Abstract

Focused crawler is a crawler to download specific web pages that match the topic specified by the user. The main task of focused crawler is to collect as more as possible relevant web pages according to the given topic. Not all web pages download in a web site, but only the web pages related to topics that will be stored, thus saving resource usage of the server.

This final project, will implement a focused crawler using the cosine similarity, link score, and traverse irrelevant page method. Cosine similarity method used to determine whether a web page is relevant to the topic or not. Link score method is used to guide crawlers which direction will approximately get a web page relevant to the topic. Traverse irrelevant page method is a technique to traversing web pages that are not relevant, to obtain relevant web pages in it.

Testing results show that the focused crawler will get the optimal value of the precision rate by using the traverse irrelevant pages method with depth level 0. Focused crawlers can also be implemented using seed urls his close association with the topic, as well as seed urls his little relevance to the topic. Performance of the focused crawler seen from the precision rate aspect and computational time aspect will also be optimal if using seed urls that his little relevance to the topic.

Keywords: *focused crawler, cosine similarity, link score, traverse irrelevant page, weight table.*