ABSTRACT:

The tremendous growth of the Web poses many challenges for all-purpose single-process crawlers including the presence of some irrelevant answers among search results and the coverage and scaling issues regarding the enormous dimension of the World Wide Web. Hence, more enhanced and convincing algorithms are on demand to yield more precise and relevant search results in an appropriate amount of time. Since employing link based Web page importance metrics within a multi-processes crawler bears a considerable communication overhead on the overall system and cannot produce the precise answer set, employing these metrics in search engines is not an absolute solution to identify the best search answer set by the overall search system. Thus considering the employment of a link independent Web page importance metric is required to govern the priority rule within the queue of fetched URLs. The aim of this paper is to propose a modest weighted architecture for a focused structured parallel Web crawler which employs a link independent clickstream based Web page importance metric. The experiments of this metric over the restricted boundary Web zone of our crowded UTM University Web site shows the efficiency of the proposed metric.