XII Jornadas de Sociología. Facultad de Ciencias Sociales, Universidad de Buenos Aires, Buenos Aires, 2017.

Complejidad y Epistemología en las prácticas de conocimiento de Big Data y Data Mining; El caso de los algoritmos de análisis de sentimientos.

Juan Pablo López Alurralde.

Complejidad y Epistemología en las prácticas de conocimiento de Big Data y Data Mining; Los casos de los algoritmos de análisis de sentimientos y de algoritmos de recomendación.

Juan Pablo López Alurralde

Eje 2: Epistemología y Metodología

Mesa 48: Complejidad y Sociología

Universidad Nacional de Tucumán / Quilmes

Juanlopezalurraldegmail.com

Abstract

Dada la ingente acumulación de datos digitales en entornos cibernéticos, se han desarrollado diferentes técnicas con las que analizar y producir nuevos conocimientos mediante la detección de patrones que sirven, la mayoría de las veces, para el diseño de esquemas de publicidad y control en la internet. Este ámbito de trabajo se ha llamado como "Big Data" y, además de perseguir fines comerciales, presenta algunas prácticas relevantes para el quehacer de la sociología. En particular, analizaremos algunos de los algoritmos utilizados para el análisis de sentimientos en redes sociales como algoritmos que utilizan datos "sociales" para la generación de recomendación en el dominio del comercio electrónico.

Palabras calves:

Big Data, Sociología, Algoritmos, Epistemología, Datos.

Introducción

Foucault ha analizado lo que ha llamado como "ortopedia social" (Focault, 1981) interesado en dar con los dispositivos modernos con los que se ha desarrollado un control disciplinario bajo cierto mecanismos penales que han configurado, a su parecer, la actual sociedad disciplinaria, y que se presentan, al mismo tiempo, como soporte epistemológico de nuevas ciencias que hacen de los sujetos su objeto de estudio.

Esta sociedad disciplinaria, expresada icónicamente en la figura del panóptico, vuelve a reconfigurarse con la aparición de los dispositivos digitales conectados a la internet. El panóptico, entendido como figura arquitectónica que remite simultáneamente a una forma de poder, disciplinamiento y a una forma de producción de conocimientos, cobra nuevos matices como analogía de las capacidades de conocimiento y control del ciberespacio digital[1].

[1] Tomo el recaudo de agregar, casi la mayoría de las veces, el prefijo "cyber" a la noción de lo "digital", pues lo "digital", en sí mismo, no refiere necesariamente al esquema socio-técnico ni al conocimiento generado por éste que aquí nos interesa. Pues una cosa es un reloj de muñeca digital marca Cassio producido en 1990, y otra bien distinta son las prestaciones que implica un reloj

Efectivamente, tras la intensificación del uso de terminales digitales conectadas a la internet, emerge una nueva forma de conocimiento y control que atañe especialmente a lo que individuos hacen, patentan y proyectan de sus propias subjetividades en el ciberespacio. Las facilidades de procesamiento de información inherentes a los medios digitales es proporcionalmente directa a la capacidad de producción de datos que los actos de los usurarios generan en sus respectivos trayectos digitales. Y en una sociedad en la que la información es la mercancía más importante, esto no pasa desapercibido. Industrias, intereses y nuevas disciplinas crecen de la mano en un mercado que recién transita por sus primeros años. Una postal ilustra los nuevos tiempos: en el podio de las firmas más ricas de la historia figuran marcas como Facebook y Google (Alphabet), cuyos bienes y servicios se fundamentan en secuencias de 1s y 0s en los que se cifran la conducta de todos sus usuarios, y no en productos físicos reducibles a átomos. La ciencia social nunca pagó tanto. Pero los métodos, datos, instrumentos y fines de esta nueva ciencia son totalmente diferentes. Podemos decir que seguimos estando frente a una "ciencia" en la medida en la que se buscar dar con conocimientos verídicos, y "social" pues las variables de análisis recaen sobre varias dimensiones de los sujetos-usuarios.

Diferentes estimaciones redondean en 3 mil millones a la cantidad de usuarios activos en la internet. A esto habría que sumarle la cantidad de dispositivos anexados a cada usuario y las miles de interacciones que surgen entre estos. Teléfonos, computadoras, tablets, Smart TV y un sinfín de nuevos componentes electrónicos que se suman a la ola de la internet de las cosas (IoT). Esta cantidad inaudita de usuarios y puertos de conexión se resume en un aluvión datos que crece día a día y en el que se patentan las elecciones, preferencias, intereses sexuales, posiciones políticas, clase social, miedos, deseos, amores, odios, etc. Nunca hubo tantos datos disponibles acerca de la subjetividad de los usuarios.

