Chapter 6
Conclusions and Future Directions of Work

6.1. Introduction

The World Wide Web, which is an up-to-date and expanding ocean of information interconnected web pages that can be accessed through the internet. According to\(^1\) Web contains about 4.56 million indexed web pages by April 2014. From this figure, it is clear that an enormous amount of data is present on the internet. In order to get value out of this data and present the information to the user in a very simple form, researchers are working hard to collate this data which is spread across multiple web pages. As the data is spread across multiple web pages, manual copy-paste is impossible at large scale. This colossal size of data on the internet makes it essential and vital to develop automatic methods for extraction of data from multiple web pages.

This chapter thus concludes the thesis, discusses its key contributions and outlines extensions for future research. This chapter is divided into three main sections. Section 6.2 presents the main contributions of the current work. Section 6.3 discusses the directions for future work, and finally section 6.4 takes the thesis to the conclusion.

6.2. Contributions

Web data is the most important source for decision making and marketing now-a-days. In this work, we have firstly presented a comparative study of the web data extraction approaches from two view points. The aim of the web data extraction approach is to convert the web data into a relational database which can be efficiently accessed. The

\(^1\)worldwidewebsize.com
extraction of data records from web pages can be highly profitable for both general users and special users. The major contributions of our work can be outlined as follows:

I). We have provided a detailed survey of the existing web data extraction approaches based on the degree of automation and degree used by each approach to generate wrapper and to the best of our knowledge we were first to outline, classify and explain these approaches in detail.

II). We have proposed a Fully-automatic method known as ESD for extraction of data records from a single webpage and each stage has been discussed in detail.

III). We have proposed a Fully-automatic method for extraction of data records from multiple e-commerce websites. Extraction of data from e-commerce websites is becoming more valuable as extracting competitor’s product information has several advantages.

IV). We have proposed a customer behaviour model, based for prognostication of the customer. Our method is comparatively good at predicting the precision for buying behaviour. By predicting customer behaviour, we can understand who are our best customers, so that we can connect with them in a friendly way.

6.3. Future Directions

In the present work, we had studied the problem of extraction of data records automatically from web pages. This thesis discusses the two main methods for extracting data from a single webpage as well as from multiple web pages. Although our thesis has made significant contributions in web data extraction, but still a number of issues are there that need to be explored and improved in future. These are as listed below:

I). Our ESD method, an automatic method for extracting data from a single page only extracts the string data items, such as product name and product price and do not extract the image data item such as product image of the data records.
So, in the near future we will improve our method to extract all of the data items.

II). As the Web is growing continuously with each passing day, there is a need for more efficient web Crawling. To handle this issue, there is a need for robust and vigorous distributed web crawler.

III). To clean the noisy and ill formatted tags of an HTML webpage, we need to improve the pre-processing phase as a cleaned HTML page has a positive effect on the accuracy.

IV). There is a big room to correctly extract the data records, so we need to improve our string calculation technique.

V). Due to the heterogeneous nature of web data, i.e. web contains structured, semi-structured and unstructured data. In near future we will try to perform extraction on other types of data as well, such as extraction from web blogs.

VI). As we know Bioinformatics is an emerging field of application of web data extraction. In this field, extracting information from scientific works to create a graph will help scientists in extracting data from genes and proteins. But such type of information is present not only in web pages, but in PDF or PostScript forms. So in near future we will extend our work in this potential area.
6.4. Publications out of this Thesis

**Journal Paper(s):**


**Communicated Paper(s):**


**International Conference Paper(s):**

National Conference Paper(s):

