ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

A Survey on Data Annotation for Web Databases

Bincy S Kalloor, Shiji C.G
PG Student, Assistant Professor, Department of Computer Science
Marian Engineering College, Trivandrum

Abstract - The Internet provides a great extent of beneficial knowledge which is usually formatted for its users, which makes it troublesome to extract relevant data from diverse sources. The World Wide Web plays a major role as all kinds of information repository and has been so far very successful in disseminating information to humans. For the encoded data units to be machine processable which is indispensible for many applications, such as deep web data collection and Internet comparison shopping, they need to be taking out and allot meaningful labels. In this paper present an automatic annotation approach, first aligns the data units on a result page into different groups, such that the data in the same group have the same meaning. Then for each group annotate it from different feature and collective the different annotations to predict a final annotation label.

***Keywords*-Data alignment, Data annotation, Web databases, Wrapper generation.**

I. INTRODUCTION

Data mining is the computational process of discovering patterns in big datasets. The overall goal of the data mining process is to extract information from a data set and convert it into an understandable structure for further use. Web mining is the one among the application of data mining techniques to discover patterns from the web. Web mining can be divided into three different types; they are web usage mining, web content mining and web structure mining. Search engines are very important tools for people to reach the vast information on the World Wide Web. Recent studies indicate that web searching; behind email is the second most popular activities on the Internet. A large portion of the deep web is database based. This type of search engine is referred as web databases. Result page returned from a Web database has multiple search result records. Each search result records contain multiple data units. Now there is a high demand for collecting data from multiple WDB's. The World Wide Web is a large source of structured data. It is a vast and rapidly growing receptable of information. There are various kinds of objects, such as products, people etc embed in both statically and dynamically generated web pages. Early applications requires large human efforts to annotate data units manually, which severely limit there scalability. In online shopping site for example eBay the databases are in unorganized manner, it is a time consuming process so automatic annotation proposed. Automatic annotation approach consists of three phases. They are alignment phase, annotation phase and annotation wrapper generation phase.phase1 is the alignment phase, in this phase first identify all data units in the search result records then organize them into different groups, with each group corresponding to its concepts. Groping data units of the same meaning can help to identify the common patterns and the features. In phase 2 introduce multiple basic annotators with each exploiting one type of features. Annotator is used to produce a label for the data units within their group, and probability model is adopted to determine the best appropriate label for each group. In phase 3 generation of an automatic annotation rule, which perform annotation quickly which is essential for online applications. As consequences, extracting data from WebPages and making it available to computer applications remains a complex and relevant task.

A. Motivation

In online shopping sites the databases are in unorganized manner, so it takes more time to get the search results. To overcome this drawback automatic annotation approach is proposed. This paper is organized as follows. Section 2 gives an overview of different recommendation techniques. Section 3 concludes the paper.

II. RECOMMENDATION TECHNIQUES

This section provides an overview of different recommendation techniques.

A. ViDE

Extracting structured data from deep web pages is a difficult task due to the underlying complex structures of such pages. Some of the limitations are web page programming dependent or precisely HTML document and incapable of handling the ever increasing complexity of HTML source code. To overcome this problem Vision-based Data Extractor is proposed. ViDE [2] W. Liu et.al is based on the visual features users can capture on the deep web pages while also utilizing some simple no visual information such as data type and frequent symbols to make the solution more robust. In earlier times labeling is done manually, it is time consuming and errors can be occurred. After that semi automatic annotation came in to exist, in this approach there is no scalability. After that automatic annotation approach came in to exist. Some of the disadvantages of the Vision-based Data Extractor, can only process deep WebPages contain one data region, while there is number of multi-data region deep Webpage, which is a time consuming process.

