International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

Annotating Search Results from Web Databases using Alignment algorithm & different Annotators.

Prof.J.Y.Kapadnis[1], Akshay.R.Gore[2], Hitesh.N.Sonawane[3],Gokul.S.Itnare[4],Ravindra.K.Balak[5]
[1]*Assistant Professor, Computer Department, P.V.G's COE Nasik, Maharashtra, India*
[2345]*Student, Computer Department, Computer Department, P.V.G's COE Nasik, Maharashtra, India*

Abstract - *The number of web databases have become web assessible through search interfaces which is in HTML form. The data units which are returned from the underlying databases are encoded to the result pages dynamically for human browsing .The encoded data units which are to be machine process able is essential for many applications that is deep web data collection & Internet comparison shopping, they need to be extracted out & assigned meaningful labels. This paper presents an automatic annotating approach which first aligns data units on a result page into different groups such as data in the same group resides same semantic. Then, for this each group we annotate it from various different aspects& aggregate the different annotations to predict a final annotation label for it. The annotation wrapper for the search site is automatically constructed & can be used to annotate new result pages from the same web databases. Our experiment is highly effective of our proposed system.*

Key Words: **Data alignment, data annotation, web database, wrapper generation**.

1. INTRODUCTION

The use of Internet has been increased widely over a period of time. Also, the use of E-Commerce has increased rapidly since a decade. The Web Databases are accessed through HTML based search engine. The result returned from web database is in the form of Search Result Record (SRR). SRR contains text nodes and data units. Here, we perform data unit level annotation. There is a high demand of data of interest from multiple Web Databases (WDBs). Now-a-days databases become web accessible, these databases are having data units encoded. To separate data units and assigns meaningful labels there is an automatic annotation approach which automatically assigns labels to data units within SRRs returned from web databases.

1.1 Problem Statement

The data on web servers are accessible through HTML level based search engines. The information mined from these web servers are mostly in unstructured format called Search Result Records (SRR's). Whenever a web user wants to search for a specific product he/she reviews or goes through many sites to obtain a relevant result. In order to avoid such inconvenience, we are presenting Annotating Search results from web databases using Alignment, Clustering algorithm and applying different annotators.

1.2 Literature Survey

Numbers of Web Databases has reached 25 million according to a recent survey.The data on web servers are accessible through HTML level based search engines. All the Web Databases make up the deep web (hidden web or invisible web).Often the retrieved information is enwrapped in Web pages in the form of data records. These special Web pages are generated dynamically and are hard to index by traditionally crawler based search engine, such as Google or Yahoo. Each data record on the deep Web pages corresponds to an object. Extracting structured data from deep Web pages is challenging problem due to underlying intricate structure of such pages. Because of the supervised training and learning process, these systems can usually achieve high extraction accuracy. But due to lack of scalability they are not suitable for application which needs to extract large data from various web sources.

The results generated from web a database is in raw data and not in structured format .Annotation allow the result the result to be displayed in structured format. The most existing approach simply assigns labels to each HTML text node, we thoroughly analyses the relationship between text nodes and data units. Each SRR extracted by ViNTs has a tag structure that determines how the contents of the SRRs are displayed on a web browser. A data unit is a piece of text that semantically represents one concept of an entity. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags. A tag node corresponds to an HTML tags surrounded by "<" and ">" in HTML source, while a text node is the text outside the "<" and ">." Text nodes are the visible elements on the webpage and data units are located in the text nodes. Instead of using HTML tag tree structure of the SRR's to align data unit's of our approach also consider other important featured shared among data units, such as their Data Units (DU), Data Contents (DC), Presentation Style(PS), Adjacency(AD)Information .We construct an annotation wrapper for any given web databases. The wrapper can

applied to any efficiently annotating the SRR's retrieved from the same WDB with new queries.

A) Data Content
Data units or text node with same concept often share certain keyword's. This is true for two reasons first, the data unit corresponding to search fields were the user enter the search condition usually contain the search keywords.

B) Presentation Style
This featured described how data unit display on web pages. It consist of six style features, font face, font size, font color, font weight, Text Declaration(Underline strike etc.) and whether it is italic.

C) Data Types
Each data unit has its own semantic type although it is just text string in HTML code. The following basic data types currently consider in our approach. Date, Time, Currency, Integer, Symbol, Decimal, Percentage, String.

D) Adjacency Information
The preceding and succeeding information from different search result record is called Adjacency Information.

