

Minerías de información

Definición sencilla de la minería:

El proceso o negocio de cavar en las minas para obtener minerales, metales, joyas, etc.

Minería de datos: ¿Qué es la minería de datos?

Visión de conjunto

En general, la minería de datos (a veces llamada de datos o descubrimiento de conocimiento) es el proceso de analizar los datos desde diferentes perspectivas y resumir en información útil, información que puede ser utilizada para aumentar los ingresos, reduce los costes, o ambos. El software de minería de datos es uno de una serie de herramientas analíticas para el análisis de datos. Permite a los usuarios analizar datos de muchas dimensiones o ángulos diferentes, lo categorizan, y resumen las relaciones identificadas. Técnicamente, la minería de datos es el proceso de encontrar correlaciones entre los patrones o los campos en grandes bases de datos relacionales.

Innovación continua

Aunque la minería de datos es un término relativamente nuevo, la tecnología no lo es. Las compañías han utilizado potentes ordenadores para tamizar a través de volúmenes de datos de escáner de supermercados y analizar los informes de investigación de mercado durante años. Sin embargo, las continuas innovaciones en la potencia de computación, almacenamiento en disco, y el software de estadística está aumentando drásticamente la exactitud del análisis al tiempo que reduce el costo.

Ejemplo

Por ejemplo, una cadena de supermercados Medio Oeste utiliza la capacidad de extracción de datos de software de Oracle para analizar los patrones de compra locales. Descubrieron que cuando los hombres compran pañales los jueves y los sábados, también tendían a comprar cerveza. Un análisis más detallado mostró que estos compradores normalmente hicieron su compra semanal los sábados. Los jueves, sin embargo, sólo han comprado algunos artículos. El minorista llegó a la conclusión de que debía comprar la cerveza para tenerla disponible para el próximo fin de semana. La cadena de supermercados podría utilizar esta información recientemente descubierta en diversas maneras de aumentar los ingresos. Por ejemplo, podrían mover la pantalla de cerveza más cerca de la pantalla del pañal. Y, podrían asegurarse de que la cerveza y los pañales se venden a precio completo de los jueves.

Los fundamentos de la minería de datos

Las técnicas de minería de datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos de negocio se almacenaban en primer lugar en las computadoras, continuó con mejoras en el acceso a los datos, y más recientemente, generó tecnologías que permiten a los usuarios navegar a través de sus datos en tiempo real. La minería de datos lleva este proceso evolutivo más allá del acceso a los datos retrospectivos y de navegación para la entrega de información prospectiva y proactiva. La minería de datos está lista para su aplicación en la comunidad de negocios, ya que se apoya en tres tecnologías que ya están suficientemente maduras:

- La recopilación de datos masiva
- Equipos con varios procesadores potentes
- Algoritmos de minería de datos

Las bases de datos comerciales están creciendo a un ritmo sin precedentes. Una reciente encuesta del Grupo META de los proyectos de almacenamiento de datos encontró que el 19% de los encuestados están más allá del nivel de 50 gigabytes, mientras que el 59% espera estar allí. En algunas industrias, tales como el comercio minorista, estas cifras pueden ser mucho mayores. La necesidad de acompañamiento para la mejora de los motores de cálculo se podrá satisfacer de manera rentable con la tecnología informática con varios procesadores en paralelo. Los algoritmos de minería de datos incorporan técnicas que han existido durante al menos 10 años, pero sólo han sido implementadas recientemente como herramientas maduras, fiables y comprensibles que superan ampliamente a los métodos estadísticos de mayor edad.

En la evolución de los datos de negocio a la información de negocios, cada nuevo paso se ha basado en el anterior. Por ejemplo, el acceso de datos dinámica es fundamental para la obtención de detalles en las aplicaciones de navegación de datos, y la capacidad de almacenar grandes bases de datos es fundamental para la minería de datos.

Los datos, información y conocimiento

Datos

Los datos son los hechos, números o texto que pueden ser procesados por un ordenador. Hoy en día, las organizaciones están acumulando vastas cantidades de datos en diferentes formatos y diferentes bases de datos cada vez mayor. Esto incluye:

- Datos operacionales o transaccionales, tales como, ventas, costos, inventarios, nómina y contabilidad.

-Los datos no operacionales, tales como ventas de la industria, los datos de pronóstico, y los datos macro económicos.

