TEMAS DE WEB SEMANTICA (III)

Procesamiento del lenguaje natural y creación de contenidos[1]

Reyna Carolina Medina Ramírez

Noviembre 2010

La evolución de la Web semántica se traduce en un mundo de innovaciones para la red. Sus efectos se multiplican a lo largo de los diferentes servicios que ya brinda la Web actual. Por eso mismo, resulta imprescindible considerar las diferentes posibilidades que tiene y sus alcances respecto a lo que conocemos actualmente.

En este artículo trataremos temas relativos al desarrollo de la Web 3.0, al mismo tiempo que revisaremos preocupaciones que han acompañado a la red desde sus inicios. Nos referimos al procesamiento del lenguaje natural y a la creación de contenidos. A lo largo de las dos partes de este texto, distinguiremos el impacto que la anotación ontológica tiene sobre estos campos.

Procesamiento del lenguaje natural

El lenguaje natural (español, francés, inglés, etc.) es el sistema más perfecto, completo y complejo creado por el hombre pues, además de ser infinitamente productivo, es capaz de adaptarse a nuevas realidades sin alterar su estructura. Se refleja en el habla, en la

[1] Este artículo fue redactado por Fernando Barajas con base en la investigación *Web Semántica y ontologías de dominio. Un enfoque para la organización y gestión de recursos gubernamentales*, cuya responsable es la Dra. Carolina Medina Ramírez, quien colabora en proyectos de investigación aplicada del Fondo de Información y Documentación para la Industria INFOTEC.

escritura y en construcciones diversas como el lenguaje de señas. El procesamiento informático del lenguaje natural se enfoca principalmente en tareas como la corrección ortográfica, la traducción automática o la creación de resúmenes. Asimismo, pretende generar documentos de forma automática (corpus, libros, periódicos), interpretar oralmente lo escrito (reconocimiento de voz) y hacerse de una comprensión inteligente del sistema lingüístico.

Todo sistema informático que aspire al procesamiento del lenguaje natural debe tomar en cuenta las distintas dimensiones lingüísticas, es decir: la morfología (estructura interna de las palabras para delimitar, definir y clasificas sus unidades), el léxico (serie de palabras utilizadas en un idioma o en un lenguaje), la sintaxis (agrupamiento y orden de elementos lingüísticos y análisis de acuerdo con el orden apropiado), la semántica (significado de las palabras, conceptos e implicaciones), la fonología (dimensión auditiva) y la pragmática (dimensión de lengua en uso, evolución e implicaciones de acuerdo al contexto).

Es en la dimensión semántica del procesamiento del lenguaje natural donde la aportación a la Web 3.0 se hace palpable. Específicamente, en la comprensión de texto para generar automáticamente metadatos, anotaciones y ontologías. Otra aplicación para la nueva Web es la IR, o recuperación de datos pertinentes para el usuario. Su evolución va desde el reconocimiento de vínculos en la red hasta el de voz.

Si bien los actuales buscadores de ontologías resultan de gran utilidad para expertos, no son la mejor herramienta para el usuario final. De ahí la necesidad de búsquedas que comprendan a cabalidad las preguntas de los usuarios y respondan de acuerdo con sus necesidades específicas. De esta forma, la Web semántica puede

responder de manera clara y óptima a los requerimientos de usuarios finales o no especializados. Finalmente se trata de comprender el lenguaje natural de la pregunta del usuario, responder en forma de ontologías y luego traducirlas a lenguaje natural.

Actualmente existen diversas herramientas que se enfocan en el análisis de un solo aspecto de la lengua. Además de ello, existen otras más complejas, como la comprensión del lenguaje natural (análisis complejo de diversos niveles de la lengua), la generación de dicho lenguaje (representación abstracta que tiene que derivar en un texto bien formado), las interfaces de este tipo de lenguaje (convierte el lenguaje informático en natural), la recuperación de información (selección de documentos pertinentes de acuerdo a lo que busca el usuario independientemente del idioma), la extracción de información (no de documentos, sino de datos específicos), los correctores ortográficos y gramaticales, el análisis cuantitativo de textos (número de aparición de una palabra, distribución, etc.) y la traducción automática.

Sin duda, estas herramientas representan una mejora significativa en la experiencia del usuario; no obstante, su aplicación sigue siendo nada más que parcial. Ante lo complejo y multidimensional que es el lenguaje natural, el reto que representa para la informática es más que considerable.

Creación semántica de contenidos

Actualmente los motores de búsqueda en la Web se enfrentan a problemáticas que dificultan mucho la recuperación de información en la enorme biblioteca de datos que es la red. La Web semántica busca resolver esos problemas anotando conceptualmente el contenido; sin embargo, sólo puede trabajar con el limitado número de páginas ya semantizadas. La anotación manual consume una gran cantidad de tiempo y esfuerzo;

por ello, es necesario crear mecanismos para la anotación automática que facilite la evolución de la Web. Sin duda, esto mejoraría por mucho la experiencia del usuario.

