International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Web Data Annotation System using Alignment Algorithm

Ms. P. P. Boraste[1], Mr. P.P. Shinde[2]

[1]*Assistant Professor, Dept. Of Computer Engineering, KBTCOE, Nashik, Maharastra, India*
[2]*Assistant Professor, Dept. Of Computer Engineering, KBTCOE, Nashik, Maharastra, India*

Abstract - *Now a day, there is a vast growth in the databases and the information technology. Rapidly grown-up databases are being accessed by means of HTML and web technology. During this process, the data units from the database are processed for various applications such as online shopping, deep web collection. Encoding may involve in this process and the data units before or after extraction. Meaningful label are assigning to the data units is also a research area under consideration. This paper focuses on the state of the art review of the methods used in the data annotation for the web databases with experiment results. Analysis of the various annotation methods are put forth. A novel technique for data annotation relating to knowledge database is proposed.*

***Key Words*: Data alignment, Data annotation, Search record results (SRR), Web database (WDB), Wrapper generation.**

1. INTRODUCTION

A huge amount of web databases are being used in various search engines. However, the search engines give out multiple records while accessing the web databases (WDB). The resulting record comes from the web databases show various data units. When we access the web databases, we receive the outcomes from the search engine. There are three phases for automatic annotation solution as mentioned by authors of [1] consists of - alignment phase, annotation phase, and annotation wrapper generation phase. The first phase of automatic annotation is alignment phase organizes all data units according to different groups where each group represents different ideas or concepts. The annotation phase groups the data to produce a meaningful label to every data units. The annotation rules are generated in annotation wrapper generation phase. The solution also uses six basic annotators; where each annotator can separately assign labels. To data units. Two main concepts primary used for annotation research are data units and text nodes. Data unit is a piece of text that defines one concept of real world entity.

A text node is different from data units. Their relationship between data units and text nodes are analyzed for assigning the label a meaningful name. Clustering algorithm can be exploited to classify the data units into different groups. This classification will help in categorizing the data with same concepts in the same group.

The rest of the paper is organized as follows. Section II reviews the proper data annotation process and different annotation phases. Section III presents the literature review of the algorithms related to labelling and annotation. Proposed algorithm explained in next section IV. Experimental results are presented in section V . Concluding remarks are given in section IV.

2. DATA ANNOTATION PROCESS

Annotation is the process that first aligns the data units on a result page into different groups in such a way that data in the same group have the same semantic. Then according to grouping, each group annotates it from different styles.

After the successful grouping, labeling is done to give meaningful names to each group of the data units. Next is the annotation wrapper generation phase which is automatically constructed for the search sites, this process is done after the successful labeling of the data units. This can be used to annotate new result pages from the same WDB.

1.1 Data Annotation Phases

Phase 1- Alignment phase

Alignment phase line up all the data into different groups where each group corresponds to a different concept. (e.g., all titles of books are grouped together).

Phase 2- Annotation phase

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.

Phase 3- Annotation wrapper generation phase

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

Three phases are used for data annotation are summarized in Table 2.1

Annotation phases		
1. Alignment Phase	2.Data Annotation Phase	3.Wrapper generation Phase
SRR(Search result records) identification of annotation rules	Some annotator with data units	Generation of features and rules
Organizing in groups by clustering based techniques	Each annotator produce a label for data units	Data retrieved from same web database

Table 2.1 Three annotation phases

1.2 Data Units and Text nodes features

Data unit's features 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. Features of data units are:

1) Data content: data unit and data node of the same concept usually shares certain keyword and have same leading label.

2) Presentation style: feature that 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 or classification of tags traversing from root to corresponding node in tree.

