International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Data Extraction and Annotation Methods Using Tag Value Structure

Tushar Jadhav[1], Santosh Chobe[2]

[1, 2] Department of Computer Engineering, DYPIET Pimpri, Savitribai Phule Pune University, India

Abstract: ***The World Wide Web generates search result pages which is based .On the user's input query. It is very crucial for many applications like data integration which requires combining more databases to automatically extract the data from the search results. A unique method for extracting the data and then aligning is implemented which uses Unsupervised duplicate detection algorithm which identifies and segments the result records first and then aligns the segmented results in a table , in which data values of similar attributes are put in same column. The new technique is implemented so as to handle the case when the search results are not adjoining which might happen because of auxiliary data such as advertisements ,comments etc. and also to handle nested tag structure which might be present in the search results. The result shows that the implemented algorithm performs well than existing methods.***

Keywords: Web Databases, wrapper generation, data annotation, data alignment

1. Introduction

The amount of information that is currently available on the net in HTML format grows at a very fast pace, so that we may consider the Web as the largest "knowledge base" ever developed and made available to the public. However HTML sites are in some sense modern legacy systems, since such a large body of data cannot be easily accessed and manipulated. The reason is that Web data sources are intended to be browsed by humans, and not computed over by applications. XML, which was introduced to overcome some of the limitations of HTML, has been so far of little help in this respect. As a consequence, extracting data from Web pages and making it available to computer applications remains a difficult task.

Extraction of data from HTML page is generally done by software known as wrappers. Previous approaches [2] for wrapping Web sites were based on manual methods. A major issue with manual wrappers is that writing them is usually a difficult and labor intensive task, and that by their nature wrappers tend to be brittle and difficult to maintain.

This paper develops a novel approach to the data extraction problem: our goal is that of fully automating the wrapper generation process, in such a way that it does not rely on any a priori knowledge about the target pages and their contents. The results returned for a query from a web database has many search result records. Every search result contains many data units every one of which depicts one prospect of a real world entity. Fig.1(a) displays three search result records on a result page from a book database. Every record comprise of single book with multiple data units, for e.g., the second book record in Fig. 1(a) has data units "Data Mining : Practical Machine Learning Tools and Techniques," "Ian H. Witten and Elbe Frank," etc.

(a) Original HTML Page

```
<form><font color="blue">Data Mining: Practical Machine
Learning Tools and Techniques </font> <BR>by Ian H.
Witten and Elbe Frank</BR><font color
="blue"><B>Paperback</B></font><BR><font color
="maroon"><B>457.00</B>
<br></font>You Save :<font color ="maroon">
68.00 (13%)</font> </BR>
</FORM>
</html>
```

(b)Simplified HTML Source Code for Second SRR

Figure 1(a ,b): Example search results from amazon.com

This paper concentrates on the issue of extracting data records automatically which are embedded in the SRR's generated by web databases. A search result page carry the actual data as well as other information, like advertisements, navigational panels, comments, etc. The aim of web database data extraction is to take out any irrelevant , unwanted, information from the search result page, extract the search records from the result page, and then align the extracted SRR's into a table so that the data units belonging to the similar attribute are put into the one table column.

2. Related Work

There is continuous research on Web information extraction and annotation in recent years. Most of the systems [13], [14] depends upon users to mark the needed information on sample pages and for labelling the marked data at the same instance, and then the system can produce a series of rules so as to derive the similar collection of data on webpages from the source. These kind of systems are known as a wrapper induction system. Due to learning process and supervised training, these systems can generally achieve high accuracy in extracting information .But, they fall apart by poor scalability and that is why they are not useful for applications [15], [16] which needs to extract information from a huge number of web sources.

For the extraction of structured data from a collection of web pages generated from a common template Arvind Arasu et al.[1] proposed EXALG , it first discovers the unknown sample which generates the pages and so as to extract the data from input pages it uses the discovered sample. EXALG uses of two methods, differentiating roles and equivalence classes.

G. Mecca et.al [2] investigated different methods for extracting data from HTML websites by using automatically generated wrappers , so as to automatize the data extraction process, and finally developed a novel approach for data extraction problem ,the idea is to use fully automating the wrapper generation process, so that it does not depend on any a knowledge about the target pages ,contents.

Luigi Arlotta et al. [4] introduced wrapper induction system. This system rely on human users and marks the label data. Actually wrappers are built automatically, the values that they extract are anonymous and a human intervention is still required to associate a meaningful name to each data item.

