Dynamic Annotation of Search Results from Web Databases

Bincy S Kalloor
PG Scholar, Department of CSE
Marian Engineering College
Trivandrum, India
bincykalloor@gmail.com

Sheeja Agustin
Assistant Professor, Department of CSE
Marian Engineering College
Trivandrum, India
sheejabinudas@yahoo.com

Abstract- The Internet provides a great extent of beneficial knowledge which is usually formatted for its users, which makes it troublesome to extract relevant data from diverse sources. The World Wide Web plays a major role as all kinds of information repository and has been very success full in disseminating information to users. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and internet comparison shopping, they need to be extracted out and allot meaningful labels. This paper deals with the automatic annotation of Search result records from the multiple web databases. Search result presents an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then for each group annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. Finally wrapper is automatically generated by the automatic tag matching weight method.

Keywords*- Data Alignment, Data Annotation, Web Database, Wrapper Generation*

____*****____

I. INTRODUCTION

Data mining is the computational process of discovering patterns in the big datasets. The overall goal of the data mining is to extract information from a datasets and convert it into an understandable structure for further use. Search engines are very important tools for people to reach the vast information on the World Wide Web. For example, if we are searching for a book on a search engine it will display huge amount of results. But we have no time to look all the results displayed by the search engine and also there will be many pages some of them will click to next page and look whether there is relevant information is displayed. This method is a time consuming method. Early application requires tremendous human efforts to annotate the data units manually, which severely limit their scalability.

To overcome this problem an automatic annotation approach is introduced. Automatic annotation approach that first aligns the data units on a result page into different groups, such that the data in the same group having the same semantic meaning, then for each group annotate it from the different aspects and aggregate the different annotations to predict the final annotation label. By using the automatic tag matching method we can reduce the searching time.

II. RELATED WORKS

In recent years web information extraction and annotation has been an active research areas. Extracting structured data from deep web pages is a challenging problem [2].Some of the limitations are webpage programming dependent, Incapable of handling ever increasing complexity of HTML source code. To overcome this problem- a vision based approach [2] vision based data extractor. ViDE is used to extract structured results from deep WebPages automatically. It can only process deep web pages containing one data region while there is significant number of multi-data region deep WebPages, which is time consuming process.[3]ODE which automatically extracts the query results records from the HTML pages. Automatic data extraction is important for many applications such as meta-querying, data integration and data warehousing. In semi automatic wrapper induction has the advantage that no extraneous data are extracted as the user can label only the data in which he/she is interested. To overcome this supervising learning methods are used [4]. Labour intensive and time consuming are the drawback and also it is not scalable to a large number of websites.[4] Technique for extracting data from HTML sites through the use of automatically generated wrappers. A key problem with the manually coded wrappers is that writing them is usually a difficult and labour intensive task and difficult to maintain [4] has a prior knowledge about the page contents. It is an daunting task for users to access numerous web sites individually to get the desired information.

[5] is a tool that perform automatic integration of web interfaces of search engines. It is to identify matching attributes. [6] ViNTS is automatically producing wrappers that can be used to extract search result records dynamically. It utilizes both the visual features on the result page displayed on browser and HTML tag structure of the source file. It helps people to locate and understand information. Existing approaches use decoupled strategies [7]. A

probabilistic model to perform two tasks simultaneously. HCRF can effectively integrate all useful features by learning their importance.

III. PROPOSED SYSTEM

In this approach first user gives the query to the search engine. In returned result page containing multiple SRR, the data unit corresponding to the same concept often share special common features. After the feature selection data alignment is done. The purpose of data alignment is to put the data units of the same concept into one group so that they can be annotated holistically. There are six basic annotators [1] to label data units, with each of them considering a special type of patterns/features. Once the data units on a result page have been annotated, and use these data units to construct an annotation wrapper for the WDBs. in the wrapper generation the tree alignment method is also used to calculate the similarity between the wrapper and the input web results. The input trees are merged into one union whose nodes recovered the statistical information such as the time a node has been aligned and the text length of the node. A heuristic method is utilized to find the most probable content block.

A similarity series was built by calculating the similarity between the input pages and the current wrapper using the tree alignment algorithm. After that change point is detected and wrapper is regenerated. 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 tag matching. One of the major contributions of our work is kind of linear regression method for getting the weight of different tag-matching. The main problem of the previous method is that they did not consider about employing different weights for various 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. Its feasible to get this kind of web pages collection automatically.