-Datos meta, datos acerca de los datos en sí, como el diseño de base de datos lógicos o definiciones del diccionario de datos

Información

Los patrones, asociaciones o relaciones entre todos estos datos pueden proporcionar información. Por ejemplo, el análisis del punto de datos de transacciones de venta al por menor puede proporcionar información sobre los productos que están vendiendo y cuándo.

Conocimiento

La información puede ser convertida en el conocimiento de los patrones históricos y las tendencias futuras. Por ejemplo, la información de resumen sobre las ventas de supermercado puede ser analizada a la luz de los esfuerzos de promoción para proporcionar el conocimiento del comportamiento de compra del consumidor. Por lo tanto, un fabricante o minorista podrían determinar qué artículos son los más susceptibles a los esfuerzos de promoción.

Almacenes de datos

Los espectaculares avances en la captura de datos, capacidad de procesamiento, transmisión de datos y capacidades de almacenamiento están permitiendo a las empresas integrar sus diversas bases de datos en unidades de almacenamiento de datos. El almacenamiento de datos se define como un proceso de gestión de datos centralizado y de recuperación. El almacenamiento de datos, como la minería de datos, es un término relativamente nuevo, aunque el concepto en sí ha existido durante años. El almacenamiento de datos representa una visión ideal de mantener un repositorio central de todos los datos de la organización. Se necesita la centralización de los datos para maximizar el acceso de los usuarios y el análisis. Los espectaculares avances tecnológicos están haciendo de esta visión una realidad para muchas empresas. Y, avances igualmente dramáticos en el software de análisis de datos están permitiendo a los usuarios acceder a esta información libremente. El software de análisis de datos es lo que apoya la minería de datos.

¿Qué puede hacer minería de datos?

La minería de datos es principalmente usada por las empresas con un fuerte enfoque del consumidor (minorista, financiero, comunicación y organizaciones de marketing). Permite a estas empresas determinar las relaciones entre los factores "internos" como el precio, el posicionamiento del producto, o las habilidades del personal, y los factores "externos", tales como los indicadores económicos, la competencia y demografía de los clientes. Y, que les permite determinar el impacto en las ventas, satisfacción del cliente y las

ganancias corporativas. Por último, les permite "profundizar" en la información de resumen para ver los datos transaccionales detallados.

Con la minería de datos, un minorista podría utilizar los registros de punto de venta de compras de los clientes para enviar promociones específicas basadas en el historial de compras de un individuo. Gracias a la minería de datos demográficos de comentario o de tarjetas de garantía, el minorista podría desarrollar productos y promociones para atraer a segmentos específicos de clientes.

Por ejemplo, la minería de Blockbuster Entertainment es su base de datos histórica de alquiler de videos para recomendar individualmente a los clientes en sus vacaciones. American Express puede sugerir productos a sus titulares de tarjetas basadas en el análisis de sus gastos mensuales.

WalMart es pionera en la minería de datos masiva para transformar sus relaciones con los proveedores. WalMart captura transacciones en puntos de venta de más de 2.900 tiendas en 6 países y continuamente transmite esos datos a su masivo 7,5 terabytes de almacenamiento de datos Teradata. WalMart permite a más de 3.500 proveedores, acceder a datos sobre sus productos y realizar análisis de datos. Estos proveedores utilizan estos datos para identificar patrones de compra de los clientes a nivel de exhibición de la tienda. Ellos utilizan esta información para gestionar el inventario de almacén local e identificar nuevas oportunidades de comercialización. En 1995, los equipos de WalMart procesan más de 1 millón de consultas de datos complejos.

La Asociación Nacional de Baloncesto (NBA) está explorando una aplicación de minería de datos que se puede utilizar en conjunción con las grabaciones de imágenes de los juegos de baloncesto. El software avanzado del explorador analiza los movimientos de los jugadores para ayudar a los entrenadores a orquestar jugadas y estrategias. Por ejemplo, un análisis de la hoja de play-by-play del juego entre los New York Knicks y los Cavaliers de Cleveland el 6 de enero de 1995, es que cuando Mark Price jugó la posición de guardia, John Williams intentó cuatro tiros en salto e hizo cada uno. Avanzados no sólo se encuentra este patrón, pero se explica que es interesante porque difiere considerablemente del porcentaje de aciertos promedio de 49.30% para los Cavaliers durante ese juego.

Al utilizar el reloj universal de la NBA, un entrenador puede usar automáticamente los clips de vídeo que muestran cada uno de los tiros intentados por Williams, sin necesidad de ir a través de horas de vídeo. Esos clips muestran un juego de pick-and-roll muy exitoso en el que Price desarma la defensa de Knick y luego encuentra a Williams para un salto de tiro abierto.

