IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. XI (Mar-Apr. 2014), PP 68-70
www.iosrjournals.org

A Survey on Data Annotation for the Web Databases

Miss.Priyanka P.Boraste 1
[1](*Department of Computer Engineering, MCOE & RC, University of pune, Maharashtra, India)*

Abstract : *Web search engines are designed to search information in the web database and to return dynamic web pages. Data unit's returns from the databases and information technology are accessible through HTML form-based interfaces and web technology. Web pages are retrieved when a query is submitted to the search interface. Each web page contains several search result records related to user query. Every SRR contains multiple data units each of which describes one aspect of a real-world entity. Then SRR get extracted and assigned meaningful labels. To reduce human efforts a multi-annotator approach is proposed to automatically extract data units and assign labels. After the successful extraction align the data units into different groups where, data inside the same group have the same semantic(meaning).Then automatically annotation wrapper can generated and used to annotate new result records from the same web database.*

Keywords: *Data alignment, Data annotation, Data unit level Annotation, Web database, Wrapper generation.*

I. Introduction

Databases are established technologies for managing large amount of data. Web is a good way of presenting information. Efficiency of searching and updating information increases by Alignment and annotation of data. Data alignment is aligning the data or arranging the data in such a way that data inside the same group have the same meaning 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. Data annotation enables fast retrieval of information in the deep web. Data units comes from the web database consists of several search result records (SRR's). A data unit is a part of text that semantically represents real world entity concepts. Dynamically for human browsing these data units are encoded into the result page and assigned meaningful labels. Annotate the data units requires lots of human efforts. Thus, lack in scalability. To overcome this, automatic assigning of data units within the SRRs is required. An automatic annotation approach that first arrange all data into different groups i.e. inside the same group have same semantic. Then each group is annotated in different aspects and aggregated to predict a final label. Finally, wrapper is generated. This automatic annotation approach is scalable and highly effective.

A clustering based shifting technique is proposed to align the data units into different groups. This paper is organized in following sections. Section 2 reviews the related works. Section 3 gives a detail explanation about automatic annotation approaches also relationship between data unit and text node. Section 4 shows the proper data unit and text node features and alignment algorithm. Finally, section 5 concluded with description.

II. Related Work

In recent years, web information extraction and annotation is an active research area. The literature proposed in [1] reports that the traditional approach takes much time to annotate the database. It also requires enormous manual efforts. Automatically assigning the meaningful labels has been introduced in [1].Also discussed three annotation phases viz. Alignment phase, annotation phase and annotation wrapper generation phase. In data extraction from large websites [1] annotates data units with their closest labels on the result page. This approach proposed that they maintain all type of relationship between the text nodes and data units. The wrapper induction system is introduced in [2][3] which mark the label data and also rely on human users. However, this system achieves higher extraction precision in the result. In addition, this system undergoes lesser scalability that does not fit in the applications mentioned by authors [4][5].

A similar approach [6] is based on ontology means, 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. Also a domain dependent annotation process has been introduced. An ontology based system insightful to the data quality has been introduced in [8].

In [9][10][11] Automatically building a wrapper has been presented. These methods are used only for the data extraction, but not for annotation. The various methods discussed by the authors W. Liu, et al. [11] assigns the labels to the data from the web databases.

2.1 Outcomes of literature survey:

Issues of relationship, scalability, wrapper induction, automatically data extraction, and the ontology based approaches are investigated.

Clustering approaches adopted in the literature are limited; hence there is scope for linking clustering based methods with data annotation approaches.

Used search result records as a Database which will change accordingly.

III. Automatic Annotation Approaches

3.1 Annotation Phases

Phase 1: Alignment phase: In alignment phase align all the data into different groups. Each group corresponds to a different concept. (e.g., all titles of books are grouped together).

Phase 2: Annotation phase: In annotation phase used several basic annotators with each exploiting one type of features. Every annotator is used to predict a label for the data units within the organized groups and label the data units.

Phase3: Annotation wrapper generation phase: In annotation wrapper generation phase an annotation rule is generated for each identified entity or concept. To annotate the data units wrapper is used which retrieved from same web database for new queries. And thus performs annotation quickly.

