Multi-Agent Sistems

1st Assignement – Web crawler

Master in Electrical Engineering
Electrical Engineering Dept.
Institute of Technology
University of the Algarve

Pedro Cardoso

Updated: March 9, 2017

Introduction

A Web crawler, sometimes called a spider, is an Internet bot which systematically browses the World Wide Web, typically for the purpose of Web indexing (web spidering).

Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently.

Crawlers consume resources on the systems they visit and often visit sites without tacit approval. Issues of schedule, load, and "politeness" come into play when large collections of pages are accessed. Mechanisms exist for public sites not wishing to be crawled to make this known to the crawling agent. For instance, including a robots.txt file can request bots to index only parts of a website, or nothing at all.

Wikipedia
https://en.wikipedia.org/wiki/Web_crawler
March 9, 2017

Assignement

For this programming assignment you are asked to programm a Web crawler. The important part is to somehow "parallelize" the process in order to (try to) beat a sequential implementation. As a suggestion use threads and a queue to process your results.

As a minimum work, you should list the pages under a certain domain. As a suggestion use the w3.ualg.pt as a root. PLEASE AVOID disrupting the domain's service (read about DoS attacks), by placing some kind of timer.

As an add to your work, the contents of the retrieved pages will be allocated in memory or stored in a database, allowing for content searches.

Your program should have a good, clean logical structure. We will also be looking at good documentation, descriptive variable names, and adherence to the coding convention.

To deliver...

A small report with:

- Brief description of the implemented methods.
- Explanation of the results.
- Code (attached or bitbucket repository)

Delivery date

Although you can deliver your work up to the exams dates, we advise you to deliver it earlier. A good date would be the 30th of April.

References

- Web Crawler, Wikipediam, https://en.wikipedia.org/wiki/Web_crawler Conjecture

Final remarks

If you have any doubts, please contact the professor directly or by email – pcardoso@ualg.pt