Information Extraction from the Web by Matching Visual Presentation Patterns

Matej Minarik, Radek Burget

Brno University of Technology, Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czech Republic

iminarikma@fit.vutbr.cz, burgetr@fit.vutbr.cz

Abstract. There is a large amount of data available on the Web. Data are often represented as text, enriched with tables, lists, images or other visual structures. These data are usually coded in HTML without any additional semantics, which makes them nigh impossible to automatically process and extract. There are approaches based on top-down document segmentation according to visual information and layout. We present a bottom-up approach which starts with the smallest consistent elements and matches the visual relationships among these elements to a pre-defined ontological structure of extracted records. This method considers not only the visual attributes of a particular segment, but also its position amongst other segments.

Keywords: web data integration, information extraction, structured record extraction, page segmentation, content classification, ontology mapping

1 Introduction

There is a great amount of documents on the World Wide Web. These documents are usually HTML documents with lots of data encoded in HTML tags. There are some generic tags available, such as *div* or *span* with no semantics at all. HTML 5 [4] introduced new tags with a generic semantics, such as *article*, *header* or *footer*. Additionally, RDFa annotations [6] (among other possibilities) allow expressing the semantics of the individual document elements by mapping to ontological concepts or properties.

Unfortunately, large amounts of data available on the World Wide Web are still only accessible through plain HTML documents with no semantic annotations at all. When we want to machine-process the contained data, it is necessary to identify the contained data records and recognize the semantics of the individual data fields. For some data sources with a regular structure (such as Wikipedia), we may create specific extraction templates. However, Web is a heterogeneous place with many different sources of data represented in countless ways.

Many methods have been developed recently for identifying and extracting structured data records from plain HTML documents [5, 7, 9]. Most of them are based on a

J. Steinberger, M. Zíma, D. Fiala, M. Dostal, M. Nykl (eds.)
Data a znalosti 2017, Plzeň, 5. - 6. října 2017, pp. 227-231.

top-down approach: first, the document is pre-processed in order to locate the most probable regions of interest (called data sections [9] or data regions [5]). Then, the individual data records are found in the regions based on the detection of repeating structures in the model by frequency measures [7] or visual pattern detection [1, 9].

In order to avoid the complex and often unreliable process of data region identification, we have recently proposed an opposite bottom-up approach [2]. Considering that the expected structure of the records to be extracted is usually known in advance, we try to find a best match among this expected structure and the presentation patterns in the whole document starting from the smallest atomic elements. In this paper, we propose possible future development that aims to bridge the gap between the plain HTML documents and the semantic web.

2 Method Overview

The key idea of the method presented in [2] is to identify visual presentation patterns that are used to represent the data we are trying to extract and then to extract these data using discovered patterns.

This method assumes only the visual representation of the documents, so it may be applied not only on HTML documents, but on PDFs and other visual formats as well. Since we are working with the visual information only, we need to create a uniform representation of the source document. This representation is a set of *visual areas* and it is the first step of the process.

The next step is the initial *tagging* step. In the initial tagging, we assign one or multiple *tags* to each discovered visual area. The idea is to identify all the visual areas in the document that *could* possibly correspond to a particular piece of information (e.g. a personal name, date or e-mail address). This tagging may be based on a variety of approaches starting with simple regular expressions and ending with a kind of *NER* classifier. We admit that there will be incorrect tags and even multiple tags assigned for some visual areas.

After the initial tagging step, we need to disambiguate the tags. This method assumes that all the data records are presented in a visually consistent way in the source documents. This allows introducing presentation constraints on the data records. The disambiguation task itself consists of finding matching record presentation and layout which meets visual constraints and has a great support in source documents. This method defines four visual relations:

- $(a_1, a_2) \in R_{side}$ when a_1 and a_2 are on the same line and a_2 is placed to the right of a_1 with no element between them
- $(a_1, a_2) \in R_{after}$ when a_1 and a_2 are on the same line and a_2 is anywhere to the right of a_1 (there may be elements between them)

- $(a_1, a_2) \in R_{under}$ when a_1 and a_2 are roughly in the same column, a_2 is placed below a_1 with no element between them; additionally, the vertical space is not larger than *0.8 em*
- $(a_1, a_2) \in R_{below}$ when a_1 and a_2 are roughly in the same column, a_2 is placed anywhere below a_1

We choose the most supported relation by trying to cover as many tagged visual areas as possible.