El Big Data, Data Mining, Knowledge Discovery in Data Bases (KDD), etc. son las disciplinas que comprenden un amplio espectro de herramientas científicas-tecnológicas que las firmas utilizan para el procesamiento de semejante cantidad de datos. El objetivo es encontrar patrones significativos en un mar de datos que se traduzcan en decisiones que, en la mayoría de los casos, buscan acrecentar la venta algo.

Foucault, en la bibliografía anteriormente citada, realiza una arqueología de las prácticas del poder como formas del saber. Entre ellas concibe a la indagación, como procedimiento para saber lo ocurrido; y al examen, como forma arquetípica del panoptismo donde la vigilancia y el control se

---

digital tipo "SmartWatch", conectado a la internet e intermediario de un sinfín de datos de sus respectivos usuarios. Me interesa lo digital en su enclave cibernético, de inter-conexión.

ejercen continua y sistemáticamente sobre los individuos y sus cuerpos. El "panoptismo digital", por otro lado, no radicaría en una figura precisamente arquitectónica, sino, más bien, en otra abstracta y virtual ¿Cuál y es y qué tiene que decir la nueva epistemología del saber social cyber-digital y de qué manera se configuran y presentan los aspectos disciplinarios del "panóptico digital"?

Problema de investigación

En este trabajo se propone indagar el modo en el que los datos y el conocimiento son tratados y producidos en el ámbito de las técnicas informáticas desarrolladas para beneficio del comercio electrónico dentro del ámbito de lo que se conoce como big data. Más específicamente, se propondrá responder a las preguntas relativas a qué y cómo se conoce en el ámbito de las nuevas técnicas de exploración social que se hacen posibles gracias al alcance y a las prestaciones del entramado cyber-digital como medio de experiencia social masivo.

Entendiendo a la acumulación sistemática de datos como algo connatural e inherente a este medio, se propondrá indagar en algunas técnicas de tratamientos de bases de datos utilizadas para desarrollar aproximaciones, predicciones, correlaciones y descripciones acerca del comportamiento, gustos, intereses, locaciones, intereses, etc. de los cyber-usuarios con el fin (en el mayor de los casos) de incrementar el rendimiento de sus ventas y o servicios.

Atendiendo al funcionamiento e implicancias de estas técnicas, se tratará de entender y comprender el modo en el que el conocimiento es captado, controlado y producido por los actores involucrados en las tareas de big data. En definitiva, se tratará de responderse a las cuestiones epistemológicas de base en lo que resulta ser el auge del nacimiento de nuevas técnicas y disciplinas de exploración social científica (pero con fines comerciales).

Desarrollo

Cuando hablamos de big data hacemos referencia a entornos de trabajo donde la acumulación de datos digitalizados supera en creces las posibilidades de procesamiento que las técnicas estadísticas tradicionales pueden ofrecer. Didácticamente, suele decirse en los manuales de la materia que big data hace referencia a tres "V": Variedad, cuando la fuente y origen de los datos con los que trabajamos son miles y variados; Velocidad, cuando las herramientas de análisis tradicionales no alcanzan una velocidad aceptable de procesamiento; y volumen, cuando el monto de datos es inconmensurable para la mente humana por sí sola siendo del todo necesaria la inserción de herramientas de procesamiento computacional (Berman, 2013).

Ahora bien, vale decir que los datos almacenados en los servidores de las diferentes compañías que registran servicios y sitios en la red de redes, pueden provenir tanto de usuarios humanos, que conectados a la internet a través de sus variados dispositivos dejan sus rastros digitales alojados en "la nube", o bien de cualquier cosa a la que, a través un sensor o lo que sea, hayamos conectado a la internet.

De esta manera, cuando hablamos de big data hablamos de montones de datos relativos no sólo a las preferencias que los usuarios humanos proyectan en sus recorridos digitales y que son de interés para las firmas que los analizan, sino que también contamos con grandes bases de datos relativas a elementos y fenómenos que nada tienen que ver con lo que a un sujeto cualquiera, por ejemplo, podría gustarle. Y es así que tenemos instituciones climáticas, por ejemplo, que analizan sus propias bases de datos, como otras firmas especializadas en el análisis de trayectoria de determinadas acciones en la bolsa de NASDAQ. No hay ningún fenómeno "digitalizable" sobre el que no sea factible practicar procesos de análisis de big data, data mining o lo que fuera. Las aplicaciones del big data son tan diversas y específicas que para darse idea de lo que puede hacerse, pude tomarse como ejemplo lo que firmas como UPS hacen con sus vehículos de transporte, a los que monitorea a través de sensores conectados a la red que miden el desempeño de las partes en orden de poder predecir y prevenir fallas mecánicas (Levis, 2011).

