METODOLOGÍA PARA LA IDENTIFICACIÓN Y CLASIFICACIÓN DE DELITOS CIBERNÉTICOS EN MÉXICO UTILIZANDO LA RED TOR

Mtro. Julio Jesús Salas Conde
Benemérita Universidad Autónoma De Puebla
Ingeniería de Lenguaje y Conocimiento
Puebla, México
jsalasconde@gmail.com

Dr. Manuel Martín Ortíz
Benemérita Universidad Autónoma De Puebla
Laboratorio Nacional de Supercómputo
Puebla, México
mmartinmx@gmail.com

***Resumen*— Hoy en día, los delitos cibernéticos son un tema de interés mundial debido a su complejidad e impacto, este protocolo expone una revisión a los delitos cibernéticos en la Dark Web que impactan en México, con el objeti-vo de desarrollar una metodología la cual descubra redes onion en Tor, extraiga su información contenida a través de palabras clave y sus sinónimos, y genere un algoritmo de clasificación que permita identificar delitos cibernéticos dentro de las redes .onion descubiertas, esto con el fin de ayudar en la prevención y el combate de los mismos en instituciones de seguridad del país y que permita al investigador obtener información relevante la cual no obtiene de la Web tradicional, impactando directamente en la sociedad que ha sufrido este tipo de delitos y requiere el estudio de nuevas tecnologías que aporten nueva información que lleve a la resolución de los casos de investigación.**

Keywords—component; formatting; style; styling; insert (key words)

I. INTRODUCTION

Las agencias gubernamentales en investigación de delitos informáticos en México necesitan de metodologías y herramientas que les permitan estudiar y comprender el fenómeno del crimen cibernético en México. Uno de los objetivos de esta disertación es analizar el contenido Web para encontrar delitos cibernéticos a partir del descubrimiento de redes onion recopilados en la red privada virtual de Tor, incluyendo sitios web .onion, foros, salas de chat, blogs, sitios de redes sociales, entre otros. Utilizando métodos de minería de datos, minería de texto y minería web, así como técnicas de minería para realizar análisis de enlaces y análisis de contenido.

En México se cuenta con una estrategia de ciberseguridad, la cual atiende delitos electrónicos en atención a denuncias de los ciudadanos a través del ministerio público o vía electrónica atendidos en las líneas “088”, el “911” o por aplicaciones electrónicas como la aplicación móvil “PF Móvil”.

Los investigadores de los delitos cibernéticos tienen las atribuciones de investigar en el ciberespacio conductas delictivas que se cometan a través de medios electrónicos como computadoras, tabletas y celulares entre otros, e interactuar con redes de comunicaciones como internet. Además reciben instrucción por parte de ministerios públicos y otras agencias de seguridad de buscar datos como nombres de personas, domicilios, placas de carros, pseudónimos, etc. en internet y otras fuentes de información.

Comúnmente estas investigaciones se llevan a cabo mediante buscadores (como Google y Bing), y redes sociales (Facebook, Twitter, Instagram, etc.), sin embargo estas redes de información aunque amplias, son limitadas en cuanto al total de información que se encuentra en todo el contenido de internet, y los delitos cibernéticos se cometen en mayor medida dentro de las llamadas redes privadas virtuales dentro de la Deep Web siendo Tor la más usada.

II. ESTADO DEL ARTE

El crecimiento sin precedentes de Internet ha dado lugar a un enfoque considerable en las técnicas de rastreo crawling/spidering en los últimos años. Los Crawlers se definen como "programas de software que atraviesan el espacio de información de la World Wide Web siguiendo los enlaces de hipertexto y recuperando los documentos web mediante el protocolo HTTP estándar" (Cheong, 1996). Son programas que pueden crear una colección local o índice de grandes volúmenes de páginas web (Cho y García-Molina 2000). Los Crawlers se pueden utilizar para los motores de búsqueda de uso general o para la construcción de la colección específica del dominio. Estos últimos se denominan rastreadores enfocados o temáticos ((Chakrabarti, 1999), (Pant, 2002), (Pant, 2002)).

Existe la necesidad de un Crawler enfocado, que pueda recopilar los foros de Dark Web. Muchos Crawlers de este tipo, se han centrado en la recopilación de páginas web estáticas en inglés desde la "superficie web". Un Crawler orientado al foro de Dark Web se enfrenta a varios desafíos de diseño. Una preocupación importante es la *accesibilidad*. Los foros web son dinámicos y a menudo requieren membresías. Son parte de la "hidden web" (Florescu et al., 1998, Raghavan y García-Molina 2001) la cual no es fácilmente accesible a través de la navegación web normal o del rastreo estándar. También hay consideraciones de minería web multilingüe. Más del 30% de la web está en idiomas no ingleses (Chen y Chau 2003). Estos foros contienen archivos de texto estáticos y dinámicos, archivos de registro y varias formas de multimedia (por ejemplo, imágenes, archivos de audio y

video). La recopilación de diversos tipos de contenido presenta muchos desafíos únicos que no se encuentran con el spidering estándar de archivos indexables (basados en texto).

