ISSN:0975-9646

Mining Annotated Search Results From Web Databases

Thushara K P
Department of Computer Science and Engineering
Malabar Institute of Technology
Anjarakandy,Kannur, Kerala

Varsha Philip
Department of Computer Science and Engineering
Malabar Institute of Technology
Anjarakandy,Kannur,Kerala

***Abstract*—A web database is a system for storing information that can then be accessed via a website. The result page obtained from web database (WDB) is Search result record (SRR). These SRRs contain multiple data units which is to be extracted and labeled semantically. The Annotation approach contains three phases: alignment, annotation and annotation wrapper generation. In this paper, we present a Searching mechanism which perform searching from the annotated search result, frequently used data units are identified by Join-Based Algorithm which include top-K ranking function. The advantage of this approach is fast operation on dataset and provides facilities to avoid unnecessary scans to the database.**

***Keywords*-SRR, Annotation, WDB.**

I. Introduction

Data mining is the computational process of discovering patterns in large data sets the overall goal is to extract information from a data set and transform it into an understandable structure for further use. A web database contains multiple SRRs ,each SRR contain several data units. Data units represent the single concept of a real-world entity. Here annotation is performed on the basis of data units.

Fig. 1 represents three SRRs on a result page of book WDB.Each SRR have one book with several data units, e.g., the first book record in Fig. 1 has data units Talking Back to the Machine: Computers and Human Aspiration, Peter J. Denning, etc.Now-a-days there is an increasing demand for retrievig data from multiple WDBs. Early applications have limited scalability due to the requirement of immense human efforts to annotate data units manually. In our System, we consider how label can be assigned auto-matically to the data units within the SRRs returned from WDBs.

In our System,initially we perform annotation and then searching based on these annotated results. In annotation, during the first phase(alignment phase)all data units in the SRRs are identified and then aligned them into groups corresponding to same concept. In the second phase basic annotators are used to label data units.In the next phase an annotation wrapper(annotation rule)is generated, which shows how to extract data units of same concept for the corresponding web database. This annotation wrapper makes the annotation faster.

Then we perform join-based algorithm [8] for keyword search to retrieve frequently used data units from annotated results. Our proposed system has following contributions:
January 6, 2015

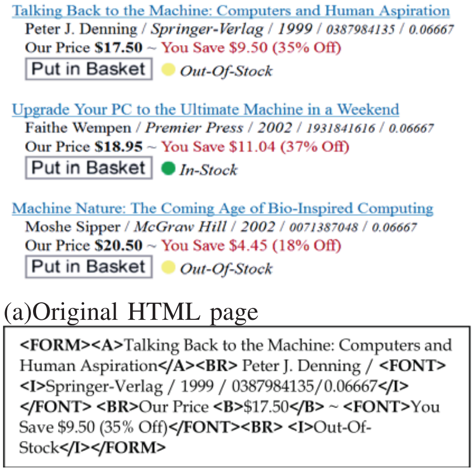
- Our System analyze the relationships between text nodes and data units and perform data unit level annotation. In most existing approaches labels are assigned to each HTML text node only.
- We put forward a clustering-based shifting technique for alignment of data units into groups of same semantic. The current methods use only the DOM tree or other HTML tag tree structures of the SRRs to align the data units. Our System considers some important features of data units:data types (DT), data contents (DC), presentation styles (PS),and adjacency (AD) information.
- To enhance data unit level annotation we use the integrated interface schema (IIS) over multiple WDBs in the same domain.
- We make use of six basic annotators; each can be used to assign labels independently to data units based on certain characteristic of the data units. We combine the results of annotators into a single label.So it is easy to change existing basic annotators and add new annotators without affecting the operation of other annotators.Thus this model is highly flexible.
- We generate an annotation wrapper [1] for given WDB. The wrapper can be applied to annotate the SRRs retrieved from the WDB with new queries.
- We propose a Join-Based Algorithm for searching frequent data unit in SRRs.

