage keeping in mind the end goal to better serve the requirements of clients of the website. Information on the web is ordered into three sorts to be specific Structured, Unstructured and Semi structured. Structured information is as rundown, tree, table and Database. An Unstructured information contains free content and does not take after a specific language structure for portrayal for a web page. Semi-Structured information don't have a pre-characterized structure. HTML and XML goes under this classification.

The paper is composed as takes after. Segment II manages Objectives and Motivation, segment III depicts the problem description, segment IV manages an algorithmic techniques for web information extraction called Quest_Support, segment V indicates the Implementation and segment VI conclude the paper.

II. RELATED WORKS

As of late, it has been seen that the consistently intriguing and up and coming distributing medium is the World Wide Web[7]. A significant part of the web content is unstructured so assembling and comprehending such information is extremely dreary. Web servers worldwide produce a huge measure of data on web clients' perusing exercises. A few researchers have examined these supposed web get to log information to better comprehend and describe web clients.

Information can be enhanced with data about the substance of went by pages and the cause (e.g., geographic, hierarchical) of the solicitations. The objective of this venture is to break down client conduct by mining advanced web get to log information. The few web utilization digging strategies for extricating valuable highlights is examined and utilize every one of these systems to bunch the clients of the area to contemplate their practices completely. The commitments of this proposition are an information advancement that is substance and root based and a treelike perception of continuous navigational groupings. This representation takes into consideration an effortlessly interpretable tree-like perspective of examples with featured significant data. The results of this undertaking can be connected on differing purposes, including showcasing, web content prompting, (re-)organizing of web locales and a few other E-business forms, similar to suggestion and publicist frameworks. It additionally rank the best significant archives in view of Top K inquiry for viable and proficient information recovery framework. It channels the web reports by giving the important substance in the search engine result page (SERP).

In this paper, they consider the issue of mining the educational structure of a news Web webpage that comprises of thousands of hyperlinked archives. They characterize the educational structure of a news Web webpage as an arrangement of file pages (or alluded to as TOC, i.e., chapter by chapter list, pages) and an arrangement of article pages connected by these TOC pages. Based on the Hyperlink Induced Topics Search (HITS)[9] calculation, it propose an entropy-based analysis (LAMIS) instrument for dissecting the entropy of grapple messages and connections to dispose of the excess of the hyperlinked structure with the goal that the mind boggling structure of a Web website can be refined. Nonetheless, to build the esteem and the openness of pages, a large portion of the substance destinations have a tendency to distribute their pages with intrasite excess data, for example, route boards, ads, duplicate declarations, and so forth. To additionally kill such repetition, it propose another component, called InfoDiscoverer, which applies the refined structure to distinguish sets of article pages. InfoDiscoverer additionally utilizes the entropy data to break down the data measures of article sets and to remove educational substance hinders from these sets. The result is valuable for search engines, data operators, and crawlers to list, remove, and explore critical data from a Web website. Investigations on a few genuine news Web destinations demonstrate that the exactness and the review of the methodologies are much better than those acquired by ordinary techniques in mining the enlightening structures of news Web locales. On the normal, the expanded LAMIS prompts noticeable execution change and builds the exactness by a factor going from 122 to 257 percent when the coveted review falls in the vicinity of 0.5 and 1. In examination with manual heuristics, the accuracy and the review of InfoDiscoverer are more noteworthy than 0.956.

This paper displays a Bayesian learning structure for adjusting data extraction wrappers with new quality disclosure, lessening human exertion in removing exact data from concealed Web destinations. The approach goes for consequently adjusting the data extraction information already gained from a source Web webpage to another inconspicuous website, in the meantime, finding beforehand concealed properties. Two sorts of content related pieces of information from the source Web website are considered. The main sort of piece of information is gotten from the extraction design contained in the already learned wrapper. The second sort of piece of information is gotten from the already extricated or gathered things. A generative model for the age of the website free substance data and the webpage subordinate format configuration of the content parts identified with trait esteems contained in a Web page is intended to bridle the vulnerability included. Bayesian learning and expectation-maximization (EM) methods are produced under the proposed generative model for recognizing new preparing information for learning the new wrapper for new inconspicuous locales[13]. Already inconspicuous traits together with their semantic names can likewise be found by means of another EM-based Bayesian learning based on the generative model. We have directed broad analyses from in excess of 30 true Web locales in three distinct spaces to exhibit the viability of our structure.