En particular, en orden de precisar los modos en los que este idea de "panóptico digital" puede funcionar, analizaremos en lo siguiente dos tipos de algoritmos altamente utilizados en el ámbito del comercio electrónico y en las técnicas de *Social Media Analytics*. Más concretamente, intentaremos dilucidar los modos en los que ciertos algoritmos producen, usan y exploran datos y conocimientos sociales cyber-digitales.

Análisis de algoritmos de recomendación en *e-commerce*

Es un lugar común del imaginario colectivo, si se quiere, afirmar que firmas como "Google", la DNA o el Estado Argentino, se hacen de datos "íntimos" de los usuarios a través de prácticas que rozan con el más vil espionaje para poder, en el mejor de los casos, "vendernos" con mayor facilidad sus ideas, bienes, servicios y o productos.

Efectivamente, como decíamos más arriba, el interés por las firmas privadas en los datos que los individuos generan en sus respectivos trayectos digitales han sido objeto de valor para estas empresas ya que a partir de estos pueden generar, por ejemplo, interfaces de ventas más personalizadas a cada cliente.

Sin embargo, tras el análisis de algunos algoritmos y métodos de venta, algunas de las presunciones, como las recientemente planteadas, tienden a desdibujarse.

Para adentrarnos a la manera en las que las firmas de comercio electrónico se acercan a nosotros para concretar sus ventas, analizaremos algunos de los algoritmos más ampliamente utilizados en el mercado. Entre ellos, los algoritmos de tipo "*Collaborative Filtering*", "CF" en adelante.

Análisis: "Collaborative Filtering Algorithms"

Producir recomendaciones personalizadas de calidad es uno de los desafíos más importantes del comercio electrónico. La cantidad de visitantes, la cantidad de bienes disponibles, la inmensa cantidad de datos alojados que los usuarios generan en cada una de sus visitas, y la necesidad de ofrecer respuestas instantáneamente (incluso aún antes de que los mismos usuario sepan qué es lo que quieren, por ejemplo, comprar) son variables que tienden a complicar la tarea de generar interfaces de ventas personalizadas efectivas que ayuden a concretar las transacciones comerciales.
Para responder a estos desafíos se hicieron uso de un cierto tipo de algoritmos llamados "Collaborative Filtering". Estos algoritmos son una especie específica dentro del género de algoritmos de recomendación. En general, estos algoritmos "aprenden" de los clientes y a partir de ello generan recomendaciones (Schafer, Konstan, & Riedl, 1998). Los algoritmos CF, en particular, funcionan de la siguiente manera:

> *Collaborative filtering works by building a database of preferences for items by users. A new user, Neo, is matched against the database to discover neighbors, which are other users who have historically had similar taste to Neo. Items that the neighbors like are then recommended to Neo, as he will probably also like them.*
>
> *The basic idea of CF-based algorithms is to provide item recommendations or predictions based on the opinions of other like-minded users. The opinions of users can be obtained explicitly from the users or by using some implicit measures.* (Sarwar, Karypis, Konstan, & Riedl, 2001)

El objetivo final de este tipo de algoritmo es sugerir nuevos ítems o predecir la utilidad de un ítem para un usuario particular basándose en datos previos de gusto y preferencia y en las opiniones de otros usuarios que han mostrado una tendencia similar en materia de gustos.

Ahora bien, debido a ciertos limitantes técnicos, se confeccionaron en la práctica dos tipos diferentes de CF.

Los primeros y mayormente utilizados fueron llamados "*Memory-based Collaborative Filtering Algorithm*"[2]. Estos utilizan toda la base de datos de usuarios e ítems disponibles para generar una predicción. Para hacerlo, lo primero que realizan es crear un set de usuarios, a los que llaman "*neighbors"*, de los que se supone que poseen un historial de búsquedas y gustos compatible para cada usuario-cliente nuevo que aparezca en el sitio de ventas. Una vez que la universo de la "vecindad" del usuario activo ha sido definido, el algoritmo arroja como resultado los ítems que su vecindad o "*nearest-neighbor"* han prefigurado (Sarwar, Karypis, Konstan, & Riedl, 2001).

Sin embargo, estos algoritmos se han enfrentado a una serie de inconvenientes técnicos que derivaron en su creciente desuso y apuesta por otros tipos de algoritmos. En primer lugar, uno de los serios inconvenientes que enfrentan estos algoritmos es la escasa información de gustos de los usuarios en función de la totalidad de los bienes (ítems) disponibles. En el mejor de los casos, los usuarios más activos no han evaluado ni comprado menos del 1% de los bienes disponibles de cada *retailer*. Dadas estas circunstancias, un algoritmo basado en "*nearest-neighbor*" se encuentra imposibilitado de hacer alguna recomendación para cada usuario en particular. En segundo lugar, este tipo de algoritmos implica computar la totalidad de los datos de usuarios disponibles, datos que crecen día a día y que complican la posibilidad de que el mismo pueda ejecutarse en un lapso de tiempo aceptable en la generación de sugerencias de compras (tengamos en cuenta que la mayoría de estos algoritmos se ejecutan *on-line* mientras los usuarios visitan una página).

