ISSN: 0067-2904

Crawling and Mining the Dark Web: A Survey on Existing and New Approaches

Mohammed Khalafallah Alshammery[1*], Abbas Fadhil Aljuboori[2]
[1]*College of Information Technology, University of Babylon, Babil, Iraq,*
[2]*University of Information Technology and Communications, Baghdad, Iraq,*

Received: 26/1/2021 Accepted: 2/5/2021

Abstract

The last two decades have seen a marked increase in the illegal activities on the Dark Web. Prompt evolvement and use of sophisticated protocols make it difficult for security agencies to identify and investigate these activities by conventional methods. Moreover, tracing criminals and terrorists poses a great challenge keeping in mind that cybercrimes are no less serious than real life crimes. At the same time, computer security societies and law enforcement pay a great deal of attention on detecting and monitoring illegal sites on the Dark Web. Retrieval of relevant information is not an easy task because of vastness and ever-changing nature of the Dark Web; as a result, web crawlers play a vital role in achieving this task. Thereafter, data mining techniques are applied to extract useful patterns that would help security agencies to limit and get rid of cybercrimes. The aim of this paper is to present a survey for those researchers who are interested in this topic. We started by discussing the internet layers and the properties of the Deep Web, followed by explaining the technical characters of The Onion Routing (TOR) network, and finally describing the approaches of accessing, extracting and processing Dark Web data. Understanding the Dark Web, its properties and its threats is vital for internet servers; we do hope this paper be of help in that goal.

Keywords: Dark Web, Deep Web, Data Mining, Crawling, TOR (The Onion Routing)

الزحف والتعدين في شبكة الوب المظلمة: مسح حول المناهج الحالية والجديدة

محمد خلف الله الشمري*[1] وعباس فاضل الجبوري[2]
[1]قسم البرمجيات, كلية تكنولوجيا المعلومات, جامعة بابل, بابل, العراق
[2]قسم هندسة الحاسوب, كلية الهندسة, جامعة تكنولوجيا المعلومات والاتصالات, بغداد, العراق

الخلاصة

شهد العقدان الماضيان زيادة ملحوظة في الأنشطة غير القانونية على شبكة الوب المظلمة. إن التطور السريع واستعمال البروتوكولات المعقدة يجعل من الصعب على الوكالات الأمنية تحديد هذه الأنشطة والتحقيق فيها بالطرق التقليدية. علاوة على ذلك ، يمثل تعقب المجرمين والإرهابيين تحديًا كبيرًا مع الأخذ في الاعتبار أن

*Email: mohammed.alshammery@student.uobabylon.edu.iq

الجرائم الإلكترونية لا تقل خطورة عن جرائم العالم الحقيقية. في الوقت نفسه ، تولي جمعيات أمن الحواسيب وإنفاذ القانون قدرًا كبيرًا من الاهتمام لاكتشاف ومراقبة المواقع غير القانونية على شبكة الوب المظلمة. استعادة المعلومات ذات الصلة ليست مهمة سهلة بسبب اتساع وتغير طبيعة شبكة الوب المظلمة ؛ نتيجة لذلك ، تلعب برامج زحف الوب دورًا حيويًا في تحقيق هذه المهمة. بعد ذلك، يتم تطبيق تقنيات التنقيب عن البيانات لاستخراج أنماط مفيدة من شأنها أن تساعد وكالات الأمن على الحد من الجرائم الإلكترونية والتخلص منها. الهدف من هذه الورقة هو تقديم مراجعة للباحثين المهتمين بهذا الموضوع. بدأنا بمناقشة طبقات الإنترنت وخصائص الوب العميق ، متبوعًا بشرح الخصائص التقنية لشبكة توجيه البصل ، وأخيراً وصف طرق الوصول إلى بيانات الوب المظلمة واستخراجها ومعالجتها. يعد فهم الوب المظلم وخصائصه وتهديداته أمرًا حيويًا لخوادم الإنترنت ؛ نأمل أن تساعد هذه الورقة في تحقيق هذا الهدف.

1. Introduction

Internet is one of the broadest accomplishments of mankind that have seen rapid development, getting attention of researchers of various specialties to develop more and more applications to make it accessible to a great variety of users, from persons to foundations, but guaranteeing confidentiality and security at the same time [1].

The internet is much larger than what we see. Available search engines, e.g. Google, search approximately 4% of the entire web only. In addition to this searchable content, there is a great deal of resources and data that present on the web and such sites are commonly called as Deep Web and Dark Web. Deep Web commonly pointing to resources and data that are not accessible with usual search engines and hyperlinks. The part of Deep Web that is largely utilized for illegal actions, such as weapon trading, child abuse, drug trafficking, etc., is called Dark Web [2].

