Arquitectura para el manejo del modelo de usuario en una biblioteca digital

Ing. Reydi Castillo Buergo

Las bibliotecas digitales son uno de los tipos de gestores de contenido más empleados en la actualidad por centros e instituciones. El volumen de datos que en ellas se almacenan, la diversidad de formatos que pueden emplearse, los múltiples dominios del conocimiento presentes en sus contenidos y la complejidad de los equipos humanos que las crean y las mantiene, provocan grandes dificultades de gestión, sobre todo, si se pretende ofertar una atención personalizada para sus usuarios. En este trabajo se propone una arquitectura para el manejo del modelo de usuario en una biblioteca digital. Dicha arquitectura se centra en la integración de tecnologías emergentes, basadas en semántica y diversas técnicas de adquisición de preferencias de usuarios, a fin de facilitar la generación del modelo de usuario que interactúa con una biblioteca digital, anotar semánticamente los objetos digitales que contiene una biblioteca digital y gestionar contenidos teniendo en cuenta su significado de acuerdo a las preferencia semánticas de los usuarios.

***Palabras clave**: personalización, modelado de usuarios, bibliotecas digitales, Web semántica, ontologías, extracción de información, anotación semántica, búsqueda semántica.*

RESUMEN

ABSTRACT

Digital libraries are one of the content managers that centers and institutions use more frequently nowadays. The data volume they can store, the diversity of formats that can be used, the many domains of knowledge that are present in their contents, and the complexity of human teams that create and maintain them result in great management difficulties, mostly if the idea is to provide users with a customized service. This paper proposes an architecture to manage the user model in a digital library. Said library is focused on integrating emerging technologies based on semantics and diverse techniques for capturing user preferences, in order to facilitate generation of the user model that interacts with a digital library, to semantically annotate the digital objects a digital library contains, and to manage contents taking into account their meaning according to the semantic preferences of users.

***Keywords**: customized, user modeling, digital libraries, Semantic Web, data mining, semantic annotation, semantic search.*

Introducción

El creciente desarrollo en las últimas décadas de las Tecnologías de la Información y las Comunicaciones, ha traído consigo un considerable aumento del volumen de información contenida en Internet, así como el surgimiento de una gran diversidad de formatos para su almacenamiento. Estas transformaciones han provocado que los mecanismos de los cuales está provista la Web actual, se vuelvan insuficientes para la correcta manipulación y comprensión de dicha información; resultando más compleja la implementación de procedimientos automáticos para su recuperación en base al significado.

En respuesta a la intensificación de este problema, a finales de los años 1990 [1] y en

un mayor desarrollo en los principios del 2000 [2], se torna prominente la necesidad de desarrollar una Web en la cual cada recurso tenga asociado una descripción de su significado. Esta nueva Web fue denominada por Tim Berners-Lee como *Web Semántica*, y tiene como principio fundamental la creación de estructuras de datos que describan semánticamente el contenido de cada recurso. El desarrollo de esta visión de la Web brinda innumerables beneficios como son: la localización, integración y reutilización de recursos entre aplicaciones, empresas y comunidades de desarrollo; la organización y búsqueda de información sobre la base de su significado y no justamente por el texto y la desambiguación conceptual, pues los sistemas semánticos pueden distinguir cuándo las palabras y frases son equivalentes, así como cuándo son utilizadas con diferentes significados. Esta iniciativa mantiene a su vez los principios que han hecho de la Web actual un verdadero éxito como son: la descentralización, el intercambio y la compatibilidad [3].

Las posibilidades que ofrece la *Web Semántica* han dado paso al incremento de las investigaciones en esta área, y aún cuando su tiempo de materialización no está claramente definido, cada día se escala un peldaño y se aplican sus principios en la solución de múltiples tareas. Las ontologías han sido consideradas como el centro de las aplicaciones para la Web Semántica, por lo cual se han desarrollado múltiples propuestas en el ámbito de la gestión semántica de contenidos basadas en ontologías; algunas de estas enfocadas a la recuperación a través de consultas en lenguaje natural [4] [5] [6], otras en la recuperación personalizada [7] [8] [9] [10]. También se han presentado aproximaciones vinculadas a la anotación semántica [11] [12] [13], a la población de ontologías y bases de conocimiento [14], y otros métodos que contribuyen a incrementar la eficiencia en la gestión de contenidos teniendo en cuenta su significado. Existen herramientas y sistemas que implementan estas propuestas en la solución de tareas específicas tales como la gestión de noticias [15] [16] [17], la gestión de contenidos multimedia [18] y la gestión de contenidos en bibliotecas digitales.

Un concepto que se ha puesto de moda es el de la Web 2.0. Entre los principios que se deben tener en cuenta en las aplicaciones orientadas a esta Web se encuentra la atención personalizada a los usuarios, tanto en la presentación visual como en la información que es presentada. Determinar el tipo de contenido por el cual se interesan usuarios particulares desde el punto de vista de su significado, es una tarea difícil de lograr por parte de sistemas informáticos, sin embargo, la introducción de las ontologías en la descripción de las preferencias semánticas de usuarios promete resultados relevantes. Las ontologías facilitan la formalización de las preferencias de los usuarios en una representación interoperable, donde los intereses pueden acoplarse con el significado de contenidos, facilitando la implementación de diversos mecanismos de gestión orientados a satisfacer la demanda de usuarios interesados en contenidos pertenecientes a áreas específicas del conocimiento.

Las bibliotecas digitales son uno de los tipos de gestores de contenido más empleados en la actualidad. En estas, el contenido almacenado es catalogado mediante algún estándar para la formalización de metadatos, lo cual ayuda en la organización y posterior recuperación de la información. Aunque el empleo de metadatos para la gestión de contenidos es de gran ayuda, esta metodología no soluciona las necesidades actuales de gestión en base al significado y orientada a la satisfacción de los intereses de usuarios específicos. Este trabajo tiene como objetivo diseñar una arquitectura para el manejo del modelo de usuario en una biblioteca digital. Dicha arquitectura está integrada por tecnologías de la Web Semántica y diversas técnicas de adquisición de preferencias de usuarios, que contribuyen a elevar su satisfacción de una biblioteca con el contenido que se le oferta.

Arquitectura para el manejo del modelo de usuario en una biblioteca digital

La arquitectura se centra en el uso de tecnologías de la *Web Semántica* y diversas técnicas de adquisición de preferencias de usuarios para:

(1) facilitar la generación del modelo de usuario que interactúa con una biblioteca digital, (2) anotar semánticamente los objetos digitales (OD) que contiene una biblioteca digital y (3) gestionar contenidos teniendo en cuenta su significado y de acuerdo a las preferencia semánticas de los usuarios.