For example consider the first two records of the web databases as shown in Figure 2.1

EXTRACTING STRUCTURED DATA FROM WEB PAGES :

N.John and H. Jack */ cPGCON Int'l Conf. Engineering of Data/2003* our price **$10**, *put in the basket*

Automatic Annotation of Data Extracted from Large Web Sites:

L.lue, V. john G. Mecca*/ Workshop the Web and Database (Web DB)/ 2003 price **$15,** put in the basket*

Figure 2.1: The original HTML page

The corresponding source code for the Figure 2.1 is shown in Figure 2.2

```
FORM><A> Extracting Structured Data from Web Pages
</A><BR> N.John and H. Jack /<FONT><I>cPGCONInt'l
Conf. Engineering of Data / 2003 </I></FONT><BR>our
Price <B>$10 </B>
```

Figure 2.2: Resulting source code of the HTML page

If consider the first record shown in figure 2.1 "Extracting Structured Data from Web Pages" is the first text node. This text node is not always identical to data units. Where, N.John and H. Jack and other three fields like *cPGCON Int'l Conf. Engineering of Data/2003*, our price $10, put in the basket are the data units of that text node. The above data is being accessed from the web databases. A text node is the text outside the "<" and "> in source code." Text nodes are the visible elements on the webpage and data units are located in the text nodes. Then comparison with the other fields is performed by using different types of relationships (one to one, one to many or many to one) in between data units and text nodes.

3. LITERATURE SURVEY

In this section, we present a literature review of the approaches used in the existing literature. 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. However, the issue of assigning a label to the data units automatically has been discussed in [1].Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng authors has discussed three phases viz. Alignment phase, annotation phase and annotation wrapper generation phase.

In the alignment phase, the data unit was organized into various groups. The grouping of the same data units facilitates to recognize feature and patterns with the data units. In the annotation phase, a label is produced for each data unit in the group and a suitable label was assigned to each group. In the annotation wrapper generation phase, the annotation rule generated and apply the annotation rapidly.

This approach proposed that they maintain all the type of relationship between the text nodes and data units. The wrapper induction system was introduced in [4][5] 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 [2][3].

A similar approach is being introduced by D. Embley, et al. [3] based on ontology that extracts data from the web documents automatically. A method to align the data has been discussed by authors S. Mukherjee, et al. [7] which

maintains only one to one relationship between the text node and the data units. In [6], the authors introduced a domain dependent annotation process. However, this process manually assigns the label to the data. An ontology based system insightful to the data quality has been introduced in [10].

Automatically building a wrapper has been presented in [5][6]. These methods are used only for the data extraction, but not for annotation. The various methods discussed by the authors W. Liu, et al. [9] assigns the labels to the data from the web databases.

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.

4. PROPOSED APPROACH

In the existing system many search result records (SRR) comes from ViNTs for the WDB. Then detect text nodes, extract data units and text nodes features. After the successful extraction of data units make alignment for data units by using alignment algorithm. Finally results are shown in the form of different groups where data in the same group have the same semantic.

But in the proposed work system check data units in knowledge database after the extraction of data units .If data units present in the knowledge DB make alignment and show result in the form of groups otherwise add data units to the knowledge database.

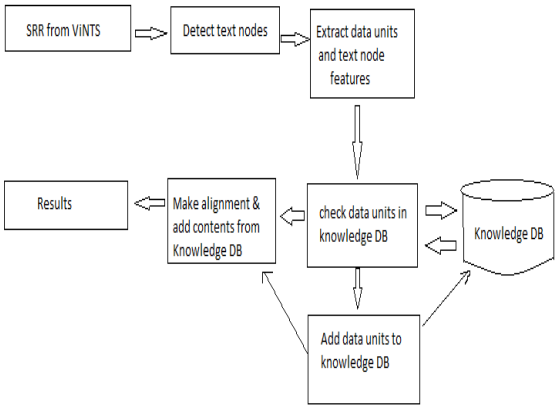

Figure 4.1. Proposed System Architecture

The algorithm for the data annotation in [1] can be summarized as:

-Data alignment algorithm is based on the assumption that attributes appear in the same order across all SRRs on the same result page; the SRRs may contain different sets of attributes. Thus SRR arranged in table format.

