CMDR: Classifying Nodes for Mining Data Records with Different HTML Structures

Fok Kar Wai, Lim Wee Yong, Vrizlynn L. L. Thing
Institute for Infocomm Research (I^2R)
{fokkw, weylim, vriz}@i2r.a-star.edu.sg

Victor Pomponiu
Universita degli Studi di Torino, Italy
pomponiu@di.unito.it

***Abstract*—This paper addresses the problem of automated structured data records extraction from web pages. In particular, we focus on the extraction of posts from online forum sites. We show that variability in the HTML structure within user generated content in forum posts can negatively affect the extraction accuracy and propose the integration of a deep learning node classifier in the popular Mining Data Regions (MDR) process proposed in prior work. Experiment on a forum web page dataset containing posts with varying HTML structures indicate the merits of the proposed modification for MDR.**

I. Introduction

Online forums in the World Wide Web provide thematic platforms for users to share content and hold discourse. Such sites are typically built using popular forum software such as vBulletin, phpBB, etc., where each user post is contained within fixed pre-defined templates. The automated traversal through such site for extraction of user generated content from online forums is useful for general intelligence gathering [15], [8], [9], considering that the content are typically already organized around pre-defined topics of interest to a significant extent.

A web page is commonly represented in a tag tree given the nested tag structure in the document. In a given thread page of a forum site, each post can be regarded as a data record to be extracted. To automate the extraction of data records in the page, manually crafted XPath information can be used for identifying individual records so that the data within the records can be stored accurately to a pre-defined data structure. However, specifying such information for multiple websites is tedious and error-prone. There exists an area of work on the automated extraction of structured data from web pages. A frequent concept underlying proposed techniques make use of similarities between data records, either within a page or across multiple similar pages, to identify characteristics that can help in the automated extraction of the data into a structured table. Earlier work consider similarity based on the HTML structure such as distance between sibling tag trees or nodes [21], [10], [16], [14], [1] or tag paths [13], [20].

For data regions meant for human consumption, visual cues are present in the pages to enable readers to quickly discriminate between the data regions and the non-data regions. Techniques such as VIPS, ViNTS [22], ViPER [18], ViDE [12] have been proposed, leveraging on visual features to help in the data extraction process.

However, regardless of whether the focus is on HTML structure, visual features or both, previous work leverage on the assumption that such features should be similar between data records in order to automatically identify their location in a web page.

The observation that HTML structures between data records are similar to each other is true to a large extent if the template for the data record allows only limited or no user specified HTML markup. However, sites especially online forums frequently give users more liberty in the HTML markup in the posts. As a result, fixed similarity thresholds used in previous work may not be flexible enough to discriminate the difference between post data records and between non-data nodes. On the other hand, there are features that can help characterise if a node or its sub-tree is within a data region.

This work introduces the use of a binary classifier for an initial identification of data and non-data nodes in a page and use such labels to improve on the automated structured data extraction process. Specifically, we built a neural network for learning features in forum posts, which is then integrated with a popular existng data extraction technique, Mining Data Region (MDR) [21]. Experiments on multiple forum sites with varying HTML structures in the posts indicate improvement in the proposed modified technique.

In the following section, we discuss the related work. In Section III, we present our approach on data records extraction for web pages. The evaluation of our framework, as well as an example illustrating the improved output of the framework compared to previous work is presented in Section IV.

II. Related Work

Prior work in information extraction are mainly on extraction from either natural language text where data is amongst other noise words, or web pages where the aim is to extract structured data records from within the HTML tags [19]. This section focus on the latter, while the interested reader is referred to the survey in [19] that compares various web page data extraction techniques based on inputs/outputs, algorithm complexity and extraction accuracy.

A framework for web page data extraction typically involves a three stage process of (i) filter, (ii) clustering of the page's DOM tree nodes based on corresponding subtrees, tag paths [13], content [16], [14] or visual information and (iii)

selection of the most likely cluster as the data region with the nodes as markers for the data records.

Earlier work on data region/record identification techniques use basic assumptions such as assuming the largest content, size [4], [2] or fan-out node [4] to be root for the only data region in the page.

