ABSTRACT

As per the tremendous usage of Internet from time to time, the information or data available over the Websites has also grown to a significant level. More and more information is provided over the Websites as per the market requirement and hence, the necessity of Web data extraction system over the Websites comes into demand.

Web Content Mining is used to extract data from Web sources according to a particular pattern. Pattern discovery means to choose a particular pattern or format so that data can be extracted efficiently from the Web source. Techniques like template matching, statistical, neural network etc. can be used for this purpose. The data extraction approaches normally uses keyword matching (meta-data) based data extraction techniques. To increase the accuracy of content extracted, content matching based extraction approaches are required but it takes significant amount of time or applicable over to small scale websites. Hence, to extract data from large scale websites using content matching becomes a great challenge to get accurate information at the expected time.

Web usage mining is used to record user behavior. Weblog consists of the accessing or navigation record of a Webpage or Website by the users. By knowing the accessibility record, valuable data can be extracted from the Web sources. It can also be used for publishing user choice at priority over the Websites. Weblogs consist of duplicate entries for the same content and hence content extracted from Weblogs must be checked for duplicate entries and these weblogs must also be cleared from time to time to preserve the required records before clearing the weblog. Secondly whenever a user extract data from a Web source, only 10% of the total results found are viewed by the user and rest is not useful in most of the cases.

A Webpage is made up of a lot of content blocks out of which only a part of the information is useful for a particular application and the remaining information is noise. The efficiency of the Web Data Extraction can be improved by extracting useful after removing the noise part. As the nature of data available on the WebPages is unstructured or semi-structured in nature, hence direct queries as in structured form cannot be applied. Hence, for noise removal web data extraction approaches should first develop the DOM tree to provide a structured Webpage.
Total time taken to extract the content is time to develop the DOM tree and time to extract the content.

In nut shell, the work presented in this thesis is an attempt to increase the efficiency of web data extraction techniques in terms of accuracy and extraction time of desired information from the WebPages irrespective of the fact that whether the data on webpage is unstructured or semi structured in nature.