International Journal of Computer Applications Technology and Research
Volume 3– Issue 7, 484 - 487, 2014, ISSN: 2319–8656

Efficient Web Data Extraction

Yogita R.Chavan
University of Pune
KKWIEER
Nashik, India

Abstract: Web data extraction is an important problem for information integration as multiple web pages may present the same or similar information using completely different formats or syntaxes that make integration of information a challenging task. Hence the need of a system that automatically extracts the information from web pages is vital. Several efforts have already been carried out and used in the past. Some of the techniques are record level while the others are page level. This paper shows the work aims at extracting useful information from web pages using the concepts of tags and values. To avoid discarding of non-matching first node that represents non auxiliary information in the data region an efficient algorithm is proposed.

Keywords: auxiliary information, data extraction, DOM Tree, record alignment.

1. INTRODUCTION

Web information extraction is one of the very popular research activities aims at extracting useful information from web pages. Such extracted information is then stored into the database that can be used for faster access to the data. Due to the assorted structure of web data, automatic discovery of target information becomes a tedious task.

In order to extract and make use of information from multiple sites to provide value added services, one needs to semantically integrate information from multiple sources. Hence the need of a system that will automatically extract the information from web pages efficiently is vital.

The work aims at studying different web page extraction strategies/techniques and to implement the technique based on tag and value similarity as well as a few enhancements, if possible.

A method of record extraction is referred from CTVS proposed by W. Su et al [6]. This method is further modified using label assignment technique mentioned in DeLa [3] partly to overcome the drawback of not considering an optional attribute found in data region which cause the loss of information. This information is stored in temporary file during data region identification step and regions are then merged using similarity technique [11]. Applying the heuristics, only one data region is selected to extract exact result records. If information stored in temporary file belongs to this selected data region, it is segmented before final record extraction because of which the optional attribute that was not considered during data region identification is considered and information loss is prevented.

2. LITERATURE SURVEY

Due to the necessity and quality of deep web data web database extraction has received much attention from the Data mining and Web mining research areas in recent years. Earlier work focused on wrapper induction methods called as non-automatic methods require human assistance to build a wrapper. Wrappers are the hand coded rules i.e. a customized procedure of information extraction. In this method an inductive approach is used where user learns or marks part or all of the items to extract the target item containing set of training pages. A system then learns the wrapper rules and uses them to extract the records from the labeled data.

Advantages

-No extra data extracted

Disadvantages

- Labor intensive and time consuming

- Performs poorly when the format of query result page changes

- Thus, not scalable to a large number of web databases.

Systems WIEN, Stalker, XWRAP and SoftMealy follow wrapper induction technique.

More recently, automatic data extraction systems like RoadRunner, IEPAD, DeLa and PickUp have been proposed. C.H. Chang et al used a method of pattern discovery for information extraction that generates extraction rules which utilize a decoded binary string of the HTML tag sequence and tries to find maximal repeated patterns using a PAT tree which then become generalized using multiple string alignment technique. At the end the user has to choose one of these generalized patterns as an extraction rule.

This method identifies and extracts the data using repeating patterns of closely occurring HTML tags. It is convenient for set of flat tuples from each page and also produces poor results for complex and nested structure data structure[5].

V. Crescenzi, et al proposed a method for automatic data extraction that extracts a template by analyzing two web pages of an equivalent category at a time. In this method one page is used to derive initial template and it then tries to match the second page with the template.

Challenges of this method are deriving the initial template needs to be done manually [10].

Kai Simon et al used visual perceptions for automatic web data extraction that project the contents of the HTML page

onto a 2-dimensional X/Y co-ordinate plane due to which it is able to compute two content graph profiles, one for each X and Y planes. These used to detect data regions by locating valleys between the peaks as the separation point between two data.

Drawback of this method lies in the assumption that the data regions are separated by the defined empty space regions. This may not always true [7].

Hongkun Zhao et al proposed a technique for fully automatic wrapper generation for search engines that extracts content line features from the HTML page, where this content line is a type of text which could be visually bounded by a rectangular box.

Several sample pages are used to extract the correct data region from the HTML page using parsing technique. But result records with irregular block structures are excluded in this method and also parsing the sample static and dynamic HTML regions become overhead.

