Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 5, Issue 04, April -2018

AN IMPROVED APPROACH OF WRAPPER GENERATION TECHNIQUES FOR WEB SOURCES

Sweta

Dept. of P.G. Studies and Research in Computer Science, GUK

Abstract— *The World Wide Web has more and more online Web databases which can be searched through Web query interfaces. All the Web databases make up the deep Web. Often the retrieved information is enwrapped in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index by traditional crawler based search engines, such as Google and Yahoo. The topic of Web data extraction has received a lot of attention in recent years and most of the proposed solutions are based on analyzing the HTML source code or the tag trees of the Web pages. Web data extraction is the process of extracting user required information from websites. The web document contains data which is not in structured format. Specific data is able to be extracted from all these Web sources in order to be used by other users or applications. The word web data extraction means the extraction of data that is present in the web documents in HTML format and removing the unwanted things such as tags, advertisements, videos and so on from web sources.*

Keywords- *APIs; HTML; web sources; web data extraction; wrapper*

I. INTRODUCTION

Web sites are an integral part of the technological world today. And the rate at which the websites are evolving over the years has been phenomenal. As per the inputs from the website Internet Live stats, there are more than one billion websites in the internet, which is clearly an indication of the exponential rate at which the websites are being added every second to the internet. The use of the web sites to derive data is also growing like never before, across countless sectors. Companies need data for various purposes like obtaining new customers, tracking industry trends, analysis for business purposes, understanding government regulations and more. Automating the web data extraction [1] is the best approach and many organizations are leading the way in finding path-breaking solutions to achieve a reliable way to achieve automation. It is extremely beneficial for harvesting structured information with specific data types and results in a reduction of redundancy, elimination of the manual errors, cost overhead. Also, website structure changes are monitored, providing access to the right data at desired intervals.

II. A CLASSIFICATION OF TYPE OF THE DATA

1. Free text: This type of text could be found in natural language texts, for example magazines or pharmaceutical research abstracts. Patterns involving syntactic relations between words or semantic classes of words are used to extract data from this type of sources.

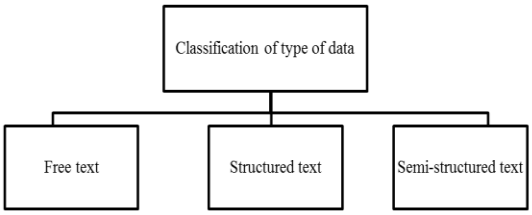

Figure1. Classification of type of data

2. Structured text: This type of text is defined as textual information in a database or file following a predefined and strict format [2]. To extract this kind of data we have to use the format description.

3. Semi-structured text: This type of text is placed in an intermediate point between unstructured collections of textual documents and fully structured tuples of typed data. To extract data we use extraction patterns that are often based on tokens and delimiters, for example the HTML-tags.

III. DIFFERENT WAYS TO PERFORM THE EXTRACTION

There are three ways to perform web data extraction as shown in "Fig.2".
1. Manual extraction of the data
2. Use a built API
3. Use of an Automatic wrapper

Manual extraction of the data
Manual extraction is the most precise option to extract data as in this approach the data fields of user interest is considered. The necessity to treat elements in an individual way takes a lot of time when treating large amount of data. It is a good option for small and concrete data extractions.

Use a built API
An API belongs to the owner of the Web page where user wants to extract data. Normally, APIs are found in few specific numbers of Web pages and its use and supply are limited by the specifications of the owner. To use APIs user have to take a look at the documentation and the method list of the owner.

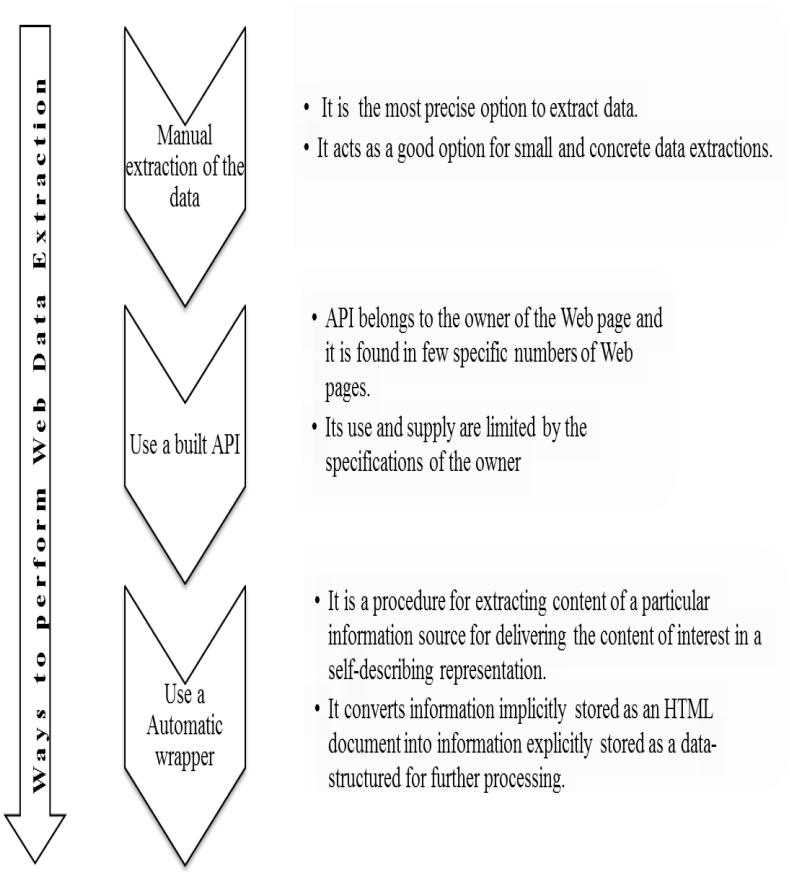

Figure 2.Ways to perform Web Data Extraction

Wrapper
A wrapper let the end-user use a set of methods without the necessity to have support of the owner of the Web page and with independence of the content. It is a procedure that is designed for extracting content of a particular information source for delivering the content of interest in a self-describing representation. Its target is to convert information implicitly stored as an HTML document into information explicitly stored as a data-structured for further processing. A wrapper for a Web source accept queries about information in the pages of that source, fetches relevant pages from the source and extracts the requested information and returns the result. The construction of a wrapper can be done manually or by using a semi-automatic or automatic approach as shown in "Fig.3".

