

La web semántica: una visión crítica¹

Por Lluís Codina

DADO QUE UNO DE LOS TÉRMINOS DE MODA para los próximos años en relación con internet va ser la web semántica, en este pequeño texto nos proponemos 3 cosas: exponer las razones que la justifican y orientan, informar sobre la infraestructura que se supone que la hará posible y dar nuestra propia interpretación sobre sus posibilidades reales a corto y medio plazo.

Qué pueden hacer los ordenadores

La web semántica (o *semantic web*) es, de momento, el nombre de una aspiración y el de un objetivo muy ambicioso que, de cumplirse, cambiaría de forma radical la web tal como la conocemos hoy. ¿En qué consiste esta aspiración? Ni más ni menos que en conseguir que las páginas que la forman dejen de ser simples cadenas de caracteres para los ordenadores y se conviertan en textos con sentido, es decir, con semántica tal como, de hecho, lo son para los seres humanos.

«Los metadatos son información sobre la información y son, en realidad, una antigua fórmula. Los catálogos de las bibliotecas son metadatos»

¿Por qué un objetivo semejante? Tal como se codifican las páginas actuales, principalmente mediante html, tienen muy poco sentido para las máquinas. En efecto, si vemos sus códigos fuentes actuales encontramos, por ejemplo, algo como lo siguiente:

...
<i>Cómo conseguir la paz mundial</i>
...

cuando el ordenador lo interprete, a través del programa navegador, aparecerá como un texto en negrita y cursiva, como éste:

...

Cómo conseguir la paz mundial

... Con esto casi se acaba todo lo que es capaz de hacer un ordenador con las páginas html. Otra cosa que pueden hacer es construir índices con las palabras que aparecen en las páginas web. Después, cuando alguien envía una pregunta a un motor de búsqueda, lo que hace este último es comparar las palabras de la pregunta con las de su índice. Por ejemplo, supongamos que a un gobernante, a punto de embarcarse en una peligrosa aventura militar, le embargan las dudas a causa de la decidida oposición ciudadana a la guerra y decide indagar en internet para ver si encuentra documentos sobre temas de guerra y paz.

Pongamos que accede a *Google* y pone la siguiente pregunta: “guerras inevitables”. Esto hará que *Google* compare las palabras de la pregunta con las de su índice. Si encuentra un documento que tenga “guerras” e “inevitables” lo devolverá como respuesta. Si no, pues nada. Ya está, ahora si que ya hemos visto prácticamente todo lo que pueden hacer los ordenadores y que tenga que ver con procesamiento de información textual en páginas web.

¿Un nuevo objetivo?

Con estas limitaciones, la búsqueda en internet, como todo el mundo sabe, está repleta de frustraciones. Si alguien busca por “caballos” no encontrará nada que trate sobre “yeguas”. Si alguien consulta sobre cómo evitar la guerra, no encontrará un documento sobre

cómo conseguir la paz, etc. La web semántica quiere solucionar esto. ¿Les suena? A mí sí, a mí me suena a inteligencia artificial. Por tanto, aunque no quieran llamarlo así, con la web semántica se está buscando el mismo objetivo, a saber, que los ordenadores entiendan que un documento sobre “yeguas” puede ser muy relevante para una necesidad de información sobre “caballos”, y que la semántica de la pregunta “¿es posible evitar la guerra?” es la misma que la de “¿es posible conseguir la paz?”.

Además, y aquí es donde suelen poner más énfasis los propagandistas de la web semántica, incluido el mismo creador de la web **Tim Berners-Lee**, se espera que los ordenadores puedan desarrollar tareas de gestión que requieran interpretar información y tomar decisiones adaptándolas al contexto. Por ejemplo, supongamos que yo sé que necesitare tomar un vuelo para, digamos una bella ciudad de Galicia el día tal dentro de dos semanas, y que necesitare regresar a Barcelona 3 días después. En lugar de buscar en la web de diversas compañías aéreas para encontrar las mejores ofertas y horarios, y después en otras páginas para buscar un hotel, hacer las reservas, etc., lo que se espera que pueda hacer gracias a la web semántica en el futuro es entrar en mi asistente digital personal y encargarle la tarea.

