Towards Automatic Structured Web Data Extraction System

Tomas GRIGALIS [a,1]
[a] *Vilnius Gediminas Technical University, Vilnius, Lithuania*

Abstract. Automatic extraction of structured data from web pages is one of the key challenges for the Web search engines to advance into the more expressive semantic level. Here we propose a novel data extraction method, called ClustVX. It exploits visual as well as structural features of web page elements to group them into semantically similar clusters. Resulting clusters reflect the page structure and are used to derive data extraction rules. The preliminary evaluation results of ClustVX system on three public benchmark datasets demonstrate a high efficiency and indicate a need for a much bigger up-to-date benchmark data set that reflects contemporary WEB 2.0 web pages.

Keywords. Information extraction, structured web data, deep web

Motivation and Research Questions

For the Web search engines to advance into a more expressive semantic level, we need tools that could extract the information from the Web and represent it in a machine readable format such as RDF [1]. Information extraction at Web scale is the first step in persuing this goal. However, current algorithmic approaches often fail to achieve satisfactory performance in real-world application scenarios due to abundant structurally complicated WEB 2.0 pages.

In this work we address the problem of automatic information extraction from structured Web data, such as lists of products in online stores. We propose a novel approach, called ClustVX, which is fully automatic, scalable, and domain independent.

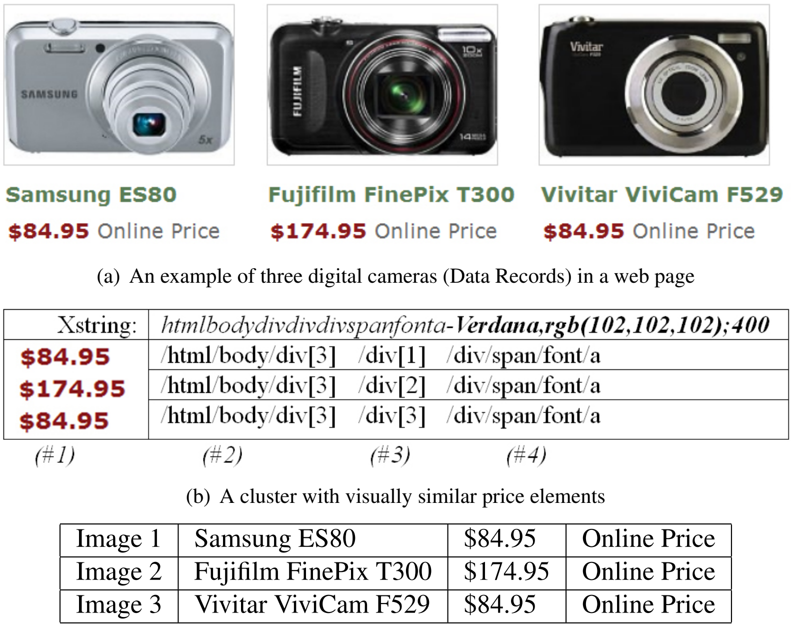
ClustVX is based on two fundamental observations. First, vast amount of information on the Web is presented using fixed templates and filled with data from underlying databases. For example, Fig. 1(a) shows three Data Records (DRs) representing information about three digital cameras in an online store. The three DRs are listed according to some unknown to us style template and the information comes from a database. This also means, that each DR has almost the same Xpath (tag path from root node in HTML tree to particular web page element), where only a few node numbers differs.

Second, although the templates and underlying data differ from site to site, humans understand it easily by analyzing repeating visual patterns on a given Web page. We hypothesize, that the data which has the same semantic meaning is visualized using the same style. Therefore humans, viewing such a web page, are able to comprehend its

[1] PhD student. Email: tomas.grigalis@vgtu.lt; Supervisor: Antanas Cenys, Email: antanas.cenys@vgtu.lt

unique structure quickly and effortlessly and distinguish items photos, titles, prices and etc. For example in Fig. 1(a) prices are brown red and bold, title is green and bold, text ”Online Price” is grey.

