# ANNOTATING USER DEFINED QUERIES

[1]Seema Bhaganagre, [2]T D Vineeth, [3]P Madhuri, [4]B Malathi
[1]Student, [2]Student, [3]Student, [4]Associate Professor
[1]Department of Information Technology,
[1]Aurora's Technological and Research Institute, JNTUH, Hyderabad, Telangana, India

***Abstract:*** Data returned by a user query is collected from a structured database termed Web Database. Data is stored in these databases in the form of data units in various Search Result Records, wherein they are labeled semantically for processing. Earlier applications, being manual, were difficult to annotate. The advance application can perform annotation independently. With the help of six series of various annotators, each annotating on the different feature of the data units, all the data units are then grouped and labeled accordingly. And with an add-on in the third annotator (i.e.); thesaurus which also helps in looking for the similarity in the data units. The annotators estimate the final label. And finally, the wrapper generation generates annotation rule (r), which assists to annotate new queries effortlessly. This indeed helps to increase the precision and recall.

***Index Terms* - Data Annotation, Web databases**

## I. INTRODUCTION

The deep web is a database for many search engines, i.e.; data returned from the resultant pages come from an underlying structured database. These search engines are generally known as Web databases (WDB). Every resultant page from a WDB has many search result records (SRRs). Each SRR contains multiple data units, each of which describes one facet of the real world entity. A Data unit is a piece of text that semantically represents one concept of a real-world entity, it also constitutes to the value of a record under an attribute. Whereas a text node is a sequence of text surrounded by a pair of HTML tags. Annotation of these data units is the most important part of the data extraction during data analysis and data mining process, in which data is collected from multiple web databases. Annotation means assigning correct semantic labels to each data unit.

For example, if we consider an information about a web page that shows result for a mobile phone, then we get to see that, this web page has multiple data units like the name of the phone, model number, company, manufacturer, software version, hardware specification, add-on features, price and so on. Thus, the system should be able to recognize the semantics of each data unit. Here arises a problem, where most of the resultant web pages have data units which lack semantic labels. Having correct semantic not only helps in linking the SRRs but also helps in storing them efficiently in the database. Due to the lack or no proper labeling to the data units, the user fails to recognize the data units easily. Hence, we propose to assign meaningful labels to the data units in the SRR returned by WDB and automatically annotate them. With every website containing a web service interface, it is much easier to annotate the data unit in the SRR as the description of the semantics of each data unit given by the Web Service Description Language (WSDL). Here, we try to perform annotation at the data unit level.

Sections [II] talks about the literature survey, which includes one base paper and two reference papers. Section [III] is the system architecture describes the overall working of the system. Section [IV] describes various approaches using six different annotators. Section [V], performance analysis, talks about precision and recall. Section [VI], Conclusion.

## II. LITERATURE SURVEY

Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Clement Yu [1], suggested various annotator approach to building an annotation wrapper to annotate the SRR fetched from a web database. Employees a series of six annotators and a probabilistic model to the connect them all. Each annotator utilizes a particular to annotate. A significant attribute while annotating is it uses both the LIS and IIS of various other web databases within the domain. Employees clustering based shifting method. Increased Precision and Recall.

Poonam V. Wankhede, Sachin N. Deshmukh [2], employed multiple annotators to generate an annotation wrapper from any web database. Ever annotator uses a specific feature over data units to annotate and label them.

Bincy S Kalloor, Sheeja Agustain [3], as the structure of the web page is complex, data retrieval becomes a difficult task. Preliminary annotations were executed manually, were time-consuming and performed low. Automatic annotation was introduced to subjugate this problem. Ultimate increasing the speed of data sets and data processing is minimized.

## III. SYSTEM ARCHITECTURE

System architecture gives an overall view of the system. When the user enters a search query, the search interface returns many multiple SRR's, based on the retrieved results, the system architecture can be divided into three phases. Where *Phase 1* is the 'Alignment Phase'. In this phase, all the data units in the SRRs are recognized and which are sorted into various groups with each group representing a different concept. Merging data units with the same semantic label into a single group helps to identify the frequently occurring patterns and features among these data units.

In *Phase 2*, which is also called as the ‘Annotation Phase', has various multiple basic annotators, which annotate the data units into groups by each annotator focusing on one feature of the data unit and thereby grouping them and labeling. A probability model is applied to each group which helps in assigning the most relevant labels to each group.

