ISSN:0975-9646

A Survey on Search Results Annotation

Rosamma K S , Jiby J Puthiyidam

Dept of Computer Science College of Engineering Poonjar,
Cochin University of Science and Technology
Kottayam, Kerala, India

***Abstract*-The use of web search engines are very frequent and common worldwide over the internet by end users for different purposes. A web search engine takes the query request from the end user and executes that query on relational database used to store the information on behalf of that web search engine. Based on input queries the dynamic response is generated by search engine, in the form of HTML based pages. These pages are supported with the web databases. Every web page generated contains many results to display for particular query, called as Search Result Records (SRRs). These SRRs may contain data units that are relevant to one common semantic. These SRRs are further required to be assigned with proper labels. The manual methods for record extraction and labeling have a worse scalability. Thus automatic annotation based method is needed to improve the accuracy as well as scalability of web search engines. This paper takes the review of such systems.**

***Key words*-Web Database, Annotation, Data alignment**

I. INTRODUCTION

Web information extraction and annotation are two important research areas in recent years. Large portion of the deep web is database based. The data encoded in the returned result pages come from the underlying structured databases, called Web databases (WDB)[12,13]. A typical search result page may contain multiple search result records (SRRs). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. It corresponds to the value of a record under an attribute.

There is a high demand for collecting data of interest from multiple WDBs. Unfortunately, the data units in the SRRs are often not provided with a semantic label. Having semantic labels for data units is important for the record linkage task and for storing collected SRRs into a database table for later analysis.

Since the data units in the SRRs are initially unstructured and unlabeled, an automatic labeling method is required. Thus many alignment and annotation methods were introduced. These methods improve the efficiency of searching and updating of data. Manual methods of labeling had less scalability. So many automatic annotation methods were introduced.

II. ANNOTATION METHODS

Web information extraction and annotation area received a fast growth in the recent years. Many of the systems developed are based on, marking the specified areas in the sample page and then use a set of rules to extract the information. There are many systems that have higher extraction accuracy through the use of supervised learning techniques.

A. *Vision Based Approaches for Web Data Extraction*

These methods [1,2,3] use visual features of deep web page for extracting the web data. There exists many systems that use this concept. ViDE is based on some common visual features of the deep web. ViDE first builds a visual block tree using the VIPS algorithm. Using the Visual Block tree, data record extraction and data item extraction are carried out based on the proposed visual features. Then a visual wrapper generated is to improve the efficiency of both data record extraction and data item extraction.

There is an another system that uses enhanced co-citation algorithm[3] .Unlike the other systems that develops a new set of APIs for the extraction of visual information, this algorithm retrieve the visual information of the deep web pages directly from the web database. The framework is processed under three different phases. The first phase involves extraction of web pages using enhanced co-citation algorithm. The algorithm follows two strategies to extract the visual information of web pages from web database, content based and link based. The former extracts the textual content of the users' query links and its siblings, and the later utilizes only the link construction among the web pages collected for the enhanced co-citation algorithm. The second phase is the data record extraction. The objective of this phase is to determine the border line of data records and remove them from deep Web pages.
It seems that the following assumptions are satisfied:

i. All data records present in multi data region are extracted
ii. For every extracted data record, no data item should be neglected and no erroneous data item be incorporated

Finally, the data record is extracted from the respective web page and the space connecting two data records are identified. For the location data region in a new page, each data record is discovered by the visual comparison with the collected visual information using enhanced co- citation algorithm. Experimental evaluation shows that this approach is better with 5-10% higher in precision rate, with 5-10 % increase in recall and time consumption reduced from 40-50% compared to the ViDE approach.

Key features: The enhanced co-citation algorithm extracts the deep web pages directly from the database instead of using APIs. So it consumes less time for web page extraction. The algorithm needs to improved, for achieving speed and to avoid noise in the extracted data.

