BAB I
INTRODUCTION

1.1. Background

The dark web may be a dangerous place where illicit things are bought and sold, as well as where malicious websites like malware and hacker sites are frequently disseminated [1]. Because the dark web is essentially a network system accessed through numerous levels of protection (encryption) under a conventional internet network, it is frequently misused [2]. Not everyone can view the websites and content found on the dark web because of this multilayer encryption. Aside from layered encryption, the anonymity mechanism that shields users' identities from discovery makes it impossible to identify the identity of those behind illicit transactions on the dark web. There are still few investigations into illicit content on the dark web because of how tough it is to obtain. Nonetheless, content analysis using crawling to gather dark web data aided by the TOR network can be used to investigate dark web content and get information on already existing content.

Crawling is a web crawling technique that involves navigating through pages connected to the main page of a website [3]. This crawling process is carried out using a URL approach that matches the page of a chosen topic. Using the URL technique, a list of URLs for related websites on a specific subject is generated based on the keywords typed in. For those looking for information or data on the subject, collecting URLs gathered from the web through the crawling approach is highly helpful. There have been numerous crawls on the visible web and a few on the hidden web. Because the dark web requires several levels of encryption to access, crawling it is more challenging than on the surface web.

Pre-processing is also utilized for hyperlinks and content on the dark web after crawling to retrieve content. Pre-processing is carried out by classifying URLs according to the keywords they include. Data and information on web content on websites selling illicit goods and selling data from official sites like education, e-commerce, and others can be known thanks to the crawling and pre-processing techniques on the dark web. Following the acquisition of content-related data, additional analytics are required due the growing number of illicit transaction abuse therefore, more research is required to identify the most prevalent material on the dark web.

Previous researchers have used the directed graph method to group content on the dark web based on keywords discovered there. The outcome of their study is that the directed graph method can group data in more than half of the cases examined [4]. To be able to classify them that correspond with the content keywords, this research still requires work, particularly in identifying the content keywords. Furthermore, Cyber Threat Intelligence (CTI) has also employed analytics techniques to identify security breaches via a basic graph [5]. It is still important to enhance data analysis because using this basic graph just offers a basic data visualization presentation. The study of the spread of online radicalization has also used graph-based analytics techniques. By comparing the original sentiment sender with the processed data using an inter-thread graph, this research focuses on the case of hate speech on online forums. More data analysis is necessary to improve the outcomes as this research is still in its early stages [6]. Social network analysis has also been done on the Twitter platform, where the spread of COVID-19 has been observed [7]. This study, however, differs from others in that it looks at how social platform dynamics differ from those of the dark web.

This graph-based analysis method has been used with variable degrees of success in several situations. Nevertheless, prior studies have demonstrated that this method can be enhanced by employing focused crawling techniques to make better use of analysis data, leading to more readable keywords for the content that has been analyzed. Furthermore, no further analysis has been done; prior research on the dark web has only included visualizations with directed graphs. This study gathers data on the dark web through targeted crawling, as opposed to earlier studies that used directed graphs. It then presents the data as an undirected graph that illustrates the relationships between each URL and other URLs that share the same terms. Because the use of directed graphs in earlier studies produced a graph shape that was divided from one cluster to another, using undirected graphs produces higher density and modularity values than using directed graphs.

Based on this visualization, content centrality is calculated to obtain the most popular URLs and content on the dark web by utilizing degree, relatedness, and closeness. Moreover, URL density analysis and modularity analysis are carried out to guarantee the standard of URL sharing in every cluster. This centrality, density, and modularity analysis enhances data analysis by producing a graph layout that complies with appropriate scientific criteria for visual analysis and significant ranking

1.2. Problem Statement

There is less security for the nation and less privacy protection, which can result in the sale of personal information, the more widespread illicit activities that are carried out on the dark web. These activities include the sale of drugs, alcohol, personal data, malware, hackers, and other items. Researchers studying network security also require access to the limited amount of information that is currently available about the dark web, in addition to security concerns. To learn more about the data on the dark web, analytics about the content can be performed based on these issues. One method suggested to determine what kind of content is most frequently found on the dark web is content analysis.

