International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34078-34083

Wrapper Annotation For Web Database Search Results

1.N.Sowndhariya

Information Technology
Saranathan College of Engineering ,Venkateswara Nagar, Panjappur,Trichy,Tamil Nadu
Email Id: sowndhariyanandu.in@gmail.com

2.T.C.R.Jeyarathika

Computer Science
Shivani Engineering College, Poolangulathupatti, Tamil Nadu
Email Id:jeya.rathika@gmail.com

3.C.Valarmathi

Computer Science
Shivani Engineering College, Poolangulathupatti, Tamil Nadu
Email Id:1993valar@gmail.com

4.P.Saranya

Computer Science
Shivani Engineering College, Poolangulathupatti, Tamil Nadu
Email Id:cutepic1993@gmail.com

ABSTRACT
Web Databases are accessing through HTML form based search interfaces. Result page from web Database has multiple records. Each of these Search Result Records (SRR) contains multiple data units which are usually encoded into the result page dynamically for human browsing. To use Encoded data units for machine processable such as deep web data collection and internet comparison shopping, they need to extract and assigned label semantically.
We present an automatic annotation approach that contain three phase to annotate the data units. First, identify data units in SRR and organize into different groups such that the data in the same group have the same semantic.
Then for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it.
An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. This approach is highly effective.

1. INTRODUCTION

1.1 OBJECTIVE

Our goal is to automatically annotating label to the extract data from web database by grouping data in the same group having same semantic. Web page from the Web site the content rests upon. Data Mining refers to extracting or mining knowledge from large amount of data. Knowing discovery in databases consists of an interactive sequence of Data cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation, and Knowledge Presentation.

Data Mining is usually defined as searching, analyzing and sifting through large amounts of data to find relationships, patterns, or any significant statistical correlations.
The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Web structure mining, one of three categories of web mining for data, is a tool used to identify the relationship between Web pages linked by information or direct link connection. This connection allows a search engine to pull data relating to a search query directly to the linking.
A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search results record. The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Search result record from web database contains multiple units; they are not used for application such as deep web data collection and internet comparison shopping. They need semantic labels for further process.
A large portion of the deep Web is database-based, i.e., data encoded in the returned result pages of many search engines come from the underlying structured databases (e.g., relational databases). Such type of search engines will be referred to as Web databases. A typical result page of a Web database consists of multiple search result records (SRRs) and each SRR corresponds to an entity. Usually, each SRR consists of multiple data units (or instances) like book title, author, publisher, price, etc. Frequently, not all data units are

encoded with meaningful labels. So, we address how to automatically annotate the data units in the SRRs returned by Web databases. And annotate data units as assigning meaningful labels to them.

The annotation problem has become a very significant problem due to the rapid growth of the deep Web and the need to query multiple Web analysis/mining, it is imperative that the data units are correctly labeled so they can be appropriately organized and stored for subsequent machine processing. Note that for search sites that have Web services interfaces, it may be easier to annotate their SRRs because the semantic meanings of their data units are more clearly described in WSDL. However, that very few search sites have Web services interfaces. Therefore, it is still necessary to extract and annotate data from the legacy HTML pages. In this system, we propose a holistic and multi-annotator approach to automatically constructing an annotation wrapper for any given Web database. Given a set of sample result pages from a Web database, we first extract the SRRs from these pages. Then the data units in all SRRs are aligned such that all data units in each aligned group semantically belong to the same attribute/concept. We then design different basic annotators to annotate data units in each aligned group holistically. The results of different basic annotators are combined to determine an appropriate label for each group of data units. Finally, with the annotated data units, an annotation wrapper is constructed for the Web database which can be used to annotate new SRRs retrieved from the Web database in response to new queries.

