FLERSA: Soporte a la Definición de Anotaciones y Búsquedas Semánticas en un CMS

José L. Navarro-Galindo, José Samos

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Granada
C/ Periodista Daniel Sucedo Aranda s/n, 18071 Granada (Spain)
(+34) 958 240 576
jlnavarr@correo.ugr.es, jsamos@ugr.es
Sitio web: http://www.scms.es

Resumen En este artículo, se presenta FLERSA (Flexible Range Semantic Annotation) como una herramienta de anotación de contenido web en lenguaje natural centrada en el usuario. La herramienta ha sido desarrollada sobre un WCMS (Sistema de Gestión de Contenidos Web) y permite tanto anotaciones como búsquedas semánticas. Para la anotación semántica manual, se usa una nueva técnica de marcado de rangos flexibles, basada en el estándar RDFa. Para la anotación semántica automática, se usa un enfoque híbrido, basado en técnicas de aprendizaje tales como el Modelo de Espacio Vectorial combinado con N-gramas. Además, la herramienta es capaz de realizar diferentes tipos de búsquedas: clásica, basadas en palabras clave; guiada por las anotaciones; guiada por conceptos y, la más interesante, basada en lenguaje natural.

Keywords: anotación semántica, búsqueda semántica, RDFa, metadatos, Web Semántica.

1. Introducción

Uno de los aspectos más importantes de cara a progresar hacia la Web Semántica es cómo convertir el contenido Web nuevo y el existente, expresado en lenguaje natural, en su equivalente semántico en el que cual el contenido se enriquece con metadatos en formato estructurado.

El marcado semántico de documentos Web es el primer paso hacia la adaptación del contenido Web a la Web Semántica. El enriquecimiento semántico se hace posible mediante el etiquetado del contenido web con metadatos, los cuales posibilitan describir las entidades que se encuentran en el contenido y las relaciones entre ellas [16]. Proporcionar un significado bien definido a los elementos que actualmente forman la Web posibilita, entre otras cosas, mejorar las capacidades de búsqueda contextual e incrementar la interoperabilidad de sistemas en contextos colaborativos [18]. Sin embargo, la mayoría del contenido de la Web permanece desestructurado debido a la dificultad que supone su marcado.

La principal contribución de este trabajo es presentar FLERSA, una herramienta de anotación semántica para sistemas de gestión de contenidos web. Entre

las principales cualidades de la herramienta destacan la creación de anotaciones semánticas manuales basadas en una nueva técnica de marcado de rangos flexibles para conseguir la evolución de los documentos anotados de forma más efectiva que XPointer; también permite realizar anotaciones semánticas de manera automatizada mediante el uso de técnicas de aprendizaje tales como el Modelo de Espacio Vectorial combinado con N-gramas para determinar los conceptos de los que trata el contenido web; además permite la realización de búsquedas contextuales basadas en la información semántica recogida en las anotaciones, donde el uso de ontologías de apoyo ayudan a la inferencia de los resultados.

El artículo comienza con una introducción al ámbito de la Anotación Semántica, continuando con un estudio de la herramienta FLERSA: sus características, requisitos de diseño, arquitectura y desarrollo. También se estudia de forma abreviada como tienen lugar los procesos de anotación semántica junto con las tecnologías asociadas. Después, se estudia cómo explotar las anotaciones semánticas a la hora de realizar búsquedas y los beneficios que éstas aportan. Finalmente, el trabajo termina con las conclusiones y las referencias bibliográficas.

2. Anotación Semántica

En el diccionario de la Real Academia de la Lengua Española se define "*anotación*" como "Acción y efecto de anotar". Asimismo, se define el término "*anotar*" como "Poner notas a un escrito, una cuenta o un libro".

En el contexto computacional, una anotación consiste en asignar una nota a una porción de texto específica. Más específicamente, en el contexto de la Web Semántica, la nota asignada contiene información semántica en forma de metadatos con el objetivo de establecer un enlace entre una ontología de referencia [9] y la parte específica del texto que está siendo marcada.

2.1. Tipos de Anotaciones Semánticas

Aunque se pueden realizar diferentes clasificaciones, nuestro foco de interés se centra en dos criterios:

- De acuerdo al lugar donde se almacenan
  - **Internas o embebidas:** Almacenadas dentro del mismo documento web donde se realiza la anotación.
  - **Externas:** Almacenadas en ficheros o en servicios distintos del documento web que se anota.
- Conforme a su nivel de automatización
  - **Directa o Manual:** El usuario realiza las anotaciones directamente en un contenido dado mediante el uso de herramientas específicas.
  - **Automática:** De alguna manera, un proceso automático genera las anotaciones, identificando entidades semánticas y sus relaciones.

