ISSN: 1138-9761 / www.cac.cat

QUADERNS
DEL CAC

Ultrapasando Babel: mediación social y sistemas inteligentes en el descubrimiento, filtraje, acreditación y personalización de contenidos digitales

Enric Plaza

Jefe del Departamento de Sistemas de Aprendizaje del Instituto de Investigación en Inteligencia Artificial (IIIA-CSIC)

enric@iiia.csic.es

Resumen

La convergencia de contenidos digitales transforma el modelo de distribución: de un modelo de difusión centralizado de contenidos a un modelo de comunicación reticular, más simétrico. Dicha transformación también afecta a la elaboración de contenidos, que está al alcance de cualquier ciudadano con un ordenador y conexión a internet. La llamada objeción Babel *critica ese efecto democratizador. En el presente artículo analizamos los distintos procesos de mediación que relacionan los contenidos con los destinatarios y que están presentes tanto en el modelo de difusión centralizado como en el de comunicación reticular. El objetivo del artículo es mostrar que es viable el desarrollo de procesos de descubrimiento, filtraje, acreditación y personalización en un modelo de comunicación reticular donde los consumidores son también contribuyentes.*

Palabras clave

Inteligencia artificial, personalización, búsqueda, mediación.

Abstract

The convergence of digital content is transforming the distribution model from the centralised distribution of content to a more symmetrical model of network communication. This transformation also affects the production of content, this now being within the scope of any citizen with a computer and internet connection. The so-called Babel objection criticises this democratising effect. In this article we analyse the different mediation processes that relate content with recipients that are present both in the centralised distribution model as well as in that of network communication. The aim of this article is to show that it is viable to develop the discovery, filtering, accreditation and personalisation processes of a network communication model where consumers are also contributors.

Key words

Artificial intelligence, personalization, research, mediation.

Introducción

La convergencia de contenidos digitales transforma el modelo de distribución: de un modelo de difusión centralizado de contenidos (de pocos centros a muchos usuarios) a un modelo de comunicación reticular (de muchos a muchos). Esa transformación también afecta a la elaboración de contenidos, que está al alcance de cualquier ciudadano con un ordenador y conexión a internet. El modelo de comunicación reticular es, en principio, simétrico, en el sentido de que cualquier nodo puede ser, a su vez, consumidor y creador de contenido, ya sean datos, información, conocimientos o cultura. Ese efecto democratizador ha sido criticado con la llamada *objeción Babel*: si todo el mundo puede hablar, nadie podrá escuchar a causa de la resultante cacofonía (la sobrecarga informacional). Si la objeción Babel es cierta, la democratización fracasará y los ciudadanos de la red dejarán de ser contribuyentes activos para pasar a ser a consumidores pasivos. Si puede organizarse un esquema que permita relacionar eficientemente y cómodamente los contenidos y sus destinatarios, podremos ultrapasar la objeción Babel.

En el presente artículo analizamos los distintos procesos de mediación que relacionan los contenidos con sus destinatarios, es decir, el descubrimiento, filtraje, acreditación y personalización. Dichos procesos están presentes tanto en el modelo de difusión centralizado como en el de comunicación reticular, que únicamente añade una dificultad cuantitativa al desarrollo de esos procesos. El objetivo del artículo es mostrar que es viable el desarrollo de procesos de descubrimiento, filtraje, acreditación y personalización en un modelo de comunicación reticular en el que los consumidores son también contribuyentes. En particular, analizaremos dos elementos básicos: *a)* los contenidos informacionales proporcionados por los propios contribuyentes sobre los procesos de mediación y *b)* el uso de técnicas de inteligencia artificial en la gestión de una gran cantidad de datos en los procesos de descubrimiento, filtraje, acreditación y personalización.

La simetría reticular y la propiedad de los medios materiales de producción y distribución

La conmoción que conlleva cualquier cambio de paradigma —presentemente la transformación de un modelo de difusión (de pocos a muchos) a un modelo de comunicación reticular (de muchos a muchos)— hace surgir dos tipos de respuestas antagónicas: la de los apocalípticos y la de los integrados. Umberto Eco (1964) caracterizó ambas tesis antagónicas (los apocalípticos y los integrados) respecto a los *mass media* de los sesenta; hoy en día podemos detectar unas respuestas parecidas. Por una parte, la de los apocalípticos/reaccionarios, que sólo encuentran problemas en el nuevo paradigma de la información en la red: cacofonía, sobrecarga informacional, falta de credibilidad, etc. Por otra, la de los integrados/revolucionarios, que sólo destacan las posibilidades positivas: mejor acceso a la información, democratización del proceso de distribución de información, más capacidad de crítica/monitorización de actuaciones de los grupos establecidos, facilidad de coordinación de un gran número de personas, etc.