1.2 EXISTING SYSTEM

- Extracted result set from web database and annotate labels after organizing them into different group corresponding to different concept.
- Data units are organize by group to align data based on their concepts. Uses HTML tags to align data units by filling them into a table through a regular expression based on data tree algorithm.
- Simply labels are assigned to attributes of Tables (i.e.) columns of tables
- For multiple WDBs, Local Interface Schema (LIS) is used for annotating

1.2.1 DEMERITS

- Simply assigns label to each HTML text node
- Local Interface Schema raises the inadequacy problem
- Inconsistent label problem because different labels are assigned to semantically identical data units returned from different web database

1.3 PROPOSED SYSTEM

We automatically assign labels to the data units extracted from Web Databases (WDBs). Given a set of search result record that have been extracted from a result page returned from a WDB, our automatic annotation solution consists of three phases.

Alignment Phase

- Identify all data units and organize into group with different concept and analyze the relationship between text nodes and data units.
- Relationship is analyzed, by identifying the features such as their data types (DT), data contents (DC), presentation styles (PS), and adjacency (AD) information.
- A clustering-based shifting technique used to align data units into different groups so that the data units inside the same group have the same semantic.
- We utilize the integrated interface schema (IIS) over multiple WDBs in the same domain to enhance data unit annotation.

Annotation Phase

- We use six basic annotators; each annotator can independently assign labels to data units based on certain features of the data units.
- We also employ a probabilistic model to combine the results from different annotators into a single label.
- It is highly flexible so that the existing basic annotators may be modified and new annotators may be added easily without affecting the operation of other annotators.

Annotation wrapper generation Phase

- We construct an annotation wrapper for any given WDB.
- The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB with new queries.

1.3.1 ADVANTAGES

- Data unit level annotation is performed.
- Uses alignment algorithm
- Integrated interface schema has potential to increases the annotation recall.
- Global attribute name given for label
- Quickly annotate same WDB without reapplying entire annotation process by using annotation wrapper.

2. DESIGN

2.1 SYSTEM ARCHITECTURE

System design is the process of defining the architecture, components, modules, and data for a system to satisfy specified requirements. One could see it as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering. Our project has been done for book database.

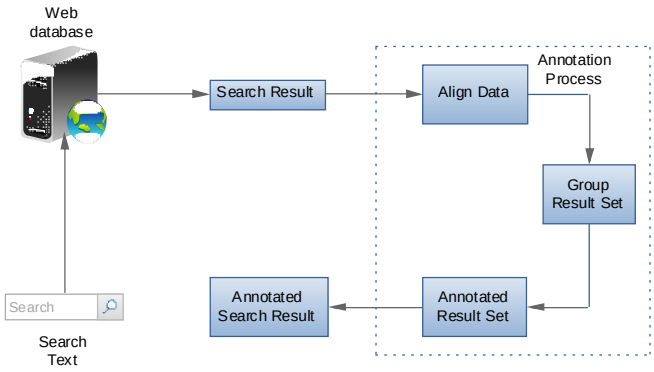


Fig 2.1 System Architecture

This architecture describes the process of annotation. The user once logged in search for the book using the search interface. The searched book is then retrieved from the database and parsed into data units in the extraction process using HTML agility package.
Then the data units are aligned using the similarity function which is based on the data content, presentation style, tag path, data type and adjacency similarity.
These aligned data units are clustered into groups such that each cluster is of similar concepts. By analyzing the features of each cluster, they are assigned with a unique name i.e., annotated. The annotated search results are given to the user.
These processes are done to generate a wrapper. Wrapper is created to retrieve the same data easily without performing alignment and grouping again.

2.2 FLOW DIAGRAM

A data flow diagram is a network that describes flow of data and process that change or transform data through the system.
Our optimizations take into account the easy retrieval of the book from the database or website for the user.
We create a label i.e., annotation of each groups which makes the search and retrieval efficient.