The data extracted by automatically generated wrappers is a unusual problem, and it represents a step towards the automatic extraction and manipulation of web data. The web pages are designed to be used by humans, and that's why mostly they contain text strings, i.e. labels, The goal is to interpret the end user the meaning of the published data. However, this system achieves higher extraction precision in the result and an increased maintenance cost. Also, this system suffer lesser scalability that does not work in the applications like extraction algorithms. To reduce the cost associated with wrapper production and maintenance cost, the researchers have concentrated on automatic generation of wrappers.

S. Mukherjee et al.[5] discussed a method which is mainly concerned with automatically annotating HTML documents. To detect semantic changes in document content, it uses structural and semantic analysis techniques. The idea is to use template-based content-rich HTML documents. This technique shows the key observation that semantically related items that display consistency in presentation style and spatial locality.

F.H.Lochovsky et al.[6] introduced a system based on ontology, a new data extraction method in which query results are extracted from HTML pages automatically by using Ontology-assisted Data Extraction method. It creates ontology for a domain according to information which is identical between the search query interfaces and search result pages from different web sites within the same domain. Then, the constructed domain ontology is used to match the query result section in a query result page and to align and label the data values in the extracted records for data extraction.

For efficient retrieval of data from the web, Y. Jiang et al.[7] made use of ideas taken from databases, however it requires structured data. Yet most web data is unstructured and cannot be queried using traditional query languages. To solve this problem, different ways for querying the web have been proposed. Basically there are two categories: querying the web with web query languages and generating wrappers for web pages.

An ontological approach is proposed for extracting and structuring data from documents posted on the web. The data extraction method is based on conceptual modelling, this approach focuses specifically on unstructured documents which are rich in data, narrow in ontological breadth, and contain multiple records of information for the ontology. So to automatically extract data in multi-record documents and label them, it employs ontologies together with several heuristics. However, it is necessary to construct ontologies manually for different domains.

Wei Liu et al.[8] proposed Vision-based Data Extractor (ViDE), which automatically extracts structured results from deep web pages. ViDE is basically based on the visual features human users can catch on the deep web pages and to make the solution more robust , it employs simple non-visual data such as data types and frequent symbols .It consists of two main components,(ViRIE) Vision based Data Record extractor and (ViDIE) Vision-based Data Item extractor. Using the visual features for data extraction, ViDE neglects the flaw of those solutions that need to analyze complicated web page source files.

H. Zhao et al.[9] proposed a technique for automatically producing wrappers, used to extract search result record from dynamically generated result page. Automatic extraction of search result record is important for many applications. ViNT employs result page features such as visual content as it is shown on a browser and the HTML tag structure of the source file. Manually generating search result record wrappers is costly, time consuming and impractical for many applications. Visual information and Tag structure which is based on wrapper generation is used to automatically produce wrappers. ViDE focuses on the issue of how to extract the dynamically generated search result pages returned by search engine. A result page contains multiple SRR's and some of the irrelevant information to the users query. Accurate wrappers entirely based on the HTML tag structure. This method makes less sensitive to the misuse of the HTML tags.

Searching is done either manually which is ineffective and

hard to maintain. It gets difficult for the users to access number of web sites individually to get the needed information. H. He et al.[10] proposed WISE-Integrator tool which performs integration of Web Interfaces of Search Engines automatically. It is used for identifying the similar attributes from distinct search interfaces for integration. WISE-Integrator is capable of automatically grouping elements into logical attributes and deriving a rich set of meta-information for each attribute.

J.Zhu et al.[11] proposed Hierarchical Conditional Random Field approach. Current approach makes use of decoupled strategies. The data record detection and attributes labelling is done in two separate phases. It gets ineffective the idea of extracting data records and attributes separately. It proposes a probabilistic model to perform both processes simultaneously. HCRF can integrate all useful features by learning by their importance, and it can also integrate hierarchical interaction. Its limitations are cost and template dependency.

J. Wang et al.[12] proposed DeLa, a method which is very similar to proposed annotation work. DeLa's alignment method is based on HTML tags, on the other hand proposed work uses other features such as text content, adjacency information, data type, proposed annotation method deals with relationships between text nodes and data units, DeLa utilizes different search interfaces of WDBs for annotation.

3. Proposed Work

This paper considers how to assign labels to the data values present in the SRRs automatically. From a collection of search result record which have been extracted from a result page returned from a web database.

