MJRET
Open Access
Multidisciplinary Journal of Research in Engineering and Technology
Journal homepage: www.mjret.in
ISSN:2348 - 6953

REVIEW ON AUTOMATIC ANNOTATION OF QUERY RESULTS FROM DEEP WEB DATABASE

Mr. Chaitanya Bhosale, Prof. Sunil Rathod
Computer Dept.
Dr. D. Y. Patil School Of Engineering(Affiliated to Savitribai Phule Pune University)
Pune, India
chaitanyab.03@gmail.com,sunil.rathod@dypic.in

Abstract: *In recent years, web database extraction and annotation has received much attention from the database and Information Extraction(IE) in research area due to the volume and quality of deep web. Many web databases are accessible through HTML form-based interface. When query is submitted to the search interface the query result page is generated. Search Result Records(SRRs) are the result pages obtained from web database(WDB) and these SRRs are used to display the result for each query. Every SRR contains multiple data units equivalent to one semantic. These search results can be utilized in other web applications such as comparison shopping, data integration,metaquerying. To make these applications successful the search pages should be annotated in a meaningful fashion. To make it effortless for human,an automatic annotation approach is suggested. Within this,we first groups the data units of result records so that the information in the group have the same meaning. After that we annotate each group in different aspects and obtain the final annotation after aggregating them. In addition,we use a new CTVS technique for extraction of QRRs from a query result page,in which we use optional labeling and dynamic tagging for the improvement. Further an annotation wrapper is generated automatically which can be used for annotation new result records from the same WDB.*

Keywords: *Data extraction, Data alignment, automatic wrapper generation, web database.*

1. INTRODUCTION

Databases are established technologies for managing huge amount of data.World Wide Web is a favourable way of presenting information across the internet. Alignment and Annotation of data improves the quality of information search and data updates. Data Alignment is the way of arranging data and accessing in computer memory.Data Annotation is the methodology for adding extra information to a document, a word or phrase, paragraph or the entire document. In other words Data Unit Annotation is the process of assigning meaningful labels to data. For example, a folder in a computer system labeled as “Trip-2014” might hold files of photographs taken in trip.

Data annotation enables in quick retrieval of useful data embedded in the deep web. Search Result Records (SRRs) are the result pages obtained from web database (WDB) and these SRRs are used to display the result for each query. Every SRR contains multiple data units equivalent to one semantic.A data unit is a part of text that meaningfully represents real world entity concepts. Dynamically for human browsing these data units are encoded into the result page and assigned meaningful labels.Human efforts are required to annotate the data units. Therefore, lacks scalability. To overcome this, automatic assigning annotations to data units within the SRRs is needed. An automatic annotation approach first arranges all data into different groups such that within same group all data units have same meaning. Then each group is annotated in different aspects and aggregated to predict a final annotation. Finally, wrapper is generated. Wrappers are commonly used as translators which annotate new result records from the similar web database. This automatic annotation approach is scalable and highly effective. A clustering based shifting technique is proposed to align the data units into different groups.

We introduce new technique called New Combined Tag and Value Similarity (NCTVS) for extraction of QRRs from a quey result page. With this, record extraction and alignment is done,in addition there is optional labeling and dynamic tagging.

- Optional labeling is the technique by which the problem of elimination of optional attribute that appears as the start node in data region, as supportive information is eliminated.This is included in the record extraction step.
- In the existing system which uses static tagging, results are far less accurate.The Dynamic tagging uses the semantic data extraction concept described below.
- First here we examine the relationship between text node and data units and perform data unit level annotation.

- To align the data units of different groups of same meaning we propose a clustering based shifting technique. Our system also considers some important features such as data contents (DC), data types (DT), presentation style (PS), and adjacency (AD) information.
- To improve the quality of data unit annotation we used the integrated interface schema (IIS) over various WDBs in the same area.
- In this we use six basic annotator which results are combine to form a single label.
- Then new annotation wrappers are constructed for any WDB. In which wrappers are used to annotate the same web database with new queries more easily.