El modelado de usuario que se propone consiste en la creación de perfiles de usuarios mediante el uso de técnicas basadas en ontologías. Las ontologías facilitan la formalización de las preferencias de los usuarios en una representación interoperable, donde los intereses de los usuarios pueden acoplarse con el significado de contenidos almacenados en bibliotecas digitales, en un nivel que permita el razonamiento conceptual. Las preferencias semánticas de un usuario pueden representarse mediante vectores de pesos que indiquen la intensidad en la que está interesado en cada concepto (en esta propuesta cuando se refiere a un concepto, puede ser una clase o una instancia) de diversas ontologías de dominio. De esta forma, si dichas ontologías son compartidas tanto en la personalización de usuarios como en la anotación semántica de OD, entonces es fácil recuperar OD que incluyan contenidos del interés de los usuarios.

Comúnmente la anotación semántica se realiza mediante el análisis de textos en lenguaje natural, con el fin de extraer estos metadatos semánticos que permitan vincularlo con las ontologías. Sin embargo, en una biblioteca digital pueden almacenarse contenidos en cualquier formato, ya sean textos en lenguaje natural, imágenes, videos, entre otros. Por esta razón, la anotación semántica que se propone parte del análisis de los metadatos que describen a los contenidos almacenados en una biblioteca digital. Aunque los metadatos semánticos constituyen un paso de avance en la descripción de la semántica de un recurso cualquiera, no permiten su descripción con toda la expresividad que se necesita, los textos en lenguaje natural que se especifican en ellos no pueden ser interpretados por los ordenadores, y no permiten el razonamiento o inferencia de nuevo conocimiento.

Para lograr la anotación semántica de los OD almacenados en una biblioteca digital, se propone analizar sus metadatos haciendo uso de técnicas de Extracción de Información Basada en Ontologías (OBIE por sus siglas en inglés) [19]. La OBIE permite el reconocimiento de entidades presentes en los textos que se procesan, en este caso los metadatos, y su vinculación con conceptos de ontologías. Partiendo de la concepción de que las entidades mencionadas en un texto constituyen una parte importante de su semántica, entonces serán vinculados los OD a los conceptos que fueron vinculadas las entidades reconocidas en sus metadatos. Al igual que como se propuso para el modelado de usuarios, se pueden crear vectores de pesos que indiquen la intensidad con la que se relaciona un OD con conceptos de diversas ontologías de dominio.

Una vez anotados semánticamente los OD almacenados en una biblioteca digital, y creados los perfiles semánticos para usuarios, las ontologías que son compartidas en ambas tareas se convierten en el medio que establece la conexión entre un usuario y los OD que incluyen el contenido de su interés. Esta arquitectura habilita una nueva capa de datos semánticos en una biblioteca digital que facilita el desarrollo de mecanismos de gestión de contenidos, ya sean teniendo en cuenta las preferencias de los usuarios, como son los métodos de recomendación, o solo sobre las descripciones semánticas de los OD. La figura 1 ilustra de forma general la arquitectura y en los siguientes epígrafes se describen en detalles todos los componentes que la integran.

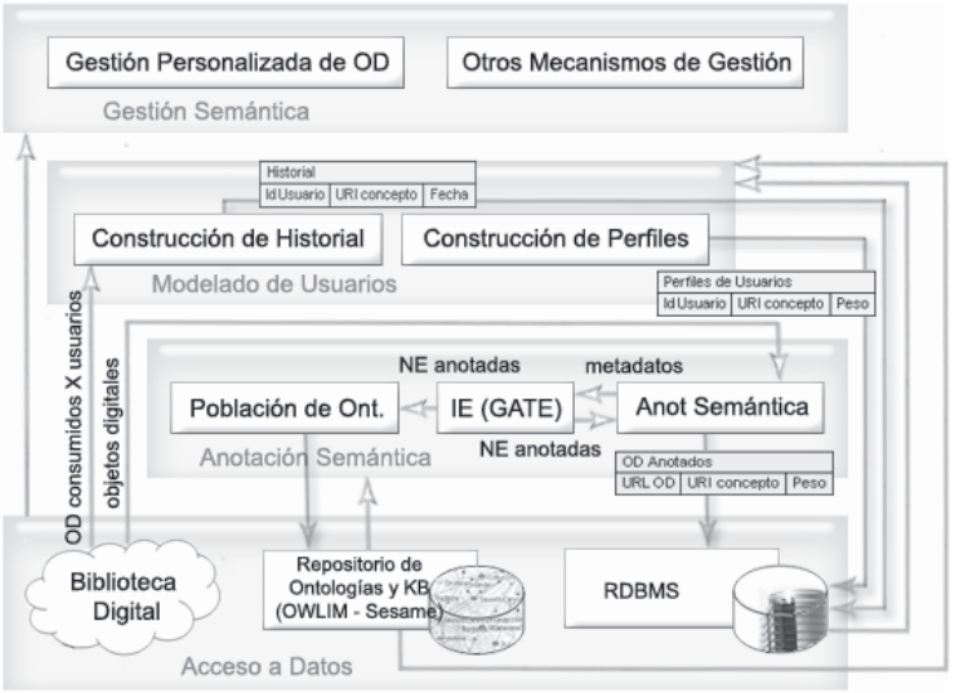

Fig. 1. Arquitectura para el manejo del modelo de usuario en una biblioteca digital.

Anotación semántica de OD

Los metadatos que describen el contenido envuelto en un OD expresan gran parte de su semántica. Dependiendo del estándar que se emplee para su formalización, estos metadatos contienen información de gran relevancia como el título del contenido, una descripción, las categorías en que se clasifica, el autor, fecha de creación, etcétera. Si haciendo uso de métodos de OBIE analizamos estos metadatos, sería posible reconocer las entidades presentes en ellos y obtener tanto la clase a la que pertenecen en una ontología como la instancia más específica con la que se identifica. De esta forma, un OD puede anotarse y vincularse a conceptos de ontologías de una manera formal a partir de las entidades reconocidas en sus metadatos.

Para representar una anotación semántica existe un conjunto de prerrequisitos [12]:

- Una ontología, o al menos una taxonomía, donde se definan las clases de entidades que serán referenciadas (ver epígrafe 2.8).

- Un identificador de entidades que permita reconocer estas y vincularlas a sus descripciones semánticas (ver epígrafe 2.2).