The Manual generation of wrapper involves the writing of ad-hoc code. It requires understanding the structure of the document and translating it into program code. The task is not trivial and hand coding could be tedious and error-prone.
The semi-automatic wrapper generation benefits from support tools to help design the wrapper. By using a graphical interface the user can describe which the important data fields to be extracted. A specific configuration of the wrapper should be done for each Web page source as the content structure varies from each other. Expert knowledge in wrapper coding is not required at this stage, and it is also less error prone.

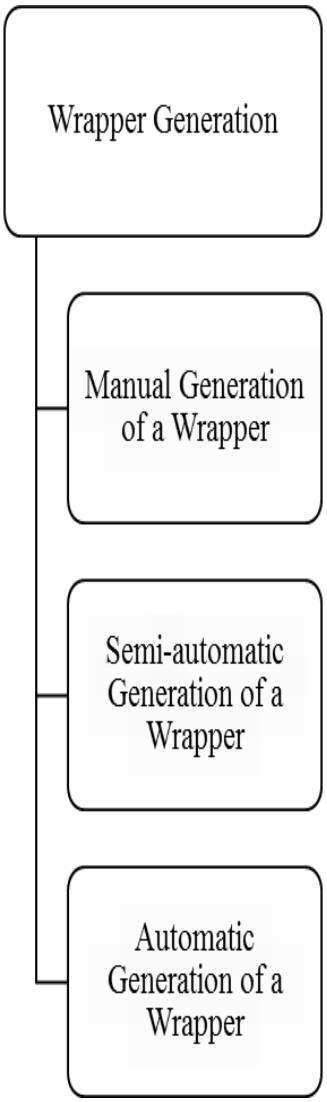

Figure 3. Wrapper Generation Techniques

The automatic wrapper generation [3] uses machine-learning techniques, and the wrapper research community has developed learning algorithms for a spectrum of wrappers. This kind of wrapper require a minimum intervention of human experts and systems which go through a training phase, where it is fed with training examples, and, in many cases, this learning has needs to be supervised.

Generally, the steps to extract information using a wrapper are the following:
1. Load the information of the source page
2. Transform the source page for its posterior treatment
3. Identify the appearing elements
4. Filter these elements
5. Export of the final data to an output format

The first and last steps are common to all types of wrappers as it needs a data Input and a data output to perform a data extraction.

IV. CONCLUSION

As the entire world is rapidly moving towards a data-centric operational model, it's high time to evaluate data requirements and get started with extracting relevant data from the web to improve the efficiency in various fields. Web data extraction is the process of automatically converting Web resources into a specific structured format. Its main purpose is to make web data available for subsequent manipulation or integration steps. This paper discuss about the classification of types of data in data extraction, different ways to perform data extraction and available techniques for wrapper generation.

V. REFERENCES

[1] EmilioFerrara, PasqualeDeMeo, GiacomoFiumara, RobertBaumgartner:" Web data extraction, applications and techniques: A survey" ACM Transactions on Computational Logic, Vol. V, No. N, June 2010, Pages 1–20.
[2] Arvind Arasu, Hector Garcia-Molina:" Extracting Structured Data from Web Pages", SIGMOD '03 Proceedings of the 2003 ACM SIGMOD international conference on Management of data Pages 337-348.
[3] Kristina Lerman University of Southern California:" Automatic Wrapper Generation and Data Extraction". August 25, 2010.
[4] Crescenzi V., Mecca G., Merialdo P. RoadRunner: towards automatic data ex- traction from large Web sites. In Proceedings of the 27th International Conference on Very Large Data Bases, Roma, 2001, 109-118.
[5] Cohen W. W., Hurst M., Jensen L. S. A flexible learning system for wrapping tables and lists in HTML documents. In Proceedings of the 11th International World Wide Web Conference, Budapest, 2002, 232-241
[6] Arlotta L, Crescenzi V, Mecca G, et al. Automatic annotation of data extracted from large Web sites. In Proceedings of the 6th International Workshop on Web and Databases, San Diego, 2003, 7-12
[7] Hui Song, Suraj Giri, Fanyuan Ma. Data Extraction and Annotation for Dynamic Web Pages.
[8] Cope J, Craswell N., Hawking D. Automated discovery of search interfaces on the Web. In Proceedings of the 14th Australasian Database Conference (ADC 2003), Adelaide, 2003, 181-189
[9] Zhang Z., He B., Chang K.C. Under- standing Web query interfaces: best-effort parsing with hidden syntax. In Proceedings of the 23rd ACM SIGMOD International Conference on Management of Data, Paris, 2004, 107-118
[10] Betreuender Hochschullehrer: Prof. Dr. Erhard Rahm Betreuer: Dr. Andreas Thor: A comparison of HTML-aware tools for Web Data extraction .Leipzig, September, 2008.
[11] Liu L, Pu C, Han W. XWRAP: An XML-enabled wrapper construction system for Web information sources. In Proceedings of the 16th International Conference on Data Engineering, San Diego, 2000, 611-621.