A problem with this research, aside from security concerns, is the dearth of knowledge regarding content on the dark web. Only topic keyword information found on the dark web is displayed by research on the topic using a directed graph [4]. Additionally, this study does not carry out any additional analysis and only employs a degree approach analysis on the 200 available data. Because all of the sides of a directed graph are one-way, not every node is in the same location, and there's a chance that some nodes are not connected because they're outside of the loop, making the graph asymmetric. Because the directed graph is asymmetric, it is challenging to precisely analyze the information flow on the dark web because the content is frequently dispersed, intricately connected, and lacks a clear direction. It can also move unexpectedly in different directions. When it comes to dark web analysis, the relationships between contents frequently matter more than the actual direction of the relationships.

For the analysis of content on the dark web, directed graph is still unsuitable. Because undirected graphs can be applied to structures with extremely complex environments where patterns and connections may not always be immediately clear, research on dark web content has been developed in response to the shortcomings of directed graphs for content analysis. Because undirected graphs can represent relationships in a variety of ways, highlight connections between nodes, detect intricate patterns, employ metrics flexibly, and analyze content's association and connectedness regardless of connection direction, they are frequently a better choice for content analysis on the dark web. In this research, research development was conducted with 1000 dark web URL node data and centrality analysis using density, modularity, betweenness, closeness, and degree centrality. Because of the nature of directed symmetry and asymmetry, which allows nodes in directed graphs to be disconnected from one another, using undirected graphs results in higher densities than directed graphs.

1.3. Objectives

Analytics are used to determine the content and information available on the dark web-based on the problem statement concerning data on the dark web. Analytics are performed using graph visualization, which seeks to offer a clear and concise picture of the unprocessed data that was gathered during the crawling procedure. Additionally, the purpose of crawling is to obtain data from the dark web for additional examination. It can then determine which content is most popular on the dark web based on the findings of the analysis that was done. The goal of this study is to:

1. Establish a dark web network and use the technique that has been developed to obtain content trends on the dark web to visualize data based on topics (keywords).
2. Creating graph-based analytics with graph centrality using degree, betweenness, and closeness to get the dominant content URL.
3. Obtain graph density values on a collection of dark web URLs with density analysis.
4. Obtain network strength value based on the graph with modularity analysis.

1.4. Scope of Problem

To achieve the research objectives, it is necessary to restrict the problems in the research based on the objectives. The following are the research problem's limitations.

1. The dark web is the subject of the investigation.
2. Utilizing the dark web's URLs and topic keywords for research
3. Focused crawling is used to create the crawling system.
4. Network performance is not covered in this research.
5. Cookies and captcha are not covered in this study.

1.5. Hypothesis

It's known from earlier crawling studies that we can use the focused crawling method to investigate the dark web. Using URL crawling to navigate from a topic's main page to its seed, one can browse pages on a specific topic using the focused crawling method. Despite being encrypted in layers this focused crawling technique is still applicable on the dark web. These crawling techniques when combined can yield the best possible dark web data, which is then grouped using graph visualization according to relevant content keywords. Analyses utilizing the graph centrality method with the visualization data display are possible because this method can yield dominant content that is prioritized on the dark web. The values of degree, betweenness, and closeness are calculated using the relationship between each node, intermediate nodes, and mapping between nodes to determine centrality.

The density of websites on the dark web and the distribution of their content data can be found in addition to centrality analysis. The dark web network shape can be obtained by applying the density and modularity methods to search for density values and data distribution. The content information on the dark web can be explained by combining methods of centrality, density, and modularity.

1.6. Research Method

Multiple research methods are employed in this analytics study. These are the research methods that were used.

