Perspectivas Contemporâneas
Revista eletrônica de ciências sociais aplicadas.
ISSN: 1980-0193

APROXIMACIÓN A LA DIFUSIÓN DE INFORMACIÓN Y CONOCIMIENTO EN DEMOCRACIA

APPROACH TO THE DIFFUSION OF INFORMATION AND KNOWLEDGE IN DEMOCRACY

Juan Agustín Franco Martínez [(1)]
Universidad de Extremadura - Espanha

RESUMEN

El objetivo general de este trabajo es analizar el proceso de difusión de la información y conocimiento científico como mecanismo para la toma de decisiones democráticas, destacando las técnicas de minería de textos. La metodología adoptada es un meta-análisis de artículos científicos indexados en Science Direct sobre técnicas de gestión de datos, en particular, la minería de textos, con el propósito de estudiar su proceso de difusión y realizar predicciones. Se concluye que los artículos de Text Mining siguen un modelo de difusión logística (de influencia interna), cuyo máximo será aproximadamente en 2017. Destaca también la ausencia de investigaciones en ciencias sociales y en especialidades interdisciplinares como la toma de decisiones en contextos democráticos.
Palabras clave: Democracia, conocimiento, minería de datos, análisis de difusión.

ABSTRACT

The main objective of this paper is to analyze the process of diffusion of information and scientific knowledge as a mechanism for democratic decision making, highlighting text mining techniques. The methodology adopted is a meta-analysis of articles indexed in Science Direct on technical data management, in particular, text mining, in order to study the diffusion process and make predictions. We conclude that the articles of Text Mining follow a logistic distribution (internal influence), whose maximum is approximately 2017. Also highlights the lack of research in the social sciences and interdisciplinary specialties such as decision making in a democratic context.
Key Words: Democracy, knowledge, text mining, diffusion analysis.

INTRODUCCIÓN

La toma eficiente de decisiones democráticas necesita en nuestros días de la asistencia de técnicas que minimicen los costes de esfuerzo y de tiempo invertidos en el proceso, además de contrarrestar los efectos perniciosos sobre el bienestar social de los intereses empresariales privados. En este sentido, las técnicas que tratan de explotar el contenido significativo de un conjunto de datos o de una ingente masa de información son cada vez más necesarias, especialmente en todo lo referente a Internet, que ha supuesto una explosión de enormes cantidades de datos e informaciones,

universalizando así el acceso a la información y al conocimiento. El objetivo general de este trabajo consiste en aproximarnos al estudio de la difusión de la información y el conocimiento como mecanismo para una mejor toma de decisiones democráticas. Los objetivos intermedios se centran, en primer lugar, en el análisis metodológico de diferentes aspectos teóricos relacionados con las preferencias individuales y colectivas, con la hipótesis de racionalidad económica, y con la influencia de las Tecnologías de la Información y la Comunicación en la toma de decisiones democráticas a nivel público o privado.

En un primer acercamiento podemos distinguir varias etapas de análisis de la información, necesarias para una mejor gestión democrática del conocimiento, por un lado, la etapa de extracción de información; por otro, la etapa de generación de conocimiento; y finalmente, la etapa de difusión del conocimiento generado en la etapa anterior. Entre las técnicas de la primera etapa (extracción de información) se encuentran las herramientas de explotación de datos (data mining), especialmente son de una relevancia singular las técnicas de explotación de textos (text mining), por su novedad y potencial en su apoyo en la toma de decisiones on-line en contextos democráticos. Su característica fundamental reside en su capacidad para generar y transmitir conocimiento relevante, algo que por sí mismo Internet es incapaz de ofrecer ni de ser gestionable individualmente. Con esta perspectiva se realiza una reflexión previa sobre las limitaciones teóricas del concepto de "individuo" que ignora la interdependencia entre bienestar individual y social. Después se enmarca el problema del bienestar social y la democracia bajo el Teorema de Arrow. Posteriormente se describen los peligros que acechan a las decisiones democráticas en sociedades donde prima la 'racionalidad económica' que arrincona la información veraz. Se enumeran las principales etapas de análisis de la información para la toma eficiente de decisiones democráticas, destacando las técnicas de Text Mining como unas de las más relevantes. Y después se plantea un meta-análisis de la literatura existente sobre minería de textos, estudiándose varios modelos de difusión de artículos científicos sobre Text Mining a partir de una muestra de la base de datos Science Direct entre 1997-2007. Finalmente se comentan los principales resultados y conclusiones del estudio.

FUNDAMENTO TEÓRICO

Las limitaciones teóricas del concepto de "individuo"

Planteamos a continuación una reflexión clave que no debería dejarse de lado con respecto al significado estadístico de "individuo" y la supuesta independencia a él asociada y a su conjunto de preferencias personales. Si nadie está aislado, y todos y cada uno de los seres humanos forman parte de más de un grupo social, entonces, se comparten preferencias. Las supuestas "preferencias individuales e independientes" son, en realidad, colectivas y dependientes. Sin olvidarnos de las diversas formas de influencia que existen en la actualidad, cuyo ejemplo más evidente son las técnicas de marketing y de los mass media con su continuo recurso a la imagen y la emoción. En consecuencia, se invalidan muchos de los resultados estadísticos asociados a la independencia de variables. Todo lo cual nos lleva a preguntarnos sobre una cuestión que

se analiza en el siguiente epígrafe: ¿no sería más adecuado trabajar en todo el proceso de generación de conocimiento con la estructura de preferencias grupal? Es decir, se plantea la siguiente hipótesis: no hay ninguna estructura de preferencias individuales que no sea la expresión de una estructura de preferencias colectiva (aunque la definición precisa de las personas que integran dicho colectivo sea desconocida).

