ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

Dynamic Annotation by Web Database Search Results

Bincy S Kalloor[1], Sheeja Agustin[2]

PG Scholar, Department of CSE, Marian Engineering College, Trivandrum, India[1]

Assistant Professor, Department of CSE, Marian Engineering College Trivandrum, India[2]

ABSTRACT: The Internet provides a great extent of beneficial knowledge which is usually formatted for its users, which makes it troublesome to extract relevant data from diverse sources. The World Wide Web plays an major role as all kinds of information repository and has been very success full in disseminating information to users. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and internet comparison shopping, they need to be extracted out and allot meaningful labels. Search result presents an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then for each group annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new results from the same wed database. Proposed system is to make the system dynamic by outlier detection technique.

KEYWORDS: Data alignment, Data annotation, Web databases, Wrapper generation

I. INTRODUCTION

Data mining is the computational process of discovering patterns in the big datasets. The overall goal of the data mining is to extract information from a datasets and convert it into an understandable structure for further use. Web mining is the one among the application of data mining techniques to discover patterns from the web. Web mining can be divided in to three types, they are web usage mining, web content mining and web structure mining. Search engines are very important tools for people to reach the vast information on the World Wide Web, large portion of a deep web is a database based. That is for many search engines, data encoded in the returned result pages come from the underlying structured databases such type of search engine is often referred as web databases. Result page returned from a web database has multiple search result records. Each search result records contain multiple data units. Now there is a high demand for collecting data from multiple WDB's. Early applications require tremendous human efforts to annotate the data units manually, which severely limit their scalability. In this paper, how to automatically assign labels to the data units within the SRR's returned from the WDBs. Automatic annotation solution consists of three phases they are alignment phase, annotation phase and annotation wrapper generation phase. In the alignment phase, first identify all the data units in the SRRs and then organize them into different groups with each group corresponding to a different concept. Grouping data units of the same semantics can help to identify the common patterns and features among these data units. In the annotation phase the introduction to multiple basic annotators with each exploiting one type of features. Every basic annotator is used to produce a label for the units within their group holistically, and a probability model is adopted to determine the most label for each group. In this annotation wrapper generation phase a annotation rule is generated that describes how to extract the data units of this concept in the result page and what the appropriate semantics label should be. The rule for all aligned groups, collectively, form the annotation wrapper for the corresponding web database, which can be used to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phase again.

II. RELATED WORKS

In recent years web information extraction and annotation has been an active research areas. Extracting structured data from deep web pages is a challenging problem [2].Some of the limitations are webpage programming dependent, Incapable of handling ever increasing complexity of HTML source code. To overcome this problem- a vision based approach [2] vision based data extractor. ViDE is used to extract structured results from deep WebPages automatically. It can only process deep web pages containing one data region while there is significant number of multi-data region deep WebPages, which is time consuming process.[3]ODE which automatically extracts the query results records from the HTML pages. Automatic data extraction is important for many applications such as meta-querying, data integration and data warehousing. In semi automatic wrapper induction has the advantage that no extraneous data are extracted as the user can label only the data in which he/she is interested. To overcome this supervising learning methods are used [4]. Labour intensive and time consuming are the drawback and also it is not scalable to a large number of websites.[4] Technique for extracting data from HTML sites through the use of automatically generated wrappers. A key problem with the manually coded wrappers is that writing them is usually a difficult and labour intensive task and difficult to maintain [4] has a prior knowledge about the page contents. It is an daunting task for users to access numerous web sites individually to get the desired information.[5] is a tool that perform automatic integration of web interfaces of search engines. It is to identify matching attributes. [6] ViNTS is automatically producing wrappers that can be used to extract search result records dynamically. It utilizes both the visual features on the result page displayed on browser and HTML tag structure of the source file. It helps people to locate and understand information. Existing approaches use decoupled strategies [7]. A probabilistic model to perform two tasks simultaneously. HCRF can effectively integrate all useful features by learning their importance.

III. MULTI ANNOTATOR APPROACH FOR ANNOTATING SEARCH RESULT RECORDS

In this approach first user gives the query to the search engine. In returned result page containing multiple SRR, the data unit corresponding to the same concept often share special common features. After the feature selection data alignment is done. The purpose of data alignment is to put the data units of the same concept into one group so that they can be annotated holistically. There are six basic annotators [1] to label data units, with each of them considering a special type of patterns/features. Once the data units on a result page have been annotated, and use these data units to construct an annotation wrapper for the WDBs. So that the new SRRs retrieved from the same WDB can be annotated using this wrapper quickly without reapplying the entire process. Finally display the results.

