Chapter 5

Dilip Kumar Sharma
G.L.A. Institute of Technology and Management, Mathura, India

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

ABSTRACT

Web crawlers specialize in downloading web content and analyzing and indexing from surface web, consisting of interlinked HTML pages. Web crawlers have limitations if the data is behind the query interface. Response depends on the querying party’s context in order to engage in dialogue and negotiate for the information. In this paper, the authors discuss deep web searching techniques. A survey of technical literature on deep web searching contributes to the development of a general framework. Existing frameworks and mechanisms of present web crawlers are taxonomically classified into four steps and analyzed to find limitations in searching the deep web.

INTRODUCTION

The versatile use of the internet has proved a remarkable revolution in the history of technological advancement. Accessibility of web pages starting from zero in 1990 has reached more than 1.6 billion during 2009. It is like a perineal stream of knowledge. The more we dig the more thirst can be quenched. Surface data is easily available on the web. Surface web pages can be easily indexed through conventional search engine. But the hidden, invisible and non-indexable contents which cannot be retrieved through conventional methods used for surface web and whose size is estimated to be thousands of times larger than the surface web is called deep web. The deep web consist of a large database of useful information such as audio, video, images, documents, presentations

DOI: 10.4018/978-1-4666-0023-2.ch005
and various other types of media. Today people really heavily depend on internet for numerous applications such as flight and train reservations, to know about new product or to find any new locations and job etc. They can evaluate the search result and decide which of the bits or scraps reached by the search engine is most promising (Galler, Chun, & An, 2008).

Unlike the surface web, the deep web information is stored in searchable databases. These databases produce results dynamically after processing the user request (BrightPlanet.com LLC, 2000). Deep web information extraction first uses the two regularities of the domain knowledge and interface similarity to assign the tasks that are proposed from users and chooses the most effective set of sites to visit by ontology inspection. The conventional search engine has limitations in indexing the deep web pages so there is a requirement of an efficient algorithm to search and index the deep web pages (Akilandeswari & Gopalan, 2008). Figure 1 shows the barrier in information extraction in the form of search form or login form.

**Contributions:** This paper attempts to find the limitations of the current web crawlers in searching the deep web contents. For this purpose a general framework for searching the deep web contents is developed as per existing web crawling techniques. In particular, it concentrates on survey of techniques extracting contents from the portion of the web that is hidden behind search interface in large searchable databases with the following points.

- Comparison of different algorithms web searching with their advantages and limitations.
- Discuss the limitations of existing web searching mechanisms in large scale crawling of deep web.

### CURRENT DEEP WEB INFORMATION RETRIEVAL FRAMEWORK

After exhaustive analysis of existing deep web information retrieval processes, a deep web information retrieval framework is developed, in which different tasks in deep web crawling are identified, arranged and aggregated in sequential manner. This framework is useful for understanding entire working of deep web crawling mechanisms as well as it enables the researcher to find out the limitations of present web crawling mechanisms in searching the deep web. The taxonomical steps of developed framework can be classified into following four major parts.

#### Figure 1. Query or credentials required for contents extraction
Related Content

**Studying and Analysis of a Vertical Web Page Classifier Based on Continuous Learning Naïve Bayes (CLNB) Algorithm**
[www.igi-global.com/article/studying-analysis-vertical-web-page/2625?camid=4v1a](www.igi-global.com/article/studying-analysis-vertical-web-page/2625?camid=4v1a)

**Virtual Web Services: Extension Architecture to Alleviate Open Problems in Web Services Technology**
[www.igi-global.com/chapter/virtual-web-services/37747?camid=4v1a](www.igi-global.com/chapter/virtual-web-services/37747?camid=4v1a)

**Medical Tourism Service Quality: An Indian Perspective**
[www.igi-global.com/chapter/medical-tourism-service-quality/140859?camid=4v1a](www.igi-global.com/chapter/medical-tourism-service-quality/140859?camid=4v1a)

**Impact of Internet Usage in Saudi Arabia: A Social Perspective**
[www.igi-global.com/article/impact-internet-usage-saudi-arabia/2628?camid=4v1a](www.igi-global.com/article/impact-internet-usage-saudi-arabia/2628?camid=4v1a)