

**Propuesta de Proyecto de Investigación
Maestría en Ciencias y Tecnologías de la Información**

18 de noviembre de 2022

1. Nombre del proyecto

Identificación de usuarios en Internet utilizando técnicas de aprendizaje maquina

2. Responsables

Dr. Miguel López Guerrero
UAM unidad Iztapalapa
Edif. T ofna. 302
milo@xanum.uam.mx

Dr. Óscar Arana Hernández
UAM unidad Iztapalapa
Posdoctorante del PCyTI
aranahoscar@gmail.com

3. Área(s) de conocimiento relacionada(s) con el proyecto

- redes de computadoras,
- algoritmos de aprendizaje maquina,
- ciencias de la computación, y
- simulación

4. Descripción del proyecto

● Contexto

En la actualidad es innegable que la Internet paulatinamente ha ido ocupando un lugar preponderante en las actividades que realizamos día con día. Una de las razones de su popularización radica en la facilidad con la que los usuarios podemos acceder a la información y a los servicios que ésta ofrece. Efectivamente, a través de la sencilla interfaz gráfica que ofrecen los navegadores, el proceso para acceder a los múltiples recursos disponibles por Internet consiste simplemente en escribir la dirección URL deseada en la barra de navegación y seguir unos cuantos hipervínculos hasta llegar al lugar deseado. Sin embargo, las direcciones que el usuario introduce no son las mismas que requieren los protocolos de encaminamiento para identificar una computadora en Internet y comunicarse con ella. Por esta razón, desde los primeros años de operación de Internet, se decidió utilizar un servicio de conversión de direcciones URL a direcciones numéricas. Dicha traducción de nombres de dominio es un servicio que se proporciona a través de servidores DNS (*Domain Name System*). Por lo tanto, la utilización del servicio DNS es una característica inherente de la navegación por Internet.

● Motivación

La cantidad de actividades que los usuarios podemos realizar en Internet se ha incrementado, en particular, durante los últimos años. Muchas actividades que anteriormente se hacían de manera presencial, se han trasladado a una interacción virtual apoyada por las TIC. Si bien esto supone una simplificación en muchas de nuestras actividades, también supone manifestar muchos de nuestros intereses y gustos como hábitos de uso de Internet. Por tanto, ésta se ha convertido en un lugar en el que todos los usuarios dejamos rastros característicos de nuestras actividades. En particular, como se mencionó previamente, cada vez que un usuario desea visitar algún sitio Web, deberá escribir la dirección deseada en su navegador, lo que desencadenará el envío automático al servidor DNS de una solicitud de conversión de una dirección URL a una dirección numérica. Evidentemente, a través del acceso –usualmente no autorizado– y análisis de la serie de consultas generadas por un usuario durante una sesión de Internet, se podría deducir fácilmente una gran cantidad de información acerca de él. Por ejemplo, qué pasatiempo tiene una persona en particular, en dónde realiza sus compras y qué compra. A este respecto, existen varios trabajos de investigación que han demostrado la viabilidad de identificar a los usuarios a través del análisis de sus patrones de uso de la Internet. Para ello se puede

analizar el historial de consultas DNS utilizando técnicas de aprendizaje maquina. Aunque existen esfuerzos incipientes en este sentido, éste es aún un campo con bastantes preguntas abiertas respecto a idoneidad de las diferentes técnicas de aprendizaje maquina. Así, esta propuesta de proyecto de investigación de maestría se enfoca en el estudio, implementación y evaluación de diversos algoritmos de aprendizaje maquina aplicados al problema de la identificación de los usuarios a través del análisis de su tráfico DNS en Internet.

- Aporte esperado al área de conocimiento

Se generará una evaluación de desempeño de varios algoritmos de aprendizaje maquina aplicados al problema de la identificación de los usuarios en Internet.

5. Objetivos

- Objetivo general

Generar una evaluación de desempeño de varios algoritmos de aprendizaje maquina aplicados a la identificación de los usuarios en Internet.