Según (Cassou, 2009), el delito informático, se entiende como toda aquélla conducta ilícita susceptible de ser sancionada por el derecho penal, consistente en el uso indebido de cualquier medio informático. Agencias internacionales como la Organización para la Cooperación y el Desarrollo Económicos (OCDE), lo define como cualquier conducta, no ética o no autorizada, que involucra el procesamiento automático de datos y/o la transmisión de datos.

(Chen, 2011) presenta diez capítulos sobre enfoques computacionales y técnicas desarrolladas y validadas en la investigación de la Dark Web. Este proyecto de Dark Web de la Universidad de Arizona es un programa de investigación científica a largo plazo que tiene como objetivo estudiar y entender el fenómeno del terrorismo internacional (jihadista) a través de un enfoque computacional centrado en los datos. Su objetivo es recopilar "TODO" el contenido web generado por grupos terroristas internacionales, incluyendo sitios web, foros, salas de chat, blogs, sitios de redes sociales, videos, mundo virtual, etc. Desarrollaron minería de datos multilingües, minería de texto y web. Técnicas de minería para realizar análisis de enlaces, análisis de contenido, análisis de métricas web (sofisticación técnica), análisis de sentimientos, análisis de autoría y análisis de video. Los enfoques y métodos desarrollados en este proyecto contribuyen a avanzar en el campo de la Informática de Inteligencia y Seguridad (ISI). Estos avances ayudarán a las partes interesadas a realizar investigaciones sobre el terrorismo y a facilitar la seguridad y la paz internacionales.

(Balduzzi M.) encontró lo siguiente:

Deep Web es cualquier contenido de Internet que, por diversas razones, no puede ser o no está indexado por los motores de búsqueda como Google. Esta definición incluye páginas web dinámicas, sitios bloqueados (como aquellos en los que necesita responder a un CAPTCHA para acceder), sitios desvinculados, sitios privados (como aquellos que requieren credenciales de inicio de sesión), contenido no HTML / contextual / Acceso a las redes.

Las redes de acceso limitado cubren sitios con nombres de dominio que han sido registrados en raíces del Sistema de Dominio de Nombres (DNS) que no son administrados por la Corporación de Internet para Nombres y Números Asignados (ICANN), como dominios .BIT, sitios que se ejecutan en DNS estándar pero tienen dominios de nivel superior no estándar, y finalmente, darknets. Las Darknets son sitios alojados en la infraestructura que requiere software específico como TOR antes de que se pueda acceder. Gran parte del interés público en la Red Profunda radica en las actividades que ocurren dentro de darknets.

Una persona inteligente que compra medicamentos de drogas recreativos en línea no querrá escribir palabras clave en un navegador normal. Él / ella tendrá que ir en línea de forma anónima, utilizando una infraestructura que nunca llevará a las partes interesadas a su dirección IP o ubicación física. Los vendedores de drogas también no quieren instalarse en ubicaciones en línea donde la policía pueda determinar fácilmente, por ejemplo, quién registró ese dominio o dónde existe la dirección IP del sitio en el mundo real.

Hay muchas otras razones aparte de comprar drogas por qué la gente quisiera permanecer anónima, o para fijar los sitios que no podían ser remontados a una localización o una entidad física. La gente que quiere proteger sus comunicaciones de la vigilancia del gobierno puede requerir la cobertura de darknets. Los denunciantes pueden querer compartir una gran cantidad de información privilegiada a los periodistas, pero no quieren dejar rastro en papel. Los disidentes en regímenes restrictivos pueden necesitar el anonimato para permitir que el mundo sepa lo que está sucediendo en su país.

Pero en el otro lado de la moneda, la gente que quiere tramar un asesinato contra un objetivo de alto perfil querrá un método que se garantiza que no se puede rastrear. Otros servicios ilegales como la venta de documentos como pasaportes y tarjetas de crédito también requerirán una infraestructura que garantice el anonimato. Lo mismo podría decirse de las personas que tienen información personal de otras personas como direcciones y datos de contacto.

Cuando se habla de Deep Web, es inevitable que la frase "Clear Web" o “Surface Web” aparezca. Es exactamente lo opuesto a la Web profunda: la parte de Internet que puede ser indexada por los motores de búsqueda convencionales y accesibles a través de navegadores web estándar sin necesidad de software y configuraciones especiales.

Hay mucha confusión entre los dos. Sin embargo, la Dark Web no es la Deep Web; Es sólo una parte de la Red Profunda. The Dark Web se basa en darknets, redes en las que se realizan conexiones entre pares de confianza. Algunos ejemplos de los sistemas Dark Web incluyen TOR y el Invisible Internet Project (I2P).