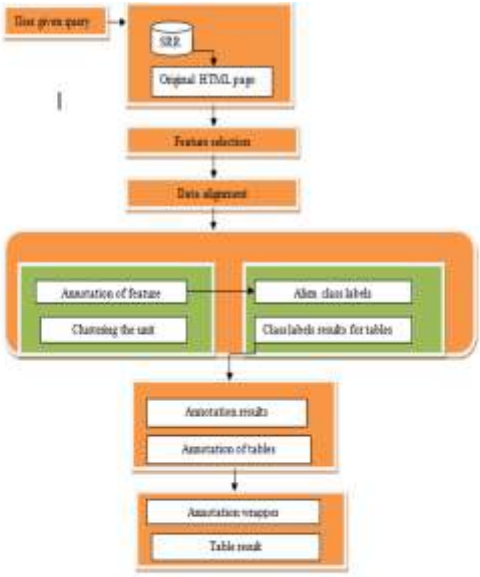


Fig 1: Architecture for annotating search results

Alignment algorithm 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 and also needs the similarity between two data unit groups. In this first create the alignment groups then the clustering is done. After clustering choose the annotation method. There are six annotation methods are there. First set table annotator then perform the function of table annotator. Table annotators first identify the column header. Then for each SRR takes a data unit. Then select the column header with maximum overlap. At last a unit is assigned and labelled. Then query based annotation is done, in this first set of query terms are there. From that find the group with largest occurrences and label is assigned. In the schema annotator attribute is identified with highest matching score. In the frequency annotator find the common preceding units then concatenated preceding units and label the group. In text prefix/suffix annotator check the data units and share the same prefix or suffix. The common knowledge annotator from the group of data units match the patterns or values and label the group. At last wrapper is generated. Each annotated group of data unit corresponds to an attribute in the SRR. The data unit groups are annotated and organised based on the order.

In automatically getting tag matching method 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) Wheret_i^{mn} is the number of wi occur in the alignment procedure.

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

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

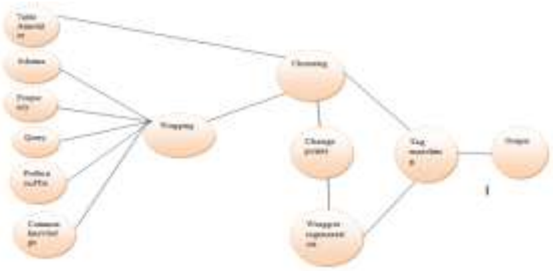


Fig 2: Flow Graph of the work

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

The group of equations is rewritten as:

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

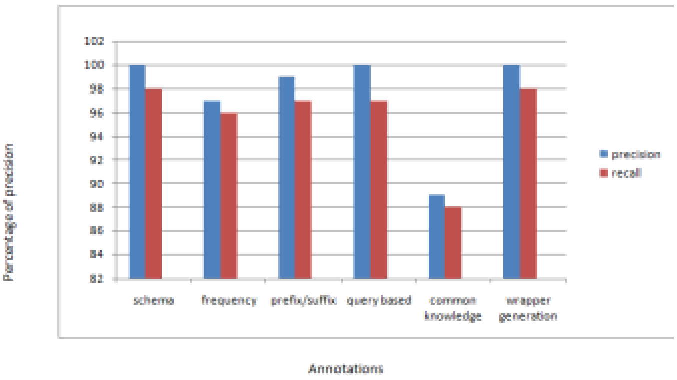
The solution of the above equations is used as the weight of each type of tag-matching (wi).

IV. PERFORMANCE ANALYSIS

Table 1: Applicabilities and Success rate of Annotations

	Applicability	Success rate
Schema value annotator	9%	1.0
Frequency-based annotator	10%	0.99
Prefix/suffix annotator	10%	0.90
Query-based annotator	9%	0.97
Common knowledge annotator	9%	0.89

From the table 1 is shown the applicability and the success rate of each annotator. In the existing system there is an drastic change in the speed of searching. In this method we use the precision and the recall was used to determine the success rate of each annotators. For annotation precision is the percentage of the correctly annotated units over all the data units annotated by the system and in the case of recall, it is the percentage of the data units correctly annotated by the system over all the manually annotated units.



From the above graph we understood that speed of the searching is increased and the processing time is reduced.

V. CONCLUSION AND FUTURE WORK

From the above work we concluded that wrapper is automatically generated. The speed of the searching is increased and the processing time is reduced. From this experiment we overcome the time consuming problem by automatic annotation by we databases. Still there is an room for an improvement that by using artificial intelligence we can make a new annotation.

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng,” Annotating Search Results from Web Databases” IEEE Transaction on Knowledge and Data Engineering, vol. 25, NO. 3, March 2013

[2] 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

[3] W. Su, J. Wang, and F.H. Lochovsky, “ODE: Ontology-Assisted Data Extraction,” ACM Trans. Database Systems, vol. 34, no. 2, article 12, June 2009

[4] V. Crescenzi, G. Mecca, and P. Merialdo, “RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites,” Proc. Very Large Data Bases (VLDB) Conf., 2001

[5] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, “Fully Automatic Wrapper Generation for Search Engines,” Proc. Int’l Conf. World Wide Web (WWW), 2005

[6] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma, “Simultaneous Record Detection and Attribute Labeling in Web Data Extraction,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2006

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