UNIVERSIDAD DE LA CIUDAD DE AGUASCALIENTES

MAESTRÍA EN CIENCIA DE DATOS

SEMINARIO DE TESIS

“Detección de anomalías mediante aprendizaje automático en tráfico de servidores web”

Alumno:

E23S-18014: MITSIU ALEJANDRO CARREÑO SARABIA

Periodo Agosto 2024 - Diciembre 2024, Aguascalientes, Ags

Variable independiente:

- remote_addr: Dirección IPv4 del cliente que inició la conexión.
- remote_usr: Nombre de usuario (solo aplica si el usuario está autenticado).
- date: Fecha en formato “dd/MM/YYYY”.
- time: Hora en formato “HH:MM:SS”.
- request: Descripción del recurso solicitado por el cliente.
- req_method: Método HTTP mediante el cual se solicitó el recurso.
- req_uri: Dirección URI del recurso solicitado.
- http_ver: Versión HTTP bajo la que se estableció la conexión cliente-servidor.
- status: Código de estatus HTTP que resolvió el servidor.
- body_bytes_sent: Cantidad de bytes enviados en la respuesta.
- http_referer: Valor de la cabecera http_referer con el que se conectó el cliente.
- user_agent: Valor de la cabecera user_agent con el que se conectó el cliente.
- dec_req_uri: Es una réplica decodificada del valor req_uri.
- clean_path: Filtrado del valor req_uri únicamente con el recurso solicitado (sin parámetros).
- clean_query_list: Listado de python de los parámetros query.
- domain: Nombre del dominio listado en el campo http_referer.
- fabstime: Valor time transformada en decimal.
- weekday: Valor numérico correspondiente al día de la semana donde Lunes (1) y Domingo (7).

Variable dependiente: Grado de anormalidad en la actividad

Capítulo 1

Consideraciones metodológicas

1 planteamiento del problema

1.1 Pregunta central de investigación

¿Estimar el grado de anomalía de tráfico de servidores web mediante técnicas heurísticas y de aprendizaje automático es una técnica confiable para

1.2 Preguntas secundarias de investigación

- ¿De qué manera se analiza el tráfico de servidores web actualmente?
- ¿Qué elementos debe tener un sistema de detección de anomalías para ser útil (falsos negativos/falsos positivos, canales de comunicación, protocolos de contingencia y respuesta)?
- ¿Actualmente cómo se ha implementado el aprendizaje automático en análisis de tráfico de servidores web?

1.3 Objetivos de investigación

Se pretende explorar la implementación de técnicas heuristicas así como de aprendizaje automático para determinar si la actividad y tráfico de un servidor web es anómala, permitiendo:

a) Analizar grandes cantidad de datos de manera automática
b) Permitir la constante actualización de patrones, ajustando el comportamiento normal y detectando nuevos vectores de ataque.

1.4 Antecedentes

Con la expansión del acceso a servicios de internet, así como la creciente disponibilidad de dispositivos de distintas categorías para conectarse a la red, la demanda y tráfico de servicios web se encuentra en constante aumento. Mucho se ha desarrollado en términos de escalabilidad de infraestructura así como adopción de soluciones distribuidas para dar servicio a la ascendente demanda desde la producción en masa de dispositivos celulares y móviles, hasta la progresiva adopción del internet de las cosas, pero derivado de dicha disponibilidad, se genera una cantidad inmensa de tráfico que cualquier servidor web disponible desde internet debe dar seguimiento, procesar, contestar.

El origen de dicho tráfico puede ser generado por peticiones de usuarios reales, peticiones de bots, peticiones automatizadas y peticiones de usuarios con intenciones maliciosas cada uno de estos grupos de clientes se comportan de maneras muy específicas, por lo que se considera una área de oportunidad el analizar estos datos para reconocer y clasificar los usuarios que interactúan con los servicios ofrecidos.

El tráfico de servidores web tiene claras tendencias como recursos solicitados, región geográfica de donde se solicita, hora en que se solicitó, cantidad de bytes enviados, por lo que identificar las tendencias y detectar las anomalías es un trabajo que puede ser automatizado y al que se le pueden aplicar distintas técnicas de aprendizaje automático que permitan analizar la información desde múltiples dimensiones y perspectivas.

Como consecuencia del acceso generalizado a servicios en internet, es común confiar en las protecciones que da el proveedor del servicio y permitirle alojar datos personales y sensibles en sus servidores, lo que aumenta la relevancia de evaluar qué y cómo se están accediendo a los recursos solicitados, así como desarrollar herramientas que faciliten filtrar las anomalías para tomar acciones correctivas.

1.4.1 Históricos

1.4.2 Contexto actual

1.5 Justificación

El tráfico a un servidor web provee datos confiables sobre la información y el contexto bajo el que se usan sus recursos, pero la cantidad de información generada es tan grande que un análisis manual no es viable. Entender los usos típicos y diferenciarlos de los

atípicos es una herramienta poderosa que aplicada en tiempo real permitirá mejorar la calidad, y resguardo de la información contenida.
Analizar los registros de tráfico web permite no solo entender la manera en que se consume la información que contiene un servidor, sino también detectar si el uso generalizado se transforma, o si existen anomalías e incluso calcular un parámetro de probabilidad de ser malintencionadas. Dado el volumen de información que se genera, y la creciente sensibilidad de los datos alojados, aplicar herramientas de aprendizaje automático permitirá agilizar y perfeccionar cualquier proceso manual.

1.6 Hipótesis. - es una aproximación a la verdad

1.7 Marco teórico

1.7.1 Marco teórico referencial. – teorías hipótesis trabajos de investigación (enumerar)

a) Tesis
b) Teorías
c) Opiniones

1.7.2 Marco jurídico referencial

1.8 Delimitaciones

a) Tiempo
b) Tipo
c)

Capítulo 2

Marco teórico

Leyes

Teoría

Capítulo 3

Desarrollo de la investigación

En la práctica cómo se comporta la variable dependiente

En la práctica cómo se comporta la variable independiente

Ventajas desventajas, cuadros comparativos

Capítulo 4

Resultados de la investigación

4.1 Resultados de la investigación

4.2 Conclusiones

4.3 Recomendaciones

4.4 Bibliografía