III. PROPUESTA

A. *Problema de investigación*

En la actualidad es de vital importancia el estudio de las redes privadas, ya que es conocido el uso de las mismas para la comisión de delitos en los ámbitos informáticos, por ejemplo se puede citar que existen delitos cibernéticos en redes TOR, los cuáles requieren de un análisis para ser identificados y proporcionar información relevante a los investigadores.

Además es importante destacar que no solamente esto sucede en la red TOR, existen otras redes como Ares, Freenet, I2P, en la Deep Web, sin embargo esta es la más usada.

Existen delitos cibernéticos en redes privadas de internet como TOR, los cuáles requieren de un análisis para ser identificados y proporcionar información relevante a investigaciones llevadas a cabo por investigadores de los mismos, ya sea para el monitoreo de este tipo de delitos o para la búsqueda de información relevante en el esclarecimiento de un hecho.

El no investigar los delitos cibernéticos en las redes privadas de la Dark Web, es un problema que afecta a la ciudadanía directamente, esto debido a que se puede encontrar información relevante la cual no se encuentre en la Surface Web, que pueda ayudar a las instituciones de seguridad pública en el esclarecimiento de un delito y llevar a la resolución del mismo.

Además, se convierte en un problema para las instituciones de seguridad pública el no realizar investigaciones en la Dark Web, debido a que deben mantenerse a la vanguardia en los avances tecnológicos cibernéticos como se hace en países de primer mundo.

Debido a que los delincuentes cibernéticos utilizan hoy en día redes privadas como TOR para cometer estos delitos, es indispensable que las autoridades en este campo conozcan y tengan el conocimiento suficiente para prevenir y combatir estos mismos con estas tecnologías.

Por lo tanto, es necesario investigar los delitos cibernéticos en la Dark Web para mejorar el proceso de investigación en un caso y dar más y mejores datos a las autoridades correspondientes, así como prevenir estos delitos al monitorear estas conductas en estas redes privadas.

B. *Preguntas de investigación*

¿Es posible desarrollar una metodología utilizando características extraídas de los documentos *.onion* que nos permitan descubrir delitos cibernéticos en redes privadas de la Dark Web como un subconjunto de la Deep Web?

¿Existen mecanismos para descubrir redes *.onion* ?

¿Existen modelos para extraer información contenida en las redes *.onion*?.

C. *Hipótesis*

Sí se desarrolla una metodología de acceso y manipulación de datos en la red Tor basada en un diccionario de palabras claves y sus sinónimos entonces se podrán descubrir delitos que ayuden en el proceso de investigación de delitos cibernéticos en México.

Sí se investiga el contenido de redes onion dentro de Tor entonces se puede encontrar información con mayor relevancia que conlleve a la resolución de algunos casos de investigación que sin estos resultados no se llegaría a resolverlos.

Sí se realiza monitoreo en Tor entonces se puede ayudar a prevenir la comisión de un delito cibernético o reducir su impacto.

Si se utilizan técnicas de recopilación de información y análisis de los datos obtenidos entonces se puede encontrar la información más relevante de una forma más rápida que sin utilizar esta técnicas y metodologías.

Si se comparan los resultados obtenidos al realizar una investigación de un delito Cibernético utilizando como fuentes de datos el contenido de la Surface Web, en contra de utilizar como fuente de información la Dark Web, entonces se puede determinar estadísticamente el aporte y relevancia de la metodología propuesta en cuanto al tiempo de respuesta y la cantidad y utilidad de los resultados entregados afines a la investigación.

D. *Objetivo general:*

Desarrollar una metodología de acceso y manipulación de los datos que ofrece la Dark Web a partir de la red TOR para descubrir delitos cibernéticos que ayuden en el proceso de investigación.

IV. PROPUESTA DE EVALUACIÓN

El procedimiento metodológico que se propone para cumplir el objetivo general y los objetivos particulares, se describe en la Figura 7.1, en la cual se describen los siguientes pasos:

1. Descubrimiento de redes .onion mediante mecanismos como “OnionScan”.
2. Extraer información contenida enlas redes .onion descubiertas a través de palabras clave y sus sinónimos (tor-browser-selenium).

Generar un algoritmo de clasificación que permita identificar delitos cibernéticos dentro de las redes .onion descubiertas.