A. Objectives

- Perform data extraction.
- Perform data alignment

B. System flow

Components of system flow includes SRR's, data extraction, data alignment, combing tag value similarity, SRR results. The system flow is shown below in fig.2,

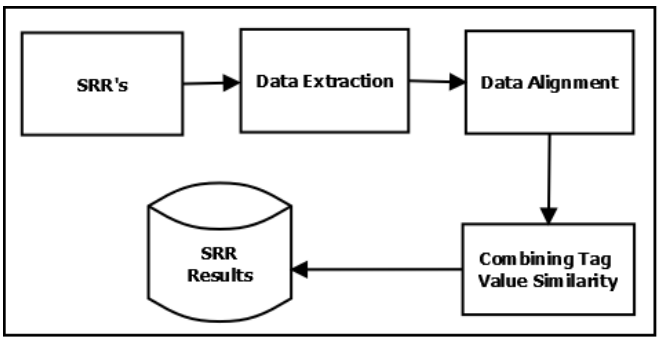

Figure 2: System Flow

4. Data Extraction

For the annotation purpose, the search result records have to be extracted from the returned result pages. So as the irrelevant information such as links, information about the hosting site and advertisements are to be discarded from the result page. It is very time consuming and tedious for extracting the data records from the result page by manually written programs and it is not practical as the search engine changes the result page display over period of time. So as to extract the records from the results page a wrapper generator which bases on Visual information and Tag structure is used. The extraction of data using ViNT is based on features such as visual content of the web page and also the HTML tag structure of the page in HTML format.

Each search result record is saved in a tag tree structure with one root and every node in the tag tree corresponds to an HTML tag in the original page.Fig.3 shows the tag tree structure of a html page, using this structure, it becomes easy to find each node in the HTML page. The information like physical position , its coordinates and area size of each node can also be obtained using ViNTs.

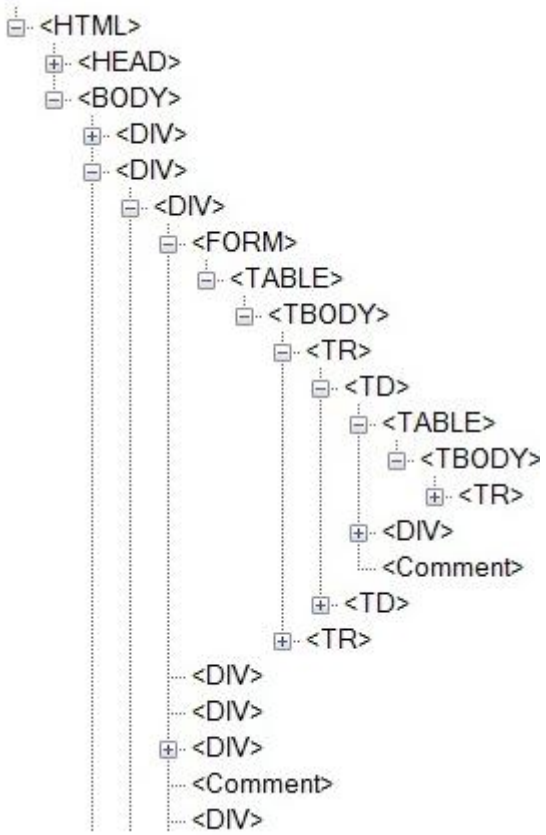

Figure 3: Tag tree structure of a webpage

5. Data Alignment

Once the extraction of search result record from the web page is done, the data units are generally not aligned. The aim of alignment of data is to place the data units from search result records into a group which has same semantic. Alignment of records makes annotation of data a lot easier. It is based on the idea that the data units in different search results of the same semantic mostly have the determined presentation and layout.

Data Alignment concentrates on five features for similarity such as Data Unit, Data Content, Presentation Style, Data Type, Tag Path Similarity which is described in [8].So as to

increase the efficiency of data grouping and alignment, a cluster based shifting technique is used.
Alignment is carried out based on same features, a group of data units which are having similar features are put in one group by aligning it. If a group contains data units of one concept and if there is no other data unit of another concept then the group is known as well aligned group. The aim of alignment is to put the data units in the table so every alignment group is well aligned.

The goal of data alignment is to put the data units of the same concept into one group so that they can be annotated comprehensively.

Alignment Steps:

a) Merge Text Nodes
It detects and removes decorative tags from every SRR, which permits the text nodes identical to the same attribute to be merged into a single one.

b) Align Text Nodes
After merging, it aligns text nodes into different groups. So that same group has the same concepts.

c) Split (Composite) Text Node
In this step the composite text nodes are splitted into separate data unit.

d) Align Data Units
This is the last step for alignment, in which every composite groups are separated in different multiple aligned groups, which contains data units of same concept.