ClustVX exploits both of these two observations by representing each web page element with a combination of its Xpath and visual features such as font, color and etc. For each visible web page element we encode this combination into the string called Xstring. Clustering Xstrings allows us to identify visually similar elements, which are located in the same region of a web page and in turn have same semantic meaning. See Fig. 1(b) where price elements are clustered together according to their Xstring. Subsequent data extraction leads to a machine readable structured data. The result of this extraction is shown in Fig. 1(c). Our preliminary evaluation on three public datasets demonstrate that the new method is able to consistently achieve high recall and precision in extracting structured data from given web pages.



(a) An example of three digital cameras (Data Records) in a web page

Xstring:	*htmlbodydivdivdivspanfonta*-***Verdana,rgb(102,102,102);400***
$84.95	/html/body/div[3] /div[1] /div/span/font/a
$174.95	/html/body/div[3] /div[2] /div/span/font/a
$84.95	/html/body/div[3] /div[3] /div/span/font/a

(#1) *(#2)* *(#3)* *(#4)*

(b) A cluster with visually similar price elements

Image 1	Samsung ES80	$84.95	Online Price
Image 2	Fujifilm FinePix T300	$174.95	Online Price
Image 3	Vivitar ViviCam F529	$84.95	Online Price

(c) Desired extraction result

Figure 1. An example of information extraction using ClustVX

In the following we present a brief review of the current related research work and then, in Sec. 2 we outline the ClustVX system. We present experimental results in Sec. 3 and, finally, outline the necessary future research directions and further aspects of experimental evaluation in Sec. 4.

1. Related Work

Data extraction systems can be broadly divided into supervised and unsupervised categories. Supervised learning approaches require some manual human effort to derive the extraction rules, while automated data extraction systems work automatically and need no manual intervention to extract data.

In this work we focus on the latter as we believe that only fully automatic systems can be applied for web-scale data extraction. Our proposed ClustVX system belongs to this category.

One widely adopted technique to automatically detect and extract DRs is to search for repetitive patterns in HTML source code by calculating the similarity of HTML tree nodes. Variations of simple tree matching algorithm [10] are employed for this task [3,8]. However, this technique finds it difficult to deal with structural irregularities amongst DRs , such as lists inside DRs [11].

Contrary to the above recent system VIDE [4] tries to not tie itself to HTML tree at all and instead depends purely on visual features of a web page. It builds a visual containment tree of a web page using patented VIPS [7] algorithm and then uses it instead of HTML tree. However if there are some unloaded images or missing style information in a web page VIPS may fail to build correct visual containment tree which leads to data extraction problems [4].

Combining previous two approaches ViNTs [5] and DEPTA [3] systems try to exploit visual features of web pages to aid structural based data extraction process. However, ViNTs system do not extract data items, it just segment DRs, and evaluation of DEPTA demonstrated, that it cannot handle contemporary pages efficiently [11].

Systems like TextRunner [6] try to extract entities and their relationships from web pages using natural language processing and machine learning approaches, but those techniques usually work on regular text and are not suitable for detecting repetitive patterns in web pages. By contrast, WebTables [2] extracts entities from structured web tables enclosed in < `table` > HTML tags. However, it would miss structured data presented in other form of html tags.

In summary, none of the systems can properly handle visual and structural features of web page to effectively extract structured web data. The ClustVX system proposed in this work fully exploits visual and structural information and achieves promising results.

2. The ClustVX Approach

The ClustVX processes a given Web page in the following steps:

1. Preprocessing the page. In this stage the web page is cleaned from all HTML text formatting tags, such as < `b` >, < `em` >, which appear in the middle of a text and may hinder the clustering process. Visual features of web page elements acquired from browser's API are embedded into HTML tags for processing in the next step.
2. Generating Xstring representation of each HTML element. Each visible web page elements is represented by a Xstring, by which the elements are later clustered. As we see in Fig. 1(b) Xstring consists of a) tag names from Xpath b) visual features of that element (font style, color, weight, etc.). Structural features (string of tag names) identifies position in HTML document. Visual features enhance understanding of semantic similarity between web page elements.
3. Clustering of web page elements. All visible web page elements are clustered according to their Xstring. Resulting clusters contain only semantically similar web page elements. In Fig. 1(b) at (#1) we see a cluster of price elements.

