Web Crawling Algorithm

San Jose State University, Spring 2016, CmpE 130 Project

by: Luis Liang (008330166), Mauricio Rivera (007700147), Anahit Sarao (008435583), Nathan Lin (009942621)

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Abstract**

In this project, we will focus on the complexity and variety of web crawling algorithms. There are different applications that may be used to crawl the web, for example: search engines, web analytics (SaaS), and even for cyber security.

The project will include a detailed report about different ways we can implement a web crawling algorithm, its complexity and future applications that may be help internet users. There will also be a basic web crawling demo that will illustrate how ‘spidering“ works around the web.

**Week by Week Proposed Schedule**

|  | **Week of** | **Task** | **Notes** |
| --- | --- | --- | --- |
| 1 | 3/7 | Research | Weekly Meeting |
| 2 | 3/14 | Research | Weekly Meeting |
| 3 | 3/21 | Research | Weekly Meeting |
| 4 | 3/28 | Program Analysis | Report on progress with programming |
| 5 | 4/4 | Continuation on programming | Prepare demo |
| 6 | 4/11 | Start of Final Report | Final details in demo |
| 7 | 4/18 | Continue with Final Report | Check Final Report |
| 8 | 4/25 | Finalizing Project Report | Turn in Final Project on 4/28 |
| 9 | 5/2 | Prepare and Finalize Presentation & Demo | Present in class |

**Sources:**

1. <http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler>
2. <http://chato.cl/papers/castillo04_scheduling_algorithms_web_crawling.pdf>
3. <http://airccse.org/journal/avc/papers/1314avc01.pdf>
4. <https://www.google.com/insidesearch/howsearchworks/crawling-indexing.html>
5. <http://homepage.cs.uiowa.edu/~psriniva/Papers/TOIT04.pdf>
6. <http://www.devbistro.com/articles/Misc/Implementing-Effective-Web-Crawler>
7. <http://www.ijritcc.org/download/Analysis%20of%20Web%20Crawling%20Algorithms.pdf>