IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AUTOMATIC ANNOTATION OF QUERY RESULTS FROM DEEP WEB DATABASE

Chaitanya Bhosale [*], **Prof. Sunil Rathod**

[*] Department of Computer Engineering, Dr. D.Y. Patil School Of Engineering
Lohgaon, Pune, India.

ABSTRACT

In recent years, web database extraction and annotation has received more attention from the database. When search query is submitted to the interface the search result page is generated. Search Result Records (SRRs) are the result pages obtained from web database (WDB) and these SRRs are used to display the result for each query. Every SRRs contains multiple data units similar to one semantic. These search results can be used in many web applications such as comparison shopping, data integration, metaquerying. But to make these applications successful the search pages are annotated in a meaningful fashion. To reduce human efforts, an automatic annotation approach is used. In which, we first aligns the data units on result records into various groups such that the information in the similar group have same meaning. After this we annotate each and every group in different domains and obtain the final annotation after aggregating them. In addition, we use New CTVS technique for extraction of QRRs from a query result page, in which we use optional labeling and dynamic tagging for the improvement. Then an annotation wrapper is generated automatically which is used for annotation new result records from the same web database.

KEYWORDS: Data alignment, data annotation, web database,wrapper generation,Information Integration,Search Result Records.

INTRODUCTION

Databases are known technologies for managing large amount of data. World Wide Web is a good way of presenting information. Alignment and annotation of data increases the quality of searching and updating data. Data alignment is the way of arranging data and accessing in computer memory. Data annotation is the methodology for adding extra information to a document, a word or phrase, paragraph or the entire document. In other words data unit annotation is the process of assigning meaningful labels to data. For example, a folder in a computer system labeled as “Trip-2015” might hold files of photographs taken in trip.

The automatic annotation solution as mentioned by authors of [1] consists of three phases- Alignment phase, Annotation phase, and Annotation wrapper generation phase. The alignment phase organizes all data units according to different groups where each group represents different concepts. The annotation phase groups the data to produce a meaningful label to every data units. The annotation rules are generated in annotation wrapper generation phase. The solution also uses six basic annotators; where each annotator can independently assign labels to data units. Two main concepts primary used for annotation research are data units and text nodes. Data unit is a piece of text that defines one concept of real world entity, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

Dynamically for human browsing these data units are encoded into the result page and assigned meaningful labels. Human efforts are required to annotate the data units. Thus, lack in scalability. To overcome this, automatic assigning of data units within the SRRs is required. An automatic annotation approach that first arrange all data into different groups i.e. inside the same group have same meaning and then each group is annotated in different aspects and aggregated to predict a final annotation. Finally, wrapper is generated. Wrappers are commonly used as translators which annotate new result records from the similar web database. This automatic annotation approach is scalable and highly effective. A clustering based shifting technique is proposed to align the data units into different groups.

In this paper, we introduce a new technique called New Combined Tag and Value Similarity (NCTVS) for the extraction of QRRs from a query result page. In existing CTVS;

1) Record extraction
2) Record Alignment

Comparing with the existing CTVS technique, new CTVS improves the data extraction accuracy in 2 ways:

1) Optional labeling &
2) Dynamic tagging

DATA ANNOTATION

Annotation is the process that first aligns the data units on a result page into different groups in such a way that data in the same group have the same semantic. Then according to grouping, each group annotates it from different styles. Labeling is done to give meaningful names to each group of the data units. After the successful labeling of the data units, the annotation wrapper is automatically constructed for the search sites. This can be used to annotate new result pages from the same WDB.