The proposed system thoroughly analyses the relationship between text node and data units performing data unit level annotation, which are then aligned and clustered using alignment and clustering algorithm to align the data unit into different group. So that the data unit inside the same group have semantic. Specific annotators such as table annotator, query-based annotator, in text prefix/suffix annotator, schema based and common knowledge annotator are applied to the aligned results in order to provide meaning to the web result.

1. Table Annotator (TA):-
Many WDBs use a table to organize the returned SRRs. In the table, each row represents an SRR. The table header, which indicates the meaning of each column, is usually located at the top of the table. Usually, the data units of the same concepts are well aligned with its corresponding column header. This special feature of the table layout can be utilized to annotate the SRRs.

2. Query-Based Annotator (QA):-
The basic idea of this annotator is that the returned SRR's from a WDB are always related to the specified query. Specifically, the query terms entered in the search attributes on the local search interface of the WDB will most likely appear in some retrieved SRRs.

3. Schema Value Annotator (SA):-
Many attributes on a search interface have predefined values on the interface. For example, the attribute publishers may have a set of predefined values (i.e., publishers) in its selection list. Our schema value annotator utilizes the combined value set to perform annotation.

4. Frequency-Based Annotator (FA):-
The data units with the higher frequency are likely to be attribute names, as part of the template program for generating records, while the data units with the lower frequency most probably come from databases as embedded values.

5. In-Text Prefix/Suffix Annotator (IA):-
In some cases, a piece of data is encoded with its label to form a single unit without any obvious separator between the label and the value, but it contains both the label and the value. Such nodes may occur in all or multiple SRRs.

6. Common Knowledge Annotator (CA):-
Some data units on the result page are self-explanatory because of the common knowledge shared by human beings. For example, in stock and out of stock occur in many SRRs from e-commerce sites. Human users understand that it is about the availability of the product. Because this is common knowledge. So our common knowledge annotator tries to exploit this situation by using some predefined common concepts.
-

2. System Design

2.1 System Architecture

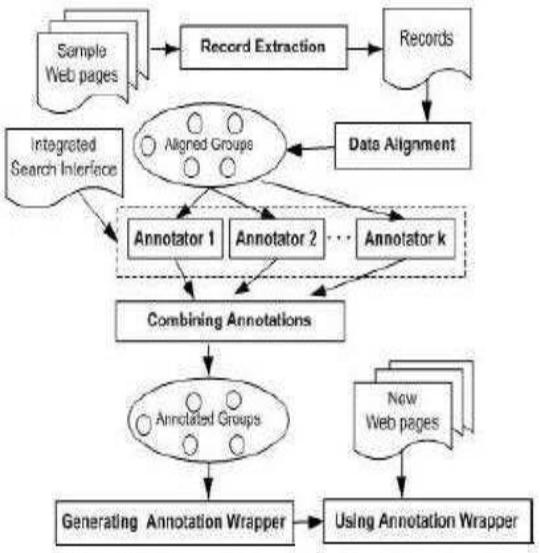

Fig- 2.1 System Architecture

2.2 Block Diagram

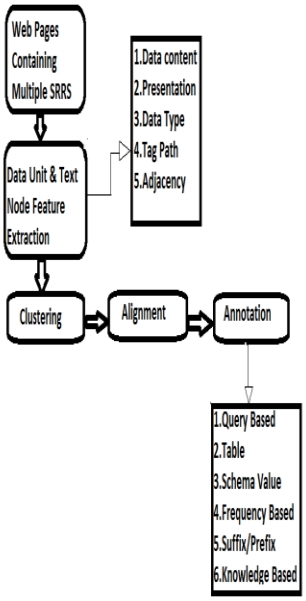

Fig- 2.2 Block Diagram

3. Proposed Algorithm

3.1 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 ← Ø;
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
```

Fig-3.1 Alignment Algorithm

3.2 Clustering Algorithm