Dado que se ha demostrado la imposibilidad de agregar preferencias individuales (Teorema de Arrow), ¿por qué trabajar al nivel de preferencias individuales? Sería factible aprovechar el uso de Internet no sólo para expresar (individualmente) y debatir cuestiones (colectivamente), sino para facilitar la formación natural de equipos que consensúen una estructura de preferencias grupal. En otras palabras, ¿qué significa "una estructura de preferencias individual"? Podría significar que cada persona puede dividir su afiliación entre diversas opciones de manera proporcional o no. De forma que se diluye el concepto de invididuo ("indivisible"). ¿Y qué es lo que puede ser divisible en una persona? Su tiempo y su dinero. Cada persona dedica en función de su grado de libertad y sentido de la responsabilidad una proporción de su tiempo y de su dinero a determinadas tareas que contribuyen a una finalidad concreta. El tiempo y el dinero sí son agregables (y ponderables). El "grado de libertad" se refiere a la independencia asociada a la madurez sicológica, grado de influenciabilidad, nivel de inteligencia (académica y emocional), capacidad de crítica constructiva, independencia económica, entre otros. Mientras que el "sentido de la responsabilidad" se refiere al cumplimiento del deber, grado de motivación, tolerancia al fracaso, asunción de las consecuencias de las propias decisiones, asunción de compromisos… En términos estadísticos diríamos que a cada opción/preferencia se le asigna un número de horas (y de euros) procedente de cada participante, de manera que al final tendremos un número total de horas y de euros que se destinan a un objetivo concreto. ¿A cuántos "individuos" equivale si cada "individuo" dispone de 24 horas diarias y un salario medio de 1500 euros mensuales? El cálculo es sencillo.

Algunas conclusiones evidentes que pueden derivarse de lo anterior: lo lógico y más probable es que haya en términos reales menos "individuos" dedicándose a una tarea que personas. ¿Por qué hablamos en términos de probabilidad? Porque la productividad (intensidad) de cada hora y de cada euro no será la misma para todas las personas. Incluso sería posible comprobar cómo también existen un pequeño número de tareas (de ejecución de poder real efectivo) en las que el número de personas destinadas a ellas será menor que su equivalente en "individuos", debido a su elevada productividad (y quizá escaso sentido democrático). Se observa, por tanto, que es preciso desvincular con claridad los conceptos de "persona" (concepto humano integral) e "individuo" (concepto técnico estadístico parcial). En el fondo de las cuestiones comentadas sobre el concepto atomístico (reducido, simplista) de "individuo" subyace un planteamiento teórico que se distancia de los enfoques probabilísticos clásicos y se aproxima más a enfoques posibilísticos asociados a modelos de conjuntos difusos (Terán, 2002).

Teorema de Arrow y democracia

Dado que Arrow (1950) establece una serie de condiciones fuertes que imposibilitan

una elección social óptima, cabría plantearse si detrás de tales condiciones no subyace una determinada concepción de "elección social óptima" que el propio Teorema de Arrow no hace explícita. Máxime cuando ya es conocido que el "principio de Pareto" (considera que el bienestar social aumenta si aumenta el bienestar de una persona, sin empeorar el de nadie) es insuficiente desde el punto de vista del bienestar social, ya que no se pregunta sobre la distribución inicial de los recursos. De hecho, consagra la desigualdad y el aumento de la brecha entre ricos y pobres. Como muy bien afirma Navarro (2014: 9): "A lo máximo que el conocimiento económico llega es al análisis de la pobreza, centrándose más en los pobres que en las causas de la pobreza. Es común oír o ver la expresión de que 'no me importan las desigualdades o que la gente sea tan rica como pueda. Lo único que me importa es la pobreza'. El problema con este dicho, muy común entre economistas liberales, es que las desigualdades y la pobreza están íntimamente relacionadas".

En conclusión, en el principio de Pareto se oculta deliberadamente el modelo de reparto inicial del poder, del estatus quo social, de suma importancia para toda concepción legítima y rigurosa que nos hagamos sobre una "elección social óptima". En definitiva, lo que implica el Teorema de Arrow es que, si no existe la imposición o el gobierno dictatorial, tampoco existirá un modelo de agregación de todas las preferencias individuales que nos sirva para definir un modelo de satisfacción general. Esto invalida el supuesto de que la conducta individual egoísta (mecanismo de libre mercado) genera decisiones sociales integrando todas las preferencias individuales. No basta con agregar las preferencias de cada individuo para determinar las preferencias de la sociedad en su conjunto.

Un ejemplo clásico es el de la paradoja del voto, donde la regla de la mayoría no da un resultado satisfactorio: sea cual sea la alternativa mayoritaria, siempre habrá alguien que prefiera una alternativa distinta. No existe ningún sistema democrático (respetando todas las preferencias individuales) que resuelva este problema cíclico. Una sociedad no puede alcanzar la satisfacción general si el criterio de partida es el de la preferencia individual. Es preciso un criterio ético externo al sistema de preferencias individuales. Es necesario un objetivo social explícito para definir la preferencia social. Esto significa que necesariamente los problemas económicos se resuelven fuera del mercado, fuera de las preferencias individuales (Torres, 1999).