Extracting Structured Data From Web Pages:
A. Arasu and H. Garcia-Molina/ Proc. SIGMOD Int'l Conf. Management of Data/2003
price $10, put in the basket
Annotating Structured Data of the Deep Web:
Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu/ Proc. IEEE

```
<FORM><A> Extracting Structured Data from Web Pages
</A> <BR> A.Arasu and H. Garcia-Molina/<FONT><I>
Proc. SIGMOD Int'l Conf. Engineering of Data/2003
</I></FONT><BR> Price <B>$15</B>
```

price $15, put in the basket

Fig.2. Resulting source code of the HTML page

From the first record "extracting Structured Data from Web Pages" is the first text node. This text node is not always identical to data units. Where, A.Arasu and H. Garcia-Molina and other three fields like Proc. SIGMOD Int'l Conf. Engineering of Data /2003, price $10, put in the basket are the data units of databases. A text node is the text outside the "<" and "> in source code." Text nodes are the visible elements on the webpage and data units are located in the text nodes. Then comparison with the other fields is performed by using different types of relationships (one to one, one to many or many to one) in between data units and text nodes.

Table 1 indicates different methods for data annotation

Table 1. Different annotation methods

Sr. No.	Methods	Description
1.	Automatic annotation [1]	With the closest labels annotate the data units on result pages.
2.	Decorative tag detector [9]	For every HTML tags detect the decorative tags.
3.	Clustering based shifting [1]	Aligned data units are splits from composite text nodes.
4.	Simple probabilistic[1]	Analysis indicates that single annotator is not capable of fully labeling, so combine annotators.

LITERATURE SURVEY

In recent years, web database extraction and annotation has received much attention from the database and Information Extraction (IE) in research area. Many systems like wrapper induction [3],[4] depends on human users to mark and label the desired data. Then they induce a series of rules called wrapper to extract the similar set of information on result pages from the same web database. Hence, the system achieves high extraction accuracy through supervised training and learning process. But they suffer from poor scalability and not suitable for online applications like metasearch engine [12].

A similar approach [5],[18] is based on ontology means, in which it automatically extracts the data from web documents. Authors S. Mukherjee, et al. [7] discussed a method to align the data units which maintains only one type of relationship i.e. one to one relationship in between data unit and text nodes. For various domains ontologies are constructed manually. The effort to automatically build a wrapper has been presented in [1], [2], [6]. But, this method is used only for the data extraction, not for the annotation. The various methods were discussed in the literature [11], [10], [14] that assign the meaningful label to the data from the web databases. Most of the previous approaches of automatic data alignment techniques are based on few features like HTML tag paths(TP) [13], ViDIE uses Visual features [6], splitting of SRR into text segments [8]. The existing technique proposed in [1] report that they maintain all the type of relationship between the text nodes and data units. While the method [7] maintains only one to one relationship between the text node and the data units and the method [8] maintains one to many relationships between the text node and the data units. The method discussed based on the similar concept is DeLa [10]. But this method uses the HTML tags. It handles only two types of relationship between the text node and data units where we use all type of relationships. Here DeLa uses only local interface schema (LIS) search interface of WDBs for annotation process. Dela does not support the non contiguous query result records. It labels columns with labels only from the query result page or query interface from the same website. Complex search forms are used by this method rather than using keywords to keep track of pages that querying back end database.

J.Madhavan et al [15] define about deep web crawl in which content hidden behind HTML form which is obtained by form submission with valid text input values. Here, an algorithm ISIT is used to select input values for text search input that accept keywords. Y.Lu [9] describe about annotating the structured data of the deep web. It is similar to [1] and our method. where in this paper they describe about four relationships between text node and data units but only two of them i.e. one-to-one and one-to-many are explained [9] in detail.

In addition, we use clustering shift algorithm for one to nothing relationship where Y.Lu et al use pure clustering algorithm.

ViNTs [17] – For learning wrapper generation it requires a set of training pages from a website which uses both visual and tag features. Firstly it utilizes the visual data value similarity without considering the tag structure to check data value similarity regularities, denoted as data value similarity lines, and then combines them with the HTML tag structure regularities to generate wrappers. Both visual and non visual features are used to weight the relevance of various extraction rules. For this technique, the result page must contain at least four QRRs, and one no-result page is required to build a wrapper. The input used in system is URL of search engine's interface. Hence, it is necessary for ViNTs to monitor format changes to the query result pages, which is a difficult problem. In contrast, CTVS requires neither training pages nor a prelearned wrapper for a website. However, unlike ViNTs, CTVS cannot handle no-result pages, since CTVS assumes there are at least two QRRs in the page to be extracted.

