Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Annotating Web Databases

Jagruti B. Boraste
Department of Computer Engineering, Savitribai Phule Pune University
Maharashtra, India

***Abstract*— The databases on different web are accessible through HTML based search engines. The information or data mined from these web servers mostly in unstructured format. Whenever a web user submitted query, the web database shows various Search Result Records (SRRs).There is no any provision of the semantic labels of data units in search result pages for the user query. Early applications require human efforts to annotate data units manually, which limit their scalability and usability.**

To reduce the human efforts, the automatic annotator is proposed to remove set of SRRs from result page returned from the web databases in structured format. SRR have various data units each of which describes one aspect of a real world entity. Arrange data units into different groups with each group corresponding to a different concept and the data units within the same group having same meaning. Automatically assign appropriate semantic label to the data units within the group by using a variety of features simultaneously such as their data types, data contents, presentation styles, and adjacency information. A machine learning technique is implemented for data alignment and data annotation. A machine learning technique aligns and annotates the data retrieved from the different Web DataBases (WDB) in response to user queries. The machine learning technique improves the classification accuracy.

***Keywords*— Data alignment, Data annotation, Web database, Search Result Record**

I. INTRODUCTION

The use of Internet has been increased widely over a period of time. Also, the use of E- Commerce has increased quickly since a decade. The different Web Databases are good for managing the large amount of different data. There are different technologies and researches are focusing on the extraction of relevant information from large web data storage. But still there is necessity of availability of automatic annotation of this extracted information into a systematic way so to be processed later for different purposes. Web information extraction and annotation has been dynamic research area in web mining. The user or customer enter the search input query in the search engine, and search engine return the dynamically search output records on Web browser. The different web databases are accessed through HTML based search engine. The result returned for the user query from web database is in the form of Search Result Record (SRR).When we extract the pages, the resulted pages returned from a WDB have various Search Result Records (SRRs). SRR holds text nodes and data units. There is a high demand for data of interest from multiple Web Databases (WDBs). For example, a book comparison shopping system collects various result records from different book sites; it needs to find out whether any two SRRs refer to the same book. The system also wants to list the prices offered by each site. Thus, the system wants to know the semantic of each data unit.

Each SRRs represents one book with some data and text units. It consists text node outside the "<" and ">" in HTML source code, tag node surrounded by HTML Tags and title, author , price, publication are the values associated with it as data units. A data unit is nothing but a piece of text that semantically represents one concept of an entity. It varies from the text node which is refers to the sequence of text surrounded by a pair of HTML tag. The relationship between the data unit and text node is very essential for the purpose of annotation because the text node are not always the same to data nodes. The Web DataBases(WDB) has multiple sites to store in it. For this task, assigning label to required data and storing the collected SRR into a data base is important. Effectiveness of searching and updating information increases by Alignment and Annotation of data. Data alignment is arranging the data in such a way that data inside the same group have the same meaning and accessing in computer memory. Data annotation is the method for adding information to a document, a word or phrase, paragraph or the whole document. Data annotation allows fast retrieval of information in the deep web [1]. Finally, wrapper is produced which offers an annotation wrapper for the search site to automatically constructed and annotate the new result pages from the same web databases. This annotation wrapper produces an annotation rule that describes how to extract the data units from result page. Once the annotation wrapper annotate the data there is unnecessary to carry out the alignment and annotation phases again [2].

The wrapper is a software concept which wraps the contents of a web page using its source code via HTTP protocols but it does not alter the original query mechanism of that web page. This circumstance assumes that every web database is having a common schema design. Therefore, the terms extractors and wrappers are interchangeably used in most of the system [3].