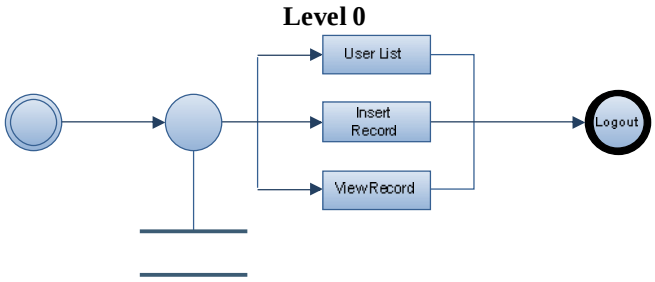


Fig 2.2 Data flow of Admin login

In Level 0 the Admin login to web database and views the user list. He can insert a new book, view the books present in the database and can modify any details if necessary.

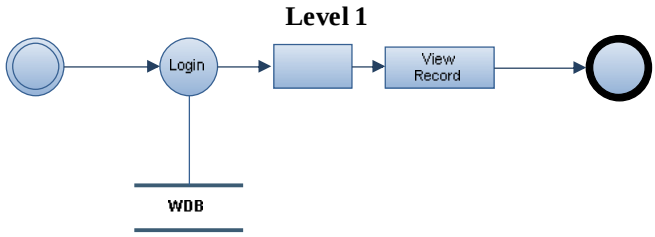


Fig 2.3 Data flow of User login

In Level 1, the registered user can login and view his profile. He can search for the book using the search interface and view the results. If the book in not present in the database, then it would be indicated to the user. For an unregistered user, first he has to register by providing the asked details.

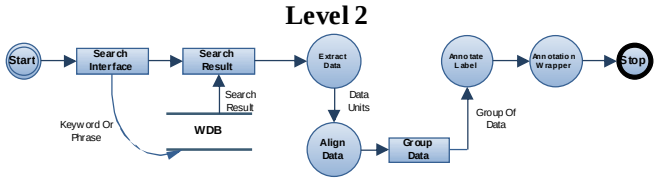


Fig 2.4 Data flow of Annotation process

The search results are extracted into data units. Then they are aligned and grouped using similarity function. The grouped data units are now annotated by analyzing the features. This process is all together known as Wrapper generation.

3. IMPLEMENTATION

3.1 INTRODUCTION

A large portion of the deep web is database based, for many search engines, data encoded in the returned result pages come from the underlying structured databases. Such type of search engines is often referred as Web databases (WDB). A typical result page returned from a WDB has multiple search result records (SRRs). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. Each SRR in our project represents one book with several data units. In this project, a data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags. In this project, we perform data unit level annotation. There is a high demand for collecting data of interest from multiple WDBs. For example, once a book comparison shopping system collects multiple result records from different book sites, it needs to determine whether any two SRRs refer to the same book. The ISBNs can be compared to achieve this. If ISBNs are not available, their titles and authors could be compared. The system also needs to list the prices offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages. For instance, no semantic labels for the values of title, author, publisher, etc., are given. Having semantic labels for

data units is not only important for the above record linkage task, but also for storing collected SRRs into a database table (e.g., Deep web crawlers) for later analysis. Early applications require tremendous human efforts to annotate data units manually, which severely limit their scalability. In this project, we automatically assign labels to the data units within the SRRs returned from WDBs.

Our automatic annotation solution consists of three phases. Phase 1 is the alignment phase. In this phase, we first identify all data units in the SRRs and then organize them into different groups with each group corresponding to a different concept (e.g., all titles are grouped together). The result of this phase with each column containing data units of the same concept across all SRRs. Grouping data units of the same semantic can help identify the common patterns and features among these data units. These common features are the basis of our annotators. In Phase 2 (the annotation phase), we introduce multiple basic annotators with each exploiting one type of features. Every basic annotator is used to produce a label for the units within their group holistically, and a probability model is adopted to determine the most appropriate label for each group. At the end of this phase, a semantic label is assigned to each column. In annotation wrapper phase, for each identified concept, we generate an annotation rule that describes how to extract the data units of this concept in the result page and what the appropriate semantic label should be. The rules for all aligned groups, collectively, form the annotation wrapper for the corresponding WDB, which can be used to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phases again. As such, annotation wrappers can perform annotation quickly, which is essential for online applications.

