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

Data Annotation and Construction Wrapper for Web Databases

Deepika D Phalak[1], H. A. Hingoliwala[2]

[1]PG Student, Computer Engineering Department, JSPM's JSCOE, Savitribai Phule Pune University, Pune, India

[2]Professor & HOD, Computer Engineering Department, JSPM's JSCOE, Savitribai Phule Pune University, Pune, India

Abstract: ***Now-a-days organizations use many more databases for their day to day operations. These tremendous amounts of databases have been accessed on the web in HTML form. When a query is submitted to the search interface these web pages are retrieved. The data units extracted from the stored database are merged into the pages dynamically and results are shown on the webpages. So for this annotation approach is used. In this it first aligns the data units into different groups in a way that data in the same group have the same meaning. Annotation label is assigned to each group on the basis of different annotators. An annotation wrapper is constructed in order to annotate the new results in same database without performing the whole annotation process. Database is going to increase in order to search more terms and to get more results.***

Keywords: Annotator, Annotation label, Annotation wrapper, Data Annotation.

1. Introduction

Databases are the technologies for dealing with large number of data. Webpage are the most efficient and easy way to represent information. Data unit represents real word entity i.e. value of a record under an attribute. Data alignment and its annotation increase the efficiency of searching information. Data alignment is the process of collecting the data into different groups in such a way that data in the same group have same meaning. Annotation is the method used for labelling i.e. adding extra information to the document, paragraph etc so that it simplifies what type of data and which document it is. In short it is the method of assigning meaningful labels. For example, in library various books are arranged in a rack. So rack labelled as "JAVA" might hold books of java programming language. Thus it enables fast retrieval of information in large amount of database.

Same on the web, result page retrieved from web database (WDB) consists of several search result records (SRRs) and each record consists of multiple data units. These data units are encoded dynamically into result pages for human browsing and converted into machine processable format and then assigned meaningful labels. Encoding of data units requires more human efforts to annotate or label data units manually. So it gives poor scalability.

In this we present an automatic annotation approach. This approach arranges the data units into different groups and make sure that each data unit in the same group have same meaning.

Each group is then annotated on the basis of different basic annotators to make a final label. Semantic labels are not only important for record linkage task but also used for storing collected SRRs into a database table. At the final stage, a wrapper is generated. It is generated in order to annotate the new results in same database without performing the whole annotation process. Automatic annotation approach is highly effective and more scalable.

Web Pages data extraction
Arasu and H. Garcia-Molina/ *SIGMOD Int'l Conf. Management of Data/2003*
Our price **$15**, put in the basket

Automatic Annotation of Data Extracted from Large Web Sites
L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, / *Workshop the Web and Databases (WebDB)/ 2003*
Our price **$15**, put in the basket

Experiments on Multistrategy Learning by Meta-Learning
P. Chan and S. Stolfo, / Proc. Second Int'l Conf. Information and Knowledge Management (CIKM)/ *1993*
Our price **$15**, put in the basket

Combining Approaches for Information Retrieval
W. Bruce Croft / *Advances in Information Retrieval: Kluwer Academic/2000*
Our price **$15**, put in the basket

RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites
V. Crescenzi, G. Mecca, and P. Merialdo, / *Proc. Very Large Data Bases (VLDB) Conf./ 2001*
Our price **$10**, put in the basket

Figure 1: Original HTML Page

```
<FORM><A> Web Pages data extraction</A><BR> A.
Arasu and H. Garcia-Molina /<FONT><I> SIGMOD
Int'l Conf. Management of Data  / 2003
</I></FONT><BR> Our Price <B>$15 </B>
```

Figure 2: Resulting Source Code of HTML Page

For wrapper construction we are specifying some rules that tell us how to extract data of concept and what is the semantic label. So it is helpful to annotate the retrieved data from the same database without applying the same annotation process.

The data of interest has been collected by an individual from various web databases. For example, an individual wants to

access and purchase the research paper from the web database. To do this activity, the data should be properly label in such a way that the data can be used for further analysis. However, in many incidents, the data access from the web database are not properly organized and or labeled. In many applications, the individual annotate manually the data units. The Figure 1 shows a sample original HTML page extracted after the query in the web databases. Figure 1 shows the first five records of the web database. The corresponding the source code for the figure 1 is shown in figure 2.