II. LITERATURE SURVEY

Web information extraction and annotation has been a dynamic research area in recent years. The traditional system takes a lot of time to annotate the web database. Early applications need tremendous human efforts to annotate data units manually, which severely limit their scalability and usability. Automatically assign labels to the data units within the SRRs arrived from WDBs has been introduced in [1].The Author presents the three phases of annotation. Phase 1 is called the alignment phase. In this phase, first identify all data units in the SRRs and then arrange them into different groups with each group corresponding to a different concept (e.g., all titles are grouped simultaneously). In Phase 2 (the annotation phase), multiple basic annotators are introduced with each utilizing one type of features. Every basic annotator is used to generate a label for the units within their group holistically, and a probability model is adopted to decide the most appropriate label for each group. In Phase 3 (the annotation wrapper generation phase), for each identified concept, generate an annotation rule that illustrates how to extract the data units of this concept in the result page and what the appropriate semantic label should be.

Wrapper induction system [3] was introduced which rely on human users to mark the desired information on sample pages and label the marked data at the same time, and then the system can induce a series of rules (wrapper) to extract the identical set of information on web-pages from the same source. This system can usually accomplish high extraction accuracy. This system had poor scalability for some applications mentioned by authors [7], [8] that need to extract information from a large number of web sources. The efforts to automatically construct wrappers are [5] but the wrappers are used for data extraction only (not for annotation). There are some works [9], [10], [4], [11] which aim at automatically assigning meaningful labels to the data units in SRRs. The author Arlotta et al. [9] mainly annotates data units with the closest labels on result pages.

The Author present a novel data extraction method, ODE (Ontology-assisted Data Extraction) [10], which automatically extracts the query result records from the HTML pages. ODE first creates ontology for a domain according to information matching between the query interfaces and query result pages from different web sites within the same domain. Then, the created domain ontology is used during data extraction to identify the query result segment in a query result page and to align and label the data values in the extracted records. One limitation of ODE is that to label attributes it is essential that the labels appear in the query interfaces or query result pages within a domain. In [12], the author introduced a domain dependent annotation process. However, this process manually assigns the label to each data.

A multi-annotator approach in [13] that first aligns the data units into different groups such that the data in the same group have the same semantics. Then for every group, we annotate it from different aspects and aggregate the different annotations to predict a final annotation label. An annotation wrapper for the any search site is automatically constructed and can be used to annotate new result pages from the same site. In this approach the author proposed one-to-one and one-to-many relationship between text nodes and data units, but this system was not capable to find many-to-one and one-to-nothing type of relationship between text nodes and data units.

Many web sites hold large sets of pages generated using a common template. For example, Amazon sets the author, title, comments, etc. in the similar way in all its book pages. The values used to produce the pages (e.g., the author, price, title,.....) frequently come from a database. The system automatically [6] extracting the database values from such template generated web pages without any learning examples or other similar human input. The Author formally defines a template, and suggests a model that describes how values are encoded into pages using a template.

A latest vision-based approach [14] that is web page programming language independent is proposed. This approach mainly utilizes the visual features on the deep web pages to implement deep web data extraction, as well as data record extraction and data item extraction. This approach consists of four core steps: Visual Block tree building, data record extraction, data item extraction, and visual wrapper generation. However, there are still a few remaining issues, ViDE can only process deep web pages containing one data region while there is significant number of multi-data-region deep web pages.

Data Extraction and Label Assignment (DeLa) [4] first uses HTML tags to align data units by filling them into a table through a regular expression based data tree algorithm. Then, it employs four heuristics to decide a label for each aligned table column. The approach is performs attributes extraction and labeling simultaneously. However, the label is predefined and contains only a small number of values. Among all existing researches, DeLa [4] is the most similar to proposed technique.

III. PROPOSED SYSTEM

A web data extraction and data annotation is a research area in the web database. The data extraction and data annotation problem, i.e assigning meaningful labels to the extracted data unit of each SRR is a challenging task. Annotating or analyzing large data in a single website may lower the processing speed. The proposed system consist of Naive Bays machine learning technique. The classifier is used for the data alignment and annotation in the web databases. First the system loads the HTML source code page for the user query and then performs preprocessing on that result page. So system will get the only code for the SRRs. After that performs the alignment phase using algorithm and then use multi annotator approach to automatically annotate the data units. After that the system construct an annotation wrapper for annotating the search result records retrieved from any given web database. This approach consists of two basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators exploits one type of features for annotation and our experimental results show that each of the annotators is useful and they together are capable of generating high quality annotation. The machine learning technique is used to automatically obtain the data units with alignment and semantic labelling.