The proportion of the Dark Web that is used for illicit actions and illegal contents is approximately 57% [3]. The Dark Web usually depends on incorporating crypto currencies such as bitcoins with anonymized access as bases in founding a marketplace for dealing weapons, drugs, and other contrabands [4]. The term Dark Web was first used in the 2000s and has been widely employed both in the media and academia since then [4]. It became well-known with the introduction of "Silk Road", a drugs market in 2011 to its closure in 2013[3].

The most important difficulties the analysts encounter whilst investigating criminal activities in the Dark Web is the anonymity offered in Dark Web service [3]. The most familiar services in the Dark Web is The Onion Routing(TOR) network that offers the ability for the individuals to privately and anonymously share data by peer to peer dealings instead a central server [3]. TOR was originally created as part of the US Naval Research Laboratory's protected communication system, to safeguard and anonymize traffic by transferring it through several layers of encrypted relays [5]. The hidden service protocol of TOR allows web services to stay anonymous by concealing the IP addresses of network servers through several relays within TOR's overlay network [6].

The scope of the current study focuses on the research presented in the second decade of this century. The main issues dealt with in this study are how to access the dark web, extract data, process it and analyze it. These issues have been collected from several studies published in accredited sites (i.e., IEEE, Google Scholar, Springer.. etc.) and many keywords have been used in the research process (i.e., crawling, mining, the onion routing, threat, monitoring the dark web .. etc.) and the number of researches used in this The survey is approximately thirty research papers, and a number of unrelated research papers were excluded. The main differences of this study from the previous studies by discussing crawling and mining in the Dark Web.

This study includes seven sections: started by introductory section, Internet Layers, The Onion Routing, Dark Web Crawler, Dark Web Mining, Discussion and Conclusion.

2. Internet Layers

Dark Web is a subset of Deep Web, the part of the Web not indexed by web browsers and cannot be reached via usual search engines. However, it could be accessed by specific software that provides entry to anonymity networks. Thus, planned steps are required to access the Dark Web, which works exclusively anonymously both for the user and the service provider [7]. Figure 1 describes the Internet layers.

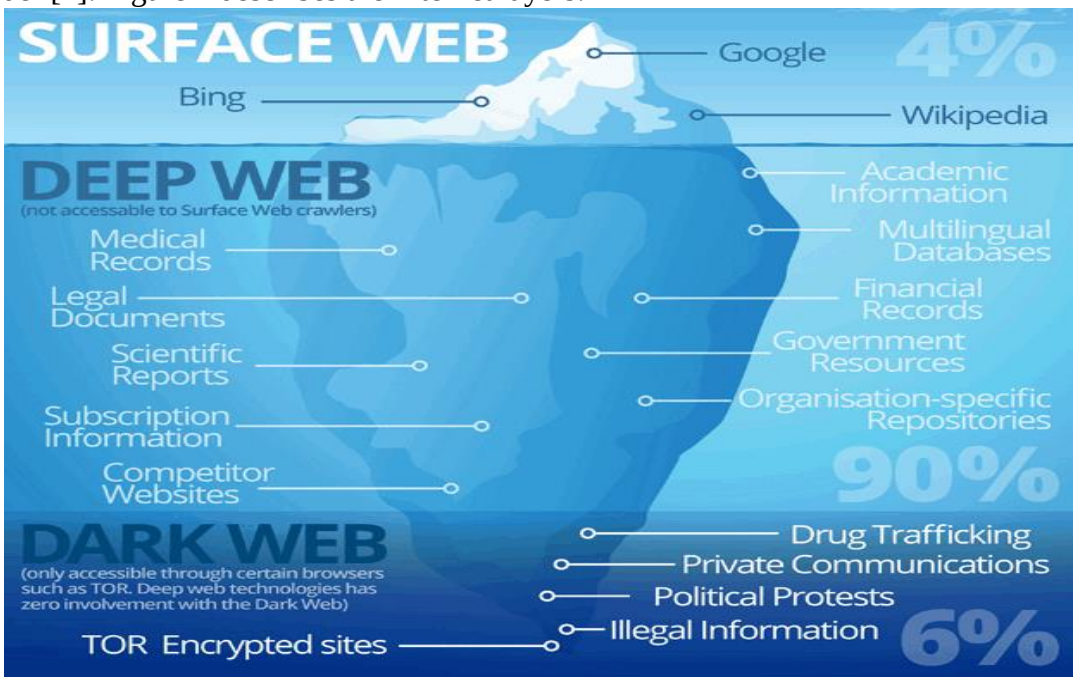

Figure 1-the Internet layers. Source: designed by the author.

2.1 Surface Web

The Surface Web is the portion of the Internet that is indexable by search engines. Other names of this portion of the Internet are Indexable Web, Visible Web, Lightnet, or Clearnet [8][9]. Surface Web is the part of the internet that is considered public and accountable. It is public because it is accessible and not restricted by authentication or payment and it is accountable as users are identifiable thus liable to law enforcement [9].