Subsequently, it is common to identify the data region/record based on the assumption that data records shared similar DOM tree structure [21] and visual characertistics [13]. However, there must exist multiple data records within a page [21], [13], [16], [14] or across pages in the website [20] to allow for the learning of the defining characteristics of the data regions or records.

An example of such work, proposed by Zhai et al., is DEPTA, which consists of an unsupervised two step approach for automated extraction of two or more data records in a web page [21]. The first step involves the identification of data region based on tree structure and text content similarity between subtrees of sibling nodes. The second step involves the alignment of attributes from the data records in a data region using a proposed *Partial Tree Alignment* algorithm. The objective is to align the same data type from each data record to the same column in a table. Specifically, the largest child subtree is initialized as the "seed tree" and other subtrees are iteratively compared against this subtree, such that unmatched child nodes are inserted into the seed tree based on the similarity of their adjacent nodes with the seed tree. Although the alignment process allows for some form of HTML variability in the data records, it can be difficult to set a common threshold across all pages in a web site.

Shi et al. [17] highlighted three cases of data misalignment in ESTM algorithm [21] and proposed an enhanced tree alignment algorithm, Directed Acyclic Graph Multi Tree Matching (DAG-MTM). Such misalignment problems in ESTM arises due to (i) complicated substructure under "terminal" nodes (ii) missing nodes in both trees and (iii) missing content in the $< td >$ nodes. To this end, the authors do away with terminal nodes and consider leaf nodes instead. In addition, they propose a similarity measure that takes into account non-leaf nodes' bounding rectangles in the web page. Nonetheless, because the techniques make no commitment on the locations of unaligned data items, variabilities in the HTML structure can still result in variability in the extraction data across different pages in the same website. An important assumption in [21] is that records in a data region are rooted in sibling subtrees, hence rendering the technique unsuitable for identifying (single) records obtained across two or more web pages.

Bing et al. [1] highlighted another limitation in the DEPTA MDR technique regarding its inflexibility in handling the grouping of subtrees of "complicated" data regions pertaining to embedded and nested regions. The authors proposed the Record Segmentation tree (RST) structure and corresponding extraction technique that is robust against repetitive and optional fields in the records, which in DEPTA can give rise to misleading edit distance leading to error in the MDR process.

It is possible to leverage on visual cues to augment the Partial Tree Alignment technique with visual characteristic [5]. An earlier work that relies on the visual information is Vision-based Page Segmentation (VIPS) that uses nodes' rendered dimension, font, colour, etc. in addition to heuristics to identify data regions in web pages [3]. Other works have since build upon VIPS for data regions and separators identification [11], [6], [7].

Unlike most work that utilize the web page's DOM tree for their data extraction algorithms, [13] identifies data regions by clustering HTML tags based on their tag paths and positions in the web page. Similarity between tag paths is quantified by proposed *offset* and *interleave* functions. Similar to [21], the effectiveness of proposed algorithm is predicated upon regular HTML structure among data records.

Vision-based approaches such as VIPS [3], ViNTS [22], ViPER [18], ViDE [12] have been proposed as well. All such approaches rely on certain assumptions on the visual characteristics of web pages. These assumptions include a certain fixed layout of the data regions within the web page as well as regularity of the type of content and the visual appearances of the content within the data records. Based on these assumptions, certain rules are created which form the basis of their extraction techniques. For example, VIPS proposed an algorithm that utilises visual characteristics of nodes in a web page and identifies visual separators to extract the data regions in the page [3]. ViNTS make use of both tag tree and visual information to generate extraction rules in the form of tag path regular expressions [22]. ViPER further considers records that may not necessarily be in one continguous region or are arranged horizonatally across the page [18].

Unsurprisingly, similarity measures are used in these work to quantify similarity, either structural or visual, between node(s). Such an approach would likely perform well on web sites where the visual structure of data records are very similar. However, for sites such as online forums, where a large proportion of the content of the data records are provided by the users with the freedom to specify certain markups to the content, the records may not have high similarity.

