Multi-aspect Document Content Analysis using Ontological Modelling

Martin Milicka and Radek Burget

Faculty of Information Technology, IT4Innovations Centre of Excellence
Brno University of Technology, Bozetechova 2, 612 66 Brno, Czech Republic
{imilicka,burgetr}@fit.vutbr.cz

Abstract. Existing methods of information extraction from web documents are usually based on a single aspect of the document or its contents such as the code, textual features or visual features. Due to the great variability of the available online documents, it seems reasonable to combine multiple kinds of analysis in order to use all the available knowledge for identifying a particular information in the document. In this paper, we propose an ontological document model that allows to integrate the results of the analysis of different document aspects. We propose a generic architecture of an information extraction system based on this model and we show its applicability on a practical example.

Keywords: document modeling, information extraction, page segmentation, content classification, ontology, RDF

1 Introduction

Information extraction (IE) from web documents is a difficult task mainly because of very loose and variable structure of the documents and lack of available metadata or annotations. Most common IE approaches analyze mainly the the HTML code (DOM), the text of the document (named entity recognition, statistical analysis of the text, etc.) or the visual presentation (page layout and visual features of the presented contents). Usually, only one of these aspects is used. However, the web is diverse: Depending on the nature of the presented information and the target users, the visual hints may be crucial for some web pages while other pages may be primarily text-oriented and the visual presentation plays a secondary role. Therefore, analyzing multiple aspects together seems to be a promising way of research.

Several models have been introduced for representing documents: DOM [5] is a standard for modelling HTML document code. Similarly, CSS [1] defines a formatting model that describes the contents of a rendered page. In the layout analysis area, the page segmentation algorithms usually use specific models for representing the segmentation results [2, 3]. E.g., VIPS [3] represents the segmented page as a hierarchy of visual blocks and separators. The mentioned models are not intended to be shared by multiple applications; there is usually no explicit representation of the model defined that would allow storing a

created model and sharing it among multiple applications or analysis methods. RDF-based models may be used for storing metadata and annotations in PDF files [4].

In this paper, we propose an ontology-based extensible model of web documents that allows to integrate the results of multiple analysis algorithms that include the visual organization of the page (layout) and other visual features (fonts, colors, etc.), results of the visual area classification based on visual features and the results of text classification including the named entity recognition (NER) algorithms. We propose an architecture of an IE system based on this model and we show how the multiple aspect analysis may be used for improving the results of information extraction in the domain of news articles.

2 Ontological Document Model

A document may be described on different levels of abstraction. We define three levels of document description where each level adds a specific knowledge about the document.

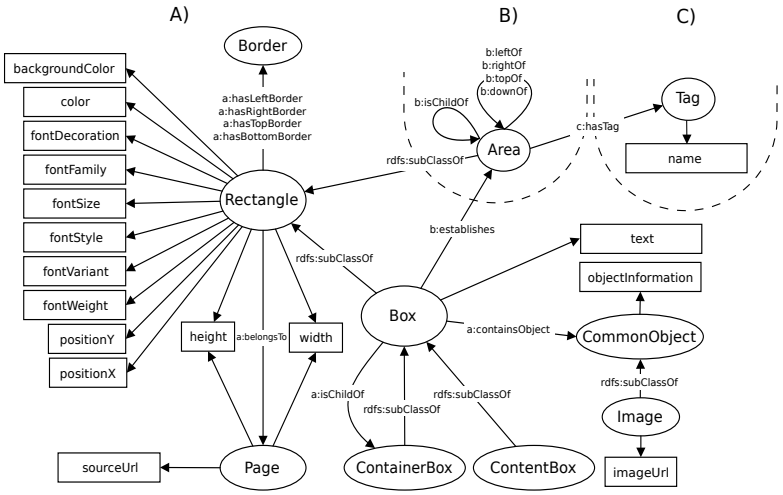

Fig. 1. A) Box model ontology B) Segmentation ontology C) Classification ontology

1. Box model description (*rendered page level description*) represents the output of the page rendering process – visual features of the individual content parts and their positions on the resulting page.
2. *Semantic level* where the box model is extended with an additional semantic information as described below.
3. *Domain description* that represents a connection to the specific domain of the processed documents.

We have designed a set of ontologies that allow representing all the information about a document using RDF. The *Box model ontology* (fig. 1A) represents the box model description. The document is represented as a set of (possibly nested) rendered content *boxes* together with their size, position and visual features.

