www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 5 May 2015, Page No. 12205-12209

Survey on Data Extraction and Annotation Methods

[1]Tushar Jadhav, [2]Santosh Chobe
Department of Computer Engg, DYPIET Pimpri
Savitribai Phule Pune University, India
tusharjadhav2611@gmail.com
sanchobe@yahoo.com

Abstract - **Web consist of vast chunks of information on Web sites the user can retrieve this information by using search input query to Web databases & obtain the relevant information. Web databases return the multiple search records dynamically on browser, these search records contain the Deep Web pages in the form of HTML pages. The conventional search engine does not index the hidden Web pages from Web databases. Several existing techniques have addressed the problem of how to extract efficient structure data from Deep Web. The deep web refers to the hidden database used by web sites. But the information extraction & annotation is key challenge in web mining. The retrieval of information should be done automatically & arrange in a systematic way for processing. Different techniques like wrapper induction is been induced. Various types of annotators are used on the basis of the data to be labeled. This paper describes the automatic annotation approach on the basis of different feature of text node and data units.**

Keywords – Web Databases, wrapper generation, data annotation, data alignment.

I. INTRODUCTION

There are various technologies & researches which are focusing on the extraction of relevant information from large web data. But there is still need of availability of automatic annotation of this extracted information into a systematic way so as to get processed later for various purposes. Web information extraction and annotation has been active research area in web mining. There is huge amount of data which is available on the web , the user inputs the query in the search engine, and search engine returns the results on Web browser. There are Many E-commerce sites which are available to users, for e.g., when a user looks for details while buying a laptop or mobile phone such as price and configuration, but this type of information is only stored in the form of hidden back-end databases of the many notepad vendors, then the user has to visit each web sites to collect needed information from various web site and distinguish all those retrieved information manually ,so he can get the required product at reasonable price. This is a time consuming process & due to human effort it leads to inaccuracy up to particular extent. So as to get data as per user requirement there is need of technique which can help in providing such desired data. The work that is done previously mainly aimed at multiple techniques in firing queries, information fetching & optimization and wrappers. Wrapper is a software which extracts the contents of a web page using its source code via HTTP protocols [8] but it does not change the original query mechanism of that web page. This infers that each web database is having a common schema design.[2] WWW is having huge amount of data available on it but there is no tools so as to extract related information from Web databases. In deep web databases, search engines are referred as a Web databases (WDB). When we extract the pages, the resulted pages returned from a WDB have multiple Search Result Records (SRRs). Each SRRs contain multiple data units each of which describes one aspect of real-world entity & text units [1]. Consider a book comparison web, we can compare SRRs on a result page from a book WDB. Each SRRs represents one book with several data and text units .It consists of text node outside the <HTML>, Tag node surrounded by HTML Tags & title, author ,price, publication and the values associated with it as data units. A data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of record under an attribute. It is different from the text node which refers to the sequence of text surrounded by a pair of HTML tag.

The relationship between the data unit and text node is very important for the purpose of annotation because the text node are not always identical to data nodes. The WDBs has multiple sites to store in it, this requires two important tasks ,first label the required data and store the collected SRR in database. Early applications require tremendous human efforts because to annotate data units it has to be done manually, which badly limit their scalability. Later approaches focus on how to automatically assign labels to the data units within the SRRs returned from WDBs. So this well reduces human involvement and increases the accuracy. E.g. in a book comparison website we want to find the details such as price from the different websites for the same book so we can decide the choice to buy the book with the reasonable price and the reliable website. The ISBNs can be compared to achieve this. If ISBNs is missing, their titles and authors could be compared.

II. RELATED WORK

There has been lot of research made on web information extraction and annotation in recent years. Most of these systems depend on human users to select the desired

information on pages and at the same time the marked data is labeled, and then to extract the same set of data from page source the system applies series of rules. These systems are known as a wrapper induction systems.

Many systems use extraction and wrapper generation techniques for extraction of data.