2.2. Infraestructura y Lenguajes de Anotación Semántica

Un lenguaje de anotación es un conjunto de etiquetas semánticas y reglas sintácticas que son usados para describir a la computadora la estructura de un documento digital (DLO, Document Like Object) para representar su significado.

RDF [10] es el marco de trabajo desarrollado en 1997 por la W3C como infraestructura estándar para describir recursos y proporcionar una base para el procesamiento de datos y para posibilitar la interoperabilidad semántica entre aplicaciones Web. RDF usa el lenguaje de marcas extensible XML [3] para codificar los metadatos.

RDFa [1] es una sintaxis propuesta por la W3C para expresar datos estructurados RDF en documentos HTML, o lo que es lo mismo para incrustar semántica en los documentos.

Annotea [6] es un proyecto de la W3C que especifica la estructura de anotación para documentos web, haciendo énfasis en el uso colaborativo de anotaciones. Permite realizar anotaciones en las páginas web sin que el documento original sufra ninguna transformación.

2.3. Métodos de Marcado

Se trata de un aspecto importante cuando se llevan a cabo anotaciones semánticas, ya que es necesario delimitar, de algún modo, el rango de texto sobre el cual se realiza la anotación.

XPointer [4] es un estándar de la W3C que proporciona un mecanismo formal para identificar de forma única fragmentos de un documentos XML con objeto de crear enlaces. La mayoría de las herramientas de anotación tienden a usar la tecnología XPointer así como patrones y expresiones regulares.

DOM Range [7] es una tecnología que ayuda en el proceso de marcado mediante la definición de rangos. Un `Rango` es una parte arbitraria de un documento HTML, definida por puntos límite que denotan el comienzo y el fin del mismo.

2.4. Sistemas de Anotación Semántica

A día de hoy, se dispone de un amplio abanico de herramientas de anotación para la producción de etiquetas semánticas. Algunas de estas herramientas (como por ejemplo Amaya [15]) proporcionan marcado semántico de páginas web pero no soportan anotación semántica automática. Tampoco permiten la correcta evolución de los documentos anotados debido a que se basan en la tecnología XPointer. Estas herramientas permiten que los usuarios realicen anotaciones semánticas de forma más o menos sencilla, pero no proporcionan continuidad en el proceso semántico de manera que tras el proceso de anotación se facilite la explotación de la información que se anota.

Otras herramientas soportan la anotación semántica automática (como por ejemplo KIM [14]), pero se están haciendo obsoletas y no siguen la filosofía de diseño centrada en el usuario, por lo que requieren de un gran esfuerzo por

parte de los usuarios finales para conseguir que se resuelvan sus necesidades. Un estudio completo de todas estas herramientas se puede encontrar en [19].

La principal motivación para desarrollar una nueva herramienta fue intentar tener en cuenta tantos aspectos de arquitectura centrada en el usuario y de recuperación semántica de información como fueran posibles. Desde nuestro punto de vista, algunas características importantes para una herramienta de anotación semántica son:

- Entorno web centrado en el usuario [13]. Se caracteriza por centrar el diseño de la herramienta en las necesidades y objetivos del usuario, en oposición a centrarlo en los diseñadores y/o posibilidades tecnológicas.
- Entorno de trabajo ligero para una infraestructura común. Las herramientas de la Web Semántica deberían ser compatibles con las aplicaciones más comunes como por ejemplo los WCMS y los portales Web.
- Anotaciones semánticas manuales que permitan la evolución del documento.
- Anotaciones semánticas automáticas usando técnicas de aprendizaje, que permitan la anotación automática de grandes repositorios de información específica de un dominio.
- Uso de ontologías en las anotaciones semánticas, tanto como infraestructura de anotación como a nivel de ayuda para establecer enlaces entre conceptos y los fragmentos de texto que se refieren a ellos.
- Mezclar técnicas de recuperación de información tradicionales con las basadas en ontologías.
- Evitar el problema de la "Web profunda" para las anotaciones en documentos. Los indexadores de los motores de búsqueda deben poder acceder a la información semántica almacenada en los documentos anotados.
- Enfoque "paga según recibas" para anotaciones semánticas automáticas. Donde el sistema comienza desde un Corpus inicial; conforme el contenido web crece y las anotaciones semánticas son validadas, el Corpus mejora en el tiempo y se consigue un aumento de la efectividad del sistema.

Que sepamos, no existíatodavía ninguna herramienta que reuniera todas las características expuestas anteriormente.

3. La Herramienta FLERSA

En esta sección se presenta la herramienta de anotación semántica llamada FLERSA (FLExible Range Semantic Annotation) que materializa las características deseables presentadas en la sección 2.4 para los sistemas de anotación semántica.

