Una mirada desde las forjas: la construcción de datos, metadatos y anotaciones para entrenar algoritmos de aprendizaje automático. El renovado potencial de los metadatos

Soria, Marcelo A.

soria@agro.uba.ar

Universidad de Buenos Aires. Facultad de Agronomía. Departamento de Biología Aplicada y Alimentos y Facultad de Ciencias Exactas y Naturales, Maestría en Explotación de Datos. Buenos Aires. Argentina.

Línea temática 4. Metadatos. Datos sobre datos
(Buscar y que nos busquen a través de nuestras palabras)

Palabras clave

Privacidad, Inteligencia Artificial, Aprendizaje Automático, Re-identificación

Resumen

La complementación de datos, metadatos y anotaciones es un paso fundamental para entrenar sistemas de aprendizaje automático. En esta presentación se revisa ese proceso con énfasis en los metadatos. El uso de metadatos en sistemas de gestión de la información es de vieja data. Históricamente se consideraba a los metadatos como los hermanos menores e inofensivos de los datos. Eso ya no es así. El enorme volumen de metadatos que se crean constantemente y el desarrollo de metodologías de explotación, evidenció una nueva dimensión de su valor y el potencial para revelar información sensible de las personas. En cualquier aplicación actual de grandes datos es necesario evaluar y balancear los riesgos y beneficios del uso de metadatos.

¿Datos o metadatos? Una mirada multidimensional

Ya no nos extraña hablar y escuchar sobre datos. Y no es una moda. Generamos y consumimos cantidades crecientes de datos y muchas aplicaciones de uso diario requieren datos para “aprender” sus tareas, como los buscadores web, los traductores automáticos, los sistemas de reconocimiento de imágenes, los recomendadores de música o de videos, entre otros.

Hay tres factores principales que explican la actual ubicuidad de los datos. El primero es que es cada vez es más barato y fácil producirlos. Por ejemplo, la mayoría de los teléfonos incluyen un sensor GPS, algo que hace unos 15 años atrás requería un dispositivo específico. El segundo factor es que se puede tratar como datos casi cualquier elemento de información: textos, imágenes, video, sonidos, movimiento. Esto permitió la creación de un sinfín de aplicaciones. Finalmente, el tercer factor es que la mayoría de los productos más exitosos y destacados de inteligencia artificial pertenecen a un campo de esta disciplina llamado Aprendizaje Automático, que en esencia consiste en aprender a partir de cantidades enormes de ejemplos, es decir, datos.

Otro elemento importante en el tratamiento y explotación de datos son las anotaciones. Para avanzar en la distinción entre datos, metadatos y anotaciones consideremos una imagen que se va a usar para entrenar un sistema de conducción autónoma de un auto. Supongamos que esta imagen es una foto de un cruce de calles con algunos autos y un semáforo. El sistema de visión artificial que guiará al auto necesita la imagen original para aprender a reconocer el ambiente donde se moverá. Además, para avanzar en el entrenamiento necesita saber dónde están los autos en esa imagen. Para esto se delimitan las coordenadas de los rectángulos que contienen a cada auto, y cada uno de estos rectángulos es una anotación. La imagen del ejemplo puede tener asociadas otras anotaciones para otros objetos, como el semáforo o peatones esperando a cruzar. Finalmente, la imagen tendrá metadatos asociados, tales como la fecha y hora en que se capturó la imagen, la geolocalización, el modelo de cámara usada, parámetros fotográficos de la imagen: valor de apertura, ISO equivalente, resolución, etc.

Un caso de metadatos más cercano son las fotos que se toman con un celular. A veces no se puede acceder fácilmente a los metadatos de las fotos desde el celular, pero al descargarlas a una computadora es fácil acceder a ellos y observar propiedades de la imagen tales como apertura, ISO equivalente, etc.

Al entrenar sistemas de aprendizaje automático no siempre se establece una separación categórica entre dato, metadato, anotación, sino que se tiende a considerar al conjunto como un dato complejo y multidimensional.