- Una Base de Conocimientos (KB por sus siglas en inglés) con las descripciones de las entidades (ver epígrafe 2.8).

Las anotaciones semánticas que se proponen aquí no son tomadas dentro de los OD, sino que se almacenan separadas de estos, con una estructura que permite recuperar los OD que han sido anotados por un concepto de una ontología. Cada anotación será almacenada en una tabla de una base de datos que contendrá la URL del objeto que se anota, la URL de un concepto de una ontología (la clase a la cual pertenece o la instancia más específica con la cual se identifica), y un peso que refleja la intensidad con la que se relaciona el OD con el concepto. La figura 2 ilustra la estructura de la anotación.

Es importante poder determinar la intensidad de la relación, o sea, el nivel de relevancia del contenido envuelto en un OD con respecto al concepto con el cual se anota. Los pesos en las anotaciones permiten discriminar o priorizar un OD en la recuperación. El algoritmo para calcular los pesos de las anotaciones es el propuesto [5], basado en una adaptación del modelo vector-espacio clásico. En esta adaptación el peso $\boldsymbol{d}_x$ de un concepto $\boldsymbol{x}$ para un documento $\boldsymbol{d}$ es calculado como:

$$d_x = \frac{freq_{x,d}}{\max_y frec_{y,d}} * \log \frac{|D|}{n_x}$$

Fórmula 1. Cálculo del peso de las anotaciones.

Donde $\boldsymbol{freq}_{x,d}$ es el número de ocurrencias en $\boldsymbol{d}$ del concepto $\boldsymbol{x}$, $\boldsymbol{max}_y\boldsymbol{freq}_{y,d}$ es la frecuencia de ocurrencia del concepto más repetido en $\boldsymbol{d}$, $\boldsymbol{n}_x$ es el número de documentos anotados con $\boldsymbol{x}$ y $\boldsymbol{D}$, es el conjunto de todos los documentos en el espacio de búsqueda. En este caso se consideran como documentos los metadatos de los OD. El rango de valores de $\boldsymbol{d}_x$ puede ir desde 0 hasta 1, donde los valores cercanos a uno indican un nivel de relevancia superior, y valores cercanos a cero menor.

Un problema que se introduce al aplicar este algoritmo para el cálculo de los pesos de las anotaciones es la aparición en el texto de una misma entidad referenciada con nombres diferentes, abreviaturas o a través de artículos. Por ejemplo, en un texto pudiera ser identificado como una instancia «Tom Cruise», y posteriormente aparecer la misma instancia como «El actor», o cualquier artículo

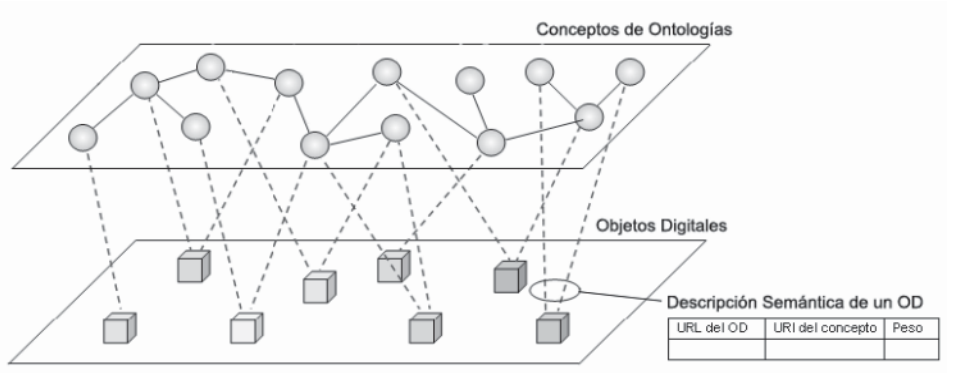

Fig. 2. Anotación semántica de objetos digitales (basado en [10]).

que haga referencia a este. Si estas referencias no son detectadas como una misma entidad, entonces se introducen imprecisiones en el método de anotación, que pueden provocar la creación de varias anotaciones para la misma entidad y un cálculo inexacto del peso de las anotaciones.

Para darle solución a este problema se pueden utilizar técnicas de Procesamiento del Lenguaje Natural (NLP por sus siglas en inglés) como la resolución de co-referencias, para identificar los artículos que hacen referencia a una entidad, o entidades referenciadas con nombres diferentes. También se propone el apoyo en la jerarquía de clases y la inclusión en la KB de todos los alias posibles que puede tener una instancia.

Extracción de información semántica sobre metadatos de OD

La Extracción de Información (IE por sus siglas en inglés) [19] se basa en la búsqueda de cinco tipos diferentes de información en textos en lenguaje natural: reconocimiento de NE (del inglés *Named Entity*), búsqueda de menciones mediante CO (del inglés *Co-reference Resolution*), Extracción de Descripciones o TE, Identificación de Relaciones o TR (del inglés *Template Relation*) y Extracción de Eventos o ST (del inglés *Scenario Template*). De estos tipos de información, el reconocimiento de NE es la tarea que mayor nivel de confiabilidad ofrece; además de constituir la base para las demás tareas mencionadas.

En el campo del NLP y en particular en la IE, se le denomina entidades nombradas o NE a las personas, organizaciones, lugares, o cualquier otra entidad que sea referenciada a través de su nombre. Para una interpretación más amplia, las NE también incluyen valores escalares (números, fechas, cantidades de dinero), direcciones, etcétera. Las NE deben tratarse de una forma diferente porque su naturaleza y semántica es diferente al de una palabra (términos, frases, etc.) Mientras las NE denotan individuos o instancias particulares, las palabras denotan conceptos, clases, relaciones o atributos universales. Las palabras pueden ser descritas a través del significado común del léxico, sin embargo, la comprensión y gestión de NE requiere de un conocimiento del mundo más específico [12].

En sistemas de reconocimiento de NE tradicionales se utilizan tipos demasiado genéricos (organización, localización, persona, etc.) para clasificar las anotaciones producidas, por lo que la semántica no queda claramente definida. Este problema puede resolverse mediante una infraestructura para la OBIE. El enfoque que se propone está basado en el reconocimiento de NE presentes en los metadatos que describen OD con respecto a ontologías de dominio. De esta forma las entidades reconocidas pueden ser clasificadas con respecto a clases de ontologías y vinculadas exactamente a un individuo que posee su descripción semántica en una base de conocimientos.

