Hans Peter Luhn y el nacimiento del Hashing

El método de hashing del ingeniero de IBM dio a las computadoras una forma de buscar rápidamente documentos, ADN y bases de datos

Por Hallam Stevens, profesor de historia en la Universidad Tecnológica de Nanyang, Singapur
Traducción: Google Translator + correcciones del profesor

IEEE Spectrum
30 de enero de 2018 | 16:00 GMT

Esto es solo un extracto (artículo completo en inglés)

Los años posteriores a la Segunda Guerra Mundial fueron formativos para los expertos en computadoras electrónicas. Varios tipos de computadoras construidas durante la guerra habían servido para realizar cálculos vitales para balística, armas atómicas y criptografía. Las tensiones de la Guerra Fría garantizaron una continua financiación del desarrollo de computadoras; como resultado, se construyeron máquinas cada vez más precisas y potentes. Pero sus usos principales -cálculo numérico intensivo y almacenamiento- apenas cambiaron.

En 1946 y 1947 Hans Peter Luhn trabajaba en un dispositivo para leer automáticamente documentos mecanografiados. El dispositivo consistía en una cinta metálica insertada en una máquina de escribir, que perforaba patrones magnéticos en papel que luego podían escanearse. Poco después, comenzó a trabajar con dos químicos del MIT, Malcolm Dyson y James Perry, en una máquina que podía buscar automáticamente compuestos químicos utilizando tarjetas perforadas. Cada tarjeta perforada era codificada con información sobre un compuesto en particular. El usuario insertaba una "tarjeta de preguntas" en la máquina, que a continuación enumeraba un conjunto de criterios con los que se podían comparar y clasificar todas las tarjetas. Aunque su escáner era altamente especializado, Luhn siguió buscando formas más generales de procesar información automáticamente.

Los años de la posguerra vieron una explosión en el número de artículos publicados en ciencia e ingeniería. A muchos expertos les preocupaba que la "sobrecarga de información" abrumara a investigadores y empresarios por igual. Con el objetivo de organizar e indexar ese aluvión de textos, Vannevar Bush, líder de la burocracia científica generada durante la guerra en Estados Unidos y uno de los arquitectos de la National Science Foundation, propuso un dispositivo electromecánico de tamaño de escritorio, el Memex, para almacenar y vincular información.

Sin embargo, la idea de Bush nunca llegó a realizarse. En cambio, las ideas de Luhn tuvieron más éxito. El 6 de enero de 1954, por ejemplo, presentó una patente titulada “Computer for Verifying Numbers” [PDF]. Este dispositivo mecánico portátil pretendía resolver un problema práctico muy sencillo. En ese momento, varios tipos de números de identificación, como números de tarjetas de crédito y números de Seguridad Social, estaban comenzando a desempeñar un papel importante en la vida pública y privada. Pero los números eran difíciles de recordar y podían ser transcritos incorrectamente o falsificados deliberadamente. Lo que se necesitaba era un medio para verificar rápidamente si un número de identificación era válido.

La computadora de mano de Luhn aplicaba un algoritmo de comprobación de suma (checksum) desarrollado por él mismo. Para un número de 10 dígitos, la computadora realizaría los siguientes pasos:

- Multiplicar por 2 cada segundo dígito
- Si el resultado es 10 o más, sumar los dígitos de ese resultado para obtener un número de un solo dígito (por ejemplo, 16 se convertiría en 1 + 6 = 7)
- Sumar los 10 dígitos del nuevo número
- Multiplicar por 9
- Tomar el último dígito de ese resultado

Esta receta producía un número de "verificación" de un solo dígito. En la formulación original de Luhn, un 0 indicaba que el número original era válido. En versiones posteriores, el dígito de verificación simplemente se agregaba al número original como un dígito final, por lo que se podía verificar fácilmente que el dígito final coincidiera con el número de verificación producido por la máquina. La secuencia subyacente de cálculos, ahora conocida como algoritmo de módulo 10, se usa todavía en muchas aplicaciones. Por ejemplo, los números internacionales de identidad de equipos móviles (IMEI) se verifican de esta manera.

Pero la cosa no acaba ahí: los engranajes y las ruedas de la máquina de Luhn se convirtieron en la base de uno de los algoritmos más importantes de la era digital: el hashing. Se trata, en realidad, de una clase de algoritmos que permite organizar la información de modo que sea fácil de encontrar para una computadora. Al igual que un picadillo (hash) de carne en conserva y patatas, un algoritmo de hashing corta y mezcla datos de varias maneras. Tal mezcla, cuando se despliega inteligentemente, puede acelerar muchos tipos de operaciones de la computadora.

A principios de 1953, Luhn había escrito un memorandum interno para IBM en el que sugería poner la información en “casilleros” para acelerar la búsqueda. Supongamos que quisiéramos buscar un número de teléfono en una base de datos y descubrir a quién pertenece (nombre, dirección, etc.). Dado el número de 10 dígitos 314-159-2652, una computadora podría simplemente ir buscando ese número en una lista hasta encontrar la entrada correspondiente. Sin embargo, en una base de datos de millones de números, esto retrasaría enormemente la operación.

La idea de Luhn era asignar cada número de teléfono a un casillero numerado, de la siguiente manera: los dígitos del número de teléfono se agruparían en pares (en este caso, 31, 41, 59, 26, 52). A continuación, los dígitos emparejados se sumarían (produciendo como resultado, en el ejemplo anterior, 4, 5, 14, 8 y 7). A partir de estos dígitos (en el caso de un resultado de dos dígitos, sólo se utilizaría el último dígito) se generaría un nuevo número (en este caso, 45487). Por último, el número de teléfono original (así como el nombre y la dirección del usuario) se colocarían en el casillero con la etiqueta 45487.

Buscar la entrada correspondiente a un número de teléfono implica calcular rápidamente el número del casillero utilizando el método de Luhn, y luego recuperar la información allí contenida. Incluso en el caso de que el casillero en cuestión contuviera múltiples entradas agrupadas en una lista, la búsqueda secuencial del número de teléfono en dicha lista sería mucho más rápida que la búsqueda del número en la lista completa.

Durante décadas, los informáticos y programadores han mejorado los métodos de Luhn y les han dado nuevos usos. Pero la idea básica sigue siendo la misma: utilizar un problema matemático para organizar los datos en casilleros fácilmente explorables. Debido a que la organización y la búsqueda de datos son problemas comunes en la informática, los algoritmos de hashing se han vuelto cruciales para la criptografía, los gráficos, las telecomunicaciones y la biología. Cada vez que envíamos un número de tarjeta de crédito a través de internet o utilizamos el diccionario de un procesador de textos, las funciones hash entran en funcionamiento.

Hoy en día, las técnicas de hashing desempeñan una gran cantidad de funciones en la gestión y protección de nuestras vidas digitales. Cuando introducimos nuestra contraseña en un sitio web, es probable que el servidor almacene una versión hash de la misma. Cuando interactuamos con un sitio web utilizando una conexión segura o cuando compramos algo con bitcoins, los métodos hash también están funcionando. Servicios en la nube como Dropbox y Google Drive se sirven de los algoritmos hash para almacenar y compartir archivos de manera eficiente. En genética y otras áreas de investigación que requieren el uso intensivo de datos, los métodos de hashing reducen drásticamente el tiempo necesario para examinar computacionalmente grandes cantidades de datos.