www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 03 Issue 07 July, 2014 Page No. 6921-6925

A New Technique Called Annotation Used To Categories The Content Of Web Pages

U. Madhavi, S. Hrushikesava Raju , P. Nirupama
M.Tech: Department of CSE SIETK, Puttur, INDIA
madhaviusurupati@gmail.com
Associate Professor
Department of CSE SIETK, Puttur, INDIA
hkesavaraju@gmail.com
Head of the department
Department of CSE SIETK, Puttur, INDIA

Abstract- Now-a-days databases become web accessible; these databases having data units are encoded. To separate data units assign meaningful labels. To assign labels there is an automatic annotation approach which automatically assign labels to data units within the search result records returned from web databases. For that it provides an annotation wrapper for the search site to automatically constructed and annotate the new result pages from the same web databases. There are six basic annotators, for every basic annotator we produce label for the data units within their group. A probability model is selected to determine the most appropriate label for each group. This annotation wrapper generate an annotation rule that describes how to extract the data units from result page .Once the annotation wrapper annotate the data there is unnecessary to perform the alignment and annotation phases again. We construct an annotation wrapper by using data types, data contents, presentation styles and adjacency information. In this paper the data annotation problem is studied and proposed a multiannotator approach to automatically constructing an annotation wrapper for annotating the search result records recover from any given web databases.

Keywords: Annotator, web databases, wrapper

I. INTRODUCTION

A large portion of the web is database based for many search engines, data encoded in the returned result pages come from the underlying structured databases. Such type of search engines is often referred as web databases (WDB). A typical result page returned from a WDB has multiple search result records (SRRs). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. There are three phases to align the data. First phase includes alignment phase it first identify all data units in the SRRs and then organize them into different groups with each group corresponding to a different concept. Second phase includes annotation phase it introduce multiple basic annotators with each exploiting one type of features. Every basic annotator is used to produce a label for the units within their group holistically, and a probability model is adopted to determine the most appropriate label for each group. Third phase includes annotation wrapper generation phase it describes how to extract the data units of this concept in the result page and what the appropriate semantic label should be. There are some contributions like:

1. Most existing approaches simply assign labels to each HTML text node, by having this we thoroughly analyse relationships between text nodes and data units.
2. We propose some important features like data units, data contents, presentation styles and adjacency information.
3. We use six basic annotators to combine the results from different annotators into a single label.

II. RELATED WORK

Web information extraction and annotation has been an active research area in recent years. Many systems rely on human users to mark the desired information on sample pages and label the marked data at the same time, and then the system can induce a series of rules (wrapper) to extract the same set of information on webpages from the same source. These systems are referred as a wrapper induction system. Because of the supervised training and learning process, these systems can usually achieve high extraction accuracy. These systems suffer from poor scalability and are not suitable for applications that need to extract information from a large number of web sources. For these problems we have some solutions in order to extract correct information.

One of the problems is finding out the exact information from the web. Browsing is not suitable for locating particular items of data because it is tedious, and it is easy to get lost. Furthermore, browsing is not cost-effective as users have to read the documents to find desired data. Keyword searching is sometimes more efficient than browsing but often returns vast amounts of data, far beyond what the user can handle. So, Embley [1] utilize ontologies together with several heuristics to automatically extract data in multi record documents and label them. Ontologies for different domains must be constructed manually. A document contains multiple records for ontology if there is a sequence of chunks of information about the main entity in ontology. Specifically, this approach consists of the following five steps.

(1) Develop an ontological model instance over an area of interest.

(2) Parse this ontology to generate a database scheme and to generate rules for matching constants and keywords.

(3) To obtain data from the Web, invoke a record extractor that divides an unstructured Web document into individual record-size chunks, cleans them by removing mark-up-language tags, and presents them as individual unstructured record documents for further processing.

(4) Invoke recognizers that use the matching rules generated by the parser to extract from the cleaned individual unstructured documents the objects expected to populate the model instance.

(5) Finally, populate the generated database scheme by using heuristics to determine which constants populate which records in the database scheme. These heuristics correlate extracted keywords with extracted constants and use relationship sets and cardinality constraints in the ontology to determine how to construct records and insert them into the database scheme. Once the data is extracted, they can query the structure using a standard database query language.

