III Conferencia Internacional de Biblioteca Digital y Educación a Distancia, Mérida, 2009.

Recuperación y clasificación automática de información, resultados actuales y perspectivas futuras.

De Giusti, Marisa, Villarreal, Gonzalo Luján, Sobrado, Ariel y Vosou, Agustín.

Marisa Raquel De Giusti[1], Gonzalo Luján Villarreal[2], Ariel Sobrado[3], Agustín Vosou[4]

Recuperación y clasificación automática de información, resultados actuales y perspectivas futuras

Abstract - **En este trabajo se presenta una herramienta de recolección de información abierta que, mediante la combinación de ontologías y tesauros, brindará información clasificada y unificada en un repositorio temático a los usuarios del Servicio de Difusión de la Creación Intelectual (SeDiCI) [1]; esta clasificación permitirá optimizar considerablemente las búsquedas dentro del portal.**

SeDiCI posee actualmente una gran cantidad de documentos con una sintaxis y catalogación correctas, pero carece de relaciones semánticas entre los mismos. Esta falta de relaciones semánticas genera mayores esfuerzos por parte de los usuarios para vincular documentos unos con otros, a fin de filtrar y clasificar los resultados de una consulta a partir de un dominio específico.

Con el objeto de ayudar a los usuarios de SeDiCI a encontrar información pertinente, se propone aquí la incorporación de una herramienta capaz de establecer relaciones semánticas entre los documentos. Esta herramienta constará de dos módulos: el primero estará encargado de recolectar información abierta de interés mediante un agente que navega recursivamente a través de las URLs de los documentos localizados; el segundo módulo será capaz de identificar las páginas marcadas junto a sus etiquetas, y proveer un conjunto de reglas para extraer la información y guardarla en un fichero RDF. A continuación se realizará un proceso de homogeneización entre los términos encontrados, clasificando la información en función de una ontología de dominio. El material recolectado poblará de este modo la ontología, sumándose al repositorio semántico. Para las primeras pruebas de esta herramienta, se utilizará el repositorio propio de SeDiCI, junto con una operación de marcado automática.

Una vez que los documentos hayan sido vinculados semánticamente, se proveerá un buscador capaz de aprovechar estas nuevas relaciones – compuestas por clases y subclases – dentro de la ontología lo cual resultará en una considerable mejora en el proceso de organización y entrega de información pertinente al usuario.

I. Introducción

El Servicio de Difusión de la Creación Intelectual (SeDiCI) [1] fue creado, inicialmente, para exponer la creación de las distintas unidades académicas de la UNLP [2] como vía de socialización de conocimientos. SeDiCI oferta sus contenidos siguiendo el protocolo de la Iniciativa Open Archives (OAI) y a la vez recolecta información académica externa libre bajo este protocolo para ponerla a disposición de la comunidad científica de toda la UNLP. Un objetivo constante del servicio es brindar a los usuarios información cada vez mayor y más pertinente. Para lograr este objetivo se ha pensado en mejorar los mecanismos de obtención de información libre desde la web perteneciente a las distintas áreas del conocimiento, verificando las fuentes y estructurando adecuadamente dicha información dentro de la biblioteca digital para ofrecer a los usuarios mecanismos de búsqueda más precisas.

Los buscadores de propósito general, como Google y Yahoo!, son actualmente el principal medio utilizado para acceder a la enorme cantidad de información volcada en Internet. Si bien estos buscadores ofrecen millones de resultados en pocos segundos, los datos obtenidos a partir las búsquedas de los usuarios son seleccionados por medio de *palabras clave* en lugar de *conceptos*: de este modo, se pierden relaciones de importancia y con ello la información obtenida es diferente incluso en el caso de que las palabras utilizadas para la búsqueda sean sinónimos [3].

Una de las alternativas actuales para organizar y localizar información en la web es la denominada web semántica [4] [5], una iniciativa del W3C liderada por Tim Berners Lee [6]. En este trabajo, dada la naturaleza de la biblioteca digital, la propuesta no está dirigida a la generación y compartición de ontologías (más allá de que las

1 M.R. De Giusti; Investigador adjunto sin director de la Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CICPBA) y Director del Proyecto de Enlace de Bibliotecas de la UNLP; email: marisa.degiusti@sedici.unlp.edu.ar

2 G.L. Villarreal; Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) y Proyecto de Enlace de Bibliotecas de la UNLP; email: gonetil@sedici.unlp.edu.ar

3 A. Sobrado; Proyecto de Enlace de Bibliotecas de la UNLP; asobrado@sedici.unlp.edu.ar

4 A. Vosou; Proyecto de Enlace de Bibliotecas de la UNLP; agustinvosou@sedici.unlp.edu.ar

consecuencias de este trabajo lleven a este logro), sino que se pretende hacer un uso combinado de ontologías [7] y de los tesauros [8] que actualmente utiliza el servicio a fin de incrementar la eficiencia de los procesos de obtención automática de información de fuentes heterogéneas (la web, determinados repositorios, etc.) por parte de SeDiCI y la búsqueda por parte de los usuarios.

