ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

Automatic Annotation Wrapper Generation and Mining Web Database Search Result

V.Yogam[1], K.Umamaheswari[2]

[1] PG student, ME Software Engineering, Anna University (BIT campus), Trichy, Tamil nadu, India

[2] Assistant professor, Department of computer science, Anna University (BIT campus), Trichy, Tamil nadu, India

Abstract*:* Search result record (SRR) is the result page obtained from web database (WDB) and these records are used to display the result for each query. Each SRR contain multiple data units which need to be label semantically for machine processable. In this paper we present the automatic annotation approach which involve three phases to annotate and display the result. In first phase the data units in result record are identified and aligned to different groups such that the data in same group have the same semantics. In the second phase, for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. In third phase, an annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. This approach is highly effective. From the annotated search result, frequently used websites are identified by using apriori Algorithm which involve pattern mining. The advantage of this new technique is fast operation on dataset containing items and provides facilities to avoid unnecessary scans to the database

Keywords— Result record, web database, data units, annotation wrapper.

I. INTRODUCTION

Data mining is processes of manipulating the large dataset and to extract some knowledge from it which can be easily understand by the human being. For many search engine the data comes in the result page is based on the some structured database which is also called as web database (WDB).

A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search results record (SRR). Web database has multiple search result record. Each SRR refer to an entity. SRR from web database contain multiple data units (or instances). Each SRR refer to an entity. Data units are different from text node. Text node is surrounded by pair of HTML tags. Data units are texts that semantically represent the single concept of an entity. These data units are not used for application such as deep web data collection and internet comparison shopping. Here the annotation is done on the basis of data units. The data units are annotated by assigning meaningful labels to them.

Annotation problem has become significant problem due to the rapid growth of the deep web and the need to query multiple web mining, it is imperative that is data units are correctly labeled so they can be appropriately organized and stored for subsequent machine processing. Note that the search sites that have web service interfaces, it may be easier to annotate their SRRs because the semantic meanings of their data units more clearly describe in WSDL. However that very few search sites have web services interfaces. Therefore it is still necessary to extract and annotate data from legacy HTML pages.

In this system we first extract the SRR page from the given web database. Then the data units are identified and aligned such that the aligned data units are belong to the same attributes or concepts. We then design different basic annotator to annotate data units of each aligned group. These different basic annotator results are combined to determine appropriate label for each data unit groups. Finally the annotator wrapper is generated for the corresponding WDBs which is used to annotate new SRRs retrieved for different queries.

Hence the automatic annotation solution of SRR consists of three phase. The first phase is the alignment phase. In this phase we first identify all the data units available and organize them in to different groups. Grouping data units of same semantics helps to identify the common patterns between these data units which are the basis for annotator. The second phase is the annotator phase. In this phase many multiple basic annotators are introduced in which some common features are followed. Each basic annotator is used to label the units of same group. It is also used for identifying most appropriate label for each group. The third phase is the annotation wrapper generation phase. In this phase an annotation rule is generated for each identified concepts which shows how to extract data units of same group. The collective annotation rule for a aligned groups is known as annotation wrapper for the corresponding web database. This annotation wrapper is used annotate the data units for different queries without generating alignment and annotation phase. So this makes the annotation quicker.

Then we perform a frequent pattern mining to retrieve the frequently used web page of annotated group. It is used find the most trustable web sites so that the result page we produce will be more effective. Hence our system has the following contribution.

- First here we analyse the relationship between text node and data units and perform data units level annotation.
- To align the data units of different groups of same semantics we propose a clustering based shifting technique. Also in our system we consider some important features such as data types (DT), data contents (DC), presentation style (PS), and adjacency (AD) information.
- To enhance the data unit annotation we utilize the integrated interface schema (IIS) over multiple WDBs in the same domain.
- Here we use six basic annotator which results are combine to form a single label.
- Then new annotation wrappers are constructed. Which is used to annotate the same web database for different queries more easily.

At last the frequent item sets are generated to identify the correlation value of the particular datasets.

II.RELATED WORK

Web information extraction and annotation is an active area in recent years. Many system like wrapper induction system are rely on human [6],[11] to generate the wrapper on the marked data of the sample page. These systems can achieve high extraction accuracy because of some supervised training and learning process. But it performs poor scalability for the application that need to extract information from large number of web source.

