# Chapter 7
# A Novel Architecture for Deep Web Crawler

**Dilip Kumar Sharma**
*Shobhit University, India*

**A. K. Sharma**
*YMCA University of Science and Technology, India*

## ABSTRACT

*A traditional crawler picks up a URL, retrieves the corresponding page and extracts various links, adding them to the queue. A deep Web crawler, after adding links to the queue, checks for forms. If forms are present, it processes them and retrieves the required information. Various techniques have been proposed for crawling deep Web information, but much remains undiscovered. In this paper, the authors analyze and compare important deep Web information crawling techniques to find their relative limitations and advantages. To minimize limitations of existing deep Web crawlers, a novel architecture is proposed based on QIIIEP specifications (Sharma & Sharma, 2009). The proposed architecture is cost effective and has features of privatized search and general search for deep Web data hidden behind html forms.*

## 1. INTRODUCTION

Traditional Web crawling techniques have been used to search the contents of the Web that is reachable through the hyperlinks but they ignore the deep Web contents which are hidden because there is no link is available for referring these deep Web contents. The Web contents which are accessible through hyperlinks are termed as surface Web, while the hidden contents hidden behind the html forms are termed as deep Web. Deep Web sources store their contents in searchable databases that produce results dynamically only in response to a direct request (Bergman, 2001). The deep Web is not completely hidden for crawling. Major traditional search engines can be able to search approximately one-third of the data (He, Patel, Zhang, & Chang, 2007) but in order

DOI: 10.4018/978-1-4666-2157-2.ch007

to utilize the full potential of Web, there is a need to concentrate on deep Web contents since they can provide a large amount of useful information. Hence, there is a need to build efficient deep Web crawlers which can efficiently search the deep Web contents. The deep Web pages cannot be searched efficiently through traditional Web crawler and they can be extracted dynamically as a result of a specific search through a dedicated deep Web crawler (Peisu, Ke, & Qinzhen, 2008; Sharma & Sharma, 2010). This paper finds the advantages and limitations of the current deep Web crawlers in searching the deep Web contents. For this purpose an exhaustive analysis of existing deep Web crawler mechanism is done for searching the deep Web contents. In particular, it concentrates on development of novel architecture for deep Web crawler for extracting contents from the portion of the Web that is hidden behind html search interface in large searchable databases with the following points:

- Analysis of different existing algorithms of deep Web crawlers with their advantages and limitations in large scale crawling of deep Web.
- After profound analysis of existing deep Web crawling process, a novel architecture of deep Web crawling based on QIIIEP (query intensive interface information extraction protocol) specification is proposed (Figure 1).

This paper is organized as follows: In section 2, related work is discussed. Section 3 summarizes the architectures of various deep Web crawlers. Section 4 compares the architectures of various deep Web crawlers. The architecture of the proposed deep Web crawler is presented in section 5. Experimental results are discussed in section 6 and finally, a conclusion is presented in section 7.

## 2. RELATED WORK

Deep Web stores their data behind the html forms. Traditional Web crawler can efficiently crawl the surface Web but they cannot efficiently crawl the deep Web. For crawling the deep Web contents various specialized deep Web crawlers are proposed in the literature but they have limited capabilities in crawling the deep Web. A large volume of deep Web data is remains to be discovered due to the limitations of deep Web crawler. In this section existing deep Web crawlers are analyzed to find their advantages and limitations with particular reference to their capability to crawl the deep Web contents efficiently.

### Application/Task Specific Human Assisted Approach

Various crawlers are proposed in literature to crawl the deep Web. One of the deep Web crawler architecture is proposed by Raghavan and Garcia-

*Figure 1. Mechanism of QIIIEP-based deep web crawler*