La herramienta está disponible en la dirección http://www.scms.es/joomla. Dispone de un usuario de pruebas (usuario y contraseña: "demo") desde el cual se pueden realizar anotaciones a los documentos web almacenados en el sitio. El sitio también dispone de videos con ejemplos sobre su uso.

Adicionalmente, las herramientas de anotación y búsqueda semántica que aquí se presentan se han implementado como una extensión de un conocido

WCMS llamado Joomla! El componente se llama `com_semantic` y se encuentra disponible para su descarga, bajo licencia GNU/GPL Affero v3, en la dirección web http://salmer.sourceforge.net.

3.1. Características

FLERSA es una herramienta de marcado semántico diseñada para generar anotaciones semánticas en el contenido de documentos web una vez que éstos han sido creados. El creador del documento será el usuario que creará el marcado; los demás usuarios se beneficiarán de la explotación del conocimiento asociado.

La herramienta es fácil de usar desde entorno web; está completamente integrada con los navegadores web de tal forma que los usuarios únicamente tienen que interactuar por medio del ratón con el texto del documento donde se hacen las anotaciones y a través de los menús que la herramienta proporciona.

La herramienta hace uso exclusivo de estándares abiertos tales como RDF, RDFa y OWL con objeto de promover la interoperabilidad y la extensibilidad.

El entorno pOWL [2] proporciona a FLERSA soporte multi-ontología, facilitando la creación de una base de conocimiento compuesta por vocabularios consensuados y taxonomías a partir de las cuales llevar a cabo las anotaciones.

La principal ontología de FLERSA se basa en Annotea, el esquema de anotación propuesto por el W3C. Esta ontología se usa como estructura base de anotación para cualquier anotación semántica en un documento web, de manera que se crea una instancia para cada anotación que se hace. La ontología subyacente también permite la posibilidad de usar vocabularios (microformatos) alternativos cuando se hace una anotación. Se realizará un estudio más profundo en la sección 3.5.

Se ha desarrollado según la arquitectura cliente-servidor, lo que posibilita que múltiples usuarios realicen anotaciones en múltiples páginas web de forma simultánea (centrada en el usuario) y lo que es más importante, que puedan explotar el conocimiento contenido en estos documentos "inteligentes" [5].

En su implementación se ha buscado en la medida de lo posible la compatibilidad cross-browser. No todos los navegadores soportan la especificación W3C DOM Range, por lo tanto sería deseable conseguir la compatibilidad multinavegador desde la implementación. Se ha conseguido la compatibilidad con los navegadores más difundidos como son: IE, Mozilla Firefox, Chrome y Opera.

Se trata de una herramienta cuyo lado cliente posee un nivel de acoplamiento débil, es decir, la implementación realizada en el lado cliente puede ser adaptada fácilmente a otro sistema. Los servicios que proporciona el lado servidor se han integrado a la infraestructura Web subyacente.

La característica principal de la FLERSA es el almacenamiento dual de anotaciones semánticas, esto es, se almacenan en la base de datos del lado servidor en lenguaje de definición de metadatos RDF y, por otro lado, se almacenan incrustadas en el mismo documento donde se anota en lenguaje RDFa de forma totalmente transparente al usuario. Esta característica une las ventajas del modelo de almacenamiento de anotaciones centralizado a las del incrustado: inferencia de nuevo conocimiento a partir de la base de datos de anotaciones,

disponibilidad de anotaciones autocontenidas en el propio documento, el libre acceso a los metadatos de los documentos Web por parte de indexadores, buscadores y otros tipos de servicios semánticos para mejorar las búsquedas y por último la posibilidad de proporcionar información sobre la estructura interna de los documentos, así como la relación entre ellos.

Las principales funcionalidades que aporta la herramienta son:

- Creación de anotaciones a nivel local o global a la página web
- Edición de anotaciones preexistentes
- Borrado de anotaciones.
- Almacenamiento permanente de las anotaciones.
- Visualización del RDF generado en la página (W3C's RDFa Distiller).
- Búsqueda inteligente de anotaciones en base a las propiedades que se han anotado.

3.2. Requisitos de Diseño

Se establecen siete requisitos de diseño para Sistemas de Anotación Semántica en el artículo de Uren et al. [19] que a su vez extienden los establecidos por Handschuh [5]; a continuación se presentan de forma resumida:

- Requerimiento 1 - Formatos estándar.
- Requerimiento 2 - Diseño colaborativo/centrado en el usuario.
- Requerimiento 3 - Soporte de ontologías.
- Requerimiento 4 - Soporte para formatos de documentos heterogéneos.
- Requerimiento 5 - Evolución de documentos.
- Requerimiento 6 - Almacenamiento de anotaciones.
- Requerimiento 7 - Automatización.

