Building a Search System Using Python and MySQL

Your Name

May 22, 2024

Abstract

This report describes the development of a search system using Python and MySQL.

1 Introduction and goals

The goal of this project is to develop a search system that can efficiently retrieve and store information from a collection of HTML files. The system is built using Python for file handling and MySQL for database management. The project also focuses on the creation of indexes to have a more efficient search. The final goal is a system that is able to execute a query on my database that returns the files where a given word is found in the name and in the content, counting the currencies.

2 Python Program

The Python program is responsible for traversing the directory structure, retrieving HTML files, and extracting relevant information. The program detects directories, retrieves file content, and gathers metadata such as file name, size, and the containing folder. While the program detects each items, it builds a batch and when the batch reaches a given size, the information is uploaded in to the database, the transaction of is committed and the data structure holding the information is flushed.
This allows the uploading of large file-systems.

3 Database design

The schema of the database is designed with two tables. The first table *files_info* holds the information of the files, like name, path, extension and size.
The other table *files_content* contains just an id and and the content of the HTML files. This is done because of order and because if we have a query

that often excludes from its select the content attribute, we might have some inefficiencies.
This is done in the search query that I have implemented.
I have created also two indexes. One on the file name attribute, of type b-tree and the other one is a full-text index. Indexes are used to speed up the search queries and their usage can be specified via index hints like *using index(name)* or *force index(name)*.
The python code is also responsible for creating tables indexes and foreign keys.

4 Search query

The search query is executed by a python script. It is the union of the result of two queries, one on the file content that use the the difference between the length of the file content with and without the target word to return if the word is found and how many times, and another one that uses the *like* operator to find the target in the name. In this query the index usage hint is added.

5 Conclusion

Once the software components are completed I was able to upload my whole file-system in the database and perform searches on it using my searching script. This is allowed by the batch structure of the population script.