Project Proposal: A Nutch-based Search Engine with Web Interface

Team members: Guanghan Ning, Nguyen Trung

1. Introduction

Internet is our everyday source for all kinds of information, like news, broadcast, music, pictures, or even movies. We are immensed in the flood of big data everyday. However, the information we need is only a small portion of it and for different individules the importance of information is hardly the same.A search engine will make it convenient for us to absorb information efficiently. The search engine will do the mapreduce for us humans to get the useful pages.

2. Objectives

We are going to use Nutch - an open-source effort to build a web search engine, as our web crawler to download the data from the internet, and we are usign HBase to store the database,Pig to perform mapreduce on the database, and we are going to build a simple web interface as front-end layer for users to search text.

3. Resources

3.1 Server

IBM cloud server(clusters we created) in Canada

OR

lewis, a cluster of multi-core compute servers operated by Uiversity of Missouri Bioinformatics Consortium.

OS：Platform OCS Linux 4.1.1 (based on Redhat Linux).

3.2 Programming Tools

HBase, Pig

3.3 Software/platform

Nutch, BigInSights

4.Roles of each member

Basically, Trung will be working mostly on Nutch to build webcrawler for the engine, and Ning will be mostly working on HBase and Pig to store and process the data. In terms of other work, like web interface, or some unpredictable small problems to fix, we will be working together.

5. Risks

Deploying Nutch on IBM cloud clusters, or other technical problems may prolong the predicted time of completion.

6. Schedule

Tasks and deadline:

6.1 Planning/Design	10/15
6.2 Deploying Nutch to download data	10/22
6.3 Use HBase to build our Database	10/29
6.4 Implementing and Testing mapReduce on the Database	11/12
6.5 Deal with the web interface	11/19
6.6 Documentation	11/26
6.7 Presentation	12/3