Arvind Arasu et al.[1] proposed EXALG, for extracting structured data from a collection of web pages generated from a common template. EXALG first discovers the unknown template that generates the pages and uses the discovered template to extract the data from the input pages. EXALG makes use of two concepts, differentiating roles and equivalence classes.

Luigi Arlotta et al. [4] introduced wrapper induction system which marks the label data and also rely on human users. Since wrappers are built automatically, the values that they extract are anonymous and a human intervention is still required to associate a meaningful name to each data item.
The data extracted by automatically generated wrappers is a unusual problem, and it represents a step towards the automatic extraction and manipulation of web data. The web pages are designed to be used by humans, and that's why mostly they contain text strings, i.e. labels, The goal is to interpret the end user the meaning of the published data. However, this system achieves higher extraction precision in the result and an increased maintenance cost. Also, this system suffer lesser scalability that does not work in the applications like extraction algorithms. To reduce the cost associated with wrapper production and maintenance cost, the researchers have concentrated on automatic generation of wrappers.

S. Mukherjee et al.[5] discussed a method which is mainly concerned with automatically annotating HTML documents. To detect semantic changes in document content, it uses structural and semantic analysis techniques. The idea is to use template-based content-rich HTML documents. This technique shows the key observation that semantically related items that display consistency in presentation style and spatial locality.

F.H.Lochovsky et al.[6] introduced a system based on ontology, a new data extraction method in which query results are extracted from HTML pages automatically by using Ontology-assisted Data Extraction method. It creates ontology for a domain according to information which is identical between the query interfaces and query result pages from distinct web sites within the same domain. Then, the constructed domain ontology is used to match the query result section in a query result page and to align and label the data values in the extracted records for data extraction.

For efficient retrieval of data from the web, Y. Jiang et al.[7] made use of ideas taken from databases, however it requires structured data. Yet most web data is unstructured and cannot be queried using traditional query languages. To solve this problem, different ways for querying the web have been proposed. Basically there are two categories: querying the web with web query languages and generating wrappers for web pages.

An ontological approach is proposed for extracting and structuring data from documents posted on the web. The data extraction method is based on conceptual modelling, this approach focuses specifically on unstructured documents which are rich in data, narrow in ontological breadth, and contain multiple records of information for the ontology. So to automatically extract data in multi-record documents and label them, it employs ontologies together with several heuristics. However, it is necessary to construct ontologies manually for different domains.

Wei Liu et al.[8] proposed Vision-based Data Extractor (ViDE), which automatically extracts structured results from deep web pages. ViDE is basically based on the visual features human users can catch on the deep web pages and to make the solution more robust , it employs simple non-visual data such as data types and frequent symbols .It consists of two main components,(ViRIE) Vision based Data Record extractor and (ViDIE) Vision-based Data Item extractor. Using the visual features for data extraction, ViDE neglects the flaw of those solutions that need to analyze complicated web page source files.

H. Zhao et al.[9] proposed a technique 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. ViNT employs result page features such as visual content as it is shown 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 applications. Visual information and Tag structure which is based on wrapper generation is used to automatically produce wrappers. ViDE 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.

Searching is carried out either manually or semi automatically which is inefficient and difficult to maintain. It gets difficult for the users to access number of web sites individually to get the needed information. H. He et al.[10] proposed WISE-Integrator tool that performs automatic integration of Web Interfaces of Search Engines. It is used for identifying the similar attributes from distinct search interfaces for integration. WISE-Integrator is capable of automatically grouping elements into logical attributes and deriving a rich set of meta-information for each attribute.

J.Zhu et al.[11] proposed Hierarchical Conditional Random Field approach. Current approach makes use of decoupled strategies. The data record detection and attributes labelling is done in two separate phases. It gets ineffective the idea of extracting data records and attributes separately. It proposes a probabilistic model to perform both processes simultaneously. HCRF can integrate all useful features by learning by their importance, and it can also integrate hierarchical interaction. Its limitations are cost and template dependency.

