SLIIT
Discover Your Future

Prevention Of Data Leakage By Malicious Web Crawlers

H.P.Somarathne
Reg. No.: MS20904128
M.Sc. in IT
Specialized in Cyber Security

Supervisor : Mr. Kavinga Abeywardena

November 2021

Department of Information System Engineering
Faculty of Computing
Sri Lanka Institute of Information Technology

Declaration

I hereby declare that the project work entitled "Prevention Of Data Leakage By Malicious Web Crawlers", submitted to the Sri Lanka Institute of Information Technology is a record of an original work done by me, under the guidance of the respective Supervisor. This project work is submitted in the partial fulfillment of the requirement for the award of the degree of MSc in IT (cybersecurity specialization). The results embodied in this report have not been submitted to any other University or Institution for the award of any degree or diploma. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

____________________ ____________________

Date Signature

Acknowledgement

I would like to express my heartfelt gratitude to Mr. Kavinga Yapa Abeywardana, For his exceptional leadership, monitoring, and consistent encouragement over the duration of this research project. Please accept my gratitude for the suggestions and for the ongoing assistance he has provided in completing this project effectively.

I would like to express my appreciation to the Sri Lanka Institute of Information Technology for providing the required resources to help students develop their abilities and skills. Finally, I'd want to express my gratitude to all of those who assisted us in a variety of ways and provided us with constant encouragement, without which this task would not have been possible.

Table of Contents

1.1 Abstract

Web crawlers are tools that are used to search for information on the internet in order to access it. Since the beginning of public use of the internet, web crawlers have made it easier for search engines to index the content on the internet. Unfortunately, Web Crawlers can be used for nefarious purposes as well as for legitimate ones. Because of the rising use of search engines and the prioritization of the need to get a higher ranking in the indexing of online sites, the threats posed by web crawlers have expanded significantly. In web crawlers, the robots exclusion standard is the regulating point. It establishes a set of criteria for the approved paths that a web crawler can take. Crawlers, on the other hand, are able to circumvent these restrictions and retrieve information from restricted web pages. Due to this, web crawlers can collect information that can be used for phishing, spamming, and a variety of other unethical and illegal activities. This has a significant impact on service providers, as web crawlers can collect information that can be used for phishing, spamming, and a variety of other unethical and illegal activities. The purpose of this study is to introduce a unique field of research into the detection and prevention of web crawlers. As a result of the low amount of traffic production, typical crawler detection methods were found to be ineffective at capturing dispersed web crawlers, which was discovered. Specifically, the research combines improved conventional web crawler prevention methods with a novel crawler detection method in which the threshold values are measured. This method adds distributed web crawlers to the restriction list, preventing them from traversing the websites, as well as to the restriction list itself. In order to measure threshold values, the LMT (Long tail threshold model) is being presented as a method of measurement. Furthermore, the detection methodology is built on the basis of the observation of crawler traffic and the identification of unique characteristic patterns of them in order to distinguish them from human-generated traffic, as previously mentioned. A limitation approach is incorporated into the system in order to reduce the influence that a crawler can have on a website.