Deep Web substance are gotten to by questions submitted to Web databases and the returned information records are enwrapped in progressively created Web pages (they will be called deep Web pages in this paper)[14]. Extricating organized information from deep Web pages is a testing issue because of the hidden complex structures of such pages. As of recently, countless have been proposed to address this issue, yet every one of them have natural restrictions since they are Web-page-programming-dialect subordinate. As the well-known two-dimensional media, the substance on Web pages are constantly shown frequently for clients to peruse. This inspires us to look for an alternate route for deep Web information extraction to beat the constraints of past works by using some fascinating normal visual highlights on the deep Web pages. In this paper, a novel vision-based approach that is Web-page-programming-dialect free is proposed. This approach principally uses the visual highlights on the deep Web pages to actualize deep Web information extraction, including information record extraction and information thing extraction. We likewise propose another assessment measure revision to catch the measure of human exertion expected to create culminate extraction. The investigations on an extensive arrangement of Web databases demonstrate that the proposed vision-based

approach is exceedingly viable for deep Web information extraction.

As a result of the snappy development of virtual learning and working up the exact information wishes of the clients, the data mining process has an imperative position to separate the accommodating information from that huge amount of information. The extraction of those information can likewise be expert the utilization of other learning mining strategies. The basic role of doing pattern mining is to improve shrewdness revelation designs for the productive benefit as much as possible from discovered pattern and tail it in space of printed content mining[1]. In learning mining gathering, so much analysis works of art fixate of consideration on making a productive pattern discovering set of standards which accompany technique similar to consecutive pattern mining basic stock mining and close successive digging for mining accommodating styles. However there's an expansive issue to discover and supplant productive pattern. In proficient pattern disclosure and utilize strategies there are essential issues. Those are:

- Low recurrence and
- Trend confusion disadvantage

The general assessment of a proposed gadget is intended to manage the issues of low recurrence and pattern distortion of pattern disclosure approach. The program endeavors to disentangle the overall strategy issues and analyse the result produced by method for slant sending and pattern organization mind slant co-commonness techniques.

III. Objectives And Motivation

The present work is planned to meet the accompanying goals:

- To plan and actualize a calculation for Web Data Extraction to download the documents and place them in their separate envelopes like XML, HTML and Text information.
- To scan for a question in light of sifting parameters and the quantity of connections to be found ought to be taken as client input. The connections acquired as yield must be downloaded onto the neighborhood machine and the documents must be ordered into envelope in light of record compose.
- To distinguish the Academic Search related capacities where Web Data Extraction calculation can be connected viably.

Most web extraction devices incorporate a web crawler. Web slithering alludes to the recursive procedure of downloading web pages from an arrangement of URL's, extricating the connection found inside them and adding them to the arrangement of connections holding up to experience a similar procedure. The crept pages are then ordered by content. These documents are then served to the clients in view of the client's watchwords/question.

Web crawling includes looking through a vast arrangement space which requires a great deal of time, hard plate space and parcel of use of assets when all is said in done. Thus utilizing existing web search tools as the back-end is more pragmatic way to deal with Web content extraction. It gives the open door for the client to store the indexed lists in a nearby server or open catalog. Thusly looks including similar watchwords can be performed locally as opposed to rehashed seeks and successive download. This spares time and assets all things considered.

IV. Problem Description

The Flow chart for the proposed framework Quest Support is appeared in Figure 1. The watchwords are taken as contribution from the client either from the document or GUI. It produces URL from the hunt inquiry given and the pursuit alternatives chose by the client. Utilizing Wget programming, the documents are downloaded. GNU Wget is a free programming bundle for recovering records utilizing HTTP, FTP. It is a non-intelligent charge line instrument that is called from different projects. The downloaded list items in HTML frame are changed over to XML utilizing HTML Cleaner. HTML on the web isn't very much shaped and not reasonable for additionally preparing in light of the fact that it contains missing statements and unclosed labels in the record. HTML Cleaner is an open source HTML parser written in Java which acknowledges HTML record and redresses singular components and delivers all around framed XML. XPath devices are utilized to discover the connections in the XML archive. A connection is picked and Wget is utilized to download the report and spare the yield messages to a record. The document augmentation of the downloaded record is resolved from the MIME write. The record is put in organizer of particular document expansion.