The data units from the first record shows that "Extracting Structured Data from Web Pages", and "Arasu and H. Garcia-Molina" are data units and text nodes. The data is being accessed from the web database. If the resulting data contain four fields, then do the comparison of the other fields. The basic objective is to compare one result with the other results. Thus, there is a need to check the semantics of the result viz. Data unit.

The following paper is organized as: Section 2 describes the previous survey of paper. Section 3 specifies the implementation part i.e. the workflow, algorithm used and contribution. Section 4 concludes the paper.

2. Related Work

Information retrieval and annotation has been populated in research area. Wrapper induction systems rely on human users for mark and labelling. They produce a series of rules known as wrapper to retrieve the same set of information on result pages from the same WDB. So the system obtains high extraction accuracy but suffers from poor scalability and is not suitable for online applications.

Old applications require large amount of human efforts to manually label the data units. We are considering how to automatically annotate data units in same search records.

Conceptual model based data extraction uses heuristics structural framework to retrieve information automatically and label them. But structural framework for various domains needs to be constructed manually.
Several works automatically gives meaningful labels to data units in results records. Data extraction from large websites annotate data units with the closest labels on result pages but its application is limited because some web databases do not encode data units with their labels in result records.

Existing automatic alignment techniques are depending on very one or very few features. Most probably used is HTML tag paths. It is assumed that subtrees related to two data units with different result records but having the same concept have the similar tag structure. But this assumption is not right as tag tree is sensitive to minor differences which cause wrong coding.

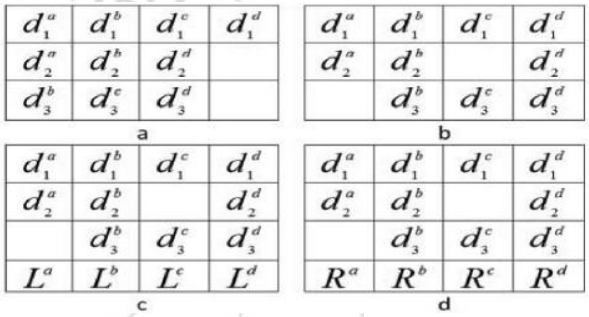
Vision based approach used visual contents on webpages to perform alignment. But it is not on data unit level rather it depends on text unit level. It is also not clear how it combines its annotation to make a single label.

A regular expression based data tree procedure uses HTML tags to align data units and filling them into a table. But the alignment is purely based on HTML tags only. Other important features like data type, content, presentation style are not taken into consideration. For each WDB wrapper is constructed only for data unit extraction.

In our approach, data alignment handles all relationships between data units and text nodes. Clustering based shifting technique is used to align the data. Here annotation wrapper is constructed for extraction as well as for assigning labels. Specifically it defines some rules.

3. Implementation

Let d_i^j represent the data unit into the $_i^{th}$ result records of concept j. The result records on a result page is shown (a) in a table format in which each row represents an SRR.



***Phase 1*- Alignment Phase:** Here we identify all data units in the result records and arrange them into different groups in a way that each group have different concept shown in (b). Data units having same semantic are grouped together in order to identify common patterns and features among them.

***Phase 2*- Annotation Phase:** In this, annotators are used to create label for the data units in their group. Fig (c) shows that semantic label L^j is allocated to each column.

***Phase 3*- Annotation Wrapper Generation Phase:** As shown in (d), for each concept an annotation rule is generated which describes how to retrieve data units in the result page.

***Alignment Algorithm*-**
1 From each result records, it detects and removes decorative tags so that text nodes with same attribute are merged in single text node.
2 Align text nodes into groups with the same semantic.
3 Then it is necessary to determine whether group needs to be again split. So here we have to identify separators.
4 After separation of composite group again they need to be aligned.

***Clustering Algorithm*-** Here for each SRR we are performing breadth first traversal on the DOM tree to remove the decorative tags. After that we will try for depth first traversal to remove decorative tags.