El segundo tipo de algoritmo CF, que ha sabido posicionarse por sobre el anterior ha sido clasificado como "*Model-based Collaborative Filtering Algorithm*". Estos proveen de recomendaciones a partir del desarrollo de un modelo probabilístico de *rating* para cada usuario a través de algoritmos basados en *machine learning*. *At it's core*, la intuición fundamental de estos algoritmos descansa en la presunción de que cada usuario encontrará útil y gustoso los ítems similares a los que ya haya demostrado cierto interés.

A diferencia de la configuración del algoritmo anterior, este último modelo "bucea" a través del set de ítems a los que el usuario ya ha evaluado (*input*) para computar luego la similitud que guardan ente los ítems que el usuario aún no ha evaluado (y posiblemente no comprado). Paso seguido, el algoritmo devuelve (*output*) el ítem que presuntamente será del gusto y utilidad del usuario.

El punto más complejo de este modelo de algoritmo reside en el modo en el que se computa la similitud entre los diferentes Ítems. Para ello se han desarrollado varios métodos: cociente de similitud, correlación de similitud, cociente ajustado de similitud, etc (Sarwar, Karypis, Konstan, & Riedl, 2001).

[2] Este tipo de algoritmo guarda estrecha relación con otro formato de algoritmos de recomendación llamados "Clustering Algorithms". Estos proceden de manera similar:

Este modelo particular ha sido llamado «"ítem-to-ítem" "Collaborative Filtering"»[3] y ha podido solucionar los problemas que otros algoritmos no. En particular, puede ejecutarse a gran velocidad (a pesar de correr *on-line*); necesita de pocos datos como insumos para ejecutarse; y, lo más importante de todo, la calidad de las sugerencias y predicciones es aceptable.

En lo que sigue, comenzaremos a analizar el funcionamiento de otros dispositivos. Luego haremos algunos comentarios "epistemológicos" respecto a los mismos.

Sentiment-Analysis

Otra forma de encarar el marketing para el acrecentamiento de ventas consiste en la exploración de las posiciones sentimentales que los usuarios de la red asocian a cada producto, idea, servicio o lo que fuere. Estás técnicas exploratorias, en realidad, pueden ser ejecutadas por cualquiera que tenga el interés de analizar los sentimientos subjetivos asociados a cualquier cosa y a gran escala. En la mayoría de los casos, son las firmas privadas las interesadas en evaluar cómo reaccionan los clientes a sus bienes o servicios con el fin de mejorar su eficiencia y rentabilidad.

Este tipo de análisis recae en algoritmos y dispositivos totalmente diferentes a los recientemente nombrados. Su observación nos permitirá entender un poco más la riqueza del espectro de mecanismos de producción, uso, exploración y descubrimiento de datos y conocimientos que alberga el mundo del big data. Para nuestro caso en particular, nos centraremos en el estudio de análisis de sentimientos realizados en Twitter.

"Lexicon-based" Algorithm for Sentiment Analysis

Este tipo de análisis forma parte del género de estudios llamado "*Natural Language Processing*" y tiene su origen en la estadística y computación lingüística (Brand, Hurwitz, Nugent, Halper, & Kaufman , 2013). Más específicamente, estos algoritmos o programas pueden entenderse como parte del subgénero de herramientas subsumidas en "*Text Analytics*".

El algoritmo en particular que tomaremos para nuestro análisis, llamado "*Lexicon-based*" *algorithm*, fue desarrollado específicamente para el análisis de datos en Twitter. Aún así, sus creadores, arguyen que el mismo funcionaría en otros campos textuales (Taboada, Brooke, Tofiloski, Voll, & Stede, 2014). Traigo a colación esta observación para que se tenga en cuenta que muchas de las propiedades que le serán predicadas a este algoritmo en particular, describen también a otros muchos algoritmos creados para el análisis de sentimientos.

[3] La gran aparición de este tipo de algoritmos sucede cuando Amazon (la tienda virtual más redituable de la historia) decide dejar a un costado el criterio de exponentes de las letras y editores en su sistema de recomendación de libros para comenzar a recomendarlos a través de este algoritmo. El tiempo le daría a la razón, al tanto que la firma comenzó a utilizar este algoritmo para la venta de todo su catálogo. (Mayer & Cukier, 2013)

Ahora bien, en un contexto de big data, donde la cantidad de datos a analizar es humanamente insondable, se hace necesario programar un algoritmo que pueda, en la medida de lo posible, operar automáticamente en la determinación de sentimientos asociados a textos. No hay, durante el proceso de ejecución de estos algoritmos, mayor participación humana en la codificación de los sentimientos ¿Cómo puede computarizarse algorítmicamente algo tan complejo y humano como los sentimientos que asociamos a las cosas?

