Date: 3/20/2008
To: Alexander Dekhytar, Associate Professor
From: Vinay Anantharaman
Subject: Senior Project Status

The implementation of LPageRank is nearly complete. I will outline the status of the following components: Web Crawler, Log Miner, Rank Algorithms, Informational Retrieval, and front-end.

Web Crawler

The web crawler implementation is complete and unit tested. In Winter quarter I have improved how a valid link is determined, pages are stored, and rate control of the pages. There exists a few improvements yet to be implemented. The crawler will be run over the spring-break using pagerank.csc.calpoly.edu.

Link Determination

A valid link must be within a specified set of domains. The domains are specified in a XML configuration file. Furthermore, the link must not begin with, or end with a set of strings. The strings are specified using a XML file. Currently links may not begin with mailto, and end with css, png, gif, and a few other image types.

Page Storage

Upon retrieving a page they are stored out to disk or memory. If the pages are stored to memory they will be processed and stored into the index. Otherwise, pages stored out to disk with an accompanying XML file describing the url, and mime-types for each file.

Rate Control

The crawler pauses for a specified number of seconds before retrieving the next page. I have set the pause at five seconds.

Improvements

Ryan Matterson recommends that the crawler read meta tags which, indicate the page should not be crawled. Also, I plan to read the meta tags for keywords. I will implement these features by April.

Log Miner

The log miner is fully complete and unit tested. I attempted to run the miner on a 3GB log file and ran for more than 36 hours without completion. Performance improvements are required to reduce the time it requires to parse through large log files.

Rank Algorithms

LPageRank, and PageRank implementation is complete and unit tested. Once the log file is parsed, and crawler ran I will run these algorithms over the data.

Informational Retrieval

The informational retrieval system is based on the Lucene.Net system. Before the pages are placed into the index the tags for HTML pages are removed. It is possible to add other pre-processors for documents like pdfs.

Front-End

The command line program which runs the offline components runs the Web Crawler, Log Miner, Information Retrieval, and Page ranking system. The online component front-end for running queries is not built yet and will be by April.