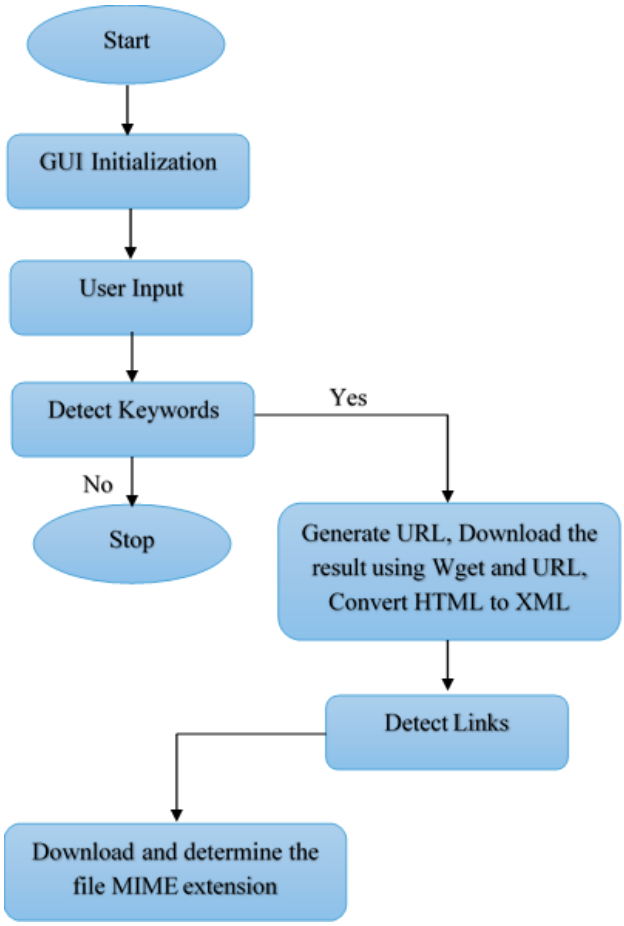

Fig.1: Quest Support Flow Diagram

V. ALGORITHM: QUEST_SUPPORT

The present work deals with implementation of an algorithm Quest_Support.

Understanding of Inputs: Keywords, Number of links.
Production of Output: Downloaded files in respective folder.

Steps Followed in Quest_Support:
Step 1: Generate the search URL from the keywords and various options provided by the user.
Step 2: Use Wget to retrieve the search results as HTML file using the search URL.
Step 3: Use HTMLCleaner (an external party utility), to convert the HTML to a clean XML document.
Step 4: Use XML XPath tools to find the links in the document.
Step 5: Choose a link and use Wget to download the documents. Also save the output messages to a file.
Step 6: Determine the MIME type of the downloaded file from the output file. And then check if the type is associated with a file extension from req/mime.txt.
Step 7: Repeat the same for other links.
Step 8: If the keywords have been specified by file input, repeat this process for other keywords (from Step 1).

A. Algorithm Working Procedure

1. User enters keyword and presses search button. A thread is spawned to do the search (so that it doesn't block the UI). All the steps here onwards happen within the thread.
2. From the GUI options, a link to Google with the search options is generated. For example the link would be:
http://www.google.com/search?hl=en&safe=active&q=JSON+s hema&num=10&start=0
   - Where 'hl=en' denotes the language preference is English.
   - 'safe=active' sets safe search.
   - 'q' = the keywords.
   - 'num' is the number of search results to show.
   - 'start' indicates the search result to start from. That is we can start from the 10th search results and display the next 20 links by setting num=20. Results 10 to 30 will be shown in that case.
3. Send the link to Wget and Wget will extract the resulting html page which contains the search results.
4. Convert HTML to XML. For the purpose a third party utility known as 'HTMLCleaner' is used to convert HTML to XML.
5. Use XPath to extract the links (search results) from the Google search page.
6. Begin iterating through the links.
7. Download the link with Wget and retrieve the MIME type of the file by Quest Support.
8. Go to step 6 and iterate through the rest of the links.
9. Go to step 1 and check if any more queries are to be searched. When no more queries remain, quit the thread.

VI. IMPLEMENTATION

The Quest Support algorithm has been implemented using DotNet Platform Architecture as a Front End and SQL Server as a Back End.

A University comprises of various sorts of clients, for example, Faculty, Students, Staff, Web Administrator and Librarian. Every client has their necessities while perusing data on the Internet. Web Mining helps in removing data as indicated by client's inclinations.

