A Survey - Web Information Extraction and Annotation

K. Anusha
PG Student, Dept.of Computer Science
V.P.M.M Engineering College for Women,
Affiliated to Anna University Chennai, India

A. John De Britto, Asst. Professor
Dept.of Computer Science
V.P.M.M Engineering College for Women,
Affiliated to Anna University Chennai, India

Abstract— The tremendous growth in the volume of data and with the terrific growth of number of web pages provides user an easy way to access the information and services. Since all the databases have become web accessible through Html form-based search interfaces. Such data is huge and for applications such as online shopping comparison, article collection etc and the data units returned from the databases are usually encoded into the result pages dynamically for human browsing and cannot be processed by machines. For the encoded data units to be machine processable they need to be extracted out and assigned meaningful labels. Annotation of such collected information leads to several advantages including fast decision making, relevant information visiting, to reduce the time of futile searches, historical data management and elimination of older searches. This paper is intended to provide an alignment of the entire data unit on the search result record and organize them into different group such that the data in the same group have same semantic and we use multiple annotators to produce label for the data units and aggregate the different annotations to determine the most appropriate label for each group. Next an annotation wrapper is constructed to efficiently annotate new result pages from the same web database

Keywords— Data alignment, Web database, data annotation, Wrapper generation, Data Extraction

I. INTRODUCTION

The tremendous growth in the volume of data and with the terrific growth of number of web pages provides user an easy way to access the information and services. With increasing information overload, we are facing new challenges for not only locating relevant information but also accessing variety of information from different resources. However most web pages are designed still for human consumption and cannot be processed by machines. Furthermore, web search engine, the most popular tool to retrieve the web pages, does not offer support to interpret the results. For that human intervention is still required. As the size of the search result is often just too big for human to interpret and finding relevant information on the web is not as easy as we would desire. Most of the traditional search engines get the answer syntactically correct but large in amount. The semantic web adds more semantics to the web pages so that they can be processed by machines as well as humans. The semantic web is a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. For many search engines the data returned from the databases are usually encoded in the returned result pages come from structured database such type of search engine is referred as (WDB). The result page from the WDB has multiple search records (SRR) has multiple data units.

In this paper, we perform automatic data unit level annotation whereas early applications require tremendous human effort to annotate data unit manually. Collect a set of SRR that has been extracted from a result page from WDB. Our automatic annotation approach consists of three steps. First, we identify all data units in the SRR and align them into different group such that data in the same group have same semantic. For example all titles are grouped together. Second for each group we annotate it from different aspects and aggregate the multiple annotations to predict a final annotation label for it. Third the annotation wrapper is generated automatically for the search site and can be used to annotate new result pages from the same web database instead of performing the alignment and annotation process again. Annotation Wrapper can perform annotation quickly which is essential for many online applications.

I. RELATED WORK

An Information extraction and annotation has been an active research area. In wrapper induction systems [4], [5] they rely on human users to mark and label the desired information. They induce a series of rules called wrapper to extract the same set of information on result pages from the same web database. Hence, the system achieves high extraction accuracy through supervised training and learning process they suffer from poor scalability and not suitable for online applications. Conceptual-model-based data extraction [3] uses ontologies with heuristics to extract information automatically from the result pages and label them. Ontologies are defined as structural framework for organizing information. Ontologies for various domains are constructed manually. Several works in [1], [8] automatically assigns meaningful labels to the data units of SRRs. In data extraction from large websites [1] annotates data units with their closest labels on the result page. This method has limited applicability since they do not encode data units with labels on result pages. In ODE [8], first ontologies are constructed using query interface and result pages from the same web database. Domain ontologies are used to label each data unit and with the same label they are aligned. This

IJERT

method is sensitive to quality and completeness attributes. Previous approaches of automatic data alignment techniques are based on few features: HTML tag paths [9], Visual feature [6], splitting of SRR into text segments [2].

II. OVERVIEW

Each SRR extracted has tag structure i.e. each node is either a tag node or text node. Tag nodes are HTML tags surrounded by "<" and ">" in Html source whereas text nodes are visible elements on web page that are outside the angular brackets. Since our annotation is at data unit level first, we need to identify data unit from text node.There are four types of relationship between data unit denote as U and textnode denoted as T.

A. *One-to-one Relationship*
In this type each text node contains exactly one data unit. It is denoted as (T=U). Each tag node surrounded pair of tags <A>and </A> is a value of title attribute such a text nodes are atomic text nodes that are equal to data unit.

B. One-to-Many Relationships
It is denoted as T ↄ U. multiple data units are encoded in one text node. The text of such kinds of nodes is called as composition text nodes.