B. Extracting Structured Data from Web Pages

Most of the web sites contain large sets of pages generated using a common template or layout. The extraction problem deals with automatically extracting the database values from such template- generated web pages. The term structured data refers to any set of data values conforming to a common schema or type. This method proposes an algorithm, EXALG[4] to solve the EXTRACT problem. The algorithm works in two stages. The first phase deals with discovering sets of tokens associated with the same type constructor in the (unknown) template used to create the input pages. The second phase called Analysis uses the above sets to deduce the template, which is then used to extract the values encoded in the pages.

The first phase computes equivalence classes-sets of tokens having the same frequency of occurrence in every page. Token may be a word or a HTML tag. The algorithm retains only the equivalence classes that are large and whose token occur in a large number of input pages, and are termed as LFEQs(for Large and Frequently occurring Equivalence classes).These LFEQs are formed by tokens associated with the same type constructor. The algorithm tries to add more tokens to LFEQs by differentiating roles of tokens using the context in which they occur. It enters the second stage when it cannot grow LFEQs, or find new ones. This stage builds an output template using the LFEQs constructed in the previous stage..

The algorithm constructs the output template by generating a mapping from each type constructor in to ordered set of strings and the output schema is produced by the type corresponding to root equivalence class.

Key features: EXALG is extremely good in extracting the data from the web pages, generated from a common template. Also it does not completely fail to extract any data even when some of the assumptions made by EXALG are not met by the input collection, that is, the impact of the failed assumptions is limited to a few attributes. But there is a chance of information loss when naive key word indexing and searching is used.

C. Web Data Extraction Based On Partial Tree Alignment

This approach studies the problem of extracting data record, segments these records, and put them into a database table. It can be achieved in a two step strategy. First phase includes identifying individual data records in a page, it uses a method based on visual information to segment data records. Second phase deals with aligning and extracting data items from the identified data records. An enhanced MDR[5] algorithm is used in the first phase. The algorithm is based on two observations:

- A group of data records that contains descriptions of a set of similar objects are typically presented in a contiguous region of a page and are formatted using similar HTML tags.
- The nested structure of HTML tags in a Web page naturally forms a tag tree.

The MDR algorithm works in the following steps:

- Building a HTML Tag Tree
- Mining Data Regions
- Identifying Data Records

A web browser renders each html element as a rectangle. Then the tag tree can be constructed based on these nested rectangles. The 4 boundaries of the rectangle of each HTML element can be found by calling the embedded parsing and rendering engine of a browser. The tree can be built based on the containment check on the identified rectangles. The second step mines every data region in a page that contains similar data records. Data regions can be mined by comparing tag strings of individual nodes (including their descendents) and combination of multiple adjacent nodes .A sequence of adjacent generalized nodes forms a data region. This concept captures the situations that a data record may be contained in a few sibling tag nodes rather than one and that data records may not be contiguous in the tag tree, but generalized nodes are contiguous.

The third step is identifying the data records. A single or a combination of tag nodes in the tag tree may not represent a single data record. The data records that are not contained in a contiguous segment of the HTML code can exist. The algorithms efficiently process these non contiguous data.

Now data extraction can be performed by using the partial tree alignment technique. It has two sub steps:

The first step composes a single tag tree structure [8,7] for each data record. Then these tag trees in each data region are aligned using partial tree alignment technique. The matching process only uses the tags. The partial alignment method [6] is based on tree matching. The STM (Simple Tree Matching) evaluates the similarity of two trees by producing the maximum matching through dynamic programming with complexity O(n1n2), where n1 and n2 are the sizes of tree. Also no node replacement and no level crossing are allowed.

The approach aligns multiple tag trees by progressively growing a seed (tag) tree. The seed tree(T_s), is initially picked to be the tree with the maximum number of data fields. The algorithm then tries to find for each node in T_i a matching node in T_s(i=s). When a match is found for node n_i, it builds a link from n_i to n_s to indicate its match in the seed tree. If no match can be found for node n_i, then the algorithm attempts to expand the seed tree by inserting n_i into T_s.The expanded seed tree T_s is then used in subsequent matching.

The resulting alignment T_s can also be used as an extraction pattern for extracting data items from other pages generated using the same template.