In existing applications data units are manually annotated which requires lots off human efforts, which limit their scalability. Now, meaningful annotations are based on correct extraction of query results. Presently automatic web data extraction has been relatively grown. In this approach, the method of automatic web data extraction defined in Combining Tag and Value Similarity (CTVS) for data extraction and alignment purpose [16] is adapted. Among the above discussed web data extraction methods, CTVS can handle both non contiguous and nested structure data but none of the method label assignment. CTVS deals the Tag and Value similarity, in which data is automatically extracted from query results result records after the very first identification and segmentation of the query result records (QRRs) in the query result page and then alignment of the segmented QRRs in a table is done where the data values of the similar characteristics are put into the identical column. In this approach we introduced new method called New Combined Tag Value Similarity (NCTVS) for the extraction of QRRs from query result page. NCTVS improves the data extraction accuracy in two ways i.e. optional labeling and dynamic tagging.

PROPOSED APPROACH

Problem Definition

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. The scope of the project is when we extract 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. In this project data can be aligned efficiently and can also be extracted by performing careful linkage of data units.

Fig.3 elaborates the proposed system architecture.

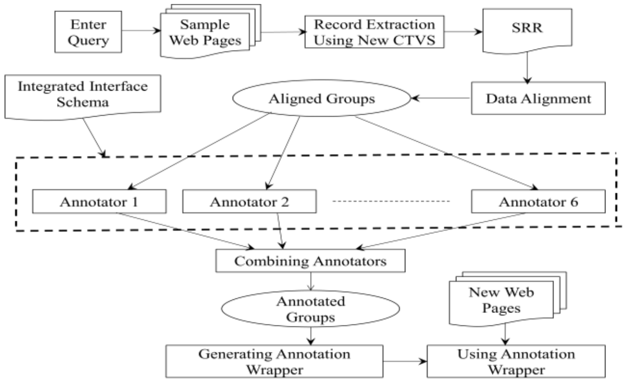


Fig.3. Proposed System Architecture

We introduce a new technique called new Combined Tag and Value Similarity (NCTVS) for the extraction of QRRs from a query result page.

- Record extraction identifies the QRRs in a query result page which involve the following sub steps: data region identification, buffering, semantic extraction and the segmentation step.
- Record Alignment where the data values for the same attribute are aligned and put in to the same column of the table

Comparing with the existing CTVS technique, NCTVS improves the data extraction accuracy in two ways:

- Optional labeling is the technique by which the problem of elimination of optional attribute that appears as the start node in a data region, as auxiliary information is eliminated. This is incorporated in the record extraction step.
- In the existing system which uses static tagging, results are far less accurate. The dynamic tagging uses the semantic data extraction concept described below.
- The existing system uses the datasets as result pages which are previously stored on local machine. It should have at least two QRRs for the extraction purpose. We overcome this in our proposed system. With this, record extraction and alignment is done, in addition there is optional labeling and dynamic tagging.

- First here we examine the relationship between text node and data units and perform data unit level annotation.
- To align the data units of different groups of same meaning we propose a clustering based shifting technique. Our system also considers some important features such as data contents (DC), data types (DT), presentation style (PS), and adjacency (AD) information.
- To improve the quality of data unit annotation we used the integrated interface schema (IIS) over various WDBs in the same area.
- In this we use six basic annotator which results are combine to form a single label.
- Then new annotation wrappers are constructed for any WDB. In which wrappers are used to annotate the same web database with new queries more easily.

The algorithm for the data annotation in [1] can be summarized as:

- Data alignment algorithm is based on the assumption that attributes appear in the same order across all SRRs on the same result page; the SRRs may contain different sets of attributes. Thus SRR arranged in table format.