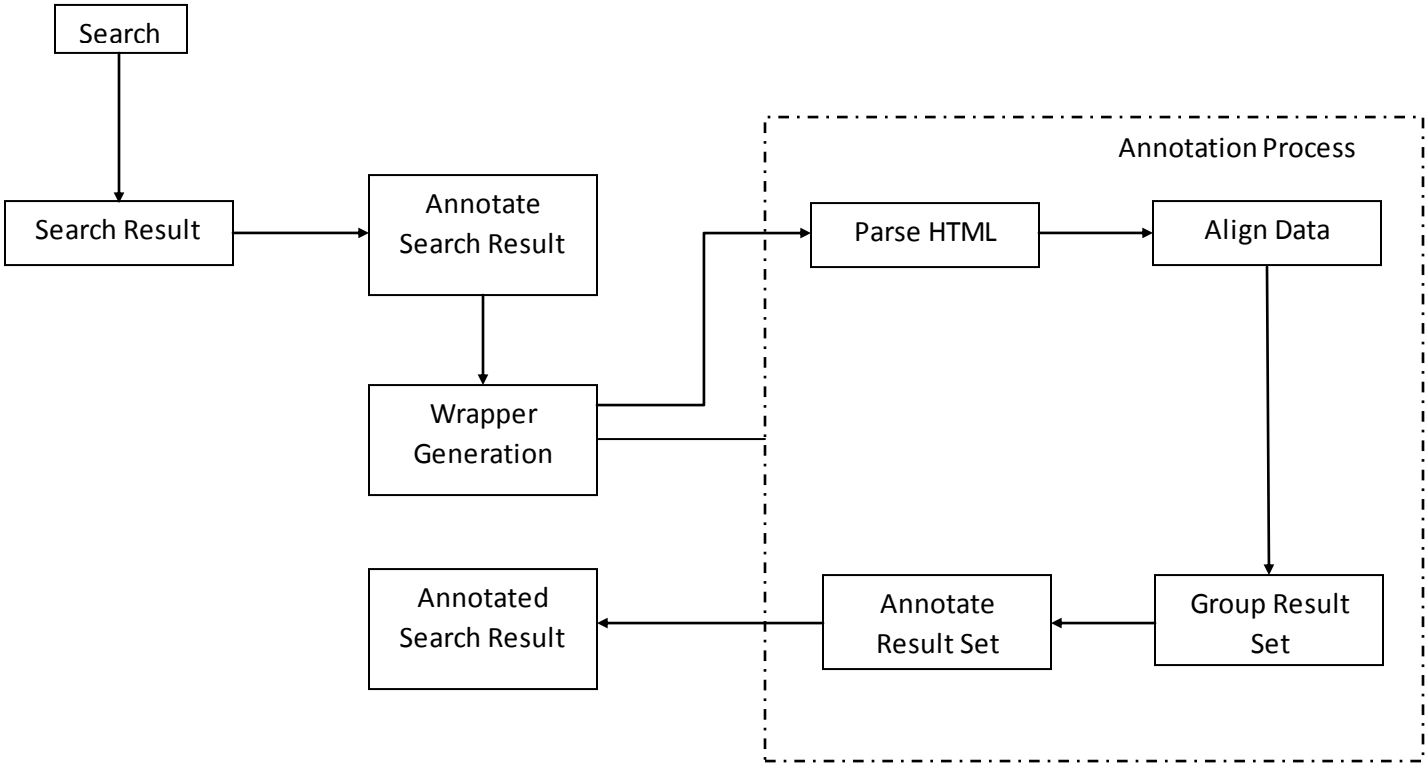

Fig.1: Architecture Diagram

2. RELATED WORK

In recent years, web database extraction and annotation has received much attention from the database and Information Extraction (IE) in research area. Many systems like wrapper induction [3],[4] depends on human users to mark and label the desired data.Then they induce a series of rules called wrapper to extract the similar set of information on result pages from the same web database. Hence, the system achieves high extraction accuracy

through supervised training and learning process.But they suffer from poor scalability and not suitable for online applications like metasearch engine [12].

A similar approach [5],[18] is based on ontology means, in which it automatically extracts the data from web documents. Authors S. Mukherjee, et al. [7] discussed a method to align the data units which maintains only one type of relationship i.e. one to one relationship in between data unit and text nodes.For various domains ontologies are constructed manually.

The efforts to automatically build a wrappers has been presented in [1], [2], [6]. But, this method is used only for the data extraction, not for the annotation. The various methods were discussed in the literature [11], [10], [14] that assign the meaningful label to the data from the web databases. Most of the previous approaches of automatic data alignment techniques are based on few features like HTML tag paths(TP) [13], ViDIE uses Visual features [6], splitting of SRR into text segments [8].