Key features: Partial alignment takes only those data fields in a pair of data records that can be aligned (or matched) with certainty, and it doesn't depend on the rest of the data fields. This method enables very accurate alignment of multiple data records. The partial tree alignment method is able to align data items in nested records, it fails to handle data records that are relatively rare in record lists.

D. Automatic Data Extraction Based On Structural -Semantic Entropy

This algorithm locates and extracts the data of interest from web pages across different sites. The structural-semantic entropy concept[9] is used to identify and locate the data-rich nodes. It measures the density of occurrence of relevant information on the DOM tree representation of web pages.

Definition: The structural-semantic entropy H(N) of a node N in the DOM tree representation of a web page can be defined as:

$$\mathrm{H(N)} = -\sum_{i=1}^{m} \mathrm{pi}\log\mathrm{pi}$$

pi is the proportion of descendant leaf nodes belonging to semantic role i of the node N.

The algorithm navigates on webpage in bottom-up manner based on DOM tree representation. Then the structural-semantic entropy discovers the data-rich nodes and list nodes from the web page by DE-SSE algorithm. This algorithm extracts the attribute-value pairs from those regions in automated manner. For all attributes, a regular expression is constructed to equivalent the keywords in order to deduce the attribute values. So that the leaf nodes can be interpreted with their semantic entropy measures. Due to structural- semantic entropies the nature of a node depends on its children. The algorithm recognizes the record boundary repeatedly from each data rich node containing a record. Regular expressions are constructed for some attributes such as price and time, to increase the precision. It is to ensure that whether the strings are valid values or not for those attributes. When the extracted string appears to not valid for the attribute, then the content of the next-next node will be extracted until meet up the node interpreted with a different semantic role. There may be circumstances in which the title of a record is not explicitly associated with a string. They can be extracted from the first leaf child of the data-rich nodes or the previous leaf of that node.

Key features: The algorithm is efficient in data extraction based on the requirements that the web pages share the similar template as well as dissimilar template. It is independent of modifications in web-page format which enable to detect false positive rate in associating the attributes of records with their respective values. In order to make this approach better, efficient method would be proposed to handle memory consumption, computational overhead, storage, fast processing etc.

E. Semantic Similarity Based Data Alignment and Best Feature Extraction Using PSO

This method proposes an efficient automatic semantic annotation of data units in semantic manner and extract features from SRRs features such as text and data unit feature using Particle Swarm Optimization (PSO)[10] methods. The following relationships are used for extraction and measuring semantic similarity.

- One-to-One Relationship: every text node include accurately one data unit.
- One-to-Many Relationship: every text node includes multiple data units.
- Many-to-One Relationship: numerous text nodes simultaneously type a data unit.
- One-To-Nothing Relationship: every text unit nodes in the example belong to the category of text unit only not data unit within SRRs.

Particle swarm optimization method is used to solve many of the real time application problems. It treats HTML data unit pages as particles input that moving from one particle HTML pages to another HTML pages from SRRs .The importance of this process is to extract important features such as Data Content (DC), Presentation Style (PS), Data Type (DT), Tag Path (TP), and Adjacency (AD) from HTML encoded pages for web search engine. The important features are:

- Data Unit (DC): similar to equal concepts with keywords.
- Type (DT): type or category of data unit belongs same concept is found
- Tag Path (TP): series of nodes present in the text node from HTML and navigates beginning the

derivation of the SRR to the resultant node in the tag hierarchy.

- Presentation Style (PS): different styles supported by web pages that are font style, size, color, text adornment, etc.
- Adjacency (AD): the results are found by keywords to search similar concepts and find from different SRRs.

The algorithm differentiate various data units and identify similar data unit concepts through roughest theory based classification which identifies similarity among data unit's results from above similarity measures formation which partitions the creation into sets of related data units called elementary sets[11]. These sets of data units in the encoded file format can be used to create much information on the data and text units along with similarity functions use of similarity among various representations go ahead to information granulation. It is based on the assumption that every data units are encoded in HTML format of the creation is connected through a definite quantity of data and text units information, characterize by a number of attributes which communicate the descriptions of data units, text units. The usage of similarity of data units and text units nodes in the pages considered as the attributes of elements for better grouping of similar concepts which is additional second-hand to conclude the similarities among the data units based on their characteristic standards.