Profundicemos, a continuación, sobre la posibilidad de mejorar las condiciones de elección democrática a partir de la crítica a los 5 principios del Teorema de Arrow. Así, algunos posibles supuestos irreales en los que se basa este teorema serían los siguientes: 1) enfoque estático, 2) negación de la dimensión social, 3) falacia de la transitividad, 4) juicios de valor inherentes y 5) dictadura de las alternativas. Para cada crítica se indicarán otros posibles enfoques de análisis.

1) El enfoque estático se refiere a la no consideración explícita del tiempo en todo proceso de toma de decisiones, al cambio permanente que sufren las alternativas y los individuos. La misma toma de decisiones es un concepto dinámico.

Más aún, las alternativas que se proponen en el momento inicial pueden variar con respecto a las alternativas finalmente realizadas (es la clásica crítica que

se hace a los programas electorales antes y después de las elecciones). Igualmente ocurre con los individuos, puede suceder que una parte de la población esté asumiendo como propias las decisiones que tomaron otros en el pasado. Pero no sólo esa interacción "inter" (inter-alternativas e inter-individuos), sino también "intra", más difícil de identificar, especialmente en intervalos de tiempo infinitesimales. La experiencia y el continuo proceso de aprendizaje en el que se ve envuelto todo ser humano hacen irrealista un enfoque estático en la toma de decisiones. He aquí, una primera implicación para la toma de decisiones sistémica, la dimensión temporal. Ignorar este fenómeno es a riesgo de continuar remando en las aguas estancadas del equilibrio general competitivo de los economistas neoclásicos (de belleza matemática irrefutable, pero inútil para el análisis económico de la realidad).

2) Negación de la dimensión social en los análisis de toma de decisiones a nivel individual. Subyace una visión de homogeneidad acerca de las condiciones en las que se toman las decisiones sobre las que se sustenta todo un mundo de múltiples posibilidades, alternativas y combinaciones. Subyace una concepción económica individualista, donde los individuos son idénticos unos a otros, que es lo que hace posible el planteamiento de una agregación individual de preferencias para la obtención de una estructura aditiva de preferencias sociales. Franco (2013) muestra que esto no es así, ni siquiera para algo tan 'científicamente' aceptado como la forma decreciente de la demanda.

3) La falacia de la transitividad. No sólo es una condición imposible por la razón aducida más arriba sobre la cuestión del cambio permanente que sufren alternativas y decisores, sino que también es erróneo universalizar ese principio para cualquier tripleta de opciones. Sea, por ejemplo, la siguiente situación: $A > B$, $B > C$, ¿ $A > C$? Si la alternativa A es el color rojo, la B el azul, y la C el amarillo, ¿qué impide preferir C antes que A, es decir, $C > A$? La situación en el caso cardinal parece clara, pero también en el caso ordinal, puesto que en la comparación pareada puede ocurrir que se comparen características diferentes de las alternativas, y así se califica como inconsistencia lo que sólo es una comparación de alternativas no homogéneas. En este caso la valoración multicriterio adquiere una importancia crucial, ya que múltiples elementos definen la decisión sobre una cuestión concreta. Y así como existen diferentes características que definen una opción, ¿por qué detener la heterogeneidad inherente a todo proceso de decisión, en todas sus dimensiones y en todos sus decisores? ¿Es creíble la ponderación independiente de diferentes elementos de una alternativa concreta: cuando estoy puntuando la característica medioambiental de un proyecto no me influye la percepción que tengo de las características económica y social, por ejemplo?

También existe otra situación que invalida el principio de transitividad, los comentados comportamientos estratégicos y no-honestos (fíjese el evidente juicio de valor que hay en esta etiqueta). De hecho, puede afirmarse que dado un bagaje de conocimientos inicial y supuesta la racionalidad, no hay inconsistencia en la emisión de juicios o en la estructura de preferencias.

4) Los juicios de valor inherentes están relacionados con la insuficiencia del comentado principio paretiano. Es decir, bajo su supuesta definición rigurosa de bienestar social, hay en realidad una apuesta por unos valores liberales muy concretos: La defensa y

alabanza del emprendedor, de la iniciativa privada, frente a la marginación y humillación del fracasado, del pobre. Otro principio, también liberal, pero más apropiado y consistente con el concepto de "elección social óptima" podría ser el principio rawlsiano (no considera que el bienestar social haya aumentado si no aumenta el bienestar de la persona en peor situación). Incrementos en el bienestar de las personas con menos recursos implican incrementos de bienestar social, mientras que el bienestar social disminuye aunque se produzcan aumentos en el bienestar de los individuos con más recursos. Obviamente, el mejor principio sería el principio no-liberal de Marx consistente en abolir la lucha de clases, donde las personas no son mercancía ni nadie puede ser explotado por el mero hecho de subsistir.

5) La dictadura de las alternativas se refiere a que el Teorema de Arrow, explícitamente, no admite la dictadura de ningún individuo. Nadie puede imponer su preferencia a otro. En cambio, implícitamente, la dictadura entra por la puerta de atrás a través de la imposición de las alternativas, camuflando la dictadura del individuo que diseña las alternativas. El principio de la mayoría es el principio de la dictadura de las alternativas, de las alternativas inamovibles. La mejor elección social debería contemplar la flexibilidad en la adaptación de las alternativas iniciales a las aportaciones de los individuos, de tal forma que la alternativa final puede no coincidir con la alternativa inicial debido precisamente al arco de flexibilidad en el que se ha movido la alternativa y también al propio arco de flexibilidad en el que se han movido los individuos. Un ejemplo clásico de amplitud máxima del arco de flexibilidad en el que se mueven los individuos lo encontramos en el argumento de la película "Doce hombres sin piedad". Otro ejemplo clásico de amplitud del arco de flexibilidad en el que se mueven las alternativas podría ser la concepción política del socialismo (razón por la cual el electorado de "izquierdas" suele estar más dividido que el de "derechas"). Un enfoque diferente a la problemática de las preferencias sería el basado en las necesidades (Guillén, 2003).

