Report programming exercise Ericsson In-Game Communication

Hossein Molazemhosseini - 850916-7956
Master Student KTH – Software engineering of the distributed systems
Mobile: 0725298657 Email: homh@kth.se

Introduction

Sequential:

It's a kind of web crawler program. I have used Depth First Search algorithm to process the websites. The depth will be defined as an input parameter. Recursive method is used for building a tree structure and processing the websites and all the WebPages which are linked from this website. For finding all the web pages which are linked from a website and counting the number of the capital letters, I have used javax.swing.text.html.HTMLEditorKit. The program accuracy in finding web pages' links is limited to this library. I have translated relative URLs into absolute URLs.

Concurrent:

A crawler incurs several delays like Resolving the host name in the URL to an IP address using DNS, Connecting a socket to the server and sending the request and Receiving the requested page in response.
Solutions to Overlap the above delays by fetching many pages concurrently, Can use multi-processing or multi-threading, Each process or thread works like a sequential crawler, except they share data structures, Shared data structures must be synchronized (locked for concurrent writes) and finally Speedup of factor of 5-10 are easy this way.
I have used hash table not hash map for synchronization and multi threading. Put() and get() methods of the hash table are synchronized and so **thread safe.** There would be no problem in concurrent accesses of the threads to the shared variables.

Limitations

Internet speed has a direct effect on the program performance. Extracting the web pages' links may be sometimes not accurate and also growing number of other formats, Flash, SVG, RSS, AJAX.

Environment/programming language/tools

Environment: Windows 7 – professional edition
Programming Language: Java 1.6
Tools: Netbeans IDE 6.9.1

How to build and run the program

This program is written in java, so you need to have the java 1.6 (JRE) on your system or if you have the net beans IDE installed, you can open the project and run it.
There are folder **[WebCrawlerParallel]** for the parallel program and [**WebCrawlerSerial**] for the sequential program. They are both Netbeans project. You can find the source java files under [**src**] folders.
If you want to compile the program without Netbeans go to the **[src]** folder and:

- Javac WebParserSerial.java
- Javac WebParserParallel.java SearchHandlerThread.java

And you can run the class files by java command and giving the argument.
You can also run the program by going to the [**Executables**] folder and:

- Java –jar WebCrawlerParallel.jar www.bbc.com 2 resultParallel.txt 30
- Java –jar WebCrawlerSerial.jar www.bbc.com 2 resultSerial.txt

In the executables folder you can also see the sample outputs of the application.

Note:

1. The Head Url should be the first page of the website, not the pages which are linked from the root.
2. The pages which are not found or any problem in connecting, are printed out to the output screen. that's because of using printstacktrace() method. But the final result will be printed out to the file. Some problems are, server not responding, file not found, and other errors. Redundant web pages will be skipped and not processed.

Procedure:

I had two days, so I just spend half of my first day on brain storming. To go deep to the problem, understand it by heart and see what would be the requirements and possible problems during the production of the software and what could be my possible solution.

I also had a course network programming with java, which two lectures where about multi threading and connecting to web resources. And regarding my previous experience in programming and searching the internet, I found the solution.

As I have been recently working for two years in java, I ,decided to choose Java.

There was the other better way to use thread pool. Dynamic spreading the load among threads.