While existing work have achieved good results for large number of extractions, it is noted that the majority of the records are uniform in their HTML structure, with techniques tolerating only relatively minor deviations between records' HTML structures. In particular, rich HTML features allowed in user generated content can invalidate the assumption that data records have similar HTML structure or visual characteristics, resulting in incomplete data extraction or misalignment of attributes in the extracted data records. For example, the presence of multiple quotes within a forum post would induce an additional column for each of the quotes, causing tables from different pages to have varying number of columns.

An example of data records with disparate HTML structures is shown in Figure 1. This work focuses on such data records with high variability in the HTML structures. To this end, we proposed the integration of a node classifier to an existing popular data extraction technique [21]. In particular,

the contributions are:

- identifying the challenges that variable HTML structure between records can pose for automated structured data records extraction,
- introducing a classifier for discriminating between data and non-data nodes in a web page,
- integration of the classifier with a popular data records extraction process [21]

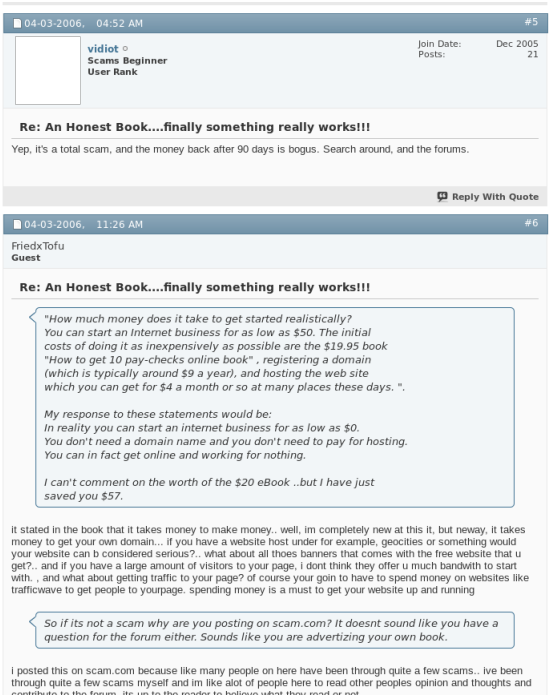

Fig. 1: Example screen capture of segment of a forum thread page containing 2 posts

III. Approach

We propose a system that is a modification of the existing Mining Data Region (MDR) process within DEPTA [21] to improve its data region extraction capabilities. We first introduce the use of deep learning to identify whether or not each HTML node in a web page belongs to the area of interest. The nodes labels are then incorporated into MDR to improve its data region extraction process.

A. Using deep learning to classify HTML nodes in a web page

A binary classification for each HTML node within a web page tag tree is first considered. Specifically, nodes within the area of a data record are annotated as `data`, while the rest of the nodes are considered `noise`.

Features listed in Table I are extracted for each node. The features range from the layout category such as size and location of the node to style category such as the Cascading Style Sheets (CSS) properties of the node. For generalisation across different sites, we note that the features do not capture the semantics of the content, but rather the physical properties such as the relative positioning and visual properties of the nodes.

Feature
1. Absolute and relative location
2. Absolute and relative size
3. Distance to parent node
4. Average distance to sibling nodes
5. Average distance to child nodes
6. HTML tag
7. Font of text belonging to node
8. Color of text belonging to node
9. Opacity of text belonging to node
10. Font of text belonging to node
11. Background color
12. Border style
13. Border radius
14. Border width
15. Border color
16. Margin width
17. Padding width

TABLE I: Features extracted from HTML nodes.

In order to train the neural network, we obtained a dataset using forum web sites. Firstly, XPaths for HTML nodes within the forum posts are manually defined for each site. Next, we randomly crawl the forum sites and extract up to 100,000 HTML nodes per site or till all pages have been crawled. HTML nodes identified by the XPaths are then annotated as `data`, while the rest of the nodes are annotated as `noise`. In total, approximately 2,400 web pages were crawled from ten web sites. Eight of these sites are live, while the other two are dark web forum sites obtained from an offline dataset[1].