Racionalidad económica, marketing y conocimiento

Hay algunas cuestiones importantes en este sentido. Hay conocimientos que son incómodos, por lo que "el mejor argumento" puede ser rechazado "democráticamente" porque la sociedad no está preparada para asumirlo. Hay una hipótesis implícita que supone racionalidad (medible a través de niveles de consistencia, estabilidad, etc.) en la estructura de preferencias. La hipótesis clásica de racionalidad quizá sea excesiva, en particular porque existen otra serie de elementos, en muchas ocasiones ocultos e ignorados por el mismo participante, que afectan a la supuesta "racionalidad". Elementos inconscientes, intereses no revelados, deseos inconfesables, tabúes, etc. En esta cuestión resultaría de interés destacar cómo el análisis científico de las motivaciones humanas basadas en lo intangible, lo subjetivo y lo emocional (por ejemplo, en marketing) realmente no busca la generación de conocimiento sino el control y la manipulación de los deseos e instintos de los consumidores (Chomsky, 2003; Cortina y Carreras, 2004). Lo cual sólo nos confirma lo que ya se sabía desde las aportaciones de la psicología cognitiva, que la mayoría de las

decisiones se basan en argumentos no-racionales. En definitiva, sería preciso articular la terna "conocimiento-razonamiento-emoción", para evitar una posible brecha en las capacidades y posibilidades de autogobierno de las sociedades en favor de la clase dirigente en su afán de mantener el poder. Los mass media como "líderes de opinión" no son necesariamente los creadores y difusores de la "mejor opinión".

En un sistema democrático ideal deberían coincidir los "mejores argumentos" con los "argumentos más seguidos" e influyentes de los líderes sociales. Pero puede haber conflicto. ¿El argumento modal es necesariamente el mejor? ¿Cuándo el "argumento modal" es preferible al "mejor argumento"? ¿Qué condiciones satisface el "mejor argumento"? Sería preciso definir el "espacio ideológico" (individual y colectivo) en el que conviven "conocimiento" y "prejuicio". Ya que no todos los sistemas ideológicos digieren igual la misma información ni engendran el mismo conocimiento a partir de un conjunto común de recursos dados.

Etapas de análisis de la información

La toma eficiente de decisiones, en particular las de carácter político y económico, necesita en nuestros días de la asistencia de técnicas que minimicen los costes de esfuerzo y de tiempo invertidos en el proceso. En este sentido, las técnicas que tratan de explotar el contenido significativo de un conjunto de datos o de una ingente masa de información son cada vez más necesarias, especialmente en todo lo referente a internet, que ha supuesto una explosión de enormes cantidades de datos e informaciones, universalizando así el acceso directo a la información e indirecto al conocimiento. Cabe distinguir varias etapas de análisis de la información: 1) extracción de información, 2) generación de conocimiento y 3) difusión del conocimiento generado.

Entre las técnicas de la primera etapa (extracción de información) se encuentran las herramientas de explotación de datos (data mining), especialmente son de una relevancia singular las técnicas de explotación de textos (text mining), por su novedad y potencial en su apoyo en la toma de decisiones on-line en contextos democráticos. Su característica fundamental reside en su capacidad para generar y transmitir conocimiento relevante, algo que por sí solo, internet es incapaz de ofrecer. En definitiva se pone de manifiesto la importancia y necesidad del protagonismo de cada persona en su propio proceso de aprendizaje, en línea con las teorías del constructivismo cognitivo, lo cual está en consonancia con las metas y propósitos del nuevo Espacio Europeo de Educación Superior. Por tanto, la generación y gestión del conocimiento asistidas por internet en sistemas de autogobierno que minimizan la delegación de las decisiones importantes supone una importante y novedosa línea de investigación en el campo de la Decisión Multicriterio, que además tiene aplicaciones, aún no suficientemente exploradas, en múltiples campos y áreas científicas (economía, medicina, historia, ingeniería, biología, química, etc.). Por ejemplo, dentro de la Economía, en la gestión óptima de recursos naturales, en los análisis de género, en temas educativos, en estudios de marketing sobre motivaciones del consumo, etc.

Los antecedentes de los trabajos sobre "explotación de palabras" (Text Mining) son los trabajos pioneros sobre "recuperación de información" (Salton y McGill, 1983; Baeza-

Yates y Ribeiro-Neto, 1999), sobre "procesos del lenguaje natural" (Manning y Schutze, 1999) y sobre estadística, inteligencia artificial y teoría de la información (Pierce, 1980). La técnica de Text Mining se define, en términos generales, como cualquier tipo de procesamiento de textos que trata de encontrar, organizar y analizar la información (Konchady, 2006). O también como la creación de nueva información (patrón, tendencia o relación) que no es obvia o fácilmente visible mediante una lectura de documentos individuales (Hearst, 1999), entendiéndose por "documento" cualquier unidad de texto, tal como una página web, un email, un artículo, etc.

