RESEARCH ARTICLE OPEN ACCESS

Data Annotation for the Web Databases

Miss. Priyanka P. Boraste
Department of Computer Science and Engineering
Matoshri College of Engineering & Research Centre
Nashik, University of Pune
Maharashtra - India

ABSTRACT
In recent years, there is an immense growth in the databases and the information technology. These databases are being accessed by using HTML and web technology. During this process, a data unit is come back from the database. The resulted data units are being encoded into the end of the resulting pages. The resulting data unit is used in various application viz. Deep web collection and Internet shopping. However, the encoded data units need to be extracted from the database and allot a meaningful label. In this paper, we presented a state of the art review of the methods used in the data annotation for the web databases. In addition, we present an analysis of the various methods and a theoretical proposal for the system.

Keywords:- Data alignment, Data Annotation, Web Database, Wrapper Generation

I. INTRODUCTION

An immense amount of web databases is being used in the various search engines. However, the search engines gives out the multiple records while accessing the web databases. The resulting records from the web databases show various data units. When we access the web database, then we received the outcomes from the search engine.

The automatic annotation solution consists of three phases- Alignment phase, Annotation phase, and Annotation wrapper generation phase. Also used six basic annotators; where each annotator can independently assign labels to data units.

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.

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: The 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 the HTML page

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.

In this paper, we presented a state of the art review of the approaches for the data unit level annotation from the web databases. As well, we presented our proposed approach for the data annotation for the web database.

The basic procedure of our proposed approach is listed bellow in the following steps. Initially, a web database returns a multiple search record result (SRR). Each SRR contain several data units. Detect text nodes from the search record result. Extract data unit and text node features. Check data unit from the knowledge database. If the data units is remains present in the knowledge database then make alignment as well add contain contents from the knowledge database. Finally display the result. If the data unit does not present in the knowledge database then add data unit to the knowledge database. The process is carried out in such a way that the data annotation will be done automatically.

The paper is organized as follows: Section 2 presents the state of the review of previous work. Section 3 presents the analysis work and the implementation details. Section 4 discusses the step used in our new proposal and results. Finally, a conclusion is presented in section 5.

II. RELATED WORK

In this section, we presented an analysis of the various approaches used in the data annotation for the web databases. Table 1 shows a summary of the approaches used in the data annotation of the web databases.

A recent literature [1] [2] reports, that the traditional approach takes much time to annotate the database. And also requires much effort manually. However, the issue of assigning a label to the data units automatically has been discussed in [1]. Author has discussed three phases viz. Alignment phase, annotation phase and annotation wrapper generation phase. In the alignment phase, the data unit is organized into various groups. The grouping of the same data units facilitates to recognize feature and patterns with the data units. In the annotation phase, a label is produced for each data unit in the group and a suitable label is assigned to each group. In the annotation wrapper generation phase, the annotation rule is generated and apply the annotation rapidly.

The wrapper induction system is introduced in [2][3] which mark the label data and also relies on human users. However, this system achieves higher extraction precision in the result. In addition, this system undergoes lesser scalability and also not fit in the application [4] [5].

A similar approach is being introduced in [6] that based on ontology and extract data from the web documents automatically. In [7], the author introduced a domain dependent annotation process. However, this process manually assigns the label to the data. The ontology based system has been introduced in [8]. However, this method is insightful to the data quality.

To automatically build a wrapper has been presented in [9][10][11]. However, this method is used only for the data extraction, but not for the annotation. The various methods were discussed in the literature [12 [13], [14] that assign the label to the data from the web databases.

The method which is used to annotate the data is introduced in the literature [12]. However, this method is useful only for the limited application. In [14], a data tree algorithm is presented that uses regular expression to align the data from the web databases. Another method [13] that focus not only on the attribute extraction, but also on the assigning the label to the data units.

A method to align the data has been discussed in [7] and [14]. The approach proposed in [1] report that they maintain all the type of relationship between the text nodes and data units. The method [7] maintains only one to one relationship between the text node and the data units. The method [14] maintains one to many relationships between the text node and the data units. A similar approach to execute the task of data alignment and the wrapper has been addressed in [11]. However, this approach is useful only for the alignment of the text node, but not for the data unit in the web database. In addition, this method uses only visual feature. The method discussed based on the similar concept is DeLa [14]. However, this method uses the HTML tags.

III. IMPLIMENTATION DETAILS

In this section, we presented a state of the art review of the approaches used in the existing literature. In recent years, web information extraction and annotation is an active research area.