Embley et al [10] utilize ontology and other heuristics to automatically extract data in multiple records and label them. But ontology for different domain needs to be constructed manually. Arasu et al [1] describe about extracting structured data from the web page. In which structured template is used to extract the information from the web page. To extract information from the unstructured page structured template pages are used. The human input is absence here so that the occurrence of error is limited and time consuming. But it does not suitable for large database also it does not say about crawling, indexing and providing support to querying structure pages in web. Information is lost when naive key word indexing and searching is used.

J.Madhavan et al [2] define about deep web crawl in which content hidden behind HTML form which is obtained by form submission with valid input values. These inputs are text inputs. Here an algorithm ISIT is used to select input values for text search input that accept keywords. Here informative test is used to evaluate query template for combination of the form input. It increases the accessibility of deep web content for search engine users. Dependencies between values in different input of a form are not considering. No annotation technique is used.

Now a day there are thousands of search engines were available in the web. But there is a demand to generate automatic tool (wrapper) to extract the selected result records from the HTML result page of search engine .Clement et al [] deals with the dynamic content of automatic extraction of select result records. Here the section extraction is focused which automatically extract all the dynamic section from search result page. Static and semi dynamic content are used to find the boundaries of different dynamic sections and it addresses the issue of correctly differentiating sections and the records. But it does not do any automatic annotation technique.

E-commerce search engine (ESE) is used by the user to search and compare products from multiple web sites. H. He et al [4] proposed an E-commerce Meta search engine (EMSE) is built fully automatically. It has many components. Here, the focus is on the interface integration step of the E-Meta base project. WISE integrator is used to do interface integration step automatically. Hence the WISE integrators also contain the interface extraction component. A comprehensive solution to the search interface integration problem. Fully automated using only general (i.e., domain-independent) knowledge while most existing works employ manual or semi-automatic techniques. Solves a rarely addressed issue. More semantic relationships are needed for attribute matching and value merging. There is a need for human integrators to involve in integration process.

DeLa [12] is closely related to our method. But DeLa alignment method is purely based on the HTML tags; it uses only two types of relationship between the text node and data units where we use all type of relationships. Here DeLa uses only LIS interface for annotation process. The feasibility of heuristic-based automatic data annotation for web databases is provided. Information discovery problem is not defined. Simply labels are assigned to attributes of Tables.

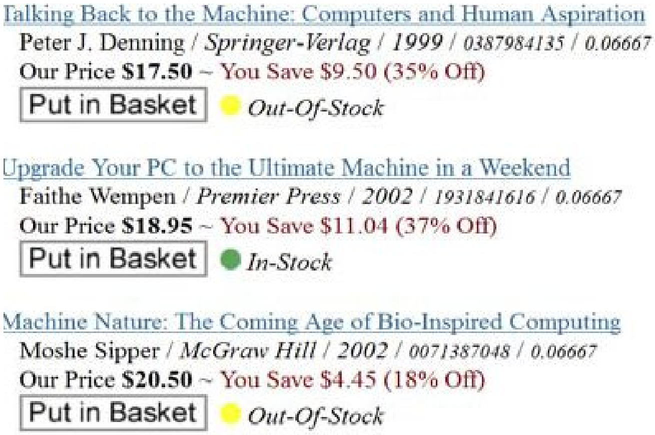
Y.Lu [5] describe about annotating the structured data of the deep web. It is similar to our method where in this paper they describe about four relationships between text node and data units but only two of them are briefly explained where in our method other two relationships also explained. Here we use clustering shift algorithm for one to nothing relationship where Y.Lu et al use pure clustering algorithm. Anyhow no frequent used web page in the annotated group is used for efficient output.

III. AUTOMATIC ANNOTATION AND WRAPPER GENERATION

Result page from web database has multiple records (SRR). Each SRR contain multiple data units each of which describes one aspects of real world entity. Our approach contains three phases. Alignment phase, Annotation phase and wrapper generation phase.

Text nodes are represented between the paired HTML tags. Tag nodes are surrounded by “<” and “>”, where text nodes are the text outside the “<” and “>”. Text nodes are the visible element on the web page and data units are located in the text node.

An example for search result is shown below with both original HTML page and the HTML source.



