The World Wide Web (WWW) continues to grow rapidly at more than a million pages per day. It makes difficult to discover topic relevant information that can be used in specialized portals and on-line search. New Web pages are mushrooming in every corner of the world, Gigabytes of information is being uploaded, deleted or modified every unit of time. 103 million new Websites were added to the Web in the year 2013. This makes the total to 759 million Websites on the Web till December, 2013. The exponential growth of information on Web demands some efficient and effective methods to organize and retrieve the information available. Two of the most popular search engines Google and Bing reports 48 billion Web pages and 14 billion Web pages respectively for indexing from a total of 14.3 trillion live Web pages on the internet. The total number of pages available is much more than the pages reported by Google or Bing for indexing. The imbalance between exponential growth and, limited network and storage resources poses basic limits of scale for today’s generic crawlers and search engines. Search engines find it hard to refresh their index often enough to keep up with the pace of change of the Web. The user has to go through selection and browsing process for finding pages most relevant to him.

The generic search engines will become less applicable to a user with specific information needs that do not fall into most popular categories. Focused crawling therefore has gained significant attention as research field of Web mining. Focused crawler traverses a subset of the
Abstract

Web for gathering documents only on a specific topic, instead of searching the whole Web exhaustively, eliminating the problems caused by size of the Web as a whole. Rather than collecting all accessible documents, focused crawler analyses its crawl boundary to find the links likely to be most relevant for the crawl, eliminating the topic specific information retrieval problem. Most of the crawlers available make use of link structure or keyword matching scores for finding relevancy. This relevancy can be improved by making use of some semantic measures along with others, to guide the crawl. In this work a focused crawler based upon WordNet semantics, Vector Space Model (VSM) and hub score learning is proposed. Crawling results for Breadth First Crawler, VSM based Best First Crawler, Naive Bayse Breadth First Crawler, Naive Bayes Best First Crawler, and the proposed Semantics and Learning based Crawler are studied for precision rate, recall rate, empirical precision rate, empirical recall rate, empirical F-measure and average page retrieval time.