Alignment algorithm is used for align the data units by using four steps. First; merge number of text nodes into single node. Then align the text nodes in different groups. Composite text nodes divide into single units. And lastly align the data units by separating composite groups into aligned groups and used clustering algorithm to clustering the text nodes.

Proposed System Algorithm

New CTVS Algorithm:
Input: Query Result Record, R
Output: Extracted Data, E
1. Input (I/P) Query
2. From the available links find the keywords
3. Store information in to a database
4. Perform structure analysis
5. Extract tags from the link
6. Store them to a temporary file
7. Match the attributes Identify the data regions
8. Segment the records ,Temp Containing optional data QRR Actual records
9. Merge QRRs
10 If the result not found then go for semantic extraction
11 Repeat steps 5
12 Final Result section is identified

Mathematical model

Let S be a system defined as,

S = ∑{I, f(x)}

I: Input Dataset or URL.

f(x): It provides a set of functions that performs on the input URL or Dataset,

defined as,

f(x) = {DT, TP, AD, DUS, Si, Se, Gi }

Step1:

DT = {fl, int, sy, dm}.

Where,

DT= Data Type, fl= First-Letter-Capitalized-String, int= Integer, sy= Symbol,

dm= Decimal

Step 2:

Tp={tagp}

Where,

TP= Tag Path, tagp = actual value of tag

Step3:

Ad= { d1, d2, dp, ds }

Where,

AD= Adjacency, dp,ds= data units

Step4:

Dus= {d1, d2}

Where,

d1, d2 = is a weighted sum of the similarities of the five features between them

Step5:

Local versus Integrated Interface Schemas

Si = {A1;A2; . . .;Ak | 1<=j<=k},

Where,

Aj= is an attribute.

Step 6: When a query is submitted against the search interface, the entities in the returned results also have a certain hidden schema, denoted as

Se = {a1; a2; . . . ; an},

Where, Each aj (j = 1 . . . n) is an attribute to be discovered.

Step 7:

Schema Value Annotator (SA): Many attributes on a search interface have predefined values on the interface.

It consist set of, Group of data units Gi = {d1; . . . ; dn},

The schema value annotator is to discover the best matched attribute to the group from the IIS. Let Aj be an attribute containing a list of values {v_1; . . . ; v_m} in the IIS.

For each data unit dk, this annotator first computes the Cosine similarities between d_k and all values in Aj to find the value (say v_t) with the highest similarity.