The rest of the paper is organized as follows:Section 2 describes related work. Section 3 introduce data alignment and wrapper generation process. Section 4 describes performance evaluation of our work and section 5 concludes the paper.

II. Related Work

Web information extraction and annotation is an active research area.The early system Wrapper Induction for Information Extraction [5] depend on human efforts for extracting data from a particular information source. A wrapper is a procedure that is specific to a single information resource. These are applicable to tabular pages and uses positions of particular strings.

The methods for extracting structured data from web pages [2] only extract database values from template generated



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

```
<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 results from Bookpool.com [1]

web page,does not consider annotation. During first stage it recognize token set associated with the same type constructor in the template which are used to create the input pages. In the Analysis stage it uses the above sets to deduce the template. This template is then used to extract the values.

The effort to automatically assign labels that is Automatic Annotation of Data Extracted from Large Web Sites [3], depends on important information about semantics of data, and assumes these are available on webpages. This approach assumes that published data are accompanied by textual description to help human user. Applicability is limited because many WDBs encode the data units without their labels.

One of the method for spliting SRR is Harvesting Relational Tables from Lists on the Web [4]. Webpages contain data structured in ”lists” these can be split into multiple coulmns. It first split the individual lines into multiple fieds and construct a table by determining single number of likely column in output table. Records having too many fields are remerged and resplit. Record with too few fields are expanded by inserting null values.

The DeLa [6] most similar to our work.The alignment process of this approach depends only on HTML tags. It uses a regular expression based gata tree algorithm for alignent purpose. It sends querry through HTML forms,automatically generate regular expression wrapper to extract data objects from result page and restore retrieved data into an annotated table.Main disadvantage is label set is predefined so only small number of values are available.

The effort for automatic wrapper generation, Fully Automatic Wrapper Generation For Search Engines [7],focuses on how to extract search result records from dynamically generated result pages returned in response to submitted queries.It uses visual content feature on the result page as displayed on a web browser and HTML tag structure of HTML source file of result page.Each SRR is stored in a tree structure.

More efficient automatic annotation, Annotating Search Results from Web Databases [1]in which a clustering-based shifting technique is used to align data units and an annotation wrapper is generated which describes how to extract data units and what semantic label should use.

Our data alignment approach differs from the previous works. We first handles all types of relationships between text nodes and data units such as

One-to-One Relationship: In this, each text node contains exactly one data unit.We refer to such type of text nodes as atomic text nodes.

One-to-Many Relationship: In this multiple data units are encoded in one text node.Such kind of nodes are known as composite text node.

Many-to-One Relationship: In this multiple text nodes together form a data unit.

One-To-Nothing Relationship: The text node belongs to this category will be displayed in a certain pattern across all SRRs.These are called template text nodes.

The existing approaches consider only one-to-one, one-to-many relationships. Second, we use a variety of features together. Third, we launch a new clustering-based shifting algorithm to perform alignment. Fourth, we use a join based algorithm for performing keyword search.

The effort for efficient keyword search, Join-Based Algorithms for Keyword Search in XML Databases [9] in which a join-based algorithm is used for scanning all SRRs and used a top-k algorithm for ranking function. In this approach a XML tree structure is used.

III. Data Alignment and Wrapper Generation

Each SRR has a tag structure which determines how the contents of the SRRs are displayed on a web browser. Each node in the tag structure is a tag node or a text node. A tag node is an HTML tag surrounded by ”<” and ”>” and a text node is the text outside the < and >. Text nodes are the visible and it contain data units.From Fig. 1,text nodes are not always identical to data units.

A. Alignment