-Alignment algorithm is used for align the data units by using four steps. First; merge number of text nodes into single node. Then align the text nodes in different groups. Composite text nodes divide into single units. And lastly align the data units by separating composite groups into aligned groups and used clustering algorithm to clustering the text nodes.

Initially, a web database returns a multiple search record result (SRR). Each SRR contain several data units.

-Detect text nodes from the search record result.

-Data unit and text node features can be extracted out.

-Check data unit from the knowledge database.

-If the data units are remains present in the knowledge database then make alignment.

-Clustering based algorithm is used for clustering the text nodes. And alignment algorithm is to align the data units into different sets.

-As well add contain contents from the knowledge database

-Finally display the result.

-If the data unit does not present in the knowledge database then add data unit to the knowledge database.

4.1 Data Alignment Algorithm

```
ALIGN(SRRs)
1.  j ← 1;
2.  while true
      //create alignment groups
3.    for i ← 1 to number of SRRs
4.      G_j ← SRR[i][j];   //j^th element in SRR[i]
5.    if G_j is empty
6.      exit; //break the loop
7.    V ← CLUSTERING(G);
8.    if |V| > 1
        //collect all data units in groups following j
9.      S ← ∅;
10.     for x ← 1 to number of SRRs
11.       for y ← j+1 to SRR[i].length
12.         S ← SRR[x][y];
        //find cluster c least similar to following groups
13.     V[c] = min_{k=1 to |V|} (sim(V[k], S));
        //shifting
14.     for k ← 1 to |V| and k ≠ c
15.       foreach SRR[x][j] in V[k]
16.         insert NIL at position j in SRR[x];
17.   j ←j+1;        //move to next group
```

Figure 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 or semantics.

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.

4.2 Clustering Algorithm

Figure 4.3 shows clustering algorithm which clusters the text node in a way that same group containing the same element of same concept only. Basically it is a clustering-shift algorithm which handles the one to nothing relationship between text nodes and data units.

```
CLUSTERING(G)
1.  V ←all data units in G;
2.  while |V| > 1
3.    best ← 0;
4.    L ←NIL; R ←NIL;
5.    foreach A in V
6.      foreach B in V
7.        if ((A != B) and (sim(A, B) > best))
8.          best ← sim(A,B);
9.          L ←A;
10.         R ←B;
11.   If best > T
12.     remove L from V;
13.     remove R from V;
14.     add L ∪ R to V;
15.   else break loop;
16. return V;
```

Figure 4.3 Clustering Algorithm

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.

5. RESULTS

This section represents the results of the proposed approach for the data annotation and annotation features in the web databases. Three data units and text nodes features are derived from five features such as data types (DT), data contents (C), presentation styles (PS), and adjacency (AD).

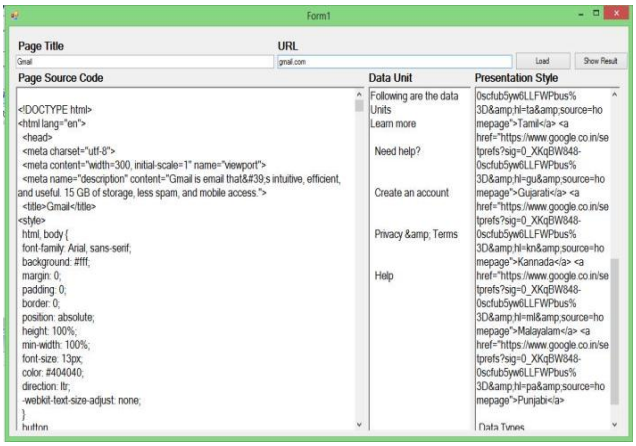

Figure 5.1 output of the user query

Figure 5.1 shows the snapshot of the results. The screen is broken in three parts: - page source code that contains the code about user query, data units and presentation style this two are features of data unit and text nodes, which show how data unit is displayed on a web page.

