Search Engine Implementation

BSCS-6AB Semester Project for Data Structures and Algorithms

Introduction

You have already summarized the paper “The Anatomy of a Large-Scale Hypertextual Web Search Engine” in Assignment 3. Picking ideas from the paper, you are required to make an elementary search engine as your semester project.

Most search engines go in four stages:

1. Crawling
2. Forward Indexing
3. Reverse Indexing
4. Searching

Crawling or web crawling is done through a tool which browses as many websites as it can and stores them in a searchable format (forward indexing). To do this, you usually start with a few websites and keep following the hyperlinks in these websites until you you have crawled all the linked pages.

Forward indexing in simple words means storing which words each page contains. So you have found a list of words against each page. But the user will search using words. For that, you require a list of documents against each word. This is called reverse index or inverted index.

Specifics

For the scope of this project, you are not required to do crawling. So you will implement forward indexing, reverse indexing and search on pages corpus.

The corpus on which you are required to demonstrate your search engine is Wikipedia Simple ~ 132 MB (link)

As you have guessed, you can download the complete text of Wikipedia from the Internet. However, the Wikipedia is very large. So you can initially test your approach on Wikipedia Simple. Wikipedia Simple is a smaller version of Wikipedia designed for children which contains articles in simple english.

The dataset is in the form of an XML file. You can parse it to get the respective data (i.e. page IDs, content etc.) which you have to index.

Grading criteria

1. Forward indexing
2. Reverse indexing
3. Single word searching
4. Multi word searching
5. Ranking results based on frequency and position
6. Making a scalable system (using proper algorithms)
7. Using GIT for team collaboration
8. System Interface

Single word search means retrieving list of all the pages which contain the word the user has searched for.

Multi word search means retrieving list of all pages which contains all the words the user has searched for. These words don't have to be present next to each other. So if a user searches for "Data structures and Algorithms", you have to retrieve all the pages which contain the four words "Data", "structures", "and" and "algorithms".

Wikipedia pages dataset has text stored in form of different level of headings (title, heading 1, heading 2, heading 3, normal text etc). So you should give higher priority to the pages which are present in the title then those in normal text and so on.

Making a scalable system means that your system should work as the size of the data increases. For example, if the user searches for word "the" which appears in maybe 1 million pages, your system should not crash or wait infinitely.

As you have already been introduced to the idea of using git (bitbucket) for collaboration between team members, you are going to use that in your project. You should make a private repository on either github or bitbucket. For github, you will require github education to get access to create private repositories for free. You should not complete the project and then upload it once on github. Instead, you should keep pushing the updates as commits as you progress through the project. All group members are required to make the commits which will show team collaboration. You will be required to upload the commit history screenshot as your final project submission.

You should also make a usable interface for your system. It should ideally contain a text box where the user can enter his search query and when he presses enter, he should get links to all the results.

Languages and Tools

You may use any language of your own choice. However, you cannot use tools which do the indexing for you, for example MySQL. So you will have to implement the indexing yourself in language of your own choice (C, C++, Java, PHP, Python etc).

Timeline

Project presentation and viva will be done in the last week of the semester so you have exactly three weeks to complete your project.

In case of any queries, please book appointment with TAs who will guide and help you through the project.