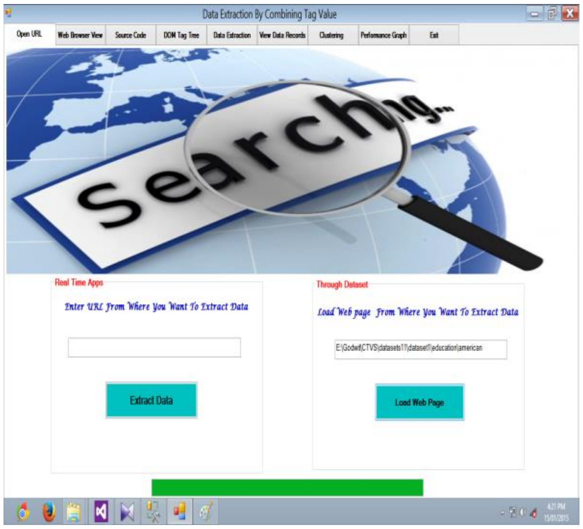
Where, Gi= Group of data units

V= list of values.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

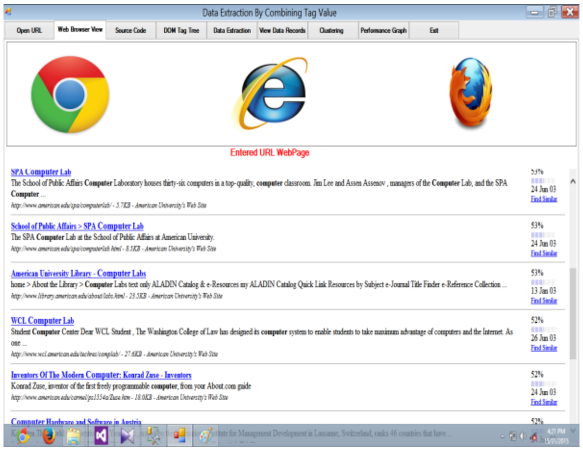
RESULTS

This section represents the results of the proposed approach for the data annotation and annotation features in the web databases.



E:\Godwit\CTVS\datasets11\dataset1\education\american

Fig. 4. Home Page (Enter URL of Web page)



SPA Computer Lab 53% 24 Jun 03
The School of Public Affairs Computer Laboratory houses thirty-six computers in a top-quality, computer classroom. Jim Lee and Assen Assenov , managers of the Computer Lab, and the SPA Computer ...
http://www.american.edu/spa/computerlab/ - 5.7KB - American University's Web Site

School of Public Affairs > SPA Computer Lab 53% 24 Jun 03
The SPA Computer Lab at the School of Public Affairs at American University.
http://www.american.edu/spa/computerlab.html - 8.5KB - American University's Web Site

American University Library - Computer Labs 53% 13 Jan 03
home > About the Library > Computer Labs text only ALADIN Catalog & e-Resources my ALADIN Catalog Quick Link Resources by Subject e-Journal Title Finder e-Reference Collection ...
http://www.library.american.edu/about/labs.html - 23.5KB - American University's Web Site

WCL Computer Lab 52% 26 Jun 03
Student Computer Center Dear WCL Student , The Washington College of Law has designed its computer system to enable students to take maximum advantage of computers and the Internet. As one ...
http://www.wcl.american.edu/techres/complab/ - 27.6KB - American University's Web Site

Inventors Of The Modern Computer: Konrad Zuse - Inventors 52% 24 Jun 03
Konrad Zuse, inventor of the first freely programmable computer, from your About.com guide
http://www.american.edu/carmel/ps1554a/Zuse.htm - 18.0KB - American University's Web Site

Computer Hardware and Software in Austria 52%
... Institute for Management Development in Lausanne, Switzerland, ranks 46 countries that have ...

Fig.5. Query related search results

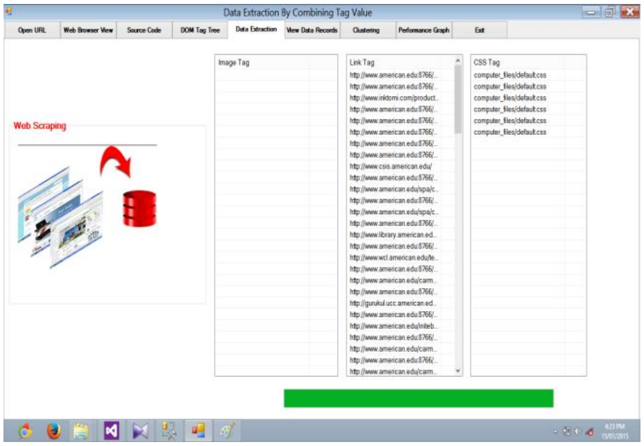
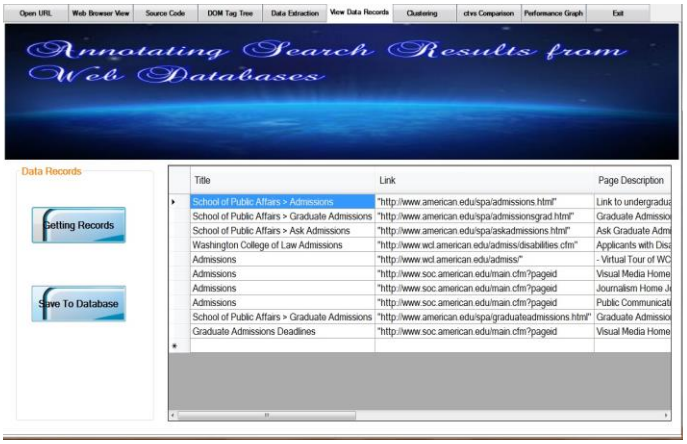