The existing technique proposed in [1] report that they maintain all the type of relationship between the text nodes and data units. While the method [7] maintains only one to one relationship between the text node and the data units and the method [8] maintains one to many relationships between the text node and the data units. The method discussed based on the similar concept is DeLa [10]. But this method uses the HTML tags.It handles only two types of relationship between the text node and data units where we use all type of relationships. Here DeLa uses only local interface schema (LIS) search interface of WDBs for annotation process.

J.Madhavan et al [15] define about deep web crawl in which content hidden behind HTML form which is obtained by form submission with valid text input values. Here, an algorithm ISIT is used to select input values for text search input that accept keywords. Y.Lu [9] describe about annotating the structured data of the deep web. It is similar to [1] and our method.where in this paper they describe about four relationships between text node and data units but only two of them i.e. one-to-one and one-to-many are explained [9] in detail.

In addition, we use clustering shift algorithm for one to nothing relationship where Y.Lu et al use pure clustering algorithm.

In existing applications data units are manually annonated which requires lots off human efforts, which limit their scalability. Now, meaningful annotations are based on correct extraction of query results. Presently automatic web data extraction has been relatively grown. In this approach, the method of automatic web data extraction defined in Combining Tag and Value Similarity (CTVS) for data extraction and alignment purpose [16] is adapted.

CTVS deals the Tag and Value similarity, in which data is automatically extracted from query results result records after the very first identification and segmention of the query result records (QRRs) in the query result page and then alignment of the segmented QRRs in a table is done where the data values of the similar characteristics are put into the identical column. In this approach we introduced new method called New Combined Tag Value Similarity (NCTVS) for the extraction of QRRs from query result page. NCTVS improves the data extraction accuracy in two ways i.e. optional labeling and dynamic tagging.

ViNTs [17] – For learning wrapper generation it requires a set of training pages from a website which uses both visual and tag features. Firstly it utilizes the visual data value similarity without considering the tag structure to check data value similarity regularities, denoted as data value similarity lines, and then combines them with the HTML tag structure regularities to generate wrappers. Both visual and non visual features are used to weight the relevance of various extraction rules.For this technique, the result page must contain at least four QRRs, and one no-result page is required to build a wrapper.The input used in system is URL of search engine's interface.

3. CONCLUSION

Assigning semantic labels to the extracted data unit of each SRR is a challenging task. The automatic multiannotator approach considers several types of data unit and text node features and makes annotation scalable and automatic. Here three phases used for automatic annotation in which alignment of the data units into different groups, labeling of each group and construction of an annotation wrapper. A new algorithm for data annotation in the web database would be proposed. The proposed technique would be implemented with the expected results by using knowledge database as a database. We presented a novel data extraction method, NCTVS, to automatically extract QRRs from query result page with optional labeling and dynamic tagging.

ACKNOWLEGEMENT

I wish to express my sincere thanks to the guide Prof. Sunil Rathod and Head of Department, Prof. Soumitra Das , as well as our principal Dr.S.S.Sonavne , also Grateful thanks to our PG Coordinator Prof.P.M.Agarkar and last but not least, the departmental staff members for their support.

REFERENCES

[1]. Hai He, Hongkun Zhao, Y. Yiyao Lu, Weiyi Meng, "Annotating Search Result Records from web databases," IEEE Transaction on Knowledge and Data Engg., volume 25, No. 3, mar. 2013

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

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

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

[5]. 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 Engg., volume 31, No. 3, pp. 227-251, 1999

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

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

[8]. H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," Procedure Very Large Databases (VLDB) Conference, 2009.

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

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

[11]. L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," Procedure Sixth Int'l Workshop the Web and Databases (WDB), 2003.

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

[13]. Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Procedure 14th Int'l Conference World Wide Web , 2005.

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

[15]. D. Ko, L. Lot, V. Ganapathy, A. Rasmussen, J. Madhavan, and A.Y. Halevy, "Google's Deep Web Crawl," Procedure VLDB Endowment, volume 1, no. 2, pp. 1241-1252, 2008.

[16]. Frederick H. Lochovsky ,Weifeng Su, and Jiying Wang "Combining Tag and Value Similarity for Data Extraction and Alignment", IEEE Transactions on knowledge and Data Engineering, Volume 24, no.7, pp. 1186--1200, July 2012.

[17]. W. Meng, Z. Wu, V. Raghavan, H. Zhao, and C. Yu, "ViNTs - Fully Automatic Wrapper Generation for Search Engines", In Proceedings of the 14th World Wide Web Conference, pp. 66-- 75, May 2005.

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