TABLE 1
Summary of the Approaches with characteristic

SR. NO.	REFERENCE NUMBERS	CHARACTERISTIC
1	[1]	MAINTAIN FULL RELATIONSHIP
2	[2][3]	LESS SCALABLE, WRAPPER INDUCTION STEP USED
3	[6]	AUTOMATICALLY EXTRACT DATA
4	[6][7][8]	ONTOLOGY BASED APPROACH
5	[9][10][11].	DATA EXTRACTION ONLY

We investigate the issue of relationship, scalability, wrapper induction, automatically data extraction, and the ontology based approaches. However, the clustering approaches used in the approaches are of limited use. Therefore, our further step is to adopt and experiment by using other clustering approaches.

Annotation phases: Automatic annotation consists of three phases such as Alignment phase, Annotation phase and annotation wrapper phase. These phases are shown in figure 3. In an alignment phase, SRR first identified all the data units. After the identification of the data units, then next step is to organize the data units into different groups. However, every group will contain different concepts.

In the annotation phase, potentially consist of several basic annotators. Each annotator consists of specifically one type of features. However, each annotator is used to produce a label for each data unit. In the third step, the annotation wrapper phase is used. This step basically generates the annotation rule. These rules are used to describe the result page and also specify the way in which the data unit is extracted. In addition, it finds out the semantic label for individual data units. To align the data units into different groups clustering based shifting technique is used. Hence, the data in the same the same group have same semantic. All this three annotation phase used to align the data units from web databases into different groups, annotating the data units with different aspects then generate an annotation wrapper rule.

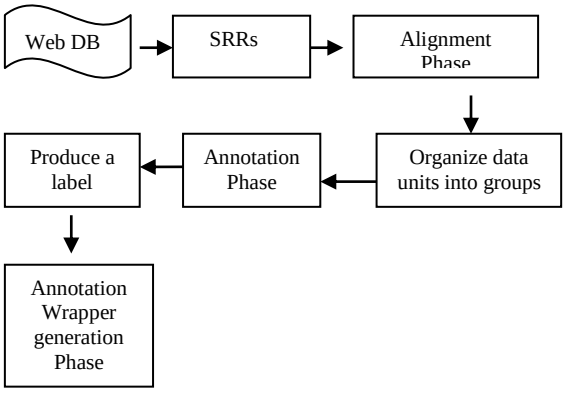

Figure 3: Annotation Phases

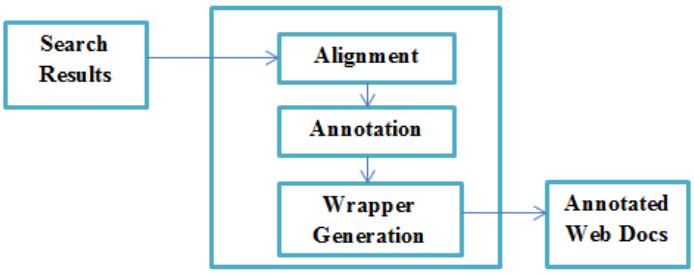

Figure 4: Proposed framework for automatic annotation

A. Alignment Algorithm

Alignment algorithm has following four steps.

Step 1: Merge text nodes: This step detects and removes decorative tags from each SRR to allow the text nodes corresponding to the same attribute merge into a single one.

Step 2: Align text nodes: After the merging aligns text nodes into different groups. So that same group has the same concepts.

Step 3: Split text nodes: In this step split the composite text nodes into separate data unit.

Step 4: Align data units: This is the last step for alignment in which separates each composite group into

Multiple aligned groups with each containing the data units of the same concept

The algorithm for data alignment assumes that the attributes of the data are in some specific order for all the rows. The assumptions make the algorithm work in that fashion. Generally this assumption is true for many search results that are presented in tabular format.

IV. RESULTS

In this section, we proposed a theoretical framework for the data annotation in the web databases. The basic procedure is listed below in the following steps.

1. Initially, a web database returns a search record result (SRR). Each SRR contain several data units.
2. Detect text nodes from the search record result.
3. Extract data unit and text node features.
4. Check data unit from the knowledge database.
5. If the data units are remains present in the knowledge database, then makes alignment as well add contain content from the knowledge database. Finally display the result.
6. It the data unit does not present in the knowledge database, then add data unit to the knowledge database.