Además de su utilización en sistemas de aprendizaje automático, los metadatos también tienen un papel destacado en los grandes sistemas de

administración de información de grandes instituciones, públicas y privadas, donde son parte fundamental de los procesos de control (Uttamchandani, 2021). Por ejemplo, supongamos el caso de un banco que cuenta con un grupo de empleados que está autorizado a hacer consultas sobre saldos de clientes en la base de datos de la empresa. En este banco se descubre un fraude que afecta a varios clientes y para determinar qué sucedió es esencial saber quién o quiénes y cuándo hicieron consultas sobre los saldos de los clientes afectados y si se realizaron cambios en lsus datos financieros. En este caso, el dato es la consulta y los metadatos son los registros de quienes y cuando hicieron consultas.

Otro ejemplo sobre diferencias entre datos y metadatos, y que es relevante para la próxima sección, es el de las llamadas telefónicas. El contenido de audio de una llamada telefónica es el dato, que requiere herramientas sofisticadas o un laborioso trabajo humano para llevar a cabo la desgrabación. Además, y más importante, en los países democráticos se requiere una autorización judicial y cumplir con una serie de salvaguardas para interceptar, grabar y desgrabar una llamada. Los metadatos típicos de una llamada son: número del originador, número del receptor, duración de la llamada, locación del originador y del receptor (torres de celular cercanas en el caso de una llamada celular). Estos datos se recolectaron desde el inicio de la telefonía para propósitos de facturación, son fáciles de recolectar y analizar, consumen poca memoria y tradicionalmente no tuvieron el grado de protección legal que tiene el contenido de la llamada porque se consideraba que, en comparación a la llamada en sí, contienen poca información de interés y que el acceso a los metadatos no constituía una amenaza de la privacidad, o al menos, una amenaza grave.

Vigilancia con metadatos: las revelaciones de Edward Snowden

En junio de 2013 Edward Snowden, un ex-analista de inteligencia de la NSA (National Security Agency, una agencia de inteligencia electrónica de Estados Unidos) inició la publicación de documentos secretos de dicha agencia que revelaban el funcionamiento de una serie de programas de recolección de información, algunos de ellos declarados ilegales con posterioridad (Szoldra, 2016). A partir de estos documentos filtrados se pudo obtener información más detallada sobre MAINWAY, un programa de recolección de metadatos derivados de llamadas telefónicas, cuya existencia se había filtrado a la prensa unos años antes.

Una de las características del sistema MAINWAY es el desarrollo de un grafo para registrar llamadas. Los grafos son estructuras matemáticas muy útiles para representar interacciones en redes sociales o llamadas telefónicas. En su forma más básica los grafos son conjuntos de “nodos” y de “vértices”. En el caso de las llamadas telefónicas, el conjunto de nodos representa a las

personas que pueden hacer llamadas telefónicas y los vértices representan la ocurrencia efectiva de llamadas y conectan a los dos nodos involucrados. Para visualizar un grafo se suelen utilizar círculos para representar nodos y líneas para los vértices formando una red más o menos densa de conexiones. El uso de grafos facilita la detección de quien se comunica con quien, descubrir comunidades de personas que se comunican frecuentemente entre sí o personas que son centrales para conectar diferentes comunidades, entre otros tipos de análisis.

En el caso de MAINWAY los nodos abarcaban casi la totalidad de personas con líneas telefónicas residentes en los Estados Unidos y aquellas personas de fuera de los Estados Unidos que mantuvieron comunicaciones con residentes. Se recolectaron metadatos que comprendían llamadas realizadas en los cinco años anteriores; este plazo más tarde ese redujo. Con estos datos se construyó un grafo o red de interacción de proporciones gigantescas que permitía establecer rápidamente con quienes se comunicó una determinada persona y, a su vez, con quienes se comunicaron estas segundas personas. El sistema originalmente permitía al analista hacer hasta tres saltos a partir de una persona de interés, luego se redujo a dos. El grafo de MANWAY estaba completamente pre-calculado y se actualizaba periódicamente, de manera que era posible establecer rápidamente las relaciones de interacción para cualquier persona, fuera objeto de una investigación autorizada o no (Gellman, 2020).

