**Computer Science & Engineering Department**

School of Engineering & Technology, HNBGU Srinagar Garhwal

(A Central University)

#### CANDIDATE’S DECLARATION

I hereby declare that the work presented in this report titled **“WEB CRAWLER in SEO (Search Engine Optimization)”** in partial fulfillment of the requirements for the award of the degree of bachelor in Technology in Computer Science and Engineering submitted to **“School of Engineering & Technology, HNBGU Srinagar Garhwal (A Central University)**, is an authentic record of my own work carried out under the supervision of **Mr. AMIT PANWAR.**The matter presented in this report has not been submitted by me for the award of any other degree or in any other Institute/University.

Himanshu Beniwal

**(Roll No. 15134501017)**

The Viva-Voce examination of Ms./Mr. **Himanshu Beniwal** has been held on

**Signature of Coordinator External Examiner**

#### ACKNOWLEDGEMENT

I express my deepest sense of gratitude towards my supervisors Faculty, Department of Computer science & Engineering, HNBGU, for his patience, inspirational guidance, constant encouragement, moral support and keen interest in my work.

I express my sincere gratitude to Dr. M.M.S. Rauthan, Head of Computer Science and Engineering Department for his kind attitude in understanding and fulfilling every need during the course of my work.

I express my sincere gratitude to Mr. Amit Panwar, Coordinator project for their constant support, enthusiasm, and affectionate attitude during work.

I owe a debt of gratitude to my parents, for their consistent support, patience, and encouragement during my education.

Last but not least, I am thankful to the Almighty who gave me the strength and health for completing the work.

Himanshu Beniwal

#### ABSTRACT

A **Search Engine Crawler** is a program or automated script that browses the World Wide Web in a methodical manner in order to provide up to date data to the particular search engine. While search engine crawlers go by many different names, such as web spiders and automatic indexers, the job of the search engine crawler is still the same.

The process of web crawling involves a set of website URLs that need to be visited, called seeds, and then the search engine crawler visits each web page and identifies all the hyperlinks on the page, adding them to the list of places to crawl.

URLs from this list are re-visited occasionally according to the policies in place for the search engine. The policies of the search engine can be different for each search engine, and may be a cautionary action to ensure that some of the pages that have been added to the index before have not become spam.

Contents

[Chapter 1: Introduction 1](#_Toc482882819)

[1.1 Concept of Search Engines: 1](#_Toc482882820)

[1.2 What is Search Engine Crawler? 1](#_Toc482882821)

[1.3 Don’t Let the Spiders and Web Crawlers Scare You Away! 3](#_Toc482882822)

[1.4 How Do Search Engine Spiders Choose Which Webpages to Visit? 3](#_Toc482882823)

[1.5 What Search Engine Spiders are NOT Looking for: 4](#_Toc482882824)

[1.6.1 On-Page factors: 5](#_Toc482882825)

[1.6.2 Site factors: 6](#_Toc482882826)

[1.6.3 Off Page factors: 6](#_Toc482882827)

[1.6.4 Domain factors: 7](#_Toc482882828)

[1.7 White hat versus black hat techniques 8](#_Toc482882829)

[1.8 As a marketing strategy 9](#_Toc482882830)

[Chapter 2: Why Python? 10](#_Toc482882831)

[2.1 Python in Brief. 10](#_Toc482882832)

[2.2 What makes for power in a programming language? 11](#_Toc482882833)

[2.2.1 Python is Efficient. 11](#_Toc482882834)

[2.2.2 Python is fast. 11](#_Toc482882835)

[2.2.3 Python is broad. 11](#_Toc482882836)

[2.2.4 Easy-to-learn: 11](#_Toc482882837)

[2.2.5 Easy-to-read: 11](#_Toc482882838)

[2.2.6 Easy-to-maintain: 11](#_Toc482882839)

[2.2.7 Interactive Mode: 12](#_Toc482882840)

[2.2.8 Portable: 12](#_Toc482882841)

[2.2.9 Extendable: 12](#_Toc482882842)