In the last phase i.e.; the *Phase 3*, also known as the ‘Annotation Wrapper Generation Phase'. In this phase, various annotation rules are generated which help in the withdrawal of data units of a concept in the result page and applying a relevant semantic label, which in response to new queries doesn't go through the first two phases but directly annotates the data received.

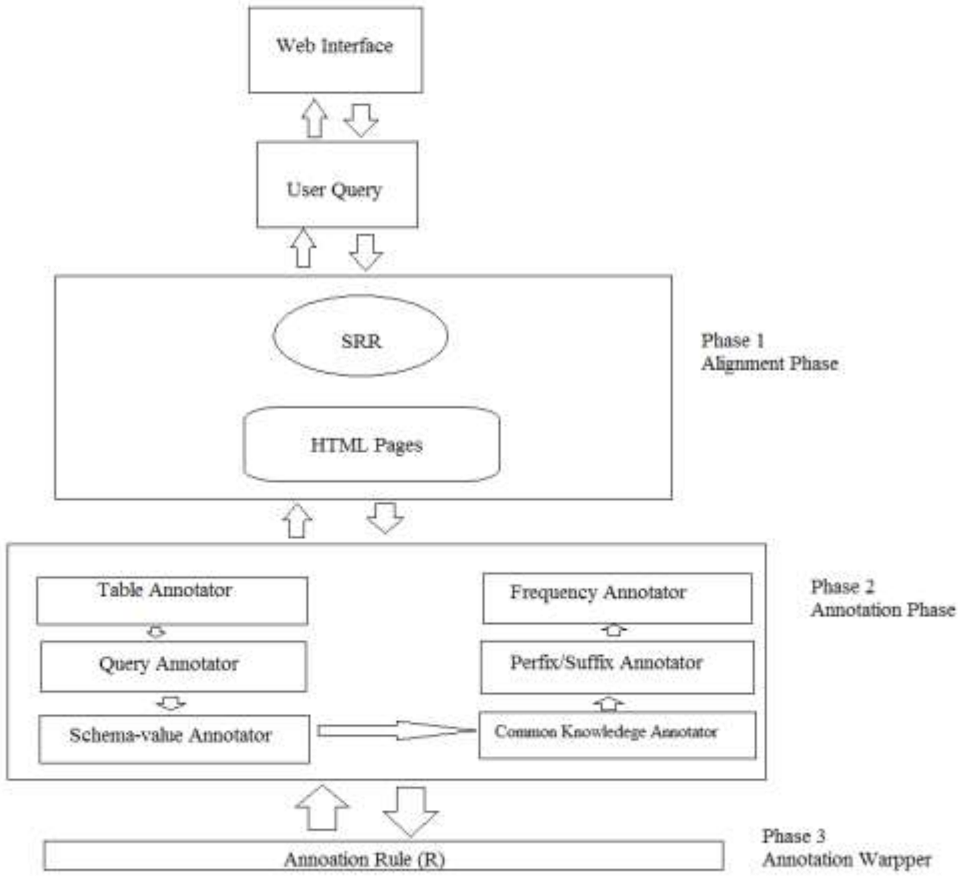

Fig: System Architecture

## IV. APPROACH

Every annotator performs annotation over five different features namely, Data Content, Presentation style, Data type, Tag path and Adjacency. Then the six annotators work in the following manner, wherein they work in a serial manner.

### Table Annotator

These are the first annotator out of the six annotators which we are been used. It is used to arrange the data units into a table format. Each row represents an SRR (Search Result Record), table header gives the meaning of the column, as the alignment of each column header table, a layout is used to annotate SRR. The physical position information of data units can be obtained during extraction of the SRR's. For every SRR, it selects a single data unit in a particular cell and even selects the column header. Then finally the selected units are assigned as column and labeled, and then the remaining data is processed similarly.

### Query Based Annotator

The second annotator is the query based annotator where the SRR is returned depending on the query issues arises after a search. Terms entered in the query are mostly present in the retrieved SRR's. Here we can even use search interface which can annotate the title value of SRR. Then it issues a query against an attribute and finds the largest total occurrence query term. As Local Interface Schema (LIS) does not have the entire attribute underlying in the database because they are not needed for specifying the query condition. Hence query annotator cannot complete annotating the SRR by itself.