En definitiva, la técnica de Text Mining forma parte de un conjunto de técnicas más amplio conocido como Data Mining, cuyas primeras aplicaciones fueron en la planificación de campañas de marketing. Las principales diferencias entre ambas técnicas se refieren no sólo a la materia prima con la que trabajan, datos y texto, sino a la estructuración de los mismos. Así, el data mining trabaja con datos numéricos estructurados, y, por tanto, su efectividad depende de la existencia previa de una base de datos fiable. Mientras que Text Mining trabaja con texto no-estructurado, es decir, trata de construir un modelo a partir de datos alfanuméricos imprecisos. Un buen método de procesamiento de este conjunto de información verbal no-estructurada será aquel que utilice la información suficiente para construir un modelo general con la máxima potencia predictiva, ignorando los datos textuales que no pueden ser utilizados o son erróneos. Algunos manuales sobre Text Mining que implementan el uso de algún software son los de Berry y Castellanos (2007), Weiss et al. (2005) o Konchady (2006).

Las áreas de conocimiento que más producción científica están realizando en los últimos años mediante la aplicación de Text Mining son las biomédicas (incluyendo las químicas, farmacéuticas y neurociencias). Sin embargo, son menos frecuentes, y por ello, más novedosos, los estudios sobre temas sociopolíticos, como es el análisis de la creación y gestión del conocimiento en sistemas democráticos apoyados en las nuevas Tecnologías de la Información y la Comunicación. Además de estudios comparativos de diversas metodologías orientadas a la toma de decisiones en situaciones de incertidumbre (Beynon y Peel, 2001; Chang y Wang, 2006; Prinzie y Van den Poel, 2008), como Text Mining, análisis Delphi, modelos logit y probit, etc.

Las aplicaciones de Text Mining se han llevado a cabo en el campo empresarial mediante la realización de encuestas a los empleados para recabar información sobre una determinada política de recursos humanos o de reparto de beneficios. O mediante el análisis de las páginas web de las empresas competidoras con el fin de identificar y planificar estrategias de inversión o conocer las prioridades financieras del sector. También se han realizado aplicaciones en el terreno de la protección de patentes o la gestión del consumidor o la contratación de personal mediante la red. Otras aplicaciones se han desarrollado en el campo de la investigación médica y legal para la gestión de enormes bases de datos electrónicas. Finalmente, también se han realizado aplicaciones en temas sociales relacionados con la opinión pública, el comercio electrónico, la investigación académica o la jerarquización de ensayos científicos en función de su calidad.

METODOLOGÍA

A continuación se comenta la metodología seguida en el meta-análisis realizado en este trabajo. Se realiza una revisión de la literatura sobre artículos indexados en la base de datos Science Direct que aplican técnicas de minería de textos, con el propósito de estudiar su senda temporal en el marco clásico de adopción y difusión de innovaciones. Así, durante el mes de abril de 2007 se accedió a la base de datos de Science Direct y se seleccionaron un total de 153 artículos desde 1997 hasta 2007 sobre text mining, que involucran a más de 400 investigadores, entre los cuales se ha detectado una veintena que ha participado en varios artículos. Se realizan varias estimaciones no lineales del proceso sigmoidal de difusión agregada de esta muestra de investigaciones, así como predicciones hasta 2017 del proceso de difusión y su horizonte de estabilización. Todas las estimaciones se han realizado con el paquete estadístico SPSS v.15.

Los modelos de difusión temporal describen la senda de adopción acumulada de una determinada tecnología o innovación, siendo también válidos para propósitos predictivos. Estos modelos analizan, por un lado, la penetración de una innovación en un sistema social a lo largo del tiempo; y por otro, el grado de saturación o nivel máximo de adopción. En nuestro caso la adopción se refiere al hecho de que los investigadores adoptan la decisión de investigar un tema nuevo (la extracción de información mediante minería de textos). La curva de difusión logística (forma de S simétrica) se ajusta bien a procesos de difusión basados en la imitación o efecto contagio, es decir, en el contacto entre los potenciales adoptantes del sistema en que se difunde. La información que se transmite entre los miembros del sistema social es de carácter interno, por lo que la difusión se produce por acumulación de información y experiencia, reduciéndose progresivamente la incertidumbre inicial, a medida que los nuevos adoptantes "contagian" a los potenciales adoptantes. El punto de inflexión en la especificación logística ocurre cuando se alcanza el 50% acumulado de adoptantes. Otro modelo de influencia interna habitual en procesos de difusión de innovaciones es el modelo Gompertz, esta función es asimétrica y el punto de inflexión ocurre antes que en la curva logística. Ambas aproximaciones, logística y Gompertz, se adaptan bien al análisis de innovaciones en contextos sociales homogéneos donde el efecto de imitación es decisivo, puesto que implica que la decisión de no adoptar produce desventajas con respecto a los adoptantes. Por otra parte, en contextos donde la adopción previa tiene una importancia mínima, donde la innovación es sencilla y no requiere etapas iniciales de aprendizaje, la difusión de innovaciones basada en información externa al sistema donde se propaga la innovación, sin que exista comunicación entre los miembros del sistema, se modeliza mejor a través de la función exponencial, muy utilizada en marketing.

La curva de difusión exponencial tiene forma cóncava, con crecimiento decreciente, con asíntota superior y sin punto de inflexión. Como puede observarse, los modelos de influencia interna parecen más apropiados para el estudio de la difusión de investigaciones aplicadas sobre text mining entre los investigadores de diversas disciplinas.

Se han calculado los modelos permitiendo que los coeficientes y el techo de adopción sean variables y se ajusten por sí

mismos a la realidad, estimándose por el método de Levenberg-Marquardt las regresiones no lineales sin restricciones, dado que aún no se ha alcanzado el techo o nivel de saturación, por lo que a partir de 2007 las estimaciones realizadas nos proporcionan una predicción hasta el momento en que el proceso alcanza su nivel de estabilización. El nivel de convergencia de la suma de los cuadrados y de los parámetros se fijó en 10-8. Las estimaciones de los parámetros son significativas al 95% de confianza.

