Visualización de Recursos Textuales en la Web Semántica

M. Pérez-Coutiño, M. Montes-y-Gómez, A. López-López, L. Villaseñor-Pineda
Laboratorio de Tecnologías del Lenguaje, Ciencias Computacionales, INAOE, México
{mapco,mmontesg,allopez,villasen}@inaoep.mx

RESUMEN

Las recomendaciones actuales para el desarrollo de la web semántica se han enfocado principalmente en el modelado de los recursos web para su entendimiento por máquinas. Estos modelos pueden tener información valiosa para los usuarios que navegan dichos recursos. Sin embargo, los aspectos de la interacción entre los usuarios humanos y estos modelos no han sido tratados explícitamente. Este artículo describe la visualización y navegación de sitios web basada en la extracción de descripciones semánticas que modelan el contenido de los recursos textuales y que cumplen con las recomendaciones de la Web Semántica.

Palabras clave:
Interfaces de usuario para la web, Visualización de Información Textual, Metadatos, Web Semántica.

INTRODUCCION

Internet se ha transformado en el medio preferido para el intercambio de información y conocimiento. Sin embargo, hoy día, esta información está diseñada para su uso por humanos y no por computadoras (Berners-Lee et al., 2000). Esto resulta en una serie de problemas debido a la naturaleza no estructurada de la información en la web. Con el objetivo de mejorar y extender el uso automático de la web su información debe ser enriquecida. Una forma de alcanzar este objetivo es incluir *meta-información*, i.e. información acerca del recurso en sí describiendo su contenido y sus relaciones con otros recursos en una manera significativa para las máquinas (Jeffery, 1998).

Los esfuerzos en la creación de esquemas de metadatos para la web están liderados por el *World Wide Web Consortium* (*W3C*) mediante su iniciativa de la web semántica[1]. Esta define la web semántica como una representación abstracta de datos en la web basada principalmente en el estándar RDF.

Otras comunidades como el *Dublin Core Metadata Initiative*[2] también utilizan RDF/XML para la publicación de datos en la web.

Con base en estas ideas se desarrolló un sistema multiagentes (SMA) que sirve de soporte en el proceso de autoría de documentos web (Pérez et al., 2003). Una característica relevante de este sistema es su doble funcionalidad. Por una parte, y de manera principal, el SMA considera la extracción de metadatos de recursos textuales, y su representación en un formato adecuado para mejorar el procesamiento automático de la web. Por otra parte, el SMA también contempla la generación de esquemas que faciliten al usuario la visualización de información descriptiva relevante para la toma de decisiones durante la navegación de colecciones de texto.

Este artículo se enfoca en la segunda funcionalidad del SMA. En él se propone que las descripciones semánticas también pueden ser usadas como mecanismo de retroalimentación al usuario para orientarlo en el contexto del contenido del recurso, y para indicarle las relaciones del recurso con otros a su alrededor. Básicamente se explica la manera en que el SMA genera un volumen de hipertexto en XHTML para la visualización de la colección por los usuarios humanos. Esta salida utiliza una plantilla que cumple con las recomendaciones del *W3C* para XHTML 1.

DESCRIPCION DEL SISTEMA

Como se mencionó anteriormente el objetivo de nuestro *Sistema Multiagentes* (SMA) es construir automáticamente las descripciones semánticas de los recursos web, en particular los recursos textuales, y utilizar dichas descripciones para ofrecer retroalimentación al usuario y orientarlo en el contexto del contenido del recurso, así como indicarle las relaciones del recurso con otros a su alrededor. Para alcanzar este objetivo el SMA realiza dos tareas principales:

- *Procesamiento de la entrada*, consiste de la identificación del lenguaje, tópicos, y otros atributos para cada documento, así como sus interrelaciones.
- *Generación de salidas*, considera la creación de un conjunto de metadatos en RDF/XML para el procesamiento automático mediante máquinas, y una colección de documentos en XHTML para la visualización y navegación por humanos.

[1] www.w3.org/2001/sw

[2] www.dublincore.org

J. Díaz de León, G. González, J. Figueroa (Eds.): Avances en Ciencias de la Computación, pp. 277-280, 2003.

La descripción detallada de la arquitectura del SMA se puede encontrar en (Pérez et al., 2003).
A continuación se presenta una breve descripción de la funcionalidad de los cuatro principales agentes del SMA. Se hace especial énfasis en el agente generador de vistas en XHTML.