Mi asistente digital, nos dicen los propagandistas de la web semántica, será un programa que conocerá mis preferencias. Sabrá, por ejemplo, que no me hace feliz tirar el dinero así que elegirá la mejor oferta económica, pero tendrá en cuenta que no soy masoquista y no me reservará un vuelo que salga a las 4 de la mañana, etc. Tomará los datos personales que necesite de mi cuenta y cerrará las transacciones

con los agentes de software de la empresa de aviación y del hotel y, por último, realizará las anotaciones correspondientes en mi agenda para que no se me olvide nada y no llegue tarde al aeropuerto. ¿Qué les parece?, ¿fácil, no? Ni hablar. Se trata, ni más ni menos que de un objetivo en el que la informática ha fracasado en los últimos 40 años, la inteligencia artificial, ¿por qué va a funcionar ahora?

Infraestructura

Los medios con los cuales se supone que se conseguirá la web semántica son los siguientes: primero, un nuevo lenguaje de codificación de páginas, un nuevo lenguaje de marcado que, como es sabido, se denomina xml. Con él se pueden diseñar lenguajes de etiquetado muy estructurados y muy explícitos en los cuales, en lugar de etiquetas como `` e `<i>`, serían `<título>`, `<subtítulo>`, `<autor>`, `<ciudad>`, etc.

Como para cada tipo de información o de documento harán falta etiquetas específicas —por ejemplo, las páginas web de las compañías aéreas necesitarán algunas como `<vuelo>`, `<hora de salida>`, `<destino>`, etc.— se ha creado un lenguaje, el xml. En realidad es un metalenguaje puesto que permite definir lenguajes específicos, es decir conjuntos de etiquetas determinados para cada necesidad de información. Por ejemplo, los editores de diarios disponen ya de su propio conjunto de etiquetas, así como los matemáticos para expresar ecuaciones, etc.

El segundo elemento con el que se cuenta son los metadatos. Es decir, tenemos aquí otro término-fetiché formado con el prefijo meta. Como saben muy bien los documentalistas, los metadatos son información sobre la información y son, en realidad, una antigua fórmula. Los catálogos de las bibliotecas son metadatos. La venerable

norma Isbd es una norma sobre metadatos, los descriptores asignados a un documento son metadatos, los tesauros y clasificaciones son lo que ahora en la jerga de los metadatos se denominan *schemes*, etc.

La cuestión es que las páginas web ya tienen metadatos. Al menos, suelen tener el metadato título, en forma de etiqueta `<title>` en una zona de las páginas web invisible para las personas, pero visible para los ordenadores. Además, algunas páginas, muy pocas, suelen tener otros como `<keyword>`, `<description>`, etc.

Como saben bien los documentalistas, existe una ambiciosa norma de alcance internacional, *Dublin Core*, que proporciona una lista unificada y normalizada de hasta 15 metadatos del tenor de los ya comentados para que los editores y autores que lo deseen las incluyan en sus páginas web. La idea es simple: si las páginas web tuvieran metadatos del tipo `<título>`, `<autor>`, `<tema>`, `<lugar de publicación>`, etc., los usuarios podríamos hacer preguntas mucho más precisas a los motores de búsqueda. Podríamos, por ejemplo, hacer peticiones de información de este tipo: “búscame documentos publicados en tal o cual lugar y que traten de este y este tema, bajo este punto de vista”.

Pero los metadatos actuales no tienen ni semántica ni sintaxis ni están unificados bajo una norma común que agrupe la diversidad de plataformas de metadatos existentes. Para dotarlos de esas 3 cosas, se han desarrollado otras normas. La más importante se denominada *rdf* (*resource description framework*), que especifica una gramática lógica para que los autores de páginas web puedan describir las propiedades semánticas de los documentos en una notación estándar y común para cualquier tipo de metadatos y basada en nociones fun-

damentales. Básicamente: hay objetos, tales como páginas web, que tienen propiedades tales como un responsable intelectual o una fecha de publicación. Así mismo, hay relaciones entre los objetos, como una página web forma parte o es una versión de otra, etc.