Forum web site	Precision	Recall	F_1 Score
renotalk.com	0.89	0.82	0.86
scam.com	0.75	0.94	0.83
forums.salary.sg	0.69	0.96	0.81
forum.brightsparks.com.sg	0.95	0.73	0.83
forums.sgtrains.com	0.92	0.68	0.78
japan-guide.com	0.96	0.71	0.82
forum.singaporeexpats.com	0.88	0.87	0.88
forums.ubi.com	0.93	0.56	0.70
ironclad	0.93	0.58	0.72
silkroad	0.94	0.82	0.87
Average	0.88	0.77	0.81

TABLE II: Neural Network Classification Results

A preliminary experiment based on neural network classifiers trained on the features extracted from the nodes in the above dataset is performed.

To evaluate if the proposed features can be generalised across forum sites, a leave-one-out approach is adopted where nodes from each site are labelled using a classifier which is trained on the remaining nine sites. Average precision, recall and F_1 score at 88%, 77% and 81% respectively is obtained across the forum sites. Detailed result for each forum site is shown in Table II.

The labelled nodes are then used in the subsequent step to enhance the data region extraction process as described in the next Subsection.

[1]https://archive.org/details/dnmarchives

B. Integrating classification results into DEPTA Mining Data Region

The mining data region process was first proposed in [10], with an enhanced version proposed in [21] taking into account the basic visual information in candidate data regions. Algorithm 1 provides a concise overview of the Mining Data Region process. The objective is to identify data regions in a web page, where each data region contains one or more data records and each data record corresponds to one or more sub-trees in the web page HTML tag tree. FindDRs() is the entry function to the data region mining process where the initial input is the root of a web page tag tree. It then traverses the tag tree depth-first and calls IdentDRs() (Line 4) to identify possible data regions under each parent node in the tag tree. Identified data regions of child nodes contained within data region(s) of the parent are discarded by UnCoveredDRS() function (Line 8), which is elaborated in greater detail in [21].

Algorithm 1 Mining Data Regions (MDR)

```
1:  procedure FINDDRS(node, K, τ_low, tau_high)
2:      D ← ∅
3:      if TreeDepth(node) ≥ 3 then
4:          D = IdentDRs(node, K, τ_low, τ_high)
5:          C = ∅
6:          for all child ∈ children(node) do
7:              C = FindDRs(child, K, τ_low, tau_high)
8:              C ← C ∪ UnCoveredDRs(C)
9:          D ← D ∪ C
10:     return D
11:
12: procedure IDENTDRS(node, K, τ_low, τ_high)
13:     R ← ∅     // initialise candidate data regions
14:     for all c_i, c_j ∈ childSetPairs(node, K) do
15:         dist = distance(c_i, c_j)
16:         if (dist ≤ τ_low) or (τ_low < dist ≤ τ_high and
    matchData(c_i, c_j)) then
17:             collectRecords(R, {c_i, c_j})
18:     D = selectMaxDRs(R)
19:     return nD
```

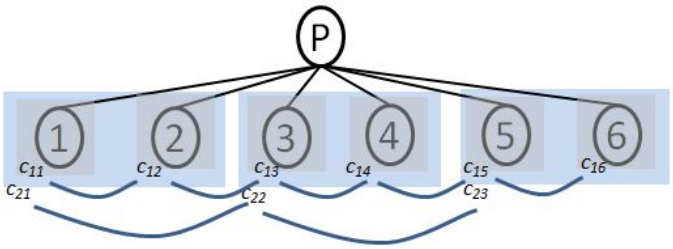

Fig. 2: Example child node sets.

Consecutive pairs of set of child nodes are enumerated in childSetPairs() (line 14, IdentDRs()). Consider the child nodes in Figure 2, example 1- and 2-combination pairs are $\{(c_{11}, c_{12}), (c_{12}, c_{13}), \cdots, (c_{15}, c_{16}\})$ and $\{(c_{21}, c_{22}), (c_{22}, c_{23})\}$ respectively. Readers are referred to [21] for the detail on how such pairs are enumerated. Tree edit distance between child node set in each pair is computed, such that pairs with distance less than τ_{low} are considered similar, and hence possibly data records. Such child node sets are then collected as part of candidate data regions (lines 16-17).

