AYER INFORMACIÓN, HOY BIG DATA

Rocha Arcos Oscar Alberto
alberto.oscar96@outlook.com
De la Cruz Sosa Carlos
cdelacruzupiita@gmail.com
Santiago Godoy Rafael

Instituto Politécnico Nacional UPIITA

Resumen

Estamos en la era de la información (data), donde ésta lo es todo. Esta en los libros, artículos, periódicos, revistas, noticias, en las notas de refrigeradores, en el celular, y hasta en nuestras identificaciones. Sin embargo, la cotidianidad con la que usamos la información nos ha hecho ser poco conscientes del valor que ésta tiene, así como de otros usos alternativos [1].

Este enfoque está cambiando poco a poco, ya que más empresas y países apuestan a la utilización de esta información y sus diversos usos, específicamente en la inteligencia artificial (IA, por sus siglas en inglés). Algunos de estos casos son: Facebook con la recolección de datos de los usuarios [2] y China con un plan estratégico nacional para analizar la información a través de la IA [3].

Entonces, para el manejo y análisis de la información, primero hay que recolectarla, y de eso se encarga el Big Data.

1. ¿Qué es el Big Data?

De acuerdo con la empresa Oracle, Big Data es un conjunto de datos más grandes y más complejos que lo habitual, los cuales llegan en mayor variedad, en volúmenes crecientes y con una velocidad cada vez mayor. A esto se le conoce como las tres V's.

Si bien el Big Data se trata de recolectar información, cabe señalar que arriba del 80 % de esta información es caótica y desestructurada, por lo que la cantidad útil, después de un proceso de clasificación se vería drásticamente reducida [4].

2. Las tres "V"

2.1. Volúmen

La cantidad da datos importa; por lo que en el Big Data se deberá procesar grandes volúmenes de datos no estructurados y sin secuencia.

2.2. Variedad

La variedad se refiere a los muchos tipos de datos disponibles, los cuales se tendrán que estructurar y adaptar a una base de datos. Cada tipo de dato (audio, video, texto, etc.) requiere un procesamiento diferente; es por eso que las bases de datos se deben de apoyar con programas especializados en el tratamiento y clasificación de los datos con los que se va a trabajar.

2.3. Velocidad

Idealmente se debe de tener un sistema que actué (procese, clasifique, etc.) a la velocidad en la que se reciben los datos. Sin embargo, esto es opcional y estará delimitado por los requerimientos del sistema, si es que necesita actuar en tiempo real.

3. Nueva V

Recientemente se ha añadido una nueva característica la cual es la veracidad de la información, esto con razón de que anteriormente se recolectaba información basura, que no ayudaba al análisis o que esta información no era real.

Con la inclusión de esta nueva característica, se tiene en cuenta la fiabilidad de los datos y a su vez se obtienen beneficios en procesamiento y ahorro de memoria de almacenamiento de la información.

4. Herramientas para el Big Data

4.1. Web Crawling

Conocido en el español como "Araña Web", es una de las técnicas empleada para rastrear información a través de un sitio web. Haciendo uso de una lista de direcciones URL's iniciales conocidas como "semillas" para después a través de ellas para encontrar nuevas URL's y así sucesivamente hasta obtener una lista de todas las URL's obtenidas en el proceso.

Este proceso se utiliza para hacer un compendio de todas las fuentes de información disponible. Una analogía sería como ir a la biblioteca y obtener todos los libros de Biología disponibles.

4.2. Web Scrapping

Tomando en cuenta la analogía anterior, ahora se requiere obtener información acerca de un tema específico de Biología, por ejemplo, división celular, por lo que ahora se extrae información de este tema de todos los libros disponibles, esta extracción de información puede realizarse mediante la búsqueda de palabras clave del tema deseado y/o frases relacionadas, en este caso, sería utilizando la frase "división celular" y conceptos que hagan relación a esta frase.

Este proceso aplicado a la Web podría ser mediante la búsqueda del contenido a través del formato HTML (del que esta conformado una página web) y relacionando los atributos de las etiquetas HTML con las palabras clave de la búsqueda.

4.3. Inteligencia Artificial

Se tiene que aprender a lidiar con toda esta inundación de información para no ahogarnos en ella; debido a esto y a posibles requerimientos de seguridad e imparcialidad (en uso bancario, por ejemplo) es más viable que nuevas herramientas sean las encargadas de manejar la información recabada.

La Inteligencia Artificial es una de las herramientas que tiene más proyección al futuro. Ésta basa su funcionamiento usando redes neuronales buscando imitar el comportamiento del cerebro humano, aprendiendo patrones y relacionando nuevas entradas a estos patrones. Entre más muestras tenga el sistema, mejor podrá aprender, y menos se equivocará al tratar de reconocer una palabra. Es muy parecido a cuando se aprende un nuevo idioma, en el cual, entre más tiempo se practique, mejor se dominará. He aquí la importancia del Big Data, entre más información se tenga el sistema aprende mejor.

5. Conclusiones

Se concluye que el Big Data siga creciendo debido al impulso del Internet de las Cosas (IoT, por sus siglas en inglés) y a los dispositivos capaces de recolectar información. Aunado a este crecimiento también vendrán nuevas técnicas y/o mejoras para el análisis de la información al igual que nuevos retos.

Las aplicaciones mencionadas solo son algunas de las tantas que usan el Big Data, el día de mañana no sabemos qué cosas nuevas se podrán crear.

Referencias

1. L. G. Salas (**2018**) *"La transformación de la comunicación en el siglo XXI", Forbes México.* [En linea]. Disponible: https://www.forbes.com.mx/comunicacion-siglo-xxi/. [Ultimo acceso: 20- Mayo-2018].

2. S. Frier (**2018**) *"¿Por qué Facebook recolecta datos hasta de quienes no son sus usuarios?.El Financiero* [En linea]. Disponible: http://www.elfinanciero.com.mx/tech/por-que-facebook-recolecta-datos-hasta-de-quienes-no-son-sus-usuarios. [Ultimo acceso: 20- Mayo- 2018].

3. "M. Díaz (**año**) *"La ambición china por controlar la IA es mayor de lo que creíamos"MIT Technology Review* [En linea]. Disponible: https://www.technologyreview.es/s/10080/la-ambicion-china-por-controlar-la-ia-es-mayor-de-lo-que-creiamos. [Ultimo acceso: 20- Mayo- 2018].

4. Autor (**2018**) *"What is Big Data? /libro/Oracle/Oracle.com* [En linea]. Disponible: https://www.oracle.com/big-data/guide/what-is-big-data.html. [Ultimo acceso: 22- Mayo- 2018].

5. J. Titcomb (**2018**) *"Facebook admits up to 270m users are fake and duplicate accounts/libro/The Telegraph/web* [En linea]. Disponible: https://www.telegraph.co.uk/technology/2017/11/02/facebook-admits-270m-users-fake-duplicate-accounts/. [Ultimo acceso: 20- Mayo- 2018].