November-December 2019
ISSN: 0193-4120 Page No. 5181 - 5184

Extraction of XML Documents using Web Data Mining

P. N. Santosh Kumar,

Assoc Prof, ECM, SNIST, Email:Pnsk47@gmail.com

V. Radha,

Prof of CSE, IDRBT, Hyd,India

Article Info
Volume 81
Page Number: 5181 - 5184
Publication Issue:
November-December 2019

Article History

Article Received: *5 March 2019*
Revised: *18 May 2019*
Accepted: *24 September 2019*
Publication: *24 December 2019*

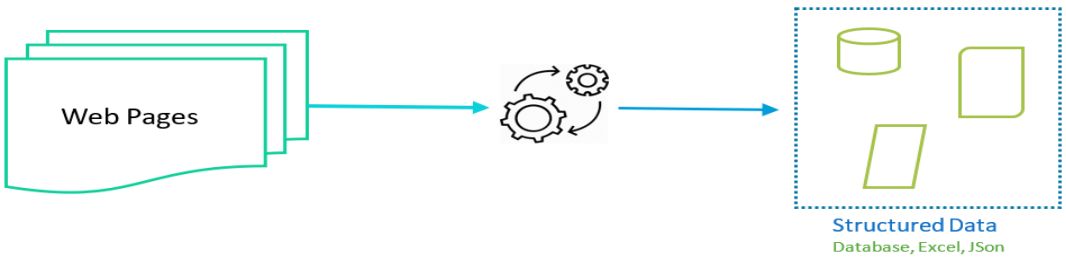
Abstract

Mining that is performed on main web data is known as web structured mining. One sort of data mining is known as web structured mining. As the usage of Internet has been increased drastically, there are many web pages. All those web pages or the web data has to be mined for further usage. Data generated in webin Day life has become need of the hour importance of web mining is growing along with the massive volumes. Based on its similarity, Web data clustering in this the collection of Web documents in to clusters is possible with this organization. So many researchers has put an eye because similar documents available for versatile applications. Many Researchers around proposed extract web contents, few times they will fail to handle dynamic data. Web content extraction algorithms are pivotal importance to extract from web sources like useful contents. Hence, proposed a new method alternate way for web content extraction.

Keywords: *XML extraction, data mining, web structured mining*

I. Introduction:

Data extraction is a process of obtaining data from the web pages in a cyclic manner without making any changes in the content along with transformation of the data that is extracted from the database or in another file format that is required for other applications.



The web scrapping is consisting of three functionalities:

1. Crawling also known as web interaction mainly involves relocation to the generally (pre-determined target) i.e. web pages consisting of the information that is required.
2. The execution wrapper generation is a process for recognizing the data that is wanted on the pages targeted as well as

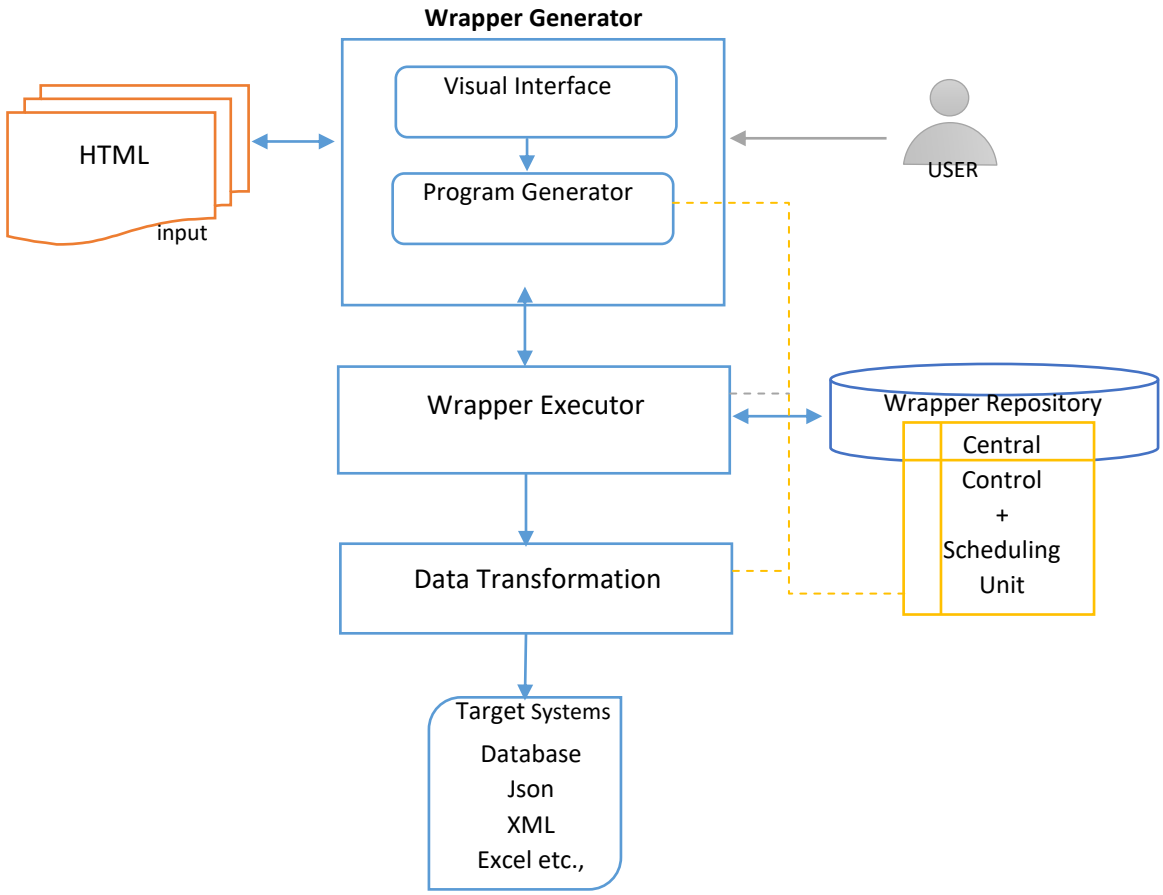
November-December 2019
ISSN: 0193-4120 Page No. 5181 - 5184