3.2 MODULE DESCRIPTION

Our project consists of five major modules. The modules are followed with their description and the sample code.

Modules

- Web Database Creation
- Extracting SRR
- Aligning And Grouping Data
- Annotating Label
- Wrapper generation

3.2.1 Web Database Creation

Database is created to provide search interface for web sites. Two types of member are able to create database, one is admin and other one is user who are authorized to create record by admin. Authorized user function is to create record, view record and search record. Admin function is to create record, view record, modify record and view user details.

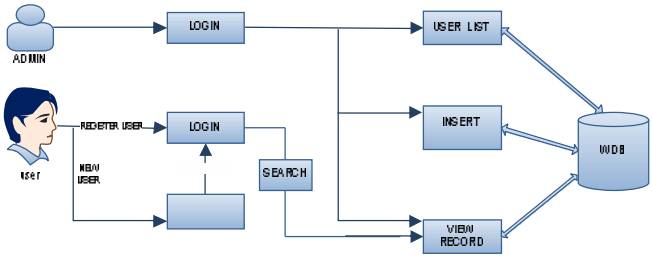


Fig 3.1 Block diagram of Web database creation

3.2.2 Extracting SRR

Search interface is used as tool to communicate with web database. Records are extracted from result of search. Keyword or phrase is given to search interface to identify and return result from web database. From result set the records are extracted. From the extracted record, parse to identify all data units and elements in search result.

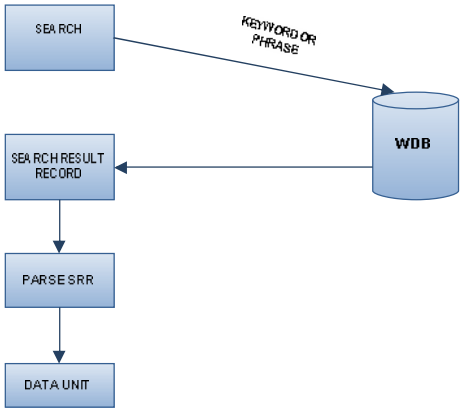


Fig 3.2 Block diagram of Extracting SRR

3.2.3 Aligning and grouping data

The purpose of data alignment is to organize data of same concept. Data alignment performed based on features such as data type, presentation style and so on. In aligned data, grouping is performed to organize data into different groups with each group corresponding to a same feature. The data alignment put the data units of the same concept into one group so that they can be annotated holistically. Whether two data units belong to the same concept is determined by how similar they are based on the features.

