# Chapter 14
# Web Harvesting:
## Web Data Extraction Techniques for Deep Web Pages

**B. Umamageswari**
*New Prince Shri Bhavani College of Engineering and Technology, India*

**R. Kalpana**
*Pondicherry Engineering College, India*

## ABSTRACT

*Web mining is done on huge amounts of data extracted from WWW. Many researchers have developed several state-of-the-art approaches for web data extraction. So far in the literature, the focus is mainly on the techniques used for data region extraction. Applications which are fed with the extracted data, require fetching data spread across multiple web pages which should be crawled automatically. For this to happen, we need to extract not only data regions, but also the navigation links. Data extraction techniques are designed for specific HTML tags; which questions their universal applicability for carrying out information extraction from differently formatted web pages. This chapter focuses on various web data extraction techniques available for different kinds of data rich pages, classification of web data extraction techniques and comparison of those techniques across many useful dimensions.*

## INTRODUCTION

The information available on the World Wide Web has grown to several zettabytes according to Richard Currier (2013). Estimated size of pages indexed in Google in the last three months is shown in Figure 1. The structured information such as lists and tables containing the target data of interest is embedded in semi-structured web pages which complicates automated extraction.

Many mining applications depend on the data available in this huge repository. The process of automatically retrieving data from websites is known as web data extraction aka Web scraping or Web harvesting. Applications include business intelligence, product intelligence, market intelligence, data analytics, data mashup, meta-search, meta-query etc. Information on WWW is available in different forms. The classification is shown in Figure 2.

DOI: 10.4018/978-1-5225-0613-3.ch014

*Figure 1. Size of pages indexed in Google (http://www.worldwidewebsize.com/)*

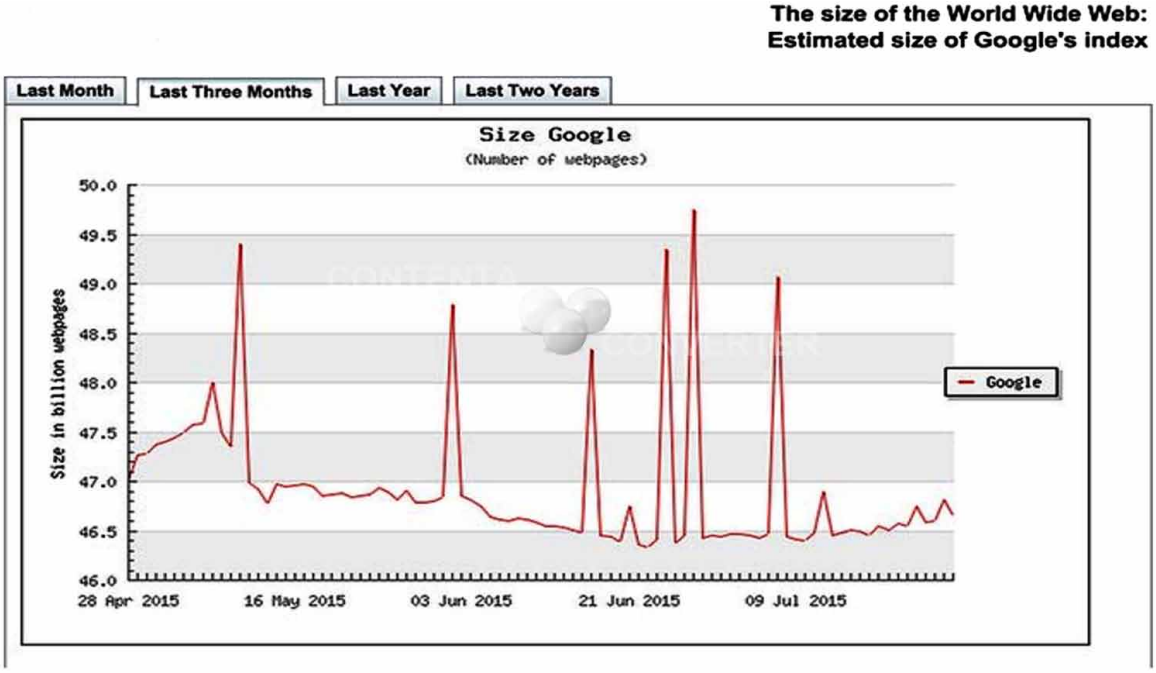
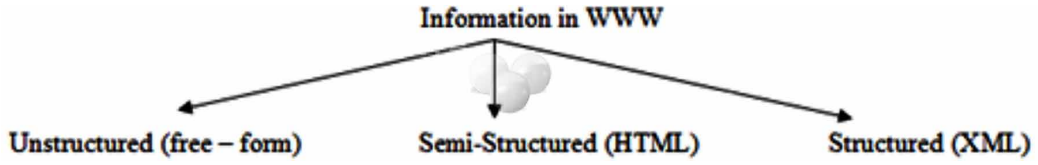



*Figure 2. Classification of different forms of information available on WWW*




Techniques used for information extraction depends upon the representation of information in WWW. Text mining is a domain which focuses on processing unstructured information. On the other end, extraction of data from structured representation such as XML document can be handled easily using several APIs for ex., JAXP. Our focus is on semi-structured information presented in the form of HTML pages. HTML (Hyper Text Markup Language) is used initially for formatting data and therefore the information is not structured which makes the extraction task cumbersome. Many techniques have been proposed to perform information extraction from HTML pages. Detailed discussion of such techniques is available in the next section.

HTML documents can be represented in various forms. 1. String/Text – Source code of HTML document where a web page is represented using its source code. Certain data extraction tools like OLERA and Trinity etc. use the source code and apply string comparison techniques for information extraction. 2. DOM tree – Many techniques in the literature such as Thresher, DELA, MDR, TPC etc. use DOM tree representation for extraction of information from HTML pages where an HTML page is represented as a tree of HTML elements. 3. Visual (CSS Box) – When the HTML documents are rendered by the browser, each tag is represented as a CSS box. The position and size of box are used by some techniques