CHAPTER 3
AIMS AND OBJECTIVES

The aim of this thesis is to propose efficient Web data Extraction approaches which provides efficient results in terms of Web data extraction parameters i.e. accuracy and data extraction time. This chapter provides the research objectives chosen for this thesis. These objectives are selected according to the aim of the thesis and issues found in Web Data Extraction.

Section 3.1 provides the objectives along with the little description and issue it solves.

3.1 OBJECTIVES
A Research objective is chosen to solve a particular issue. Following objective were selected as solution for various issues which evolves in web data extraction:

i) To propose and implement pattern discovery approach for Web Content Mining.

Web Content Mining is used to extract data from the Web sources according to a particular pattern. Pattern discovery means to extract data according to a particular pattern from a Web source so that data can be extracted efficiently. Techniques like statistical, neural network etc. can be used for this purpose. This objective is chosen according to issue Rate of growth of information over the Websites. A huge amount of information is available over the Website hence to provide accuracy level in considerable amount of time is a great challenge. This objective chooses a particular pattern from the Website to get solution of this issue.

ii) Develop a Web Usage Mining based approach for efficient data extraction.

Weblog consist of the accessing record of a Webpage or Website by the users. By knowing the accessibility record, useful data can be extracted from the Web sources. It can also be used for publishing user choice at priority over the Websites.
This objective is chosen according to issue *publishing data according to user choice*. It checks the user choice from the Website and provides priority to this data as comparable to other data of the Website.

iii) **To propose a noise reduction approach for improving the efficiency of data extraction from WebPages.**

All elements of a Webpage except the desirable information are known as noise (like links, advertisements, scroll bars etc.). The reduction or removal of noise from the WebPages leads to efficient Web data extraction. To remove the noise webpage must be transformed from unstructured form to structured form. This objective is chosen according to issue *Noise in WebPages*. The noise removal from WebPages helps in Search Engine Optimization (SEO).

iv) **Performance comparison of various existing data extraction techniques in terms of parameters relevant to Web environment.**

A lot of different methods are proposed by different researchers on the basis of techniques available for Web data extraction. Some of these methods will be chosen and compared on the basis of efficiency parameters available for web data extraction.

This objective implements already existing techniques and compare them to check which technique can be applied in which scenario or situation as per the requirement.