A SURVEY OF TOOLS FOR EXTRACTING AND ALIGNING THE DATA IN WEB

SureshKumar.T
Assistant Professor
K.S.R College of Technology
Namakkal, Tamil Nadu, India
Suresh_technology@yahoo.co.in

Sivaranjani.S
PG Scholar
K.S.R College of Technology
Namakkal, Tamil Nadu, India
sivaranjanitech@gmail.com

Dr.Shanthi.N
Professor & Dean
Nandha Engineering College
Erode,TamilNadu,India

ABSTRACT

The world-wide web is rapidly growing day by day in all fields, mining the data from multiple websites is necessary to filter the relevant contents. Although many approaches developed for extracting the data, there were some difficulties found when using such tools. In this paper, we survey web data extraction and alignment process in two dimensions: record extraction and alignment. The first dimension explains the extracting data records from multiple query result pages automatically. The second one measures similarity between the data records for aligning the records by pairwise and holistically and then nested structure processing. We believe these criteria enhance the performance measures to check existing data extraction methods.

Keywords-Data extraction, automatic wrapper generation, data record alignment

I INTRODUCTION

Online database, called web database, generates query result page relevant to user's query in search engine. Web search engine needs to cooperate with multiple web databases to answer a user query. All web applications including web users need to co-operate with search engines is necessary. Such data from different web databases are either structured or semi-structured. It is important to drop the irrelevant data from result pages automatically for extracting data for user's quick view. Data Extraction is the way to mine data from huge databases. Information extraction in web, mainly involves extracting information from structured data in web pages which includes non contiguous information. Obviously, the search engine consists of static as well as dynamic pages, static pages generally does not make any problems when display but dynamic pages have lots of issues.

In this paper, we analyze the extraction and the alignment method in three tools namely DeLa, Viper and CTVS.

II OVERVIEW OF EXTRACTION TOOLS

In olden days the wrapper needs human knowledge for extracting the relevant item and it was unable to act automatically. Recently, methods developed for automatic extraction of records from huge web databases. In wrapper induction, extraction based on inductive learning since it has merits on extracting the data it needs human knowledge and lots of time consuming. This was under two major critical problems: continuous navigation and maintaining a wrapper when periodic changes occur. The Table [1] refers summarization of three tools efficiency to extract the records from the data regions.

To solve existing tools problem, some learning methods proposed by researches such as DeLa [3], RoadRunner [5] for automatically extracting the web pages.

DeLa [3] splits the data region into more subparts and they consider only the data region with largest subparts and discard the others and it entirely depends on tag structure. It calculates the similarity before the data records

are aligned in a manner, this leads to irregularity in making optional attributes. Moreover, this has good precision value when wrapping the records.

Table 1
Data Extraction Method Summarization

Tools	Nested Structure Processing	Single Result Page	Non-Contiguous Data Regions
CTVS	Yes	Yes	Yes
DeLa	Yes	Yes	No
Viper	No	Yes	No

Viper [2] (**Vi**sual **P**erception-based **E**xtraction of **R**ecords) is one of the methods which is able to extract and separate relevance of different repetitive information contents or patterns with respect to the user's visual perception along with tags that are embedded with corresponding web page. It uses both human visual data perception value and the HTML tag structure to find rank and weight the patterns. Although Viper offers good results for single page, it does not handle when pages with nested structured data. CTVS handles nested structured data more efficiently than the Viper.

ViNTs [4] is an another tool like Viper for extracting the data, it uses visual as well as non-visual features to rank and weight the relevance of different extraction rules but has some drawbacks. First, it depends only on major data regions where the data records are highly reported than the other pages. Second, web users needed to collect the training pages or labels from the websites. Third, it needs continuous navigating for periodically updating dynamic changes. In contrast, CTVS [1] (Combining Tag and Value Similarity for data extraction and alignment) requires neither training page nor a pre-learned wrapper. Table 1 [1] summarizes some characteristics of the extracting tools compared in CTVS paper.

III SURVEY OF DIMENSION

QRR Extraction

DeLa (Data Extraction and Label Assignment for Web databases) automatically wrap the contents in a page using labels and put them into a table. DeLa extract the pages from the web with the help of HTML tags. Suppose a web page contains more than an instance, it assumes as a sequence of token and produces a rules or regular expression for each page. This leads to production of too many expressions for a single QRR. DeLa assumes a nested structure as a flat structure this leads to fail in accurate similarity calculation. And it is difficult to manage those rules in each instance. Label Assignment for each data value is upon some heuristics. It gives accurate data extraction and implementation of this tool is available now.

