Automatic Data Extraction from Deep Web Page

Sagar G.R[1] Mr. Rampur Srinath[2]
[1,2]Department of Information Science and Engineering
[1,2]The National Institute of Engineering, Mysuru

Abstract— There is large volume of information available in the World Wide Web. The information on the Web is contained in the form of structured and unstructured objects, which is known as data records. Our paper mainly concentrate on mined the data from the deep web pages, because most of data unit returned from the database are usually encoded into the result pages dynamically for human browsing. Some of the approaches used to solve this problem are manual approach, supervised learning, and automatic techniques. The manual method is not suitable for large number of pages. It is a challenging work to retrieve appropriate and useful information from Web pages. Currently, many web retrieval systems called web wrappers, web crawler have been designed. For the encoded data units to be machine process able, this is essential for many applications such as deep web data collection. Then most importantly our method displays the result as single page output. More fast and accurate at the same time, however, extracting the content from the original HTML document is complicated by the large amount of less informative and typically unrelated material such as navigation menus, forms, user comments, and ads. 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.

Key words: data annotation, web database, wrapper generation

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

Extracting information from web database and annotating them has been an active research area in recent years. Many systems confide in human users to mark the desired information on sample pages and label the marked data at the same time, and then the system can instigate a series of rules (wrapper) to extract the same set of information on web pages from the same source. These systems are often referred as a wrapper induction system. Because of the supervised training and learning process, these systems can usually achieve high extraction accuracy.

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

Extracting HTML document is complicated by the large amount such as navigation menus, forms, user comments, and ads. The Internet is widely used as a get necessary information sharing across the world. People across the world, 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.web databases that can be used further in applications like price comparison, data collection, and other related applications. When the search word "MACHINE" is given in Google, it returned more result pages.

It is evident that there are many search results that are associated with different web pages. The search results are to be made machine process able 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. With many processing techniques, the search engines are presenting the results in meaningful way.

II. RELATED WORK

In Online databases respond to a user query with result records encoded in HTML files. Data extraction, which is important for many applications, extracts the records from the HTML files automatically. We present a novel data extraction method, ODE (Ontology-assisted Data Extraction), which automatically extracts the query result records from the HTML pages.

ODE first constructs an 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 constructed domain ontology is used during data extraction to identify the query result section in a query result page and to align and label the data values in the extracted records Earlier the case was different. The results needed much human effort in order to annotate it manually. Recently Lu et al. 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.

In order to overcome these challenges, our system design transforms the page into a format called Extensible Hypertext Mark-up Language (XHTML) then; we make use of the DOM tree hierarchy of a web page and regular expressions are extracted out using the Extensible Style sheet Language (XSL).

The first phase is known as alignment phase where data units are organized into groups based on different concepts. The second phase is known as annotation phase which takes care of making annotators that annotate web documents automatically. The third phase 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 to improve the process of annotation to provide a best result to the user.

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 project, data unit level annotation is performed. 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.

There are several challenges in extracting information from a semi-structured web page e.g.

- Lack of a schema.
- Ill formatting.
- High update frequency and semantic heterogeneity of the information

Demerits

- In existing system. If ISBNs are not available means it is very difficult to retrieve a single book. No cost comparison is done here.
- Using only single label we cannot obtain the exact data which we required.

III. PROBLEM STATEMENT

Basically in every search engines just shows the web content and web links related to our input in the search box. It is just a text node which refers to a sequence of text surrounded by a pair of HTML tags. There is no the relationship between text nodes and data units. There is no cost comparison is done. Here, data unit level annotation and cost comparison is performed.

IV. OBJECTIVE

The main purpose of the paper is to provide multiple options for the user to retrieve the related data from the web database and also to avoid the replicate copies. The user can obtain the same data by using multiple labels like Title of the book, author name, url, ISBN and year of publishing.

It also help the user to overcome from confusion for e.g. if a user want to search a book name called Android. In existing search engines many result will be displayed due to this user get confused by thinking which book to select and which not to select. This problem can be overcome by using this project.

SCOPE: When we search any content in a search engine, it will group the content into different category related to what we are searching about and also provides data unit level annotation which means order or group the content which belongs to our wish.

V. PROPOSED SYSTEM

Here it will not only extract the data in efficient manner but this system will also display as a single page output and our system design transforms the page into a format called Extensible Hypertext Mark-up Language (XHTML) Then, we make use of the DOM tree hierarchy of a web page and regular expressions are extracted out using the Extensible Style sheet Language (XSL) technique, with a human training process. The relevant information is extracted and transformed into another structured format— Extensible Mark-up Language (XML)

The main focus has been done towards 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. The alignment phase, the annotation phase, the annotation wrapper generation phase. After completing this cost comparison will be performed.

Steps in proposed system are as follows

- The alignment phase.
- The annotation phase.
- The annotation wrapper generation phase.