Ante los cuestionamientos que se produjeron a medida que se conocía más sobre MAINWAY, las respuestas de las autoridades giraban alrededor de la idea de que no se estaba violando la privacidad de los ciudadanos porque el sistema no recolectaba llamadas, sólo los metadatos. Pero como veremos en la próxima sección la explotación avanzada de metadatos y su cruce con otras fuentes de información, permite inferir información personal y potencialmente sensible.

Es razonable suponer que este tipo de actividades de vigilancia basadas en metadatos no son exclusivas de los Estados Unidos, sino que podrían ocurrir también en otros países. También hacen análisis sobre grafos de interacciones personales las empresas que recolectan grandes cantidades de datos, como las plataformas de redes sociales. Y no se pueden dejar de lado a las organizaciones criminales que aprovechan las crecientes oportunidades de contar con bases de datos masivos obtenidas de intrusiones y hackeos a instituciones públicas y privadas.

Tus metadatos pueden delatarte

Como contrapartida de lo analizado en la sección anterior, los metadatos son muy útiles para diferentes tipos de análisis con impacto social positivo. Por ejemplo, los datos de uso de tarjetas de transporte público permiten mejoras en

la planificación de los servicios y la explotación de datos de geolocalización de teléfonos móviles es fundamental para estimar la efectividad de las cuarentenas sanitarias. Por este motivo es frecuente que las instituciones que recolectan metadatos los pongan a disposición de investigadores o planificadores. Previo a compartirlos, los registros de los datos se anonimizan; esto es, se elimina cualquier información que pueda ser personalmente identificable: nombre, dirección, teléfonos y si están presentes, identificadores en redes sociales. Los nombres y apellidos se reemplazan por un identificador único y arbitrario que se mantiene constante para el mismo individuo en toda la base de datos. Esto es necesario para poder seguir a los individuos en el tiempo y en el espacio, sin necesidad, ni posibilidad, de conocer su identidad.

Sin embargo, se han documentado diversos casos que demuestran que las técnicas de anonimización no son apropiadas porque si se cuenta con información extra las personas se pueden re-identificar. Es decir, que aún en los casos en que se preparan los datos para distribuirlos sin exhibir información personal, existen riesgos de identificación.

En un trabajo ya clásico, De Montjoye y cols. (2013) utilizaron un conjunto de datos de movilidad de celulares que consistía en trayectorias de movilidad de un millón y medio de personas durante quince meses expresadas como datos espacio-temporales. Para la geolocalización se contaba con la ubicación de la torre de telefonía celular más cercana y la localización temporal estaba redondeado a horas. Con esta abundancia de datos, extensión de tiempo y localizaciones de baja resolución se podría suponer que no sería posible identificar las trayectorias de un único individuo. Esto es, dada una localización y hora se podría suponer que habría un número considerable de individuos que la satisfacen. Sin embargo, los autores determinaron que con solo cuatro datos es posible identificar al 90% de los individuos de la muestra. Puesto en términos cinematográficos, supongamos que una detective sigue a un sospechoso por la Ciudad de Buenos Aires, de quien no posee ningún dato, pero anota en su libreta que lo vio en la zona de Estación Liniers a las 9 de la mañana, en Corrientes y Pueyrredón a las 12, en las cercanías del Teatro Colón a las 14 y de nuevo en nuevo en Corrientes y Pueyrredón a las 17. Con estos datos y una consulta a la base de datos espacio-temporales, nuestra detective tiene una probabilidad mayor del 90% de determinar el identificador anonimizado de su sospechoso y con eso reconstruir todas sus trayectorias en los últimos 15 meses y así establecer cuál es su locación en las noches, con lo que tendrá una ubicación aproximada de dónde vive o determinar qué lugares recorre los fines de semana.

Entre las observaciones que se le hicieron a este trabajo se mencionó que en ciudades con varios millones de habitantes, o con bases de datos a escala nacional, la probabilidad de identificación caía tanto que la re-identificación era prácticamente imposible. Sin embargo, más tarde Farzanehfar y cols. (2021)

demostraron que la disminución de la probabilidad de re-identificación con grandes bases de datos es mucho menor que la supuesta anteriormente. Además, se realizaron publicaron varios trabajos similares con metadatos derivados de otras fuentes, como ser, consumos con tarjetas de créditos (De Montjoye, 2015) y cercanía a redes wifi (Boutet y cols, 2016) con resultados similares: con las técnicas actuales de anonimización es relativamente fácil re-identificar individuos.