However, data records with different HTML structures can exist in sites such as online forums where users are allowed HTML-rich inputs, possibly leading to large variations in the edit distances between the records. Hence there may exist no clear tree edit distance threshold to discriminate between the data and noise regions.

We propose a modification to the MDR process [21], using $\{\texttt{data}, \texttt{noise}\}$ labels from a node classifier to address the above weakness in the mining data region process in dealing with records with varied HTML structures. We refer to this MDR enhanced using the Classification labels as CMDR. We introduced the use of a more lenient threshold τ_{high}, where $\tau_{high} > \tau_{low}$. For a pair of child node sets c_i, c_j, with tree edit distance greater than τ_{low}, but less than τ_{high}, further evaluation is performed in matchData() to determine if c_i and c_j should be considered as candidate data records (line 16).

Let the nodes in node set c_i be c_i^x, and nodes in set c_j be c_j^y, where both $x, y \in \{1, 2, \cdots, K\}$ and K is the number of nodes in both node sets. If there exists a pair of nodes c_i^x, c_j^y, where $x = y$, that are both classified as `data` by a model trained using the web sites listed in Table II, matchData() then outputs *True* to allow for the pair of child node sets to be collected as candidate data regions. Lastly, selectMaxDRs() refers to the process in [21] for selecting the maximum covering set of child nodes to be the extracted data regions for $node$.

IV. Performance Studies

A. Dataset

DEPTA is able to fully extract the majority of data records with similar HTML structures. However, this work targets pages containing considerable HTML structure differences amongst data records. To this end, a manually curated dataset is used in our evaluation. The dataset contains 31 web pages from eight different web sites. Data records are the forum posts in these pages. There is a total of 335 posts. It is noted that the focus of this work is on records with different HTML structures in the user generated content, such as the example shown in Fig 1. Hence, higher number of data records from prior datasets are not used to prevent skewing the evaluation on posts which have very similar HTML structure.

Of the eight web sites, five of them were used in the training of our neural network model given in Table II. However, the web pages used for training are different from those used for evaluation. In addition, pages from the remaining three web sites were not used for training in order to evaluate the generalization ability of our model to classify nodes in pages from new web sites.

Forum site	#Pages	#Posts	Extracted posts		False regions	
			MDR	CMDR	MDR	CMDR
renotalk.com	5	68	36	62	9	4
scam.com	5	45	24	40	4	6
forum.brightsparks.com.sg	1	10	6	10	0	0
forums.sgtrains.com	5	46	25	39	15	12
forum.singaporeexpats.com	2	24	15	23	0	4
towerofsaviorsforum.com	5	45	24	40	7	5
epicgames.com	3	27	13	23	0	2
forums.hardwarezone.com.sg	5	63	22	45	0	14
Total	31	335	167 (49.9%)	289 (86.3%)	35	47

TABLE III: Forum posts data record extraction results.

B. Experimental Results

The total number of web pages and posts as well as experimental results for correctly extracted posts by DEPTA using the original MDR and our proposed CMDR modification is shown in Table III.

A strict criteria is used in the evaluation.

- A forum post is considered as correctly extracted only if all data fields in the post (e.g. user account name, date, title, content, etc.) have been extracted within a single record.
- Output from the data extraction process are tables of data records. If forum posts are present in more than one table, only the table containing the largest number of correctly extracted forum posts is considered for the results in Table III. This is because such "split" extraction is due to error in the mining data region process and hence should be penalized accordingly.
- When using DEPTA with MDR, it is observed that it may be possible for multiple posts to be extracted into a single data record. In such cases, we consider the number of extracted posts to be the number of records extracted by the algorithm.

From the results, we can observe that for all the web sites in the dataset, we are able to improve the number of correctly extracted posts. Out of the total 335 posts, we improved the number of posts extracted from 167 to 289 which is 49.9% of posts to 86.3% of them.

Falsely extracted data regions are incomplete extracted data from within the forum posts. Such cases occasionally arise due to low similarity between forum posts and the presence of valid data regions within the post content. For example, a $<ul>$ list within the post content can arguably be a data region, with the list items as the data records. However for evaluation in this work, extracted data regions containing content from forum post(s), but with data records *not* corresponding to complete posts are considered as "false regions".

