ISSN 2321-6905
October-2014

International Journal of
Science Engineering and Advance Technology

Vigorous Component Built Data Controlling System

Chunduri Raghavendra#1, M Ayyappa Chakravarthi#2
#1 Student of M.Tech Department Of Computer Science Engineering G.V.R&S college of Engineering & Technology, GUNTUR
2 Department of Computer Science Engineering G.V.R&S college of Engineering & Technology, GUNTUR

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 Internet super highway is widely used as a vehicle to information sharing across the globe. People across the globe, of all walks of life, are accessing Internet resources through search engines. The search engines provide web based interface for information search. Search engines return huge amount of data which is presented in encoded format through web pages. However, the data comes from underlying database. The search results from web databases that can be used further in applications like price comparison, data collection, and other related applications. When the search word "LAP TOP PRICES" is given in Google, it returned 24,60,00,000 result pages. Some of the results obtained are as follows.

Figu Some of the

As can be seen in Figure 1, it is evident that there are many search results that are associated with different web pages. The URL associated with each search result is different and the results came from many underlying databases in the web. The search results are to be made machine processable in order to use them further in real world applications. With the annotations, it is possible to process the web pages returned by search engines. For instance, the prices of various companies pertaining to a product can be compared. The comparison web sites can exist that derive data from across the pages returned by search engines. By providing price comparison, the web sites over Internet can help netizens to make well informed decisions. With many processing techniques, the search engines are presenting the results in meaningful way. Earlier the case was different. The results needed much human effort in order to annotate it manually. Recently Lu et al. [1]presented various ways of annotating the search results. They developed a mechanism that will automatically annotate the search results getting rid of manual labeling of web pages. Their solution

ISSN 2321-6905
October-2014

contains three phases. They are illustrated in Figure 2.

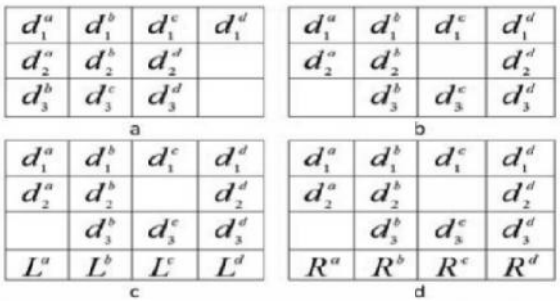


Figure 2 – Three phase annotation solution [1]

The first phase is known as alignment phase where data units are organized into groups based on different concepts. The phase 2 is known as annotation phase which takes care of making annotators that annotate web documents automatically. The phase 3 is known as annotation wrapper generation phase where an annotation rule is generated for each identified concept. Annotation wrapper is the collection of all the rules for all groups which have been aligned. Annotation wrappers help improve the process of annotation. A clustering based scripting technique is used to achieve this.

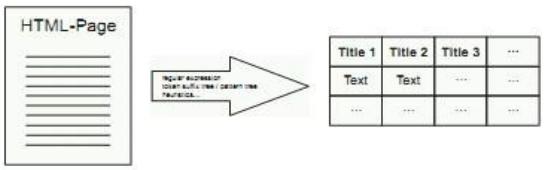
In this paper we implemented few annotators presented by Lu et al. [1]. We built a prototype application that takes care of automatic annotations of search results. The research results are obtained through Google search. The results are automatically annotated using the mechanism proposed in [1]. The empirical results revealed that our prototype is useful and can be used in the real world. The remainder of the paper is structured as follows. Section II reviews literature that focuses on the prior work pertaining to annotation of search results. Section III presents the proposed approach to achieve automatic annotation of web search results. Section IV presents experimental results while section V concludes the paper.

II.IMPLEMENTATION

Our automatic annotation solution consists of three Our data alignment algorithm is based on the assumption phases as that attributes appear in the same order

fig: Phases of automatic annotation solution

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 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.

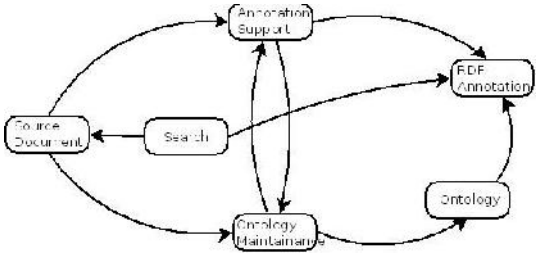
III.EXPERIMENTALRESUTLS

We have made experiments data from various domains with respect to two annotators only. The annotators used include table annotator and query – based annotator. Both the annotators are supported by the prototype application and it is extensible so as to support more annotators in future. The performance of data alignment and annotation are presented in Table 1.

