Master's Thesis

Design and Implementation of a Configurable Generic Search Engine Indexing using Scalable Crawlers

Alhajras Algdairy

Examiner: Prof. Dr. Hannah Bast

Advisers: M. Sc. Natalie Prange

Albert-Ludwigs-University Freiburg

Faculty of Engineering

Department of Computer Science

Chair of Algorithms and Data Structures

April 25th, 2021

Writing Period

01. 12. 2020 – 25. 04. 2021

Examiner

Prof. Dr. Hannah Bast

Second Examiner

Prof. Dr. Thomas Brox

Advisers

M. Sc. Natalie Prange

Declaration

I hereby declare, that I am the sole author and composer of my thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work.
I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Place, Date

Signature

Abstract

Web search indexing is an essential system that powers modern search engines. It automates the process of collection and organization of data from web pages to create an updated index of the web that can be optimally searched. Web search indexing consists of two essential components, a web crawler, in which search engine bots systematically traverse the web to find new or updated content based on rules declared beforehand, followed by the second component which is the indexing of the collected data. The process of web search indexing comes with its own challenges, including performance, managing dynamic content, and answering the question of what is the most relevant content. As the web continues to evolve and grow, the task of web search indexing will remain a key focus of search engine technology and research.The aim of this thesis is to design and implement a generic configurable web search indexing that can be used as a basic tool on different websites and can be further expanded and improved, and scaled. The approach included a simple UI design that allows users to configure and create crawlers and index the generated data.

Acknowledgments

First and foremost, I would like to thank:

- My parents for supporting me during the master's program.
- My wife for her love and support.
- Prof. Dr. Hannah Bast for accepting my topic and for her guidance and supervision.
- M. Sc. Natalie Prange for her thoughtful ideas and suggestions.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

The World Wide Web (WWW) contains an enormous amount of data; this data is increasing each day rapidly. The amount of total data created and replicated is expected to grow to more than 180 zettabytes by 2025 according to Statista. The growth is expected to continue as more smartphones are more and more affordable, and more people can reach the internet. Moreover, due to the COVID-19 pandemic, more companies started offering work remotely, more shops created online stores, and more services switched to cloud-based. This change in society during the last few years has made the internet a vital part of our day-to-day life.

Although the data is available, making a helpful meaning is a challenge. Search engines, for example, try to organize and index that information to make them easily searchable by the end user. Furthermore, collecting data can help spot competitors and have a deeper meaning in the market. Additionally, data scientists are now playing essential roles in most organizations and enterprises to understand consumer needs by collecting and analyzing data from the web.

Although some websites provide APIs to provide organized information about their services, for example, some airline companies provide API that serves information about their flight schedules, other online shops also provide a documented API to get helpful information about their available products. This is not a guaranteed approach to gathering data, as not all websites offer an excellent documented API. For example, social media websites are reluctant to give information about their users, which is understandable. What if you would like to go through all comments and classify them as spam or not? Depending only on the assumption of having an API for each website is a fragile approach.

Information retrieval (IR) is a term introduced in 1951 by Calvin Mooers. It is accessing and retrieving data from a vast pool of unstructured information. One of the most practical applications of IR is to collect information from the internet; therefore, implementing a generic algorithm to gather the needed information and index them is a valid approach. Crawlers or Spiders are bots programmed to follow specific roles defined by the user to automate fetching and extracting data from the internet.

One form of IR is a web search engine. A web search engine is a system engineered to index the internet. Users can search for articles, documents and pages by entering keywords. The search will provide a list of the most related result that matches the search query. Using the crawlers explained earlier; the engine can index the collected information and optimize the search process using different algorithms and techniques.

Almost everyone nowadays uses Google, Bing or DuckDuckGo search engines for personal usage for research or enterprise to do market research. Search engines are so important that they make Search Engine Optimization SEO position merge and vital to any business. Harvesting, manipulating, and analysing data are essential, making information almost the new currency.

1.2 Problem Statement

Although the available search engines are too efficient in crawling and indexing the Internet, businesses like E-commerce are primarily interested in knowing the lowest price for a specific product to understand their marketplace among their competitors. Achieving this by using Google, for example, will not solve the issue as the search directly for the product will rank the products based on some criteria predefined by the vendor Google. Those criteria can include best brands, geo-location close to the user, how well the developers optimize the SEO in the website and more. Note that the lowest price criteria are not included in page ranking. The second issue is the result format; each search engine provides a different list of results based on their implementation. This is only suitable if one is only interested in comparing prices and does not care about the various templates used on each website.

Search engines need to be tweaked and configured to match the domain of interest as E-commerce in the previous example and to match a specific use case like the price comparison mentioned.

The main problem is that businesses are often interested in only a portion of the internet that interset with their domain and expertise. Furthermore, the criteria for indexing and page ranking depend heavily on their use case and is vital to their business to take control of it and configure it as fit.

Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025 Published by Petroc Taylor , Sep 8, 2022 https://www.statista.com/statistics/871513/worldwide-data-created/: :text=The

1.3 Contributions

1.4 Chapter Overview

2 Background

Explain the math and notation.

3 Related Work

Give a brief overview of the work relevant for your thesis.

4 Approach

The approach usually starts with the problem definition and continues with what you have done. Try to give an intuition first and describe everything with words and then be more formal like ‘Let g be ...’.

4.1 Problem Definition

Start with a very short motivation why this is important. Then, as stated above, describe the problem with words before getting formal.

4.2 First Part of the Approach

4.3 N-th Part of the Approach

5 Datasets

6 Experimental Evaluation

7 Summary of Results

8 Conclusions and Future Work

Bibliography

[KimathiKimathi2020] Kimathi2020Kimathi, G. 2020June. What Is Ray Tracing Technology and How It Works in GPUs. What is ray tracing technology and how it works in gpus. `https://www.dignited.com/62084/how-ray-tracing-works`