Date: 12/7/2007
To: Alexander Dekhtyar, Associate Professor
From: Vinay Anantharaman
Subject: Senior Project Status

The status of implementing the LPageRank search engine is on schedule. Three components of the search engine were to be built in Fall Quarter. I will outline the status of the Web Crawler, Web Traffic Log Parser, and LPageRank Algorithm.

Web Crawler
I have implemented the web crawler in C#. The web crawler retrieves the starting URL from the component's XML configuration file. The links on the page are extracted and placed in a queue if and only if they do not point to images, the link does not point to a previously visited page, and are part of the same domain as the starting URL. A URL graph is built while traversing the domain. Edges are added between the page and the links found on the page. After adding the links and edges for the current page another page is retrieved from the link queue. The process continues till no more links are available in the queue.

The web crawler requires unit testing for individual components to ensure expected results. I will unit test the sub components and the full crawler over the Winter break.

Web Traffic Log Parser
I nearly have finished implementing the Web Traffic Log Parser in C#. A few small components related to iterating over a data structure that is partially stored on disk. The parser reads the input log file, which contains records of user activity for Cal Poly's web servers. Each record is read in and the requested URL is placed into a data structure representing that user's session. Records must satisfy the following conditions the status of the request must be 'OK' and must not represent a request for an image. The records of user's are placed into data structure that stores the records on disk while maintaining a small set in memory. Once the file is exhausted of records each user session is stored by access time. For each link accessed in a given time, 30 minutes currently, an extra weight of one is added to the edge between the previous link and the current one. The weights are added to the URL graph that was constructed in the Web Crawler.

The Web Traffic Log Parser requires unit testing for individual components and the full component to ensure accurate results. I will complete the unit testing over the Winter break.

LPageRank Algorithm
I have not started implementation of the algorithm. This component will be completed and unit tested by the end of the Winter Break.

Once I finish implementation and unit testing of the three components I will gather experiment results to show within the first two weeks of the Winter Quarter.