In **Viper**, it considers both the user's perception and HTML tags for extraction. It extracts only the contiguous page in a website and it fails to perform nested structure effectively. It follows three step processes for data extraction that is data preprocessing, data segmentation in visual basis and finally weighting the data regions after extract the region from the web page. It performs good data extraction but implementation is not available.

In **ViNT** (Visual Information and Tag structure based Wrapper Generated) considers both the HTML tags and visual data for tag tree construction as like Viper.Fig.1[12] shows the ViNT architecture. The tool uses both the visual as well as non visual features for extraction. If the data records contain more than one region ViNT select the major one and discards the others. It needs user's knowledge for collecting the training pages and labels and also needs continuous monitoring.

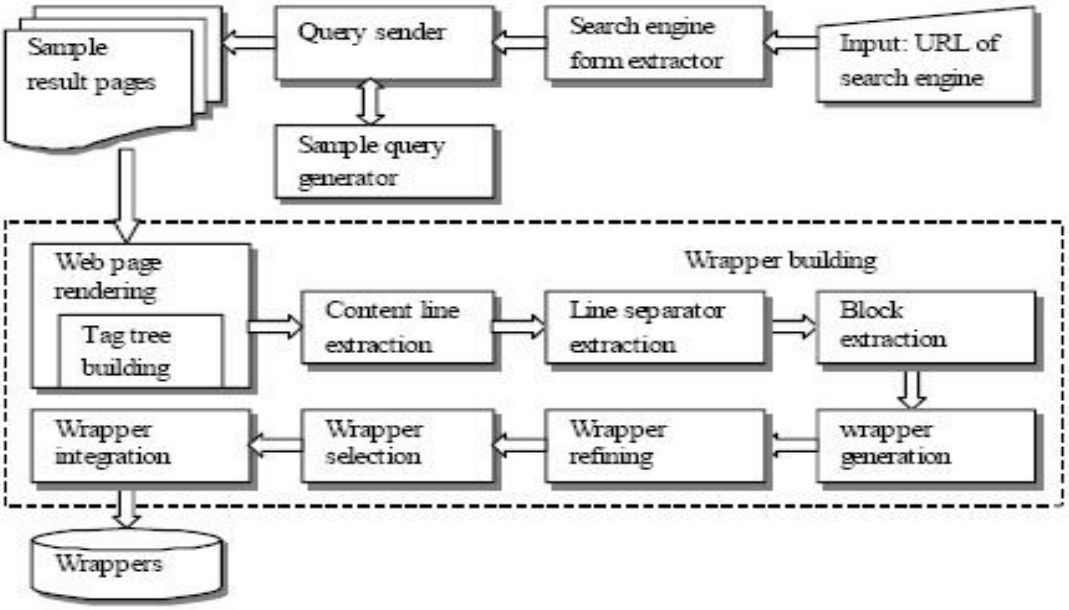

Fig. 1 ViNT Architecture

In **CTVS [1**] (Combining Tag and Value Similarity for Data Extraction and Alignment**)**, extraction is done in five steps.

1. Tag tree construction
2. Data region identification
3. Record segmentation
4. Data region merge
5. Query result section identification

It is a modal that combine tag and their corresponding value for data similarity calculation and according to that result it wraps the data from the query result pages. It constructs the tree first and identifies the data records from the QRRs then segments them into a table. It can handle the nested structured data records more effectively than the existing methods. CTVS process the non-contiguous pages from a website while the existing method does not do.

QRR Alignment

Alignment is effective only when the nested structure processing is done before the data records are aligned. In **DeLa**, aligning the nested structure is happened after all the data records are aligned this tends to make vulnerability in selecting the optional attributes from the records. It falsely assumes flat structure for nested structured records and put all the records into a single parent node. This results in making hard to do aligning the records.

In **Viper**, Multiple Sequence Alignment method is followed for aligning the data records. It aligns globally that is, it considers all the records for alignment once extraction is completed. It uses Suffix tree for global alignment. Viper fails to align the non contiguous page and gives poor performance for nested structure processing.