Además de los requisitos anteriores, a la hora de diseñar la herramienta se plantearon los siguientes requisitos más específicos:

- Requisito 8 - Integración con la infraestructura de los WCMS.
- Requisito 9 - Interfaz de Usuario Web.
- Requisito 10 - Compatibilidad Cross-browser.

3.3. Arquitectura

La arquitectura del sistema sobre la que se ha desarrollado la herramienta consta de cuatro niveles: el nivel de gestión de información, el nivel núcleo, el nivel semántico del servidor y el nivel Web. Véase a continuación la figura 1.

A `nivel de núcleo` nos encontramos con el sistema operativo y los servicios de red que este aporta, el servidor Web y la infraestructura para trabajar con ontologías.

El `nivel de gestión de información` está formado por los componentes del sistema encargados del almacenamiento tanto de contenido de los documentos Web, como de las anotaciones sobre los mismos y la base de conocimiento compuesta por las ontologías del sistema.

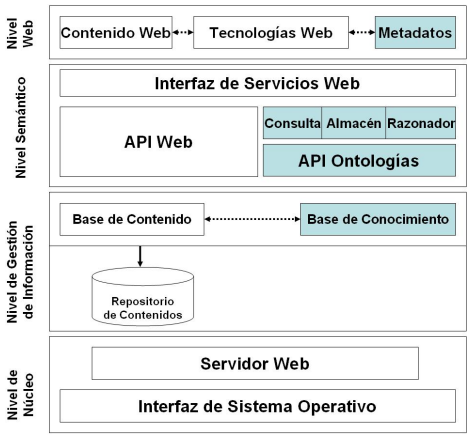

Figura 1. Arquitectura de FLERSA

En el `nivel semántico` de servidor es donde se desarrollan los servicios de aplicación. Todo el tráfico de mensajes entre los clientes Web que demandan servicios y los programas que los proporcionan tienen lugar en este nivel. Aquí se ha realizado la implementación de los programas que dan servicio a la interfaz Web. Los programas implementados aquí hacen uso de librerías de programación o APIs que ofrecen las capas subyacentes. Entre las funciones más usadas que ofrecen estas APIs caben destacar las facilidades de almacenamiento y recuperación de información, las facilidades para trabajar con objetos visuales en la programación del Front-End y las facilidades para trabajar con ontologías.

Por último contamos con un `nivel de interfaz Web`, situado en el nivel superior de la arquitectura del sistema, desde donde el usuario realiza toda la interacción con la herramienta de anotación semántica. En este nivel conviven los contenidos de los documentos Web, los metadatos vinculados a éstos y las tecnologías Web encargadas de modificar en tiempo de ejecución los documentos Web para dotarlos de anotaciones semánticas en forma de metadatos, así como también de realizar la gestión de mensajes oportuna, mediante el uso de servicios del lado Servidor, para aportar la funcionalidad de la herramienta.

3.4. Técnica de Definición de Rangos Flexibles

XPointer es una tecnología robusta como método de localización del componente de texto al que se refiere una anotación, pero presenta problemas cuando se realizan modificaciones sobre un documento sobre el que existen anotaciones. Normalmente, cuando editamos un documento solemos añadir, borrar y/o alterar el orden de los párrafos que lo componen, lo que provoca que se desajusten los puntos de anclaje definidos en XPointer para las anotaciones, por lo que sería necesario repetir el proceso de anotación cada vez que se modifica sustancialmente un documento.

La Técnica de Definición de Rangos Flexibles [11] para documentos web es una alternativa a la tecnología XPointer basada en el estándar RDFa. La herramienta FLERSA usa esta técnica para la delimitación e identificación de fragmentos de texto sobre los que se realizarán las anotaciones semánticas. Su principal objetivo es permitir que las anotaciones semánticas definidas siguiendo esta técnica soporten la evolución del documento web donde se encuentren de forma más efectiva que otras técnicas. La técnica también funciona bien cuando se usa para definir anotaciones sobre fragmentos de texto que se solapan. El término *Rangos Flexibles* indica que las anotaciones pueden ser definidas sobre diferentes rangos de texto y sobre elementos multimedia.

La técnica usa la tecnología DOM Range del W3C ya que la funcionalidad que proporciona permite la identificación y delimitación de fragmentos de texto sobre los que se realizarán las anotaciones semánticas. Esta delimitación se lleva a cabo usando elementos HTML y por lo tanto se almacena incrustada dentro del mismo documento que se anota.

Además, se usa RDFa. Este lenguaje de marcas, unido a la capacidad de la tecnología DOM Range para delimitar fragmentos de texto sobre las que tendrá lugar las anotaciones, permite insertar las anotaciones que se hacen dentro del documento donde se anota, resolviendo los problemas que presenta XPointer.

Se pueden encontrar ejemplos ilustrativos de la técnica en la sección 3.6, cuando se usa en el proceso de marcado manual. También se puede encontrar una completa descripción en [11].