RESULTADOS Y DISCUSIÓN

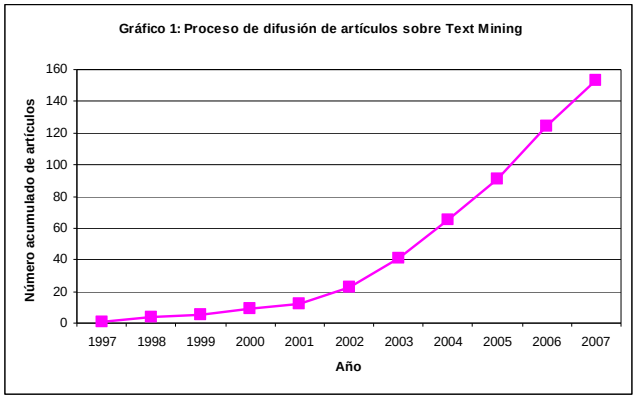
Se analizó el proceso temporal de difusión desde una óptica doble, por un lado, desde la producción científica (número acumulado de artículos); y por otro lado, desde los autores de los trabajos revisados (número acumulado de investigadores). La Tabla 1 muestra la distribución absoluta y porcentual del número de artículos y autores según disciplinas científicas. Se observa cómo el área de conocimiento que más ha trabajado en temas relacionados con la extracción de información mediante minería de textos desde finales de la década de los 90 hasta nuestros días ha sido la biomédica. Además, en esta área se observa un fenómeno característico del comportamiento investigador habitual entre los científicos de esta disciplina, ya que en comparación con otras áreas y en términos relativos, la medicina agrupa más número de autores que artículos. Sucede al contrario en las áreas de ingeniería y tecnología. El número medio total de autores por artículo es de 2,74, siendo de 3,28 autores por cada publicación médica, y de 1,88 por cada artículo en ingeniería.

En el grupo de "otros" se encuentran la mayoría de artículos publicados en la revista "Expert Systems with Applications", y también artículos sobre text mining con desarrollos matemáticos, y aplicaciones en geología, turismo, contabilidad y patentes. Destaca la ausencia de investigaciones en ramas de ciencias sociales como la economía, la sociología, y particularmente en especialidades interdisciplinares como puede ser la toma de decisiones en contextos democráticos.

TABLA 1 - Clasificación de los artículos revisados según disciplina

Disciplinas	Artículos	%	Autores*	%
Medicina y neurociencia	36	23,53	118	28,10
Información	21	13,73	54	12,86
Informática	17	11,11	49	11,67
Tecnología	10	6,54	20	4,76
Biología	9	5,88	28	6,67
Ingeniería	8	5,23	15	3,57
Química	8	5,23	25	5,95
Decisión	7	4,58	20	4,76
Estadística	3	1,96	6	1,43
Otros	34	22,22	85	20,24
TOTAL	153	100	420	100

(*): Se han excluido las repeticiones
FUENTE: Elaboración propia a partir de Science Direct.



FUENTE: Elaboración propia

FUENTE: Elaboración propia

Los gráficos 1 y 2 muestran el proceso de difusión seguido desde 1997 a 2007 por el número de investigaciones e investigadores sobre minería de textos indexadas en la base de datos Science Direct. La fase inicial antes del despegue del proceso de difusión abarca desde 1997 hasta 2001, tanto para investigaciones como para investigadores. Se observa cómo la mitad de las publicaciones en el periodo analizado se alcanzaron antes de llegar al año 2005, en menos de 8 años, momento que coincide con la etapa de expansión del sendero de difusión.

La Tabla 2 recoge los tres modelos estimados para cada nivel de análisis considerado (artículos y autores). Los mejores ajustes los proporcionan las especificaciones logística y Gompertz dado el elevado coeficiente de determinación, lo que significa que el "efecto imitación o contagio" es el motor principal en la decisión de los investigadores de estudiar temas relacionados con la extracción de información en sus áreas de conocimiento. En contraposición, el ajuste del modelo exponencial es menor, lo que indica que la decisión de investigar sobre text mining no se debe a elementos externos al sistema social de los investigadores, lo cual explica en parte lo que antes comentábamos sobre la escasez de estudios sobre minería de textos en ciencias sociales, y particularmente, sobre la toma democrática de decisiones. Predomina el comportamiento imitador en la investigación de estos temas y no el comportamiento innovador.

