Chapter 1.19
Web Information Extraction via Web Views

Wee Keong Ng
Nanyang Technological University, Singapore

Zehua Liu
Nanyang Technological University, Singapore

Zhao Li
Nanyang Technological University, Singapore

Ee Peng Lim
Nanyang Technological University, Singapore

ABSTRACT

With the explosion of information on the Web, traditional ways of browsing and keyword searching of information over web pages no longer satisfy the demanding needs of web surfers. Web information extraction has emerged as an important research area that aims to automatically extract information from target web pages and convert them into a structured format for further processing. The main issues involved in the extraction process include: (1) the definition of a suitable extraction language; (2) the definition of a data model representing the web information source; (3) the generation of the data model, given a target source; and (4) the extraction and presentation of information according to a given data model. In this chapter, we discuss the challenges of these issues and the approaches that current research activities have taken to revolve these issues. We propose several classification schemes to classify existing approaches of information extraction from different perspectives. Among the existing works, we focus on the WICCAP system — a software system that enables ordinary end-users to obtain information of interest in a simple and efficient manner by constructing personalized web views of information sources.

INTRODUCTION

The World Wide Web has become such a successful channel in delivering and sharing information that people are getting used to searching the Web as the first resort for information. As the amount of data accessible via the Web grows rapidly, the weaknesses of traditional ways of browsing and searching the Web become more and more apparent (Laender, 2002a). Browsing requires users to follow links and to read (usually) long web pages, thus making it tedious and difficult to find a particular piece of information. Keyword searching usually returns massive irrelevant information, along with some useful information hidden in the long list of search results. Even with improved search engines, such as Google, that return accurate results, a large number of web pages cannot be indexed by these engines. Therefore, users surfing the Web with these traditional facilities have been facing the information overload problem; they are overloaded with too much irrelevant information.

As HTML web pages are designed to be viewed by humans, most of the HTML syntax is for presentation purposes and does not contain much semantic meaning; this makes automatic access by software applications difficult. However, there is an increasing demand to turn web data into structured and machine-readable formats so that further processing, such as integration, filtering and customized visualization, can take place.

To address the problems mentioned above, over the past few years, some web information extraction (IE) systems (mainly in the form of wrappers) (see Adelberg, 1998; Ashish, 1997; Baumgartner, 2001; Crescenzi, 2001; Hammer, 1997; Kushmerick, 2000; Liu, 2000; Liu, 2002; Liu 2002a; Mecca, 1999) have been developed to automatically extract target information from the Web and convert the extracted data into some structured format. The approaches taken by these systems differ greatly, ranging from Natural Language Processing (NLP) to machine learning to database techniques. Despite the differences in approaches, there are several common issues that these systems need to address: (1) the definition of a suitable extraction language; (2) the definition of a data model representing the web information source; (3) the generation of data models, given a target source; and (4) the extraction and presentation of the information according to a given data model.

Objectives

This chapter aims to provide an in-depth analysis of the above issues and of how the existing approaches address them. This chapter is not intended to be a simple survey of existing web IE systems, which has been done in Laender (2002a), where a brief description of those systems is given and a simple classification is proposed. The focus of this chapter is to look into the details of each important issue mentioned above, to discuss how the issue can be handled, and to analyze approaches taken by current systems and how effective they are in solving the problems. In addition, several classification schemes are proposed in order to classify these existing systems and to help understand the issues that they try to resolve.

To further illustrate the issues, a detailed description of one of the systems called WICCAP (see Li, 2001; Liu, 2002) is provided. The aim of the WICCAP system is to enable ordinary users to create their own views of the target web sites in a simple and easy manner so that information extraction from web sites can be performed automatically.

It should also be pointed out that the focus of this chapter is on academic research projects. For a brief survey on related commercial products, the readers are referred to Kuhlins (2002).