¿Cómo funciona la minería de datos?

Si bien la tecnología de información a gran escala ha ido evolucionando, los sistemas analíticos y transacciones separadas, la minería de datos proporciona el vínculo entre los dos. El software de minería de datos analiza las relaciones y patrones en los datos de transacción almacenados en base a consultas de los usuarios indefinidos. Existen varios

tipos de software de análisis que están disponibles: estadística, aprendizaje automático, y redes neuronales. En general, cualquiera de los cuatro tipos de relaciones es buscado:

Clases: Los datos almacenados se utilizan para localizar los datos en grupos predeterminados. Por ejemplo, una cadena de restaurantes podría extraer datos de compra del cliente para determinar cuando visitan los clientes y lo que normalmente ordenan. Esta información podría ser utilizada para aumentar el tráfico, para tener especiales del día.

Grupos: Los elementos de datos se agrupan de acuerdo a las relaciones lógicas o preferencias de los consumidores. Por ejemplo, los datos pueden ser extraídos para identificar segmentos de mercado o afinidades de consumo.

Asociaciones: Los datos pueden ser extraídos para identificar asociaciones. El ejemplo cerveza-pañal es un ejemplo de la minería asociativo.

Patrones secuenciales: Los datos se extraen de anticipar tendencias y patrones de comportamiento. Por ejemplo, un minorista en equipo al aire libre podría predecir la probabilidad de una mochila que se compra basada en la compra de un consumidor de sacos de dormir y zapatos para caminar.

La minería de datos se compone de cinco elementos principales:

- Extraer y transformar datos de transacciones de carga en el sistema de almacenamiento de datos.
- Almacenar y gestionar los datos en un sistema de base de datos multidimensional.
- Proporcionar acceso a los datos, a los analistas de negocios y profesionales de la tecnología de la información.
- Analizar los datos con la aplicación de un software.
- Presentar los datos en un formato útil, como un gráfico o una tabla.

Diferentes niveles de análisis están disponibles:

- Las redes neuronales artificiales: modelos predictivos no lineales que aprenden a través de la formación y se asemejan a las redes neuronales biológicas en la estructura.
- Los algoritmos genéticos: técnicas de optimización que utilizan procesos tales como la combinación genética, mutación y selección natural en un diseño basado en los conceptos de la evolución natural.
- Los árboles de decisión: estructuras en forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos. Los métodos específicos árbol de decisión incluyen árboles de clasificación y regresión (CART por sus siglas en ingles) y la interacción Chi cuadrado de Detección Automática (CHAID por sus siglas en ingles). CART y CHAID son técnicas de árbol de decisión

utilizadas para la clasificación de un conjunto de datos. Proporcionan un conjunto de reglas que se pueden aplicar a un nuevo conjunto de datos (sin clasificar) para predecir qué registros tendrán un resultado dado. Los segmentos de CART son un conjunto de datos mediante la creación de un camino de 2 divisiones, mientras que los segmentos CHAID utilizan pruebas de chi cuadrado para crear vías múltiples divisiones. CART normalmente requiere menos preparación de datos que CHAID.

-El método del vecino más cercano: Una técnica que clasifica cada registro de un conjunto de datos basado en una combinación de las clases k de las ficha (s) más similares a él en un conjunto de datos históricos (donde $k > 1$). A veces se llama la técnica vecino k -más cercano.

-Inducción de reglas: La extracción de las reglas útiles de los datos basados en la significación estadística.

-La visualización de datos: La interpretación visual de las relaciones complejas de datos multidimensionales. Las herramientas gráficas se utilizan para ilustrar las relaciones de datos.

¿Qué infraestructura tecnológica se requiere?

Hoy en día, las aplicaciones de minería de datos están disponibles en todos los sistemas de tamaño para plataformas mainframe, cliente / servidor, y PC. Los precios de los sistemas van desde varios miles de dólares para las aplicaciones más pequeñas, hasta \$1 millón el terabyte para el más grande. Las aplicaciones en toda la empresa en general, varían en tamaño de 10 gigabytes a más de 11 terabytes. NCR tiene la capacidad de ofrecer aplicaciones de más de 100 terabytes. Hay dos factores tecnológicos críticos:

-Tamaño de la base de datos: cuantos más datos se procesa y se mantiene, más potente es el sistema que se requiere.