La IE que se propone está basada en GATE (del inglés *General Architecture for Text Engineering*) [20], la cual ofrece una infraestructura para el desarrollo de componentes para el NLP. Además, cuenta con ANNIE (del inglés *A Nearly-New Information Extraction System*), una herramienta especializada en la IE a la cual se le puede incorporar información ontológica con el fin de incorporar semántica en el proceso. ANNIE está integrada por un conjunto de recursos básicos para el NLP tales como Tokeniser, diccionarios, separador de sentencias, marcador de partes del discurso, reconocedor de NE mediante patrones gramaticales y resolución de co-referencias ortográfica, nominal y pronominal; que pueden ser utilizados en cualquier tarea que necesite IE. También GATE incorpora otros componentes que soportan el trabajo con ontologías. En la figura 3 se muestra la secuencia por la que transitan los metadatos en la IE sobre esta arquitectura, y cómo cada proceso puede ser apoyado con información ontológica.

El primer proceso que puede ser apoyado con información ontológica es el reconocimiento de NE mediante el mapeo con diccionarios. En sistemas tradicionales de IE, esta tarea se apoya en diccionarios planos que contienen un conjunto de NE de interés general y las anotaciones que se producen son marcadas con un tipo genérico. Para incorporar semántica en este proceso se crean para cada concepto de una ontología, un diccionario que almacena sus instancias, y un fichero de mapeo que permite identificar a qué concepto de cuál ontología pertenece un diccionario. De esta forma las NE reconocidas en un metadato pueden ser clasificadas con clases de ontologías y vinculadas a una instancia en la base de conocimientos que describe su semántica. Otra ventaja que ofrece el uso de ontologías en el trabajo con diccionarios es que las entidades pueden ser reconocidas a partir de cualquiera de sus alias presentes en los metadatos, a través del mapeo de los alias de las entidades almacenadas en la base de conocimientos a los diccionarios.

Otro proceso al que puede incluírsele semántica es el reconocimiento de NE mediante patrones gramaticales. Este proceso permite el reconocimiento de NE a partir de reglas gramaticales basadas en JAPE (del inglés *Java Annotation Patterns Engine*), un procesador que implementa un lenguaje para definir reglas a partir de las cuales reconoce expresiones regulares en anotaciones de documentos e infiere nuevas anotaciones. Estas reglas pueden construirse basadas en las clases de una ontología y no en un conjunto plano de tipos de entidades. De esta forma habría mucha más flexibilidad en la creación de reglas, dando la posibilidad de crearlas tanto para tipos de NE más específicas como para tipos generales.

Para el caso de la resolución de co-referencias, que identifica las menciones a NE reconocidas, ya sea a través de su mismo nombre, un alias o artículos que hagan referencia a esta, es posible incluir los alias de cada instancia almacenada en la KB para ampliar las posibilidades de reconocimiento de NE.

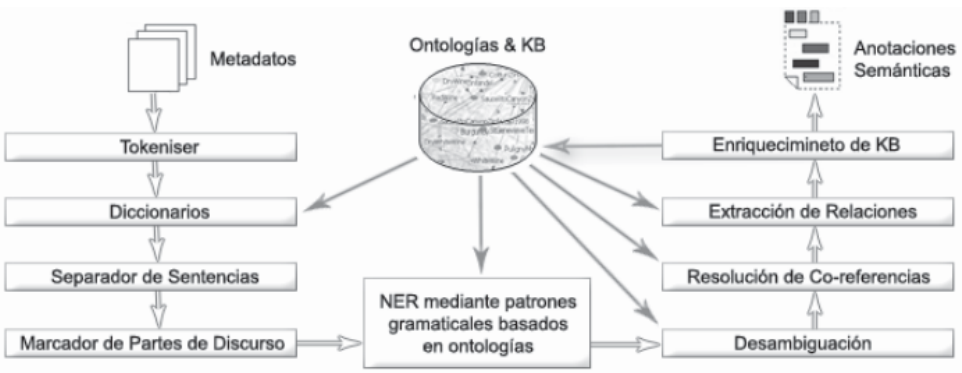

Fig. 3. Extracción de información semántica (basado en [15]).

La desambiguación es otro de los procesos que puede ser apoyado con el uso de ontologías. Al reconocer una NE en el texto, esta puede ser anotada mediante varios tipos de entidades. Sin embargo, si utilizamos una ontología para IE, aún cuando una NE sea anotada mediante varios conceptos de diferentes ontologías donde sean iguales sus nombres, pero diferentes sus significados, a través de las descripciones semánticas de las instancias y el contexto en el metadato que se procesa, pudiera determinarse cuál es el tipo al que realmente pertenece.

Gestión personalizada de contenidos en bibliotecas digitales

Durante varias décadas se han desarrollado numerosas investigaciones en el campo de la Recuperación de Información (IR por sus siglas en inglés) orientada a usuarios con intereses específicos, la mayoría basadas en técnicas clásicas como la búsqueda de palabras claves en el texto. La introducción de las ontologías, como medio de representación formal del conocimiento, ha creado nuevas posibilidades en la descripción de los intereses semánticos de los usuarios. Una representación basada en ontologías es más rica, más precisa y menos ambigua que un modelo basado en palabras claves, además de brindar la posibilidad de recuperar recursos en cualquier formato: texto plano, imágenes, multimedia, etcétera. Estas proveen las bases adecuadas para la representación de los intereses de los usuarios de forma detallada en un modelo jerárquico.

En un enfoque basado en ontologías, las preferencias semánticas de usuarios pueden ser representadas como vectores de peso (con un rango de valores entre 0 y 1) que indican la intensidad con la cual un usuario se interesa por cada concepto. Valores cercanos a cero indican poca atracción por el concepto y valores cercanos a 1 indican gran interés por el concepto [7]. Como fue descrito en el epígrafe 2.1, los OD almacenados en una biblioteca digital son anotados semánticamente contra clases o instancias de ontologías de dominio, y son asignados pesos (con un rango de valores entre 0 y 1) a cada anotación que indica la relevancia del contenido del OD para el concepto de la ontología con el cual se anota. Las ontologías conforman una capa intermedia (ver figura 4), siendo compartidas tanto en la personalización de usuarios como en la anotación semántica de OD. Si un usuario está interesado en un concepto determinado, es posible que el contenido de los OD que hayan

