Web Search Engine

PHASE I REPORT
COURSE: 03-60-654-01

University of Windsor

Guided By: Dr. Luis Rueda
UNIVERSITY OF WINDSOR

Declaration

"We confirm that we will keep the content of this project description confidential. We confirm that we will work as a group, in equal parts, and that we have not received any unauthorized assistance in preparing for or writing this project. We acknowledge that a mark of 0 may be assigned for copied work."

- Siddharth Khobare (104303348)
- Shehraj Gupta (104484157)
- Mustafa Ali Misri (104485709)
- Dhawal Rank (104340181)

Preface

The project involves developing a Web search engine that uses three to five different concepts learned in class as well as some useful topics out of the course. Today almost every answer can be found from a simple Google Search. As a part of the term project, we are developing a simple yet efficient search engine. The aim is to cover main features that any search engine should have which are content, speed and quality.

Table of Contents.

Description

Search engines have become an inevitable part of our life. There are 3.5 billion searches in google everyday. Almost every information can be found on the World Wide Web and search engines makes it easier to access particular information.

For our search engine we are going to concentrate on the following:

- Content (how wide the search results are)
- Speed (how fast search occurs)
- Quality (Are results of good quality)

All the above qualities of the search engine are addressed by features like Crawling, Indexing and Page ranking which are explained in detail in the following section. As an initial step, few domains will be crawled, indexed and ranked. Then they are sorted and stored in a HashMap which in turn is written to either a .txt or a .dat file and stored on a hard disk. When user performs a search it just has to look into the file by pattern matching and display the results. This helps in saving time as the web has not to be crawled every time. The whole search engine is to be developed in Java.

Features

Crawling

In layman terms crawling means exploring the web. Crawlers are programs that crawls the web. Crawling starts with list of URLs to be crawled. During this phase, crawler adds the URLs to a list of URLs also known as Crawl Frontier. There are 60 trillion individual web pages in world. Crawling each one of them in our machine is impossible due to the time complexity. So we are going to select 1-2 specific domains to be crawled. For our project we are going to use Apache Nutch which is an open source crawler project. Crawlers are editable for example we can manually set the frequency of crawling. This is useful in some high priority websites where data changes every second.

Indexers

Indexing is similar to the index page of any book. There are keywords followed by the location those keywords can be found on the book. The indexer will collect the URLs crawled. We are planning to use Apache Lucene library for indexing the URLs and their contents. Main reason of using Apache Lucene is we can sort data by any field which can give us ranked search. We will be storing the indexed results on a file so that there is no need of crawling and indexing every time. This can actually make the search a lot faster.

Page Ranking

Page Ranking addresses the quality factor of a search engine. The results displayed should be relevant. This is where page ranking holds an importance. Ranking through KMP or other occurrence related rankings are inappropriate as its not necessary that the page where a particular keyword was found is more important. To address this, we are going to rank pages using random surfer model. In this model, the priority of a web page is determined on basis of number of incoming links to that web page. Thus a web page that has been referred to in many other webpages, then that web page will have a higher rank.

Search Interface

A search engine is incomplete without an interface. We are going to develop a simple yet attractive search page. For now, we are going to use applet for creating a web page and twitter bootstrap for designing if we have some time left.

Task Distribution

We have tried to distribute the tasks among every one in a way that each gets a fair share of the work.

Siddharth Khobare: Crawler & Indexing

Shehraj Gupta: Indexing & Page Rank

Mustafa Ali Misri: Search Interface & Page Rank

Dhawal Rank: Search Interface & Crawler

All the documentations are the collective efforts of the team.