The efforts to automatically construct wrappers are extracting structured data from web pages, towards automatic data extraction from large websites and a vision based approach for deep web data extraction, but the wrappers are used for data extraction only. These aim to automatically assign meaningful labels to the data units in search result records. Arlotta [2] basically annotate data units with the closest labels on result pages. Data extraction from web pages is performed by software modules called wrappers. Recently, some systems automatically generate the wrappers. These systems are based on unsupervised inference techniques: taking as input a small set of sample pages, they can produce a common wrapper to extract relevant data. However, due to the automatic nature of the approach, the data extracted by these wrappers have anonymous names. In this framework the ongoing project Roadrunner have developed a prototype, called Labeller that automatically annotates data extracted by automatically generated wrappers. Although Labeller has been developed as a companion system to old wrapper generator, its underlying approach has a general validity and therefore it can be applied together with other wrapper generator systems. The experimented prototype over several real-life web sites obtaining encouraging results. They analyzed about 50 automatically generated wrappers that work on pages from several web sites: a large majority of the data extracted by the wrappers are accompanied with a string representing a meaningful name of the value. The domain ontology is then used to assign labels to each data unit on result page. After labelling, the data values with the same label are naturally aligned.

W. Liu, X. Meng, and W. Meng [3] align data units and annotate the ones within the same semantic group holistically. Data alignment is an important step in achieving accurate annotation and it is also used in automatic annotation of html documents .Most existing automatic data alignment techniques are based on one or very few features. The most frequently used feature is HTML tag paths (TP). ViDE (A Vision-Based Approach for Deep Web Data Extraction) uses visual features on result pages to perform alignment and it also generates an alignment wrapper. But its alignment is only at text node level, not data unit level. The method in harvesting relational tables from the lists on the web first splits each SRR into text segments. The most common number of segments is determined to be the number of aligned columns (attributes).

J.wang and F.H.lochovsky [4] is the most similar to Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng work. But Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng approach is significantly different from DeLa’s approach. First, J.wang alignment method is purely based on HTML tags, while Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng uses other important features such as data type, text content, and adjacency information. Second, their method handles all types of relationships between text nodes and data units, whereas J.wang deals with only two of them (i.e., one-to-one and one-to-many). Third, J.wang and their approach utilize different search interfaces of WDBs for annotation. They uses an IIS of multiple WDBs in the same domain, whereas J.wang uses only the local interface schema (LIS) of each individual WDB. Their analysis shows that utilizing IISs has several benefits, including significantly alleviating the local interface schema inadequacy problem and the inconsistent label problem .Fourth, they significantly enhanced J.wang annotation method. Specifically, among the six basic annotators in their method, two (i.e., schema value annotator (SA) and frequency-based annotator (FA)) are new, three (table annotator (TA), query-based annotator (QA) and common knowledge annotator (CA)) have better implementations than the corresponding annotation heuristics in J.wang, and one (in-text prefix/suffix annotator (IA)) is the same as a heuristic in J.wang.

Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng [5] proposes the development of search engines databases through web reaches all the way through HTML based search boundary. Now day’s analysis of data in deep manner from database or web search engines also important to return exact information in search result web pages. In generally the data units received from web accessible search engine databases are frequently prearranged into the result pages energetically for individual browsing. Here, consideration of automatic data assignment for SearchResultRecord pages returned from original web search engine databases. To conquer these problems proposed an automatic semantic annotation approach through semantic similarity measure for data units and text unit’s results from features for Search results records. From search results records important feature are extracted and then semantic similarity based measurement are measures are performed to each and every data, text unit nodes. Ontology based system measures semantic similarity between terms in the pages and then aligns the data units in efficient manner. In this work provide the proficiently analysis of the data and most excellent alignment of Search Result Records. To annotation of new search result from web search engines for various domains in databases use annotation wrapper. The basic experimentation results are estimated based on the parameters like precision and recall for various topics.

III. BASIC ANNOTATORS

1. Table annotator

Web databases used to organise in tables. In the table each row represents an search result record .Data units of the same concepts are aligned with its corresponding common header. Working of table annotator is as follows: first it identifies all the column headers of the table. Second for each search result record it takes a data unit in a cell and selects the column header whose area has the maximum vertical overlap with the cell. We use HTML tag <TH> and <THEAD> for table annotator.

2. Query-Based annotator

This annotator defines that the returned search result record from a web database are always related to a specific query. The working of query based annotator is as follows: Given a query with a set of query terms submitted against an attribute A on local search interface, first find the group that has the largest total occurrences of these query terms and then assign gn(A) as label to the group.

3. Schema value annotator

The schema value annotator first identifies the attribute Aj that has the highest matching score among all attributes and then use gn (Aj) to annotate the group Gi.

4. Frequency based annotator

The frequency based annotator intends to find common preceding units shared by all data units of the group Gi. All founded preceding units are concatenated to form the label for the group Gi.

5. In-Text prefix/suffix annotator

The in-text prefix/suffix annotator checks whether all data units in the aligned group share the same prefix or suffix. If the same prefix is confirmed and it is not a delimiter, then it is removed from all the data units in the group and is used as the label to annotate values in it. If the same suffix is identified and if the number of data units having the same suffix is used to annotate the data units inside the next group.

