

HyLiEn: A Hybrid Approach to General List Extraction on the Web

Fabio Fumarola Tim Weninger[†] Rick Barber[†]
Donato Malerba Jiawei Han[†]

[†]University of Illinois at Urbana-Champaign
Università degli Studi di Bari “Aldo Moro”

ffumarola@di.uniba.it, weninge1@illinois.edu, barber5@illinois.edu,
malerba@di.uniba.it, hanj@illinois.edu

ABSTRACT

We consider the problem of automatically extracting general lists from the web. Existing approaches are mostly dependent upon either the underlying HTML markup or the visual structure of the Web page. We present **HyLiEn** an unsupervised, **Hybrid** approach for automatic **List** discovery and **Extraction** on the Web. It employs general assumptions about the visual rendering of lists, and the structural representation of items contained in them. We show that our method significantly outperforms existing methods.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database applications—*data mining*; H.3.1 [Content Analysis and Indexing]: [structured data extraction]

Keywords

Web lists, Web mining, Web information integration

1. INTRODUCTION

This work focuses on extracting information from lists on the Web. Lists are interesting because they present information in a condensed, well structured way. The characteristics of Web lists vary widely. Consequently, a great variety of computational approaches have been applied to discover and extract the information embedded in lists. These existing approaches mostly rely on the underlying HTML markup and corresponding DOM structure of a Web page [6, 1, 7]. Unfortunately, HTML was initially designed for rendering purposes and not for information structuring (like XML). As a result, a list can be rendered in several ways in HTML, and it is difficult to find an HTML-only tool that is sufficiently robust to extract *general* lists from the Web.

Visual information extraction approaches move the focus of the problem from the HTML and its corresponding DOM

tree structure to a visual pattern recognition problem [2, 4]. These visual-based methods are still inadequate to be used for general list extraction. They tend to focus on sub-problems, such as the extraction of tables where each data record contains a link to a detail page [3], or discovering tables rendered from Web databases [4] (deep web pages).

In this paper, we propose HyLiEn, a novel hybrid based approach for *automatic* discovery and extraction of general lists on the Web. Although there are already some works that pay attention on visual information on Web pages [2, 4] and on the DOM-structure of a Web page [6, 1, 7], to the best of our knowledge, only a few methods use both visual and DOM information to perform Web list discovery [2, 5] the most recent of which, VENTex [2], is used in our case study for comparison.

2. VISUAL-STRUCTURAL METHOD

HyLiEn is primarily based on the visual alignment of list items, but it also utilizes non-visual information such as the DOM structure and the size of visually aligned items. The result of the Web page rendering process can be regarded as a set of *boxes*. Figure 1(a) and 1(b) show an example of *web page* and *box segmented page*. Each rendered box has a position and size, and can either contain content (*i.e.*, text or images) or more boxes. Generally, there is a box created for each DOM element. Starting from the box representing the entire Web page, usually the HTML tag, we recursively consider inner boxes, and extract lists boxes which are visually aligned and structurally similar to other boxes.

To extract lists on the basis of visual cues, the following basic assumption is made:

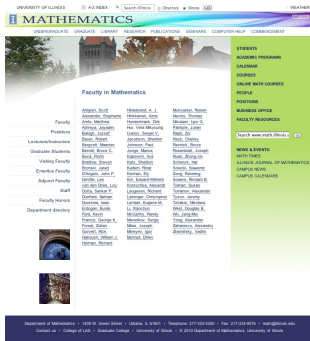
Definition 1. A list candidate $l = \{l_1, l_2, \dots, l_n\}$ on a rendered Web page consists of a set of vertically and/or horizontally aligned boxes.

As shown in [8], this assumption alone is sufficient to outperform all existing list extraction methods. However, it does not cover Web pages such as the one in Figure 1(a) where, all of the orange boxes inside Box $A_{2.2}$ correspond to a single list in the page, but there are many pairs of elements in this list which are not visually aligned. Therefore, we extend the assumption in definition 1 as following:

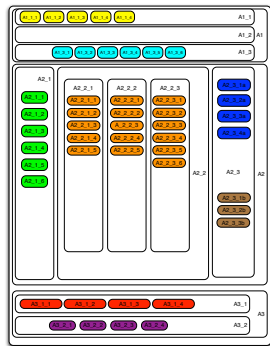
Definition 2. Two lists l and l' are *related* ($l \sim l'$) if they have an element in common. A set of lists \mathcal{S} is a *tilted structure* if for every list $l \in \mathcal{S}$ there exists at least one other list $l' \in \mathcal{S}$ such that $l \sim l'$ and $l \neq l'$. Lists in a tiled structure are called *tilted lists*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$10.00.