Overall algorithm of the system is as follows:
Input: Q // User query
Output: DU // Data unit with label
algoABWDST:

1. User select domain, after that select query type for particular website.
2. User enters the query in query field.
3. The search result page is loaded for the query and saves in text file.
4. Cleaning of collected data means extract useful information from collected data, then storing of useful information in text file for future purpose.
5. Perform classification of collected data units or clustering of data units on the basis of similar characteristics, using K-Means algorithm and Naïve Bays classifier.
6. Finding the suitable label for the data unit.
7. Display the data units with alignment and annotation. i.e display all data units in tabular form with appropriate label.

Following figure shows the architecture of proposed Annotating Web Databases system

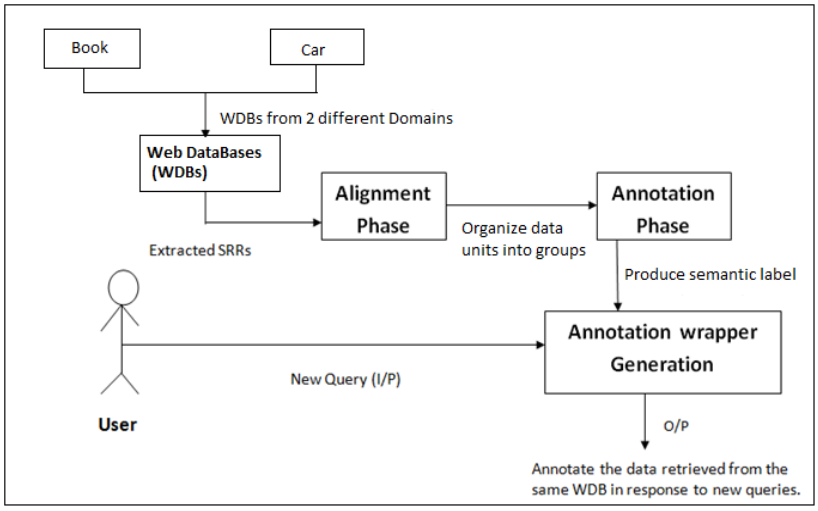

Fig. 1 Architecture Diagram

IV. RESULTS

This system is implemented and tested on 4 WDB select from Book domain and 2 WDB select from car domain. The WDB are used for training and testing. Data set DS1 is formed by obtaining one sample result page from all site and used for training. Data set DS2 and data set DS3 are generated by collecting one sample result page from each site for the testing.

TABLE I DATASET

Sr No.	Domain	Web Site
1	Book	www.bookadda.com
		www.uread.com
		www.shimply.com
		www.crossword.in
2	Car	www.cartrade.com
		www.motortrend.com

A. Performance Measures

For alignment, precision is defined as the percentage of the correctly aligned data units over all the aligned units by the system and recall is the percentage of the data units that are correctly aligned by the system over manually aligned data units by the expert.

For annotation, precision is the percentage of the correctly annotated units over all the data units annotated by the system and recall is the percentage of the data units correctly annotated by the system over all the manually annotated units.

B. Result comparison of K-Means Algorithm and Naive Bays classifier

The proposed system is implemented using Naive Bays classifier and the existing system is implemented using Clustering algorithm.

Performance for Alignment-

Fig. 2 shows the performance of data alignment for Bookadda, CrossWord, Shimply and Uread websites of Book domain. The precision and recall are calculated for query Q1="Computer Network". The query is applied on sites for both the techniques. The fig shows that Naive Bays classifier i.e. proposed system improves the data alignment than the clustering algorithm i.e. existing system.

