

Abstract

Currently, web pages are growing fast and evolving rapidly and become one of the means of dissemination of information by personal, social and commercial. The more people who need information on certain topics such as on sports, but find it difficult to obtain relevant information. That requires a Web Crawler specifically to help Internet users find relevant pages. Own Web crawler is a program that does the scanning process to all internet pages to be made indexnya and support a search engine.

Unlike the crawler that used by commercial search engines which generally aim to collect many web page, focused crawler (called as topical crawlers oft) browse and retrieve web pages relevant to a particular topic selectively. In this bachelor thesis, Naive Bayes classifier is used to distinguish the web page instead of sports and non sports web page, and using Best First Search as the crawling algorithm of the queue. Selection of the best value was done by comparing the calculation's results of Cosine Similarity.

Shown that the best-first search algorithm and the Naive Bayes classifier will help browse the relevant pages first.

Keywords: focused crawler, sports web, naive bayes, best first search