3.5. Ontología FLERSA

En general, algunas de las grandes ventajas que ofrece el uso de ontologías son: permiten definir vocabularios consensuados, separan el conocimiento del dominio del operacional, permiten la reutilización de conocimiento dominio.

En FLERSA se ha reutilizado, adaptado y enriquecido la estructura por defecto definida por el W3C en Annotea. De aquí en adelante nos referiremos a la ontología base para la anotación semántica con el nombre FLERSA-ontology.

El esquema básico de FLERSA-ontology está compuesto por cinco clases OWL: `Annotation`, `AnnnotationType`, `GranularityType`, `AnnotationSection` y `ReferenceOntology`.

Las anotaciones semánticas, desde un punto de vista ontológico, son consideradas elementos o individuos que representan anotaciones concretas en documentos siguiendo el modelo formal definido en la ontología base. Estos individuos corresponden a hechos concretos de la clase `Annotation` que se define en la ontología base. Las propiedades de las que dispone son las siguientes:

- **Annotates**: Asocia una anotación con el recurso que se anota.
- **Author**: El nombre del usuario responsable de la creación de la anotación.
- **Body**: Fragmento de texto de la página Web que se anota objeto de la anotación semántica. Se almacena para facilitar búsqueda.
- **Context**: Corresponde con la URI que delimita la posición donde se encuentra el texto u objeto multimedia que se anota.

- **Created**: Sirve para indicar la fecha y hora de creación de la anotación.
- **Modified**: Ídem con la fecha y hora de última modificación de la anotación.
- **Related**: Asocia una anotación con el concepto del que se habla en ella.
- **Granularity**: Asocia a la anotación un concepto dentro de la taxonomía que ofrece la clase `GranularityType` indicando el tipo de granularidad de la anotación: *carácter, palabra, frase, párrafo o texto libre.*
- **Section**: Asocia a la anotación un concepto dentro de la taxonomía que ofrece la clase `AnnotationSection` indicando la sección dentro de la página Web donde se ha realizado de la anotación: *texto e imagen.*
- **Type**: Asocia a la anotación un concepto dentro de la taxonomía que ofrece la clase `AnnotationType` indicando el tipo de anotación que se ha realizado: *example, advice, change, seealso, explanation, question y comment.*

En principio, las anotaciones estándar no proporcionan ventajas semánticas por sí mismas, son anotaciones simples en las que sólo se definen propiedades básicas. Es necesario un proceso adicional de edición y asociación a ontologías que permita enriquecerlas y dotarlas de funcionalidad semántica. Las posibilidades que ofrece la ontologia FLERSA son:

- Añadir sentencias que describan de que temática se habla en el fragmento de texto asociado. Para realizar esta tarea se utiliza la propiedad **Related** y se usan como valores los conceptos que proporcionan las taxonomías de la Base de Conocimiento.
- Definición de individuos pertenecientes a distintos conceptos en base a las anotaciones semánticas realizadas en un documento.

3.6. Proceso de Marcado Manual

El proceso manual de anotación semántica fue presentado en [12]. Se basa en el uso de la técnica de definición de rangos flexibles explicada en la sección 3.4. Los pasos del proceso son los siguientes:

1. Selección del fragmento de texto sobre el cual se desea realizar la anotación. En el contexto Web, se realizará normalmente mediante el ratón haciendo una selección de texto.

```
1 <p>Lorem ipsum dolor sit amet, consectetur adipiscing elit.</p>
2 <ul><li>Duis orci tellus, dignissim ac laoreet sit amet, porttitor et purus. </li></ul>
3 <p>Mauris congue ultrices sodales. Vivamus dignissim tristique leo, sit amet posuere ipsum
      hendrerit id.</p>
```

Ejemplo 1.1. Código HTML correspondiente a una selección.

2. Se genera un identificador global que se usa para el marcado del fragmento de texto seleccionado y la posterior definición de la anotación semántica. También es necesario seguir una estrategia que permita la identificación local de los elementos HTML que pertenecen al fragmento de texto seleccionado y la definición de la relación de pertenencia. El identificador global se

usará posteriormente, como punto de referencia (anchor-anclaje) en el resto de labores de anotación semántica. Los identificadores globales serán usados para asociarles metadatos.

```
<p>Lorem ipsum dolor sit amet, <span id="654-1">consectetur adipiscing elit.</span></p>
    <ul><li><span id="654-2">Duis orci tellus, dignissim ac laoreet sit amet,
        porttitor et purus.</span></li></ul>
<p><span id="654-3">Mauris congue ultrices sodales.</span> Vivamus dignissim tristique leo
    , sit amet posuere ipsum hendrerit id.</p>
```

Ejemplo 1.2. Marcado de un fragmento de texto.