- Objetivos particulares

a) Familiarizarse con los trabajos de investigación existentes para identificar usuarios de Internet usando diferentes algoritmos de aprendizaje maquina.

b) Generar una evaluación de desempeño de una selección de algoritmos de aprendizaje maquina en algún software computacional.

c) Comunicar adecuadamente los resultados de la investigación.

6. Metodología

1.- Estudiar el funcionamiento del protocolo de traducción de nombres de dominio (DNS)

2.- Estudiar la forma en la que operan los algoritmos de aprendizaje maquina aplicados al problema de la identificación de los usuarios de Internet.

3.- Seleccionar un conjunto de algoritmos de aprendizaje maquina que, de acuerdo con sus características intrínsecas, puedan ser aplicados a la identificación de los usuarios.

4.- Implementar los algoritmos seleccionados en algún software.

5.- Evaluar los algoritmos seleccionados.

6.- Escribir un documento que compile los resultados de la investigación y remitirlo para arbitraje y posible publicación a un foro especializado.

7. Integrar los resultados de la investigación en la Idónea Comunicación de Resultados y defenderla en examen de grado.

7. Calendarización de actividades

Trimestre	Curso	Actividades
23-I	Proyecto de Investigación I	<ul style="list-style-type: none">● Investigación documental acerca de la problemática de la identificación de usuarios en Internet (medidas y contramedidas)● Revisión de diferentes algoritmos de aprendizaje maquina● Escritura de avances● Aprendizaje de una herramienta de pruebas● Presentación de avances

23-P	Proyecto de Investigación II	<ul style="list-style-type: none"> • Pruebas de desempeño • Depuración de programas y resultados • Actualización de la investigación documental • Escritura de avances • Presentación de avances en forma oral y escrita (en formato de artículo)
23-O	Proyecto de Investigación III	<ul style="list-style-type: none"> • Pruebas de desempeño complementarias • Actualización de la investigación documental • Escritura de un reporte de investigación y su envío a un foro especializado con arbitraje estricto • Integración de la primera versión de la Idónea Comunicación de Resultados (ICR) • Presentación de avances
24-I	Inscripción en blanco	<ul style="list-style-type: none"> • Depuración y finalización del escrito de la ICR • Preparación y presentación del examen de grado

8. Infraestructura necesaria y disponible

Para este proyecto se cuenta con una computadora de escritorio en las instalaciones de la universidad. Sin embargo, se prefiere que el alumno utilice su propia computadora personal.

9. Lugar de realización

El lugar de realización puede ser en la universidad, si las condiciones lo permiten, o en el domicilio del alumno. En cualquiera de los casos, el equipo de trabajo tendrá una reunión presencial o virtual cada semana.

10. Entregables

- Una revisión documental del tema
- Un reporte de avances en formato de artículo
- La Idónea Comunicación de Resultados en formato de tesis

11. Referencias bibliográficas básicas

- Chang, D., Chen, J. Q., Li, Z., & Li, X. (2022, junio). Hide and Seek: Revisiting DNS-based User Tracking. *In 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)* (pp. 188-205). IEEE.
- Sun, M., Xu, G., Zhang, J., & Kim, D. W. (2017, noviembre). Tracking you through DNS traffic: Linking user sessions by clustering with Dirichlet mixture model. *In Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems* (pp. 303-310).
- Kirchler, M., Herrmann, D., Lindemann, J., & Kloft, M. (2016, October). Tracked without a trace: linking sessions of users by unsupervised learning of patterns in their DNS traffic. *In Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security* (pp. 23-34).
- Herrmann, D., Kirchler, M., Lindemann, J., & Kloft, M. (2016, December). Behavior-based tracking of Internet users with semi-supervised learning. *In 2016 14th Annual Conference on Privacy, Security and Trust (PST)* (pp. 596-599). IEEE.
- Herrmann, D., Banse, C., & Federrath, H. (2013). Behavior-based tracking: Exploiting characteristic patterns in DNS traffic. *Computers & Security*, vol. 39, (pp. 17-33).