Talking Back to the Machine: Computers and Human Aspiration
Peter J. Denning / *Springer-Verlag* / *1999* / *0387984135* / *0.06667*
Our Price **$17.50** ~ You Save $9.50 (35% Off)
Put in Basket ● *Out-Of-Stock*

Upgrade Your PC to the Ultimate Machine in a Weekend
Faithe Wempen / *Premier Press* / *2002* / *1931841616* / *0.06667*
Our Price **$18.95** ~ You Save $11.04 (37% Off)
Put in Basket ● *In-Stock*

Machine Nature: The Coming Age of Bio-Inspired Computing
Moshe Sipper / *McGraw Hill* / *2002* / *0071387048* / *0.06667*
Our Price **$20.50** ~ You Save $4.45 (18% Off)
Put in Basket ● *Out-Of-Stock*

a. Original HTML page

ISSN: 2319-8753

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2014

```
<FORM><A> Talking Back to the Machine: Computers and
Human Aspiration</A><BR> Peter J. Denning / <FONT>
<I>Springer-Verlag / 1999 / 0387984135 / 0.06667</I>
</FONT><BR>Our Price<B>$17.50</B>~<FONT>You
Save $9.50 (35% off) </FONT><BR> <I>out-of-
Stock</I></FORM>
```

b. Simplified HTML source for the first SRR

Fig 1. Example search result from Bookpool.com

A. Alignment phase

In this phase the data units are identified and organized in to groups with different concepts. For this the relationship between the data units and text node are analyzed. The relationship may be One-to-one relationship, One-to-many relationship, Many-to-one relationship and One-to-nothing relationship. Relationship is analyzed, by identifying the features such as their data types (DT), data contents (DC), presentation styles (PS), and adjacency (AD) information.

After the relationship is analyzed. Alignment algorithm is used for aligning the data units. It consists of four steps. First the text node under same concepts is merged by removing the decorative tags. Second the text nodes need to be aligned. Third the data unit is to be identified by splitting the text nodes. At last the data units are to be aligned.

A clustering-based shifting technique used to align data units into different groups so that the data units inside the same group have the same semantic. We utilize the integrated interface schema (IIS) over multiple WDBs in the same domain to enhance data unit annotation.

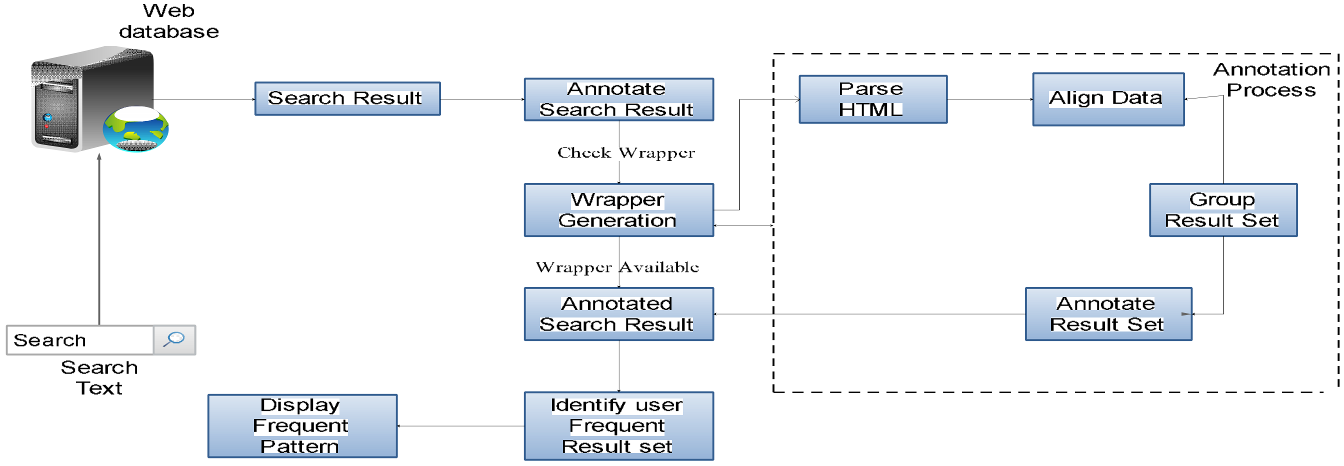


Fig 2. ARCHITECHTURE DIAGRAM

B. Annotation phase

In this phase we use six basic annotators; such as table annotator(TA), query-based annotator(QA), schema value annotator(SA),frequency-based annotator (FA), in-text prefix/suffix annotator(IA), common knowledge annotator(CA) are used to label the data unit group.

In table annotator the aligned data units are arranged in the table format and the column name is used to label the group. In schema value annotator uses the schema value such as publisher author and title for labelling. In frequency based annotator frequently available data units is used to label. In in text prefix suffix annotator the prefix or the suffix of the data units are used for labelling. Common knowledge annotator uses the basic knowledge for labelling.