1. Review of Literature

   The gathering of reference materials following the research that needs to be done is done at this point in the literature study. Data on the dark web, crawling, TOR, data visualization, and graph analytics are gathered at this stage.
2. Analysis of System Requirements

   An analysis of the requirements needed to conduct this research is done using the data gathered from the case study approach. It is possible to prepare the requirements needed to ensure the research proceeds without a hitch.
3. System Design

   In addition, the design of the system under study is completed following a review of the literature and an analysis of the system's requirements. Using centrality, density, and modularity, the focused crawling system on the dark web is designed and examined at this point in the process.
4. Execution

   Research is started at this implementation stage by examining the current system architecture. Currently, the process begins with crawling the dark web and producing analytics to obtain the most popular content.
5. System Assessment and Testing

   Testing is done on the created system after it has been put into operation. The employed analytics method is then put to the test to see if it is appropriate and accurate, and if the system yields the best results or not. One can assess analytics based on the test results.

1.7. Methodology

An automatic computer program known as a "crawler" or "web crawler" searches web pages using hyperlinks; this technique is also frequently called "web spider" or "web robot." Web crawling also referred to as spidering, is a method by which web crawlers gather information from the internet [5]. Crawling focused is a crawling technique that looks for related keywords to narrow the focus of a web page search. This targeted crawling technique will examine each hyperlink related to the specified page or keyword. The bioinformatics web has been searched for information sources using focused crawling.

Additionally, data can be transparently gathered by crawling through security websites on the clear web, social media security forums, and dark web hacker marketplaces and forums. Data harvesting in the proposed crawling architecture is done in two stages. First, websites of interest are identified through machine learning-based crawling. Next, the information is represented in a latent low-dimensional feature space and ranked according to its possible relevance to the task using sophisticated statistical language modeling techniques [6].

Many criminal elements use the dark web to commit crimes because of its anonymity, even if they only use traditional network survey techniques based on IP addresses. To help with deeper exploration, a visual dark web forum posting analysis system is created to represent the relationship between different forum messages and posters visually. Law enforcement officials and researchers studying dark networks can use the research framework of forum association visualization, which is investigated through forum news, to examine unlawful and criminal activity [7].

To discover valuable cyber threat intelligence, descriptive analytics and predictive analytics using machine learning on a dataset of dark web forum posts were used to perform analytics on dark web data. Watson Analytics and WEKA from IBM are also used, with Watson Analytics displaying trends and relationships in the data and WEKA providing machine learning models to classify the types of exploits targeted by hackers from form posts. Furthermore, machine learning aids in the development of classifiers for exploit types [8].

Using a degree graph, web archive analytics methods have also been carried out with the concept of 'blind spots' (live web features that were not included during archive creation). This analytics method research can identify content that is relevant to the object of research, but it may not be possible to do better with the complexity of web archives [19]. In addition, the analytics method was applied to discuss data from six dark web forums using social network analysis metrics and analysis methods. The large-scale structure and influential nodes in this network can be identified using this research methodology, but subnetwork analysis based on network division into discussion topics is required [20].

Content keywords that can be found on the dark web are required to analyze content from the dark web using specific content topics. keywords found in research articles that contain text on the first page. Topic data on the dark web is roughly divided into six categories, according to research on dark web content analysis: hacking, drugs, development, porn, news, and casinos. According to the study's findings, half of the content on the dark web can be mapped, with news and hacking accounting for the largest percentages at 13% and 8%, respectively, followed by drugs [4].

Furthermore, it is also well known that, through 2021, Bitcoin transactions will be extensively conducted on the dark web. The dark web contains more than the dark 31 markets, where these transactions are primarily conducted [29]. A merger was performed using keywords from the categories of drug, hacking, and bitcoin-related content, based on the content that had been discovered by earlier researchers. Because this content doesn't contain any links to illicit websites.

1.8. Writing Systematic

The following topics are systematically arranged into several discussion sections within this research:

Chapter 1 INTRODUCTION

Background, issues, goals, constraints, research techniques, and writing systematics are all included in this chapter.

Chapter 2 BASIC THEORY

Research-related theories and fundamental concepts are covered in this chapter.

Chapter 3 DESIGN SYSTEM AND MODEL

Process flow and system design flow are covered in this chapter.

Chapter 4 RESULTS AND ANALYSIS

The test results and the analysis of the test results obtained are included in this chapter.

Chapter 5 CONCLUSION

The thesis' recommendations and conclusions are presented in this chapter.