-La complejidad de la consulta: cuanto más complejas las consultas y mayor el número de consultas que se están procesando, más potente es el sistema requerido.

El almacenamiento de base de datos relacional y la tecnología de gestión son adecuados para muchas aplicaciones de minería de datos de menos de 50 gigabytes. Sin embargo, esta infraestructura necesita ser mejorada significativamente para soportar las aplicaciones más grandes. Algunos proveedores han añadido amplias capacidades de indexación para mejorar el rendimiento de las consultas. Otros utilizan nuevas arquitecturas de hardware, tales como procesadores masivamente paralelos (MPP) para lograr mejoras de orden de magnitud en el tiempo de consulta. Por ejemplo, los sistemas MPP de NCR enlazan cientos de procesadores Pentium de alta velocidad para alcanzar niveles de rendimiento superiores a las de los mayores superordenadores.

Minería de textos

La minería de texto es un nuevo campo emergente que intenta extraer información significativa del texto natural de la lengua. Puede ser caracterizado en términos generales como el proceso de análisis de texto para extraer información que es útil para fines particulares. En comparación con el tipo de datos almacenados en bases de datos, el texto es estructurado, amorfo, y difícil de tratar de forma algorítmica. Sin embargo, en la cultura moderna, el texto es el vehículo más común para el intercambio formal de información. Los campos de la minería de texto por lo general se ocupan de los textos cuya función es la comunicación de los hechos, informaciones u opiniones, y la motivación para tratar de extraer información de dicho texto automáticamente es convincente, incluso si el éxito es sólo parcial.

La frase "la minería de texto" se utiliza generalmente para referirse a cualquier sistema que analiza grandes cantidades de texto y lenguaje natural y detecta los patrones de uso de léxico o lingüísticos en un intento de extraer la información probablemente útil.

La minería de texto y minería de datos

Al igual que la minería de datos puede describirse en términos generales como la búsqueda de patrones en los datos, minería de texto se trata de buscar en patrones de texto. Sin embargo, la similitud superficial entre las dos oculta verdaderas diferencias. La minería de datos puede ser más plenamente caracterizada como la extracción de implícita, de información previamente desconocida, y potencialmente útil de datos. La información está implícita en los datos de entrada: es oculta, desconocida, y apenas se puede extraer sin necesidad de recurrir a las técnicas automáticas de minería de datos. Con la minería de texto, sin embargo, la información que se extrae es de forma clara y precisa en el texto. No está escondido para nada, la mayoría de los autores se aseguran de que ellos se expresan claramente y sin ambigüedad y, desde una perspectiva de un punto de vista humano, el único sentido en el que es "hasta ahora desconocido" es que las restricciones de recursos humanos hacen que no resulte factible que la gente lea el texto ellos mismos. El problema, por supuesto, es que la información no está formulada de una manera que es susceptible de procesamiento automático. La minería de texto se esfuerza por llevar el texto en una forma que es adecuada para el consumo por las computadoras directamente, sin necesidad de un intermediario humano.

Aunque hay una diferencia clara filosóficamente, desde el punto de vista de la computadora los problemas son bastante similares. El texto es tan opaco como los datos en bruto cuando se trata de extraer lo más detallado.

Otro requisito que es común para ambos, los datos y la minería de texto es que la información extraída debe ser "potencialmente útil." En un sentido, esto significa accionable-capaz de proporcionar una base de acciones que deben tomarse de forma automática. En el caso de la minería de datos, esta noción se puede expresar en una manera relativamente independiente del dominio: los patrones procesables son los que permiten hacer predicciones no triviales que se harán en los nuevos datos de la misma

fuente. El rendimiento puede medirse mediante recuento de éxitos y fracasos, las técnicas estadísticas se pueden aplicar para comparar diferentes métodos de minerías de datos en el mismo problema, y así sucesivamente. Sin embargo, en muchas situaciones de minería de texto es mucho más difícil caracterizar lo que "procesable" significa de una manera que sea independiente del dominio particular. Esto hace que sea difícil encontrar medidas justas y objetivas de éxito.

En muchas aplicaciones de minería de datos, "potencialmente útiles" se le da una interpretación diferente: la clave para el éxito es que la información extraída debe ser comprensible, ya que ayuda a explicar los datos. Esto es necesario cuando el resultado está destinado al consumo humano en lugar de una base de acción automática. Este criterio es menos aplicable a la minería de texto porque, a diferencia de la minería de datos, la entrada en sí es comprensible. La minería de texto con la salida comprensible es equivalente a resumir características más destacadas de un gran cuerpo de texto, que es un sub-campo por derecho propio: el texto de resumen.