2.2 Deep Web

The Deep Web is the part of the Internet that is not indexed by search engines and not connected to pages on the Surface Web. It started in 1994 and was called Hidden Web pages by then, the term Deep Web was introduced for the first time in 2001 [10]; although, occasionally the term Deep Web is falsely used to indicate specifically to the Dark Web. It has other names such as Deep net and Invisible web. According to researches, the Deep Web and Dark Web account for approximately 96% of the whole WWW, while the Surface Web constitutes the remaining 4% [11]. In the Deep Web there are websites that are more complex and there are websites that contain information about research data and confidential information [8]. Measuring the size of the Deep Web is not possible; the ongoing change in accessing and presenting information means that the Deep Web is growing rapidly and at a frequency that challenges quantification.

The difficulty indexing a webpage could be ascribed to many reasons. Firstly, use of passwords so that the crawlers cannot access the web page. In addition, use of limited number of accessing times restricts accessibility, when the page becomes inaccessible prior the crawler can get to it. Similarly, "the robots.txt file", which is located on the root of the web site, is designated to inform the crawler not into crawl into that website or parts of it. Finally,

the page is unreachable unless the whole URL is well known, because it is either hidden or unlinked to other pages on that site or other websites [1].

According to the activities that users engage in on the Deep Web, it can be divided into areas of legal activities and areas of illegal activities. Legal activities include virtual academic libraries and databases, or libraries of research papers, or just browsing anonymously or when users prefer not to be tracked. To maintain its privacy, many parties carry out their activities on the Deep Web, such as the security and military forces, as well as the press, media and others. On the other hand, illegal activities involve actions that are classified as criminal or illicit, so this part constitutes the Dark Web [11].

2.3 Dark Web

We noticed that there are various definitions for the Deep Web and Dark Web. However, Deep Web can be defined as the part of the internet that constitutes sites which web browsers can't find or indexed, whereas the Dark Web is composed of invisible nets that depends on special programs and protocols [12].

While there are legal uses of Dark Web (for example, New York Times employs it to allow confidential communication with its sources) [13], it is the site where most of the criminal activities happen. Such activities are drug deals, human organ trafficking, weapon trade, child pornography, trading critical data, malicious and spyware programs; commute data that hackers detect in computer systems, or renting a Bot net, a full equipped net linked to the web that hacktivists can use to carry out a broad spectrum security violation. In addition to fake IDs, trading documents, patients' medical records, stolen credit cards, and any other personally identifiable information. It also includes financial fraud, disseminate criminal ideologies, a funded assassination market where people pay to having somebody assassinated, and a lot more. Dark Web Hidden Services are conducted in the Dark Web as well, they are specific services that deal with cyber security attacks, and act as the milieu for malwares [4][14].

From a social perspective, Dark Web activities rely on the powerful society framework that the members of Dark Web sites concentrate on. The websites on the Dark Web, including virtual markets, require somebody who directs them and keeps their privacy and security to permit clients just to focus on their deals. This agent is in charge of running the websites, advertising products, handling traffics, and frequently acting as third parties during trade transactions, where credibility has a vital role [1].

The most popular electronic marketplace on the Dark Web is "Silk Road". Founded by Ross William Ulbricht in 2011, this marketplace specializes in drug trade and electronic products such as malware, piracy services, hacked multimedia, fraud, passports, and social card fraud. In September 2013, the FBI closed the site, and in October of that year, Ulbricht was arrested. He was sentenced to life in prison in 2015 after he collected more than thirteen million dollars from his trading and commission, while the site achieved more than (1.2) billion dollars from sales between 2011 and 2013, according to the US Federal Court [4][5][15].

3. The Onion Routing

When you get in to the Dark Web, the websites and most services can be reached via a browser in the similar way such as the Surface Web. However, there are several web sites in the Dark Web that are deliberately hidden, that means that they have not been conventionally indexed by a search engines and for this reason such web sites can only be accessed if you specifically has the URL of these sites [14].

Tools such as TOR, the Invisible Internet Project (I2P), and Free net, are needed to access the Dark Web. In addition, some levels are entered with permission and password verification. The programs used to access the Dark Web provide the privacy of the data source as well as

the privacy of the people who access the target data. Thanks to this feature, people also prefer to move data to the Dark Web to hide it [16][17].

TOR uses specially configured computers to pass requests across a net of linked nodes; consequently, it gives a specific degree of privacy and anonymity. While message goes from one relay to the other, it is encrypted in such a way that each relay only knows about the computer that sent the request and the computer it is suggested to send to [11][14].

