ABSTRACT

The World Wide Web (WWW), the largest and most frequently accessed public repository of information ever developed, contains large number of web pages interconnected through hyperlinks. It has grown from a few thousand pages in 1993 to more than a couple of billion pages at present. It is easily accessible through internet connection, though it is often more difficult to locate the resources that are relevant to the given need. Due to this problem of *information overkill*, the development of newer and powerful information retrieval tools (search engines) assume importance.

A Web search engine searches for information consisting of web pages, images and other types of files lying on the WWW. In order to maintain index at the search engine's side, a search engine, generally, employs a web crawler to download web pages.

The crawler, on its turn, is given a starting set of URLs corresponding to which pages are downloaded from the web. Before storing the web pages, the crawlers extract URLs appearing in the retrieved pages and pass them to a crawler control module which determines what links to visit next and gives the selected links back to the crawler and so on. The downloaded pages are stored into a page repository.

The Web can be divided into two parts: Surface Web and Deep Web. The Surface Web refers to the static Web pages that can be crawled and indexed by popular search engines, also termed as Publically Indexable Web (PIW). On the other hand, the Deep Web refers to the contents stored in Web databases and published by dynamic Web pages wherein people access web databases through specified query interfaces.

Infact, there are more than 300,000 Deep Web databases and 450,000 query interfaces available in the hidden web and the two figures are still increasing quickly. Besides the scale of Web databases, the contents in Web databases span well across all topics ranging from agriculture to nuclear domain. Some Deep Web portal services provide Deep Web directories that classify Web databases in some taxonomies, contain large amount of high

quality information. However, these sites hidden behind search interfaces can not be crawled by traditional crawlers. Infact, crawling hidden Web is a very challenging problem especially because of following two fundamental reasons:

- Access to these databases is provided only through restricted search interfaces, intended to be filled manually.
- Besides the access through search interfaces, the shear size of the hidden web is too large i.e. about 400 to 500 times larger than the size of the *Surface Web*. As a result, it is not prudent to attempt comprehensive coverage of the hidden Web and therefore, there is need to develop a domain-specific crawler for hidden Web.

In this thesis, design and development of a novel framework for an Extensible and Scalable Domain-Specific Hidden Web Crawler (DSHWC) is being reported. It is not only capable of crawling the hidden web but can also efficiently deal with databases hidden behind search interfaces containing single and multiple attributes as well.

CONTENTS