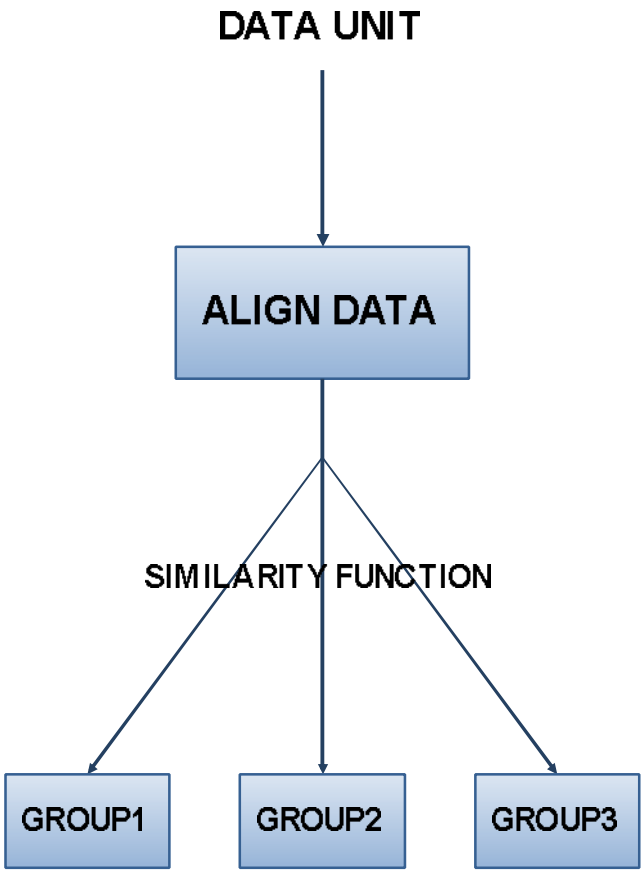


Fig 3.3 Block diagram of Aligning and grouping data

3.2.4 Annotating Label

On the grouped data, label is annotated for purpose of machine processing and deep web data extraction. Automatically label is annotated by analyzing the features on the same group. Different labels are assigned to semantically identical data units returned from different web database.

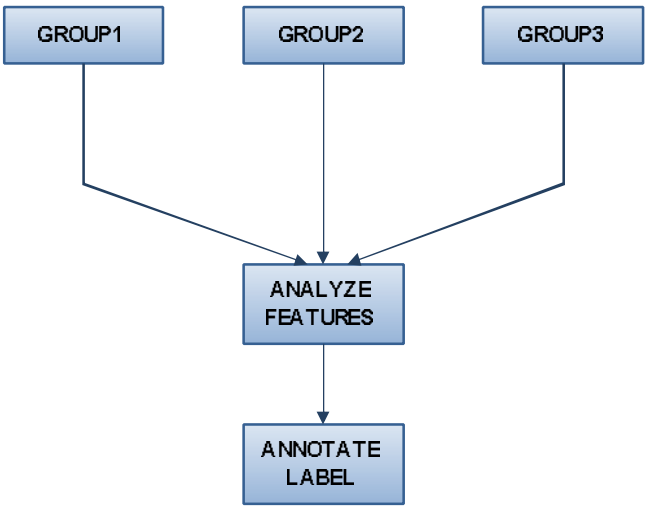


Fig 3.4 Block diagram for Annotating label

5.2.5 Wrapper Generation

The annotation wrapper for the corresponding WDB is created to annotate the label, which can be used directly to annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phases again. The efforts to automatically construct wrappers are used for effective data extractions.

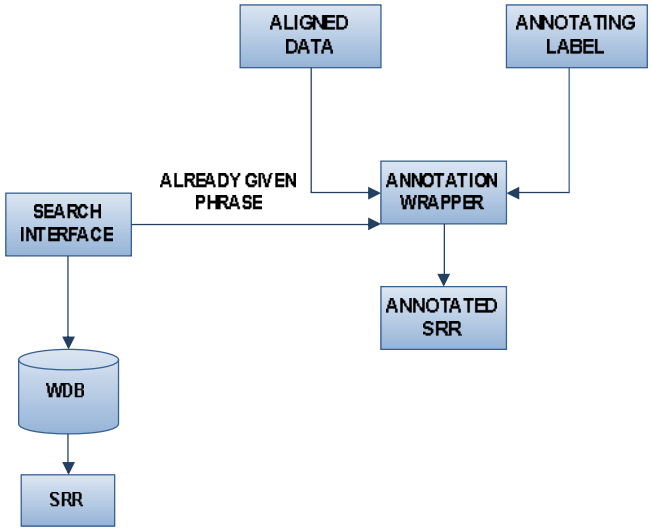
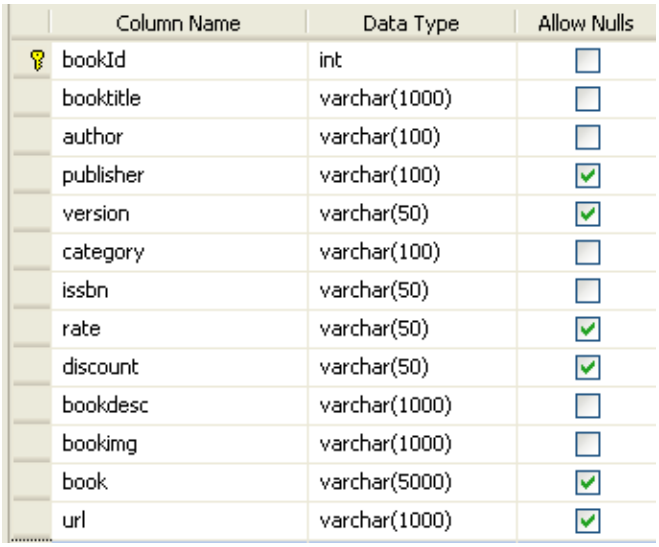