La minería de texto y procesamiento del lenguaje natural

La minería de texto parece abarcar la totalidad del tratamiento automático del lenguaje natural y, posiblemente, mucho más, además de, por ejemplo, el análisis de las estructuras de vinculación como referencias bibliográficas en la literatura académica y los hipervínculos en la literatura Web, tanto de fuentes útiles de información que se encuentran fuera del dominio tradicional de procesamiento del lenguaje natural. Pero, de hecho, la mayoría de los esfuerzos de minería de texto rechazan conscientemente los más profundos y cognitivos aspectos del procesamiento del lenguaje natural clásico en favor de las técnicas más superficiales afines a los utilizados en la recuperación de información práctica.

La razón se entiende mejor en el contexto del desarrollo histórico del tema de los recursos de procesamiento naturales del lenguaje. Las raíces del campo se encontraban en proyectos de traducción automática a finales de 1940 y principios de 1950, cuyos aficionados asumieron que las estrategias basadas en la traducción palabra por palabra lo haría proporcionar traducciones ásperas dignas y útiles que podrían ser fácilmente perfeccionadas en algo más precisas, utilizando técnicas basadas en el análisis primario sintáctico. Pero el único resultado de estos proyectos de alto perfil, financiados en gran medida, fue la clara constatación del lenguaje natural, incluso a la altura de los niños analfabetos, es un medio increíblemente sofisticado que no sucumbe a técnicas simplistas. Depende fundamentalmente de lo que consideramos como el conocimiento de "sentido común", que a pesar de la causa de su naturaleza, todos los días es excepcionalmente difícil de codificar y utilizar en forma algorítmica.

Como resultado de estos fallos embarazosos y muy publicitados, los investigadores retiraron a " mundo de juguete", en especial el "mundo de bloques" de los objetos geométricos, formas, colores, y apilado (operaciones cuya semántica es clara y explícita, posible codificar). Pero gradualmente se convirtió en éxito, los mundos de juguete, aunque en un principio impresionante, no se traduce en el éxito de las piezas realistas de

texto. Las técnicas de juguetes del mundo se ocupan bien de frases construidas artificialmente de lo que podríamos llamar la variedad "Dick y Jane" después de la conocida serie del mismo nombre de cuentos infantiles. Pero fracasan estrepitosamente cuando se enfrentan con el texto verdadero, ya sea con esmero, construido y editado o producido en restricciones de tiempo real (como conversación informal).

Mientras tanto, los investigadores en otras áreas, simplemente tuvieron que lidiar con el texto real, con todos sus caprichos, idiosincrasias, y los errores. Los esquemas de comprensión, por ejemplo, deben trabajar bien con todos los documentos, cualquiera que sea su contenido, y evitar un fallo catastrófico, incluso cuando se procesan escandalosamente archivos desviados (como archivos binarios o de entrada completamente al azar). Los sistemas de recuperación de información deben indexar documentos de todo tipo y les permiten estar ubicados de manera efectiva en cualquiera que sea su materia o corrección lingüística. La clave de la extracción y de los algoritmos del resumen de texto es que tienen que hacer un trabajo decente en cualquier archivo de texto. Los sistemas de trabajo y las prácticas en estas áreas son temas independientes, ya que la mayoría son independientes del lenguaje. Operan mediante el tratamiento de la entrada como si fueran datos, no lenguaje.

La minería de texto es una consecuencia de esta forma de pensar "texto real". Aceptando que es probable que no es mucho, lo que se puede hacer con la entrada sin restricciones, ¿La capacidad de procesar grandes cantidades de texto puede compensar técnicas relativamente simples?

Es interesante que la minería de datos también evolucionara a partir de una historia de relaciones difíciles entre disciplinas, en este caso de aprendizaje de la máquina, arraigado en la ciencia informática experimental, con metodologías especiales de evaluación y estadísticas bien fundamentadas teóricamente, pero en base a una tradición de probar hipótesis indicadas explícitamente en lugar de buscar nueva información. Los primeros investigadores del aprendizaje automático sabían o se preocupaban poco de las estadísticas; los primeros investigadores de hipótesis estadísticas estructuradas permanecieron ignorantes del trabajo paralelo en el aprendizaje de la máquina. El resultado fue que las técnicas similares (por ejemplo, la construcción de árboles de decisiones y el vecino más cercano) surgieron en paralelo de las dos disciplinas, y sólo más tarde hicieron un acercamiento equilibrado.

