Hacia una Web de Datos.

Herrera, Carlos.

Hacia una Web de Datos

Carlos Herrera

UVa

Abstract. La Web de Datos es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet puede encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida. Al dotar a la Web de más significado se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual, es posible compartir, procesar y transferir información de forma sencilla. Esta Web extendida y basada en el significado, se apoya en lenguajes universales que resuelven los problemas ocasionados por una infraestructura carente de semántica en la que, en ocasiones, el acceso a la información se convierte en una tarea difícil y frustrante.

Keywords: Web de Datos

1 Introduction

La World Wide Web (WWW) ha cambiado la gestión de la documentación electrónica. La WWW cuenta en la actualidad con billones de documentos estáticos, que son accedidos por cerca de 300 millones de usuarios en el ámbito internacional. Sin embargo, esta enorme cantidad de datos, de información, ha hecho crecer de forma exponencial la dificultad para encontrar, acceder, presentar y mantener la información que requieren determinados colectivos de usuarios. Este problema se produce principalmente porque la información sólo está pensada para ser presentada como lenguaje natural, es decir, para que la puedan procesar las personas.

Pero la web no es sólo para usuarios humanos. Otros usuarios no humanos: aplicaciones informáticas denominadas agentes, recorren la red, recopilando información, interaccionando con otros agentes, realizando tareas programadas [1]. Las empresas e instituciones no sólo tienen una cara hacia el público en forma de aplicación web, sino que, además, tienen que realizar sus procesos internos de la manera más eficiente; tienen que comunicarse con otros organismos para realizar transacciones, ser auditadas, compartir recursos con sus filiales y actualizar sus servicios con el menor impacto posible.

2 Antecedentes

A pesar de que la aparición de la Web supuso un hito espectacular en cuanto a que la facilitan a las personas el acceso a la información, la ingente cantidad de

documentos y su crecimiento exponencial, sumados a su falta de estructuración lógica, ha dado lugar a graves problemas en diversos campos [3].

Por ejemplo, los actuales buscadores, basados en el procesamiento automático de los datos han mostrado su incapacidad para mostrar una precisión realmente aceptable en los resultados que ofrecen al usuario [2].

El problema de la precisión en la recuperación de información puede ser visto como consecuencia de la falta de significado que para los ordenadores tienen los documentos Web, en su amplia mayoría formateados mediante HTML, lenguaje de etiquetado que únicamente es capaz de expresar la forma de presentación de los contenidos.

En lo relativo al desarrollo de aplicaciones distribuidas, a finales de la década de los 90, CORBA, RMI, DCOM y demás técnicas, abiertas o propietarias, dominaban el mercado para la integración de las aplicaciones. Estas tecnologías se basaban en el intercambio síncrono de mensajes binarios y en el uso de ciertos puertos predefinidos, dependiendo de forma crítica de los detalles de la comunicación. La política comercial fue un problema también: DCOM era propiedad de Microsoft; CORBA era una creación de OMG que, aunque neutral, creó una serie de especificaciones cada vez más complejas que hicieron que sólo empresas privadas fueran capaces de construir versiones fiables y de pago; RMI, de Sun, fue el ultimo intento y, aunque en 1999 el recién aparecido J2EE permitía el disfrute de esta tecnología, la revolución de los Servicios Web ya se estaba gestando.

3 Cómo funciona la Web de Datos

Tradicionalmente, los datos han estado prisioneros dentro de aplicaciones propietarias. Esta visión que la informática había otorgado a la información presentaba una gran dependencia entre el proceso de tratamiento y los datos. Pero con el tiempo, esta filosofía fue siendo sustituida por otra que otorga la primacía a los datos. El proceso de tratamiento ha pasado a un papel secundario, lo que realmente importa ahora son los datos; Data as King es el lema utilizado por desarrolladores para hacer sus productos más competitivos [4].

La Web de Datos pretende hacer de los datos un recurso independiente de la aplicación, con posibilidades de clasificación y composición e integrado dentro de un ecosistema de información. Esta idea no está pensada sólo para Internet, es más, alguna de sus más singulares ventajas pueden ser explotadas en sistemas de información empresariales.