Alignment Algorithm:

1) Read Source HTML file which contains the records.
2) Process each record in html nodes in the source html file one by one.
3) For every "Node" in the "Root", check if the element contains data or empty node.
4) If element contains "data node", then we are going to consider them as fully qualified records, which can be used in accessing for search process.
5) If element doesn't contain "data node", which might be missing in construction of document which is of no use we are going to eliminate them for further processing.

Clustering algorithm :

1) Read all fully available records from the annotation stage.
2) For each record evaluate all the "child node", and if child nodes contain full data then those records will be taken high distance records.
3) Non-available "child nodes" will be pushed in to the last part of SRR generation.4.
4) When user performs "search" SRR's will be processed according to fully available data.

6. Data alignment, labelling and wrapper generation:

Automatic annotation is based on alignment approach which aligns the data units by using different types of relationship in between data units and text nodes. A cluster-based shifting algorithm is used in alignment process. After the successful alignment label the data units and automatically construct an annotation wrapper for the search site.

7. Experimental Results

The fig.4 shows the performance graph of the system, the experiments performed on the various html datasets and results that are emerged ,the results shows that the implemented technique outperforms the previous work done.

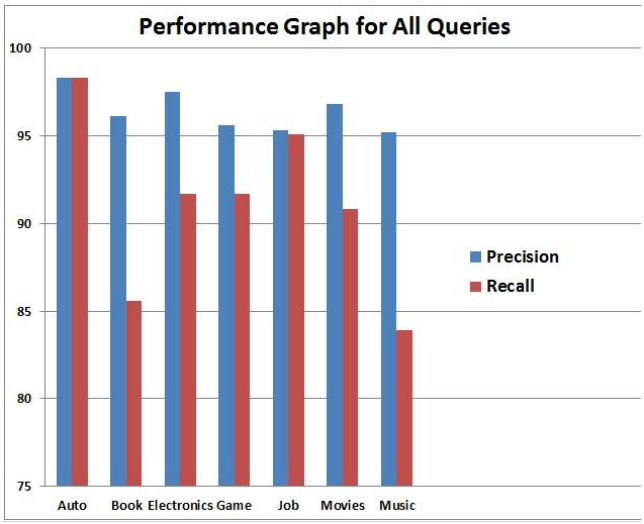

Figure 4: Performance Results for HTML datasets

8. Conclusion

In this paper, the data annotation problem is mentioned and Implemented a multi-annotator approach to annotate the SRR'S, an automatic annotation wrapper is used to search result records retrieved from web database. A new data extraction method is implemented so as to extract search result records automatically from a result page. It uses two steps for this task, first it includes identifying and segmenting the search result records. Existing methods are improved by allowing the SRR'S in a data region to be non-adjoining. In second step it aligns the data values among the SRR's. A unique alignment method is implemented in which the alignment is performed pairwise and nested structure processing. Experimental result shows that CTVS is more accurate and performs better.

References

[1] Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.
[2] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc.Very Large Data Bases (VLDB) Conf., 2001.
[3] J. Wang, J. Wen, F. Lochovsky, and W. Ma, "Instance-BasedSchema Matching for Web Databases by Domain-Specific Query Probing," Proc. Very Large Databases (VLDB) Conf., 2004.
[4] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large

Web Sites," Proc. Sixth Int' Workshop the Web and Databases (WebDB), 2003.

[5] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "BootstrappingSemantic Annotation for Content-Rich HTML Documents," Proc.IEEE Int'l Conf. Data Eng. (ICDE), 2005.

[6] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2,article 12, June 2009.

[7] D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31,no. 3, pp. 227-251, 1999.

[8] W. Liu, X. Meng, and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 3, pp. 447-460, Mar. 2010.

[9] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. Int'l Conf. World Wide Web (WWW), 2005.

[10] H. He, W. Meng, C. Yu, and Z. Wu, "Automatic Integration of Web Search Interfaces with WISE-Integrator," VLDB J., vol. 13,no. 3, pp. 256-273, Sept. 2004.

[11] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma, "Simultaneous Record Detection and Attribute Labeling in Web Data Extraction," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006.

[12] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. WorldWideWeb,2003.

[13] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.

[14] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE 16th Int'l Conf. Data Eng. (ICDE), 2001.

[15] W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.

[16] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine,"