### Schema Value Annotator

The third is the Schema Value Annotator. Attributes on search interface might have some predefined values on schema interface. For e.g.: Attributes data may have a set of predefined values. IIS (Integrated Interface Schema) may contain attributes with predefined values and these attributes have some similar values like those in LIS because as the attributes from various interfaces are combined together even they values are combined and checked over with the thesaurus for their similarity.

The schema value annotator looks for the best match attribute to the group of IIS. Normally the schema annotator adds the similarities and multiplies the total sum by the number of non-zero similarities present. Then the final result is considered as the matching score. This is retrieved by effective combination system present in IR tool which could be used to annotate web.

**Frequency Based Annotator**

This is the fourth annotator where the Data units contain both high frequency and low frequency that his data units of high frequency are attributes name and data units of low frequency are attribute value. Here the data is grouped into one group which might have high frequency and low frequency. By this, the calculation gets complete for cosine similarities for both attributes and data unit. As a result attribute name is assigned as a group label.

**Prefix/Suffix Annotator**

Fifth is the prefix and suffix annotator. A data unit contains labels, values, and separator. Few node might even occur in multiple SSR's. Once data alignment is complete all the similar data units are grouped. This annotator mostly useful for checking the data unit alignment which mostly contains same suffix and prefix. If it consists of the same prefix then it is removed from the data units present in the group and is used to label and annotate values. Where if same suffix and number of a data unit having the same suffix are equal then the number of data units present inside next group, therefore suffix is used to annotate the data unit inside the next group.

**Common Knowledge Annotator**

Sixth is common knowledge annotator. In this, we have applied to self-explanatory data units because of common knowledge shared by human beings.

E.g.: "Packing" and "Delivered" these are same predefined common concept method. Each and every common concept has a label, set of patterns or a value.

E.g.: email.id. If a group of data units from the alignment step match the patterns or value and labels which is assigned to the data unit of this group.

**Combining Annotators**

Here all the six annotators are combined together. All the above annotator are independent of each other. Based upon all these features, a simple probabilistic method it is been applied to combine different annotator with each other.

**Annotation Wrapper**

Once the complete data unit on the result page is annotated, these can be used to construct annotator wrapper. Where this can help the new SRR that is retrieved from same WDB annotator the data uses wrapping quickly without even reapplying the entire annotation process once again. The annotation wrapper itself is a description of annotation rule. Here set us to consider a group where each SRR contain its own tag node i.e., HTML tags, Testing tags etc. Then both backward and forward obtain prefix and suffix of the particular data unit. This method can be stopped at a point where a unit found is valid data unit which as a meaningful label. The prefix of all the data units is compared to obtain a common prefix for the particular data units as the similar even common suffix is found.

## V. RESULTS AND DISCUSSION

The accuracy of the results retrieved for user-defined query retrieved over the internet is calculated on the bases of how precise the result is and how easily can the query be recalled on the query. The precision and recall are used for the measure of relevance.

Precision is the ratio of the total number of relevant retrieved to the total number of relevant.

A recall is the ratio of the total number of the relevant retrieved to the total number of retrieved.

Precision = Relevant Retrieved / Relevant

Recall= Relevant Retrieved/ Retrieved

Where the precision and recall for every individual annotator is to be determined, so as to draw their accuracy.

## VI. ACKNOWLEDGMENT

We would like to express our gratitude to **Mr. J. SRIKANTH, Director**, **Aurora's Technological and Research Institute** for providing us congenial atmosphere and encouragement. We express our sincere thanks to Head of the Department **Ms. A. Durga Pavani**, for giving us the support and her kind attention and valuable guidance to us throughout this course. We convey thanks to our

Project Guide **Ms. B Malathi** of Department of CSE & IT for providing encouragement, constant support and guidance which was of a great help to complete this work successfully.

## REFERENCES

1. Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu, Senior Member - “Data alignment and Wrapper generation”. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013
2. Poonam V. Wankhede and Sachin N. Deshmukh - “Data annotation for web database” International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)” Volume 4 Issue 6, June 2015
3. Bincy S Kalloor1, Sheeja Agustin. - "Web databases and mining the application". International Journal of Innovative Research in Computer and Communication Engineering. Vol. 3, Issue 3, March 2015.