Extractor de Tópicos

El objetivo de este agente es identificar y extraer los tópicos más importantes de cada documento de la colección de entrada. La identificación de los tópicos consiste de dos pasos:

- Extraer todos los nombres de entidades (lugares, personas, organizaciones, etc.) de cada documento. Para ello se usan técnicas superficiales de procesamiento de lenguaje natural tal como el etiquetado de partes de la oración y algunas reglas heurísticas relativas al Español (Arévalo et al., 2002).
- Enriquecer el conjunto de entidades añadiéndole una lista de palabras clave obtenidas mediante técnicas tradicionales de indexado de documentos (Baeza-Yates y Ribeiro-Neto, 1999).

A partir de los tópicos identificados se construye una representación formal del contenido de los documentos, la cual será usada como base para la generación de los metadatos.

Extractor de relaciones

La meta del agente extractor de relaciones es identificar el conjunto de documentos relacionados temáticamente con cada ítem de la colección. Para alcanzar dicha meta, el agente realiza el cálculo de similitud para cada par de documentos en la colección original, y entonces determina las conexiones más importantes. La medida de similitud está basada en el coeficiente de Dice, mientras que el criterio para determinar el conjunto de documentos relacionados está basado en el valor de la similitud promedio obtenida. Este criterio permite producir relaciones independientemente de cuan heterogénea u homogénea sea la colección.

Generador de Metadatos

Este agente se basa en la meta-información extraída (i.e., los tópicos y las relaciones) junto con otra información tomada de los documentos, como la última fecha de modificación, el idioma, y el formato.

Entonces, estos elementos se codifican de acuerdo a las recomendaciones para generación de metadatos Dublin Core en RDF/XML (Becket and Miller, 2002). Los metadatos resultantes sirven como información entendible por máquinas, permitiendo su procesamiento automático por agentes web y máquinas de búsqueda.
Los metadatos correspondientes a un documento de prueba se muestran en la sección de experimentos.

Generador de Vistas en XHTML

Además del conjunto de matadatos descrito en la sección anterior, el SMA genera un volumen de hipertexto en XHTML para la navegación y visualización de la colección por los usuarios humanos. Esta salida utiliza una plantilla que cumple con las recomendaciones del *W3C* para XHTML 1.0.
Las páginas web generadas por este agente se dividen en 4 regiones (ver la figura 1):

- Una *región de metadatos* que se utiliza para mostrar los metadatos generados por el sistema. Esta información sirve como una ficha para el usuario donde puede visualizar la siguiente información del documento en cuestión: título, autor, editor, fecha, y los 5 temas y 5 documentos relacionados más importantes.
- Una *región de detalles* que se usa para desplegar información adicional sobre el documento. Por ejemplo, en ella se puede mostrar la lista completa de temas del documento, la lista completa de documentos relacionados, e incluso los metadatos asociados a cada uno de estos documentos relacionados.
- Una *región de contenido* que despliega el contenido propiamente dicho del documento actual.
- Una *región de pie de página* que sirve para desplegar información adicional al usuario, tal como derechos de propiedad, información de contacto, etc. Por ejemplo, hemos agregado una liga al código fuente del archivo que contiene los metadatos asociados al documento.

Cabe señalar que el usuario puede navegar la colección de documentos seleccionando con el ratón las ligas que relacionan los documentos entre sí. El contenido del documento actual y la información sobre sus metadatos son actualizados de acuerdo a cada documento recorrido.

EXPERIMENTOS

Para la realización de los experimentos de funcionalidad del SMA se utilizaron dos colecciones de prueba. Estas colecciones están en formato de texto plano. Su tamaño total es de aproximadamente 5 Mb. Las principales diferencias entre ellas son sus temáticas y el tamaño promedio de documento.

Por ejemplo, la colección News94 comprende noticias de diferentes temáticas y el tamaño promedio de cada documento es de 3.44 Kb. Por su parte, la colección ExcelNews consiste de 1,357 documentos de noticias nacionales e internacionales de 1998 a 2000, y también contiene notas culturales acerca de literatura, ciencia y tecnología. El tamaño promedio de documento es de 3.52 Kb.