extracting the data along with converting them into format.

Since, output is generated through program, it can also be utilized as a source of input for another system

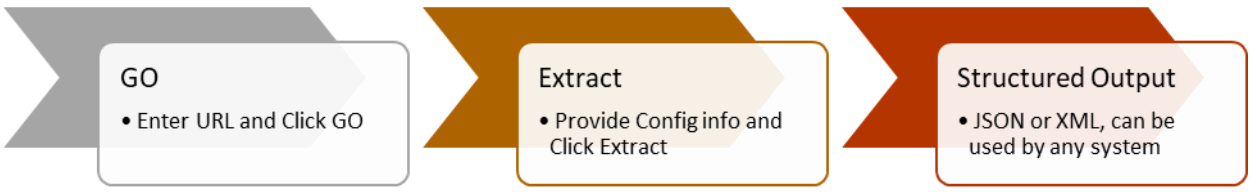
II. Implementation:

2.1 HIGH LEVEL ARCHITECTURE



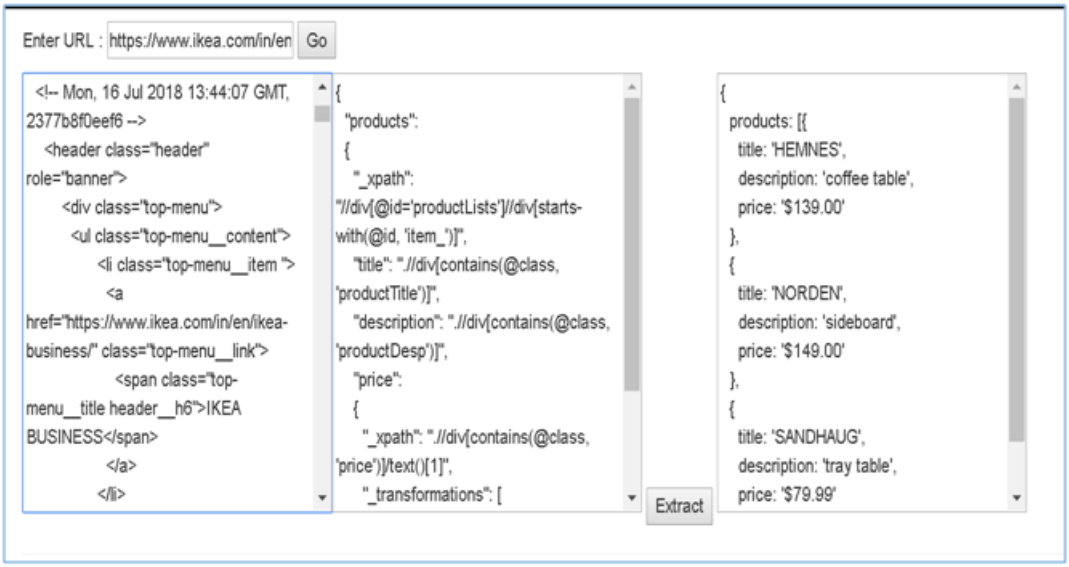
2.2 PROJECT EXECUTION

This is a simple web website developed in C# using Visual Studio which gets html, provided correct URL. The below screenshot describes the same.



