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

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