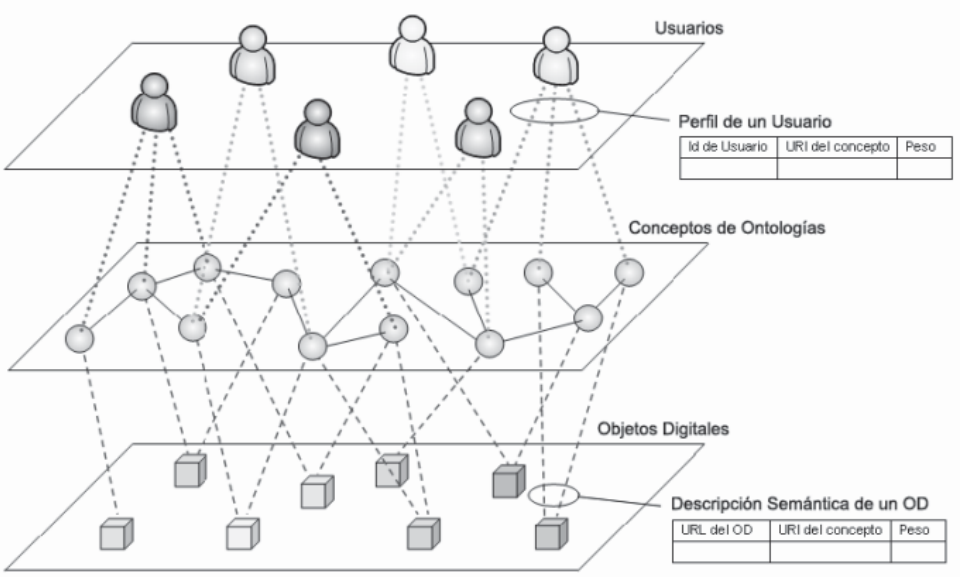

Fig. 4. Asociación de usuarios y objetos digitales (basado en [10]).

sido anotados con dicho concepto sea de su interés. Además, teniendo en cuenta la intensidad de los vectores de pesos tanto del usuario como de los OD con relación a conceptos, pueden gestionarse los OD que mejor se ajusten a su perfil. Las ontologías, además, habilitan nuevos mecanismos como la inferencia que pueden utilizarse para mejorar sustancialmente la personalización. Por ejemplo, si un usuario está interesado en impresoras, pero también está interesado en monitores, pudiera considerarse el interés de dicho usuario por cualquier tipo de periféricos, a partir de que tanto las impresoras como los monitores son subclases de periféricos. De igual forma un usuario interesado en periféricos de cualquier tipo, pudiera estar interesado por contenidos de impresoras y monitores.

Para la gestión personalizada de OD se hace uso del modelo propuesto [10]. En este modelo se define un algoritmo para determinar una medida de relevancia personal (**PRM** por sus siglas en inglés) de un objeto ***d*** para un usuario particular ***u***, acorde a sus preferencias semánticas. Esta medida es calculada como una función entre las preferencias semánticas de ***u*** y las anotaciones semánticas de ***d***. En este cálculo las preferencias de un usuario y las anotaciones de un OD, son vistas como dos vectores en un espacio vectorial K-dimensional, siendo ***K*** el número de elementos en el universo ***O*** de términos de ontologías, y las coordenadas de los vectores son los pesos asignados a términos de ontologías en las preferencias de usuarios y las anotaciones de OD. El **PRM** es representado como la similitud algebraica entre las preferencias de un usuario y los vectores de los OD. Usando el modelo vectorial para la IR clásica [21], esta similitud puede ser medida mediante el coseno de la función, donde el vector de la preferencia juega un rol equivalente al vector de la consulta en la clásica IR. La figura 5 representa la similitud entre dos objetos diferentes $\boldsymbol{d_1}$ y $\boldsymbol{d_2}$ y las preferencias semánticas del usuario ***u***.

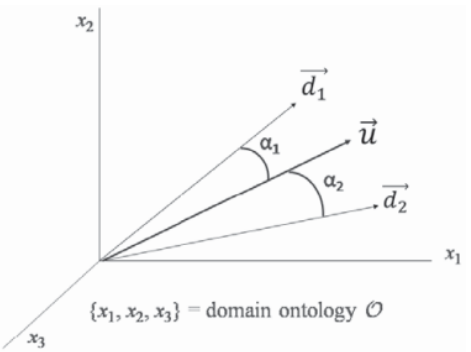

Fig. 5. Representación de la similitud entre los vectores de preferencia y anotaciones [10].

El algoritmo para el cálculo del **PRM** casa dos vectores de pesos asignados a un concepto y produce un valor entre 0 y 1. Los valores cercanos a 0 indican que las preferencias del usuario no coinciden con el OD, y valores cercanos a 1indican que los interese del usuario coinciden con el OD.

Creación y actualización automática de preferencias semánticas

Un importante aspecto en la gestión personalizada de OD es la creación y actualización automática de las preferencias

semánticas de los usuarios. Como fue descrito en el epígrafe anterior, las preferencias semánticas de un usuario son representadas mediante vectores de pesos que indican la intensidad del interés de un usuario por diversos conceptos de ontologías. Por esta razón es necesario determinar, de forma automática, cuáles conceptos son del interés del usuario, en qué medida y en qué momento, pues las preferencias varían con el transcurso del tiempo. Los conceptos que son del interés de un usuario en un período de tiempo pueden determinarse analizando los OD consumidos por este en dicho período, a través de las anotaciones semánticas podemos obtener los conceptos por los cuales han sido anotados los OD consumidos. Almacenando estos en una pila contamos entonces con un historial que nos permite crear perfiles para los usuarios, así como actualizarlos con el transcurso del tiempo.

La estrategia a seguir para la creación y actualización de los perfiles es la descrita [10]. Este modelo propone la inserción de un concepto en el perfil de un usuario teniendo en cuenta la consistencia (si un contenido envuelto en un OD consumido por un usuario está vinculado o no con conceptos relacionados semánticamente) y persistencia (cuán estables y recurrentes son los conceptos de los OD) de estos en el historial del usuario. El método a seguir es ligeramente diferente para los casos de que existan o no, los conceptos en las preferencias de un usuario. En el modelo se presentan varias situaciones a tener en cuenta:

- Un concepto ocurre una vez y su ocurrencia es confirmada en el tiempo aproximadamente con el mismo nivel, este concepto puede ser introducido después de un período como una preferencia a largo plazo del usuario.

- Un concepto ocurre una vez y su ocurrencia es muy alta en un período de tiempo corto, pero desaparece rápidamente. El concepto puede ser considerado una preferencia durante un período de tiempo, pero debe ser removido rápidamente una vez que el interés del usuario desaparece.