Given that the Deep Web is multiple networks, websites, and databases that need specific protocols to gain access to and are not easily available to everyone via conventional Web browsers, the TOR Network has been developed as the most popular Deep Web technology. It is so called because it employs multiple layers (as an onion) of encryption. Currently, the TOR Project is an open source non-profit organization with a big society. As it is an overlay network, TOR uses the already existing TCP/IP infrastructure [18][19][20].

In spite of the main reason of development of such web, TOR has been the optimal tool for many websites to perform illegal activities and at the same time maintaining anonymity of their operators and clients. Examples of popular virtual markets are Silk Road, and Agora. Some of these sites can be readily found on the network by accessing some pages that work as references of lists of links to these websites like HiddenWiki; or by utilizing specialized search engines available on TOR network like TOR Search, Duck Duck Go, and Grams. However, all these techniques can only gain access to a narrow range of hidden services on the network [1].

Once joining the network through TOR Browser, the TOR user, also known as a source, will be connected to a virtual circuit of randomly selected TOR nodes (commonly 3 computers that run the TOR Browser will be selected). After approximately ten minutes, this virtual circuit will be replaced by a new one [12].

The virtual circuit is composed of three kinds of nodes:

1. EntryNode: this node receives arriving traffic.
2. IntermediateNode: transfers information from one to the following node.
3. ExitNode: the last one which conveys traffic to the Surface Web (destination).

Exit relays may make requests on behalf of hundreds of users to make them anonymous; which exit relay is used is determine by randomizing algorithms. Worldwide, there are about seven thousand computers that work as relays. As a result, each user is hidden among multiple layers of the onion [20]. Figure 2 highlights Components of the TOR Network:

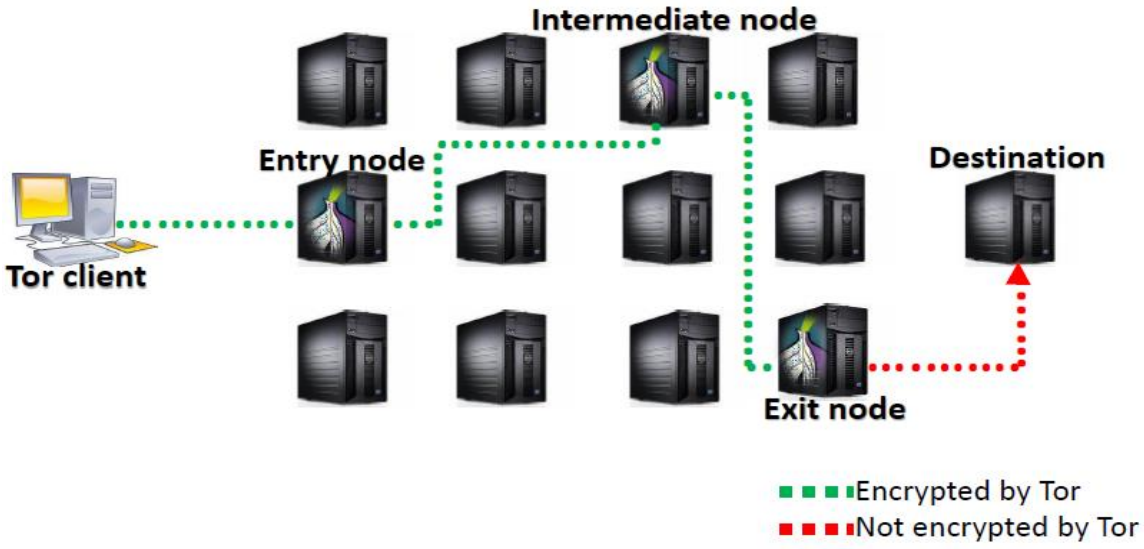

Figure 2- Components of TOR network. Source: designed by the author.

4. Dark Web Crawler

Web crawlers can be employed to collect websites automatically after determining the Dark Web forums and markets, a custom web crawler is used to detect a primary seed site [21]. The web crawler works by reaching the internet to gather and store data into database for subsequent assortment and analysis. The procedure of web crawling includes collecting pages from the net [22], and then automatically downloading them while following the hyperlinks encountered and consistently catching new webpages. Properly downloaded illicit deal data can then be processed and categorized for longer-term storage [21]. Figure 3 illustrates flow char of crawling into data.