C. Many-to-One Relationship
It is denoted as T C U. Multiple text nodes together form a data unit. The value of author attribute is contained in multiple text nodes.

D. One-to-Nothing Relationship
It is denoted as T≠ U. text nodes belonging to this type are not a part of any data unit instead they are semantic labels describing the meaning of corresponding data unit. Example: Author, Publisher.

III. DATA ALIGNMENT

The purpose of data alignment is to align the data unit of same concept into one group so that they can be annotated easily. Our data alignment method is based on assumption that attributes appear in same order because SRR from web database are generated by same template program. In our work we consider the SRR of result page in a table format such as each row holds SRR and each column represent a data unit. Each table column is referred as a alignment group. The goal of alignment group is to move data unit in the table so that it is well aligned. This method consists of four steps:

A. Merge text nodes
In this step we first detect and removes decorative tags from the result records and merge same attribute into single text node. The many-to-one relationship between text node and the data unit has decorative tags. We need to remove them to restore the integrity of data unit. To remove the decorative tag we use breadth-first traversal over DOM tree of SRR. If the decorative tag is if decorative tag is identified delete it and immediately move child node as right sibling.

B. Align text nodes
Align text nodes into group so that each group contains same concept. For example book search result, a book would not have discount price if it is not on sale. Such element may have different concepts. We merge two clusters that have highest similarity else the elements with least similarity are shifted to next group.

C. Split text node
Split (composite) text nodes. This step aims to split the "values" in composite text nodes into individual data units. This step is carried out based on the text nodes in the same group holistically. A group whose "values" need to be split is called a composite group.

D. Align data unit
Align data units. This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept.

The SRR for the given search engine are usually formatted based on a template.By Template Mining algorithm greatly help in extraction and annotate the data unit within each record correctly. Template mining can identify both template text and non template text nodes and also addresses the mismatches between tag structure and data structure.

IV. LABEL ASSIGNMENT

A. *Integrated Interface Schema*
When a query is submitted for a search interface the entities in the returned results have hidden schema. The hidden attribute discovered in the search result schema does not have matching attribute this causes local interface schema inadequacy problem and also raises inconsistent label problem. Using LIS (local interface schema) different labels are assigned to semantically identical data units returned from WDB. This is potential problem as LIS give different names to same attribute. In our approach we use Integrated Interface Schema (IIS) with WISE- Integrator for multiple web databases. The IIS combines all the attributes of the LIS. The matched attribute from different LIS are combines as values of IIS. Each attribute has a global name and an attribute mapping table is created. Our annotation method uses both IIS and LIS to annotate the retrieved data from web database. By using IIS there is a potential to increase the annotation recall and also we can alleviate both local interface schema inadequacy and inconsistent label problem.

B. *Multiannotator Approach*
We introduce 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.

If multiple labels are predicted for a group of data unit by different annotator, we compute the combined probability model to determine the label by selecting the label with largest combined probability. The advantage of this model is highly flexible.

C. *Wrapper Generation*
Once the data units are annotated on the result page. We use annotated data units to construct annotation wrapper for web databases so that new data search result record can be automatically annotated using wrapper without reapplying entire annotation process. Annotation wrapper is a description of annotation rules for all attributes on result page. Every SRR has tag node sequence and text. We can

IJERT

scan sequence to obtain both prefix and suffix of data unit. The scan stops when the data unit encountered is valid with meaningful label assigned. Prefix/suffix method is difficult to extract data unit package inside composite node due to the fact that there is no Html tags within a text node but separators. So to overcome this problem we use position index to split unit vectors. To use the wrapper to annotate new result record annotation wrapper is applied.

V. CONCLUSION

Multiannotator based data extraction from web database is proposed for automatic annotation and wrapper generation of search result records from web database. This Multiannotator approach and probabilistic method combine different annotators to produce highly flexible and high quality annotation. A special feature of our method is that when annotating results retrieved from a web database it utilizes both Local interface schema and integrated interface schema of multiple web database from same domain. Our clustering based shifting method also capable of handling different relationships between text node and data units.

.

REFERENCES

[1] 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.
[2] H. Elmeleegy, J. Madhavan, and A. Halevy, "Harvesting Relational Tables from Lists on the Web," Proc. Very Large Databases (VLDB) Conf., 2009.
[3] 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.
[4] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.
[5] 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.
[6] 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.
[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] Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. 14th Int'l Conf. World Wide Web (WWW '05), 2005.
[10] Y. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, "Annotating Search Result Records from web databases," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 3, mar. 2013.

IJERT