La respuesta no es el feliz punto medio, sino la aceptación de que existen aspectos negativos y positivos, y el análisis de cómo podemos ayudar, y con qué mecanismos, a alcanzar las posibilidades positivas y a amortiguar los efectos negativos. Es en ese sentido que la tecnología no es neutral, como tampoco lo es la legislación que restringe las posibles opciones: los mecanismos utilizados pueden hundir algunas de las posibilidades positivas o mantener algunos de los efectos más negativos.

Por ello es preciso analizar, en primer lugar, los efectos del cambio tecnológico no sólo en los ámbitos sociales y de costumbres, sino también en el económico y productivo. Desde el punto de vista más abstracto, el cambio de paradigma da lugar a un medio más similar a la red telefónica (donde todo el mundo puede comunicarse con todo el mundo) que al modelo basado en empresas editoriales/emisoras de contenido. La simetría es una característica de la estructura reticular: todos los nodos son miembros iguales de la red, todos reciben y transmiten contenido. Esa simetría también se encuentra en la red de redes, internet, pero no es suficiente para explicar el cambio de paradigma. El segundo factor es el ordenador personal, que, a diferencia del teléfono, es un medio de creación, elaboración y producción de contenidos (ya sean datos, información, conocimientos o cultura) y, sobre todo, un medio de producción altamente descentralizado, es decir, propiedad de ciudadanos individuales y no de empresas o del Estado.

Es la conjunción del medio de producción digital (el ordenador) y de la infraestructura de distribución digital (internet) en un esquema de propiedad descentralizada lo que transforma la economía política de una economía industrial de la información en una *networked information economy*, una economía reticular de la información (Benkler 2006). Un ejemplo histórico del cambio económico es el coste de la creación de diarios a inicios la era de la economía industrial. Según Benkler (2006), lanzar un nuevo diario en Estados Unidos durante los años 1835-1850, costaba al principio 10.000 dólares (en dólares actuales), un coste que llegó a los 2,5 millones de dólares (en dólares actuales). Ese brutal cambio de costes aniquiló un ecosistema de pequeños diarios con distintos tipos de organización y financiación (con una circulación semanal superior en Europa en unos EE.UU. de sólo 17 millones de habitantes).

En nuestra experiencia, vivida bajo la economía industrial de la información, parece que las dos únicas alternativas a la producción de contenido son las (grandes) empresas basadas en el mercado y las empresas estatales; nos resulta difícil imaginar alternativas "serias" fuera de esos dos modelos. A pesar de ello, la propiedad y los costes económicos de producción y difusión de contenidos han disminuido extraordinariamente (ordenadores más conexión en red). Es lo que Yochai Benkler (2006) llama el "modo social de producción", que se añade a las organizaciones basadas en el mercado o el Estado. Por lo tanto, el ecosistema de creación, elaboración y producción de contenidos que cabe esperar en el futuro próximo estará, en comparación con el industrial, mucho más descentralizado.

Descubrir y filtrar

El hecho de encontrar nuevos contenidos se ha hecho, desde siempre, de modo "formal" con guías y catálogos, pero también de modo "informal" usando las redes sociales: un amigo o conocido os comenta que tal programa de radio emite un tipo de música que os puede gustar. Internet ha añadido la posibilidad proactiva de que una persona use buscadores web (como Google) para encontrar nuevos contenidos. Cabe decir que la primera propuesta para descubrir contenidos fue "formal" y desarrollada por Yahoo, intentando realizar una guía/catálogo de las páginas web. Dicho catálogo se hacía manualmente, algo que no era escalable a causa del gran número de páginas existentes.

La alternativa fue usar buscadores web, aplicaciones basadas en técnicas de recuperación de la información que se adaptaron al análisis, la indexación y la recuperación de páginas web, por ejemplo Aliweb en 1993 y Altavista en 1995. Hoy en día Google es el buscador más popular, pero debemos analizar la razón tecnológica de su éxito: el análisis y el uso del contenido aportado por el usuario (CAU). La idea central del algoritmo PageRank usado por Google se basa en el análisis de un contenido particular aportado por el usuario: los hiperenlaces que relacionan dos páginas web. En efecto, el usuario declara que (el contenido de) la página que escribe se relaciona con (el contenido de) las páginas con las que enlaza. PageRank analiza la red de relaciones aportadas por los usuarios como enlaces con el fin de asignar a cada página P un grado de importancia determinado por (la importancia de) las páginas $P_1 \ldots P_n$ que apuntan a la página P. Dicho algoritmo se basa en anteriores trabajos realizados en bibliometría sobre el análisis de citaciones; la innovación de PageRank es que se centra en el análisis y la explotación de un tipo concreto de CAU, los hiperenlaces,

con el fin de filtrar o distinguir el contenido más "importante" del que lo es menos.