```
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;
```

Fig 3.2 Clustering Algorithm

The Data Alignment Algorithm as shown in fig 3.1 is based on the assumption that attributes occurs in the same order across all Search Result Records, although the Search Result record may contain different sets of attributes. Data Alignment is an important step in achieving Data Annotation. Basically, the process of identifying all the data units in the Search Result Records and then organizing them into different groups with each group corresponding to different concepts is called Data Alignment. The Data Alignment step consists of four steps as follows

Step1. Merge Text Nodes: Detecting and removing the decorative tags from SRRs to allow the text nodes corresponding to the same attributes to be merged into single text nodes are done in this step.
Step2. Align Text Nodes: Aligning text nodes into groups so that so that each group contains same text nodes havinf same concept is done in this step.
Step3.Spilt Text Nodes: Splitting the values in composite text nodes into individual data units is the aim of this step.
Step4. Align Data Units: Separating each composite group into multiple aligned groups with each group containing the same concept of Data Units.

The Fig 3.2 shows Clustering Algorithm which clusters the text nodes in the same group containing the elements of same concept only. Basically, Data Clustering is the process of collecting the text nodes with same specification and each cluster contains the elements of same concept only. The Clustering Algorithm is basically a Clustering- Shift Algorithm which handles the one-to nothing relationship between text nodes and data units than the previous Clustering Algorithm.

In terms of precision and recall, We conducted experiments to evaluate the precision and recall. We evaluate the significance of each feature on the performance of our alignment algorithm. So, we compare the performance when a feature (i.e. annotation) is used with that when it is not used. Each time one annotators is not used and compared the performance (i.e. precision and recall) when it is used. The below bar graph shows the higher percentage of precision and recall when we used the annotation approach (i.e. using different annotators).

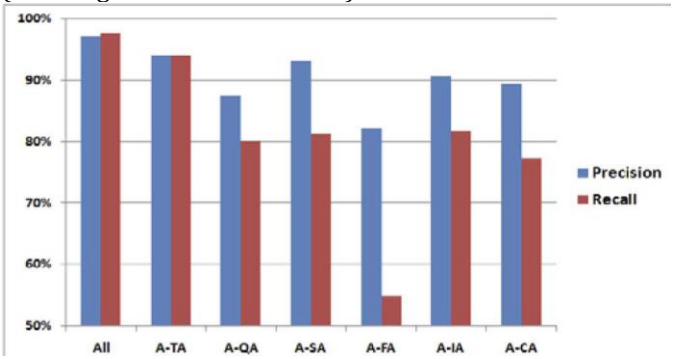

Fig: Precision and Recall

Hence, here we studied the automatic annotation of Search Result Records (SRRs) with the help of Alignment Algorithm and Clustering Algorithm and by using different annotators. Here, our method is a clustering based shifting method utilizing richer yet automatically obtainable annotation. The precision and recall of our approach is approx 98% which can be improved further years.

3. CONCLUSIONS

Annotating or analyzing large data in a single website may lower the processing speed. Hence, we encountered Data Annotation problem and we proposed a multi annotators approach to automatically construct an annotation wrapper for annotating the SRRs retrieved from the Web Databases. Accurate alignment is critical to achieving holistic and accurate annotation. So here we used clustering based shifting method utilizing richer yet automatically obtainable features which is made capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing. Hence, by using clustering based shifting algorithm for each group we annotate it from different aspects and aggregate the different annotations to predict final Annotation label for it.

ACKNOWLEDGEMENT

With deep sense of gratitude we would like to thanks all the people who have lit us path with their kind guidance. Firstly, we would like to thank project coordinator Prof.J.Y.Kapadnis and our Head of the Department Prof. M. T. Jagtap of Pune Vidyarthi Griha's College Of Engg,Nasik for being an invaluable source of inspiration & ideas and. We owe our all success to them, wo indeed directly or indirectly helped us on this topic.

REFERENCES

[1] Annotating Search Results from Web Databases, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013 .

[2] S. Handschuh, S. Staab, and R. Volz, On Deep Annotation, Proc. 12th Intl Conf. World Wide Web (WWW), 2003.

[3] J. Wang and F.H. Lochovsky, Data Extraction and Label Assignment for Web Databases, Proc. 12th Intl Conf. World Wide Web (WWW), 2003.

[4] Z. Wu et al., "Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine," Proc. IEEE/WIC Int'l Conf.Web Intelligence (WI '03), 2003.

BIOGRAPHIES

Gore Akshay Rajendra
Student of B.E Computer of
P.V.G's COE Nasik
Maharashtra, India.

Sonawane Hitesh Nivruti
Student of B.E Computer of
P.V.G's COE Nasik
Maharashtra, India.

Itnare Gokul Sunil
Student of B.E Computer of
P.V.G's COE Nasik
Maharashtra, India.

Balak Ravindra Kiran
Student of B.E Computer of
P.V.G's COE Nasik
Maharashtra, India.