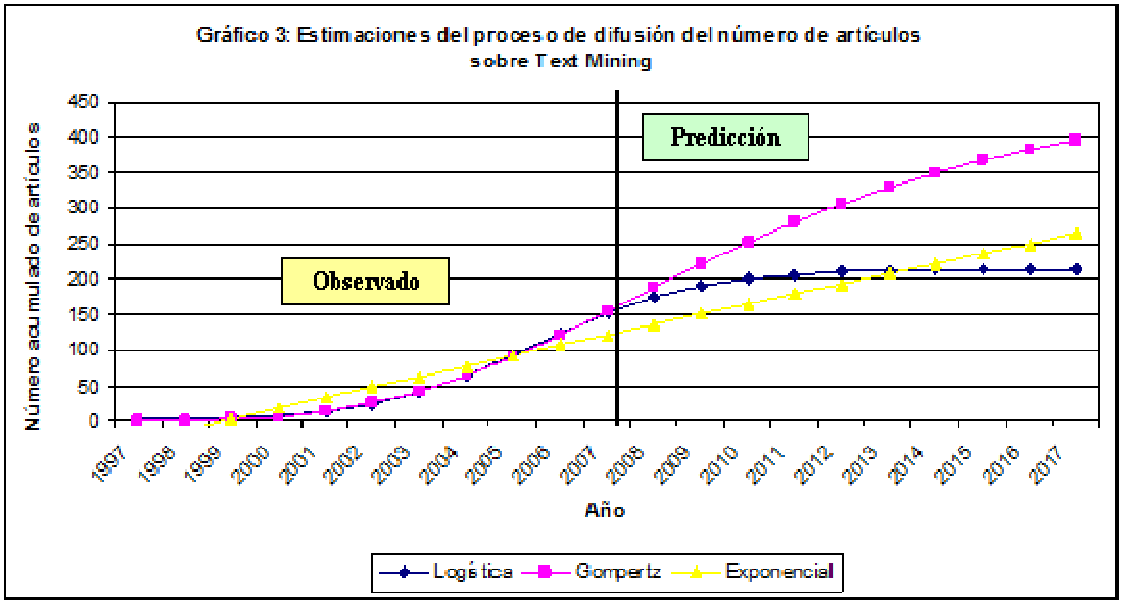
TABLA 2 - Parámetros estimados de los modelos de difusión

Artículos	**Especificación funcional de los modelos**		
	M/(1+exp(a-b*T))	**M*exp(-exp(a-b*T))**	**M-exp(a-b*T)**
	Logístico	**Gompertz**	**Exponencial**
M (nivel de saturación)	214,003 (8,203)	462,593 (91,218)	3740,291 (94117,528)
a (constante)	5,627 (0,110)	2,244 (0,083)	8,238 (25,117)
b (tasa de difusión)	0,595 (0,019)	0,196 (0,023)	0,004 (0,104)
R² (bondad de ajuste)	0,999	0,999	0,855
Autores	**Logístico**	**Gompertz**	**Exponencial**
M (nivel de saturación)	569,031 (26,567)	1110,118 (210,450)	15460,925 (583308,46)
a (constante)	5,791 (0,159)	2,333 (0,110)	9,654 (37,726)
b (tasa de difusión)	0,621 (0,027)	0,216 (0,027)	0,003 (0,104)
R² (bondad de ajuste)	0,999	0,998	0,856
Recorrido (años)	18,650	18,054	-

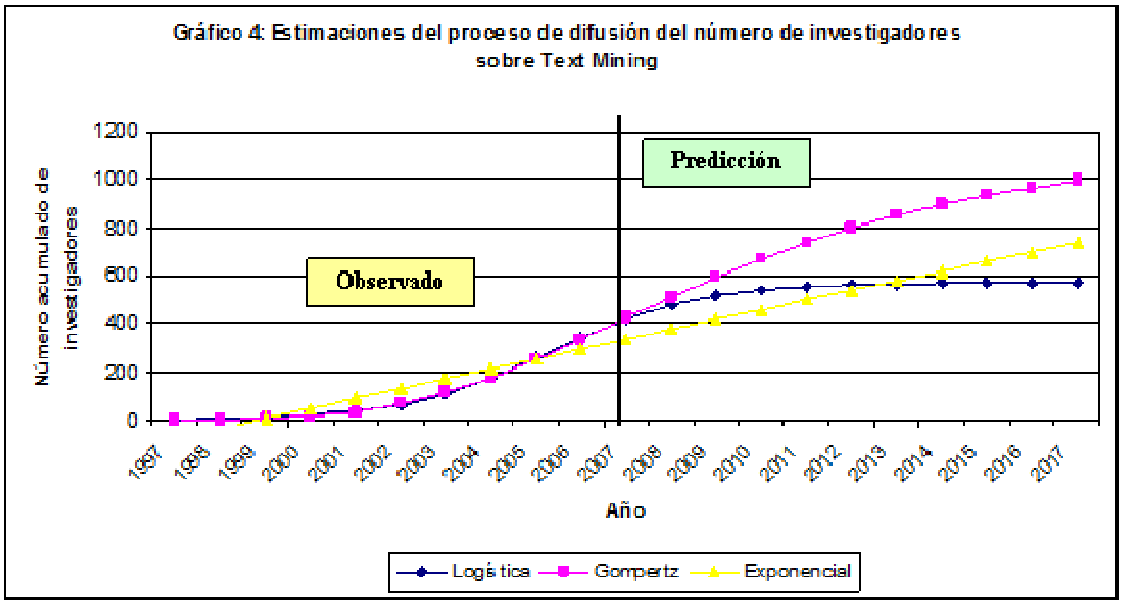
Nota: Estimación de parámetros significativa al 95% de confianza.
Erro típico de los parámetros ***M***, ***a*** y ***b*** entre paréntesis.

Sin embargo, la decisión de estudiar la información mediante la aplicación de minería de textos en las ciencias sociales es un tema más complejo, transversal e interdisciplinar, que necesita desarrollos previos en otras áreas científicas (como la estadística, la informática, la ingeniería) e incluso aplicaciones previas en sistemas sociales acotados de alguna manera (por ejemplo, en medicina).