Domain	Data Alignment Performance		Annotation Performance		Annotation with Wrapper	
	Precision	Recall	Precision	Recall	Precision	Recall
Auto	97.2%	97.4%	96.3%	96.5%	93.5%	90.2%
Book	97.1%	95.2%	96.2%	95.4%	92.6%	91.3%

Table 1 –Experimental results

As presented in Table 1, it is evident that more



than 90% precision and recall were recorded for both the performances such as data alignment and annotations. The table also shows the performance of annotation with wrapper. The results are presented in the following graphs.

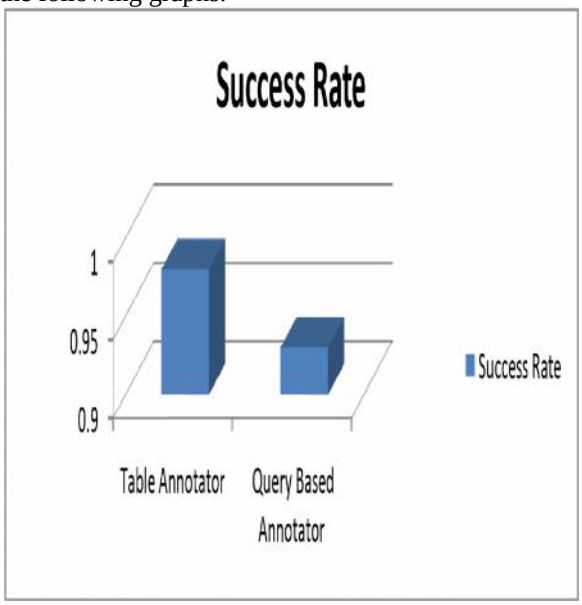


Figure 5– Success rate for table annotator and query based annotator

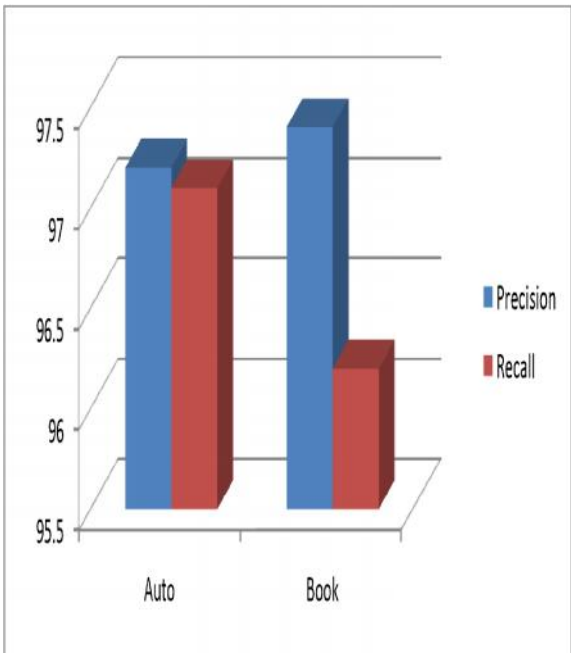


Figure 6 –Performance of data alignment for two domains

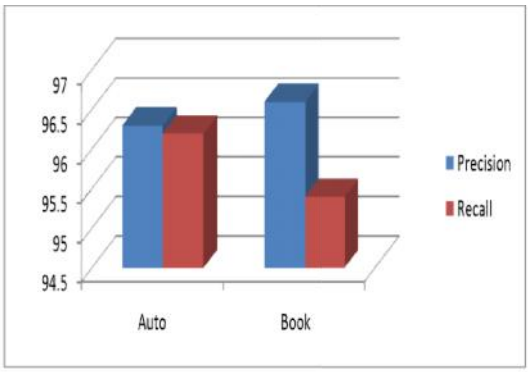


Figure 7 – Performance of data annotation for two domains

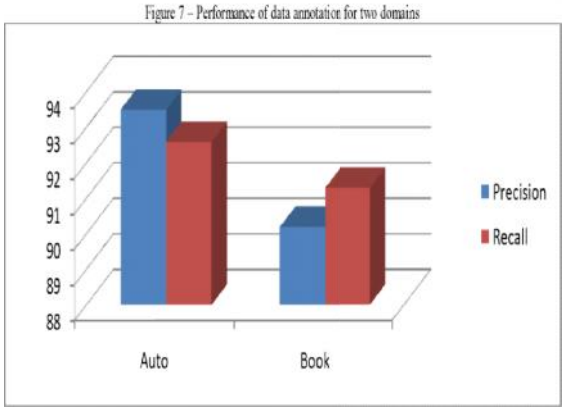


Figure 8 – Performance of data annotation with wrappers for two domains

As shown in Figure 5, Figure 6, and Figure 7, it is evident that the prototype application is capable of producing annotations automatically given search results of Google. The performance of the application is encouraging and the application can be used in the real world applications.

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, Weiyimeng, 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, DimitriosSkoutas, 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, MohebRamzyGirgis, 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

[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