Minería de sentimientos

Las computadoras pueden ser buenos en trabajar con números, pero pueden crujir sentimientos?

El surgimiento de los blogs y las redes sociales ha generado un mercado en torno a la opinión personal: opiniones, valoraciones, recomendaciones y otras formas de expresión en la red. Para los informáticos, esta montaña de rápido crecimiento de los datos es la apertura de una ventana tentadora en la conciencia colectiva de los usuarios de Internet.

Un campo emergente conocido como análisis de los sentimientos está tomando forma alrededor de una de las fronteras inexploradas del mundo informático: la traducción de los caprichos de la emoción humana en datos duros.

La teoría de la "cognición encarnada" sugiere que una variedad de actividades mentales se reflejan en los estados del cuerpo, tales como las posturas, los movimientos del brazo y expresiones faciales. Un estudio investiga el grado en que los perfiles de los usuarios de computadoras - su sexo, sentimientos y experiencias emocionales - pueden evaluarse a partir de los movimientos de los cursores de ordenador.

En un experimento, los participantes (N = 372) vieron a tres fragmentos de películas durante dos minutos cada uno, calificaron sus sentimientos después, y llevan a cabo tareas de percepción simples, tres veces, nuestro programa trazó la trayectoria del cursor de los participantes cada 20 milisegundos. Se investigó el grado en que las características extraídas de la trayectoria del cursor podrían revelar los perfiles de los participantes. Los resultados indicaron que un pequeño número de variables de trayectoria fueron útiles para identificar qué película vieron los participantes, cómo se sintieron durante la visualización de la película, y su género. Se sugiere que los movimientos del cursor proporcionan amplia información para la minería de un perfil de usuario dinámico.

Esto es más que un ejercicio de programación interesante. Para muchas empresas, la opinión en línea se ha convertido en una especie de moneda virtual que puede hacer o quebrar un producto en el mercado.

Sin embargo, muchas empresas luchan por dar sentido a la caja o baúl de quejas y felicitaciones que ahora giran en torno a sus productos en línea. Como herramientas de análisis de emociones que comienzan a tomar forma, no sólo podrían ayudar a las empresas a mejorar sus resultados finales, sino también con el tiempo transformar la experiencia de búsqueda de información en línea.

Varias nuevas empresas de análisis de emociones están tratando de aprovechar el creciente interés de las empresas en lo que se dice en línea.

"Los medios sociales solían ser este proyecto para los consultores 25 años de edad", dijo Margaret Francis, vicepresidente de producto en los laboratorios del explorador en San Francisco. Ahora, dijo, los altos ejecutivos lo "están reconociendo como una increíblemente y rica vena de inteligencia de mercado."

Scout Labs, que está respaldado por la firma de capital de riesgo iniciada por el fundador de CNet Halsey Minor, ha introducido recientemente un servicio de suscripción que permite a los clientes monitorear blogs, artículos de noticias, foros en línea y sitios de redes sociales para las tendencias de opiniones sobre productos, servicios o temas en las noticias.

A principios de mayo, la venta de entradas StubHub utiliza herramienta de monitorización del explorador Labs para identificar un aumento repentino del sentimiento negativo del blog después de la lluvia retrasando un juego de los Sox Yankees-Red.

El estadio oficial le dijo erróneamente a cientos de fans que el juego había sido cancelado y StubHub negaron las peticiones de los aficionados a las restituciones, con el argumento de que el juego en realidad había sido jugado. Pero después de detectar problemas en línea de la cerveza, la compañía ofreció descuentos y créditos a los aficionados afectados. En la actualidad esta re-evaluando su política de mal tiempo.

"Este es un canario en una mina de carbón para nosotros", dijo John Whelan, director de servicio al cliente de StubHub.

Jodange, con sede en Yonkers, ofrece un servicio dirigido a los editores en línea que les permite incorporar los datos de opiniones procedentes de más de 450.000 fuentes, incluidas las fuentes de la corriente principal de noticias, blogs y Twitter.

Basado en una investigación por Claire Cardie, un ex profesor de informática de Cornell, y Jan Wiebe, de la Universidad de Pittsburgh, el servicio utiliza un sofisticado algoritmo que no sólo evalúa sentimientos sobre temas particulares, sino que también identifica los titulares de opinión más influyentes.

