PROJECT

Group Members:

Mainak Datta	**dattam**
Parul Bhalla	**pbhalla**
Aakriti Singla	**aakritis**
Krishna Chaitanya Daliparthy	**krishnad**

Goals

The objective is to build a web indexer/crawler which is an academic clone of Google Search Engine. While Google presented it as a prototype in the year 1998, there have been many changes to the implementation of The Anatomy of a Large-Scale Hypertextual Web Search Engine over time. We will now be designing the search engine to incorporate the latest methodologies favoring performance and scalability.

High-level approach

Our design approach will involve the following components working together:

1. Mercator style crawler, parse HTML docs and crawl a decent portion of the web
2. Index the result of (1) by content
3. Run PageRank algorithm MapReduce job to rank each page
4. An interface for the users to search for key terms view results ranked based on PageRank and TF/IDF metrics

Features Implemented

Mercator - A Distributed Multi-threaded crawler using **Berkley DB** and **S3**
Indexer - Calculating Inverted Index using EMR
Page Rank - Calculating Page Rank using EMR
Storing Indexer and Page Rank results in **DynamoDB**
Search Engine - Basic Multi threaded Search Engine implementing Ranking Algorithm to display search results

Extra Features Implemented

Planning to implement the following features :

a. Crawling PDF/Image Files in Mercator
b. Handling META tags while creating Inverted Index