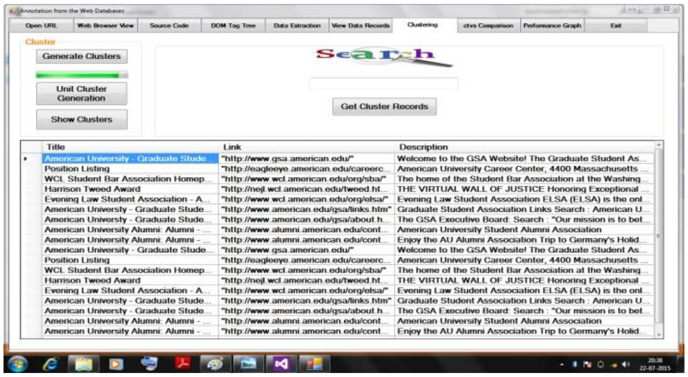
Image Tag	Link Tag	CSS Tag
	http://www.american.edu:8766/...	computer_files/default.css
	http://www.american.edu:8766/...	computer_files/default.css
	http://www.inktomi.com/product...	computer_files/default.css
	http://www.american.edu:8766/...	computer_files/default.css
	http://www.american.edu:8766/...	computer_files/default.css
	http://www.american.edu:8766/...	computer_files/default.css
	http://www.american.edu:8766/...	
	http://www.american.edu:8766/...	
	http://www.csis.american.edu/	
	http://www.american.edu:8766/...	
	http://www.american.edu/spa/c...	
	http://www.american.edu:8766/...	
	http://www.american.edu/spa/c...	
	http://www.american.edu:8766/...	
	http://www.library.american.ed...	
	http://www.american.edu:8766/...	
	http://www.wcl.american.edu/te...	
	http://www.american.edu:8766/...	
	http://www.american.edu/carm...	
	http://www.american.edu:8766/...	
	http://gurukul.ucc.american.ed...	
	http://www.american.edu:8766/...	
	http://www.american.edu/inteb...	
	http://www.american.edu:8766/...	
	http://www.american.edu/carm...	
	http://www.american.edu:8766/...	
	http://www.american.edu/carm...	

Fig.6. Data Extraction



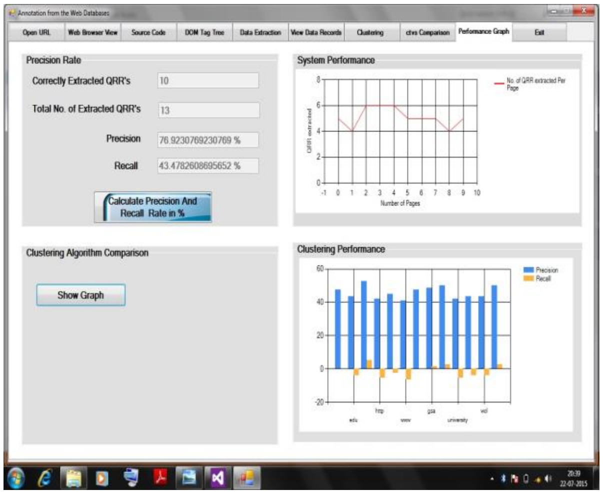
Title	Link	Page Description
School of Public Affairs > Admissions	"http://www.american.edu/spa/admissions.html"	Link to undergradua
School of Public Affairs > Graduate Admissions	"http://www.american.edu/spa/admissionsgrad.html"	Graduate Admissio
School of Public Affairs > Ask Admissions	"http://www.american.edu/spa/askadmissions.html"	Ask Graduate Adm
Washington College of Law Admissions	"http://www.wcl.american.edu/admiss/disabilities.cfm"	Applicants with Disa
Admissions	"http://www.wcl.american.edu/admiss/"	- Virtual Tour of WC
Admissions	"http://www.soc.american.edu/main.cfm?pageid	Visual Media Home
Admissions	"http://www.soc.american.edu/main.cfm?pageid	Journalism Home J
Admissions	"http://www.soc.american.edu/main.cfm?pageid	Public Communicati
School of Public Affairs > Graduate Admissions	"http://www.american.edu/spa/graduateadmissions.html"	Graduate Admissio
Graduate Admissions Deadlines	"http://www.soc.american.edu/main.cfm?pageid	Visual Media Home

