ISSN: 2321-7782 (Online)
Volume 2, Issue 9, September 2014
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

A Survey of Web Data Extraction Techniques

Vidya.V.L
PG Scholar, Department of Computer Science
Mohandas College of Engineering and Technology, Thiruvananthapuram
Kerala – India

Abstract: ***Web data extraction has a wide range of applications, but web data extraction is a major problem that has been studied by means of different scientific tools. This survey provides an overview of the web data extraction techniques, mainly focuses on automatic web data extraction techniques and also compares them based on the methods used for web data extraction.***

Keywords: data extraction, wrapper induction, Data alignment, pattern mining, Tree merging.

I. INTRODUCTION

Internet is the biggest information source on the planet .It is difficult to edit the huge amount of data on the web manually. So the concept of web data extraction system was introduced. Web data extraction system is a software system that automatically extracting data from a website. After extracting the data from the web page that extracted data are delivered to a database or some other application. Web data extraction has a wide range of applications such as bioinformatics, analysis of text based documents available to the company, business and competitive intelligence etc. By analyzing the web, we can compare the products, market trends, price details etc.

There are different ways to perform web data extractions. In the earlier stages, manual extraction techniques are used. In that technique, manually writing the programs called wrappers or extractors to extract the data from the web page. Manual extraction technique use some built in rules to extract the data. The working of this technique is based on some prior knowledge of the format of the web page. TSIMMIS, Minerva, Web-OQL, W4F and XWRAP are the examples of manual data extraction [1]. The problem with this technique is that it is a labor intensive task and maintaining wrappers can be expensive and impractical. Therefore, automatic web data extraction techniques are introduced. First supervised techniques are introduced later unsupervised technique are introduced. WIEN, STALKER and SoftMealy are the examples of supervised techniques. In supervised techniques, wrapper construction system output the extraction rules based on the training examples provided by the designers of the wrapper. The problem with this technique is that designers must manually label the training examples for generating the rules also labelling the training example is time consuming and not efficient enough, so unsupervised techniques are introduced. The advantage of this technique is that no users training examples are needed for web data extraction. IEPAD and OLERA are the some examples of the semi supervised technique. RoadRunner, EXALG, NET, FivaTech are the examples of unsupervised technique.

This paper focus on various information extraction techniques like SoftMealy, OLERA, IEPAD, RoadRunner, EXALG, NET, FivaTech .Section II gives more details about above techniques, section III provides a comparison of these techniques and section IV concludes the paper.

II. WEB DATA EXTRACTION TECHNIQUES

2.1 SOFTMEALY

SoftMealy is the approach to wrapping semi structured web pages. SoftMealy is based on the based on FST (Finite State Transducer) and contextual rules. One of the advantages of SoftMealy is that it handle missing attributes and attributes permutations in the input. The FST consists of two parts such as body transducer and a tuple transducer. The body transducer extract the part of the page that contains the tuples, the tuple transducer iteratively extracts the tuples from the body and it also accepts a tuple and returns its attributes.The main drawbacks of Softmealy are it is not able to generalize overseen separators and it need many error recovery steps for unseen separators[2] .

2.2. OLERA

OLERA is a semi supervised information extraction system and it produces extraction rules from semi structured Web documents without requiring detailed annotation of the training documents**.** OLERA's core technique is a well-known technique called string alignment also OLERA is designed with visualization support [3]. The two limitations of OLERA are it is sensitive to the ordering of input information and extraction failure could occur when the templates for each attribute are similar.

2.3. IEPAD

IEPAD is an information extraction system that automatically identifies the extraction rule by repeated pattern discovery techniques. The repeated patterns are identified using the data structure called PAT trees. PAT tree is a PATRICIA tree (Practical Algorithm to Retrieve Information Coded in Alphanumeric).In previous work, extraction rules are learned from training examples [4]. But in IEPAD, an unsupervised technique is introduced for pattern discovery.

The IEPAD system has three components such as extraction rule generator, pattern viewer and extraction module. Extraction rule generator accepts input web page . The pattern viewer is the graphical user interface which shows repetitive pattern discovered and an extractor module is used to extracts desired information from similar Web pages according to the extraction rule. The extraction rule generator consists of a translator, a PAT tree constructor, a pattern discoverer, a pattern validator, and an extraction rule composer. The function of the translator is to receive the HTML page and translate it into a string of abstract representations known as tokens. Each of this token is represented as binary code of length *l*. The PAT tree constructor receives these binary file and construct the PAT tree. The pattern discoverer then uses these PAT trees to discover repetitive patterns. Then these repetitive patterns are forwarded to validator. The validator filters out undesired patterns and produces candidate patterns. Finally, the rule composer generates the extraction rule in regular expression based on these candidate patterns. The identified in the IEPAD is that could not handle complex and nested structured data.

2.4. ROAD RUNNER

Road Runner is the technique for automatically extracting the data from the HTML sites. In Road Runner, data is extracted through the use of automatically generated wrappers. Wrappers are generated based on the similarity and difference between the web pages. The previous approaches to wrapping websites are based on the manual techniques. But manually writing these wrappers are difficult and labor intensive, so the concept of Road Runner is introduced. Road Runner starts with first input page as its initial template. Then match each successive sample pages and checks if the match occurs. If it cannot be, it modifies the current template of the page. Advantages of the Road Runner are it does not require any interaction with the user during the wrapper generation process, it has no prior knowledge about the schema of the web page and also it is not restricted to the flat records, but it can handle nested structures also [5]. The limitations of the Road Runner are number of errors in the input documents affect it's the effectiveness and they don't handle disjunction cases.

