Efficient Search Result Alignment with Annotation of Content And Querying Value

M. N. Praneswara Rao [1], Sandhya [2]

[1] M.Tech, Assistant professor of Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh, India.

[2] MCA, Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh, India.

***Abstract*— The Internet presents a huge amount of useful information which is usually formatted for its users, which makes it difficult to extract relevant data from various sources. Therefore, the availability of robust, flexible Information Extraction (IE) systems that transform the Web pages into program-friendly structures such as a relational database will become a great necessity .The motivation behind such systems lies in the emerging need for going beyond the concept of "human browsing." The World Wide Web is today the main "all kind of information" repository and has been so far very successful in disseminating information to humans[5]. The Web has become the preferred medium for many database applications, such as e-commerce and digital libraries.**

These applications store information in huge databases that user's access, query, and update through the Web. Database-driven Web sites have their own interfaces and access forms for creating HTML pages on the fly. Web database technologies define the way that these forms can connect to and retrieve data from database servers[3]. 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. And then we assign labels to each of this group.

***Keywords*—Data alignment, data annotation, web database, wrapper generation.**

I. INTRODUCTION

The Web has become the preferred medium for many database applications, such as e-commerce and digital libraries. These applications store information in huge databases that user's access, query, and update through the Web. Database-driven Web sites have their own interfaces and access forms for creating HTML pages on the fly. Web database technologies define the way that these forms can connect to and retrieve data from database servers.[3] The number of database-driven Websites is increasing exponentially, and each site is creating pages dynamically—pages that are hard for traditional search engines to reach. Such search engines crawl and index static HTML pages; they do not send queries to Web databases. 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. The explosive growth and popularity of the World Wide Web has resulted in a huge amount of information sources on the Internet. However, due to the heterogeneity and the lack of structure of Web information sources, access to this huge collection of information has been limited to browsing and searching. Sophisticated Web mining applications, such as comparison shopping robots, require expensive maintenance to deal with different data formats. To automate the translation of input pages into structured data, a lot of efforts have been devoted in the area of information extraction (IE). Unlike

information retrieval (IR), which concerns how to identify relevant documents from a document collection, IE produces structured data ready for post processing, which is crucial to many applications of Web mining and searching tools.

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 typical result page returned from a WDB has multiple search result records(SRRs). Each SRR contains multiple data unit search of which describes one aspect of a real-world entity. In this paper, 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 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[1]. We propose 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. IMPLEMENTATION

Our automatic annotation solution consists of three phases as

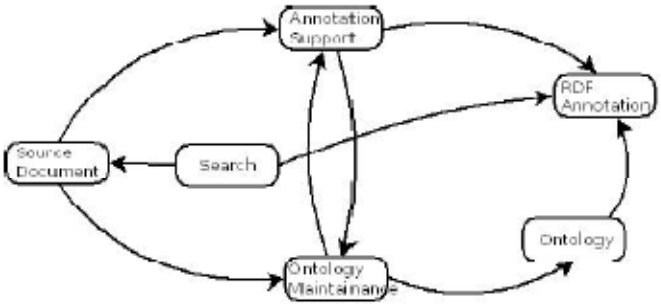

fig: Phases of automatic annotation solution

1) Extracts (automatically) text from a web-page into a table

2) Assigns labels in a table.

Phase 1 is the alignment phase, In this phase, we first identify all data units in the search records and then organize them into different groups with each group corresponding to a different concept the result of this phase with each column containing data units of the same concept across all search records. Grouping data units of the same meaning can help identify the common patterns and features among these data units. These common features are the basis of our annotators. Phase 2 is 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. Phase 3 is the annotation wrapper generation ,in this phase we generate an annotation rule that describes how to extract the data units of this concept in the result page and what the appropriate meaning annotation should be. The rules for all aligned groups, collectively, form the annotation wrapper for the corresponding WDB, which can be used to directly assign label the data retrieved from the same WDB in

response to new queries without the need to perform the above tow phases again. As such, annotation wrappers can perform annotation quickly, which is essential for online applications.[1]

Alignment Algorithm

Our data alignment algorithm is based on the assumption that attributes appear in the same order across all SRRs on the same result page, although the SRRs may contain different sets of attributes (due to missing values). This is true in general because the SRRs from the same WDB are normally generated by the same template program. Thus, we can conceptually consider the SRRs on a result page in a table format where each row represents one SRR and each cell holds a data unit (or empty if the data unit is not available). Each table column, in our work, is referred to as an alignment group, containing at most one data unit from each SRR. If an alignment group contains all the data units of one concept and no data unit from other concepts, we call this group well-aligned. The goal of alignment is to move the data units in the table so that every alignment group is well aligned, while the order of the data units within every SRR is preserved. Our data alignment method consists of the following four steps. The detail of each step will be provided later [1].

Step 1: Merge text nodes. This step detects and removes decorative tags from each SRR to allow the text nodes corresponding to the same attribute (separated by decorative tags) to be merged into a single text node.

Step 2: Align text nodes. This step aligns text nodes into groups so that eventually each group contains the text nodes with the same concept (for atomic nodes) or the same set of concepts (for composite nodes).