Para determinar los sentimientos expresados en textos, lo que estos algoritmos hacen es calcular la orientación semántica de las expresiones textuales que en cada caso se analiza. De esta manera, la determinación de la orientación semántica es traducida como una unidad de medida de un estado subjetivo y o de opinión expresados en textos.

A su vez, para calcular la orientación semántica, se realiza una doble operación por la que se tratan de medir dos variables diferentes: por un lado se calcula un factor evaluativo (positivo o negativo) y, por el otro, la intensidad y potencia de estas posiciones. Esta operación puede realizarse sobre palabras, frases, sentencias o documentos enteros en función de su posición hacia una persona, marca, producto o lo que sea el caso.

A un nivel más específico, lo que estos algoritmos hacen es utilizar, mayoritariamente, a los adjetivos como indicadores de orientación semántica. Para ello se compila previamente un diccionario de adjetivos a los que se les adjudica una polaridad (positiva o negativa) y un valor numérico de intensidad (del -5 al +5 en nuestro caso de estudio [siendo el 0 un indicador numérico de neutralidad]). Una vez que las palabras del texto son analizadas por el tamiz de este diccionario de valores, se adjudica (como output) un *score*.

Como se desprende de lo dicho, el cálculo de sentimientos expresados en textos a través del cálculo de orientación semántica de estos, se respalda en dos supuestos: por un lado, en la creencia de que las palabras tienen una suerte de "polaridad primordial", esto es, una orientación semántica independiente del contexto; y, en segundo lugar, que estas orientaciones semánticas pueden ser cuantificadas en valores numéricos.

Muchos son los inconvenientes técnicos que surgen tras la aceptación de estos supuestos. Para "afinar" los resultados de estos algoritmos, en este caso en particular, los autores tomaron en cuenta también el cálculo de orientación semántica latente en sustantivos, verbos, adverbios y en lo que llaman "intensificadores" y "*polarities shifter*". Los primero son expresiones que como "muy" o "mucho" potencian la intensidad semántica de los adjetivos y adverbios a los que se asocia. Las otras expresiones, por otro lado, hacen referencias a las funciones de reversión semántica que palabras o frases como "no", "ni" o "de ningún modo" operan en el seno de los enunciados en los que aparecen.

Además de este tipo de algoritmos, se han desarrollados otros que, funcionando con principios parecidos, realizan un cálculo semántico a través del análisis en el uso de emoticones y hashtags (Kouloumpis, Wilson , & Moore, 2010).

Sin duda pueden encontrarse miles de planteos al modo en que los principios de esta técnica funcionan. Sin embargo, mientras tanto, las firmas de análisis de sentimientos afinan sus algoritmos mientras se llenan los bolsillos. Pues poder dar con muestras globales acerca del posicionamiento de marcas o productos en la subjetividad de las personas es algo por lo que muchos pagan.

Consideraciones epistemológicas de los casos de estudio

Quisiera aclarar que lo que me motivó a la hora de seleccionar estos tipos de algoritmos fue la presunción de creer que a través de estos se generaban o se usaban grandes cantidades de datos "sociales" para la persecución de fines comerciales. De esta manera, me vería posibilitado de responder las preguntas que aquí me interesaba plantear: ¿Qué se conoce y cómo se conoce en el ciberespacio o "panóptico digital"? Esto es, si entendemos y tomamos la idea de "panóptico", tal como lo hace Foucault, como mecanismo de poder y saber, para describir tales capacidades en lo que son las prácticas de big data o data mining en el ciberespacio, entonces lo que me interesaba ver era justamente de qué manera estas funcionalidades operan en este ámbito. Seguí mi intuición según la cual la generación, explotación y usos de datos sociales a nivel masivo ocurre por medio de algoritmos. Paso seguido me detuve en los algoritmos analizados.

Ahora bien, tras finalizar el análisis estos algoritmos, observamos que los mismos representan y son parte de técnicas muy diferentes. Así las cosas, sería difícil dar una respuesta unívoca (si es que siquiera da lugar en cada caso el plantearlas) a las preguntas recientemente expresadas.

Para empezar a responder estas preguntas, es necesario volver a nuestra definición de big data. Allí hicimos referencia a la variedad de datos que implica una práctica del estilo. Y, efectivamente, las fuentes de datos disponibles para su análisis y procesamiento son tan diversos que difícilmente puedan ser tamizados todos por los mismos tipos de técnicas y algoritmos.