Resultados

La tabla 1 muestra los resultados obtenidos tras el análisis de las colecciones de prueba, estos resultados consideran tres aspectos principales:

1. La distribución temática de las colecciones.
2. El tiempo requerido para su análisis.
3. El nivel de conectividad del conjunto de documentos en hipertexto.

Como se observa en la tabla 1, el número de tópicos encontrados en promedio para cada documento varía considerablemente. Desde 18 hasta 27, por lo que el uso de la región de detalles en la interfaz para la visualización de este elemento resulta muy apropiada. Lo mismo sucede con el número de documentos relacionados a cada documento, desde 5 hasta 35.

La tabla 2 muestra un ejemplo de los metadatos obtenidos para un documento de entrada.

La figura 1(a) muestra la interfaz de hipertexto obtenida a partir del mismo documento de prueba. Observe que la región de detalles está ocupada por los temas del documento actual. Por otra parte, las figuras 1(b) y 1(c) muestran respectivamente el uso de la región de detalles para desplegar la lista completa de documentos relacionados y los metadatos de un documento relacionado.

Colección	Tópicos	Instancias de tópicos	Promedio de tópicos por documento	Tiempo de indexado (seg)	Tiempo de búsqueda (seg)	Documentos conectados	Relaciones	Promedio de documentos relacionados
News94	2,571	4,874	27	0.26 seg	0.55	90	459	5
ExcelNews	24,298	72,983	18	3.56	230.59	1350	47,486	35

Tabla 1. Resultados principales del análisis de las colecciones de prueba.

```
<?xml version="1.0"?>
<!DOCTYPE rdf:RDF SYSTEM "http://dublincore.org/2000/12/01-dcmes-xml-dtd.dtd">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description>
<dc:creator>AcreS, Multi-Agent System for web document authoring </dc:creator>
<dc:publisher>Language Technologies Lab, Csc, Inaoe </dc:publisher>
<dc:subject>Presidente Ortiz Rubio, selección de candidato, PRI, Partido Socialista Fronterizo, PNR, Poncho Martinez
Domínguez, fuerza caciquil, Supuso Madrazo, Polo Sánchez Celis, Javier Romero, derrota automática, Madrazo, partido
callista, PPS de Lombardo, ... , poca politica.
</dc:subject>
<dc:identifier>010698-1Lunes</dc:identifier>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/020598Sabado.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/050698-1Viernes.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/180698-1Jueves.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n94/200698Sabado.xhtml </dc:relation>
<dc:relation>http://ccc.inaoep.mx/~mapco/acres/n4/280698Domingo.xhtml </dc:relation>
<dc:format>xhtml</dc:format>
<dc:date>06-01-1998</dc:date>
<dc:language>es</dc:language>
</rdf:Description>
</rdf:RDF>
```

Tabla 2. Metadatos obtenidos a partir de un documento de entrada.