Each annotator can independently assign labels to data units based on certain features of the data units. Moreover different annotator may produce different label for the obtain group of data unit.

Hence to select more suitable label for the group a probabilistic model is applied to combine the results from different annotators into a single label. It is highly flexible so that the existing basic annotators may be modified and new annotators may be added easily without affecting the operation of other annotators.

C. Wrapper generation

Annotation wrapper is a description of the annotation rules for all the attributes in the result page. After the annotation is completed wrapper is generated automatically for the annotated result group. The wrapper can be applied to efficiently annotating the SRRs retrieved from the same WDB with new queries.

D. Frequent Item set Generation

Also the frequently used websites are identified from the annotated group. So that the mostly used websites are known and they can be displayed first in the result page. This makes the process more efficient. Also it avoids unnecessary scan of the database.

The architecture diagram of the system is shown in the figure in which the search result is obtained from the web database. Before annotating the result it checks the wrapper whether it is annotated earlier or not. If not annotation process takes place and then wrapper is generated and the frequent available websites are identified and then the result is displayed.

IV. PERFORMANCE EVALUATION

The proposed system performance is evaluated on the basis of two factors that is precision and Recall. The precision and recall is calculated for performance of alignment and performance of annotation. The precision for performance of alignment is as follows.

$$precision = \frac{correctly\ aligned\ data\ units}{aligned\ data\ units} \times 100$$

$$Recall = \frac{data\ units\ that\ are\ correctly\ aligned}{manually\ aligned\ data\ units} \times 100$$

Table 1 represent the performance calculation for alignment in which the average value for precision and recall is about 98%. And for each domain it is more than 96%.

Domain	Alignment	
	Precision	Recall
Book	98.4	97.3
Game	98.7	98.0
Music	99.0	99.1
Average	98.7	98.1

Table 1
Performance of Alignment

The performance of each alignment features as mentioned in alignment phase is given below in which over all alignment give the best result than the individual one. Here the tag path gives accurate result next to overall result. That means while calculating individually tag path give more accurate result than other features.

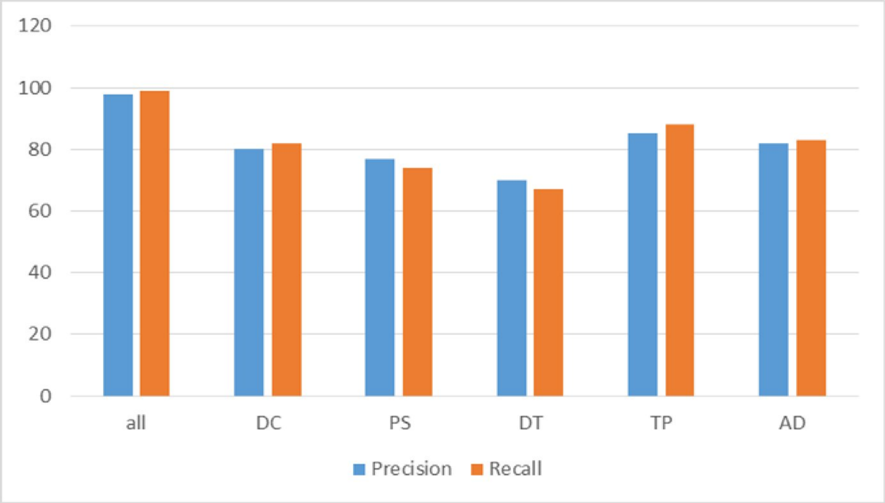


Fig 3. Performance of Alignment Features

The basic formula used to calculate precision and recall for annotation is as follows

$$precision = \frac{correctly\ annotated\ data\ units}{data\ units\ annotated} \times 100$$

$$Recall = \frac{data\ units\ correctly\ annotated}{manually\ annotated\ data\ units} \times 100$$

The table 2 shows the Performance of annotation face in which the average precision and recall is nearly 97%. And for each domain it results more than 95%.

Domain	Annotation	
	Precision	Recall
Book	97.4	96.3
Game	97.7	97.0
Music	97.0	97.7
Average	97.3	97.0

Table 2
Performance of Annotation

The performance of the basic annotator are compared and shown in the fig 4. The evaluation shows the combination of all Annotators give the most accurate result than finding each one individually. Comparing others table annotator gives nearly an accurate result.