[2.2.10 Databases: 12](#_Toc482882843)

[2.2.11 GUI Programming: 12](#_Toc482882844)

[2.2.12 Scalable: 12](#_Toc482882845)

[Chapter 3: PROJECT METHODOLOGIES 13](#_Toc482882846)

[3.1 Introduction. 13](#_Toc482882847)

[3.2 Literature Survey 13](#_Toc482882848)

[3.3 GENERAL CRAWLING STRATEGIES: 14](#_Toc482882849)

[3.3.1Breadth-First Crawling 15](#_Toc482882850)

[3.3.2 Depth First Crawling 15](#_Toc482882851)

[3.3.3 Targeted Crawling 15](#_Toc482882852)

[3.3.4 Page Rank Algorithm 15](#_Toc482882853)

[Table 1: Comparison on Crawling Algorithms 15](#_Toc482882854)

[**Algorithm** 15](#_Toc482882855)

[**Search Pattern** 15](#_Toc482882856)

[**Benefits** 15](#_Toc482882857)

[**Drawbacks** 15](#_Toc482882858)

[3.4 CRAWLING ALGORITHMS 16](#_Toc482882859)

[3.4.1 Naive Best-First Crawler 16](#_Toc482882860)

[3.4.2 Shark Search 16](#_Toc482882861)

[3.4.3 Focused Crawler 17](#_Toc482882862)

[3.4.4 Context Focused Crawler 17](#_Toc482882863)

[3.4.5 Info Spiders 17](#_Toc482882864)

[3.5 DESIGN IMPLEMENTATION 18](#_Toc482882865)

[Chapter 4: Result (Output) 20](#_Toc482882866)

[4.1 Indexation of EBAY (http://www.ebay.in/) 20](#_Toc482882867)

[4.2 Robots.txt file has been found. 20](#_Toc482882868)

[4.3 Sitemaps has been found in HTML not in XML. 21](#_Toc482882869)

[4.4 Good amount of TEXT found on Web Page. 22](#_Toc482882870)

[4.5 Links found while crawling. 22](#_Toc482882871)

[4.6 Main Title (specific Class) which were Crawled. 23](#_Toc482882872)

[4.7 What a Crawler See VS What a User See? 23](#_Toc482882873)

[4.7.1 What a Crawler Crawls? 23](#_Toc482882874)

[4.7.2 What a User is Able to See? 24](#_Toc482882875)

[4.8 Accessibility Overview for Web Page (using Tool). 24](#_Toc482882876)

[Chapter 5: Conclusion 25](#_Toc482882877)

[5.1 Overview of Search Engine Optimization. 25](#_Toc482882878)

[5.2 Reasons to Choose SEO as Your Career: 25](#_Toc482882879)

[5.2.1 High demand for SEO services 25](#_Toc482882880)

[5.2.2 A LOT of people have made a successful SEO career 25](#_Toc482882881)

[5.2.3 Search Engine Optimizers make Good Money 25](#_Toc482882882)

[5.2.4 Only Web–Designing MAY NOT be enough. 26](#_Toc482882883)

[5.2.5 Logical step ahead if you come from marketing or advertising 26](#_Toc482882884)

[5.2.6 Lots of Learning 26](#_Toc482882885)

[5.2.7 SEO is already recognized as a career 26](#_Toc482882886)

[5.3 Ugly face of SEO 26](#_Toc482882887)

[5.3.1 Dependent on search engines 26](#_Toc482882888)

[5.3.2 No fixed rules 27](#_Toc482882889)

[5.3.3 Rapid changes in rankings 27](#_Toc482882890)

[5.3.4 SEO requires Patience 27](#_Toc482882891)

[5.3.5 Black hat SEO 27](#_Toc482882892)

[Why SEO industry is growing? 27](#_Toc482882893)

[5.4 SEO Salary & Benefits 27](#_Toc482882894)

[5.5 Scope of SEO 28](#_Toc482882895)

[5.6 SEO Job Categories 28](#_Toc482882896)

[References: 29](#_Toc482882897)

[Annexure 30](#_Toc482882898)