ISSN:0975-9646

Retrieve Records from Web Database Using Data Alignment

V.kalyan Deepak[1], N.V.Rajeesh Kumar[2]

M.E Computer Science and Engineering[1], Faculty of Computing[2],
Sathyabama University[1,2]
Chennai-600119, India

Abstract-An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page 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. Our experiments indicate that the proposed approach is highly effective.

Keywords:HTML

I. INTRODUCTION

A large portion of the deep web is database based, i.e., 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 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. There is a high demand for collecting data of interest from multiple WDBs. While most existing approaches simply assign labels to each HTML text node, we thoroughly analyze the relationships between text nodes and data units. We perform data unit level annotation. A clustering-based shifting technique to align data units into different groups so that the data units inside the same group have the same semantic. Instead of using only the DOM tree or other HTML tag tree structures of the SRRs to align the data units (like most current methods do), our approach also considers other important features shared among data units, such as their data types (DT), data contents (DC), presentation styles (PS),and adjacency (AD) information.

II.RELATED ARTICLES

W. Liu, X. Meng, and W. Meng et al.[1] developed a paper for extracting structured data from deep Web pages is a challenging problem due to the underlying intricate structures of such pages. A large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language-dependent. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction. It is also proposed as new evaluation measure revision to capture the amount of human effort needed to produce perfect extraction.

J. Madhavan, D. Ko, L. Lot, V. Ganapathy et al[2] developed a paper for content hidden behind HTML forms,has long been acknowledged as a significant gap in search engine coverage. The paper describes a system for surfacing Deep-Web content, i.e., pre-computing submissions for each HTML form and adding the resulting HTML pages into a search engine index. The results of our surfacing have been incorporated into the Google search engine and today drive more than a thousand queries per second to Deep-Web content.

S. Mukherjee, I.V. Ramakrishnan, and A. Singh et al[3] developed a paper for identifying and annotating the semantic concepts implicit in such documents makes them directly amenable for Semantic Web processing. The paper describes a highly automated technique for annotating HTML documents, especially template-based content-rich documents, containing many different semantic concepts per document. Starting with a (small) seed of hand-labeled instances of semantic concepts in a set of HTML documents we bootstrap an annotation process that automatically identifies unlabeled concept instances present in other documents. The bootstrapping technique exploits the observation that semantically related items in content-rich documents exhibit consistency in presentation style and spatial locality to learn a statistical model for accurately identifying different semantic concepts in HTML documents drawn from a variety of Web sources. We also present experimental results on the effectiveness of the technique.

Y. Zhai and B. Liu et al[4] developed a paper for the problem of extracting data from a Web page that contains several structured data records. The first class of methods is based on machine learning, which requires human labeling of many examples from each Web site that one is interested in extracting data from. The process is time consuming due to the large number of sites and pages on the Web. The second class of algorithms is based on automatic pattern discovery. These methods are either inaccurate or make many assumptions.

III. SUMMARY OF EXISTING SYSTEM

In this existing system, 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. It describes the relationships between text nodes and data units in detail. In this paper, 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.

Disadvantages Of Existing System

- If ISBNs are not available, their titles and authors could be compared.
- The system also needs to list the prices offered by each site.
- 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, butalso for storing collected SRRs into a database table.

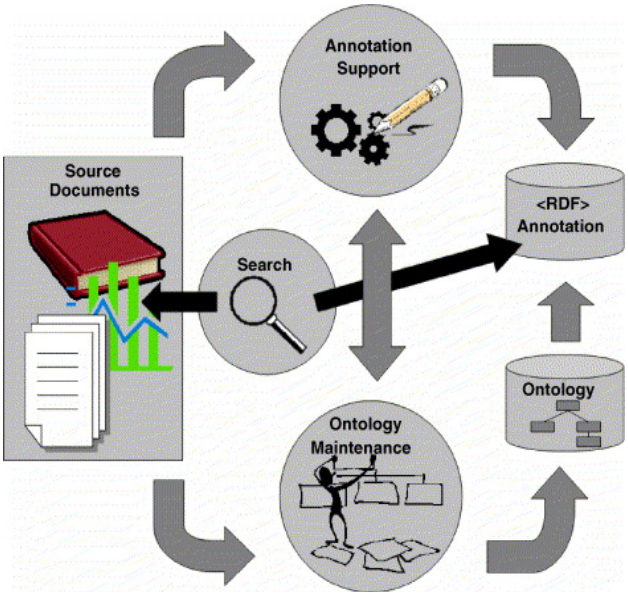
IV. PROPOSED SYSTEM

In this paper, we consider how to automatically assign labels to the data units within the SRRs returned from WDBs. Given a set of SRRs that have been extracted from a result page returned from a WDB, our automatic annotation solution consists of three phases.