Jodange, cuyos primeros inversores incluyen la Fundación Nacional de Ciencia, está trabajando actualmente en un nuevo algoritmo que podría utilizar los datos de opinión para predecir la evolución futura, como pronosticar el impacto de editoriales de periódicos en una empresa de precio de las acciones.

En una línea similar, el Financial Times ha introducido recientemente Newssift, un programa experimental que realiza el seguimiento de sentimientos sobre temas de negocios en las noticias, junto con un motor de búsqueda especializado que permite a los usuarios organizar sus consultas por tema, organización, lugar, persona y el tema.

Usando Newssift, una búsqueda de Wal-Mart reciente revela que el sentimiento sobre lo que la compañía está ejecutando es positivo en una proporción de un poco mejor de dos a uno. Cuando esa búsqueda se refina con el término sugerido "Fuerza y los sindicatos," sin embargo, la proporción de positivos a los sentimientos negativos está más cerca de uno a uno.

Estas herramientas podrían ayudar a las empresas a puntualizar el efecto de cuestiones concretas sobre las percepciones de los clientes, ayudándoles a responder con estrategias de marketing y relaciones públicas adecuadas.

Para los internautas casuales, encarnaciones más simples de análisis de sentimientos están surgiendo en forma de herramientas ligeras como Tweetfeel, Twendz, y Twitrratr. Estos sitios permiten a los usuarios tomar el pulso de los usuarios de Twitter sobre temas particulares.

Una búsqueda rápida en Tweetfeel, por ejemplo, revela que el 77 por ciento de los Twiteros les gusta la película "Julie & Julia". Sin embargo, la misma búsqueda en Twitrratr revela un par de fallos de encendido. El sitio le asigna una puntuación negativa a un tweet leído "Julie y Julia era verdaderamente encantador" Ese mismo mensaje terminaba con

"todos nos sentimos muy hambriento después de esto" - y el sistema tomó la palabra "hambre" para indicar un sentimiento negativo.

Mientras que los algoritmos más avanzados utilizados por los laboratorios de exploración, Jodange y Newssift emplean análisis avanzados para evitar este tipo de trampas, ninguno de estos servicios funciona perfectamente. "Nuestro algoritmo es de aproximadamente 70 a 80 por ciento de precisión," dijo Francis, quien añadió que sus usuarios pueden reclasificar los resultados inexactos, para que el sistema aprenda de sus errores.

Traducir el material resbaladizo del lenguaje humano en valores binarios siempre será una ciencia imperfecta, sin embargo. "Los sentimientos son muy diferentes de los hechos convencionales", dijo Seth Grimes, el fundador de la firma de los suburbios de consultoría Maryland Plana Alta, que apunta a los muchos factores culturales y matices lingüísticos que hacen difícil para convertir una cadena de texto escrito en un simple sentimiento en pro o en contra. "Pecador, es una buena palabra cuando se aplica a la torta de chocolate", dijo. El trabajo de los algoritmos más simple es escanear por palabra clave para clasificar una declaración como positiva o negativa, sobre la base de un simple análisis binario ("amor" es bueno "odio" es malo). Sin embargo, este enfoque no logra captar las sutilezas que traen el lenguaje humano a la vida: la ironía, el sarcasmo, la jerga y otras expresiones idiomáticas. El análisis de los sentimientos fiable requiere analizar muchos tonos de gris en la lingüística.

"Se trata de la confianza que puede ser expresado de forma sutil," dijo Bo Pang, un investigador de Yahoo que co-escribió "Minería de opinión y análisis de sentimientos", uno de los primeros libros académicos sobre el análisis de los sentimientos.

Para llegar a la verdadera intención de un comunicado, Pang desarrolló un software que analiza varios filtros diferentes, incluyendo la polaridad (es la declaración positiva o negativa), la intensidad (¿cuál es el grado de emoción que se expresa?) y la subjetividad (la forma parcial o imparcial es la fuente).

Por ejemplo, la preponderancia de los adjetivos a menudo indica un alto grado de subjetividad, mientras que las declaraciones verbales y sustantivos tienden hacia un punto de vista más neutral.

Mientras los algoritmos de análisis de emociones se vuelven más sofisticados, deberían comenzar a producir resultados más precisos que pueden llegar a señalar el camino a los mecanismos de filtrado más sofisticados. Podrían convertirse en una parte de uso de la Web todos los días.