The returned results are best aligned and then perform annotation approach using wrapper methods for search results returned from whichever specified web database.

Key features: The PSO-DA optimization based feature extraction with efficient semantic similarity measurement best data alignment is helpful and they simultaneously are proficient of creating high-quality explanation of several web databases in the equivalent field.

III. Results And Discussions

To evaluate the performance of above methods, we adopt the precision and recall measures from information retrieval. For alignment ,the precision is defined as the percentage of the correctly aligned data units 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 manually aligned data units by the expert. Table1 table shows the comparison of the above mentioned systems based on the precision, recall and the technique used in the system.

The above defined system were evaluated using many datasets, the table1 only shows their maximum cases.

Table1

Comparison Of Annotation Methods

System	Technique used	Precision	Recall	Time Consum-ption
ViDE	VIPS algorithm	98.7	97.2	high
VBEC	Enhanced Co-citation algorithm	94	92	low
DEPTA-MDR2	Partial tree Alignment Technique	99.68	98.18	medium
DE-SSE	Structural-Semantic Entropy measurement	99.47	98.93	low
PSO-DA	Particle Swarm Optimization method	98.7	97.2	low

IV. Conclusion

The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. A large number of techniques have been proposed in this area, but most of them have some inherent limitations. In this paper we discussed some of the techniques. To overcome these limitations an automatic multi annotator approach that utilizes the Integrated Interface Schema (IIS) of the web database is proposed. But there is still room for improvement. By trying different machine learning techniques and more sample pages from training set we can identify the best technique to the alignment problem.

Acknowledgement

I sincerely express my gratitude to everyone who supported me to prepare this paper. I am thankful for their aspiring guidance, invaluably constructive criticism and friendy advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the work.

References

[1]. Wei Liu, Xiaofeng Meng, Member, IEEE, and Weiyi Meng, Member, IEEE, ViDE: A Vision-Based Approach for Deep Web Data Extraction

[2]. D. Raghu, V. Sridhar Reddy, Ch. Raja Jacob, Dynamic Vision-Based Approach in Web Data Extraction

[3]. R.Vijay1, Dr. K. Prasadh, A Vision Based Approach for Web Data Extraction using Enhanced Co-citation Algorithm

[4]. Arvind Arasu, Hector Garcia-Molina, Extracting Structured Data from Web Pages, Stanford University

[5]. Liu, B., Grossman, R. and Zhai, Y. "Mining data records from Web pages." KDD-03, 2003

[6] Yanhong Zhai, Bing Liu, Department of Computer Science University of Illinois at Chicago, Web Data Extraction Based on Partial Tree Alignment

[7]. Tai, K. The tree-to-tree Correction Problem. J. ACM, 26(3):422–433, 1979

[8]. Valiente, G. Tree Edit Distance and Common Sub trees. Research Report LSI-02-20-R, Universitat Politecnica de Catalunya, Barcelona, Spain, 2002.

[9]. P.V.Praveen Sundar1 1Research Scholar, Hindusthan College of Arts &Science, Coimbatore, India, Towards Automatic Data Extraction Using Tag and Value Similarity Based on Structural - Semantic Entropy

[10]. T.Seeniselvi1, N.Thangamani.2 1Associate Professor, Hindusthan college of Arts and Science, India, Semantic Similarity Based Data Alignment and Best Feature Extraction using PSO for Annotating Search Results from Web Databases

[11]. S. Greco, B. Matarazzo, and R. Slowinski, “Rough sets theory for multi criteria decision analysis,” Eur. J. Oper. Res., vol. 129, no. 1, pp. 1–47, 2001.

[12]. Hongjun Lu, Integrating Database and World Wide Web Technologies, National University of Singapore

[13]. Jungwha Hong, How to Build a Web Database: A Case Study, The University of Texas at Austin.