- Enter URL and click **Go** button. It fetches the HTML and shows on the below section (left most division)
- Take appropriate JSON configuration file, which is designed to extract required information from the HTML script and paste in second section of division
- Now click **Extract** button, it parses the information generates JSON output which can be inserted into any database or can be used by any other system.

November-December 2019
ISSN: 0193-4120 Page No. 5181 - 5184



1.3. CODE EXPLANATION

2.3.1 CRAWLING WEBSITE

We should first obtain a WebRequest object to open a Stream. Create function of the HttpWebRequest class will be called by the object URI is nothing but a URL.

The below code is demonstrated how this can be accomplished as follows:

```
private string GetHTMLCode(string url)
{
    HttpWebRequest myRequest = (HttpWebRequest)WebRequest.Create(url);
    myRequest.Method = "GET";
    WebResponse myResponse = myRequest.GetResponse();
    StreamReader sr = new StreamReader(myResponse.GetResponseStream(), System.Text.Encoding.UTF8);
    string result = sr.ReadToEnd();
    sr.Close();
    myResponse.Close();

    return result;
}
```

2.3.2. EXTRACT DATA FROM CONFIG

After getting the html from desired website, parse it with JSON configuration based on the requirement.

```
private string GetExtractedData(string html, string configJson)
{
    var config = StructuredDataConfig.ParseJsonString(configJson);

    var openScraping = new StructuredDataExtractor(config);
    var scrapingResults = openScraping.Extract(html);

    return JsonConvert.SerializeObject(scrapingResults, Formatting.Indented);
}
```

2.3.3 CALLING ORDER

On HttpWebRequest and HttpWebResponse classes, different operations can be performed. A precise order must be followed for all of these operations. Beforestarting to work with HTML response, the information desired must be set initially.

Order required to be followed is as follows:

I. HttpWebRequest object to obtain
II. HTTP headers to Set, if required
III. POST data, POST request
IV. HttpWebResponse object to obtain
V. HTTP response headers reading is required
VI. HTTP response data reading is required

2.3.4. DATA MINING APPLICATIONS

Web extractions done by us are used primarily in mining of web data that can be utilized further in the following sectors as follows:

- Pharmaceutical, Research along with Healthcare.
- Price comparisons for Business improvements.
- Media & Publishing.
- Information Technology Sector.
- Financial markets Banks, Insurance.
- Customer Retention, Market Analysis, Fraud Detection.
- Social Web Mining like Twitter.
- Public Administration along with documents that are legal.

III. Conclusion:

This paper concentrates on extract valuable information from the web through a combination of various methods for different text mining and information retrieval.

Also covered C# language and its application in web miningalong with general web data extraction system. This program ecommerce, Stock Exchange or Forex Trading etccan hop (or scrape) and extract real time information from web specially required in case.

References:

[1] S. Mahesha, Dr. M S Shashidhara, and Dr. M. Giri, “An Implementation of Web Content Extraction Using Mining Techniques”, IFRSA International Journal of Data Warehousing & Mining |Vol 2|issue4|November 2012.

[2] ErdinçUzun, HayriVolkanAgunTarıkYerlikaya,” A hybrid approach for extracting informative content from web pages”, E. Uzun and al. / Information Processing and Management 49 (2013) 928–944.

[3] OlatzArbelaitz, IbaiGurrutxaga, AizeaLojo, Javier Muguerza, Jesús Maria Pérez, IñigoPerona, “Web usage and content mining to extract knowledge for modelling the users of the BidasoaTurismo website and to adapt it”, O. Arbelaitz and al. / Expert Systems with Applications 40 (2013) 7478–7491.

[4] ZHANG Bin, WANG Xiao-fei,, “Content extraction from Chinese web page based on title and content dependency tree”, The Journal of China Universities of Posts and Telecommunications. October 2012, 19(Suppl.2): 147–151, www.sciencedirect.com/science/journal/10058885.

[5] Markus Schedl, GerhardWidmer, Peter Knees, TimPohle, “A music information system automatically generated via Web content mining techniques,”M. Schedl and al. / Information Processing and Management 47 (2011) 426–439.