Development of Syllabus Based Web Content Extractor (SBWCE)

Saba Aziz Hilal

Dr. S.A.M. Rizvi (Ph.D. Supervisor), **Department of Computer Science, Jamia Millia Islamia, Delhi**

Abstract

Huge amount of data is stored and maintained on the web and Internet has become a powerful means for accessing this data. There has always been a constant need of extracting the information and knowledge effectively from the Web, and today there are techniques and terms from the various fields of data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc that help in achieving this goal. On the other side, due to the heterogeneity and lack of structure of Web data, getting knowledge from the Web is challenging and the users often find it hard to find the right content at the right time from the Web. In recent years, Web Content Mining has come up as one of the most promising research areas that aids in effective retrieval of information and data from the web. Researchers are continuously working for finding new ways for Web Content Extraction. Resource identification is an important aspect of Web Content Mining. It is the process of retrieving the intended web documents. It is done by web search engines and meta-search engines, or by crawlers. Choosing the appropriate search engine for resource identification is dependent on various factors, like it should be readily available, and is designed for fast and accurate retrieval of valuable information. Even if a popular and successful search engine is to be used, the different formats for inputting queries, the presentation formats of the retrieved results and quality of retrieved information are also analyzed. Once the intended Web documents are retrieved, the filtering of the relevant results in required formats becomes important for most of the web searches performed. It is highly desirable in the case of educational material searching.

Our research work based on the development of Syllabus Based Web Content Extractor (SBWCE) is a step forward to extract and mine useful knowledge from the Web and to make the searching, filtering and selection process of Syllabus Based Content from the Internet easy, efficient and effortless.

Syllabus Based Web Content Extractor (SBWCE) introduces a new technique of Syllabus Based Web Content Mining. It makes the Syllabus Based Web Content Extraction easy and creates an instant online book view based on the links relevant to the given Syllabus. Three important contributions are made by the current work. First, as multiple format educational information is needed for Syllabus based content; the technique used makes the finding of such content easier. Second, a new approach for capturing and recording the heuristics involved during searching by experts is used. Third, the grouping of Syllabus Words for precise extraction is exploited.

The evaluation of the filter tokens was done through experimental case studies. Precision and Recall calculations were done for comparative analysis between SBWCE's and Google's subject results. For this the subject categories - General Subjects, Subjects difficult to consolidate, Subjects having too much Web Based Material and Critical Subjects that are not essentially Web Based, were used. The results were found to be more relevant in the case of SBWCE.

SBWCE can work for Syllabi of different types and for different content formats required by the user. It is useful for the different types of users, including kids. It can also be used for on demand creation of Syllabi and the related Web Content Extraction and also for user selected Syllabus topics. It covers different subject areas but uses English Language as the base language. The Syllabus Based Web Content Extractor is an important tool for researchers. It can be used for education / learning and business purposes. Automatic Syllabus Content Extraction is to help teachers / students in getting recent updated information even if the syllabus is an old one. It is useful for all those involved in educational content finding process and can become an important part of future educational systems.