"Veo el análisis de opiniones convertirse en una característica estándar de los motores de búsqueda", dijo Grimes, que sugiere que este tipo de algoritmos podrían comenzar a influir tanto para fines generales de búsqueda en la web y las búsquedas más especializadas en áreas como el comercio electrónico, las reservas de viajes y críticas de cine.

Pang prevé un motor de búsqueda que especifica en detalle los resultados para los usuarios basado en la confianza. Por ejemplo, podría influir en el orden de los resultados de búsqueda para ciertos tipos de consultas como "mejor hotel de San Antonio".

Así como los motores de búsqueda comienzan a incorporar más y más datos de opinión en sus resultados, la distinción entre hecho y opinión puede empezar a borrarse hasta el punto que, como David Byrne dijo una vez, " todos los hechos vienen con puntos de vista."

Sentimientos contradictorios sobre el negocio de la minería y la manipulación de las emociones

En la encantadora nueva película animada, "Inside Out", que se toma dentro de la cabeza de Riley, una niña de 11 años de edad, para cumplir con los personajes que representan a cinco de las seis emociones que los psicólogos han caracterizado como universales: alegría, tristeza, miedo, la ira y disgusto. (La sexta emoción: la sorpresa, se omitió, tal vez porque los productores de películas, como la mayoría de la gente de negocios, odia las sorpresas.) Sin revelar ningún spoiler, basta con decir que, en Riley, como en las cabezas de la mayoría de las chicas reales de su edad, Joy presenta algunas imágenes de su mente a la tristeza, ira, miedo y los demás miembros, menos lindos del círculo emocional.

En esta película y en películas como "Avatar" y "Toy Story", los animadores fueron informados e inspirados por el trabajo pionero de psicólogo Paul Ekman en la cartografía de los pequeños cambios en la expresión facial. Toda esa información sobre las acciones a tomar en cuenta en la película fue dada en base a la minería de comportamientos y sentimientos de las personas. Pero los cineastas no son los únicos profesionales que recurren a Ekman en busca de inspiración y guía. La CIA, TSA y otras organizaciones, preocupados por la seguridad emplean la actividad facial de codificación para erradicar a los mentirosos y personas con malas intenciones. Y los anunciantes, deseosos de entrar en las cabezas de los consumidores y dar forma a nuestras decisiones antes de que estemos siquiera conscientes de hacerlas, ven el lavado de oro en la comercialización de máquinas de resonancia magnética funcional y en la detección de cámaras de nuestras pequeñas sonrisas, muecas y movimientos de los ojos. Ellos están tratando de probar cómo los anuncios nos hacen sentir, microsegundo a microsegundo, para garantizar que se minimizan las barreras emocionales a su mensaje y maximizar la alegría u otro incentivo emocional que genera.

Todas las decisiones que hoy en día toman las empresas están basadas en una gran base de datos que han ido llenando mediante la observación del individuo, la razón por la que ofrecen cierto tipo de productos está dada por la facilidad que le proporciona a dichas empresas la minería de sentimientos.

El Internet es una parte cada vez más importante en nuestras vidas. Los usuarios de Internet comparten información y opiniones en las redes de medios sociales donde expresan sus sentimientos, juicios, emociones personales fácilmente. La minería de textos y técnicas de recuperación de información nos permiten explorar toda esta

información y descubrimos qué tipos de opiniones, reclamos, o afirmaciones son las que hacen los autores.

En resumen la minería en el área de recopilación de datos sirve para determinar qué tipo de información están buscando los usuarios, facilitar el uso de grandes cantidades de información, de textos, clasificar características, conocer las preferencias de los clientes de una empresa. Todo esto con el objetivo que a los fines del interesado convengan. Generalmente las empresas recopilan todo este tipo de información para saber qué productos o servicios presentarle al cliente, de qué forma va a reaccionar, en que estará interesado.

Por otro lado la clasificación de información ha venido a dar una gran ayuda a aquellas personas que manejan grandes cantidades de datos, gracias a sistemas cada vez más veloces en el procesamiento de dichos datos.

Referencias:

La minería de datos prácticos, máquinas y herramientas de aprendizaje y técnicas con las implementaciones de Java (2000). Ian H. Witten, Eibe Frank. Editorial Morgan Kaufmann

La percepción basada en la minería de datos y toma de decisiones en economía y finanzas (2007). Ildar Batyrshin, Leonid Sheremetov, Lofti A. Zadeh. Editorial Illustrated

Correlaciones neuronales decisiones y acciones, opinión actual en neurobiología (2010). B. Pesaran.