ABSTRACT

Web mining is an emerging research area due to the rapid growth of websites. Web mining is classified into Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM). Extraction of required information from web page content available on World Wide Web (WWW) is Web Content Mining. The WCM is further classified into two categories first category is to directly mine the content on documents and second category is to mine the content using search engine. The mining method focuses on the information extraction and integration.

The content of Web may be text, image, audio, video. Web pages typically contain a large amount of information that is not part of the main contents of the pages, like banner advertisements, navigation bars, copyright notices, etc. Such noises on Web pages usually lead to poor results in Web mining. This research focuses on the problem of Noise free Information retrieval on web pages, which means the pre-processing of Web pages automatically to detect and eliminate noises.

This research work proposes an approach for eliminating noises from web pages for the purpose of improving the accuracy and efficiency of web content mining. The main objective of removing noise from a Web Page is to improve the performance of the search. It is very essential to differentiate
important information from noisy content that may misguide users’ interest. This approach mainly concentrates on removing the following noises in stages: (1) Primary noises—Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and other Uninteresting Data such as audio, video, multiple links. (2) Duplicate Contents and (3) Noise Contents according to block importance. The removal of these noises is done by performing three operations. Firstly, using the Block Splitting operation, primary noises are removed and only the useful text contents are partitioned into blocks. Secondly, using simhash algorithm, the duplicate blocks are removed to obtain the distinct blocks. For each block, three parameters namely Keyword Redundancy ($K_R$), Linkword Percentage ($L_P$) and Titleword Relevancy ($T_R$) calculated. Using these three parameters block importance value ($b_I$) is calculated, which is called Simhash algorithm. The importance of the block is then calculated using simhash algorithm. Based on a threshold value the important blocks are selected using sketching algorithm and the keywords are extracted from those important blocks.

The performance of the proposed approach is evaluated with several web pages and the results ensure the effectiveness of the proposed approach in identifying the important blocks, which are relevant for knowledge extraction from web pages.