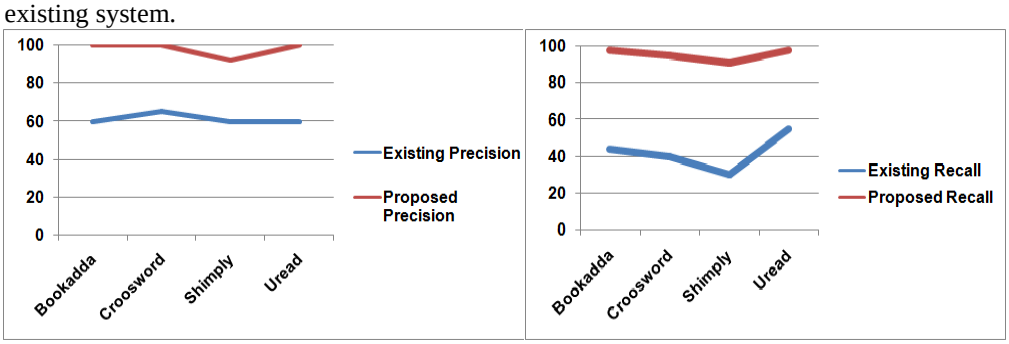

Fig. 2 A Graph of Precision and Recall for Existing system and Proposed system of Query-1

Performance for Annotation-

Fig. 3 shows the performance of data annotation for Bookadda, CrossWord, Shimply and Uread websites of Book domain. The precision and recall are calculated for query Q1="Computer Network". The query is applied on sites for both the techniques. The fig shows that Naive Bays classifier i.e. proposed system improves the data annotation than the clustering algorithm i.e. existing system.

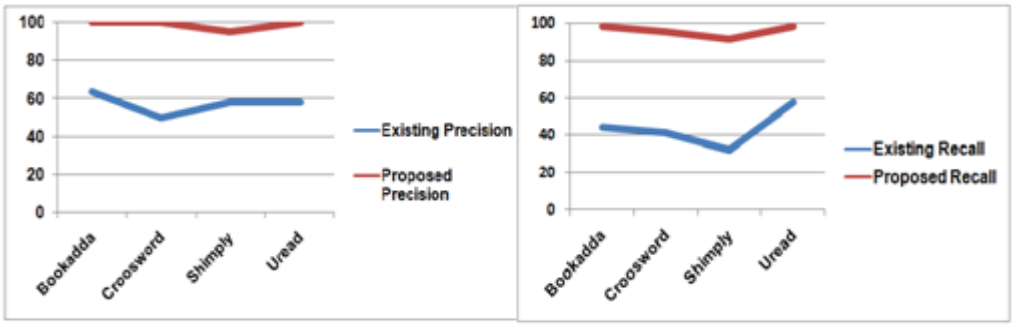

Fig. 3 A Graph of Precision and Recall for Existing system and Proposed system of Query-1

V. CONCLUSIONS

This system reveals the solution to the data extraction and data annotation problem. This system is useful for assigning meaningful labels to the extracted data unit of each search result record for the user query. The project consists of Naive Bays Classifier, which is used for the data annotation in the web databases. System performs the alignment phase in which data is classifying using classifier and then annotation phase is done. In annotation phase, multi annotator approach is used to labelling the classify data. After that, the construction of an annotation wrapper is introduced for annotating the search result records retrieved from any given web database. This system shows the data units retrieved from the web data bases are properly aligned as well as annotate. Our experimental results show that the precision and recall of using the classifier is improved than using the clustering approach.

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, Vol. 25, No.3, 2013.

[2] Boraste Jagruti Balkrishna,Swati A. Bhavsar, "Annotation based web databases search technique," In (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (1) , 2015, 388-391.

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

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

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

[6] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.

[7] W. Meng, C. Yu, and K. Liu, "Building Efficient and Effective Metasearch Engines," ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.

[8] 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.

[9] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. Sixth Int'l Workshop the Web and Databases (WebDB), 2003.

[10] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2, article 12, June 2009.

[11] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma, "Simultaneous Record Detection and Attribute Labeling in Web Data Extraction," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2006.

[12] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2005.

[13] 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.

[14] 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.