A *grosso modo*, la naturaleza de los datos, en el ámbito del big data, puede clasificarse dentro de dos segmentos totalmente diferentes. Por un lado tenemos *structured data* y, por el otro, *unstructured data*. El primer término refiere generalmente a los datos que tienen longitud y formato definido, en tanto que con el segundo término se refiere a aquellos datos que no poseen un formato definido. Vale agregar que la mayoría de los datos existentes caen bajo esta segunda clasificación.

Para que nos terminemos de dar una idea más clara, dos son los tipos de fuentes de donde provienen los datos estructurados. Estos datos pueden generarse tanto a través de operaciones ejecutadas por computadoras (automáticamente) o por seres humanos (manualmente). Más específicamente, este

tipos de datos son los que pueden provenir de sensores (de GPS, por ejemplo), de registros automatizados de visitas y ventas en páginas web, etc.

Al igual los datos estructurados, los datos no estructurados pueden provenir también de generaciones automáticas por medio de computadoras y o por accionar manual humano. Caen en este dominio todos los datos contenidos en los archivos de texto, imagen, video, audio. Aquí, no hay que malinterpretar, el formato de los documentos que contienen los datos en cuestión sí se encuentran estructurados, pero no así los datos al interior de los mismos.

Ahora bien, en lo que respecta a los algoritmos que hemos analizados, cabe decir que los algoritmos del tipo *Collaborative Filtering* hacen uso de data estructurada, mientras que los de análisis semántico de texto se ejecutan sobre data no estructurada.

Esta no es una observación menor, en la medida en la que podemos situar la finalidad de este segundo tipo de algoritmos como un proceso que comienza con data no estructurada (como *input*) para dar al final con data estructurada (como *output*). Así, podemos decir que el procesamiento de datos no estructurados tiene como fin transformarlos en información al traducirlos mediante alguna estructuración. En el caso del *Lexicon-based algorithm*, por ejemplo, lo que tenemos en un comienzo son millones de publicaciones en twitter a los que se analiza en función de detectar tendencias estadísticas sentimentales asociados a marcas o productos. Lo que comienza entonces siendo un montón de textos, termina siendo traducido (y estructurado) a un valor numérico, que en este caso medirá la intensidad de un sentimiento (positivo o negativo) respecto a algo.

En cambio, los algoritmos de recomendación que analizamos poseen un camino diferente. Estos dispositivos tienen como insumos datos estructurados (relativo a los usuarios y a los ítems de un catálogo) que terminan siendo procesados por un algoritmo que arroja como resultado otro dato estructurado. Pero lo llamativo de este caso, es que, a diferencia de los algoritmos de análisis de sentimientos, no generan un dato con carácter representativo de alguna información predicable a algún sujeto o usuario.

De los dos modelos de algoritmos que vimos de este especie, unos, los *Memory-based CF*, toman datos estructurados de la bases de los usuarios (en su conjunto) para generar un esquema matemático de predicción y probabilidad en materia de compatibilidad de preferencias sobre ítems previamente valorados. Este esquema de predicción, a su vez, termina arrojando como resultado un ítem específico para un usuario que aún no los ha evaluado ni presuntamente comprado[4].

[4] La misma dinámica de funciona en los algoritmos de recomendación de canciones (en *Spotify* por ejemplo) o de vídeos (en plataformas como *Netflix*).

Tenemos entonces datos "sociales"[5], si se quiere, que hacen referencia a los gustos de un universo de usuarios que son segmentados en función de tendencias para generar esquemas de predicción de compatibilidad de preferencias que puedan, a su vez, generar recomendaciones específicas de ítems particulares para usuarios particulares en momentos singulares.

El proceso, sin embargo, es diferente en el segundo tipo de *CF*. Los bien llamados "*ítem-to-item" CF*, generan también un esquema de predicción de compatibilidades en materia de preferencias, pero, a diferencia de los anteriores, los *Model-based CFA* toman como insumo solo datos concernientes a los ítems en venta y no los relativo a los usuarios a los que se trata de vender. No hay aquí un procesamiento de "datos sociales" en la recomendación de productos.

Es importante volver destacar que éste es el modelo de algoritmos más vigentes en los sistemas de recomendaciones (Linden , Smith, & York, 2003). En este gran universo de operaciones del ciberespacio o "panóptico digital", no sucede entonces un procesamiento de datos sociales para la generación de intercambios comerciales o para la recomendación de consumo de productos culturales como libros, canciones o películas que podríamos disfrutar. Para ello, en cambio, sí hay un análisis de los mismos objetos, pero no de los sujetos a los que se les trata de acercar.