Fig.7. View Data Records



Title	Link	Description
American University - Graduate Stude...	"http://www.gsa.american.edu/"	Welcome to the GSA Website! The Graduate Student As...
Position Listing	"http://eagleeye.american.edu/careerc...	American University Career Center, 4400 Massachusetts ...
WCL Student Bar Association Homep...	"http://www.wcl.american.edu/org/sba/"	The home of the Student Bar Association at the Washing...
Harrison Tweed Award	"http://nejl.wcl.american.edu/tweed.ht...	THE VIRTUAL WALL OF JUSTICE Honoring Exceptional ...
Evening Law Student Association - A...	"http://www.wcl.american.edu/org/elsa/"	Evening Law Student Association ELSA (ELSA) is the onl...
American Universty - Graduate Stude...	"http://www.american.edu/gsa/links.htm"	Graduate Student Association Links Search : American U...
American Universty - Graduate Stude...	"http://www.american.edu/gsa/about.h...	The GSA Executive Board: Search : "Our mission is to bet...
American University Alumni: Alumni - ...	"http://www.alumni.american.edu/cont...	American University Student Alumni Association
American University Alumni: Alumni - ...	"http://www.alumni.american.edu/cont...	Enjoy the AU Alumni Association Trip to Germany's Holid...
American University - Graduate Stude...	"http://www.gsa.american.edu/"	Welcome to the GSA Website! The Graduate Student As...
Position Listing	"http://eagleeye.american.edu/careerc...	American University Career Center, 4400 Massachusetts ...
WCL Student Bar Association Homep...	"http://www.wcl.american.edu/org/sba/"	The home of the Student Bar Association at the Washing...
Harrison Tweed Award	"http://nejl.wcl.american.edu/tweed.ht...	THE VIRTUAL WALL OF JUSTICE Honoring Exceptional ...
Evening Law Student Association - A...	"http://www.wcl.american.edu/org/elsa/"	Evening Law Student Association ELSA (ELSA) is the onl...
American Universty - Graduate Stude...	"http://www.american.edu/gsa/links.htm"	Graduate Student Association Links Search : American U...
American Universty - Graduate Stude...	"http://www.american.edu/gsa/about.h...	The GSA Executive Board: Search : "Our mission is to bet...
American University Alumni: Alumni - ...	"http://www.alumni.american.edu/cont...	American University Student Alumni Association
American University Alumni: Alumni - ...	"http://www.alumni.american.edu/cont...	Enjoy the AU Alumni Association Trip to Germany's Holid...

Fig.8. Clustering



Correctly Extracted QRR's 10

Total No. of Extracted QRR's 13

Precision 76.9230769230769 %

Recall 43.4782608695652 %

Fig.9. Performance Graph

CONCLUSION & FUTUREWORK

Assigning semantic labels to the extracted data unit of each SRR is a challenging task. The automatic multiannotator approach considers several types of data unit and text node features and makes annotation scalable and automatic. Here three phases used for automatic annotation in which alignment of the data units into different groups, labeling of each group and construction of an annotation wrapper. 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. We presented a novel data extraction method, NCTVS, to automatically extract QRRs from query result page with optional labeling and dynamic tagging. In Future work the first direction is to develop techniques for crawling, indexing and providing querying support for them structured pages in the web. Clearly lots of information in these pages is lost when naive key word indexing, and searching is used. Secondly we need to improve our method to split composite text node when there is no explicit separator.