3. IMPLEMENTATION
Algorithm

1. Query Result Page DOM Tree Construction

2. Data Regions Identification

3. Query Result Records Extraction

4. Records Alignment Pair Wise

5. Nested Structure Processing

6. Final Database Table

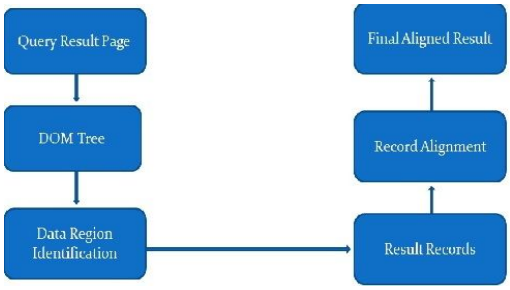

Figure 1: Block Diagram of the System

Implementation further divided into two main parts.

1. Records Extraction

2. Records Alignment

3.1 Record Extraction

3.1.1 Query Result Page DOM Tree Construction

From the source code associated with the page (that is HTML code), a DOM (Document Object Model) tree is to be constructed. Let us start with an example, the query result page for query- Apple Notebook Figure 2 shows a page with two images Apple iBook Notebook and Apple Powerbook Notebook after firing query Apple Notebook along with some non-useful information of links of advertisements. From the HTML source code, DOM tree shown in figure 3 will be generated.

Figure 2: Query Result page for query - Apple Notebook

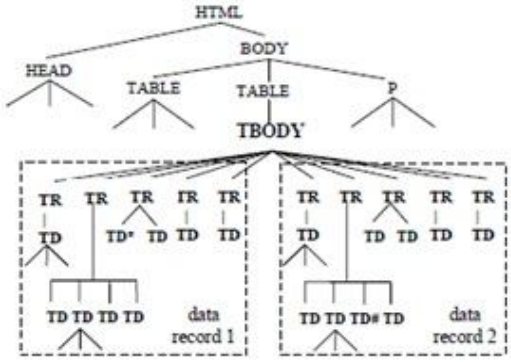

Figure 3: DOM Tree for query page- Apple Notebook

3.1.2. Data Region Identification

By calculating similarity of nodes, data regions are identified. Using node similarity algorithm where each node represents information and all of them form a graph. Finding their adjacency matrix, incidence matrix similarity between them is calculated. To calculate similarity the edit distance between the nodes is considered and two nodes n1 and n2 are similar if the edit distance between them is greater than or equal to threshold value 0.6 suggested by Simon and Lausen. Similar nodes are then recognized and nodes with the same parent form a data region. Multiple data regions may be formed during this step.

3.1.3. Query Result Records Extraction

Applying heuristics that the search result section usually located at the centre of the query result page and it usually occupies large space in query result page, a data region is identified to extract the information.

3.2 Records Alignment

For record alignment a novel method consisting of three consecutive steps for alignment proposed by W. Su [6] is referred. This method includes two steps

C1 / N1	C2 / N2	Similarity between nodes N1 and N2
4 July 2013 3.15/datetime	15.Aug.2013/date	0.5
123/ int	5 / int	1
Fast and furious part 2/ string	Fast and furious/ string	0.825
Fast and furious / string	234/ int	0

3.2.1. Pairwise Alignment

The similarity between two data values f1 and f2 with data type nodes n1 and n2 is defined as where p(n i) is the parent node of ni in the data type tree, sample of which is shown in figure 4 below.

In the first row of above table, in column C1, information is of type date time where as in column C2 only date is mentioned. Referred formula and data type tree, datetime is the parent node of date node.
Condition N1 = p(N2) and N1 is not string satisfies and similarity 0.5 in considered in column 3. In the second row of the example, both values are of integer type, so similarity one is entered. In the third row, both values are of string type so cosine similarity is taken into consideration. In the last as both values are of different types zero similarity is considered.

3.2.2. Nested Structuring Alignment

After the first step of pairwise alignment all data values of the same attribute are put into the same table column. The logic of finding connected components of an undirected graph is used for this purpose. In nested structuring multiple data values of the same attribute are put in the different row of the table.
In the end, the information will be stored in the form of a table.

4. RESULTS and DISCUSSION

4.1 Data set

1. E- COMM contains 100 deep websites E-commerce in six well-liked domains such as hotel, job, movie, automobile, book and music whereas each domain has 10 to 20 websites.
2. PROFUSION contains 100 websites collected from profusion.com

Above datasets can be used to evaluate the working of the system.