Volviendo a los algoritmos de análisis de sentimientos, estos implican, al contrario de los anteriores, la generación genuina de información o conocimiento social a escalas realmente significativas. Si somos capaces de generar información relativa a las reacciones a nivel subjetivo que nos producen películas, o zapatillas, podemos hacer el mismo tipo de análisis para prácticamente lo que sea. Y de hecho sucede. El problema u observación a tener en cuenta es que estos análisis persiguen el interés privado de las firmas que pagan para tener conocimiento acerca del posicionamiento sentimental que sus respectivos productos generan en su audiencia. Con esto quiero decir, simplemente, que científicos sociales (de formación tradicional) podrían tomar estar herramientas para ejecutar análisis con directrices, si se quiere, más epistémicamente nobles.

De todas maneras, la información de contenido social que estas técnicas producen, no se erige como una instancia de representación "teórica" relativa a los usuarios (humanos) a los que trataría de captar en su "naturaleza" social o individual. Al contrario, se puede situar esta información sólo a los efectos de ser útil a una instancia de mejoramientos de rendimientos de ventas, eficacia en servicios, etc. Saber, pero no para conocer; saber para vender.

Asimismo, esta información social que generan técnicas como éstas, no se fundamentan sobre la supuesta trascendencia de un cúmulo determinados de sujetos a quienes se predican ordenadamente

[5] Estos "datos sociales" que son, simultáneamente, de naturaleza estructurada, se generan y alojan automáticamente a partir de las acciones de calificación y compra de productos que los usuarios registran en sus respectivas trayectorias de navegación en la red. Ésta es una manera en la cual puede entenderse las operaciones que suceden en el medio "panóptico digital" o ciberespacio en lo relativo a la captación de datos "sociales" o adjudicarles a individuos.

los valores contenidos en la información generada por los algoritmos. No hay, o no interesa patentarse la figura de un *substratum* al que le predicamos o no ciertos *accidens*.

Asimismo, vale decir, en comparación con los sistemas de conocimientos tradicionales, que quien "ostenta" la "representación", es decir, el conjunto de datos que informan acerca de interacciones pasadas e intereses potenciales de los potenciales clientes, no son, precisamente, personas, científicos, sino algoritmos.

El conocimiento que terminan arrojando como *outputs* las diversas prácticas como las analizadas, no termina siendo tampoco, a la postre, un conjunto o corpus de enunciados universales ni sobre la "sociedad en general" (si es que existiese algo por el estilo) ni sobre los individuos (entendidos como "cosa en sí") que la componen.

Al contrario, el conocimiento descubierto y explotado en medio de las bases de datos se corresponde con representaciones fugaces y volátiles relativas a la interacción de dos o más variables que pueden predicarse de la correlación entre sujetos y objetos, o entre propiedades de objetos y sujetos. Para ser más claros, queremos decir no hay una idea en sí, platónicamente entendida, de la sociedad o las sociedades que pueda al menos ser apresada por estas técnicas de explotación. No hay una verdad "sociológica universal" que nos espere del otro lado. Por el momento, en lo que depara las posibilidades de estas prácticas del big data, sólo nos queda la posibilidad de dar con apreciaciones casuales de individuos y propiedades sociales que cuentan posiblemente con más utilidad que verdad (entendida ésta como *adaequatio rei et intellectus*).

Habrá quienes dicen que tampoco existen cosas como "sociedades" y que, en su lugar, existen interacciones de actantes heterogéneos vinculados por circuitos aún más heterogéneos (Latour, 1998); actantes que son definidos y redefinidos a partir de sus interacciones (Callon, 1986); interacciones, muchas de las cuales, se vuelven rastreables y analizables en la medida en la que suceden o son traducidas al medio cyber-digital (Venturini, Jensen, & Latour, 2015). Estas presunciones, propias de la teoría del actor-red, encontrarían en las prácticas del big data el ideario de una ciencia del "colectivo", entendido este concepto como la superación del dualismo naturaleza-sociedad en un mismo plano dimensional. Este interesante planteamiento merece otro lugar de análisis. Volvamos a lo nuestro.

Por último, volvamos un segundo sobre la idea de panóptico. Al mismo se lo define como un mecanismo dual, como una infraestructura tecnológica que permite tanto la fundación de campos epistémicos (como la psicología y la psiquiatría) y, al mismo tiempo, como un mecanismo de control (como el que se observan cárceles yo escuelas).