4 Tecnologías

4.1 eXtensive Markup Language (XML)

XML permite la codificación para la distribución de documentos complejos por Internet. Para explicar el origen de XML, es preciso hacer referencia a SGML

(Standard Generalized Markup Language) que es una norma que pretende establecer una manera genérica de especificar, definir documentos, la cual permitiese a su vez usar formatos de mayor flexibilidad y portabilidad. XML es un subconjunto de SGML, y define un formato de texto diseñado para la transmisión de datos estructurados. Al ser un subconjunto de SGML mantiene sus características de validación, estructurado y especialmente facilita la extensibilidad, porque es un metalenguaje que permite describir lenguajes de marcas, tanto la definición de etiquetas como la relación estructural que existen entre ellas.

4.2 Resource Description Framework (RDF)

RDF son las siglas que español indican marco de descripción de recursos. Como su nombre indica el área en la que está enmarcado es la descripción de recursos de la red, entendiendo por recurso todo lo que nos permita definir cualquier cosa, páginas, personas, dispositivos, etc. RDF permite que las condiciones que se quieren preguntar sobre un recurso sean definidas como un conjunto de propiedades que componen el esquema.

- RDF ofrece una estructura semántica inambigua que permite la codificación, el intercambio y el procesamiento automático de los metadatos normalizados.
- RDF proporciona también reglas para facilitar técnicamente la manera de explicar conceptos de modo que los ordenadores puedan procesarlo rápidamente y proporciona un medio que posibilita la edición de vocabularios con propiedades definidas para la descripción de los recursos de una comunidad.
- RDF usa la sintaxis del lenguaje XML para el intercambio y procesamiento de metadatos, las condiciones se recogen en los rdf:description de los elementos XML.

4.3 Platform for the Internet Content Selection (PICS)

Los PICS nos indican lo adecuado o conveniente de determinados ficheros de datos según la comunidad en la que se encuentre el usuario. Es una infraestructura para asociar las etiquetas con los contenidos de Internet. Aunque en un principio estaba destinado al control del acceso de los niños a Internet, su uso se puede extender a otras etiquetas que incluyan privacidad, licencias, etc. PICS es una plataforma sobre la cual se han construido otros servicios de clasificación que no sólo define una manera de construir etiquetas sino que es un mecanismo para realizar las valoraciones.

Para que el filtrado de documentos no deseados se lleve a cabo, también es necesario un software cliente y otro servidor que tengan implementado el sistema de valoración. Estas funciones se pueden realizar por separado, lo cual permite que por un lado los desarrolladores de software puedan realizar una aplicación informática sin suministrar un sistema de valoración mientras que por otro una organización puede crear sistemas de valoración sin tener que desarrollar el software.

4.4 OWL

OWL es un mecanismo para desarrollar temas o vocabularios específicos en los que asociar esos recursos. Lo que hace OWL es proporcionar un lenguaje para definir modelos estructurados que pueden ser utilizados a través de diferentes sistemas. Estos modelos, que se encargan de definir los términos utilizados para describir y representar un área de conocimiento, son utilizados por los usuarios, las bases de datos y las aplicaciones que necesitan compartir información específica, es decir, en un campo determinado como puede ser el de las finanzas, la medicina, el deporte, etc.

5 Conclusiones

La Web de Datos es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida. Al dotar a la Web de más significado y, por lo tanto, de más semántica, se pueden obtener soluciones a problemas habituales en la búsqueda de información gracias a la utilización de una infraestructura común, mediante la cual, es posible compartir, procesar y transferir información de forma sencilla. Esta Web extendida se apoya en lenguajes universales que resuelven los problemas ocasionados por una Web tradicional en la que, en ocasiones, el acceso a la información se convierte en una tarea difícil y frustrante.

References

1. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T. Linked Data on the Web. LDOW 2008.
2. Cilibrasi, R., Vitanyi, P. M. B. The Google Similarity Distance. IEEE Trans. Knowl. Data Eng. 19(3): 370-383 (2007).
3. Martinez-Gil, J. Automated knowledge base management: A survey. Computer Science Review 18: 1-9 (2015).
4. Shadbolt, N., Berners-Lee, T., Hall, W. The Semantic Web Revisited. IEEE Intelligent Systems 21(3): 96-101 (2006).