International Journal of Engineering Research-Online
A Peer Reviewed International Journal
Articles available online http://www.ijoer.in

Vol.3., Issue.2, 2015

REVIEW ARTICLE

ISSN: 2321-7758

AUTOMATIC ANNOTATION WRAPPER GENERATOR

LEKSHMI S.S[1], SURYAPRIYA S[2]

[1]PG Scholar, Sarabhai Institute of Science &Technology, vellanadu, Kerala, India.
[2]Assistant professor, Sarabhai Institute of Science &Technology, vellanadu, Kerala, India.

Article Received: 17/03/2015 | Article Revised on:01/04/2015 | Article Accepted on:04/04/2015

LEKSHMI S.S

ABSTRACT

The HTML form-based interfaces make a large number of database web accessible. Whenever a query is given to the search engine, data units are retrieved from the available databases. These data units must be extracted and they should be assigned with meaningful labels for the effective use of machine processable applications. Multi-annotator approach is focused in this paper which first aligns the data units in the result page into groups of similar semantic, after that various annotations are performed for each group and these annotations are combined to predict the final annotation. The construction of annotation wrapper is completely automatic. This wrapper can be used for the new queries.

Key Words: Data extraction, Web data annotation and Wrapper induction, Data Extraction, Data annotation, Annotators, Text nodes, Data Units and Wrapper

1. INTRODUCTION

The search engine that fetches the results from the underlying structured databases and displays in the result page will be referred to as the Web Databases (WDB) in this paper. A result page corresponding to a WDB contains multiple search result records (SRR). Each SRR consists of different data units (or instances). Each data unit refers to a single concept of an entity. A text node contains a piece of text surrounded by a pair of HTML tags. It is different from the data units referred in this paper. This paper focuses on the data unit level annotation. Annotating data units refers to assigning meaningful labels. Annotation of web pages is necessary for applications such as comparison book shopping, deep web collection etc. Also annotation is essential for easier storage of data into tables and for quick retrieval or mining of data. The result page consists of many SRRs. Once the search result records have been extracted from the result page, through three different phases annotation has been performed. The first phase is the *alignment phase* which organizes the data units into groups of the same semantic. The second phase is the *annotation phase* which performs different types of annotations for the assignment of labels. The third phase is the *annotation wrapper generation phase* which constructs annotation wrapper. Set of rules for all

aligned groups of data is the wrapper. This wrapper can be used for the new queries without repeating the whole process again.

2. LITERATURE SURVEY

The World Wide Web is having vital data in numerous formats the users have to deal with this data by using a search based form. The user will retrieve the information by firing the query. In traditional approach the search base form is design to fire the queries & required data is fetched. HTML form is containing the plain text. Querying, Integration etc. are used. These techniques are not effective to produce accurate search result record from web databases, because of human involvement and poor quality of the data extraction output. Two main problems arise during extracting the relevant information First: to categorized the unstructured view of data such as search engine. Second: categorized structure and semi-structure view of data. The web sites are also having heterogeneous nature due to language independent. The e commerce website or the information portals are updating their content on a regular basic. The web data is now machined process able so, we require the relevant information extraction with the semantic grouping. The semantic grouping means the data with similar meaning can form group with same concept. XML/RDF has been widely used for representing semantic web that required annotation for recognition of semantic web. These techniques provide manual mapping of unlabeled document segment to ontological concepts. In bootstrapping semantic labeling is addressed in semantic web annotation. The presentation style & spatial locality in the HTML tag is focused [3].The sites like educational, news portal and e-commerce are dynamically update contents on a regular basis so called as content-rich web sites contents management software that creates HTML pages by populating templates from databases. The structural analysis technique is used to group together related elements in a HTML pages into unlabeled tree. The algorithm can use the hand-labeled concept instances from HTLM pages for identification of unlabeled concept instances in HTML pages and assigns semantic labels to them. Hand-crafted ontology is not used in this algorithm. For determining the consistency in presentation style we can use the feature extraction. So the data belong to same concept or set of concepts lie under similar group.

