18

Web Document Analysis

Apostolos Antonacopoulos and Jianying Hu

18.1 Introduction

The Web has now become a very popular information repository where content providers and end-users routinely publish and access, respectively, documents on-line. The availability and extent of this vast and varied information gives rise to the need for automated content analysis. Such content analysis is crucial for applications such as information extraction, web mining, summarization, content re-purposing for mobile and multi-modal access and web security.

There are two broad categories of web documents, those that are intended and designed as web services and those that are created as web publications, where visual appearance (to human users) is critical. The development of XML and the new initiatives on the Semantic Web aim to improve the machine-readability of web documents. This semantic tagging makes content analysis rather straightforward for the documents that conform to such standards. However, while such conformance is generally true for the former type of documents, it is far less common in the latter (web publications). This fact is unlikely to change as web publications are produced by a multitude of the widest variety of authors (practically anyone can create web documents and publish them on their website). It is, therefore, correspondingly unlikely that the challenges for content analysis will diminish significantly.

The emerging issues pose new challenges (and opportunities) for Document Analysis, in a number of traditional as well as new areas. This chapter presents an overview of a number of diverse and interdisciplinary areas that reflect current research directions. The following section discusses the broad fundamental topics of web content extraction, repurposing and mining. Section 18.3 focusses on issues related to images (and their

content) encountered in web documents. Finally, the complementary areas (to analysis) of web document modelling and annotation are discussed in Section 18.4, before the chapter concludes with Section 18.5.

18.2 Web Content Extraction, Repurposing and Mining

Web content extraction refers to the process of identifying and retrieving any specific information from a web page. It goes beyond traditional information extraction which focusses on text analysis using Natural Language Processing techniques [30] and aims to make use of the structural information embedded in the markup language. On the one hand, web pages are always encoded in a markup language such as HTML or XHTML which could potentially contain rich structural information. On the other hand, for many web pages, particularly those created as on-line publication, markup tags are primarily used to create a specific display and thus do not directly correspond to any semantics or formal relations. Thus web documents are often referred to as "semi-structured" documents, and much effort in web content extraction has been focussed on inferring semantics from the markup tags.

One of the earliest works in this area was a wrapper induction system developed in 1997 [24]. The system generates extraction rules using a designated set of HTML tags as delimiters for document head, tail and data tuples for specific content such as weather or restaurant. Since then, much effort has been made by various researchers to develop wrapper induction systems that can accommodate more variations in formatting (e.g. [14, 19, 30, 31]). One limitation of this line of systems is that they all treat a marked up document as a sequence and thus the analysis is inherently local, making it difficult to infer structural relations between items places far apart in the text stream.

In order to take advantage of document structures at higher levels, researchers started analysing the HTML parse tree, or directly the DOM tree of a web document [1, 11, 13, 29, 33]. While the DOM tree provides a much more global view of the document structure, it is still insufficient because there are often multiple ways of arranging different markup tags to achieve the same appearance. As a result, a visually regular document may be highly irregular in its DOM structure, or multiple structural patterns of the DOM tree may correspond to the same visual pattern. Cohen et al. identified the problem and used pre-processing to normalize some of the common variations [11]. Others used features designed to directly infer visual characteristics of documents. Wang and Hu developed layout features to measure the visual coherence among potential table cells [45]. Yang et al. introduced visual similarity measures based on attributes such as size and colour of various DOM objects [46]. Because the rendering of a web