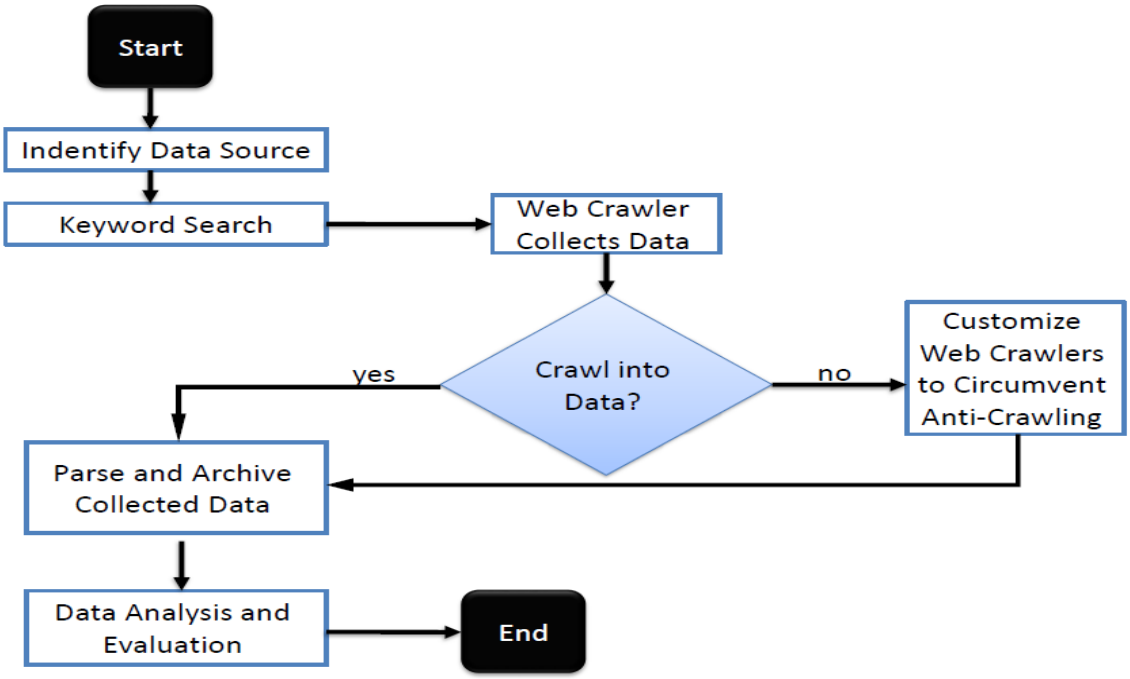

Figure 3- Flow chart of dark web data crawling [21].

A web crawler, also known as web spider or shortly crawler, is an Internet bot that crawls through an HTML website and collects information regarding that site such as the page titles, websites URL, metatags, web page contents, and most significantly links that are found in the page. Through the links it has collected in the initial page, the crawler then visits and stores the same data of the subsequent pages. Web crawler operates through sending a source, as robots.txt, which then will deliver all the information to the server [8][11].

In the last twenty years, crawling program development have seen an increasing concern in Dark Web, but with the multiple challenges involved (which we have previously mentioned), developing such programs requires additional techniques, so that the crawlers would be capable of discovering malevolent websites, accessing them, and storing their data for future processing [1].

As previously stated, the size of the Deep Web is markedly large, and it specially involves high quality and important data in a broad spectrum of semantic domains. Accordingly, designing Deep Web crawlers that automatically access such data is an interesting research area [23].

5. Dark Web Mining

The Dark Web has a large source of data of unstructured type related to illicit activities. In order to discover and represent knowledge, this preliminary information needs to be delivered to the local system (crawling), mined and retrieved data requires further processing, i. e.,

cleaning, transformation, normalization. Then it is analyzed via data mining techniques or machine learning techniques.

Dark Web mining can be divided into the following steps:

1) **Data Extraction**

The process of extracting data from internet sources is known as web data extraction. A web data extraction system commonly accesses a web source and extract the information stored in it, after which the extracted data is analyzed, transformed into a more useful structured format and saved for future use [24].

2) **Data Pre-processing**

Data pre-processing is a data mining technique that includes preparing and transforming data into an appropriate format for the mining process. The goal of data pre-processing is to reduce the size of the data, identify relationships between data, normalize data, delete outliers, and extract features. Multiple strategies such as data cleaning, integration, transformation and reduction are involved [25].

The aim of cleaning and preparing data is to increase productivity and effectiveness in the mining process. Pre-processing methods will cut up to 80% of the total mining operation. Text pre-processing solves the feature space's high dimensionality problem, in which features (or terms) will number in the tens or hundreds of thousands. It also improves the accuracy of text analysis while saving time and space [26].

Text enters a sequence of steps that may or may not include all of more of the following: [27]

1. Text tokenization through extraction
2. Lowercase Conversion
3. Removal of Special Characters
4. Stop words Elimination
5. Lemmatization and stemming
6. Pruning rare words (as they lead to noise in data) using Document Frequency
7. TF-IDF Weighting or Bag of Word

3) **Mining The Dark Web**

Using different strategies, data mining methods or machine learning techniques are applied on clean data to extract patterns. These patterns can be completely helpful for many institutions in gaining data regarding products, sellers and their marketing styles.