The SRRs on a result page will be in a table format where each row represents one SRR and each cell carries a data unit. Each column of the table is known as alignment group, which contains at most one data unit from each SRR. A well-aligned group is an alignment group which contains all the data units of same concept and no data unit of other concepts. The aim of alignment is to shift the data units in the table to make every alignment group well aligned.The order of the data units within every SRR will be preserved.

Alignment process consist of following steps:

Step 1: Merge text nodes: This step merges ext nodes corresponding to the same concept into a single text node by detecting and removing decorative tags from each SRR.

Step 2: Align text nodes:In this steptext nodes are aligned into groups so that each group contains the text nodes with the same attribute.

Step 3: Split (composite) text nodes: This step splits the values in composite text nodes into individual data units. This step is carried out based on the text nodes in the same group.

Step 4: Align data units: This step separates each composite group into multiple aligned groups which contains the data units of the same concept.

B. Annotation

The data units of the same concept often share some common features,and these common features are associated with the data units.we utilize six basic annotators to label data units:

1) Table Annotator (TA) The SRRs will be in a table. Initially, it identifies all the column headers of the table.Then, it takes a data unit in a cell and selects the column header whose area has the maximum vertical overlap with the cell. This unit is then assigned with this column header and labeled by the header text. The remaining data units will be processed similarly.

2) Query-Based Annotator (QA) On the local search interface a query with a set of query terms are submitted against an attribute A, first it finds the group that has the largest total occurrences of these query terms and then assign global name of (A) as the label to the group.

3) Schema Value Annotator (SA) The schema value annotator is used to discover the best matched attribute to the group from the IIS. The schema value annotator first identifies the attribute that has the highest matching score among all attributes and then uses global name of that attribute to annotate the group.

4) Frequency-Based Annotator (FA) The frequency-based annotator finds common preceding units shared by all the data units of the group. This is found easily by following their preceding chains recursively until the met data units are different.To form the label for the group all found preceding units are concatenated.

5) In-Text Prefix/Suffix Annotator (IA) The in-text prefix/suffix annotator checks if all data units in the aligned group have the same prefix or suffix. If there is same prefix and it is not a delimiter, then it is removed from the group and is used to annotate values following it. If the same suffix is identified and if the number of data units having the same suffix and the number of data units inside the next group matches, the suffix is used to annotate the data units inside the next group.

6) Common Knowledge Annotator (CA) Human users can understand some data units on the result page because of the common knowledge. Common concept contains a label and a set of patterns or values these can be used for annotation.
Applicabilities and Success Rates of Annotators are shown in Figure 2.

TABLE I. APPLICABILITIES AND SUCCESS RATES OF ANNOTATORS [1]

	Applicability	Success Rate
Table Annotator	6%	1.0
Query-Based Annotator	35%	0.95
Schema Value Annotator	14%	0.5
Frequency-Based Annotator	41%	0.86
In Text Prefi/Suffix Annotator	7%	0.85
Common Knowledge Annotator	25%	0.84

C. Annotation Wrapper

Once the data units are annotated,we generate an annotation wrapper by using these annotated data units. The new SRRs retrieved from the same WDB can be annotated using this wrapper quickly without repeating the entire annotation process. This annotation wrapper describes how to extract data units and what should be the label.

To annotate a new result page the wrapper can be usde,the annotation rules are applied on each data unit in an SRR, one by one depending on the order in which they appear in the wrapper.The rule is matched if this data unit has the same prefix and suffix as stated in the rule, and then unit is labeled with the given label in the rule.

D. Frequent Data Unit set Generation

In this we perform a keyword searching by using join based algorithm. The important feature of keyword search in semi-structured and structured data is that keywords can spread over multiple elements or tuples in the databases. The result of the query will be a set of elements or tuples that contain all the keywords. Here we assumes a XML tree structure.

Nodes in the XML tree can be represented by Id, e.g. 1.1.1 for Title node. A stack is used to process all the nodes in the document order.Nodes are pushed into the stack one by one. When some node v is to be pushed,all the nodes in the stack that are not ancestors of v are popped out. Joins are performed column by column, within each list.A ranking function is used to rank each constraints and then the joining is performed on the basis of this score and newly generated result is put into result set.