2.5. EXALG

EXALG is an algorithm for the extracting the structured data from a collection of web pages generated from the common template. EXALG consist of two stages such as equivalent class generation stage (ECGM) and analysis stage. In ECGM stage, find the sets of tokens having the same frequency of occurrence in every page which are known as equivalence classes [6]. EXALG retains only the equivalence classes that are large and whose tokens occur in a large number of input pages, such type of equivalence classes are known as LFEQs (for Large and Frequently occurring EQuivalence classes).The analysis stage constructs the template using the LFEQs. The problem identified in the EXALG is that it is not clear whether EXALG can work on malformed input document or not.

2.6. NET

NET is the Nested data Extraction using Tree matching and visual cues. NET is the effective method to extract data from Web pages that contains a set of flat or nested data records automatically. This method is based on a tree edit distance method and visual cues. It works in two stages. First stage is building a tag tree of the page and second stage is identifying data records and extracting data from them. The algorithm performs a post-order traversal of the tag tree to identify data records at different levels. A tree edits distance algorithm and visual cues are used to perform these tasks. Advantage of this NET technique is that it enables accurate alignment and extraction of both flat and nested data records. One of the limitations identified in the NET technique is that it incorrectly identifies a flat structure as nested one [7].

2.7. FIVATECH

FivaTech is a page-level web data extraction technique, which automatically detect the schema of a Website. FivaTech introduce a new structure, called fixed/variant pattern tree. The fixed or variant pattern tree is used for to identify the template and detect the data schema. This technique is the combination several techniques such as alignment, pattern mining. FiVaTech contains stages. First stage is merging input DOM trees to construct the fixed/variant pattern tree. In second stage schema and template are detected based on the pattern tree [8]. Limitations of the FivaTech are searching the longest repeating patterns is time consuming process and also it does not work on the malformed input document without correcting them.

III. COMPARISON

Among the webpage extraction techniques discussed above, some techniques are supervised and some other are semi supervised and unsupervised, some techniques extracts flat records and some other techniques are trying to extracts nested record.NET and RoadRunner will find out nested records in addition to flat records. EXALG, FivaTech and IEPAD produce flat records. RoadRunner and OLERA using string alignment for extracting the records. FivaTech uses tree merging technique whereas NET using tree matching. EXALG is based on equivalence class generation. SoftMealy uses ad-hoc (bottom up) learning algorithm. EXALG and FivaTech consider multiple pages of website and other techniques considers only single page. Summary of comparison is shown in Table I.

Technique	Type	Learning Algorithm	Single page/Multiple pages
SOFTMEALY	Supervised	Ad-hoc (bottom-up)	Single
OLERA	Semi supervised	String Alignment	Single
IEPAD	Semi supervised	Patter Mining, String Alignment	Single
ROADRUNNER	Unsupervised	String Alignment	Multiple
EXALG	Unsupervised	Equivalence class generation	Multiple
NET	Unsupervised	Tree matching	Single
FIVATECH	Unsupervised	Tree merging and schema detection	Multiple

Table I: Comparison of various web data extraction techniques.

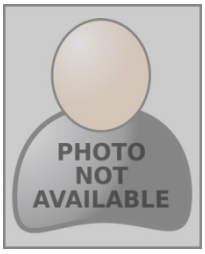
IV. CONCLUSION

In this paper studied various approaches to extract structured data from web pages. Among the above discussed web data extraction techniques, some techniques extract flat records and some other techniques are trying to extracts nested records also. Some of these techniques are either inaccurate or make many strong assumptions.

References

1. C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1411–1428, Oct. 2006.
2. C.-N. Hsu and M.-T. Dung, "Generating finite-state transducers for semi-structured data extraction from the web," Inform. Syst.,vol. 23, no. 8, pp. 521–538, Dec. 1998.
3. C.-H. Chang and S.-C. Kuo, "OLERA: Semi supervised web-data extraction with visual support," IEEE Intell. Syst., vol. 19, no. 6,pp. 56–64, Nov./Dec. 2004.
4. V. Crescenzi, G. Mecca, and P. Merialdo, "Road runner: Towards automatic data extraction from large web sites," in Proc. 27th Int. Conf. VLDB, Rome, Italy, 2001, pp. 109–118.
5. C.-H.Chang and S.-C. Lui, "IEPAD: Information extraction based on pattern discovery," in Proc. 10th Int. Conf. WWW, Hong Kong, China, 2001, pp. 681–688.
6. A. Arasu and H. Garcia-Molina, "Extracting structured data from web pages," in Proc. 2003 ACM SIGMOD, San Diego, CA, USA, pp. 337–348.
7. B. Liu and Y. Zhai, "NET: A system for extracting web data from flat and nested data records," in Proc. 6th Int. Conf. WISE, New York, NY, USA, 2005, pp. 487–495.
8. M. Kayed and C.-H. Chang, "FiVaTech: Page-level web data extraction from template pages," IEEE Trans. Knowl. Data Eng., vol. 22, no. 2, pp. 249–263, Feb. 2010.
9. J. Wang and F. Lochovsky. "WrapperInduction based on nested pattern discovery." , Technical Report HKUSTCS-27-02, Dept. of Computer Science, Hong Kong U. of Science andTechnology, 2002
10. Tai, K. The tree-to-tree correction problem. J. ACM,26(3):422–433, 1979

AUTHOR(S) PROFILE



Vidya.V.L received the B.Tech degree in Computer Science and Engineering from UKF College of Engineering and Ttechnology,Kollam, keralauniversity. Currently pursuing M.Tech in Computer Science and Engineering from Mohandas College of Engineering and technology, Thiruvananthapuram, kerala university.