Table 1 shows the applicability of Quest Support Algorithm in a University Environment.

Table 1: User Categories and Tasks

User Categories	Tasks
Student	Literature Review.
Faculty	Lecture Notes.
Librarian	Building of Academic Related Data Warehouse.
Web Administrator	Usage Analysis from Log files.
Staff	Search on specific topics.

The experimental results for Web Extraction and Storage functions are illustrated in Table 2.

Table 2: Experimental Results for Web Extraction and Storage functions

Keywords	No. of links	Size of downloaded files in KB	Total time taken in milliseconds
Web Mining	48	15094	296688
Web Structure Mining	50	28059	595757
Web Content Management	50	2125	175498
Web Content Mining	50	22410	147736
Masters in Computer Engineering-Europe	10	409	27232
Post Graduate Degrees USA	10	632	23732
IEEE Conference UAE	5	335	19493
IETF Conference	5	264	13100

VII. CONCLUSION

The present work dissects the methods for extricating web information and putting away them utilizing DotNet and SQL Server. The application has been tried effectively utilizing Academic Search watchwords. This work can be stretched out further to deal with picture, sound, video and information.

REFERENCES

[1] U. Moulali, V. Sasidhar, "*Competent pattern innovation designed for textual content mining*", 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 572 – 577.

[2] Farman Ali, Pervez Khan, Kashif Riaz, Daehan Kwak, Tamer Abuhmed, Daeyoung Park, Kyung Sup Kwak, "*A Fuzzy Ontology and SVM–Based Web Content Classification System*", IEEE Access, Vol. 5, pp. 25781 – 25797.

[3] Yeongsu Kim, Seungwoo Lee, "*SVM-based web content mining with leaf classification unit from DOM-tree*", 2017 9th International Conference on Knowledge and Smart Technology (KST), pp. 359 – 364.

[4] Tak-Lam Wong, Wai Lam, "*Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach*" IEEE Trans. on Knowledge and Data Engineering, Vol. 22, No. 4, pp. 523 – 536, 2010.

[5] Charu C. Aggarwal, Yuchen Zhao, Philip S. Yu, "*On the Use of Side Information for Mining Text Data*", IEEE Trans. on Knowledge and Data Engineering, Vol. 26, No. 6, pp. 1415 – 1429, 2014.

[6] Kaveh Hassani, Won-Sook Lee, "*Adaptive animation generation using web content mining*", 2015 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 1 – 8.

[7] G. Dhivya, K. Deepika, J. Kavitha, V. Nithya Kumari, "*Enriched content mining for web applications*", 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1 – 5.

[8] Tao Jiang, Ah-hwee Tan, Ke Wang, "*Mining Generalized Associations of Semantic Relations from Textual Web Content*", IEEE Trans. on Knowledge and Data Engineering, Vol. 19, No. 2, pp. 164 – 179.

[9] Hung-Yu Kao, Shian-Hua Lin, Jan-Ming Ho, Ming-Syan Chen, "*Mining Web informative structures and contents based on entropy analysis*", IEEE Trans. on Knowledge and Data Engineering, Vol. 16, No. 1, pp. 41 – 55, 2004.

[10] F. de la Rosa Troyano, S. del Pozo Hidalgo, R. Martinez Gasca, "*Analysis and Visualization of Scientific Communities with Information Extracted from the Web*", IEEE Latin America Transactions, Vol. 3, No. 1, pp. 56 – 61.

[11] I-Jen Chiang, Charles Chih-Ho Liu, Yi-Hsin Tsai, Ajit Kumar, "*Discovering Latent Semantics in Web Documents Using Fuzzy Clustering*", IEEE Trans. on Fuzzy Systems, Vol. 23, No. 6, pp. 2122 – 2134.

[12] Hao Ma, Irwin King, Michael R. Lyu, "*Mining Web Graphs for Recommendations*", IEEE Trans. on Knowledge and Data Engineering, Vol. 24, No. 6, pp. 1051 – 1064, 2012.

[13] Tak-Lam Wong, Wai Lam, "*Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach*", IEEE Trans. on Knowledge and Data Engineering, Vol. 22, No. 4, pp. 523 – 536.

[14] Wei Liu, Xiaofeng Meng, Weiyi Meng, "*ViDE: A Vision-Based Approach for Deep Web Data Extraction*", IEEE Trans. on Knowledge and Data Engineering, Vol. 22, No. 3, pp. 447 – 460.