Los gráficos 3 y 4 muestran la senda temporal estimada según los modelos logístico, Gompertz y exponencial para la muestra de investigaciones e investigadores sobre text mining. Observándose cómo el nivel máximo de saturación es superior en la estimación Gompertz que en la logística. Entre ambas especificaciones, el modelo Gompertz parece el más adecuado dado que el techo máximo es superior que en el modelo logístico, lo cual es debido a que la velocidad de adopción es inferior en el modelo Gompertz lo que implica una dilatación mayor en el tiempo del proceso de difusión, y por tanto, la posibilidad de alcanzar un nivel de saturación mayor. Para contrastar la idoneidad de ambos modelos se ha revisado el número total de nuevos artículos publicados en la base de datos, Science Direct, de la que se ha extraído la muestra.



FUENTE: Elaboración propia



FUENTE: Elaboración propia

Así entre 2007 y 2011 el número de artículos supera los 1.300, que en valor acumulado se encuentra en el intervalo de predicción para la muestra estimada que representa entre el 14% y 16% de artículos totales. Para 2017 se estima que se habrán publicado entre 2.800 y 3.300 artículos según la predicción Gompertz para un intervalo entre 12% y 14% a partir de la estimación muestral, situada entorno a 400 artículos según se ve en el Gráfico 3. Si establecemos que el porcentaje de representación de la muestra en 2017 será entre el 7% y 10%, supondría un intervalo que oscilaría entre 4.000 y 5.700 artículos totales acumulados desde 1997. Conforme disminuya el porcentaje de representación del valor muestral con respecto al valor real, mayor será el volumen total de artículos.

CONCLUSIONES

El estudio realizado muestra la importancia de replantear los conceptos tradicionales de análisis del comportamiento individual, especialmente en lo que se refiere a la toma de decisiones democráticas. En este sentido se ha definido un marco teórico amplio, apoyado en los principales resultados y debates sobre el principio de Pareto, el Teorema de Arrow y la búsqueda del máximo bienestar social. En este marco teórico juega un papel crucial el análisis de la información para generar conocimiento sensible orientado a la construcción democrática del bien común. Una de las principales técnicas para extraer información es la minería de textos.

Se ha encontrado que los artículos indexados en Science Direct que publican investigaciones sobre Text Mining siguen una pauta de difusión típica de sociedades homogéneas donde el efecto imitación es decisivo, puesto que implica que la decisión de no adoptar la innovación produce desventajas con respecto a los adoptantes.

En esta línea, extrapolando a la sociedad en general, puede concluirse que en el contexto actual de crisis económica europea y de triunfo del neoliberalismo, la desaparición de la clase media y, por consiguiente, la polarización y homogeneización de la sociedad provoca las condiciones para la construcción de sistemas con mayor calidad democrática, lo que indudablemente pone en peligro el poder establecido y afirma la predicción marxista según la cual las propias contradicciones del capitalismo anuncian su final.

REFERÊNCIAS

ARROW, K. A Difficulty in the concept of social welfare. The Journal of Political Economy, v.58, n. 4, p. 328-346, 1950.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern information retrieval. ACM Press Books, 1999.

BERRY, M. W.; CASTELLANOS, M. Survey of text mining: clustering, classification and retrieval. Springer. New York, 2007.

BEYNON, M. J.; PEEL, M. J. Variable precision rough set theory and data discretisation: an application to corporate failure prediction. Omega v. 29, p. 561–576, 2001.

CHANG, L.; WANG, H. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis and Prevention, v. 38, p.1019–1027, 2006.

CHOMSKY, N. El beneficio es lo que cuenta. Neoliberalismo y orden global. Ediciones VSV. Madrid, 2003.

CORTINA, A.; CARRERAS, I. Compro… luego existo. Ed. CyJ. Barcelona, 2004.

FRANCO, J. A. Principles of econometrics from the giffen demand. Technological and Economic Development of Economy, 1v. 19, Supplement 1, p. 144–163, 2013.

GUILLÉN, M. Hacia una revisión crítica del análisis neoclásico del consumo: una alternativa basada en las necesidades. Revista de Economía Crítica v.1, p. 95-111, 2003.

HEARST, M.A. Untangling text data mining. Proceedings of the ACL´99, University of Maryland, 1999.

KONCHADY, M. Text mining application programming. Charles River Media y Thomson Delmar Learning. Boston, Massachusetts, 2006.

MANNING, C.D.; SCHUTZE, H. Foundations of statistical natural language processing. MIT Press. Massachusetts, 1999.

NAVARRO, V. La ideología que reproducen las ciencias económicas. Diario Público, Acesso em jan/2014. Disponible en <www.vnavarro.es>PIERCE, J.R. An introduction to information theory symbols, signals and noise. Dover Publications. Suffolk, UK, 1980.

PRINZIE, A.; VAN DEN POEL, D. Random Forests for multiclass classification: Random Multinomial Logit. Expert Systems with Applications, v.34, n.3, p.1721-1732, 2008.

SALTON, G.; MCGILL, M.J. Introduction to modern information retrieval. McGraw-Hill. New York, 1983.

TERÁN, P. Teoremas de aproximación y convergencia para funciones y conjuntos aleatorios. Tesis doctoral. Universidad de Oviedo, 2002.

TORRES, J. Manual de economía política. Pirámide. Madrid, 1999.

WEISS, S.M.; INDURKHYA, N.; ZHANG, T.; DAMERAU, F.J. Text mining: predictive methods for analyzing unstructured information. Springer. New York, 2005.

NOTAS

(1) Doutorado em Economia Agrária, Córdoba. Pós-Graduação em Economia Aplicada, Granada. Graduação em Economía, Badajoz. Universidad de Extremadura. Professor. Avenida Universidad, s/n. 10003, Cáceres, Espanha. E-mail: franco@unex.es.

Enviado: 20/06/2009
Aceito: 27/01/2014
Publicado: 27/02/2014