4.2 Result Set

The implementation is done in java using Netbeans where user is allowed to enter a query result page as an input. Above mentioned datasets can be used for the same purpose. And results are then compared with existing systems results. Proposed methodology is used to provide the better results preventing loss of information. When applied on first 10 web pages, the table of results is, where precision metrics= Cc/Ce and recall metrics= Cc/Cr

Name of the web page	Count of correctly extracted and aligned QRRs (Cc)	Count of extracted QRRs (Ce)	Actual count of QRRs in Query result page (Cr)	Precision Metrics (%)	Recall Metrics (%)
Bed.html	2	2	2	100	100
Car.html	4	4	4	100	100
Equipment.html	2	2	2	100	100
Pulsar.html	4	4	5	100	80
Bedsheet.html	2	3	3	66.66	66.66
Coffee.html	4	4	4	100	100
Apparatus.html	2	2	2	100	100
Equipment.html	5	5	6	83.33	83.33

5. CONCLUSION AND FUTURE WORK

An efficient method for web data extraction is proposed that includes finding data regions and also considering optional attribute (non-auxiliary information) node value and further add it in the final database table. This overcomes the drawback of loss of information in a data region. This increases the performance of the system by extracting the information more effectively.
In this research, it is aimed to obtain extraction of web page's information accuracy by using the efficient algorithm. For this purpose, several fully automatic web extraction approaches are investigated. Extensive studies are done on these approaches to explain why they do not achieve satisfactory data extraction outcome. After performing several experiments as described in the result table, it is observed that efficiency of the system has increased. Second, any optional attribute that appears as the start node in a data region will not be treated as auxiliary information.
This research has found that the system outperformed the existing web data extraction systems.
Along with the advantages, this method has shortcomings like it requires at least two query result records in the result page as for forming a template at least two result records expected and the other is while selecting a single data region depending on heuristics discussed other data regions are discarded which may contain useful information needs to be stored in database table. These drawbacks would be tried to be removed in the future.

6. REFERENCES

[1] A. Arasu and H. Garcia-Molina, *Extracting Structured Data from Web Pages*, Proc. ACM SIGMOD International Conference Management of Data, Pp. 337-348, 2003

[2] R. Baeza-Yates *Algorithms For String Matching: A Survey*, ACM SIGIR Forum, Vol. 23, Nos. 3/4, 34-58, 1989

[3] J. Wang And F.H. Lochovsky *Data Extraction and Label Assignment For Web Databases*, Proc. 12th World Wide Web Conference Pp. 187-196,2003.

[4] Y. Zhai And B. Liu *Structured Data Extraction From The Web Based On Partial Tree Alignment*, IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 12 Pp.1614-1628, Dec. 2006

[5] C.H. Chang and S.C. Lui *IEPAD: Information Extraction Based On Pattern Discovery*, Proc. 10th World Wide Web Conference Pp. 681-688, 2001.

[6] Weifeng Su, Jiying Wang *Combining Tag and Value Similarity For Data Extraction and Alignment*, IEEE Transaction On Knowledge And Data Engineering, Vol.24, No.7, July 2012

[7]] K. Simon And G. Lausen *VIPER: Augmenting Automatic Information Extraction With Visual Perceptions*, Proc. 14th ACM International Conference Information and Knowledge Management Pp. 381-388,2005.

[8] Y. Zhai And B. Liu *Structured Data Extraction From The Web Based On Partial Tree Alignment*, IEEE Trans. Knowledge and Data Eng., Vol. 18, No. 12 Pp.1614-1628, Dec. 2006.

[9] D. Buttler, L. Liu and C. Pu *A Fully Automated Object Extraction System for the World Wide Web*, Proc. 21st International Conference Distributed Computing Systems Pp. 361-370, 2001

[10] V. Crescenzi, G. Mecca and P. Merialdo *Roadrunner: Towards Automatic Data Extraction from Large Web Sites*,

Proc. 27th International Conference Very Large Data Bases Pp. 109-118, 2001

[11] Miklos Erdelyi, Janos Abonyi *Node Similarity-Based Graph Clustering and Visualization*, 7th International Symposium of Hungarian Researchers on Computational Intelligence