«¿Por qué razón, millones de creadores de páginas web se van a poner a publicar sus documentos en el lenguaje xml, difícil, farragoso y absurdamente abstracto si pueden publicar en el sencillísimo html?»

Para describir el contenido de una página web, entonces, se puede utilizar la norma *rdf* mediante el procedimiento de etiquetado xml para expresar los temas de un documento entre otras cosas. Así que la gran esperanza de la web semántica se basa, al menos, en 3 cosas: xml para hacer los documentos más explícitos; metadatos (expresados también en xml) para hacerlos más fáciles de representar, indizar y buscar; y finalmente (se desprende de lo anterior, aunque nunca se dice) una nueva generación de software (programas y métodos de representación del conocimiento) que sepa explotar las dos cosas precedentes. Esta última necesitará procedimientos normalizados para representar conocimiento, ya sea complejo o de sentido común, las cuales suelen denominarse ontologías. Un campo interdisciplinario donde suelen confluír diversas disciplinas cognitivas, desde la inteligencia artificial hasta la lingüística.

¿Cuál es el problema? Pues que en el majestuoso esquema de la web semántica se supone que los metadatos los ponen (y aquí está el detalle) los propios autores de los documentos. ¿Y qué pasa con los autores de los documentos? Varias

cosas: primero, no están entrenados para poner metadatos y se necesita mucho entrenamiento para saber elegir buenas palabras clave.

En segundo lugar, los autores (no todos, ni mucho menos) mienten. Así de sencillo. Quieren que sus páginas web den muy alto en los buscadores, de manera que coloquen 30 veces la misma palabra, con pequeñas variantes, para que obtengan un buen lugar en los rankings de los motores de búsqueda en los temas que a ellos les interesa, aunque su página no tenga en realidad mucho (o nada) que ver con él.

En tercer lugar, las personas nos equivocamos, y los autores de las páginas web se equivocan: se olvidan de poner metadatos, los ponen mal, lo hacen en unas páginas sí y en otras no, se equivocan en la ortografía, etc. Conclusión: casi ningún motor de búsqueda se fía de los metadatos para generar los resultados de sus rankings.

Posibilidades reales a corto y a medio plazo

El lector ya habrá deducido que, según la opinión de quien esto escribe, las posibilidades a corto y medio plazo de la web semántica son muy reducidas. Efectivamente. Una cosa es que se trate de un objetivo que vale la pena perseguir y otra que sea factible. Permítanme un ejemplo muy significativo. Las personas, los gobiernos y las ONG deben perseguir erradicar la pobreza en el mundo y la instauración plena de los derechos humanos en todos los rincones del planeta. Es un ejemplo de un fin loable, con el que todos debemos comprometernos, pero no parece alcanzable ni a medio ni corto plazo. ¿Debe por ello abandonarse? Ni mucho menos. Todo lo contrario. Debe perseguirse con ahínco, porque es la única forma de conseguir progresos en tales terrenos, aunque sean parciales.

El problema con la web semántica tal como la presentan algunos de sus publicistas es la inmensa cantidad de ingenuidad o de ignorancia que destila (descartamos la mala fe). En comparación, los programas contra la pobreza y a favor de los derechos humanos son obras maestras de pragmatismo (y sabiduría). Se marcan objetivos ambiciosos pero realistas y, sobre todo medibles; se buscan alicientes para los actores implicados, se cuenta con las limitaciones reales del mundo real y no con comportamientos imaginarios de seres imaginarios; etc. En resumen: se realiza un esfuerzo basado en el compromiso y no en la mera propaganda. De este modo, los progresos, aunque muy parciales, son posibles, sostenidos y constatables y cientos de miles de personas con nombres y apellidos se han beneficiado en todo el mundo.