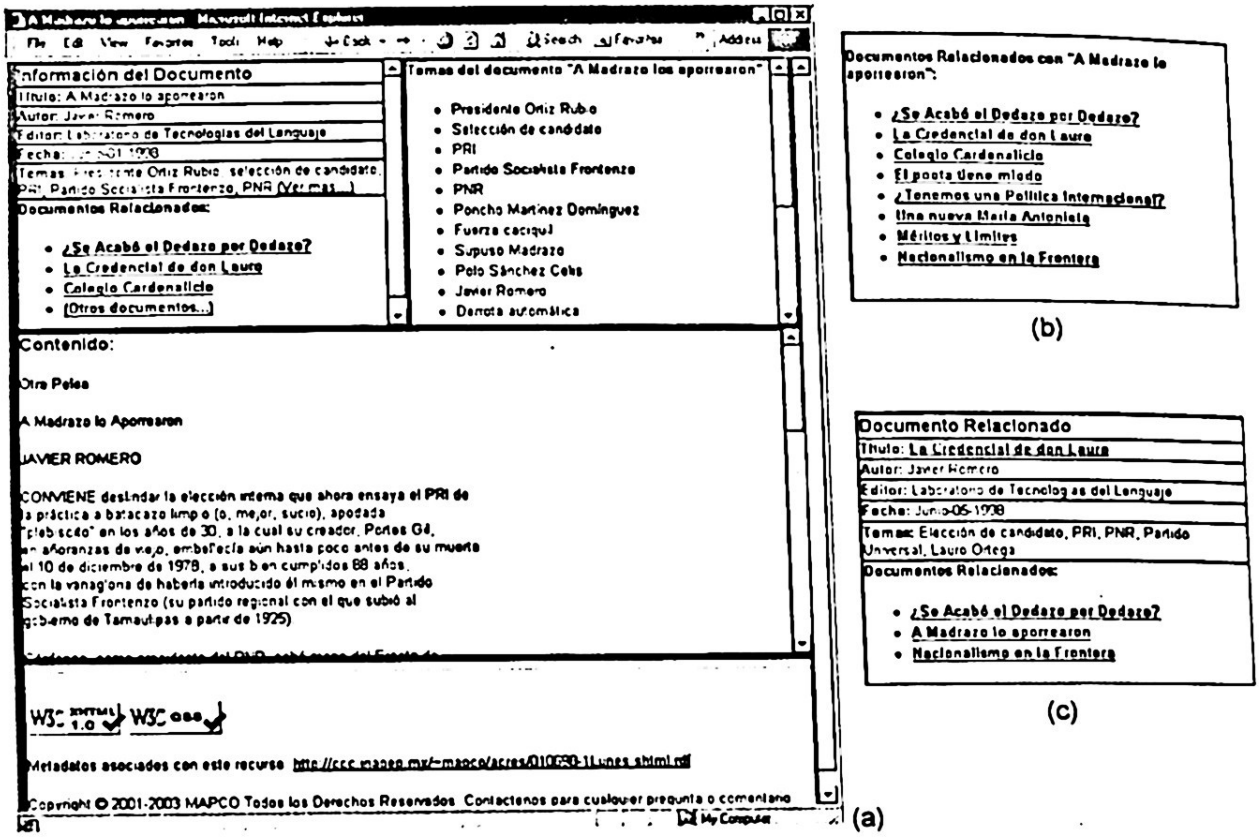

Figura 1. (a) Página de ejemplo generada por el SMA. (b) Detalles con la lista de documentos relacionados. (c) Detalles con los metadatos de un documento relacionado

CONCLUSIONES

Se ha propuesto un sistema multiagentes que automatiza parcialmente la autoría de documentos web. Una característica relevante de este sistema es la producción de dos tipos de salidas, una diseñada para su procesamiento automático por máquinas (metadatos Dublin Core en RDF/XML), y otra para humanos (el hipertexto en sí).

Los experimentos realizados nos permitieron concluir que los metadatos pueden ser usados como un mecanismo importante de retroalimentación al usuario que le ayuda en dos tareas principales. Por un lado le permite visualizar información suficiente sobre los recursos en la colección sin tener la necesidad de recorrerlos uno a uno. Por otro lado lo ubica en el contexto temático del contenido de los documentos.

El trabajo en proceso incluye:

- Una extensión al esquema de Dublin Core para capturar la semántica de las relaciones y los tópicos.
- La aplicación del SMA sobre a los resultados devueltos por máquinas de búsqueda para poder recorrerlos en orden arbitrario o por temas de interés al usuario.

REFERENCIAS

1. Baeza-Yates R., y B. Ribeiro-Neto. Modern Information Retrieval, Addison-Wesley, 1999.
2. Beckett D., y E. Miller. *Expressing Simple Dublin Core in RDF/XML*, Institute for Learning and Research Technology (ILRT) University of Bristol; W3C, 2002-07-31.URL:http://dublincore.org/documents/2002/07/31/dcmes-xml.
3. Berners-Lee T., J. Hendler y O. Lassila. The Semantic Web, *Scientific American*, May 2001.
4. Pérez Coutiño M., López López A., Montes y Gómez M, Villaseñor Pineda L., A Multi-Agent system for Web Document Authoring. *Proceedings of theAtlantic Web Intelligence Conference AWIC-03 LNCS 2663 Springer Verlag*, Madrid, España, 2003.
5. Jeffery K. Metadata: An Overview And Some Issues. ERCIM News, (35), 1999.
6. Arévalo, M., X. Carreras, L. Màrquez,, M.A. Martí, L. Padró, y M.J. Simón. A Proposal for Wide-Coverage Spanish Named Entity Recognition., vol 28. Edited by *Journal "Procesamiento del Lenguaje Natural"*SEPLN. May 2002.