A. System Architecture

Design is the process of converting a user oriented description of the data into a computer based system. This design is an important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

The input selection will be decided by the user like which parameter has to be given as label to search the information. For example User can select the book name, author name, year of publishing, url, ISBN as a label. By using any one of the parameter, he can draw the information.

In query result set huge amount of information will be stored, as the preprocessing unit receives the label, it will search the information related to the query. Multiple result will be obtained these information will be transferred to annotation result set block. The profile is used to establish the connection with the web database.

In Annotation result set block labeling will be carried out, in other word it can understand as an annotation is happening. An annotation means descriptive format of text or any web resources in other word it is considered as an alias labeling for a data units.

When the result set will be received from the annotation result set, it checks whether the given label is matching with the data present in the Meta database. Like the label is matching with the book name, author name, year of publishing and url. If any one of the field is matched means will be annotated with the same as the query submitted and rest of the related data will be also annotated.

After completing the annotation cost comparison will be carried out. Initially cost will be compared with different books, which ever the book having least cost, it will be displayed.

If many books have the same cost then we consider the number of view of the book. If the number of views also has the same number means whichever the data present at the first will be displayed. One copy of the result will be stored in final result set and the result will be displayed

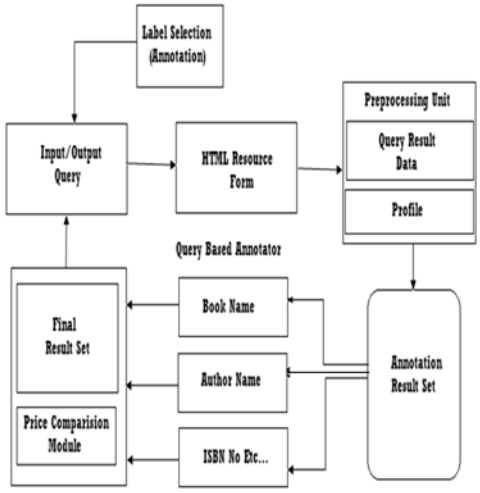

Fig. 1: System architecture.

B. Advantages

1. It consumes less time.
2. Ambiguity will be reduced. Since single information will be displayed.
3. Providing multiple options to user, to retrieve related information from the web database.

VI. CONCLUSIONS

In this paper mainly focused on the problem of annotating search results. The search results of search engines form web databases which can be used for further processing in order to leverage them in various applications like content comparison, data extraction and so on.

HTML tags are used to process the pages while annotating them. The annotated results are further useful efficiently in real world applications. There is no manual effort required it is automatic In addition, it is able to discover non-contiguous data records, which cannot be handled by existing techniques.

VII. FUTURE ENHANCEMENTS

Still this can also be improved by introducing best technique to the data alignment problem. Here now it is able to retrieve only a single book, it can be improved by displaying in each subjects, each streams etc. and find the data records that are not in the HTML form related tags.

REFERENCES

[1] M.Faheem and P. Senellart. Intelligent and adaptive crawling of Web applications for Web archiving. In ICWE, 2013
[2] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, Nirav ShahCrawling Deep web entity pages . In proceeding of WSDM' 2013.
[3] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi g Search Results from Web Databases, IEEE transactions on knowledge and data engineering, Vol. 25, No. 3, March 2013.
[4] Jain and M. Pennacchiotti. Open entity extraction from web search query logs. In Proceedings of ICCL, 2010.
[5] 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.
[6] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, a nd A. Halevy. Google's deep web crawl. In Proceedings of VLDB, 2008
[7] J. Madhavan, D. Ko, L. Lot, V. Ganapathy, A. Rasmussen, and A.Y. Halevy, "Google's Deep Web Crawl," Proc. VLDB Endowment, vol. 1, no. 2, pp. 1241-1252, 2008.
[8] H. Zhao, W. Meng, and C. Yu, "Mining Templates form Search Result Records of Search Engines," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2007.
[9] Y.Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. 14th Int'l Conf. World Wide Web (WWW '05), 2005.
[10] 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.
[11] 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.
[12] Yu, S., Cai, D., Wen, J.R. and Ma, W.Y. "Improving Pseudo-Relevance Feedback in Web Information Retrieval Using Web Page Segmentation", WWW2003
[13] 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.
[14] http://www.internetworldstats.com/Miniwatts Marking Group
[15] http://www.w3.org/TR/xslt.html XSL Transformations, W3C Recommendation
[16] http://www.w3.org/TR/xmlschema-0/ XML Schema Primer, W3C Working Draft
[17] http://www.w3.org/TR/xhtml1/ XHTML, W3C Recommendation
[18] http://www.w3.org/TR/xslt.htmlXSL Transformations, W3C Recommendation
[19] http://www.w3.org/TR/xmlschema-0/XML Schema Primer, W3C Working Draft