En un trabajo más preocupante, Mayer y cols (2016) realizaron un relevamiento con voluntarios anonimizados de quienes recopilaron metadatos de llamadas telefónicas durante algo más de un año. Enriquecieron ese conjunto de datos a partir de información pública sobre la contraparte de las llamadas. Una contraparte de una llamada podía resultar identificable porque su teléfono estaba publicado en una red social o si era el número de un comercio. También se enriqueció la geolocalización por coordenadas agregando información sobre el tipo de lugar. El objetivo aquí no era re-identificar, sino demostrar que con metadatos de llamados y el complemento de información pública se puede establecer un perfil más detallado del individuo. En efecto, los autores pudieron identificar entre los voluntarios del estudio individuos con problemas cardiacos y detalles de su atención médica, un caso de embarazo no deseado, otro de posesión -aparentemente legal- de un arma de fuego de alto calibre y otro caso de un participante que realizaría cultivo hogareño de marihuana.

En síntesis, en el pasado se consideraba a los metadatos como los hermanos menores e inofensivos de los datos y con limitado o poco valor. Eso ya no es así, gracias a su volumen y a la existencia de técnicas sofisticadas de explotación, se descubrió el enorme valor que tienen y el potencial para revelar información sensible.

Conclusiones

Los metadatos son componentes esenciales de sistemas de inteligencia artificial y de sistemas de gestión de datos. No es posible descartar su utilización sin afectar el funcionamiento de esos sistemas. En el campo de la ciencia de datos se los aprovecha para planificación de servicios de transporte y planificación urbana. Las empresas los utilizan para promociones, ubicación de locales y planificación de logística. Pero también se observan usos no deseados, como aplicaciones ilegales de vigilancia o de dudosa legalidad. Como ocurre con otros campos tecnológicos, su uso está abierto a abusos y es necesario ajustar y actualizar los controles. Sería deseable una mayor participación ciudadana para estimular a los diferentes estamentos gubernamentales a mantenerse activos en este sentido. Pero las complejidades tecnológicas y jurídicas no facilitan esa participación.

Desde el punto de vista estrictamente tecnológico se están realizando investigaciones y desarrollos para lograr que las técnicas de anonimización sean más efectivas y que al mismo tiempo los metadatos conserven su valor o pierdan solo una pequeña fracción.

La creciente disponibilidad de datos y los avances analíticos y en hardware permiten aprovechar los metadatos en formas inimaginables hace un par de décadas. Es de esperar que los avances continúen y surjan nuevos desarrollos, pero también nuevos desafíos, sobre todo en cuanto a riesgos en la preservación de la privacidad.

Bibliografía

Boutet, A.; Ben Mokhtar S, Primault V (2016) Uniqueness Assessment of Human Mobility on Multi-Sensor Datasets. [Research Report] LIRIS UMR CNRS 5205. Recueprado el 01/07/21 de: https://hal.inria.fr/hal-01381986

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports, 3(1), 1–5.

De Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. Science, 347(6221), 536–539.

Gellman, B (2020) Inside the NSA's Secret Tool for Mapping Your Social Network. *Wired*. Recuperado el 01/07/2021 de: https://www.wired.com/story/inside-the-nsas-secret-tool-for-mapping-your-social-network/

Farzanehfar, A., Houssiau, F., & de Montjoye, Y. A. (2021). The risk of re-identification remains high even in country-scale location datasets. Patterns, 2(3), 100204.
Mayer, J., Mutchler, P., & Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. Proceedings of the National Academy of Sciences of the United States of America, 113(20), 5536–5541.

Uttamchandani, S (2021) Why is reliable metadata becoming important?. *Towards Data Science*. Recuperado el 01/07/2021 de: https://towardsdatascience.com/why-reliable-metadata-is-becoming-important-f29e01b01d4d

Zsoldra P (2016) This is everything Edward Snowden revealed in one year of unprecedented top-secret leaks. *Business Insider*. Recuperado el 01/07/2021 de: https://www.businessinsider.com/snowden-leaks-timeline-2016-9