3. TYPES OF ANNOTATORS

The returned result page contains many SRRs. The data units corresponding to the similar concept (attribute) often share special common features in certain patterns. Based on this, in this paper we used the six basic annotators have been defined to label data units, where each of them considers a special type of patterns/features. Each annotator are play unique role in labeling the name to the data units are extracted by the wrapper. Four of these annotators (i.e., table annotator, query-based annotator, in text prefix/suffix annotator, and common knowledge annotator) are similar to the annotation heuristics used by DeLa but there different implementations for three of them (i.e., table annotator, query-based annotator, and common knowledge annotator) [1] [6] [2].

3.1 Table Annotator

The resulted page fetch from multiple website consist of various SRR. These information can be stored in the form of table .A table consist of different column header &rows. The cell of this table indicates the data unit. We can store the multiple data units. The table annotator used in Dela [2] Approach mainly focus on the <TD> tag elements. The information stored in <TD>elements is stored in the annotator table. But few websites contain the <TD> tag elements. So the table annotator is modified .The row is considered as SRR & the column is considered as attribute. The data unit having same features can be aligned under header & the column header. By considering the special feature we can annotate the SRR. Firstly we have to identify all the values of column then as per SRR we have to fill the data. In such way the limitation of Dela [2] is improved.

3.2 Query-Based Annotator

The SRR is always returned from WDB on the basis of fired query. When the user submits the data in the text box or select field from the list box on the search form, the query is fired on the WDB. Then the SRR is identified & the data is stored under the column header. The no of occurrences of matching the column header will decide the group & we can label it. The Dela uses only the local labels in the query. However, DeLa uses only local schema element names, not element names in the IIS [2].so,

the new approach is use to utilize the global schema.

3.3 Schema Value Annotator

Many attributes on a search interface have predefined values on the interface. More attributes in the IIS tend to have predefined values and these attributes are likely to have more such values than those in LIS. When values from different LIS are integrated then we have to modify the schema values to perform annotation.

3.4 Frequency Based Annotator

The adjacent units have different occurrence frequencies. The data units are always associated with the higher frequency & lower frequency. The attribute names are the higher frequencies, while the data units with the lower frequency most probably come from databases as embedded values. Suppose there is a group of lower frequency then we can easily find its preceding values shared by all data units in the group .We can analysis the data unit until it is different & map its preceding. Then we can combine the preceding to form the label.

3.5 In-Text Prefix/Suffix Annotator

In some cases, the data unit is aligned with its label. The data unit consists of the comma separated vales & the labels associated with it. These lie in a particular sequence separated from each other in all multiple SRR. After alignment it will form a group. The in text prefix/suffix will check for data unit. If the same prefix is there &not a deliminator then it is removed from all data units but if the number of data nodes match with the same suffix to the data node within next group then the suffix is used for the annotation. Any group whose data unit texts are completely identical is not considered by this annotator.

3.6 Common Knowledge Annotator

Some data units on the result page are self-explanatory because of the common knowledge shared by human beings For example, “in stock” and “out of stock” occurs in many SRRs from e-commerce sites. Human users understand that it is about the availability of the product because this is common knowledge. Each common concept contains a label and a set of patterns or values.

4. WRAPPER GENERATION

We use the tree alignment methods to calculate the similarity between input web pages and build a wrapper on tree alignment results. The tree alignment method is also used to calculate the similarity between wrapper and the input web results. The input trees are merged into one union tree whose nodes record the statistical information such as the times a node has been aligned, the text length of the node. A heuristic method is employed to find the most probable content block. The alignment algorithm is utilized again to detect the repeating patterns on the union tree. The wrapper is generated based on the most probable content block and the repeating patterns. A similarity series was built by calculating the similarity between the input web pages and the current wrapper using the tree alignment algorithm. The similarity series is in the order of the input web pages' timestamp.Change point detection and wrapper regeneration. A log likelihood ratio test is utilized to detect the change points on the similarity series. The wrapper generation method is applied again to generate a wrapper once a change point is detected. In this work, we focus on setting the weight (cost) of different node mapping (tag-matching). One of the major contributions of our work is a kind of linear regression method for getting the weight of different tag-matching.

Automatically getting tag-matching weight

The main problem of the previous method is that they did not consider about employing different weights for various tag-matching. For example, The block elements are elements that usually, but not always, contain other elements. They normally act as containers of some sort. The inline elements normally mark up the semantic meaning of something. Furthermore, the level of the different nodes should also be considered. The higher-level nodes should have higher weight as the higher-level nodes usually act as bigger structure block. Different weight should be assigned to different type of tag-matching.

