Abstract

This paper deals with Web Content Mining. While browsing the web, the user has to go through many pages of the Internet, filter the data and download related documents and files. This task of searching and downloading is time consuming. Sometimes the search queries call for specific option, say, limiting search to few links. To reduce the time spent by users, a web
extraction and storage tool has been designed and implemented in Java, that automates the downloading task from a given user query. The Test Scenario has been presented with various keywords. The present work can be a useful input to Web Users, Faculty, Students and Web Administrators in a University Environment.

Reference

- HTMLCleaner Team, HTMLCleaner, http://htmlcleaner.sourceforge.net/
- James Clark and Steve DeRose, W3C XPath Specifications, http://www.w3.org/TR/xpath/

- XPath Tutorial, http://www.w3schools.com/xpath/default.asp

Index Terms
Computer Science
Information Retrieval

Key words
Content Mining
Data Extraction
HTML
Web Data Retrieval
Web Information Extraction.