3. Se generan las sentencias necesarias para definir una nueva instancia o individuo del concepto anotación definido en la ontología FLERSA. Cada una de las sentencias describen uno de los atributos de FLERSA presentados en la sección 3.5. Las sentencias pueden expresarse bien en lenguaje RDF y almacenarse en la Base de Conocimiento, o expresarse en RDFa y almacenarse de forma incrustada en el mismo documento que se anota.

```
<div xmlns:r="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:a="http://www.w3.org/2000/10/annotation-ns#"
    xmlns:t="http://www.w3.org/2000/10/annotationType#"
    xmlns:dc="http://purl.org/dc/elements/1.1/">

    <span id="654" about="http://w3.ex.org/p.htm#654" rel="rdf:Seq">
      <span rel="rdf:li" resource="http://w3.ex.org/p.htm#654-1"/>
      <span rel="rdf:li" resource="http://w3.ex.org/p.htm#654-2"/>
      <span rel="rdf:li" resource="http://w3.ex.org/p.htm#654-3/></span>

    <span typeof="a:Annotation"
          about="http://ex.org/p.htm#654"></span>
    <span resource="http://ex.org/p.htm" rel="a:Annotates"
          about="http://ex.org/p.htm#654"></span>
    <span content="José Luis Navarro" property="dc:creator"
          about="http://ex.org/p.htm#654"></span>
    <span content="20/7/2009" property="a:Created"
          about="http://ex.org/p.htm#654"></span>
</div>
```

Ejemplo 1.3. Anotación semántica incrustada en HTML.

3.7. Proceso de Marcado Automático

El proceso automático de anotación semántica también fue presentado en [12]. Es posible entrenar la herramienta FLERSA para que, automáticamente, se establezcan relaciones entre las anotaciones semánticas de un documento y los conceptos que proporcionan distintas taxonomías de la Base de Conocimiento.

En el Modelo de Espacio Vectorial (VSM) se considera que cada documento, perteneciente a una colección, es un vector de pesos en un espacio vectorial de T dimensiones, donde T es el número de términos diferentes que aparecen en la colección.

$$D_i = (d_{i1}, d_{i2}, d_{i3}, ..., d_{in}). \quad (1)$$

Un n-grama es una subsecuencia de n-elementos pertenecientes a una secuencia dada. En la herramienta FLERSA se usan monogramas, bigramas y trigramas a modo de términos; es por ello que se habla de un enfoque híbrido VSM + N-gramas.

Cuando se anota un documento web en modo automático, la herramienta FLERSA es capaz de trabajar tanto a nivel global como a nivel local. A nivel global, se considera todo el texto del documento web. A nivel local, el documento es dividido en fragmentos de texto a nivel de párrafo. Después, se lleva cabo un proceso de categorización del texto para cada fragmento.

El proceso de anotación de documentos comienza manualmente, asociando conceptos pertenecientes a ontologías de referencia (usadas a modo de taxonomías) a nivel de documento, sección, párrafo, frase y otros niveles. En esta etapa, el ingeniero del conocimiento define las anotaciones básicas que formarán el Corpus. Esta es la fase de entrenamiento; se necesita, al menos, una anotación manual para cada concepto de la taxonomía que se quiere entrenar para ser usada en anotaciones automáticas. Una vez entrenado el sistema, el proceso automático de anotación semántica consiste en los cuatro pasos que se describen a continuación:

1. Un nuevo texto de entrada -un fragmento de texto o un documento completo- llega al sistema para su clasificación. El sistema tiene, al menos, una anotación manual para cada concepto con el que trabaja el sistema y que forma el Corpus del concepto.
2. Se calculan los pesos de los N-gramas que componen el texto de entrada siguiendo la ecuación 2. El sistema ha precomputado previamente los perfiles de cada concepto modelado en las anotaciones del Corpus.

$$w_i = tf_i * log\left(\frac{D}{df_i}\right). \tag{2}$$

   donde:
   - `tf` es la frecuencia de ocurrencia en el documento [8]. Las palabras más repetidas dentro de un documento son, en principio, más relevantes que las menos usadas.
   - `df` es el número de documentos en la colección en los cuales el término aparece [17].
3. El sistema compara el perfil del texto de entrada con respecto a los perfiles precomputados de cada uno de los conceptos modelados en el sistema. Se usa la medida de la distancia entre perfiles que proporciona la ecuación 3 y que se calcula fácilmente. Se calcula la similaridad de monogramas, bigramas y trigramas separadamente, de manera que se necesita un valor de similaridad ponderado, como en la ecuación 4.

$$Sim(Q, D_i) = \frac{\sum_i w_{Q,j} w_{i,j}}{\sqrt{\sum_j w_{Q,j}^2}\sqrt{\sum_i w_{i,j}^2}}. \tag{3}$$

$$Sim_{Global}(Q, C_i) = 0{,}7 * Sim_{Tri}(Q, C_i) + 0{,}2 * Sim_{Bi}(Q, C_i) + 0{,}1 * Sim_{Mono}(Q, C_i). \quad (4)$$

donde:

- W_{ij} muestra los pesos del concepto "i" para el término j-ésimo.
- W_{Qj} muestra los pesos de los términos de consulta.