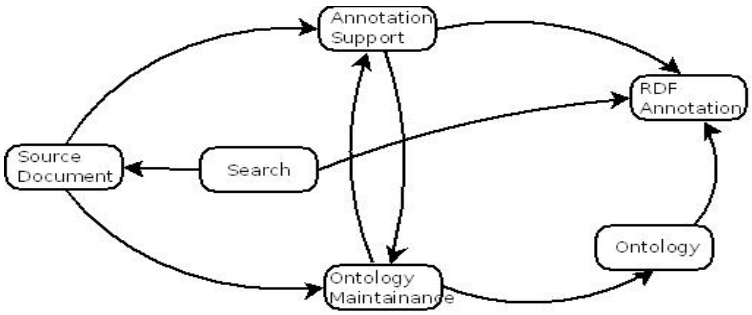

Fig: Phases of automatic annotation solution

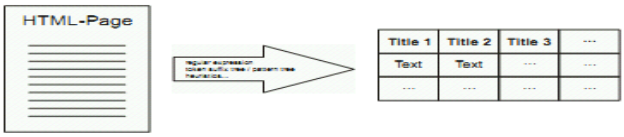

Fig: Extracts (automatically) text from a web-page into a table

1.2 Data unit and text node relationships:

Data unit is a piece of text that semantically represents concept of real world entity. Data unit is totally different from text node where, text node is a sequence of text surrounded by pair of HTML tags. Text node is visible element on the web page and data unit located in the text nodes.

Relationships between text node and data unit features are

1.2.1 One-to-One Relationship: (referred as atomic text nodes). Text node containing only 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 HTML tags <A> and </A>.

1.2.2 One-to-Many Relationship: (referred as composite text nodes) A text node consists of multiple data units i.e. multiple data units are encodes into single text nodes.

1.2.3 Many-to-One Relationship: (referred as decorative tags) multiple text nodes are encoded into single data unit.

1.2.4 One-To-Nothing Relationship: (referred as template text nodes) Text nodes are not part of any data unit inside SRRs.

This relationship for text nodes and data units are represents the relation in between them.

IV. Data Unit And Text Node Features

4.1 Features shared by data units.

4.1.1 Data content: To search information quickly data unit or text node of same concepts shares certain keywords.

4.1.2 Presentation style: This feature describes how a data unit is displayed on the web page by using few styles are out face, font size, color, text decoration etc.
4.1.3 Data type: These features are predefined characteristics that have their own meaning. Basically used data types are date, time, currency, integer, decimal etc.
4.1.4 Tag path: Sequence of tags traversing from root to corresponding node in the tree.
4.1.5 Adjacency: Adjacency refers to the data units that are immediately before and after in the SRR.

4.2 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.

4.3 Data alignment, labeling and wrapper generation:

Automatic annotation is based on alignment approach in which aligns the data units by using different types of relationship in between data units and text nodes. A cluster-based shifting algorithm is used in alignment process. After the successful alignment label the data units and automatically construct an annotation wrapper for the search site.

V. Conclusion

The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. Basically three phases used for automatic annotation in which aligns the data units into different groups, labels each group and construct an annotation wrapper. In this work not all data units are encoded with the meaningful labels. 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.

Acknowledgement

I feel great pleasure in submitting this paper "A SURVEY ON DATA ANNOTATION FOR THE WEB DATABASES" **"**. I wish to Thank IOSR Journals for giving us such a wonderful opportunity.

References

[1]. Y. Lu, H. He, H. Zhao, W. Meng, C. Yu, "Annotating Search Result From Web databases" In IEEE Transaction on Knowledge and
[2]. Data Engineering, Vol. 25, No.3, 2013 N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997\ L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrappe Construction System for Web Information Sources," Proc. IEEE
[3]. Conf. Data Eng. (ICDE), 2001W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48- 89,2002
[4]. Z. Wu et al., "Towards Automatic Incorporation of Search Engine into a Large-Scale Met search Engine," Proc. IEEE/WIC Int'l
[5]. 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
[7]. Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31 no. 3, pp. 227-251, 1999
[8]. S. Mukherjee, I.V. Ramakrishna, and A. Singh, "Bootstrappin semantic Annotation for Content-Rich HTML Documents," Proc.
[9]. IEEE Int'l Conf. Data Eng. (ICDE), 2005
[10]. 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.
Arasu and H. Garcia-Molina, "Extracting Structured Data from Web 5 Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003
[11]. V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: TowardsAutomatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.
[12]. W. Liu, X. Meng, and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans. Knowledge and
[13]. Data Eng., vol. 22, no. 3, pp. 447-460, Mar. 2010