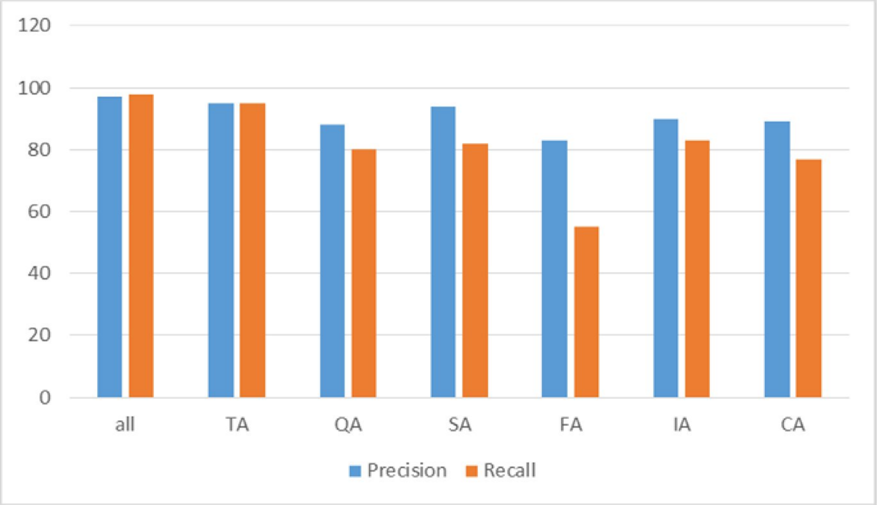


Fig 4. Performance of Basic Annotator

V. CONCLUSION

For the automatic annotation problem, a multi annotator approach is proposed which automatically construct an annotation wrapper for annotating the search result records retrieved from any given web database. In this approach six basic annotators were used and a probabilistic method to combine these basic annotators. Each annotator exploits one type of features for annotation. Each annotator results are useful and combination if these annotator are capable of generating high quality annotation. One of our main features is while annotating the results retrieved from the web database, it utilize both LIS of the web and the IIS of the multiple web databases in the same domain. IIS is used to reduce the local interface schema, inadequacy problem and the inconsistent label problem. In automatic aligned problem accurate alignment is critical to achieving holistic and accurate annotation. But by using a clustering based shifting method we obtain automatically obtainable features. This method is capable of handling variety of relationship between HTML nodes and data units such as one-to-one, one-to-many, many-to-one, one-to-nothing. By creating annotation wrapper makes the annotation easy for the new queries for the same WDB without performing alignment and annotation phase again. Using wrapper the annotation become efficient for even a new queries. Here we also use the frequent item set retrieval to know the result set which is more in annotator group. It is also used to list down the trusted sites in the data base.

REFERENCES

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

[2] 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.

[3] H. He, W. Meng, C. Yu, and Z. Wu, “Automatic Extraction of Dynamic Record Sections From Search Engine Result Pages,” VLDB J., vol. 13, no. 3, pp. 256-273, Sept. 2012.

[4] 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.

[5] 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

[6] L. Liu, C. Pu, and W. Han, “XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources,” Proc. IEEE 16th Int’l Conf. Data Eng. (ICDE), 2001.

[7] 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.

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

[9] 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.

[10] 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.

[11] N.Kruhmerick, D. Weld, and R.Do orenbos, “Wrapper Induction for Information Extraction”, Proc Int’1 joint conf.Artificial Intelligence (IJCAI), 1997.

[12] J. Wang and F.H. Lochovsky, “Data Extraction and Label Assignment for Web database”, proc. 12th int’l conf. World Wide Web (WWW), 2003.

[13] H. Zhao, W.Meng, C.Yu, “Mining Templates from Search Result Records of Search Engines”, proc. ACM SIGKDD int’l conf. Knowledge discovery and Data Mining, 2007.

[14] J.Wang and F.H. Lochovsky, and W.Ma, “Instance-Based Schema Matching For Web Database By Domain Specific Query Probing,” proc. Very Large Databases (VLDB) conf. 2004.

[15] L.Kaufman and P. Rousseau, Finding groups in Data: An Introduction to Cluster Analysis. John Wiley & sons, 1990.

[16] S. Dill et al., “SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation”, proc. 12th int’l conf. World Wide Web (WWW), 2003.

[17] S. Handschuh and S.Staab, “Authoring and Annotation of Web Pages in CREAM”, Proc. 11th int’l conf. World Wide Web (WWW), 2003.