Parallel web crawler architecture for clickstream analysis

Abstract :

The tremendous growth of the Web causes many challenges for single-process crawlers including the presence of some irrelevant answers among search results and the coverage and scaling issues. As a result, more robust algorithms needed to produce more precise and relevant search results in an appropriate timely manner. The existed Web crawlers mostly implement link dependent Web page importance metrics. One of the barriers of applying this metrics is that these metrics produce considerable communication overhead on the multi agent crawlers. Moreover, they suffer from the shortcoming of high dependency to their own index size that ends in their failure to rank Web pages with complete accuracy. Hence more enhanced metrics need to be addressed in this area. Proposing new Web page importance metric needs define a new architecture as a framework to implement the metric. The aim of this paper is to propose architecture for a focused parallel crawler. In this framework, the decision-making on Web page importance is based on a combined metric of clickstream analysis and context similarity analysis to the issued queries.