International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Survey on Data Annotation for Search Results from Web Databases

Priyanka Ashok Jadhav
Department of Information Technology,
RMD Sinhgad school of Engineering
Pune, India

Prof. Snehal Nargundi
Department of Information Technology,
RMD Sinhgad school of Engineering
Pune, India

***Abstract*— A large portion of web is database based. This portion is accessible through html form based interface. For many search engines data result come from structured database called Web database(WDBs).Results returned from WDBs are called Search Result Records (SRRs).Data unit returned from WDBs are encoded into result pages dynamically for user browsing. These data unit are need to be extracted and assigned a meaning full labels. For encoded data unit to be machine process able the proposed system presents dynamic annotation approach which align data unit on result pages in different pages. Data in same group will have same semantic. This paper provides annotation for each group from different aspects and aggregate different annotation to predict final annotation label for it. An annotation wrapper is used to annotate new result pages from same database.**

Keywords— Data alignment, Data annotation, Web database,

I. INTRODUCTION

Data unit is piece of text that represent one concept and which corresponds to Value of record under attribute .This paper presents data unit level annotation. In day to day life there is high demand of retrieving data of interest from multiple databases. For example, vehicle comparison shopping system collect multiple result records from different vehicle sites ,it need to determine whether two SRRs refer to same vehicle. The system also need to list prices offered by different sites. Thus system needs to know the semantic of each data unit which are not provided in result pages. This paper presents how to automatically assign label to records returned from WDBs.

Having only semantic label for data unit is not only important but for later analysis storing collected SRRs into database table is also important. WDBs return the result pages from which SRRs has been extracted, the automatic annotation solution consist of three phases. First phase is *alignment phase* [3]. In this phase we first identify all data unit in SRRs and organize them into different group where each group corresponds to different concept [2]. Grouping data unit of same semantic help to identify common pattern and features among these data unit .These features are basis of our annotators. Second phase introduce multiple basic annotators with each exploiting one type of feature. Here every basic annotator is used to produce label for the units within their groups. In third phase for each identified concept rules are generated which describes how to extract data unit of this concept in the result page and what should be appropriate semantic label .The rules for all aligned group collectively form annotation wrapper for corresponding WDB which is used to directly annotate the data retrieve same from WDB.

II.DATA UNIT AND TEXT NODE FEATURES

Some features of data units are considered to group data units according to their semantics. Some time visual information of web pages such as layout, position, appearance is also considered to group the data units [2] [4]. Features of data units are:

1. Data content: nodes or data unit of the same concept usually shares certain keyword and have same leading label
2. Presentation style: This feature describes how data units are presented on web pages.
3. Data Type: Every data unit has predefined characteristics which have its own meaning. Commonly used data types are integer, decimal, date, time etc.
4. Tag Path: These are sequence of tags traversing from root to corresponding node in tree.
5. These are sequence of tags traversing from root to corresponding node in tree [1].

III. SYSTEM ARCHITECHTURE

System architecture contains four main parts:

1. Web Crawling

Input to the system is taken from different structured web databases. Required SRRs are extracted from result pages of structured databases. So web crawling is done to extract result pages from WDBs.

2. Data Alignment

Data alignment is done to put data unit of same concept into one group so that they can be annotate holistically. Two records belonging to same concept is determined by how similar they are based upon features such as data content, presentation style, data type.

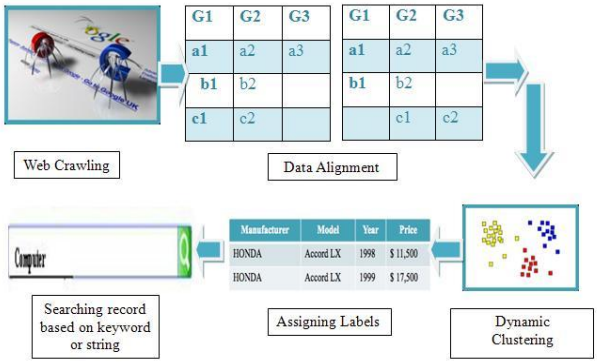

Figure.1 System Architecture

Here data alignment is based upon the assumption that attribute appear in the same order across all SRRs on the same result page although SRRs contain different set of attributes. We can conceptually consider the SRRs on result page in table format where each row represents one SRRs and each cell represent a data unit. Here we consider each column as alignment group which contain at most one data unit from each SRR. If an alignment group contain all data unit from same concept and no data unit from other concepts we call it as *well aligned* group. Goal of this group is to move data unit in the table so that every alignment group is well aligned while order of aligned group is preserved.

