World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:8, No:12, 2014

Critical Review of Web Content Mining Extraction Mechanisms

Authors : Rabia Bashir, Sajjad Akbar

Abstract : There is an inevitable demand of web mining due to rapid increase of huge information on the Internet, but the striking variety of web structures has made required content retrieval a difficult task. To counter this issue, Web Content Mining (WCM) emerges as a potential candidate which extracts and integrates suitable resources of data to users. In past few years, research has been done on several extraction techniques for WCM i.e. agent-based, template-based, assumption-based, statistic-based, wrapper-based and machine learning. However, it is still unclear that either these approaches are efficiently tackling the significant challenges of WCM or not. To answer this question, this paper identifies these challenges such as language independency, structure flexibility, performance, automation, dynamicity, redundancy handling, intelligence, relevant content retrieval, and privacy. Further, mapping of these challenges is done with existing extraction mechanisms which helps to adopt the most suitable WCM approach, given some conditions and characteristics at hand.

Keywords : content mining challenges, web content mining, web content extraction approaches, web information retrieval
Conference Title : ICCSSE 2014 : International Conference on Computer Science and Software Engineering
Conference Location : Sydney, Australia
Conference Dates : December 15-16, 2014