ABSTRACT

This project primarily exists as my personal desire to create a web crawler application for browsing the web and store the information without much interaction from human. And Java is choses as development base because it runs very well on many platform (including my Linux rig) and it has a good collection of network or Internet ready API which fulfill my needs. The great threading ability of Java also becoming a major advantage for this project because web crawler ability to process multiple link simultaneously must be implemented using threads.

Web crawler means any application which capable of automatically process any hyper links found on a HTML page and process it to gather more links which will be processed again, inside the application this behavior will take place until a stop condition is met (usually a certain link depth or number of files retrieved). Although there are many purpose of a web crawler, the main purpose of using it is for information gathering or archiving. The contents within this paper will describe the process of developing a web crawler used for archiving files gathered from HTML paged to local storage for off line browsing.

And for the final words, this project has meets it's goal though not 100% perfect because there's still some minor glitch in few of the code. Although rare, this problem will cause inability to retrieve some type of relative links. In the end I just going to say thanks to all who has support this project, I'm looking forward to fix the glitches and make the application better.

Keywords: Web Crawler, Java Network Programming, Spider.

DAFTAR ISI

DAFTAR GAMBAR