Las técnicas de inteligencia artificial pueden mejorar los procesos de descubrimiento y filtraje en el marco de la llamada Web Semántica. La Web Semántica, propuesta por Tim Bernes-Lee, el creador de la primera página web, se basa en la "anotación" de los contenidos web usando los términos de una ontología, de modo que el contenido producido por los humanos pueda ser entendido por sistemas inteligentes automáticos. Sin embargo, esa nueva tecnología web es "sectorial": cada sector requiere de una ontología propia (una descripción formal del significado de los términos que se usan en ese sector). Por ejemplo, los contenidos de cariz legal dispondrían de una ontología legal donde se definirían términos como *fraude*, mientras que los contenidos de cariz médico necesitan una ontología médica.[1] Respecto a los contenidos multimedia, la ontología musical (<http://musicontology.com>) es la más desarrollada actualmente y la BBC ha empezado a aplicarla en su sitio web.

Otra forma de mejorar el descubrimiento y filtraje es el análisis del comportamiento de comunidades de usuarios cuando realizan búsquedas y aprender a realizar un filtraje más inteligente que permita averiguar qué contenidos son realmente interesantes para esa comunidad. El University College Dublin trabaja en ese sentido: en vez de desarrollar una ontología para cada tema, el sistema aprende observando lo que hacen grupos de usuarios interesados en el fútbol, la fotografía o los iPod. Las técnicas utilizadas son parecidas a las de los sistemas de recomendación, como los sencillos pero conocidos sistemas para recomendar libros en Amazon o música en el AppleStore. El análisis de las acciones de los usuarios, en el descubrimiento y la selección de lo que es de su interés, permite un resultado mucho más personalizado para cada usuario.

Acreditación y personalización

Mientras que el descubrimiento y el filtraje se ocupan principalmente de la relevancia de ciertos contenidos respecto al usuario, también es importante una segunda dimensión, la credibilidad de los contenidos y la reputación de sus orígenes (o "fuentes", como suele traducirse *sources*). Seguramente la supuesta "falta de acreditación" de los contenidos, además de la gran cantidad existente, es uno de los factores con más peso en la opinión pesimista respecto a la hipótesis Babel. Ese pesimismo sobre la posibilidad de un mecanismo descentralizado y eficiente para la distribución de contenidos viene dado por el modelo establecido por los grandes *mass media*, en que esas grandes organizaciones consideran que su papel es el de jerarquizar los contenidos, por ejemplo qué contenidos son de primera página y qué contenidos tienen poco o nulo espacio. En ese modelo, la multiplicidad de organizaciones ofrece, a su vez, diversidad de jerarquizaciones y acreditación de contenidos (a partir de la reputación de las organizaciones). Sin embargo, la crítica a la actual situación es clara: el número de organizaciones *mass media* es reducido para garantizar la diversidad, y los contenidos a menudo se publican sin contrastarlos demasiado con la realidad en razón de la inmediatez.

Desde el punto de vista del ciudadano y el usuario, la acreditación proporcionada por los *mass media* es bastante relativa: hay gente que confiará en ciertas organizaciones y no en otras. Esa confianza se debe a los modelos de reputación que la gente se hace de organizaciones y personas concretas. Para ultrapasar Babel, por lo tanto, son precisos la creación y el mantenimiento de sistemas de valoración de la reputación de los autores/distribuidores de contenidos a través de mecanismos descentralizados que sustituyan los mecanismos jerárquicos de las organizaciones de *mass media*.

Dado que la reputación y acreditación social son también bienes informacionales, ambos pueden tratarse como cualquier otro contenido. Por lo tanto, la reputación y la acreditación social pueden ser creadas de una forma descentralizada por los propios usuarios/productores/consumidores (CAU). De hecho, conocemos el ejemplo del sitio web Slashdot (<http://slashdot.org>), que permite hacer exactamente eso y se ha convertido, hoy por hoy, en uno de los principales boletines de noticias tecnológicas (*News for Nerds*). El principio de funcionamiento es muy sencillo: los usuarios aportan la URL de una noticia o contenido en general y añaden un comentario sobre su interés. Los otros usuarios añaden comentarios, que a menudo llegan a centenares. Slashdot utiliza la revisión entre iguales *ex post* para evaluar la credibilidad o calidad de los comentarios. Ese método es una variación del sistema de publicación científica (la revisión entre iguales previa a la publicación), en que la revisión se realiza *a posteriori*.