- Un concepto ocurre una vez, pero la ocurrencia no es muy alta ni confirmada en el tiempo, este no constituye un interés del usuario por lo que nunca se convertirá en una preferencia.

- Un concepto ocurre y se convierte en una preferencia como en el primer caso, pero desaparece con el tiempo. Para este caso debe ser removido en un determinado período de tiempo.

La decisión de insertar una nueva preferencia está basada en la comparación entre un valor de ocurrencia C_{occ}

$$C_{occ} = N_{occ} / (D - d)$$

Fórmula 2. Cálculo del valor de ocurrencia de un concepto.

y un umbral $P_{thd,}$ para que un concepto candidato del historial se convierta en una nueva preferencia, donde N_{occ} es la cantidad de veces que ocurre un concepto en el conjunto de metadatos de un contenido consumido, D es la fecha de lanzamiento del proceso y d es la fecha de la primera aparición del concepto dentro del conjunto de metadatos de un contenido consumido. Los conceptos candidatos para los cuales $C_{occ} > P_{thd}$ son introducidos como nuevas preferencias y el peso dentro del perfil es inicializado con un valor por defecto neutro. El valor del umbral P_{thd} debe ser determinado mediante experimentos con datos reales.

La eliminación de conceptos de un perfil puede obedecer a dos criterios: a la definición de un tamaño límite para el perfil de un usuario, donde una vez incorporado un nuevo concepto, si es rebasado el tamaño, se elimine el concepto de menor C_{occ}, o al establecimiento de un umbral R_{thd} donde sean eliminados los conceptos que queden por debajo de este umbral. El valor de R_{thd} también debe ser determinado mediante experimentos con datos reales.

Las preferencias semánticas de un usuario no solo consisten en adicionar o remover preferencias de sus perfiles. Es importante poder determinar el peso que indica la intensidad de la relación entre el usuario y los conceptos que conforman su perfil en determinado momento, pues esta intensidad es variable en el tiempo. La actualización de los pesos de los conceptos se hará mediante el modelo matemático propuesto [17] [10], donde el peso es calculado como:

$$W_{new} = W_{old} + fd * ContentRating * e^{-\beta * x * y} * \log \frac{time}{\log length}$$

Fórmula 3. Actualización del peso de los conceptos en los perfiles.

El factor W_{old} representa el peso actual del concepto. Fd es el factor de feedback dado a través del análisis del contenido consumido y puede tomar un valor buleano o ser multi-valor. $ContentRating$ es la clasificación asignada al contenido por el sistema de recuperación personalizada y puede determinarse usando una medida de similitud que puede ser calculada por el coseno entre el contenido y el perfil de usuario. La expresión $log\ (time / log\ length)$ incorpora el tiempo de demora en leer u observar el contenido y actúa como una factor de normalización. El factor $e^{-\hat{a}*x*y}$ es usado para atender los cambios no lineales de los pesos de los conceptos de acuerdo al historial de consumo del usuario, donde x representa el término medio de consumo de contenidos del usuario por día, y representa el número de contenido consumido para los cuales aparece el concepto en sus metadatos, y el factor $\hat{a}$ es una constante que toma valores diferentes teniendo en cuenta si el contenido ha sido o no consumido. Para contenidos no consumidos el índice de decrecimiento debe ser lento, dado que el hecho de que un contenido no haya sido consumido no constituye evidencia de que no sea del interés de un usuario, y para contenidos consumidos el índice de crecimiento debe ser rápido, pues el hecho de que un contenido haya sido consumido si constituye evidencia de que sea del interés de un usuario.

Explotación de relaciones entre conceptos en el aprendizaje de preferencias

La personalización de usuarios puede mejorarse significativamente haciendo uso de los beneficios que ofrecen las ontologías, como son las relaciones que envuelven a los conceptos. Se proponen dos mecanismos de gran utilidad donde son explotados los vínculos semánticos entre los conceptos [10] en:

- *El perfeccionamiento en la adquisición de intereses*

- *La actualización de preferencias teniendo en cuenta la expansión*

Perfeccionamiento en la adquisición de intereses

La gestión del historial de conceptos puede mejorarse significativamente si tenemos en cuenta las relaciones establecidas entre estos. Como ha sido descrito anteriormente, en el historial solo aparecen aquellos conceptos con los cuales están anotados los contenidos consumidos por los usuarios. Sin embargo, el historial puede ser complementado con intereses adicionales deducidos por medio de las relaciones semánticas expresadas en las ontologías, pudiéndose considerar tanto las relaciones de herencia como las relaciones semánticas entre conceptos, también llamadas propiedades. De esta forma la adquisición de preferencias sería mucho más rápida y abundante.

La utilización de las relaciones de herencia puede ser explicada mediante el siguiente ejemplo: si un usuario está interesado en impresoras, mediante la expansión de este concepto pudiera asumirse que también está interesado en cualquier tipo de periférico, pero tal asunción debe ser comprobada mediante la ocurrencia de otros subtipos de periféricos, de esta forma, si un nuevo concepto C_{new} aparece en el contenido consumido, entonces su superclase también es introducida como un interés potencial en el historial, con un valor de seudo-ocurrencia proporcional a la ocurrencia de ***C***:

$$N_{occ}(C_{supertype}) = \gamma_1 * N_{occ}(C_{subtype})$$

Fórmula 4. Proporción entre el valor de ocurrencia de un concepto y su superclase.

donde $\tilde{a}_1 < 1$ y debe ser determinada empíricamente. De esta forma la superclase es añadida a los conceptos del historial y puede convertirse en una preferencia del usuario cuando haya sido confirmada por otras subclases, a tal punto que la seudo-ocurrencia sobrepase el umbral P_{thd}.

Al igual que la herencia, otros tipos de relaciones semánticas pueden brindar un valor significativo. Si un concepto ***c*** aparece en el contenido consumido por un usuario, entonces pudiera adicionarse al historial todos los conceptos $C_{related}$ que se relacionen con él. De esta forma, todos los conceptos relacionados directamente con intereses de un usuario (conceptos), pueden llegar a convertirse en preferencia del usuario. El valor de la seudo-ocurrencia de conceptos relacionados puede determinarse como:

$$N_{occ}(C_{related}) = \gamma_2 * N_{occ}(C)$$

Fórmula 5. Proporción entre el valor de ocurrencia de un concepto y los conceptos relacionados a este.

donde $\tilde{a}_2 < 1$ y debe ser determinada empíricamente.