Advantages Of Proposed System

- While most existing approaches simply assign labels to each HTML text node, we thoroughly analyze the relationships between text nodes and data units. We perform data unit level annotation.
- A clustering-based shifting technique 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.
- The 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.
- Construction of annotation wrapper for any given WDB. The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB with new queries.

V. BLOCK DIAGRAM



VI. IMPLEMENTATION

There are four modules:

- Data Alignment
- Data Annotation
- Web Database
- Wrapper Generation

A. Data Alignment

Data alignment is to 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 described. This data alignment achieve by the following process:

- Data content similarity
- Presentation style similarity
- Data type similarity
- Tag path similarity
- Adjacency similarity

B. Data Annotation

Data annotation method uses both the local interface schema (LIS) of the WDB and the IIS of the domain to annotate the retrieved data from this WDB. Using IISs has two major advantages. First, it has the potential to increase the annotation recall. Since the IIS contains the attributes in all the LISs, it has a better chance that an attribute discovered from the returned results has a matching attribute in the IIS even though it has no matching attribute in the LIS. Second, when an annotator discovers a label for a group of data units, the label will be replaced by its corresponding global attribute name (if any) in the IIS by looking up the attribute-mapping table so that the data units of the same concept across different WDBs will have the same label.

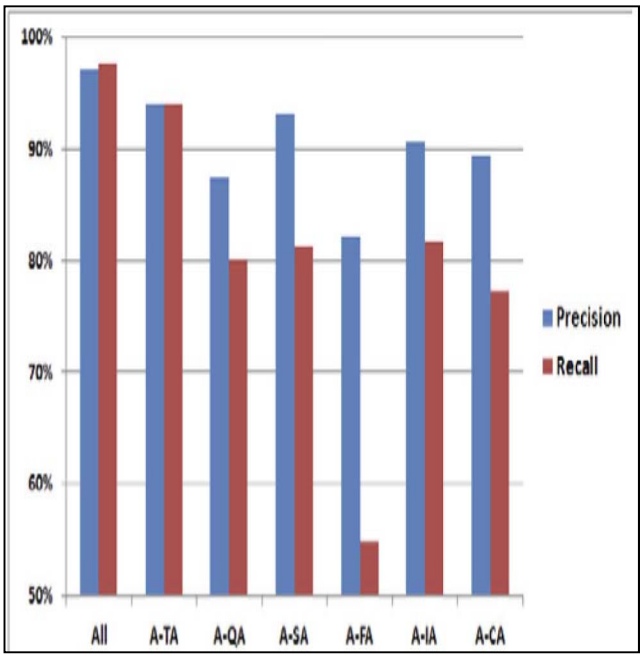
C. Web Database

A web database is an organized listing of web pages. It's like the card catalog that you might find in the library. The database holds a "surrogate" (or selected pieces like the title, the headings, etc.) for each web page. The creation of these surrogates is called "indexing", and each web database does it in a different way. Web databases hold surrogates for anywhere from 1 million to several billion web pages. The program also has a search interface, which is the box you type words into (like in Alta Vista or Google) or the lists of directories you pick from (like in Yahoo). Thus, each web database has a different indexing method and a different search interface.

D. Wrapper Generation

Once the data units on a result page have been annotated, we use these annotated data units to construct an annotation wrapper for the WDB so that the new SRRs retrieved from the same WDB can be annotated using this wrapper quickly without reapplying the entire annotation process. The annotation wrapper is a description of the annotation rules for all the attributes on the result page. After the data unit groups are annotated, they are organized based on the order of its data units in the original SRRs.

VII PERFORMANCE EVALUATION



VIII CONCLUSION

In this paper, we studied the data annotation problem and proposed a multiannotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. 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. We also explained how the use of the IIS can help alleviate the local interface schema inadequacy problem and the inconsistent label problem. In this paper, we also studied the automatic data alignment problem. Accurate alignment is critical to achieving holistic and accurate annotation. Our method is 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.

ACKNOWLEDGMENT

We would like to thank Dr.B.Bharathi, Head of the Department, Department of Computer Science and Engineering, Mr Rajeesh Kumar for her encouragement and support.

REFERENCES

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

[2] 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.

[3] P. Chan and S. Stolfo, “Experiments on Multistrategy Learning by Meta-Learning,” Proc. Second Int’l Conf. Information and Knowledge Management (CIKM), 1993.

[4] W. Bruce Croft, “Combining Approaches for Information Retrieval,” Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000.

[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] S. Dill et al., “SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation,” Proc. 12th Int’l Conf. World Wide Web (WWW) Conf., 2003.

[7] H. Elmeleegy, J. Madhavan, and A. Halevy, “Harvesting Relational Tables from Lists on the Web,” Proc. Very Large Databases (VLDB) Conf., 2009.

[8] 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.

[9] D. Freitag, “Multistrategy Learning for Information Extraction,” Proc. 15th Int’l Conf. Machine Learning (ICML), 1998.

[10] D. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, 1989.