The remaining ontologies (fig. 1B and 1C) belong to the semantic level. The *Segmentation ontology* extends the Box model ontology by the possibility of representing larger visual areas. Its basic *Area* class represents the visual areas detected during page segmentation. Finally, the *Classification ontology* allows to add a number of classes (tags) to the individual visual areas. The class assignment may be produced by a classification algorithm based on different features (e.g. text classification or visual classification) or manually, e.g. when creating a training set of documents.

3 Model Application for Information Extraction

The architecture of an IE system based on the proposed model is built around a central RDF repository that stores the information about all the processed documents as shown in the figure 2.

During the *model initialization* the source document is transformed to a format-independent RDF model based on our box model ontology. In the *model building* phase, further analysis steps such as content classification or page segmentation are applied on the model (in any order, possibly in several iterations); the results are represented using our semantic level ontologies. Alternatively, some information such as manually annotated classes may be added by the user using an interactive tool (visual editor). Finally, based on the results of the previous analysis steps, we map certain parts of the created model to a domain ontology which is actually the extraction step.

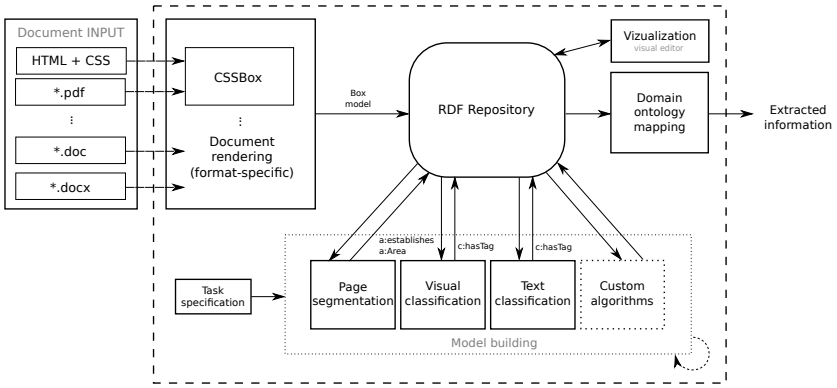

Fig. 2. A generic architecture of an IE system based on the ontological model

For testing the proposed concept, we have chosen the domain of online news articles where the task is to recognize a published article within a larger web page and to distinguish its individual parts such as heading, date of publication, paragraphs, etc. The whole IE process consists of the following steps:

For the model initialization, we have used our CSSBox[1] rendering engine that produces a box model that is later serialized to the RDF description.

The model building phase includes page segmentation, text classification based on NER (for recognizing names, places, dates, etc.) and visual classification based on the visual features of the content as described in [2]. During these steps, each detected visual area is assigned a set of *tags* that indicate the probability that the given area represents a certain part of the article. Based on the assigned tags, the areas are finally mapped to a simple Article domain ontology that models an article and its its individual parts.

Our preliminary experiments run on the *reuters.com* and *cnn.com* news portals show, that the combination of several classification methods may increase the IE precision in comparison to a single-aspect classification published in [2].

4 Conclusions

We have proposed an ontological document model suitable for the description of different aspects of web documents on several levels of abstraction. The model allows sharing all the knowledge about the document and its contents among multiple analysis methods and combine their results. We have also shown the general architecture of an IE system based on this model and we have shown its applicability in a particular domain. The actual precision of the information extraction depends on the quality of results of the individual analysis methods, the way they are combined and the used method of domain ontology mapping.

This work was supported by the BUT FIT grant FIT-S-14-2299 and the IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

References

1. Bos, B., Lie, H.W., Lilley, C., Jacobs, I.: Cascading Style Sheets, level 2, CSS2 Specification. The World Wide Web Consortium (1998)
2. Burget, R., Rudolfová, I.: Web page element classification based on visual features. In: 1st Asian Conference on Intelligent Information and Database Systems ACIIDS 2009. pp. 67–72. IEEE Computer Society (2009)
3. Cai, D., Yu, S., Wen, J.R., Ma, W.Y.: VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Research (2003)
4. Eriksson, H.: The semantic-document approach to combining documents and ontologies. Int. J. Hum.-Comput. Stud. 65(7), 624–639 (Jul 2007)
5. Hors, A.L., Hgaret, P.L., Wood, L., Nicol, G., Robie, J., Champion, M., Byrne, S.: Document Object Model (DOM) Level 3 Core Specification. The World Wide Web Consortium (2004)

[1] http://cssbox.sourceforge.net/