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to the guide Prof. Sunil Rathod and Head of Department, Prof. SoumitraDas,as well as our principal Dr.S.S.Sonavne ,also Grateful thanks to our PG Coordinator Prof.P.M.Agarkar and last but not least, the departmental staff members for their support.

REFERENCES

[1] Hai He, Hongkun Zhao, Y. Yiyao Lu, Weiyi Meng, Annotating Search Result Records from web databases, IEEE Transaction on Knowledge and Data Engg., volume 25, No. 3, mar. 2013

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

[3] N. Krushmerick, D. Weld, and R. Doorenbos, Wrapper Induction for Information Extraction, Procedure Intl Joint Conference Artificial Intelligence (IJCAI), 1997.

[4] L. Liu, C. Pu, and W. Han, XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources, Procedure IEEE 16th Intl Conference Data Engg. (ICDE), 2001.

[5] 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 Engg., volume 31, No. 3, pp. 227-251, 1999

[6] W. Liu, X. Meng, and W. Meng, ViDE: A Vision-Based Approach for Deep Web Data Extraction,IEEE Transaction Knowledge and Data Engg., volume 22, no. 3, pp. 447-460, Mar. 2010.

[7] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, Bootstrapping Semantic Annotation for Content-Rich HTML Documents, Procedure IEEE Intl Conference Data Engg. (ICDE), 2005.

[8] H. Elmeleegy, J. Madhavan, and A. Halevy, Harvesting Relational Tables from Lists on the Web, Procedure Very Large Databases (VLDB) Conference, 2009.

[9] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, Annotating Structured Data of the Deep Web, Procedure IEEE 23rd Intl Conference Data Eng. (ICDE), 2007.

[10] J. Wang and F.H. Lochovsky, Data Extraction and Label Assignment for Web Databases, Procedure 12th Intl Conference World Wide Web (WWW), 2003.

[11] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, Automatic Annotation of Data Extracted from Large Web Sites, Procedure Sixth Intl Workshop the Web and Databases (WDB), 2003.

[12] Z. Wu et al., Towards Automatic

Incorporation of Search Engines into a Large-Scale Metasearch Engine, Procedure IEEE/WIC Intl Conference Web Intelligence (WI 03), 2003.

[13] Y. Zhai and B. Liu, Web Data Extraction Based on Partial Tree Alignment,Procedure 14th Intl Conference World Wide Web , 2005.

[14] J. Wen, B. Zhang, J. Zhu, Z. Nie, and W.-Y. Ma, Simultaneous Record Detection and Attribute Labeling in Web Data Extraction, Procedure ACM SIGKDD Intl Conference Knowledge Discovery and Data Mining, 2006.

[15] D. Ko, L. Lot, V. Ganapathy, A. Rasmussen, J. Madhavan, and A.Y. Halevy, Googles Deep Web Crawl, Procedure VLDB Endowment, volume 1, no. 2, pp. 1241-1252, 2008.

[16] Frederick H. Lochovsky ,Weifeng Su, and Jiying Wang Combining Tag and Value Similarity for Data Extraction and Alignment, IEEE Transactions on knowledge and Data Engineering, Volume 24, no.7, pp. 1186–1200, July 2012.

[17] W. Meng, Z. Wu, V. Raghavan, H. Zhao, and C. Yu, ViNTs – Fully Automatic Wrapper Generation for Search Engines, In Proceedings of the 14th World Wide Web Conference, pp. 66– 75, May 2005.

[18] F.H. Lochovsky, W. Su, J. Wang ODE: Ontology-Assisted Data Extraction, ACM Transaction Database Systems, volume 34, no. 2, article 12, June 2009Apr. 2000.