CS 453: Project 1 - Crawling the Web

Project Report

Creepy - Web Crawler

Travis Hall
trvs.hll@gmail.com

Brittany Miller
miller317@gmail.com

Bhadresh Patel
bhadresh@wsu.edu

Washington State University Vancouver
September 12, 2010

Abstract

The main goal of this project is to design and implement a *web crawler*. The name of our web crawler is *Creepy*. *Creepy* is simple web crawler written in Python that takes a set of seeds (URLs) and begins crawling the web.

1 Overview

2 Automation

3 Document Storage

4 Politeness

When implementing the politeness standards, one of the things we really wanted to do was create a nice, reusable and generic library for the politeness standards. To this end, the module is designed such that you simply create a *Robot* object for a domain and then query whether or not you are allowed to access a URI. In order to ensure that *robots.txt* does not go stale, the programmer can pass along an 'expires_in' value. When the *Robot* is queried, it will then check whether the file is expired and automatically re-fetch and parse it when that is the case.

However, we also needed to design towards the state of the project as a whole, and one of the concerns was how to handle delays. In order to avoid duplicating a domain-based hash for each delay, reusing our *RobotStorage* class (and thus our *Robots*) seemed a natural choice. Unfortunately the result feels a little awkward, in that you have to update the *Robot* and inform it when the last request was made. By preference, this is something that would occur naturally while fetching the page.

Sadly, the *Robot* does not obey the extended standards for *robots.txt*. Though it certainly will parse out that information, it will be stored in the 'other' field and is unused unless specifically programmed for. This means that information like 'Visit-time', 'Request-rate', and 'Comment' are largely left ignored. This is not to say that it cannot be extended to obey them, but we had not discovered the extended standard until after the parser was already written and in use.

5 Duplicate Detection

6 Roles

Travis Hall Politeness standards, project management (Git/Github)

Brittany Miller ...

Bhadresh Patel ...