Table III also shows an increase of extracted false regions from 35 to 47 due to our modifications. However, such increase is not consistent across all the forum sites. In fact, the number of extracted false regions decreased for three of the sites when CMDR is used instead of MDR. The decrease in false regions is due to the inclusion of the MDR false regions into the correct data regions in CMDR, hence eliminating them. On the other hand, when posts are missed out, CMDR may increase the probability of detecting false regions from within those posts.

C. Example

In this subsection, we elaborate on the example in Figure 1.

Region A

...	...	...	...	...	...	...	...	...	...	...	...
04-03-2006	4:52 AM	#5	vidiot	Scam Beginner	User Rank	Join Date	Dec 2005	Posts	21	RE: An Honest Book.....	...

Region B

04-03-2006	1:47 PM	#7	martin77	Silver Scams Member	User Rank	Join Date	Aug 2004	Posts	110	RE: An Honest Book.....	...
...	...	...	...	...	...	...	...	...	...	...	...

Fig. 3: Partial output of extraction with MDR for example in Fig 1

...	...	...	...	...	...	...	...
04-03-2006	4:52 AM	#5	vidiot	Scam Beginner	User Rank	Join Date Dec 2005 Posts 21 Yep, it's a total scam...	...
04-03-2006	11:26 AM	#6	-	FriedxTofu	Guest	How much money does it take...	...
04-03-2006	1:47 PM	#7	martin77	Silver Scams Member	User Rank	Join Date Aug 2004 Posts 110 Nobody uses Geocities...	...
...	...	...	...	...	...	...	...

Fig. 4: Partial output of extraction with CMDR for example in Fig 1

There are two main differences in the structure of the two posts shown in Figure 1:

1) The second post does not contain information such as user join date and post count which exists in the first post (top right corner). This is because the author for the second post is a guest who has not registered an official account and hence does not have those information.
2) The main body of the second post contains multiple quotes whereas the first post does not.

Figure 3 shows the partial output produced by DEPTA using MDR when the web page of the sample in Fig 1 is given as input. Due to the dissimilar structure between the two posts, the second post was not extracted successfully. This resulted

in the list of posts in the web page being split into two data regions. Region A contains the post records belonging to the earlier portion of the list while Region B contains the post records belonging to the later portion.

Figure 4 shows the partial output produced by DEPTA using CMDR. Only a single region of data records was extracted for the list of posts. The output contains the data items for both the posts shown in Fig 1 as well as the following post that is not shown in the figure. Our classification labels had helped DEPTA recognize that the two posts were records that should belong to the same data region, resulting in the correct output.

In particular, it is observed that for the second record (corresponding to the second post in Figure 1), the username "FriedxTofu" is not aligned with the username field of the other records. This is due to the difference in structure where other posts have a total of three fields around the area of the username as compared to this post which has only two. Also, fields such as the join date, post count and part of the main body are correctly combined within a single data item whereas they are separated for DEPTA's output (Figure 3). This is due to accommodating the second post which do not have the join date and post count fields. However, it is important to point out that no data was lost, instead some were just misaligned. We accept this low degree of data item misalignment as a necessary trade-off for the inclusion of records which would otherwise not appear in the output, causing loss of information.

DEPTA's data alignment algorithm which is responsible for extracting the individual data items of each record and sorting them into the appropriate categories, depends on the similarity in structure of the records. Since we have fixed less similar records into the same data region, it may have an impact on the data item alignment in the output.

V. Conclusion

In this paper, we proposed a framework for the extraction of web data records with variable HTML structure by leveraging on the use of deep learning. Our framework learns the characteristics of HTML nodes located within a region that contains the data records of interest in the web pages. Our framework utilises prior work DEPTA's record extraction capabilities and improves the data region mining process with our deep learning classification results. We performed experiments on a forum web page dataset containing records with largely varying HTML structure. From the results, we showed that our framework substantially improved extraction performance on the dataset.