Table 1. Publicly available benchmark data sets for structured web data extraction

Data Set	TBDW [9]	ViNTs-2 [5]	Alvarez [8]
Sites	51	102	200
Pages per site	5	11	1
AVG records per page	21	24	18
Total records (1st page per site)	1052	2489	3557

4. Extraction of structured data. By analyzing Xpath of clustered visually similar web page elements, extraction rules are induced and data is extracted. In Fig. 1(b) at (#2) is a Xpath of the region in a page where Data Records are located. Each Data Record is enclosed in a DIV tag (#3). The final path of price elements inside a Data Record is (#4).

3. Research Methodology

To evaluate ClustVX approach we use the following three publicly available benchmark datasets containing in total of 7098 data records: 1) TBDW Ver. 1.02 [9], 2) ViNTs dataset 2 [5], 3) M. Alvarez et al. [8]. See Tab. 1 for details. These data sets contain search result pages generated from databases. Following the works of other authors [8, 11,4,3,5] in structured data extraction we use three evaluation metrics which come from information retrieval field: precision, recall and F-score.

The positive preliminary results showing that ClustVX can achieve higher than 0.98 F-Score encourage further development and evaluation of ClustVX on real-world data. We see a must to have a data set containing thousands of pages from different web sites. To create such a huge data set we are planning to exploit the power of crowdsourcing by the help of Amazon Mechanical Turk service [12]. The sercvice lets to present simple human intelligence requiring tasks, such as labeling data or telling if extraction was successful, to thousands of voluntary workers, which are paid on per hour or per task basis.

4. Conclusions and Research Directions

In this paper we presented ClustVX system, which, by exploiting visual and structural features of web page elements, extracts structured data. The preliminary evaluation of ClustVX on three publicly available benchmark data sets demonstrated, that our method can achieve very high quality in terms of precision and recall. Our future work will be concentrated on creating a new huge benchmark data set, dealing with extremely malformed HTML source code and comparing ClustVX system to competing approaches.

References

[1] Weikum, G., Theobald, M.: From information to knowledge: harvesting entities and relationships from web sources. In: *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, ACM (2010), 65–76.

[2] Cafarella, M., Halevy, A., Wang, D., Wu, E., Zhang, Y.: Webtables: exploring the power of tables on the web. *Proceedings of the VLDB Endowment* **1**(1) (2008), 538–549.

[3] Zhai, Y., Liu, B.: Web data extraction based on partial tree alignment. In: *Proceedings of the 14th international conference onWorld WideWeb*, ACM (2005) 76–85.

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

[5] Zhao, H., Meng, W., Wu, Z., Raghavan, V., Yu, C.: Fully automatic wrapper generation for search engines. In: *Proceedings of the 14th international conference on World Wide Web*, ACM (2005), 66–75.

[6] Banko, M., Cafarella, M., Soderland, S., Broadhead, M., Etzioni, O.: *Open information extraction for the web*, University of Washington (2009).

[7] Cai, D., Yu, S., Wen, J., Ma, W.: Vips: a visionbased page segmentation algorithm. Tech. rep., Microsoft Technical Report, MSR-TR-2003-79 (2003).

[8] Alvarez, M., Pan, A., Raposo, J., Bellas, F., Cacheda, F.: Extracting lists of data records from semistructured web pages. *Data and Knowledge Engineering* **64**(2) (2008), 491–509.

[9] Yamada, Y., Craswell, N., Nakatoh, T., Hirokawa, S.: Testbed for information extraction from deep web. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers and posters*, ACM (2004), 346–347.

[10] Yang, W.: Identifying syntactic differences between two programs. *Software: Practice and Experience* **21**(7) (1991), 739–755.

[11] Jindal, N., Liu, B.: A generalized tree matching algorithm considering nested lists for web data extraction. In: *The SIAM International Conference on Data Mining* (2010), 930–941.

[12] Alonso, O., Rose, D., Stewart, B.: Crowdsourcing for relevance evaluation. In: *ACM SIGIR Forum*, ACM, Vol. 42 (2008), 9–15.