In **CTVS,** alignment is done by three step process. First, Pairwise Alignment is for comparing a pair of QRRs to find the similarity and put them into a table for global alignment. Second, holistic alignment is for comparing the records globally from the result of pairwise alignment. Finally, nested structured processing is done effectively by specially designed data region identification and merging algorithm.

IV OVERALL COMPARISON

Two set of evaluation metrics are used to compare the performance of the tools. The first is at the record level which includes,

$$\textbf{Precision} = \textbf{Cc / Ce}$$

$$\textbf{Recall} = \textbf{Cc / Cr}$$

Where Cc refers to count of correctly extracted records, Ce refers to count of extracted records, and Cr refers to actual count of records.

The second is Page level Metric, namely, page level precision,

Page level precision = Cp / Na

where Cp refers to count of correctly extracted pages, and Na refers to count of all the pages

Table 2
Data extraction Performance for the AUXI dataset

AUXI data set	Contiguous pages			Non contiguous pages		
QRRs	510			543		
Method	CTVS	Vint	DeLa	CTVS	Vint	DeLa
Extracted QRRs	506	502	503	530	432	437
Correctly extracted QRRs	499	486	479	510	418	415
Record level precision	98.6%	96.8%	95.2%	96.2%	96.8%	95.4%
Record level recall	97.8%	95.2%	93.9%	93.9%	77%	76.5%
Page level precision	92.5%	90%	85%	85%	35%	37.5%

Table 2 [1] shows performance of CTVS, ViNT, and DeLa over the AUXI data set. From the table, we see that performance of CTVS is higher than all the existing methods compared here. It shows higher precision for both contiguous and non-contiguous pages than the existing data extracting methods.

V CONCLUSION

In this paper, we survey extraction tools and compare their performance metrics for both contiguous and non-contiguous pages. CTVS achieves higher precision than the existing methods. In general, all extraction methods follow ontology concepts for wrapping the user demanded information from the web. Although DeLa and ViNT perform well on data extraction, they fail to cover the nested structure processing in most cases whereas CTVS is able to cover the nested structured pages in web. Viper has good performance results but it does not handles non-contiguous pages and as well as implementation is not available. All the tools mentioned here have same drawback that is, when a data region has more than one record they give priority to major one and discards the other records.

REFERENCES

[1] Weifeng Su, Jiyang wang, Frederick H.Lochovsky, "Combining Tag and Value Similarity for Data Extraction and Alignment", IEEE Transactions on Knowledge and Data Engineering, vol.24, No.7, July 2012
[2] K. Simon and G. Lausen, "ViPER: Augmenting Automatic Information Extraction with Visual Perceptions", Proc. 14th ACM Int'l Conf .Information and Knowledge Management, pp. 381-388, 2005.
[3] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases", Proc. 12th World Wide Web Conf.,pp. 187-196, 2003
[4] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines", Proc. 14th World Wide Web Conf., pp. 66-75, 2005.
[5] V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites", Proc. 27th Int'lConf. Very Large Data Bases, pp. 109-118, 2001.
[6] R. Baeza-Yates, "Algorithms for String Matching: A Survey" , ACM SIGIR Forum, vol. 23, nos. 3/4, pp. 34-58, 1989.
[7] Duhan, N. "Page Ranking Algorithms: A Survey" Advance Computing Conference, 2009. IACC 2009. IEEE International Conf.
[8] N. Kushmerick, D.S. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction", Proc. 15th Int'l Joint Conf. Artificial Intelligence, pp. 729-737, 1997.
[9] B. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in WebPages", Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003.
[10] Y. Zhai and B. Liu, "Structured Data Extraction from the WebBased on Partial Tree Alignment", IEEE Trans. Knowledge and Data Eng.,vol. 18, no. 12, pp. 1614-1628, Dec. 2006.
[11] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "FullyAutomatic Wrapper Generat ion for Search Engines", Proc. 14thWorld Wide Web Conf.,pp. 66-75, 2005.
[12] "Web data extraction and alignment tools:A survey" , Shridevi A. Swami, Pujashree Vidap-International Journal of Scientific Engineering and Technology ,Volume No.2, Issue No.6, pp : 573-578 1 June 2013