IV. Performance Evaluation

To evaluate the performance of our methods We use the precision and recall measures from information retrieval.

For alignment, the precision is defined as the percentage of the data units which are correctly aligned over all the aligned units by the system. Recall is the percentage of the data units that are correctly aligned by the system over all data units which are manually aligned by the expert. A result data unit is taken as "incorrect if it is mistakenly extracted for e.g. failed to be split from composite text node. The performance calculation for alignment in which the average value for precision and recall is about 98%.

$$precision = \frac{correctly\ aligned\ data\ units}{aligned\ data\ units} * 100 \quad (1)$$

$$recall = \frac{data\ unit\ that\ are\ correctly\ aligned}{manually\ aligned\ data\ units} * 100 \quad (2)$$

For annotation, the precision is defined as the percentage of correctly annotated units over all the data units annotated by the system. The recall is the percentage of the data units correctly annotated by the system over all the manually annotated data units. A data unit is known as correctly annotated if its system-assigned label has the same meaning as its manually assigned label. the Performance of annotation face in which the average precision and recall is nearly 97%.

$$precision = \frac{correctly\ annotated\ data\ units}{data\ units\ annotated} * 100 \quad (3)$$

$$recal = \frac{data\ units\ that\ are\ correctly\ annotated}{manually\ annotated\ data\ units} * 100 \quad (4)$$

In top-K algorithm has a great advantage in terms of space because core operation of this algorithm is hash join, and thus does not require any additional index. The other algorithms like index-based algorithm is large and uses a single B-tree and RDIL, requires both inverted lists and B+-tree. The execution time for index-based algorithm provides a cache mechanism. Our implementation relies on the file system cache directly. The stack-based algorithm,does not use a lot of cache, because it always needs to scan all the input lists.

V. Conclusion

In this paper, we studied the data annotation problem. To solve the automatic annotation problem, we proposed a multi annotator approach which automatically construct an annotation wrapper for annotating the SRRs retrieved from any given web database. Our approach uses six basic annotators,which exploits one type of features for annotation. One of our important features is we utilize both LIS and IIS of the multiple web databases in the same domain during annotating the results retrieved from the web database. The IIS is used to reduce the local interface schema, inadequacy problem and the inconsistent label problem. By using a clustering based shifting method we obtain automatically obtainable features. The method is able to handle 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, the annotation made easy because it describes how to extract data units and which label should be used. So no need of repeating alignment and annotation phases. By Using the wrapper the annotation become efficient for even a new queries. Here we also use the frequent data unit set retrieval. We can perform a keyword based search. We utilize join based top-k algorithm for this purpose. This algorithm makes more efficient and fast searching of annotated search results from web databases.

Acknowledgment

The authors would like to express their gratitude to all who have taken particular interest on our work and supported throughtout. The authors would also like to express their gratitude to the anonymous reviewers for providing very constructive suggestions to improve the manuscript.

References

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu "Annotating Search Results from Web Databases"

[2] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages" *Proc. SIGMOD Intl Conf. Management of Data, 2003.*

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

[4] H. Elmeleegy, J. Madhavan, and A. Halevy, Harvesting Relational Tables from Lists on the Web,*Proc. Very Large Databases (VLDB) Conf., 2009.*

[5] N. Krushmerick, D. Weld, and R. Doorenbos, Wrapper Induction for Information Extraction,*Proc. Intl Joint Conf. Artificial Intelligence (IJCAI), 1997.*

[6] J. Wang and F. H.Lochovsky, Data Extraction and Label Assignment for Web Databases,*Proc. 12th Intl Conf. World Wide Web (WWW), 2003.*

[7] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, Fully Automatic Wrapper Generation for Search Engines,*Proc. Intl Conf. World Wide Web (WWW), 2005.*

[8] V. Yogam, K. Umamaheswari "Automatic Annotation Wrapper Generation and Mining Web Database Search Result"

[9] Liang Chen ,Yannis Papakonstantinou "Join-Based Algorithms for Keyword Search in XML Databases"