IV. ALIGNMENT ALGORITHM

```
Collect dataset
Clustering
Choose the annotation method
If annotation=table annotator
Perform table annotator
Elseif
If annotation=query based annotator
Perform query annotator
Elseif
If annotation= schema based annotation
Perform schema annotation
Elseif
If annotation=frequency based annotation
Perform frequency annotation
Elseif
If annotation=intext prefix/suffix annotator
Perform intext prefix/suffix annotation
Elseif
If annotation =common knowledge annotator
Perform common knowledge annotation
Elseif
Annotation wrapper
End
```

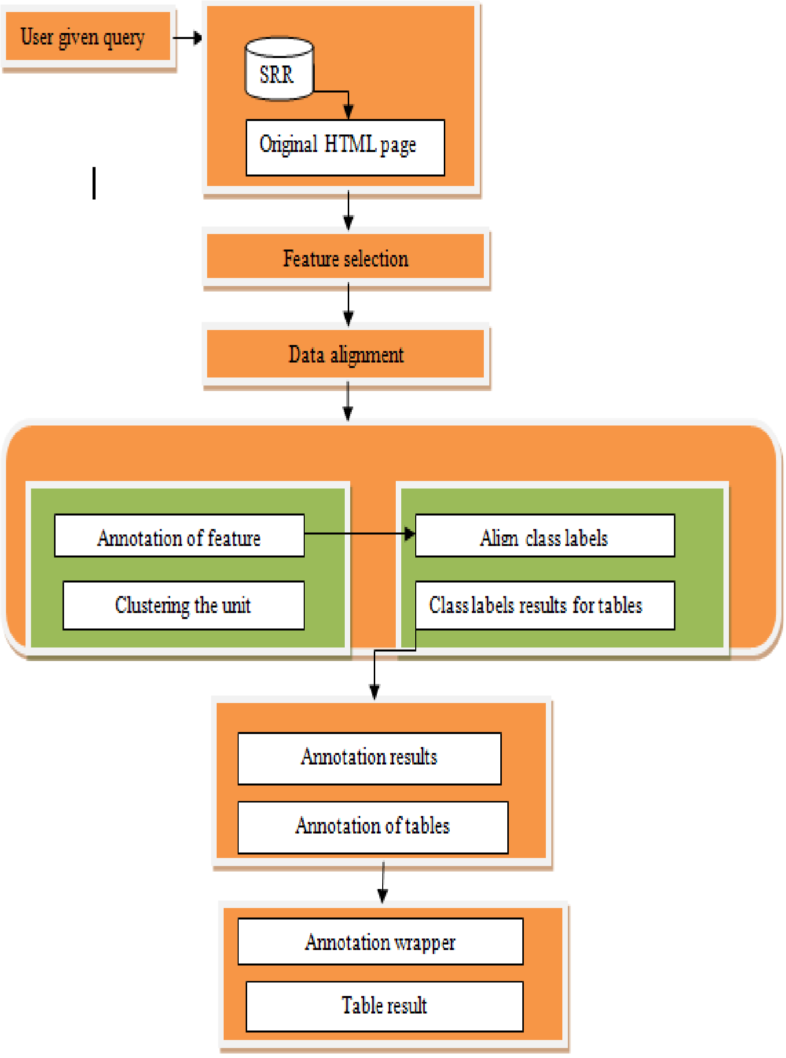

Fig 1: Architecture for annotating search results from web databases

Alignment algorithm is to move the data units in the table so that every alignment group is well aligned, while the order of the data units within every SRR is preserved and also needs the similarity between two data unit groups. In this first create the alignment groups then the clustering is done. After clustering choose the annotation method. There are six annotation methods are there. First set table annotator then perform the function of table annotator. Table annotators first identify the column header. Then for each SRR takes a data unit. Then select the column header with maximum overlap. At last a unit is assigned and labelled. Then query based annotation is done, in this first set of query terms are there. From that find the group with largest occurrences and label is assigned. In the schema annotator attribute is identified with highest matching score. In the frequency annotator find the common preceding units then concatenated preceding units and label the group. In text prefix/suffix annotator check the data units and share the same prefix or suffix. The common knowledge annotator from the group of data units match the patterns or values and label the group. At last wrapper is generated. Each annotated group of data unit corresponds to an attribute in the SRR. The data unit groups are annotated and organised based on the order.

V. PERFORMANCE ANALYSIS

The performance of the annotating search Results from web database can be analysed basis of the speed. Early annotation is done manually. It is a time consuming process and also not suitable for large number of websites.

Automatic annotation approach is proposed. From this approach the speed of the data set is increased and the processing time is reduced.

VI. CONCLUSION

Extracting structured data from deep web pages is a challenging problem due to the underlying intricate structure of web pages. Until now a large number of techniques have been proposed to address this problem but all of them have inherent limitations. Early annotations are done manually; it is a time consuming process and low efficiency. To overcome this problem automatic annotation is proposed. In this approach speed of the dataset is increased and processing is reduced.

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng," Annotating Search Results from Web Databases" IEEE Transaction on Knowledge and Data Engineering, vol. 25, NO. 3, March 2013

[2] 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

[3] 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

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

[5] 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

[6] 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

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

BIOGRAPHY

Bincy S Kalloor, pursuing Master of Technology in Computer Science and Engineering from Marian Engineering College, Kerala. She published paper in international journal. Her research areas are Data mining and Information Security.

Sheeja Agustin is an Assistant Professor in Computer Science and Engineering Department, Marian Engineering College, Kerala. She is pursuing PhD in Computer Science and Engineering from Noorul Islam University, Tamil Nadu. She receives Master of Technology in Computer Science and Engineering from MS University, Tirunelveli. She receives Bachelor of Technology from Kerala University. She published many number of papers in reputed International and National journals. Her research interests are Image Processing, Neural Networks, Fuzzy Set Theory and Applications.