International Journal of Information and Management Sciences
Volume 24, Number 3, pp. 249-263, 2013

Hybrid Focused Crawling Based Upon VSM Similarity, WordNet Semantics and Hub Score Learning

Mukesh Kumar and Renu Vig
Panjab University

Abstract

New Websites, together with new Web pages, are mushrooming in every corner of the world and gigabytes of information is being uploaded, deleted or modified every unit of time. None of the existing search engines is able to cover the complete Web as a whole for indexing due to the ever increasing size and hence is not able to provide complete and latest information all the times. Users still have to sequentially browse the search results to get the desired information. Also sometimes the search results are biased by wiling full access of an unrelated page more times than a related page for some query. Focused crawler provides the solution for growing size of the Web by browsing the portion of the Web that is related to the specific domain. It covers the maximum Web space looking for the contents related to the domain and provides the more recent and exact information. In this paper we present a focused crawler architecture based upon WordNet semantics, Vector Space Model (VSM) and hub score learning. Crawling results for breadth first crawler, VSM based best first crawler, Naive Bayes breadth first crawler, Naive Bayes best first crawler, and crawler based upon WordNet semantics, Vector Space Model (VSM) and hub score learning, are shown. The results show that the proposed crawler outperforms the others in terms of the precision and also outperform all but Naive Bayes breadth first crawler, which produces the worst precision among all the competitors, in terms of average time taken for collecting 1000 domain related pages.

Keywords: Information retrieval, World Wide Web (WWW), data mining, search engines.