J. Wang et al.[12] proposed DeLa, a method which is very similar to proposed annotation work. DeLa's alignment

method is based on HTML tags, on the other hand proposed work uses other features such as text content, adjacency information, data type, proposed annotation method deals with relationships between text nodes and data units, DeLa utilizes different search interfaces of WDBs for annotation.

Sr.No	Approaches	Methods	Tools	Limitation
1.	Manual	Data Extraction using Wrappers	Minerva WebOQL	Less Efficient Less Scalable
2.	Semi-Automatic	Tree Based	XWrap	Manual Labeling of Data Time Consuming
3	Automatic	Data Records Extraction	RoadRunner	Only text node level Annotation

Table 1

As per the analysis from Table 1, the limitation of manual approach had overcome by inducing sequence based and tree based techniques. In Road Runner[2] comparison between HTML pages and generate wrapper based on their similarity and differences. The Labeller is used for the automatic wrapper generation [5] Due to problem of human efforts and low efficiency, the unsupervised approach is an active research area in data extraction. Automatic data extraction approach is mainly categorized into three techniques data records extraction, HTML tag tree structure, Tree and pattern matching. But this approaches not suitable for the dynamic Web databases. ViDE is the Visual data extraction system which is independently works without HTML tag tree structure. ViDE is focused on the Visual features of the Web pages. ViDE is primarily based on the visual features human users can capture on the deep Web pages while also utilizing some simple non visual information such as data types and frequent symbols to make the solution more robust. [8]

III. PROPOSED WORK

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.

A. Objectives

- Perform data unit level annotation.
- Analyze the relationships between text nodes and data units.
- To utilize the integrated interface schema (IIS) over multiple WDBs in the same domain to enhance data unit annotation.

B. Data Extraction and Annotation

1. Automatic Annotation Approach :

An annotation is an online annotation associated with a web resource, typically a web page. With annotation system, a user can add, modify or remove information from a Web resource without modifying the resource itself. The annotations can be thought of as a layer on top of the existing resource, and this annotation layer is usually visible to other users who share the same annotation system.

Automatic annotation approach is a three phase approach.
These three phases are:
Phase 1: Alignment phase : This phase identifies all data units from the result records and then are grouped into distinct groups with every group identical to a distinct concept. The aim is to find the common patterns and features among the data units.

Phase 2: Annotation phase : In this phase multiple basic annotators are imported with each annotator employing one type of features. An annotator is generally used to build a label for the units in their group, then a probability method is implemented to resolve the most relevant label for each group..

Phase3: Annotation wrapper generation phase : In annotation wrapper generation phase an annotation rule is generated for each identified entity or concept. Wrapper is used to annotate the data units which retrieved from same web database for new queries which results in performing annotation quickly.

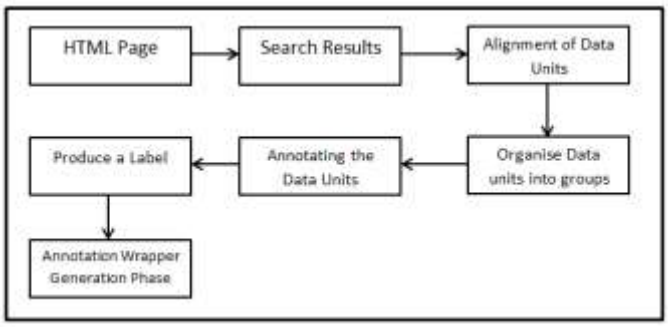

Fig 1: Phases of automatic annotation solution.