As per our humble observation in the existing literature [1], we observed that the performance of the algorithm could be improved by injecting the step number 5 and 6 in the algorithm [1]. Thus, the algorithm presented in the literature [1] can be improved.

When we have to submit a query to the page title and URL and load the page then it shows ten search records related to the same query the extract the SRR and detects the data unit and text node features. After the successful extraction of the data units calculated the tag path and data type.

After this load the ViNTS from all five search engines and finally group that data in such a manner that data inside the same group that have the same semantic and annotate it with different groups.

Following figure 5 shows five different SRRs from five different search engines.

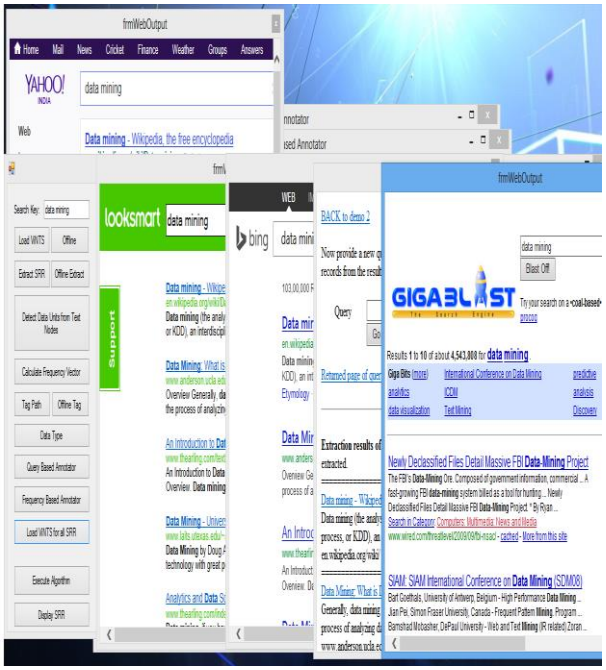

Figure 5: Query related search results

Figure 6 indicate that different types of groups can be formed, where data takes from all five search engines.

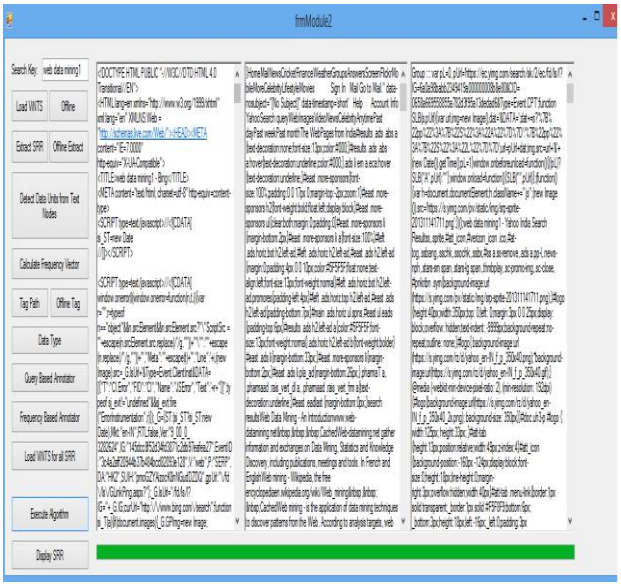

Figure 6: Group formation

V. CONCLUSION

A web data extraction and data annotation is a prime research area in the web database. In this paper, we discussed a state of the approaches used in the data annotation search problem

in the web databases. The prime research issue in the web database is data annotation and data alignment. Therefore, we propose a new proposal for the data annotation in the web database.

Thus, our future work is to use other clustering approaches to automatically obtain the data units with annotation and labelling.

REFERENCES

[1] Y. Lu, H. He, H. Zhao, W. Meng, C. Yu, " Annotating Search Results from Web databases" In IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No.3, 2013.

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

[3] 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.

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

[5] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large- Scale Metasearch Engine," Proc. IEEE/WIC Int'l Conf. Web Intelligence (WI '03), 2003.

[6] 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.

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

[8] 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.

[9] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.

[10] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.

[11] 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.

[12] 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.

[13] 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.

[14] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.

AUTHOR

First Author – Miss Boraste Priyanka P., Post Graduate Student, Department of Computer Engineering, MCOE&RC, Nashik, Pune University. Maharashtra, India. (E-mail id: dokhalepriyanka@gmail.com). date of birth 16th March 1990.Complited engineering in Information Technology from BAMU University, Jalna, Maharashtra.