Ahora bien, con respecto a los últimos tipos de algoritmos de recomendación analizados, como decíamos, estos no generan una información con valor social. Sino, simplemente, la recomendación de un mero producto, o "ítem", para ser aún más genéricos. Esta práctica, entonces, no podría ser catalogada como una práctica de producción epistémica. La pregunta que vale plantearse entonces es si estos mecanismos pueden ser considerados como dispositivos de control al constreñir y "diseñar" científicamente nuestras conductas comerciales ¿No sería ésta, acaso, la manifestación del carácter disciplinario que el "panóptico digital" opera sobre los usuarios/individuos? Esto, además, habría que plantearlo a la luz de que estos algoritmos no sólo intermedian nuestras compras, sino también los insumos culturales que elegimos para nuestro placer y entretenimiento, como los artículos periodísticos que elegimos para in-formarnos.

Consideraciones epistemológicas finales

A pesar de haber analizado aquí un mínimo de las prácticas de uso, explotación y generación de información o conocimiento social a partir de grandes bases de datos, podemos explayarnos en algunas consideraciones en torno a ciertos caracteres transversales propios del mundo del big data. En las palabras de sus más entusiasmados voceros, el abanico de herramientas de big data no sólo emerge como un mercado de crecimiento exponencial. Sino que, simultáneamente, comienza a plantear la utilidad de un nuevo paradigma científico que, a diferencia de la "ciencia tradicional", no intenta dar con un "porque", sino en un "que"; que no se maneja con hipótesis, sino con datos; y que no se maneja con muestras, sino con datos exhaustivos. Más precisamente, este paradigma científico se caracteriza por fundamentarse en operaciones que, más que causalidad, ofrece correlaciones (Mayer & Cukier, 2013).

Las correlaciones, a decir verdad, son relaciones entre variable que el instrumental básico de la estadística clásica puede figurar tranquilamente. Sin embargo, algo nuevo emerge en las posibilidades del big data. Y es que las correlaciones que permite superan el esquema lineal propio de las correlaciones de la estadística clásica. De repente, con semejante cantidad de datos disponibles para su análisis, comenzamos a ser capaces de percibir, big data mediante, correlaciones complejas entre variables no-lineales y, con ello, la posibilidad de cierta comprensión de fenómenos y dinámicas complejas (como la de fenómenos sociales).

En definitiva, el corazón epistemológico de este presunto nuevo paradigma científico, inaugurado por la emergencia del big data, descansa en sus capacidades de predicción y correlación. Las variables a las que se puede someter a este tipo de análisis, son miles y variadas. El alcance explicativo potencial de esta nueva ciencia, también.

Bibliografía

Anderson, C. (2008). The End of Theory: The Data Deluge makes the Scientific Method Obsolete. *Wred*.

Berman, J. (2013). *Principles of Big Data* (Primera ed.). Londres: British Library Cataloguing-in.Publication Data.

Brand, W., Hurwitz, J., Nugent, A., Halper, F., & Kaufman , M. (2013). *Big Data for Dummies* (Primera ed.). Hoboken: John Wiley $ Sons, Inc.

Callon, M. (1986). *Algunos elementos para una sociología de la traducción.*

Deleuze, G. (1992). *Postcript on the Societies of Control* (Primera ed.). Paris: MIT Press.

Focault, M. (1981). *La verdad y las formas jurídicas* .

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinsky , M., & Brilliant , L. (2009). Detecting influenza epidemics using engine query data. *Nature*.

Kouloumpis, E., Wilson , T., & Moore, J. (2010). *Twitter Sentimen Analysis: The Good the bad and the OMG!*

Latour, B. (1998). Tecnología es sociedad hecha durable. En M. Doménech, & F. J. Tirado, *Sociología simétrica* (págs. 109-142). Barcelona: Gedisa.

Levis, J. (2011). *Data on UPS'savings* (Primera ed.). Institut for Operations and the Management Sciences.

Linden , G., Smith, B., & York, J. (2003). *Amazon.com Recommendations, item-to-item Collaborative Filtering.*

Mayer, V., & Cukier, K. (2013). *Big Data: A Revoluton that will transform the way we think, lve and work.* New York: Houghton Mifflin Harcourt Publishing Company.

Palmas, K., & Kullenberg, C. (2009). *Contagiontology* (Primera ed.). Oslo: Glänta.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). *Item-based Collaborative Filtering Recommendation Algorithms* (Primera ed.). Hong Kong: GroupLens Research Group.

Savage, M., & Burrows, R. (2007). *The Coming Crisis of Empirical Sociology* (Primera ed.). Los Angeles: SAGE Publications.

Schafer, B., Konstan, J., & Riedl, J. (1998). *Recommender Systems in E-Commerce* (Primera ed.). Boston: GroupLens Research Project.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2014). *Lexicon-Based Methos for Sentiment Analysis.*

Venturini, T., Jensen, P., & Latour, B. (2015). Fill in the gap. A new alliance for social and natural sciences. *Journal of artificial societies and social simulation*, 1-4.