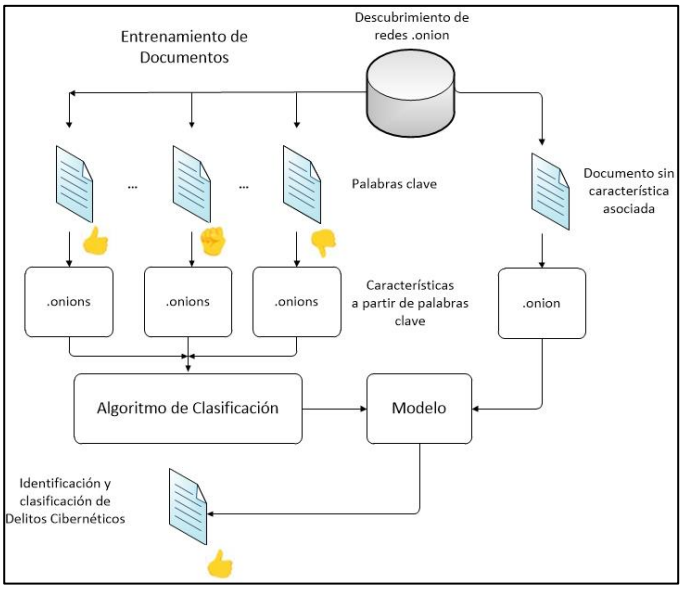

Fig. 1. Metodología para la identificación y clasificación de delitos cibernéticos utilizando la red TOR.

Se considera principalmente a la red privada de TOR como población para el análisis de información en la búsqueda de delitos cibernéticos.

Para infiltrarnos en la red Tor, se planea utilizar un software de proxy HTTP local de terceros llamado Privoxy, para conectar el Crawler a la red Tor.

Privoxy es un proxy web sin almacenamiento en caché que proporciona opciones avanzadas de filtrado para controlar los encabezados y datos HTTP. El "Crawler" opera al iniciar el proceso de rastreo en sitios web especificados por el usuario, conocidos como sitios de siembra. Recupera el contenido HTML de las páginas de sitios web de cebolla (sitios web en la Web oscura), los analiza y sigue recursivamente los enlaces salientes en las páginas. En el proceso, las estadísticas y el contenido de la página se almacenan en la base de datos.

El rastreo web avanzado de Tor presenta algunos desafíos inherentes, uno de ellos es la “Escalabilidad”. Se utilizará un rastreo distribuido, es decir, medidas tomadas para garantizar que el rastreo se distribuya en varias máquinas para aumentar el rendimiento colectivo, mediante la partición del espacio URL, de modo que cada máquina o nodo rastreador es responsable de un subconjunto de las URL en la web.

Para la búsqueda de palabras clave se utilizará Sphinx en la aplicación web. Sphinx es un motor de búsqueda independiente de texto completo que proporciona búsquedas de palabras clave más rápidas y eficientes (texto completo) mediante la creación de índices basados en el contenido de la base de datos de texto.

APORTACIONES ESPERADAS

Se espera aportar en el diseño de una herramienta que permita descubrir redes onion y sean almacenadas en una base de datos para su registro y posterior uso.

La implementación de un crawler que permita descargar contenido de redes en Tor para su posterior análsis.

Aplicar una algorítmo de clasificación que permita identificar delitos cibernéticos a partir de palabras claves dentro de las redes .onion descubiertas.

En conjunto la metodología desarrollada en este proyecto debe contribuir a avanzar en el campo de la seguridad cibernética en México. Estos avances ayudaran a las agencias gubernamentales a realizar investigaciones en delitos cibernéticos y a prevenir y combatir el crimen cibernético en México.

Se espera que esta investigación ayude a la próxima generación de analistas, agentes de inteligencia, de justicia y expertos en Internet.

Proporcionar una visión general de la Dark Web, sugerir un enfoque sistemático y computacional para entender la problemática y mostrar el progreso de la investigación a través de técnicas, métodos y casos de estudio.

BIBLIOGRAFÍA

[1] Balduzzi M., C. V. Cybercrime in the Deep Web. Black Hat EU, Amsterdam 2015. Trend Micro.

[2] Cassou, R. J. (2009). Delitos informáticos en México. Revista Número 28 del Instituto de la Judicatura Federal. http://www.ijf.cjf.gob.mx/publicaciones/revista/28/Delitos_inform%C3%A1ticos.pdf.

[3] Chen, H. (2011). Dark web: Exploring and data mining the dark side of the web (Vol. 30). Springer Science & Business Media.

[4] Cheong, F. C. (1996). *Internet Agents: Spiders, Wanderers, Brokers, and Bots.* Indianapolis, IN: New Riders Publishing.

[5] Chakrabarti, S., Van Den Berg, M., and Dom, B. (1999). *Focused Crawling: A New Approach to Topic-Specifi c Resource Discovery. In Proceedings of the Eighth World Wide Web Conference.*

[6] Pant, G. S. (2002). Exploration versus Exploitation in Topic Driven Crawlers. *In Proceedings of the WWW Workshop on Web Dynamics.*

[7] Florescu, D. L. (1998). Database Techniques for the World-Wide Web: A Survey. (2. 5.–7. SIGMOD Record, Ed.)