Acknowledgment

This material is based on research work supported by the Singapore National Research Foundation under NCR Award No. NRF2014NCR-NCR001-034

References

[1] Lidong Bing, Wai Lam, and Yuan Gu. Towards a unified solution: Data record region detection and segmentation. 2011.

[2] David Buttler, Ling Liu, and Calton Pu. A fully automated object extraction system for the world wide web. *IEEE International Conference on Distributed Computing Systems*, 2001.

[3] Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. Extracting content structure for web pages based on visual representation. In *Proceedings of the 5th Asia-Pacific Web Conference on Web Technologies and Applications*, 2003.

[4] D. W. Embley, Y. Jiang, and Y.-K. Ng. Record-boundary discovery in web documents. *SIGMOD Rec.*, 28(2), June 1999.

[5] Siwu Fan, Xinjun Wang, and Yongquan Dong. Web data extraction based on visual information and partial tree alignment. In *Web Information System and Application Conference*, pages 18–23, 2014.

[6] Jinbeom Kang and Joongmin Choi. Recognising informative web page blocks using visual segmentation for efficient information extraction. 14(11):1893–1910, 2008.

[7] Longzhuang Li, Yonghuai Liu, and A. Obregon. Visual segmentation-based data record extraction from web documents. In *Information Reuse and Integration*, 2007.

[8] Wee Yong Lim, Vyjayanthi Raja, and Vrizlynn L. L. Thing. Generalized and lightweight algorithms for automated web forum content extraction. In *IEEE International Conference on Computational Intelligence and Computing Research*, 2013.

[9] Wee Yong Lim, Amit Sachan, and Vrizlynn L. L. Thing. A lightweight algorithm for automated forum information processing. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013.

[10] Bing Liu, Robert Grossman, and Yanhong Zhai. Mining data records from web pages. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–606, 2003.

[11] Wei Liu, Xiaofeng Meng, and Weiyi Meng. Vision based web data records extraction. In *In Ninth International Workshop on the Web and Databases (WebDB)*, 2006.

[12] Wei Liu, Xiaofeng Meng, and Weiyi Meng. Vide: A vision-based approach for deep web data extraction. *Knowledge and Data Engineering, IEEE Transactions on*, 22(3):447–460, March 2010.

[13] Gengxin Miao, Junichi Tatemura, Wang-Pin Hsiung, Arsany Sawires, and Louise E. Moser. Extracting data records from the web using tag path clustering. In *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[14] Konstantinos Raftopoulos, Dimitrios Skoutas, Theodora A. Varvarigou, and Nikolaos K. Papadakis. Stavies: a system for information extraction from unknown web data sources through automatic web wrapper generation using clustering techniques. *Knowledge and Data Engineering, IEEE Transactions on*, 17(12):1638–1652, 2005.

[15] Amit Sachan, Wee Yong Lim, and V. L. L. Thing. A generalized links and text properties based forum crawler. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012.

[16] Yuan Kui Shen and David R. Karger. U-rest: An unsupervised record extraction system. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[17] Shengsheng Shi, Chengfei Liu, Chunfeng Yuan, and Yihua Huang. Multi-feature and dag-based multi-tree matching algorithm for automatic web data mining. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences on*, 2014.

[18] Kai Simon and Georg Lausen. Viper: Augmenting automatic information extraction with visual perceptions. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005.

[19] Hassan A. Sleiman and Rafael Corchuelo. A survey on region extractors from web documents. *IEEE Trans. on Knowl. and Data Eng.*, 25(9), September 2013.

[20] Jiying Wang and Fred H. Lochovsky. Data extraction and label assignment for web databases. In *Proceedings of the 12th International Conference on World Wide Web*, 2003.

[21] Yanhong Zhai and Bing Liu. Structured data extraction from the web based on partial tree alignment. *IEEE Trans. on Knowl. and Data Eng.*, 18(12), 2006.

[22] Hongkun Zhao, Weiyi Meng, Zonghuan Wu, Vijay Raghavan, and Clement Yu. Fully automatic wrapper generation for search engines. In *Proceedings of the 14th International Conference on World Wide Web*, 2005.