¿Qué sucede con la web semántica tal como la presentan sus defensores más dados a la fantasía o a la repetición tipo “la voz de su amo”? Pues que no hay por donde cogerla si uno se empeña en dotar de sentido al discurso oficial, léase el discurso del, por otro lado admirable *W3 Consortium*, dirigido por el creador de la web, **Tim Berners-Lee**. Empecemos por el etiquetado xml. ¿Por qué razón, millones de creadores de páginas web se van a poner a publicar sus documentos en el lenguaje xml, difícil, farragoso y absurdamente abstracto si pueden publicar en el sencillísimo html?

Los contenidos de **El profesional de la información** están protegidos por copyright. Pueden ser reproducidos hasta un máximo de dos por número (total o parcialmente), siempre que se cite la procedencia.

Sigamos con los metadatos: si casi nadie usa metadatos ahora, ¿por qué razón, de pronto, todo el mundo va a enloquecer de deseos de ponerlos en sus páginas? Para peor, si los autores de páginas web han demostrado su incapacidad para usar una norma relativamente simple como era la primera versión de *Dublin core*, ¿por qué van a hacerlo ahora que ha llevado su complejidad al límite de lo impracticable?

Por último, respecto a las ontologías. Si la inteligencia artificial suma ya varias décadas de fracasos, por lo menos en la hipótesis fuerte, o sea en lograr que los ordenadores se acerquen a algo semejante a pensar, ¿por qué va a tener éxito ahora, así, de repente?, ¿cuál es el cambio de paradigma que se ha producido en las ciencias de la computación y del que, por lo visto nadie se ha enterado, para suponer que los ordenadores ya poseen sentido común? Si usted revisa libros o revistas sobre inteligencia artificial de los años 70 y 80 corre peligro de sufrir un ataque de risa incontenible a la vista de lo que daban por cierto en aquellos años y los magros resultados de ahora. Pero mejor, no lo haga, porque corre el riesgo de perder el respeto a una noble ciencia que tantos logros reales y tanto bienestar ha aportado a la humanidad como es la informática.

Por tanto, las posibilidades de que la web semántica sea una realidad, sin que se produzca antes, al menos un cambio de paradigma de gran calado en las ciencias de la computación, son ridículas. Además, necesitaremos en paralelo cambios no menos importantes en otras áreas incluyendo, por supuesto, en las ciencias de la documentación.

Pero, no se preocupen, gracias a la forma absurdamente triunfalista como se está presentando la web semántica, en los próximos años

dispondremos a cambio de un test muy eficiente para detectar a quiénes gusta hablar por hablar.

Sin embargo, no nos engañemos: el objetivo de la web semántica es magnífico, producirá importantes avances en algunos o en todos los terrenos relacionados con la representación y el acceso al conocimiento y todos debemos apoyarlo. Pero, aunque solamente fuera por estética, ni siquiera ya por ética, habría que evitar volver a la irracionalidad de los primeros años de la web. Fueron unos tiempos de plomo en lo que se refiere al pensamiento crítico: no había día que alguien no anunciara una supuesta

ley histórica, económica, social o política que internet no rompiera. Eso produjo, entre otras cosas, la burbuja de internet, mucha especulación y muchos recursos tirados por la ventana.

Pero, sobre todo, fue un pequeño fracaso de la razón. No volvamos a caer otra vez en lo mismo. No es necesario. Principalmente porque hace décadas que los documentalistas ya estamos construyendo la web semántica.

Nota

1. Versión ampliada y adaptada para *EPI*. Versión anterior publicada en *Biomedica*, 2003, febrero.

Fuentes seleccionadas

Berners-Lee, T.; Hendler, J.; Lassila, O.

“The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities”. En: *Scientific American*, 2001, May. Se puede consultar a través de la página web de la revista: <http://www.sciam.com>

Geroimenko, V.; Chen, C. *Visualizing the semantic web: xml-based internet and information visualization*. London: Springer, 2002.

Semantic web.

<http://www.semanticweb.org/>

W3 Consortium. Semantic web.

<http://www.w3.org/2001/sw/>

Lluís Codina, profesor titular de ciencias de la documentación en la *Universitat Pompeu Fabra* y miembro del *Observatorio de la Comunicación Científica*.

lluis.codina@cpis.upf.es