We mention a number of related works and methods used by researchers in this field:

Concerning the Deep Web classification, Noor et al. [28] addressed the basic techniques for information extraction from Deep Web data sources known as "Query Probing", which is widely used for supervised learning algorithms, and "Visible Form Features"(Xian et al., 2009).

Kaur [29] introduce an informative survey covering many algorithms for classifying web content, emphasizing their relevance in data mining. In addition, the survey provided pre-processing methods that could aid feature discovery, such as removing HTML tags, punctuation marks, and stemming.

Graczyk et al. [30] suggested a pipeline to identify the goods of Agora, a well-known Dark net black market, into 12 groups with 79 percent accuracy. The TF-IDF is used for text attribute extraction, the PCA for feature collection, and the SVM for feature classification in their pipeline architecture.

Moore et. al. [31] have proposed a recent analysis focused on TOR secret services to explore and identify the Dark net. Initially, they gathered 5K TOR onion page samples and classified them into 12 classes using an SVM classifier.

Baravalle et al. [4] concentrated on Dark Web e-markets, primarily "Agora," an e-market for selling drugs and false IDs. The crawler simulates the authorization method for user login to the market before collecting data with the traditional web development tool LAMP Stack.

Rahayuda and Santiari [8] crawled the TOR Dark Web, focusing on nine domain types and defining the information or service they hosted. The researcher discovered how certain domains purposefully isolate themselves from the rest of TOR, among other things. As a classification tool, fuzzy K-Nearest Neighbor (fuzzy-KNN) was used. The crawling system results that were stored in the database were categorized using a fuzzy-KNN method. The crawling framework then produced data in the form of URL addresses and page information. The crawling and sample data processes were compared.

M. W. Al Nabki et. al. [32] published a recent report that classified TOR HS's criminal activities using two text representation systems, TF-IDF and BOW, as well as three classifiers, SVM, LR, and NB. They created dataset DUTA, which contains 7K samples labeled manually into 26 categories, including the Others class, which is only concerned with illicit activities such as drug trafficking and child pornography. They discovered that integrating the TFIDF text representation with the Logistic Regression classifier would achieve 96.6 percent precision and 93.7 percent macro F1 score over ten folds of cross validation.

Figure 4 illustrates the process of crawling and mining the Dark Web:

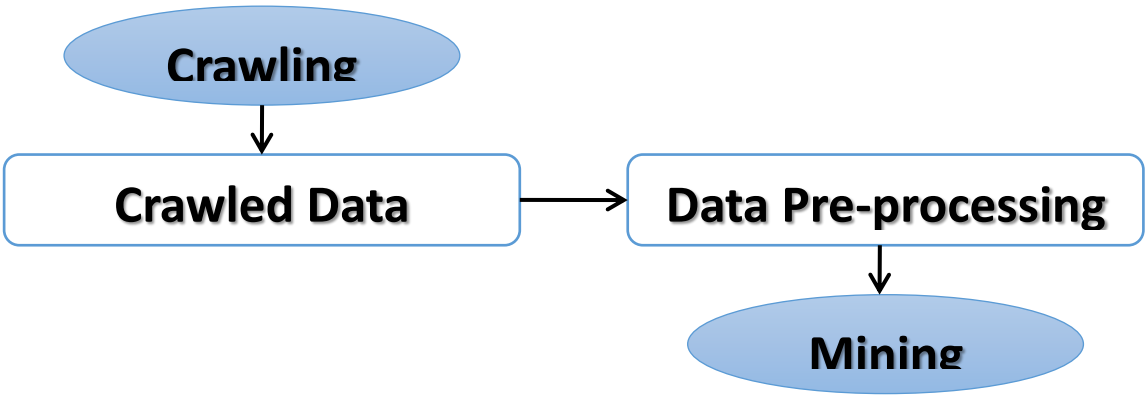

Figure 4-Crawling and Mining the Dark Web. Source: designed by the author.

6. Discussions

Crawler missions can be theoretically simple: start with seed URLs, download all pages under the selected addresses, extract hyperlinks from the pages and add them to the list of addresses, crawl on the extracted links iteratively, and so on.

Though it does not seem to be as simple as it seems, web crawling faces many challenges. These challenges are caused by TOR network features, especially uncorrelated websites, in which connections between sites are sparse, rendering it difficult for the crawler to follow.

The current study addressed the most significant challenges as:

1. Websites hosted on a private encrypted network have a shorter lifecycle than those on the Surface Web because they transfer regularly across multiple addresses, making their durability and operability time untrustworthy. Furthermore, web administrators rely on shifting websites among multiple web addresses, especially in Dark Web electronic markets, to avoid surveillance. It is worth noting that platforms operating on encrypted networks face technological challenges such as bandwidth limitations, making their connectivity much less secure than that of websites hosted on the Surface Web, and the tunnel-like transportation across multiple nodes makes loading websites hosted on TOR take longer than those with direct connections.

