INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

A SURVEY ON ANNOTATING SEARCH RESULTS FROM WEB DATABASES

Y. Pauline Jeba[1], Mrs. P. Rebecca Sandra[2]

[1]PG Full Time Student, CSE Department, Karunya University, Coimbatore

[2]Assistant Professor, CSE Department, Karunya University, Coimbatore

[1]ypaulinejeba@gmail.com, [2]rebecca@karunya.edu

Abstract

Web search engines are designed to search information in the web database and to return dynamic web pages. Many databases are web accessible through HTML form-based interfaces. When a query is submitted to the search interface these web pages are retrieved. Each web page contains several result records that are relevant to the query. Every search result records (SRRs) contains multiple data units corresponds to one semantic. These search result records are needed to be extracted and assigned meaningful labels. Previously result records from web pages are extracted and assigned labels manually thus resulted in poor scalability. To reduce human efforts a multi-annotator approach is proposed to automatically extract data units and assign labels. Extracted data units are aligned into groups and ensured that each data unit under a group has same semantic concept or meaning. Then an annotation wrapper is generated automatically and used to annotate new result records from the same web database.

Keywords: Data alignment, Data unit level Annotation, Wrapper generation

1. Introduction

Databases are established technologies for managing large amount of data. Web is a good way of presenting information. Alignment and annotation of data increases the efficiency of searching and updating information. Data alignment is the way of arranging data and accessing in computer memory. Data annotation is the methodology for adding 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. For example, a folder in a computer system labeled as “Holiday-2013” might hold files of photographs taken on holiday. Data annotation enables fast retrieval of information in the deep web. A result page retrieved from a web database consists of several search result records (SRRs) and each result records consist of multiple data units. A data unit is defined as the values that represent real world entities. These data units are encoded dynamically into result pages for human browsing and converted into machine process able unit and assigned meaningful labels. The encoding of data units requires lot of human efforts to annotate data units manually. Thus, lack in scalability. To overcome this, automatic assigning of data units within the SRRs is required.

An Automatic annotation approach [10] is proposed. This approach first arranges the data units into different groups. And ensures that each data unit within a group has same semantic i.e., meaning. Each group is then

Annotated in different aspects and aggregated to predict a final label. Finally, a wrapper is constructed. Wrappers are commonly used as translators which annotate new result pages from the same web database. This automatic annotation approach is highly effective and more scalable. The section 2 of this paper explains about the related works. Section 3 gives an detail explanation about automatic annotation approaches and finally section 4 concluded with the description of automatically assigning labels using automatic annotation approach.

2. Related works

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 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].

3. Automatic annotation approaches

Consider a set of SRRs that are extracted from a result page returned from the web database. The Automatic annotation approach has three major phases as illustrated in Figure. 1. Let d_i^j denote a data unit, belonging to the i^{th} SRR of concept j. Figure 1a represents SRR in table format.

Phase 1: Alignment phase

In alignment phase first the data units are identified in the SRRs and organized into different groups. Each group corresponds to a different concept. (e.g., all titles of books are grouped together). Figure 1b shows the result of phase 1 with each column containing data unit of sane meaning across all SRRs. This phase is used to identify common patterns and features among data units.

Phase 2: Annotation phase

In annotation phase several basic annotators are used with each exploiting one type of features. Every annotator is used to predict a label for the data units within the organized groups. A probability model is used to determine the most appropriate label. Figure 1c shows the result of phase 2 where every group assigned with a semantic label L^j.

Phase 3: Annotation wrapper generation

In annotation wrapper generation phase an annotation rule R^j is generated for each identified entity or concept. This rule describes how to extract the data unit and what semantic label should be and collectively forms a wrapper. A wrapper is used to annotate the data units retrieved from same web database for new queries and thus performs annotation quickly. Figure 1d shows the result of phase 3 with a rule denoted as R^j.

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	

(a)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b		d_2^d
	d_3^b	d_3^c	d_3^d

(b)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	
L^a	L^b	L^c	L^d

(c)

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	
R^a	R^b	R^c	R^d

(d)

Figure 1: Illustration of three phase annotation approach

3.1 Data unit and text node

The visible elements on the web page represent a text node and the data units are located in the text nodes. Relationships between text node and data unit features are,

- **One-to-One Relationship**

  Text node containing exactly one data unit, i.e. the text of this node contains the value of a single attribute. Each text node surrounded by the pair of tags <A> and </A>. This type of text nodes are referred as atomic text nodes. An atomic text node is equivalent to a data unit.

- **One-to-Many Relationship**

  Multiple data units are encoded in one text node. This type of text nodes are referred as composite text node.

- **Many-to-One Relationship**

  Multiple text nodes together form a data unit. This type of text nodes is referred as decorative tags because they are used for changing the appearance of part of the text node.

- **One-To-Nothing Relationship**

  Text nodes are not part of any data unit inside SRRs. This type of text node is referred as template text node.

There are five common features shared by the data units

- Data content
- Presentation style
- Data type
- Tag path
- Adjacency

3.1.1 Data content

Data unit or text node of same concept shares certain keywords which are used to search the information quickly. For e.g., keyword “machine” will return the information that are relevant to word machines.

3.1.2 Presentation style

Presentation feature describes how a data unit is displayed on a web page. Few of the styles are font face, font size, colour, text decoration etc.

3.1.3 Data type

Data types are predefined characteristics that have their own meaning. Basically used data types are date, time, currency, integer, decimal etc.

3.1.4 Tag path

A Tag path is a sequence of tags traversing from the root of the SRR to the corresponding node in the tree. Each node contains two parts a tag name and a direction indicating whether the next node is a sibling or the first child node.

3.1.5 Adjacency

Adjacency refers to the data units that are immediately before and after in the SRR. They are termed as preceding and succeeding data unit.

3.2 Data alignment and labelling

The existing works in [2], [7] differs when compared with the automatic annotation approach. They are based on one or few features. In automatic annotation [10] the alignment approach first handles the relationship between data units and text nodes and utilizes different types of data unit features. And a cluster-based shifting algorithm is used in alignment process.

Label assignment is performed using IIS (Integrated Interface Schema) and LIS (Local Interface Schema). IIS contains the attributes in all the LIS and thus eliminates label inadequacy and inconsistent label problems. Few basic annotators in [10] are introduced to annotate the aligned groups and a probability model is used to combine the results of multiple annotators. This approach is called multi-annotator approach.

4. Conclusion

Assigning meaningful labels to the extracted data unit of each SRR is a challenging task. The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. Multiple annotators of different features are used to annotate the extracted information from the result pages. Each annotators exhibit one special type of feature and they are together used to automatically construct a high quality annotation wrapper. Uses both LIS and IIS for label assignment and alleviates local interface schema inadequacy and inconsistent label problem.

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.