4. El sistema clasifica el fragmento de texto como perteneciente al concepto que tenga la similaridad más alta. También se considera la posibilidad de no clasificar el texto de entrada cuando su valor de similaridad es inferior al de un valor umbral. Cuando un fragmento de texto supera el valor umbral, es correctamente clasificado y se le asigna una anotación semántica similar a las estudiadas en la sección 3.6, en la que se incluye la información de categorización (propiedad `Related`).

3.8. Búsqueda Semántica

La principal ventaja que supone el enriquecimiento con metadatos de los contenidos web es la mejora en la calidad de los resultados que se obtienen cuando se realizan tareas de recuperación de información gracias a la explotación de las anotaciones semánticas introducidas.

En particular, en la herramienta FLERSA se combinan las técnicas tradicionales de recuperación de información con técnicas basadas en la semántica introducida por las anotaciones. Los tipos de búsqueda que permite realizar son los siguientes:

- **Basada en palabras clave**. Se trata del tipo de búsqueda tradicional en la que a partir de unas palabras clave introducidas por el usuario a modo de términos de búsqueda, se presentan como resultados los contenidos web donde se han localizado. Su principal problema es que no ofrecen buenos resultados cuando se usa el lenguaje natural como términos de búsqueda. Por ejemplo, un usuario no será capaz de obtener buenos resultados al utilizar los términos "viajes a Ámsterdam para la primera quincena de Agosto".
- **Basada en las propiedades de las anotaciones**. La forma más simple de explotar las anotaciones semánticas introducidas en un contenido web es realizar consultas basadas en cualquiera de las propiedades de la clase `Annotation` estudiadas en el apartado 3.5. La propiedad `Related` es la más importante puesto que en ella se especifica el concepto del que trata la anotación, y es muy útil para obtener mejoras en las búsquedas ya que además ofrece la posibilidad de extender las búsqueda mediante la inferencia nuevos conceptos específicos a partir de un concepto más genérico, todo ello dentro del ámbito de las ontologías de referencia que se usen a modo de taxonomías.
- **Basada en conceptos**. Es ella se combinan las dos búsquedas anteriores. La idea es realizar un mapeo de palabras clave con respecto a los conceptos de una taxonomía. Cuando el usuario introduce unos términos de búsqueda, el sistema determina automáticamente los conceptos que están asociados

a ellos y ofrece como resultados los contenidos web en cuyas anotaciones semánticas figuren los conceptos que son objeto de búsqueda. En este tipo de búsqueda es posible asociar al mismo concepto tanto términos equivalentes como sinónimos de forma que se mejora la búsqueda tradicional. También se permite inferir conceptos específicos a partir de más generales, ofreciendo así mejores resultados.

- **Consultas en lenguaje natural**. Es el tipo de búsqueda más acorde con la filosofía de Web Semántica. El usuario realiza una consulta expresada en lenguaje natural y el sistema la analiza, siguiendo el método híbrido VSM + N-gramas estudiado en el apartado 3.7, para determinar los conceptos objeto de búsqueda. Finalmente se presentan los contenidos web en cuyas anotaciones semánticas figuran éstos conceptos.

Está claro que los tipos de búsqueda explicados anteriormente presentan muchas ventajas, aunque se usan sólo para realizar búsqueda contextual de información contenida en un WCMS particular. Llegado este punto, podemos cuestionarnos cómo contribuye la herramienta FLERSA a realizar una aproximación de la Web actual en la dirección que marca la Web Semántica. La respuesta es que en caso de que FLERSA recibiera una amplia difusión por todos los sitios web donde se usan WCMS, gracias a su principal característica, la de almacenar en RDFa las anotaciones semánticas incrustadas en los documentos que se anotan, los indexadores de los motores de búsqueda Web tendrían a su disposición un cantidad ingente de metadatos que se podría explotar para mejorar las búsquedas. Los buscadores web deberían incorporar una infraestructura para trabajar con ontologías, así como ontologías de dominio que funcionaran a modo de "piedra roseta" permitiendo el mapeo de distintos conceptos, motores de inferencia, etc; aunque este estudio queda fuera del ámbito del presente trabajo.