2. Accessibility: To access these pages, most need user authentication and approval of their group laws. To avoid automatic logins or Denial of Service (DoS) attacks, authentication and

login procedures often involve solving CAPTCHA, interactive challenges, or quizzes, which both include manual handling.

3. Professionalism and the effectiveness of the electronic environment in which they work are essential to web managers. This may involve developing a social layering structure based on the activity level of their participants, their talents, and their technical level. They still have a system in place that terminates accounts of inactive users in order to avoid attempts at secret surfing, which they deem questionable behavior.

As a result, methods and algorithms for information extraction, clustering, and text classification of data from unstructured and structured sources must be developed and implemented.

7. Conclusion

Deep Web represent the major part of the World Wide Web, it is the layer of the internet that is not indexable by search engines, and this fact is not well known by general public. Deep Web is where most of the illegal activities take place; however, it is not entirely harmful, it has many legal applications including maintaining privacy while browsing. This part of the web can be accessed by web crawling approaches to extract data which then would be processed and analyzed using data mining techniques. Notably, web mining depends on the nature of the website and the quality of data, whereas web crawler design differs from one website to another. The development of integrated crawling and mining Approaches which is our next goal, would provide a helpful method in accessing the Dark Web and contribute to limiting the illegal activities.

References

[1] B. AlKhatib and R. Basheer, "Crawling the Dark Web: A Conceptual Perspective, Challenges and Implementation," *J. Digit. Inf. Manag.*, vol. 17, no. 2, p. 51, 2019, doi: 10.6025/jdim/2019/17/2/51-60.

[2] X. Zhang and K. P. Chow, "A framework for dark web threat intelligence analysis," *Int. J. Digit. Crime Forensics*, vol. 10, no. 4, pp. 108–117, 2018, doi: 10.4018/IJDCF.2018100108.

[3] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, "Evolution of Dark Web Threat Analysis and Detection: A Systematic Approach," *IEEE Access*, vol. 8, pp. 171796–171819, 2020, doi: 10.1109/access.2020.3024198.

[4] A. Baravalle, M. S. Lopez, and S. W. Lee, "Mining the Dark Web: Drugs and Fake Ids," *IEEE Int. Conf. Data Min. Work. ICDMW*, vol. 0, pp. 350–356, 2016, doi: 10.1109/ICDMW.2016.0056.

[5] D. R. Hayes, F. Cappa, and J. Cardon, "A framework for more effective dark web marketplace investigations," *Inf.*, vol. 9, no. 8, pp. 1–17, 2018, doi: 10.3390/info9080186.

[6] S. Ghosh, P. Porras, V. Yegneswaran, K. Nitz, and A. Das, "ATOL: A framework for automated analysis and categorization of the dark web ecosystem," *AAAI Work. - Tech. Rep.*, vol. WS-17-01-, pp. 170–178, 2017.

[7] M. Schafer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders, "BlackWidow: Monitoring the Dark Web for Cyber Security Information," *Int. Conf. Cyber Conflict, CYCON*, vol. 2019-May, pp. 1–21, 2019, doi: 10.23919/CYCON.2019.8756845.

[8] I. G. S. Rahayuda and N. P. L. Santiari, "Crawling and cluster hidden web using crawler framework and fuzzy-KNN," *2017 5th Int. Conf. Cyber IT Serv. Manag. CITSM 2017*, 2017, doi: 10.1109/CITSM.2017.8089225.

[9] M. Ali, "Electronic Crime Investigation," vol. 3429, no. 1, 2019.

[10] V. V. Mahale, M. T. Dhande, and A. V. Pandit, "Advanced web crawler for deep web interface using binary vector page rank," *Proc. Int. Conf. I-SMAC (IoT Soc. Mobile, Anal. Cloud), I-SMAC 2018*, pp. 500–503, 2019, doi: 10.1109/I-SMAC.2018.8653765.

[11] A. Khare, A. Dalvi, and F. Kazi, "Smart Crawler for Harvesting Deep web with Multi-Classification," *2020 11th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2020*, 2020, doi: 10.1109/ICCCNT49239.2020.9225369.

[12] A. T. Zulkarnine, R. Frank, B. Monk, J. Mitchell, and G. Davies, "Surfacing collaborated networks in dark web to find illicit and criminal content," *IEEE Int. Conf. Intell. Secur.*

Informatics Cybersecurity Big Data, ISI 2016, pp. 109–114, 2016, doi: 10.1109/ISI.2016.7745452.

[13] P. L. Dordal, "The dark web," *Adv. Sci. Technol. Secur. Appl.*, pp. 95–117, 2018, doi: 10.1007/978-3-319-97181-0_5.