B. ODE

ODE means Ontology-Assisted Data Extraction which automatically extracts the records from the HTML files.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

ODE [3] .Su et.al is accurate in identifying the query result section ,segmenting the query result section into query result records and aligning and labeling the data values in the query result records. Automatic data extraction is important for many applications such as meta-querying, data warehousing etc. Data extraction is fully automatic and understanding of the query result page semantically. In Ontology Assisted Data Extraction, in semi automatic annotation there is no extra data are extracted the user can label only the data in which user is interested. The drawbacks are time consuming and labor intensive, hence it is not applicable to the large websites. To overcome the problems of the semi automatic wrapper induction some of the unsupervised learning methods are being used such as Roadrunner [4] V. Crescenzi et.al, Omni etc. Dela [1] Yiyao et.al is fully automatically extracting the data from the query result page based on the tag structure that exist on HTML pages. To overcome these visual features has been used for data extraction.

C. RoadRunner

Technique for extracting the data from HTML sites through the use of automatically generated wrappers. This technique is to automate the wrapper generation and the data extraction process to compare the HTML pages and introduce a wrapper based on their similarities and differences. Data is extracted by software modules called wrappers [4] V. Crescenzi et.al. Manually coded wrappers is quite difficult, labor intensive task and difficult to maintain. The goals of fully automatic wrapper generation are,

- Works by using additional information.
- Assumption that the wrapper induction system has some priori knowledge.
- Generation of a wrapper by examining one HTML page at a time.

D. WISE-Integrator

Searching is carried out either manually or semi automatically which is inefficient and difficult to maintain. It is a difficult task for users to access numerous Web sites individually to get the desired information. [5] H. He et.al it is a tool that performs automatic integration of Web Interfaces of Search Engines. It is used for identifying the matching attributes from different search interfaces for integration. WISE-extractor [5] H. He et.al is capable of automatically grouping elements into logical attributes and deriving a rich set of meta-information for each attribute.

E. ViNTs

A technique [6] H. Zhao et.al for automatically producing wrappers, used to extract search result record from dynamically generated result page. Automatic extraction of search result record is important for many applications.ViNTs [6] H. Zhao et.al utilizes both the visual content features on the result page as displayed on a browser and the HTML tag structure of the source file. Manually generating search result record wrappers is costly, time consuming and impractical for many application. Visual information And Tag structure based wrapper generator is a tool for automatically producing wrappers. In this paper focuses on the issue of how to extract the dynamically generated search result pages returned by search engine. A result page contains multiple SRR's and some of the irrelevant information to the users query. Accurate wrappers entirely based on the HTML tag structure. This method makes less sensitive to the misuse of the HTML tags.

F. HCRF

It is a Hierarchical Conditional Random Field [7] J. Zhu et.al. Existing approaches use decoupled strategies. That is attempting to data record detection and attributes labeling in two separate phases. Separately extracting data records and attributes is highly ineffective and proposes a probabilistic model to perform both processes simultaneously. HCRF [7] J. Zhu et.al can integrate all useful features by learning by their importance, and it can also incorporate hierarchical interaction. It is a template dependent. Expensive is the one of the main limitation.

III. CONCLUSION AND FUTURE WORK

From the above papers understood that extracting structured data from deep web pages is a challenging problem. Until now a large number of techniques have been proposed but some of them have inherent limitations. In online shopping sites databases are in unorganized manner, it is quite difficult to users. To overcome this automatic multi-annotator approach is proposed. There is still room for improvement in the automatic annotator which makes the annotator dynamic. In the dynamic annotator we can add features by using powerful classification technique.

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng "Annotating Search Results from Web Databases "IEEE Transaction 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] 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.

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

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 4, Issue 3, September 2014

[5] H.He, W.Meng, C.Yu, and Z. Wu," Automatic Integration of Web Search Interfaces with WISE-Integrator,"VLDB J., vol.13, no.3, pp.256-273, Sept.2004.

[6] H.Zhao, W. Meng, Z.Wu, V.Raghavan, and C.Yu,"Fully Automatic Wrapper Generation for Search Engines," Proc. Int'I Conf. World Wide Web, 2005.

[7] 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'I Conf. Knowledge Discovery and Data Mining, 2006.