Step 3: Split (composite) text nodes. This step aims to split the "values" in composite text nodes into individual data units. This step is carried out based on the text nodes in the same group holistically. A group whose "values" need to be split is called a composite group.

Step 4: Align data units. This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept [1].

III. RELATED WORK

Chai-hui chang present a survey of information extraction system Data extraction based on partial tree alignment .the DOM tree is generated for each page. The web pages into program friendly structure such as relational database will generate Limitation:-manual labeling needs to make for extracted data[2]. ViDE A vision based approach. shen et.al approach used two-dimensional patches(i.e. one for spatial and one for the temporal dimension) but, if a patch contain both spatial and temporal dimensions as they can't be handled at the same time leads to motion discontinuity or an incomplete structure.)[3]. Embley et al. Utilize ontologies together with several heuristics to automatically extract data in multirecord documents and label them. However, ontologies for different domains must be constructed manually. Mukherjee et al.[25] exploit the presentation styles and the spatial localityof semantically related items, but its learning process for annotation is domain dependent. Moreover, a seed of instances of semantic concepts in a set of HTML documents needs to be hand labeled. These methods are not fully automatic[4]. LingLiu XWRAP: an XML-enabled wrapper construction system for Web information sources. The paper describes the methodology and the software development of XWRAP, an XML-enabled wrapper construction system for semi-automatic generation of wrapper programs. By XML-enabled we mean that the metadata about information content that are implicit in the original Web pages will be extracted and encoded explicitly as XML tags in the wrapped documents. In addition, the query based content filtering process is performed against the XML documents. The XWRAP

wrapper generation framework has three distinct features. First, it explicitly separates tasks of building wrappers that are specific to a Web source from the tasks that are repetitive for any source, and uses a component library to provide basic building blocks for wrapper programs. Second, it provides a user friendly interface program to allow wrapper developers to generate their wrapper code with a few mouse clicks. Third and most importantly, we introduce and develop a two-phase code generation framework. The first phase utilizes an interactive interface facility to encode the source-specific metadata knowledge identified by individual wrapper developers as declarative information extraction rules. The second phase combines the information extraction rules generated at the first phase with the XWRAP component library to construct an executable wrapper program for the given Web source. We report the initial experiments on performance of the XWRAP code generation system and the wrapper programs generated by XWRAP[5]. Chia-Hui ChangShih-Chien Kuo Olera: semi supervised Web-data extraction with visual support Olera is a semisupervised information-extraction system that produces extraction rules from semistructured Web documents without requiring detailed annotation of the training documents. It performs well for program-generated Web pages with few training pages and limited user intervention Crescenzi,V Efficient Techniques for Effective Wrapper Induction everal studies have recently concentrated on the generation of wrappers for extracting data from Web data sources. The ROADRUNNER system aims at automating the tedious and expensive process of writing wrappers in an unsupervised, domain-independent, and scalable manner. The system is based on a grammar inference algorithm, called MATCH, which has been designed in a sound theoretical framework. However, in its original definition MATCH lacks in expressivity; that is, in many cases when MATCH runs over real-life Web pages, it is not able to produce a solution. In this paper we address the challenging issue of developing techniques that allow us to build upon MATCH an effective and efficient system, without renouncing to the original formal background. First, we analyze the main limitations of MATCH; then we illustrate the techniques we have developed to overcome such limitations. Finally we report on the results of some experiments that show the efficacy of the introduced techniques and demonstrate the improvements of the overall system.

IV. CONCLUSION

In this paper, we automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database.

REFERENCES

[1] Annotating Search Results from Web Databases Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu, Senior Member, IEEE IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013

[2] 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

[3] 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.

[4] STAVIES: A System for Information Extraction from Unknown Web Data Sources through Automatic Web Wrapper Generation Using Clustering Techniques Nikolaos K. Papadakis, Dimitrios Skoutas, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 12, DECEMBER 2005

[5] A Survey of Web Information Extraction Systems Chia-Hui Chang, Member, IEEE Computer Society, Mohammed Kayed, Moheb Ramzy Girgis, Member, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 10, OCTOBER 2006

[6] Wang Computer Science Department University of Science and Technology Clear Water Bay, Kowloon Hong Kong Computer Science Department University of Science and Technology Clear Water Bay, Kowloon Hong Kong

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

[8] 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.

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

[10] W. Bruce Croft, "Combining Approaches for Information Retrie- val," Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000.

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

[12] 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.

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

[14] 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.

[15] D. Freitag, "Multistrategy Learning for Information Extraction," Proc. 15th Int'l Conf. Machine Learning (ICML), 1998.

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

BIOGRAPHY

Author 1: Details

M. N. Praneswara Rao is working as Assistant professor of Rajeev Gandhi Memorial College of Engineering & Technology, Nandyal, Andhra Pradesh, India. He has received B.E Degree in Computer Science and Engineering and his M.Tech in Computer Science and Engineering. His main research interest includes Networking. pranenaga@gmail.com

G.Sandya Rani, Persuing MCA Final Sem in Rajeev Gandhi Memorial College Of Engg and Technology, Dept. of MCA, JNTUAnantapur, Nandyal. sandyagangavaram@gmail.com