Features the project:

1. When the user enters any URL and query to the webpage, click on the load button on this web page.
2. The record related to query is searched using different presentation styles like font face, font size and colors etc.
3. According to presentation style the other features like data types, data contents and tag path are derived. Detect the text nodes from SRR which contains data units. Then checking for the data units from knowledge database and extract the features of data units and text nodes.
4. A comparison of data units with knowledge databases is done. In knowledge database all data arranged according to user mostly search. Align that data and then labeling with the suitable names.
5. In previous work no specific database is use. Means, SRR used as a database is change accordingly. But in proposed system, it will store result data into a database in terms of data units and the database named as knowledge database by using it can

provide more generalize form of data.

6. Also in previous work used clustering base shifting algorithm. This has limited applications. So, in proposed approach we would like to use k-means algorithm which gives better accuracy and results that previous one.

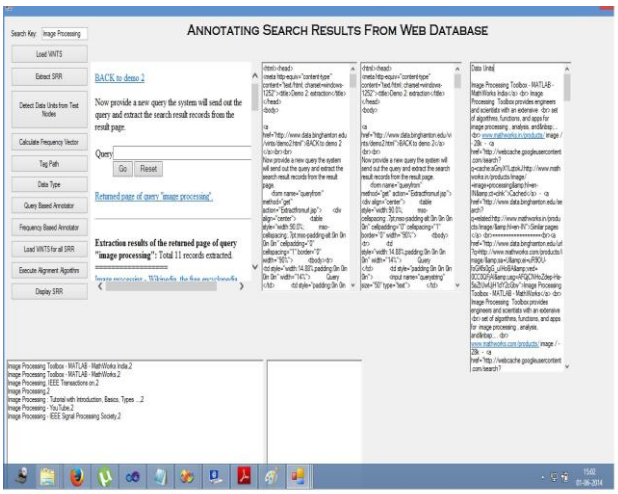

Figure 5.2 Annotating Search Results from Web Databases

Above snapshot shows the proper annotation process for the search web data. By using alignment algorithm annotate the data units and display result in the form of search result record (SRR) groups. Numbers of steps are performed like load ViNTs, extract SRR and detect data unit, text node features, etc.

6. CONCLUSION

This paper elaborates various approaches used in the data annotation search problem in the web databases. Since the prime research issue in the web database is data annotation and data alignment.

A new technique for data annotation in the web database was implemented with the expected results.

REFERENCES

[1] Y. Lu, H. He, H. Zhao, W. Meng, C Yu, "Annotating Search Results from Web databases" In IEEE Transaction on Knowledge and Data Engineering M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] A. Arasu and H. Garcia-Molina,Extracting Structured Data from Web 5 Pages," Proc. SIGMOD Int'l Conf.

[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] H. Zhao, W. Meng, and C. Yu "MiningTemplates formSearchResult Records of Search Engines," Proc. ACM SIGKDD Int'l ConfKnowledge Discovery and Data Mining, 2007 an J and Kamber M. Data mining:concepts and techniques, Morgan Kaufmann

[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] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997

[7] S. Mukherjee, I.V. Ramakrishna, and A. Singh, "Bootstrapping Semantic Annotation Jain A, Murty M and Flynn P. Data clustering: A review ACM Computing Surveys, 31(3), pp. 264–323, 1999

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

[9] W. Liu, X. Meng, and W. Meng, "ViDE:d Approach for Deep WebData ExtractidKnowledge and Data Eng., vol. 22, no. 3, p447-460, Mar. 2010.

[10] W. Su, J. Wang, and F.H Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. DB

[11] Z. Wu et al.,"Towards Automatic Incorporation Engines into a Large-Scale Metasearch Engine," Proc. IEEE/WIC Int'l C Websystem