Data alignment method consists of following steps:

1. *Merge text nodes*: This step detect and remove all decorative tags from each SRR to allow the text nodes corresponding to the same attribute to be merged into single text node.
2. *Align text nodes*: Here text node is aligned into groups so that each group contain text node with the same concept.
3. *Split text nodes*: Splitting the 'values' into composite text node into individual data units. A group whose 'values' need to be split is called composite group.
4. *Align data units*: Separating each composite group into multiple aligned group with each containing the data units of same concept

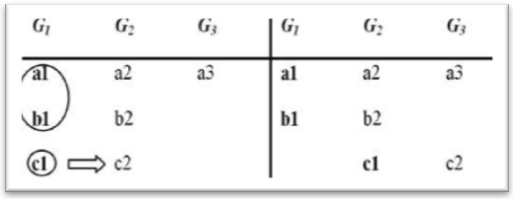

Figure 2. Step 2 of Data Alignment

3. *Dynamic Clustering*

Since particular SRR may have no value for particular attribute or may contain element of different concept. We can apply agglomerative clustering algorithm to cluster inside the group. First each text node form separate cluster of its own, then we repeatedly merge two cluster which having highest similarity value until no two cluster have similarity above threshold T. After clustering we obtain set of cluster V and each cluster contains the elements of the same concept only.

4. *Assigning Labels*

Result pages returned from WDBs contain multiple SRRs, The data units corresponding to same concept often share special features and such common features usually associated with data units in the result page in certain pattern. Based on these features some basic annotators are defined to label data unit with each of them considering a special type of pattern or feature. Four basic annotators are defined [1]. These are as follows:

- Table Annotator

SRRs returned from many WDBs are organized into table format where each row represents an SRR. In this table meaning of each column is represented by table header. Figure 3 shows example of SRR presented in table format. Data unit with same concept are well aligned with its corresponding column header. Table annotator first identify all column headers of table, then for each SRR it take data unit in cell and select the column header whose area has maximum vertical overlap within the cell. Similarly remaining data units are also processed.

Manufacture	Model	Class	Year	City	State	Price
HONDA	accord LX	4 DOOR	1998	playa del rey	CA	$11,500
HONDA	ACCORD LX	4 DOOR	1994	Spokane	WA	$ 7,500
HONDA	Accord Lx	4 DOOR	1997	Winona ake	ID	$ 8,700
HONDA	Accord LX	4 DOOR	1994	Cave Creek	AZ	$ 5,999
HONDA	Accord	4 DOOR	1999	Pomona	CA	$17,500

Figure 3. SRRs in table format

- Query Based Annotator

In this annotator SRR returned from WDB are always related to specific query. The query strings entered in search attribute of WDB my appear in some results of SRRs. Query based annotator work as follows: Given a query with a set of query terms submitted against an attribute A on the local search interface, First the group having largest total occurrences of query term are searched and gn(A) label is assigned to the group [1] DeLa [2] also uses query terms to match the data unit texts and use the name of the queried form element as the label.

- Schema Value Annotator

Schema value annotator first identify the attributes which having highest matching score among all attributes. After that these values are used to annotate the group. Attribute having more matches have first preference. This preference is given by multiplying the numbers having nonzero similarities over that having fewer matches. This improves the information retrieval effectiveness of combination systems [4].

- Frequency-Based Annotator

There are some attribute which having different values in different records. Every adjacent data units with the higher frequency are considered to be attribute names while the data units with the lower frequency come from databases as embedded values. If we consider group G_a having data units of lower frequency. This annotator tries to find common preceding units shared by all the data units of the group G_a. This can be done recursively by preceding chains until the encountered data units are different. These all units are concatenated to form labels of group G_a.

Annotation Wrapper

When annotation for data units on result pages is completed, these annotated data unit are used to construct annotation wrapper for the WDB. This wrapper is used to annotate the new SRRs retrieved from WDB. By using this wrapper annotation can be done quickly without reapplying the completed annotation process.

IV. CONCLUSION

We have studied the data annotation and annotation wrapper for annotating the search result records retrieved from web databases. In this paper we have also studied data alignment and cluster bases shifting. This method is handles relationship between HTML text node and data units. We can reduce the data alignment problem by using better machine learning technique.

REFERENCES

[1] Y. Lu, Hai He, H. Zhao, Weiyi Meng. "Annotating Search Result from Web Database" IEEE transactions 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] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.

[4] A. Arasu and H. Garcia-Molina, Extracting Structured Data from Web Pages, Proc. SIG- MOD Intl Conf. Management of Data, 2003