Slashdot no intenta evitar que se publiquen contenidos iracundos o falsos, sino que tan sólo facilita su contrastación con elementos que los corroboren o los desmientan. Los usuarios habituales suman "puntos de karma" por su buena actuación (o les son sustraídos por una mala actuación). Así se crea de forma neutral y automática un mecanismo de reputación que ayuda a ponderar a los usuarios en posiciones conflictivas. El resultado es una ordenación de los contenidos, es decir, una jerarquización, que ha sido producida, sin embargo, de una forma descentralizada por la propia comunidad de los interesados en noticias y contenidos tecnológicos. Actualmente se lleva a cabo investigación en modelos de reputación más sofisticados en nuestro Instituto de Investigación en Inteligencia Artificial (IIIA), entre otros, con el objetivo de crear plataformas de acreditación de gran alcance.

Finalmente, la personalización se caracteriza por ser un proceso que pone en correspondencia ciertos contenidos con las afinidades (de intereses o gustos) de un usuario. Una de las técnicas más usadas es el filtraje colaborativo, usado, por ejemplo, por Amazon para la recomendación de libros, películas y, como realiza también AppleStore, música. El filtraje colaborativo realiza una predicción sobre los elementos que pueden ser más afines a una persona, comparando los elementos que

son afines a otras personas "parecidas". La forma de determinar que dos personas son parecidas puede variar, pero esencialmente se compara la conducta registrada de los usuarios (en el caso de Amazon o de AppleStore, qué elementos compra cada persona). Aparte de esa técnica, actualmente se lleva a cabo bastante investigación para el desarrollo de sistemas de recomendación más depurados. Por ejemplo, una compañía *spin-off* del IIIA, MyStrands (<http://www.MyStrands.com>), desarrolla tecnologías sociales de recomendación, particularmente en el mundo de la música. Los sistemas de recomendación y personalización son un campo nuevo y muy activo dentro de la inteligencia artificial, cuyo primer congreso internacional se celebró en 2007, y es probable que a corto plazo se consoliden como una tecnología tan ubicua como lo es ahora la búsqueda de contenidos

Conclusiones

Los procesos de descentralización y automatización que actúan sobre el descubrimiento, el filtraje, la acreditación y la personalización de contenidos tendrán a buen seguro consecuencias que no podemos prever, pero para finalizar querría mencionar la importancia del fenómeno llamado "la cola larga". El término *The Long Tail* fue creado por Chris Anderson (2006) para argumentar que, en la nueva estructura de costes de internet, los productos con pocos clientes o ventas conjuntamente pueden llegar a un volumen de mercado superior a los productos con más clientes o ventas. Esas curvas se conocen en estadística como colas de Pareto, pero a menudo se llaman curvas 80/20 de distribución de ventas de un catálogo, en que el 20% de los productos suma el 80% de las ventas y "la cola" es el 80% restante de productos, que suma el 20% de las ventas. Estudios actuales muestran que en internet esta curva se transforma en 72/28, un cambio considerable a efectos prácticos. Así, por ejemplo, Amazon puede tener un catálogo amplio que incluye muchos productos con poca salida, artículos "de nicho", pero que en conjunto generan buena parte del negocio.

Eso viene al caso por el hecho de que la llamada "fragmentación" de los contenidos es un fenómeno que seguirá amplificándose a causa del efecto cola larga: cada vez se crearán más contenidos por "nichos", es decir, por mercados que no son de masas. Actualmente ya se produce la transición de los *mass media* a una miríada de servicios y contenidos dirigidos a grupos de interés de tamaño medio o pequeño, y seguirá produciéndose por la acción de las nuevas tecnologías y estructuras de costes. Los apocalípticos pueden temer a Babel, pero he intentado mostrar que hay ideas y técnicas que podrán organizar la nueva galaxia de internet de una nueva forma, descentralizada y social. Sin embargo, los usos y costumbres cambiarán, y eso, no puede negarse, producirá desazón. Personalmente, creo que la nostalgia de los tiempos en los que todos veíamos la misma película en la única tele es un error.

Nota

1 Un ejemplo del uso de ontologías para la búsqueda puede verse en <http://www.cognition.com>.

Bibliografia

Benkler, Y. *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. Yale: Yale University Press, 2006.

Eco, U. *Apocalittici e Integrati*. Milán: Bompiani, 1964.

Anderson, C. *The Long Tail: Why the Future of Business is Selling Less of More*. Nueva York: Hyperion, 2006.