The last step is the *record extraction* itself. The methods results in a set of matches which identify the visual areas that contain data we are interested in. By simply concatenating these areas we can obtain the text content.

3 Further Development Directions

The presented method is based on quite simple visual relationships discovered in the documents and it expects small and regular data records being present in the documents. For adapting the methods for more complex documents and data records, several ways of extensions may be considered.

3.1 Additional Visual Relationships

For more complex documents, we would like to consider more ways the relationships among elements may be presented in the documents [3]. We may try to represent more high-level relations, such as *heading-subheading*, or *paragraph* which is basically a couple of lines separated from other paragraphs in some visual way (usually with a newline or a starting tab). Other interesting relation might be the same row or same column inside a table. Same row might tell us that those values are referring to the same entity. Same column usually represents different values of some metric. On top of that we might try to identify grouped columns, which is frequent for timetables.

3.2 Advanced Tagging

DBPedia might be used for improving the tagging step for example by integrating DBPedia Spotlight, which is a tool for automatically annotating mentions of DBpedia resources in text. This tool is able to identify entity mentions and select the best candidate based on the user-provided configuration. Other than that, we may introduce some NLP tasks as a tagging enhancement. The obvious one is a *named entity recognition* tagger.

3.3 Semantic Relationship Representation

In our recent work [8], we proposed using RDF for representing the visual model of the processed documents. Similarly to this approach, we would like to represent even the discovered visual relationships using a RDF graph. The expected benefits are greater efficiency in discovering frequent visual patterns (for example using SPARQL queries) and the possibility to directly map the identified data fields to ontological properties.

4 Conclusion

In this paper, we have discussed a method of structured record extraction from web documents. The method is purely vision-based and unlike most existing approaches, it does not rely on a complex document pre-processing steps for identifying the data regions in the documents. Instead, it uses an opposite approach that marks all the possible (even incorrectly identified) occurences of the given information in the documents and later, the data records are identified by finding the best match between the expected record structure and the visual presentation patterns discovered in the document. We have proposed possible extensions of this method that are expected to allow processing more complex documents and integrate the extracted results with semantic web resources.

References

1. N. Anderson and J. Hong. Visually extracting data records from query result pages. In Web Technologies and Applications: 15th Asia-Pacific Web Conference, APWeb 2013, Sydney, Australia, April 4-6, 2013. Proceedings, pages 392–403, Berlin, Heidelberg, 2013. Springer.
2. R. Burget. Information extraction from web documents based on visual and semantic relationship alignment. In *NLP&DBpedia 2016*. Springer International Publishing, 2017. *To appear*.
3. R. Burget and P. Smrz. Extracting visually presented element relationships from web documents. *International Journal of Cognitive Informatics and Natural Intelligence*, 2013(2):13–29, 2013.
4. S. Faulkner, I. Hickson, S. Pfeiffer, E. D. Navara, R. Berjon, T. O'Connor, and T. Leithead. HTML5. W3C recommendation, W3C, Oct. 2014.
5. P. L. Goh, J. L. Hong, E. X. Tan, and W. W. Goh. Region based data extraction. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on*, pages 1196–1200, May 2012.
6. I. Herman, S. McCarron, M. Birbeck, and B. Adida. RDFa core 1.1 - third edition. W3C recommendation, W3C, Mar. 2015.
7. J. L. Hong, E.-G. Siew, and S. Egerton. Information extraction for search engines using fast heuristic techniques. *Data Knowl. Eng.*, 69(2):169–196, Feb. 2010.
8. M. Milicka and R. Burget. Multi-aspect document content analysis using ontological modelling. In *Proceedings of 9th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2014)*, pages 9–12. Vydavateĺstvo STU, 2014.

9. D. Weng, J. Hong, and D. A. Bell. Automatically annotating structured web data using a SVM-based multiclass classifier. In *Web Information Systems Engineering – WISE 2014: 15th International Conference, Thessaloniki, Greece, October 12-14, 2014, Proceedings, Part I*, pages 115–124, Cham, 2014. Springer International Publishing.

Acknowledgement: This work was supported by the BUT FIT grant FIT-S-17-39