[14] M. F. Bin Rafiuddin, H. Minhas, and P. S. Dhubb, "A dark web story in-depth research and study conducted on the dark web based on forensic computing and security in Malaysia," *IEEE Int. Conf. Power, Control. Signals Instrum. Eng. ICPCSI 2017*, pp. 3049–3055, 2018, doi: 10.1109/ICPCSI.2017.8392286.

[15] W. Lacson and B. Jones, "The 21st century Dark Net market: Lessons from the fall of silk road," *Int. J. Cyber Criminol.*, vol. 10, no. 1, pp. 40–61, 2016, doi: 10.5281/zenodo.58521.

[16] A. Montieri, D. Ciuonzo, G. Bovenzi, V. Persico, and A. Pescape, "A Dive into the Dark Web: Hierarchical Traffic Classification of Anonymity Tools," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1043–1054, 2020, doi: 10.1109/TNSE.2019.2901994.

[17] I. Karunanayake, N. Ahmed, R. Malaney, R. Islam, and S. Jha, "Anonymity with Tor: A Survey on Tor Attacks," 2020, [Online]. Available: http://arxiv.org/abs/2009.13018.

[18] E. Nunes *et al.*, "Darknet and deepnet mining for proactive cybersecurity threat intelligence," *IEEE Int. Conf. Intell. Secur. Informatics Cybersecurity Big Data, ISI 2016*, pp. 7–12, 2016, doi: 10.1109/ISI.2016.7745435.

[19] A. Elgzil, C. E. Chow, A. Aljaedi, and N. Alamri, "Cyber anonymity based on software-defined networking and Onion Routing (SOR)," *2017 IEEE Conf. Dependable Secur. Comput.*, pp. 358–365, 2017, doi: 10.1109/DESEC.2017.8073856.

[20] M. Chertoff, "A public policy perspective of the Dark Web," *J. Cyber Policy*, vol. 2, no. 1, pp. 26–38, 2017, doi: 10.1080/23738871.2017.1298643.

[21] H. YU, Y. YANG, L. YANG, and G. ZHU, "Dark Web Threat Intelligence and Market Analysis," *DEStech Trans. Environ. Energy Earth Sci.*, no. iccis, pp. 470–477, 2019, doi: 10.12783/dteees/iccis2019/31697.

[22] M. Manke, K. K. Singh, V. Tak, and A. Kharade, "Crawdy : Integrated crawling system for deep web crawling," vol. 4, no. 9, pp. 389–393, 2015, doi: 10.17148/IJARCCE.2015.4984.

[23] N. Agrawal and S. Johari, "A Survey on Content Based Crawling for Deep and Surface Web," *Proc. IEEE Int. Conf. Image Inf. Process.*, vol. 2019-Novem, pp. 491–496, 2019, doi: 10.1109/ICIIP47207.2019.8985906.

[24] E. Ferrara, P. De Meo, G. Fiumara, and R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-Based Syst.*, vol. 70, pp. 301–323, 2014, doi: 10.1016/j.knosys.2014.07.007.

[25] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *J. Eng. Appl. Sci.*, vol. 12, no. 16, pp. 4102–4107, 2017, doi: 10.3923/jeasci.2017.4102.4107.

[26] P. C. Gaigole, L. H. Patil, and P. M. Chaudhari, "Preprocessing Techniques in Text Categorization," *Natl. Conf. Innov. Paradig. Eng. Technol.*, pp. 1–3, 2013.

[27] A. I. Kadhim, "An Evaluation of Preprocessing Techniques for Text Classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 6, pp. 22–32, 2018, [Online]. Available: https://sites.google.com/site/ijcsis/.

[28] U. Noor, Z. Rashid, and A. Rauf, "A Survey of Automatic Deep Web Classification Techniques," *Int. J. Comput. Appl.*, vol. 19, no. 6, pp. 43–50, 2011, doi: 10.5120/2362-3099.

[29] P. Kaur, "Web Content Classification: A Survey," *Int. J. Comput. Trends Technol.*, vol. 10, no. 2, pp. 97–101, 2014, doi: 10.14445/22312803/ijctt-v10p117.

[30] M. Graczyk and K. Kinningham, "Automatic Product Categorization for Anonymous Marketplaces," *Comput. Sci.*, pp. 1–6, 2015.

[31] D. Moore and T. Rid, "Cryptopolitik and the darknet," *Survival (Lond).*, vol. 58, no. 1, pp. 7–38, 2016, doi: 10.1080/00396338.2016.1142085.

[32] M. W. Al Nabki, E. Fidalgo, E. Alegre, and I. De Paz, "Classifying illegal activities on tor network based on web textual contents," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 1, pp. 35–43, 2017, doi: 10.18653/v1/e17-1004.