(a)



(b)

Figure 1: The Illinois Mathematics Web Page and its box structure

Three tiled lists ($A_{2.2.1}$, $A_{2.2.2}$ and $A_{2.2.3}$) are shown in Figure 1(b). The notion of tiled list is useful to handle more problematic cases, such as $A_{2.2}$, by merging the individual lists of a tiled structure into a single tiled list.

Once visual aligned elements are extracted, we prune out false positives under the hypotheses that the DOM-subtrees corresponding to the elements of the list must satisfy a structural similarity measure ($structSim$) to within a certain threshold α and that the subtrees not have a number of DOM-nodes ($numNodes$) greater than a threshold β :

Definition 3. A candidate list $l = \{l_1, l_2, \dots, l_n\}$ is a genuine list if and only if for each pair (l_i, l_j) , $i \neq j$, $structSim(l_i, l_j) \leq \alpha$, $numNodes(l_i) \leq \beta$, and $numNodes(l_j) \leq \beta$.

This assumption, which is shared with most other DOM-centric mining algorithms, is made to determine whether the visual alignment of a certain boxes can be regarded as a real list or it should be discarded.

3. EXPERIMENTS

HyLiEn requires two parameters which are empirically set to $\alpha = 0.6$ and $\beta = 50$. We used a subset of 100 pages from dataset used in VENTex[2]. This dataset is general and not biased; we take advantage on this dataset to show that our method is robust and the result are not biased from the selected test set.

	Ground truth	VENTex	HyLiEn
# tables	224	82.6%	79.5%
# records	6146	85.7%	99.7%

Table 1: Recall for table and record extraction on the VENTex data set.

Table 1 shows that there are over 224 tables and 6146 data records in the ground truth. VENTex did extract 8 more tables than HyLiEn. We believe this is because HyLiEn does not have any notion of element distance that could be used to separate aligned but separated lists. Despite the similar table extraction performance, HyLiEn extracted many more records (*i.e.*, rows) from these tables than VENTex.

We did judge the precision score here for comparison sake (see Table 2). We find that, from among the 100 Web pages

	Recall	Precision	F-Measure
VENTex	85.7%	78.0%	81.1%
HyLiEn	99.7%	99.9%	99.4%

Table 2: Precision and Recall for record extraction on the VENTex data set.

we achieve 99.9% precision. VENTex remained competitive with a precision of 85.7%. We see that HyLiEn consistently and convincingly outperforms VENTex.

4. CONCLUSIONS

In this paper, a novel fully automatic hybrid approach for extracting Web lists from Web pages is presented. Experimental results show that it outperforms the existing approaches for Web lists and tables extraction. Interesting avenues for future work involve the usage of extracted lists to annotate and discover relationships between entities on the Web or to index the Web, the query answering from lists, and the entity discovery and disambiguation using lists.

5. ACKNOWLEDGMENTS

The first and fourth authors are supported by the Project DIPIS funded by Apulia Region. The second and fifth authors are supported by NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and by the U.S. Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA).

6. REFERENCES

- [1] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang. Webtables: exploring the power of tables on the web. *Proc. VLDB Endow.*, 1(1):538–549, 2008.
- [2] W. Gatterbauer, P. Bohunsky, M. Herzog, B. Krüpl, and B. Pollak. Towards domain-independent information extraction from web tables. In *WWW*, pages 71–80, New York, NY, USA, 2007. ACM.
- [3] K. Lerman, L. Getoor, S. Minton, and C. Knoblock. Using the structure of web sites for automatic segmentation of tables. In *SIGMOD*, pages 119–130, New York, NY, USA, 2004. ACM.
- [4] W. Liu, X. Meng, and W. Meng. Vide: A vision-based approach for deep web data extraction. *IEEE Trans. on Knowl. and Data Eng.*, 22(3):447–460, 2010.
- [5] K. Simon and G. Lausen. Viper: augmenting automatic information extraction with visual perceptions. In *CIKM*, pages 381–388, New York, NY, USA, 2005. ACM.
- [6] S. Tong and J. Dean. System and methods for automatically creating lists. In *US Patent: 7350187*, Mar 2008.
- [7] R. C. Wang and W. W. Cohen. Language-independent set expansion of named entities using the web. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, pages 342–350, Washington, DC, USA, 2007. IEEE.
- [8] T. Weninger, F. Fumarola, R. Barber, J. Han, and D. Malerba. Unexpected results in automatic list extraction on the web. *SIGKDD Explorations*, 12(2), 2010.