Tema-4-Gestion-de-la-Informacion...

juanfrandm98

Sistemas de Informacion Basados en Web

3º Grado en Ingeniería Informática

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Tema 4. Gestión de la Información

- Conocer diversas tecnologías para almacenar la información en plataformas Web.
- Ser capaz de interpretar documentos XML y valorar su uso en soluciones Web.
- Conocer los aspectos básicos de la Web Semántica y su influencia en los sistemas de información basados en Web.

Tipos de información en la WWW:

- Datos desestructurados: archivos de texto, vídeo, audio, imágenes, presentaciones... Tienen difícil tratamiento y almacenaje en BD.
- Datos semi-estructurados: con cierta estructura pero no muy rígida, como un Curriculum Vitae.
- Datos estructurados: todos los registros tienen una estructura fija y fuertemente tipada.

4.1. Gestión de Datos Estructurados

Los datos estructurados se gestionan con un modelo relacional, que consiste en el almacenamiento de datos en tablas compuestas por filas (tuplas) y columnas (campos). Los Sistemas de Gestión de Bases de Datos (SGBD) relacionales usualmente implementan la mayoría de las funcionalidades necesarias. Destacan MySQL, PostgreSQL, SQLite u Oracle.

PHP no puede proteger una base de datos por sí mismo. Hay que lograr una protección en profundidad tomando acciones para aumentar la protección de una base de datos en el mayor número de sitios posibles.

La inyección directa en SQL es una técnica donde un atacante crea o altera comandos SQL existentes para exponer datos ocultos, sobrescribir los valiosos o incluso ejecutar comandos peligrosos a nivel de sistema.

PHP utiliza conexiones persistentes a bases de datos, que son enlaces que no se cierran al finalizar la ejecución de un script, lo que le proporciona una gran eficacia.

4.2. Gestión de Datos Semiestructurados

La gestión de datos semiestructurados puede ser más compleja que los datos estructurados. Los posibles formatos para estos datos son json y XML; y el SGBD para ellos es noSQL.

XML (*Extensible Markup Language*) es un metalenguaje de marcas para definir lenguajes de marcas, siendo una simplificación de SGML (*Standard Generalized Markup Language*) para definir la gramática de lenguajes específicos para estructurar documentos grandes.

Se forma como un árbol de contenidos etiquetado, ilimitado y ordenado. Sintáxis:

- Elemento: etiqueta de apertura contenido textual y etiqueta de fin:
  <etiqueta> contenido textual </etiqueta>
- Atributos: en la etiqueta de apertura aparecen pares clave-valor:
  <persona nombre=”Zerjillo” apellidos=”Alonso”/>
  Los atributos no están ordenados y pueden ser complejos, mientras que el contenido está ordenado y pueden ser sub-árboles.

- Prólogo: versión de XML, codificación, localización de recursos externos:
  <?xml version=”1.0” encoding=”utf-8” ?>
- Entidades: referencias o contenidos asignados a constantes:
  <!ENTITY tema1 “Tema 1: Intro”>
  <!ENTITY tema2 SISTEM “tema2.xml”>
  <report> &tema1; &tema2; </report>
- Document Type Definition (DTD): mecanismo simple de definición de documentos XML.
- XML-Schema: lenguaje de descripción de tipos XML propuesto por el W3C. Muy versátil pero bastante más complicado que los DTDs. Permiten definición de tipos y de atributos:
  <name>Yo</name>
  <age>34</age>
  <since>1968-03-27</since>

  <book isbn=”314-2322-22”>...</book>
- MathML: lenguaje para describir fórmulas matemáticas.

XSL (*eXtensible Stylesheet Language*) se utiliza para dar estilo a los documentos XML Consta de varias partes:

- XSLT: lenguaje para transformar XML.
- Xpath: para navegar por los documentos XML.
- XSL-FO: para formatear los documentos XML.

La web semántica es un conjunto de actividades desarrolladas en el seno de W3C con tendencia a la creación de tecnologías para publicar datos legibles por aplicaciones informáticas. Se basa en la idea de añadir metadatos semánticas y ontológicos a la WWW. Esas informaciones adicionales (que describen el contenido, significado y la relación de los datos) se deben proporcionar de manera formal, para que así sea posible evaluarlas automáticamente por máquinas de procesamiento.

La ontología es la descripción formal que proporciona a los usuarios humanos un conocimiento compartido sobre un dominio concreto. Es una definición formal de tipos, propiedades y relaciones entre entidades que realmente o fundamentalmente existen para un dominio de discurso en particular.

Las ontologías son útiles para organizar datos, mejorar las búsquedas e integrar información. Ejemplo de Ontología “Universidad”:

- Clases: :Profesor, :Alumno, :Asignatura, :Departamento
- Instancias de clases: :Zerjillo es instancia de :Profesor
- Relaciones: :Imparte(:Zerjillo, :SIBW)
- Herencia: :Profesor es subclase de :Personal
- Restricciones: no :Imparte(:Alumno, :Asignatura)
- Restricción de cardinalidad: :Departamento solo tiene un :Director

Lenguajes para la web semántica:

- RDF (*Resource Description Framework*): familia de especificaciones de la W3C originalmente diseñado como un modelo de datos para metadatos. Permite describir anotaciones sobre recursos web (asociados a una URI).
- RDF Schema: extensión semántica de RDF. Un lenguaje primitivo de ontologías que proporciona los elementos básicos para la descripción de vocabularios.
- OWL (*Web Ontology Language*): lenguaje de marcado para publicar y compartir datos usando ontologías en la WWW. OWL tiene como objetivo facilitar un modelo de marcado construido sobre RDF y codificado en XML.

4.3. Gestión de datos desestructurados: búsqueda de información

Es difícil gestionar fuentes de información tan variada como la que hay en general en la web. ¿Cómo se apañan entonces los buscadores?

Un rastreador es un robot que navega por las webs indexando y clasificando los contenidos de las mismas. En el fondo hacen una búsqueda en un grafo (por anchura, por profundidad o esquemas mixtos).

Los mapas de sitio son archivos XML que describen de manera jerárquica la estructura del sitio facilitando la vida a los buscadores.

Los archivos robots.txt limitan a los buscadores impidiéndoles que partes de un sitio (o todo) sea indexado ni muestren sus resultados.

Fases habituales de un buscador:

- Tokenización: extraer palabras.
- Limpieza: eliminar tokens no útiles.
- Análisis semántico: se relacionan términos similares.
- Indexación: asociar términos de búsqueda y términos en el artículo.
- Evaluar la relevancia del artículo frente a las palabras de búsqueda.

Las técnicas SEO (*Search Engine Optimization*) son un conjunto de acciones orientadas a mejorar el posicionamiento de un sitio web en la lista de resultados de los buscadores de Internet. Trabaja aspectos técnicos como la optimización de la estructura y los metadatos de una web, pero también se aplica a nivel de contenidos, con el objetivo de volverlos más útiles y relevantes para los usuarios.