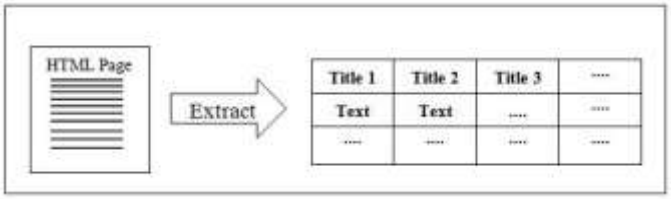

Fig 2: Extracts text from a web-page into a table

2. Data unit and text node :

Data unit is a chunks of data or text which semantically represents concept of real world entity. Data unit is totally different from text node where, text node is a sequence of text surrounded by pair of HTML tags. Text node is visible element on the web page and data unit located in the text nodes. Relationships between text node and data unit features are

a) One to One Relationship:
In this type of relationship, every text node consist of exactly one data unit.

b) One to Many Relationship:
In this type of relationship, many data units are enclosed in single text node.

c) Many to One Relationship:
It is referred as decorative tags, multiple text nodes are encoded into single data unit.

d) One To Nothing Relationship:
It is referred as template text nodes. Text nodes are not included in data unit of Search result record's.

3. Data and text node features :

There are five common features shared by the data units belonging to the same concept across all SRRs, and all of them can be automatically obtained.

a) Data content: To search information quickly data unit or text node of same concepts shares certain keywords.
b) Presentation style: Presentation style specifies how a data unit is shown on the web page by using few styles are out face, font size, color, text decoration etc.
c) Data type: These features are predefined characteristics that have their own meaning. Basically used data types are date, time, currency, integer, decimal etc.
d) Tag path: Sequence of tags traversing from root to corresponding node in the tree.
e) Adjacency: Adjacency refers to the data units that are immediately before and after in the SRR.

4) Alignment Algorithm:

Alignment is done based on same features, a group of data units having similar features are placed in one group by aligning it. If a group which consist of data units of one concept and if there is no data unit of other concepts then the group is known as well aligned group. The aim of alignment is to put the data units in the table so every alignment group is well aligned.

The goal of data alignment is to put the data units of the same concept into one group so that they can be annotated comprehensively.

a) **Merge Text Nodes**
It detects and removes decorative tags from every SRR, which permits the text nodes identical to the same attribute to be merged into a single one.

b) **Align Text Nodes**
After merging, it aligns text nodes into different groups. So that same group has the same concepts.

c) **Split (Composite) Text Node**
In this step the composite text nodes are splitted into separate data unit.

d) **Align Data Units**

This is the last step for alignment, in which every composite groups are separated in different multiple aligned groups, which contains data units of same concept.

5) *Data alignment, labelling and wrapper generation:*

Automatic annotation is based on alignment approach which aligns the data units by using different types of relationship in between data units and text nodes. A cluster-based shifting algorithm is used in alignment process. After the successful alignment label the data units and automatically construct an annotation wrapper for the search site

Conclusion

In this paper, the data annotation problem is mentioned and proposed a multi-annotator approach to annotate the SRR'S automatic annotation wrapper is used to search result records retrieved from web database.

The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. Basically three phases used for automatic annotation in which aligns the data units into different groups, labels each group and construct an annotation wrapper. In this work not all data units are encoded with the meaningful labels. A new algorithm for data annotation in the web database would be proposed. The proposed technique would be implemented with the expected results by using knowledge database as a database.

REFERENCES

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

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

[3] J. Wang, J. Wen, F. Lochovsky, and W. Ma, "Instance-BasedSchema Matching for Web Databases by Domain-Specific Query Probing," Proc. Very Large Databases (VLDB) Conf., 2004.

[4] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. Sixth Int' Workshop the Web and Databases (WebDB), 2003.

[5] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "BootstrappingSemantic Annotation for Content-Rich HTML Documents," Proc.IEEE Int'l Conf. Data Eng. (ICDE), 2005.

[6] 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.

[7] 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.

[8] 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.

[9] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. Int'l Conf. World Wide Web (WWW), 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] 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'l Conf. Knowledge Discovery and Data Mining, 2006.

[12] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web,2003.