Actualización de preferencias teniendo en cuenta la expansión

En el mecanismo descrito anteriormente los conceptos son adquiridos aisladamente y la actualización de los pesos de los conceptos solo se basa en pesos previos, sin tener en cuenta la influencia de las relaciones con otros conceptos. Por ejemplo, si queremos actualizar el peso de un concepto ***C*** en las preferencias de un usuario y conocemos que este concepto está relacionado semánticamente con al menos otro concepto, el nuevo peso de ***C*** puede ser calculado como fue descrito anteriormente, pero el nuevo peso de cada concepto $C_{related}$ relacionado a ***C*** puede ser calculado mediante la fórmula:

$$W_{new}(C_{related}) = W_{old}(C_{related}) + sf_{c,c_{related}} * W_{new}(C)$$

Fórmula 6. Actualización del peso de los conceptos relacionados en los perfiles.

donde $W_{new}(C_{related})$ es el nuevo peso del concepto, visto como un concepto relacionado al concepto ***C***, $W_{old}(C_{related})$ es el valor antiguo de peso del concepto, $sf_{C,Crelated}$ es un factor semántico que depende del tipo de relación semántica existente entre $C_{related}$ y ***C***, y $W_{new}(c)$, es el nuevo valor de peso del concepto actual. Lo que acaba de plantearse describe el efecto semántico que el concepto ***C*** tiene sobre el concepto $C_{related}$.

El factor semántico puede decrecer con el nivel de la proximidad semántica entre $C_{related}$ y ***C***, para lo cual hay que considerar los siguientes niveles:

Nivel 1: $C_{related}$ es parte de la definición de ***C***. Las relaciones consideradas son de cualquier tipo (transitiva, inversa, etcétera).

Nivel 2: $C_{related}$ está relacionado con ***C*** por medio de una combinación de la misma propiedad transitiva y $C_{related}$)» ***C*** '« Clase, lo cual significa que estos tienen una superclase en común.

Nivel n: $C_{related}$ está relacionado con ***C*** por medio de una combinación n de la misma propiedad transitiva y $C_{related}$)» ***C*** '« Clase, lo cual significa que estos tienen una superclase en común.

Lenguaje para la representación del conocimiento

Después de un estudio acerca de los diferentes lenguajes y formatos para la representación del conocimiento se decidió utilizar RDFS (del inglés *Resource Description Framework Schema*) [22]. En la actualidad existen una gran cantidad de repositorios, APIs y herramientas de gran madures para el desarrollo de sistemas utilizando RDFS como lenguaje para la representación de los recursos ontológicos y de conocimiento. A pesar de que el estándar OWL (del inglés *Ontology Web Language*) [23] ofrece más expresividad y la capacidad de utilizar en una nueva versión el trabajo de versiones anteriores, este carece de herramientas que lo soporten suficientemente. RDFS provee suficiente expresividad para el propósito que será empleado en este trabajo, donde será utilizado para la definición de ontologías de peso ligero y la descripción de entidades, ya que OWL Lite, el primer nivel de OWL, cubre definiciones útiles (relaciones transitivas y simétricas, igualdad, etcétera), se sugiere evitar el uso de definiciones de RDFS no compatibles con OWL, con el objetivo de facilitar la migración hacia este lenguaje de formalización.

Repositorio de ontologías y KB

Una KB no es más que un cuerpo de conocimiento formal sobre entidades. El término KB refleja mejor la representación de conocimiento formal no ontológico [12]. Mientras en las ontologías se definen todas las clases, sus relaciones, atributos, restricciones y dependencias, en una base de conocimientos se almacenan las descripciones de las instancias y las relaciones entre estas. Las ontologías son consideradas como un tipo de esquema para la KB, manteniéndose tanto las ontologías de dominio utilizadas como la base de conocimientos dentro del mismo repositorio.

La KB puede prepoblarse importando conocimiento desde fuentes de confianza e ir enriqueciéndose con la información descubierta mediante la IE en el proceso de anotación semántica. La IE permite tanto la incorporación de nuevas entidades como el establecimiento de relaciones entre estas. La población de la KB con instancias de importancia general, constituye un gran soporte para el proceso de anotación semántica. También es importante contar con todos los posibles alias para cada instancia, ya que constituye una gran ayuda en el reconocimiento de NE por parte de procesos de IE en la anotación semántica automática.

Conclusiones

Con la realización de esta investigación se puede concluir que la incorporación de los principios y tecnologías de la *Web Semántica* en la generación del modelo de usuario y en la descripción semántica de contenidos en una biblioteca digital, permite el desarrollo de mecanismos de gestión orientados a satisfacer las necesidades específicas de los usuarios, y por consecuencia a elevar su nivel de satisfacción. Brindar una atención que responda a los intereses específicos de comunidades de usuarios con diversas preferencias, constituye una de las más altas prioridades de los sistemas actuales para la

gestión del conocimiento, y de lograr este propósito dependerá en gran medida el éxito de tales sistemas.

El principal aporte de este trabajo es el diseño de una arquitectura para el manejo del modelo de usuario en una biblioteca digital, a partir de la integración de múltiples principios y tecnologías surgidas con el desarrollo de la *Web Semántica*, de técnicas de probada eficacia para el modelado de usuarios, para la Extracción de Información, para la anotación semántica de contenidos, y de mecanismos de gestión basados en la semántica de los contenidos y en las preferencias de los usuarios.

La arquitectura propuesta marca las bases para:

- La generación del modelo de usuario que interactúa con una biblioteca digital.

- La descripción formal de la semántica de los contenidos en una biblioteca digital.

- Facilitar el desarrollo de diversos tipos de mecanismos para la gestión de contenidos, ya sean combinando la semántica de los contenidos con las preferencias de los usuarios, o solo sobre las descripciones semánticas de los contenidos, pues habilita una nueva capa de datos semánticos que puede ser explotada con múltiples propósitos.

- Proveer interoperabilidad y compatibilidad con diferentes módulos dentro de una biblioteca digital y con sistemas externos, mediante el empleo de ontologías como medio para la representación formal del conocimiento.

Referencias bibliográficas

1) Berners-Lee, Tim. Weaving the Web. Orion Business Books. 1999.

2) Berners-Lee, Tim; Hendler, James and Ora Lassila The Semantic Web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 2001. 5, 284. pp. 34-43. Mayo, 2001.

3) Castells, Pablo. «Aplicación de técnicas de la Web Semántica» [en línea]. Universidad Autónoma de Madrid. 2002. <http://giig.ugr.es/~mgea/coline02/Articulos/pcastells.pdf> [Consultado: 3 de febrero de 2009].

