Chapter 2
Models and Approaches for Web Information Extraction and Web Page Understanding

Ruslan R. Fayzrakhmanov
Vienna University of Technology, Austria

ABSTRACT

This chapter discusses the main challenges addressed within the fields of Web information extraction and Web page understanding and considers different utilized Web page representations. A configurable Java-based framework for implementing effective methods for Web Page Processing (WPP) called WPPS is presented as the result of this analysis. WPPS leverages a Unified Ontological Model (UOM) of Web pages that describes their different aspects, such as layout, visual features, interface, DOM tree, and the logical structure in the form of one consistent model. The UOM is a formalization of certain layers of a Web page conceptualization defined in the chapter. A WPPS API provided for the development of WPP methods makes it possible to combine the declarative approach, represented by the set of inference rules and SPARQL queries, with the object-oriented approach. The framework is illustrated with one example scenario related to the identification of a Web page navigation menu.

INTRODUCTION

Information is an inalienable part of today's life. This fact is clearly evident in the ongoing development and expansion of the World Wide Web (the Web)—a huge information platform that has provided vast opportunities for people by making it possible to effectively solve various tasks in business, education, science, and our everyday lives. With the help of the Web, a person can pay bills, buy products and services, complete university degrees online, search for and read articles, keep contact with their friends and so much more.

The Web contains a vast amount of information represented mainly on web pages in unstructured and semi-structured forms. Web resources (i.e. web pages) are primarily intended for human consumption and thus their information content is not accessible for automatic processing. The necessity of developing methods for *web page understanding (WPU)* and wrappers for *web information extraction (WIE)* is based on the need for

DOI: 10.4018/978-1-4666-7262-8.ch002

this information in computer-aided systems (e.g. web form understanding for the meta-search or extraction of prices and sentiments for the competitive intelligence) and the implementation of different aspects of the Semantic Web and relevant use cases (e.g. connecting recognized entities on web pages with open data sources through Linked Data technology, improving the performance of information retrieval and query answering systems). Many methods and approaches for WPU and WIE (hereinafter referred to as *methods*) have been developed that target different forms of web page representation: the source code (X/HTML, XML), DOM tree (or tag tree), and visual representation rendered by the web browser engine (e.g. Firefox's Gecko, Chrome's Blink or Internet Explorer's Trident). Each of these aspects of a web page has its purpose, advantages and disadvantages; however, consideration of web page visual models is known to ensure the development of more robust and effective methods which can be applied over a wider range of web pages (Fayzrakhmanov, 2013, sec. 2.4). This is due to the fact that merely the visual representation analyzed by the user exclusively reflects the semantics and logical structure of a web page. Furthermore, the analysis of visual cues also gives a unique possibility to leverage certain principles and laws of Gestalt theory which in turn reflects processes of human object recognition (Krüpl-Sypien, Fayzrakhmanov, Holzinger, Panzenböck, & Baumgartner, 2011; Xiang, Yang, & Shi, 2007) for developing more robust methods.

A conceptual gap between the source code (i.e. the XML, X/HTML code and thus the DOM tree) and layout structure has been growing even larger (Oro, Ruffolo, & Staab, 2010) in recent years, forcing developers and researchers increasingly often to focus on visual features rather than the source code. This tendency is related to the use of various front-end technologies from the open web stack, such as X/HTML, CSS and JavaScript, in the web development process. These technologies thus impart a property of application with rich functionality to the contemporary web pages. Therefore, their automatic analysis should be performed on their rendered state, taking into account their visual and functional aspects.

In the absence of a standard to describe a web page's visual appearance suitable for WPU and WIE, the development of new methods generally encounters the challenge of defining necessary features and relationships. To overcome this problem and provide a convenient means for developing new methods and approaches, a *Web Page Processing System (WPPS)* was developed along with the underlying *Unified Ontological Model (UOM)* which formalizes the most required aspects of the web page conceptualization introduced in this chapter. The proposed UOM describes different aspects of a web page—its interface (web forms, links, images, etc.), layout, perceptible visual features, DOM tree, and logical structure—in one consistent and easily extensible model. During the development of the UOM, different styles of representing the layout of web pages (Kong, Zhang, & Zeng, 2006; Kovacevic, Diligenti, Gori, & Milutinovic, 2004; Oro et al., 2010), PDF (Hassan, 2010), and scanned documents (Aiello et al., 2002) were considered. WPPS is a means for developing new, effective and robust methods analyzing different forms of web page representations and profiting from both declarative and object-oriented approaches by employing the introduced bridged adapter software design pattern.

The analysis presented in this chapter is mainly based on the work of Fayzrakhmanov (2013), where the interested reader can not only find a detailed description of various concepts and aspects presented, but an underlying theory as well.

This chapter first introduces the term *web page processing (WPP)* and its relation to the fields of WIE and WPU, then conducts a comparative analysis of different approaches in terms of leveraged web page models. It presents a conceptualization of the web page and the UOM as a formalization of selective aspects of the conceptualization. The