6. Common Knowledge annotator

The common knowledge annotator considers both patterns and certain value sets as the set of countries. First the common concepts are domain dependent. second, they can be obtained from existing information resources with little additional human effort.

IV. ALGORITHM

Alignment Algorithm:

Data alignment algorithm is based on the assumption that attributes appear in the same order across all SRRs on the same result page, although the SRRs may contain different sets of attributes. Each table column, in this work, is referred to as an alignment group, containing at most one data unit from each SRR. If an alignment group contains all the data units of one concept and no data unit from other concepts, this group is called well aligned. The goal of alignment is to move the data units in the table so that every alignment group is well aligned, while the order of the data units within every SRR is preserved.

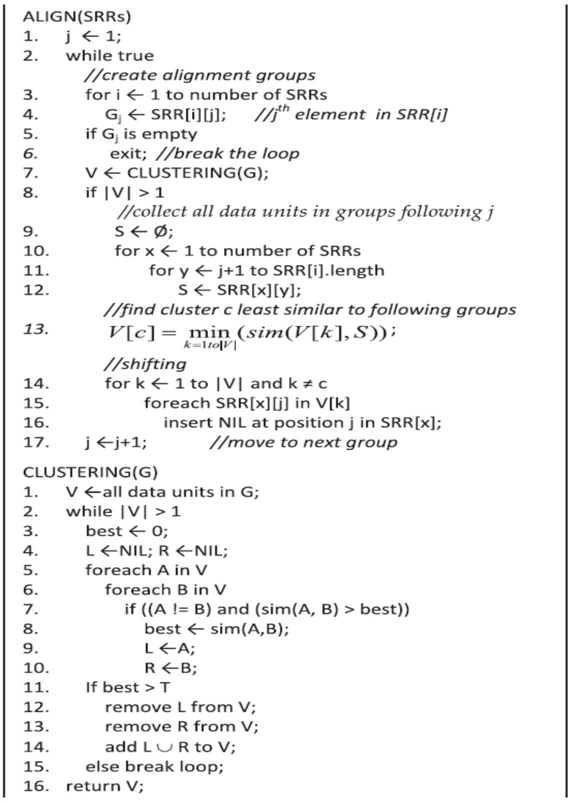
Data alignment method consists of the following four steps. The detail of each step is described below:

Step 1: Merge text nodes. This step detects and removes decorative tags from each SRR to allow the text nodes corresponding to the same attribute (separated by decorative tags) to be merged into a single text node.

Step 2: Align text nodes. This step aligns text nodes into groups so that eventually each group contains the text nodes with the same concept (for atomic nodes) or the same set of concepts (for composite nodes).

Step 3: Split (composite) text nodes. This step aims to split the “values” in composite text nodes into individual data units. This step is carried out based on the text nodes in the same group holistically. A group whose “values” need to be split is called a composite group.

Step 4: Align data units. This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept.



```
ALIGN(SRRs)
j ← 1;
while true
    //create alignment groups
    for i ← 1 to number of SRRs
        $G_j$ ← SRR[i][j];     //$j^{th}$ element  in SRR[i]
    if $G_j$ is empty
        exit; //break the loop
    V ← CLUSTERING(G);
    if |V| > 1
        //collect all data units in groups following j
        S ← Ø;
        for x ← 1 to number of SRRs
            for y ← j+1 to SRR[i].length
                S ← SRR[x][y];
        //find cluster c least similar to following groups
        $V[c] = \min_{k=1 to|V|}(sim(V[k], S))$;
        //shifting
        for k ← 1 to |V| and k ≠ c
            foreach SRR[x][j] in V[k]
                insert NIL at position j in SRR[x];
    j ←j+1;          //move to next group

CLUSTERING(G)
V ←all data units in G;
while |V| > 1
    best ← 0;
    L ←NIL; R ←NIL;
    foreach A in V
        foreach B in V
            if ((A != B) and (sim(A, B) > best))
                best ← sim(A,B);
                L ←A;
                R ←B;
    If best > T
        remove L from V;
        remove R from V;
        add L ∪ R to V;
    else break loop;
return V;
```

V.CONCLUSION

The features of data and text units are obtained from Particle Swarm Optimization (PSO) methods. From search results records important features are extracted and then semantic similarity based measurement are measures are performed to each and every data, text unit nodes. Ontology based system measures semantic similarity between terms in the pages and then aligns the data units in efficient manner**.** In this work we proficiently analysis the data and most excellent alignment of SRR records.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] J. Wang and F.H. Lochovsky, “Data Extraction and Label Assignment for Web Databases,” Proc. 12th Int’l Conf. World Wide Web (WWW), 2003.

[5] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE, and Clement Yu, Senior Member, IEEE –“Annotating search results from webdatabases” IEEE transactions on knowledge and data engineering, vol. 25, no. 3, march 2013.