4. Conclusiones y Trabajo Futuro

En este artículo, se ha presentado FLERSA, una herramienta de anotación semántica en la que se han desarrollado todas las características deseables estudiadas en la sección 2.4. Éstas son: centrada en el usuario, entorno de trabajo ligero, anotaciones manuales y automáticas, enfoque "paga según recibas", evita el problema de la "Web profunda" y permite la recuperación de información usando tanto técnicas tradicionales como semánticas.

Cabe destacar: la técnica de definición de rangos flexibles que permite la definición de anotaciones semánticas incrustadas en el mismo documento que se anota y la evolución del mismo frente a cambios, los diferentes tipos de búsqueda semántica local que explota la semántica de las anotaciones del sistema y la posibilidad de acceso a los metadatos que ofrece a los indexadores de los motores de búsqueda.

Como trabajo futuro, se está trabajando en el diseño de un método híbrido y unificado de búsqueda que proporcione una combinación flexible de los métodos de búsqueda tradicionales, basados en palabras clave, con los métodos de búsqueda semánticos, basados en anotaciones semánticas y metadatos.

Referencias

1. Adida, B., Birbeck, M.: RDFa primer 1.0 embedding RDF in XHTML. W3c working draft, W3C (October 2007), http://www.w3.org/TR/2007/WD-xhtml-rdfa-primer-20071026/
2. Auer, S.: Powl - a web based platform for collaborative semantic web development. In: Proceeding of 1st Workshop Scripting for the Semantic Web (SFSW'05), Hersonissos, Greece, May 30. CEUR Workshop Proceedings (May 2005)
3. Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., Yergeau, F.: Extensible markup language (xml) 1.0 (fifth edition). World Wide Web Consortium, Recommendation REC-xml-20081126 (November 2008)
4. DeRose, S., Maler, E., Daniel, R.: Xml pointer language (xpointer) version 1.0. Tech. rep., W3C (2001), candidate Recommendation 11 September 2001
5. Handschuh, S., Staab, S., Studer, R.: Leveraging metadata creation for the semantic web with cream. In: Gunter, A., Kruse, R., Neumann, B. (eds.) KI. Lecture Notes in Computer Science, vol. 2821, pp. 19–33. Springer (2003)
6. Kahan, J., Koivunen, M.R.: Annotea: an open rdf infrastructure for shared web annotations. In: WWW '01: Proceedings of the 10th international conference on World Wide Web. pp. 623–632. ACM Press, New York, NY, USA (2001)
7. Kesselman, J., Robie, J., Champion, M.: Document object model (dom) level 2 traversal and range specification. W3C Recommendation (November 2000), http://www.w3.org/TR/DOM-Level-2-Traversal-Range
8. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2, 159–165 (April 1958), http://dx.doi.org/10.1147/rd.22.0159
9. Maedche, A., Motik, B., Stojanovic, L., Studer, R., Volz, R.: Ontologies for enterprise knowledge management. IEEE Intelligent Systems 18(2), 26–33 (2003)
10. Manola, F., Miller, E.: Rdf primer. W3c recommendation, W3C (February 2004), http://www.w3.org/TR/2004/REC-rdf-primer-20040210/
11. Navarro-Galindo, J.L., Samos, J.: Flexible range semantic annotations based on rdfa. 27th BNCOD: Data security and security data (June 2010)
12. Navarro-Galindo, J.L., Samos, J.: Manual and automatic semantic annotation of web documents: The flersa tool. 12th iiWAS 2010 1, 540–547 (November 2010)
13. Norman, D.A., Draper, S.W.: User Centered System Design; New Perspectives on Human-Computer Interaction. L. Erlbaum Inc., Hillsdale, NJ, USA (1986)
14. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: Kim - a semantic platform for information extraction and retrieval. Nat. Lang. Eng. 10, 375–392 (September 2004), http://portal.acm.org/citation.cfm?id=1030318.1030327
15. Quint, V., Vatton, I.: An introduction to amaya. World Wide Web J. 2, 39–46 (April 1997), http://portal.acm.org/citation.cfm?id=275062.275068
16. Sheth, A., Bertram, C., Avant: Managing semantic content for the web. IEEE Internet Computing 6, 80–87 (July 2002)
17. Sparck Jones, K.: A statistical interpretation of term specificity and its application in retrieval, pp. 132–142. Taylor Graham Publishing, London, UK, UK (1988)
18. Tsai, T.M., Yu, H.K., Liao, P.Y., Shih, H.T.: Semantic modeling among web services interfaces for services integration. In: Proceedings of the 14th International Workshop on Database and Expert Systems Applications. pp. 579–. DEXA '03, IEEE Computer Society, Washington, DC, USA (2003)
19. Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., Ciravegna, F.: Semantic annotation for knowledge management: Requirements and a survey of the state of the art. Web Semant. 4, 14–28 (January 2006)