4) Vallet, D.; Fernández, M.; Castells, Pablo. An Ontology-Based Information Retrieval Model. 2nd European Semantic Web Conference (ESWC 2005). Heraklion, Greece, Mayo 2005. Gómez-Pérez, A.; Euzenat, J. (Eds.), Springer Verlag Lecture Notes in Computer Science, Vol. 3532, ISBN: 3-540-26124-9, 2005, pp. 455-470.

5) Castells, Pablo.; Fernández, Miriam; Vallet, David. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Transactions on Knowledge and Data Engineering 19(2), Special Issue on Knowledge and Data Engineering in the Semantic Web Era, Febrero 2007, pp. 261-272.

6) Fernández, M.; López, V.; Sabou, M.; Uren, V.; Vallet, D.; Motta, E.; Castells. Semantic Search meets the Web. 2nd IEEE International Conference on Semantic Computing (ICSC 2008). Santa Clara, CA, USA, Agosto 2008.

7) Castells, Pablo.; Fernández, M.; Vallet, D.; Mylonas, P.; Avrithis, Y. Self-Tuning Personalized Information Retrieval in an Ontology-Based Framework. 1st IFIP WG 2.12 & WG 12.4 International Workshop on Web Semantics (SWWS 2005), Noviembre 2005. Z. Tari, P. Meersman; Herrero, P. (Eds.), Springer Verlag Lecture Notes in Computer Science, Vol. 3762, ISBN: 3-540-29739-1, 2005, pp. 977-986.

8) Vallet, D.; Castells Pablo.; Fernández, M.; Mylonas, P.; Avrithis, Y. Personalized Content Retrieval in Context Using Ontological Knowledge. IEEE Transactions on Circuits and Systems for Video Technology 17(3), special issue on the convergence of knowledge engineering, semantics and signal processing in audiovisual information retrieval, Marzo 2007, pp. 336-346.

9) Cantador, Iván; Castells, Pablo.; Bellogín, Alejandro. Modelling Ontology-based Multilayered Communities of Interest for Hybrid Recommendations. 1st Workshop on Adaptation and Personalization in Social Systems: Groups, Teams, Communities (SociUM 2007) at the 11th International Conference on User Modeling (UM 2007). Corfu, Greece, Junio 2007.

10) Cantador, I.; Fernández, M.; Vallet, D.; Castells Pablo.; Picault J.; Ribière M. A Multi-Purpose Ontology-Based Approach for Personalised Content Filtering and Retrieval. Wallace, M.; Angelides, M.; Mylonas, Ph. (Eds.), Advances in Semantic Media Adaptation and Personalization. Springer Verlag Studies in Computational Intelligence, Vol. 93, ISBN 978-3-540-76359-8, Febrero 2008, pp. 25-52.

11) Kiryakov, Atanas; Popov, Borislav; Ognyanoff, Damián; Manov, Dimitar; Kirilov, Angel; Goranov, Miroslav. Smantic Annotation, Indexing, and Retrieval. 2nd International Semantic Web Conference (ISWC2003), 20-23 Octubre 2003, Florida, USA. LNAI Vol. 2870, pp. 484-499, Springer-Verlag Berlin Heidelberg 2003.

12) Kiryakov, Atanas; Popov, Borislav; Terziev, Ivan; Manov, Dimitar; Ognyanoff, Damyan. Semantic Annotation, Indexing, and Retrieval. Extended and updated version of [11]. Elsevier's Journal of Web Semantics, Vol. 2, Issue (1), 2005.

13) Popov, Borislav; Kiryakov, Atanas; Kitchukov, Ilian; Angelov, Krasimir; Kozhuharov, Danail. Co-occurrence and ranking of entities based on semantic annotation. In Int. J. Metadata, Semantics and Ontologies, Vol.1, 2008 (to appear).

14) Ruiz-Casado, M.; Alfonseca, E.; Castells, Pablo. Automatising the Learning of Lexical Patterns: an Application to the Enrichment of WordNet by Extracting Semantic Relationships from Wikipedia. Data and Knowledge Engineering 61(3), Junio 2007, pp. 484-499.

15) Popov, Borislav; Kiryakov, Atanas; Kirilov, Angel; Manov, Dimitar; Ognyanoff, Damián; Goranov, Miroslav. KIM – Semantic Annotation Platform. 2nd International Semantic Web Conference (ISWC2003), 20-23 Octubre 2003, Florida, USA.

LNAI Vol. 2870, pp. 834-849, Springer-Verlag Berlin Heidelberg 2003.

16) Dowman, Mike; Tablan, Valentin; Cunningham, Hamish; Popov, Borislav. Web-Assisted Annotation, Semantic Indexing and Search of Television and Radio News. In Proc. of the 14th International World Wide Web Conference. Chiba, Japan, 2005.

17) Papadogiorgaki, M., Papastathis, V., Nidelkou, E., Waddington, S., Bratu, B., Ribière, M. and Kompatsiaris, Y. (2007). Distributed User Profile Management and Adaptation for Personalised News Content Delivery, submitted to the special issue «Data Mining for Personalisation» of User Modelling and User Adapted Interaction (UMUAI) journal.

18) Vallet, D.; Mylonas, P.; Corella, M. A.; Fuentes, J. M.; Castells, Pablo.; Avrithis, Y. A Semantically-Enhanced Personalization Framework for Knowledge-Driven Media Services. IADIS WWW/Internet Conference (ICWI 2005). Lisbon, Portugal, Octubre 2005.

19) Davies, John; Studer, Rudi; Warren, Paul. Semantic Web Technologies Trends and Research in Ontology-based Systems. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, England, 2006.

20) Cunningham, Hamish; et al. «Developing Language Processing Components with GATE» [en línea]. The University of Sheffield. Version 3. Mayo 2006. <http:// gate. ac.uk/sale/tao/> [Consultado: 10 de enero de 2009].

21) Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, 1999.

22) W3C «Resource Description Framework» [en línea]. 2004. <http://www.w3.org/RDF/> [Consultado: 8 de diciembre de 2008].

23) W3C «OWL Web Ontology Language» [en línea]. Febrero 2004. <http:// www.w3.org/TR/owl-features/ > [Consultado: 5 de enero de 2009].

Recibido: 16 de agosto de 2009.
Aprobado en su forma definitiva:
3 de noviembre de 2009

Ing. Reydi Castillo Buergo
Universidad Agraria de La Habana
Correo electrónico:
<reydi@isch.edu.cu>