Fig 3.5 Block diagram of Wrapper generation

3.3 SAMPLE DATABASE

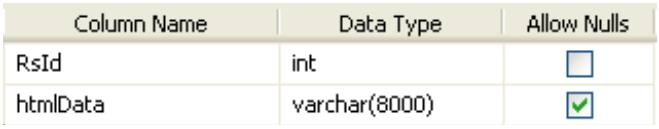
Column Name	Data Type	Allow Nulls
uname	varchar(50)	☐
password	varchar(50)	☐
dob	varchar(200)	☐
gender	varchar(50)	☑
eid	varchar(50)	☐
address	varchar(250)	☑
dateofreg	varchar(50)	☐

Fig 3.6 User registration



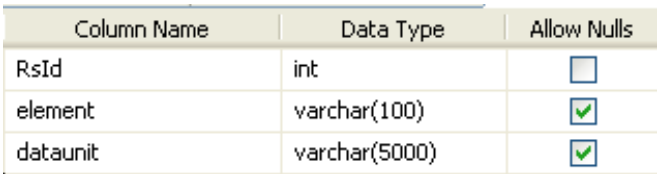
Column Name	Data Type	Allow Nulls
bookId	int	☐
booktitle	varchar(1000)	☐
author	varchar(100)	☐
publisher	varchar(100)	☑
version	varchar(50)	☑
category	varchar(100)	☐
issbn	varchar(50)	☐
rate	varchar(50)	☑
discount	varchar(50)	☑
bookdesc	varchar(1000)	☐
bookimg	varchar(1000)	☐
book	varchar(5000)	☑
url	varchar(1000)	☑

Fig 3.7Book upload



Column Name	Data Type	Allow Nulls
RsId	int	☐
htmlData	varchar(8000)	☑

Fig 3.8 Url list



Column Name	Data Type	Allow Nulls
RsId	int	☐
element	varchar(100)	☑
dataunit	varchar(5000)	☑

Fig 3.9 Extracted data units

4.1 CONCLUSION

For the automatic data annotation problem, a multi annotator approach is proposed which automatically construct an annotation wrapper for annotating the search result records retrieved from any given web database. This approach consists of six basic annotators and a probabilistic method to combine the basic annotators.

Each of these annotators exploits one type of features for annotation and results show that each of the annotators is useful and they together are capable of generating high quality annotation. A special feature of our method is that, when annotating the results retrieved from a web database, it utilizes both the LIS of the web database and the IIS of multiple web databases in the same domain.

The use of IIS can help to alleviate the local interface schema inadequacy problem and the inconsistent label problem. For automatic aligned problem, accurate alignment is critical to achieving holistic and accurate annotation. By a clustering based shifting method utilizing richer yet automatically obtainable features.

This method is 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. By creating annotation wrapper is easy to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the alignment and annotation phases again. The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB even with new queries.

REFERENCES

[1] A.Arasu and H. Garcia-Molina, “Extracting Structured Data from Web Pages,”
Proc. SIGMOD Int’l Conf. Management of Data, 2003.

[2] H. He, W. Meng, C. Yu, and Z. Wu, “Automatic Integration of Web Search Interfaces with WISE-Integrator,” VLDB J., Sept. 2004.

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