In this study, a kind of linear regression method is employed to get the weight of different tag-matching. First, we found a collection of similar web pages belong to the same "class". It's feasible to get this kind of web pages collection automatically. Next, we will use this web pages collection for getting the optimal weighting schema. Let w_i be the weight of tag-matching and $w_i > w_j$ for $i < j$. Let D_{mn} be the sum of the gains in the best alignment between the trees T_m and T_n.

$$D_{mn} = \sum_i w_i t_i^{mn}$$

(1) Where t_i^{mn} is the number of wi occur in the alignment procedure.

(2) The sum of the gains in the collection is:

$$f = \sum_{m,n} D_{mn} = \sum_{m,n} \sum_i w_i t_i^{mn} = \sum_i w_i \sum_{m,n} t_i^{mn}$$

Because the collection is the similar web pages belonging to the same "class", a set of w_i is selected which makes the maximum f.

To get $argmax_w \sum_i w_i \sum_{m,n} t_i^{mn}$, a constraint $\sum_i w_i^2 = 1$ is added.

The group of equations is rewritten as:

$$f = \sum_i w_i C_i + \lambda\left(\sum_i w_i^2 - 1\right), \quad C_i = \sum_{m,n} t_i^{mn}, \quad \sum_i w_i^2 = 1$$

The solution of the above equations is used as the weight of each type of tag-matching (wi).

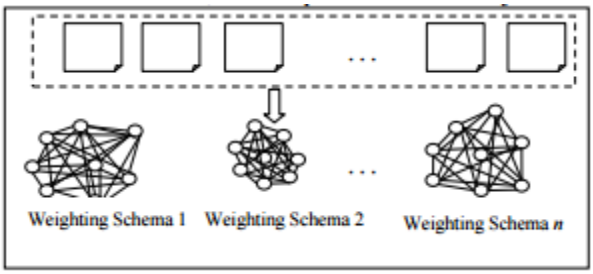

Fig: Weighted sum method

The figure illustrates an example of the weight setting method. For one collection of similar web pages belong to the same "class", we calculate the sum of the alignment gains (or the similarity) for each weighting schema. The best weighting schema is the one maximize the sum of the gains. That means to find a set of w_i that output the maximum f in the equations

5. CONCLUSION

In this paper we reviewed that various data extraction techniques as well as automatic annotation approach using multiple annotators from different Web data bases. We also surveyed that how the data extraction from the various web pages but the traditional approach is having many drawbacks like human interference, the inaccuracy in result and poor scalability. Some approach are used the different feature extraction techniques such as sequence based Tree edit distance, DOM tree, pattern matching and HTML tag structure. In visual data extraction approach is the language independent. This approach mainly focus on the presentation style of and extract the visually information from the template. But still there is need to identify the best technique for data annotation problems.

ACKNOWLEDGMENT

I would like to thank the University Authorities to provide basic facilities for carrying out the research work. I would like to thank my guide Mrs.Suryapriya S. and my parents for most support and encouragement, valuable advices on grammar and theme of the paper.

7. REFERENCES

[1] Y. Lu, H. He, H. Zhao, W. Meng, C.Yu "Annotating Search Results from Web Databases", IEEE Knowledge and Data Engg"., vol. 25, March-2013.

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

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

[4] Davi de Casto Reis, Paulo B. Golgher and Altigran S. da Silva, "Automatic Web News Extraction Using Tree Edit Distance", Proc. ACM World Wide Web (WWW), 2004.

[5] 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.

[6] 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.

[7] W. Liu, X Meng and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," *IEEE Trans. Knowledge and Data Engg.*, vol. 22, no. 3, pp. 447-460, March 2010.

[8] H. He, W. Meng, C. Yu and Z. Wu, "Automatic Integration of Web Interface with WISE-Intigrator," VLDB J., vol. 13, no. 3 pp.256-273, Sept 2004.

[9] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis and Khaled Shaalan "A Survey of Web Information Extraction Systems" IEEE, TKDE-0475-1104.R3.

[10] 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.

[11] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites," *Proc. Int'l Conf. Very Large Data Bases(VLDB),*pp.109-118,2001.