La parte fundamental de la Web semántica es la anotación. Mediante ella se crean ontologías que pueden ser entendidas tanto por usuarios como por otros sistemas de cómputo. Sin embargo, la Web sintáctica o Web actual tienen poca anotación conceptual, por lo que el reto es convertir ese gran contenido en semántico. Hacerlo manualmente resulta difícil porque requiere de una gran cantidad de esfuerzo y de tiempo. De ahí la necesidad de implementar mecanismos de anotación automática. El asunto se complica porque los sitios de Web sintáctica tienen pocas etiquetas que expliquen su contenido, o no las tienen en absoluto. En suma, se trata de darle significado a la Web tradicional para mejorar la interacción hombre-máquina y máquina-máquina.

El aprendizaje ontológico para la Web tradicional incluye la extracción, generación y adquisición de una ontología. En la actualidad, esto se lleva a cabo de manera manual o semi-automática. Hoy en día existen propuestas de auto-semantización que se enfocan principalmente a la anotación de servicios Web, el reto es plantear una forma da automatizar de manera integral el proceso de anotación.

Para la creación de contenido semántico se necesitan tecnologías que permitan la representación del significado de palabras, datos y conceptos en lenguajes estándares que puedan ser usados de manera independiente a la lengua o lenguaje que contiene la información. Existen tres tecnologías que conforman la anotación semántica:

- *Tesaurus*: inventario de términos estructurado de acuerdo con su significado. Los tesauros ponen el acento en las diversas relaciones que pueden entablarse en el lenguaje y buscan ordenarlos de acuerdo con una estructura específica. La

organización se construye según relaciones jerárquicas, de asociación o de sinonimia.

- *Topic Maps*: a diferencia del *tesaurus*, estos clasifican conceptos en lugar de términos. Así, un concepto puede estar rodeado de diversos términos, además de que la estructuración es más proclive a cambiar. Finalmente, el objetivo de los *topic maps* es representar conceptos, relaciones entre conceptos y recursos de información vinculados a esos conceptos.
- Ontologías: son capaces de formalizar los términos de un área específica de conocimiento. Con ello, es mucho más sencilla la interacción entre sus conceptos. A final de cuentas, se trata de una estructura subjetiva de relación porque la construye una sola persona desde un sólo punto de vista. Las características de esta herramienta son la posibilidad de tener varias ontologías enfocadas en lo mismo; además, pueden representar un objeto de diferentes formas, es viable identificar el nivel de abstracción en una red de ontologías y pueden ser clasificadas de acuerdo con diferentes criterios.

Finalmente, existen tres herramientas necesarias para la Web semántica: un editor de anotaciones que soporte diferentes esquemas de metadatos, la creación de servicios Web que soporten el uso de ontologías y herramientas para automatizar el proceso. Básicamente, no puede haber Web semántica si no hay anotadores automáticos, pues es prácticamente imposible pensar en anotar manualmente el inmenso contenido de la red.

Acaso podremos resumir las necesidades de la Web semántica en términos de lenguaje. Si comprendemos que su intención es generar anotaciones plenas de significado, pero exentas de lenguaje natural, entenderemos que el procesamiento del lenguaje natural y la

creación automática de contenidos son imprescindibles. De ahí lo importantes que resultan estos temas.

Si te interesó el artículo, también puedes consultar:

- Artículos de Divulgación INFOTEC
- Investigación *"Web Semántica y ontologías de dominio.- un enfoque para la organización y gestión de recursos gubernamentales"*
- Proyectos de Investigación aplicada en INFOTEC
- Proyecto *"Gobiernos Locales Digitales"*

Reyna Carolina Medina-Ramírez
cmed@xanum.mx

Doctora en Informática por la Universidad de Nice Sophia-Antipolis, Francia. Su tesis doctoral titulada *"Contribución a la búsqueda semántica de información: capitalización de conocimientos en una memoria de interacciones genéticas",* fue realizada en el Instituto Nacional de Investigación en Informática y Automatización (INRIA) Sophia Antipolis, Francia. La Dra. Medina realizó una estancia posdoctoral en la Escuela Superior de Ciencias Informáticas (ahora École Polytechnique Universitaire), Francia sobre el tema *"Mecanismos para la capitalización y la difusión de conocimientos en una memoria de proyecto"*. Tiene el reconocimiento "profesor de Tiempo Completo con Perfil Deseable" otorgado por la SEP-PROMEP.