Label assignment- Labeling is performed with the help of LIS and IIS i.e. local and integrated interface schema. LIS faces the problems like label inadequacy and label inconsistency. To overcome these problems we are designed integrated interface schema. This schema collects all the attributes of LIS and creates a global attribute which has unique name.

Annotation Wrapper Construction- Annotation wrapper is the rule defined to retrieve result records from the same database for new search terms without applying the whole annotation process. The annotation rule consists of 5 parameters: Label, prefix, suffix, separators, and unitindex. If the prefix and suffix of data are same as specified in the rule it is correct and the label is assigned at the position unitindex. To split the data units separators are given.

A. **Five features of data units:**

a) ***Data content-*** Data content are the keywords of same concept used to search the information quickly. For e.g. keyword "java" returns the relevant information of word java.
b) ***Presentation style-*** It describes how data is displayed on the webpage. Some styles are font, color, font size etc.
c) ***Data type-*** Date, time, integer, percentage etc. are the data types considered in approach.
d) ***Tag path-*** It is the sequence of tags traversed from root to corresponding node in tree. Every node has its 2 parts: tag name and direction.
e) ***Adjacency-*** It means preceding and succeeding data unit.

B. Relationship between data unit and text node:

Data unit is value of a record and text nodes are visible elements on webpage. To determine how many data units are contained in text node some relationships are defined.

- One to one relationship- Each text of the node contains exactly the value of single attribute means only one data unit.
- One to many relationship- Each text node contains multiple data units.
- Many to one relationship- In this multiple text node form a data unit.
- One to nothing relationship- Text node are not part of any data unit.

So these 5 common features of text node and data unit and their relationships are used to remove decorative tags from data units. They are helpful to separate decorative tags and text node.

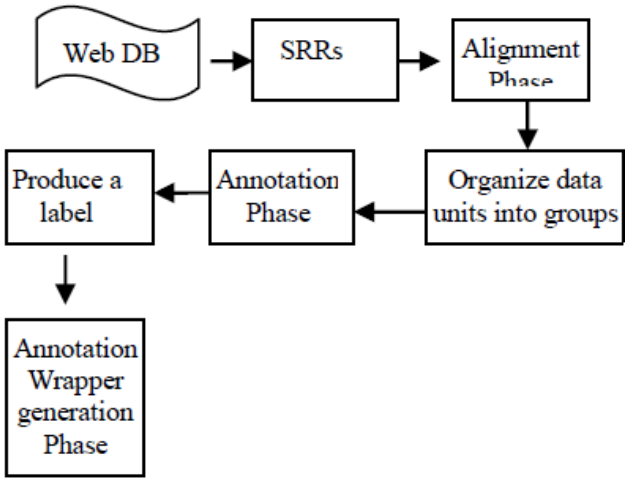


Figure 3: Annotation Phases

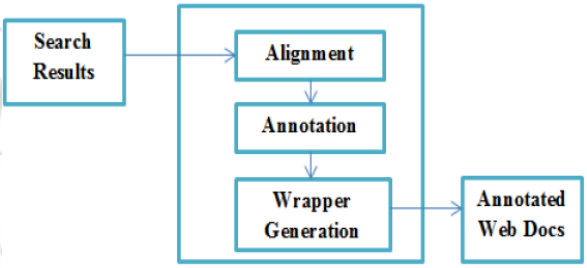


Figure 4: Proposed Framework for Annotation

4. Conclusion

Data extraction and data annotation is primarily research area in web database. Several types of data unit and text node features makes annotation scalable and automatic. Three phases used for annotation in which aligns the data units into different groups, labels each group and construct annotation wrapper.

Acknowledgement

It is with deep sense of gratitude that authors acknowledge the sincere help of concerned people for providing very constructive suggestions to improve the method.

References

[1] 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.
[2] 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.
[3] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," *Proc. Very Large Databases (VLDB)Conf.*, 2009.
[4] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," *Proc. SIGMOD Int'l Conf. Management of Data*, 2003.
[5] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large

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

[6] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," *Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE)*, 2007.

[7] 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.

[8] Y. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, "Annotating Search Results from Web Databases," IEEE Trans. Knowledge and Data Eng., vol 25, no .3, mar.2013.