Cuando se habla de fuentes heterogéneas, como se advierte en el párrafo anterior, y aunque se esté buscando información en el mismo dominio de interés, las distintas fuentes utilizan distintas convenciones para representar los mismos conceptos. Con este fin es que se propone la creación de una plataforma capaz de extraer datos de páginas marcadas semánticamente, pertenecientes a distintos portales. La información obtenida se almacena, luego de un proceso de normalización, en un repositorio ontológico sobre el cual los usuarios pueden buscar ya no con palabras claves sino semánticamente.

II. Descripción de la plataforma

La plataforma que aquí se presenta, cuyos componentes se detallan a lo largo de esta sección, se divide en dos grandes sistemas: sistema de recolección y sistema de búsqueda (Fig. 1 y 2). El sistema de recolección tiene como objetivo analizar las paginas web almacenadas en una base de datos y detectar aquellas que están marcadas, extraer su contenido y almacenarlo en una ontología previamente definida para una temática dada, seleccionando términos de un tesauro. El sistema de búsqueda es el encargado de guiar al usuario a realizar una búsqueda inteligente sobre un repositorio que ha sido poblado por el sistema de recolección.

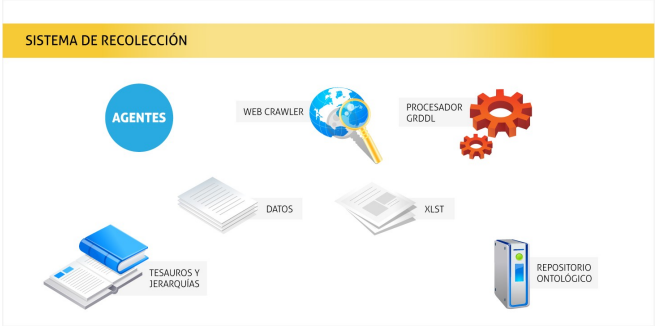

Fig. 1: Sistema de Recolección

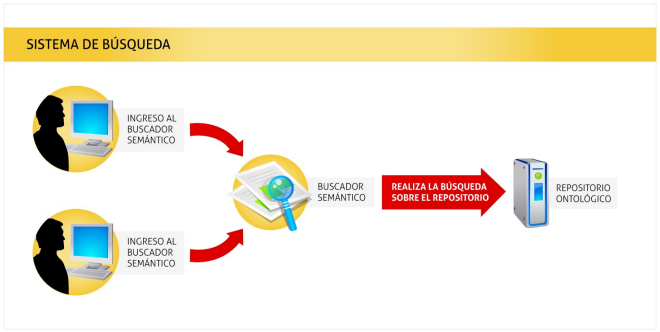

Fig. 2: Sistema de Búsqueda

Sistema de Recolección

Componentes del Sistema:

Entre los componentes más importantes se destacan la base de datos que contiene las URLs a visitar, un agente que selecciona una a una las URLs que aún no han sido visitadas, un web crawler capaz de obtener URLs embebidas dentro del código HTML de otra URL, otro agente que procesa las URLs y puebla la ontología, un procesador GRDDL que aplica transformaciones a los documentos (X)HTML, un tesauro con los términos homogeneizados y una ontología para representar los datos obtenidos.

Todos los componentes mencionados se han separado en dos módulos bien definidos: búsqueda y procesamiento. Esta separación permite ejecutar en paralelo ambos módulos asegurando que las acciones de un módulo no interrumpan al otro, aportando tanto robustez como eficiencia ambas partes.

El módulo de búsqueda requiere de una serie de pasos para completar su ejecución de manera satisfactoria, que corresponden a:

1. El Robot 1 toma las URLs no visitadas de la Base de Datos.
2. El Robot 1 envía al web crawler la URL obtenida. También la marca como visitada.
3. El web crawler verifica dicha URL, busca las URLs embebidas dentro del código HTML (identificadas dentro de la etiqueta <a>) y las agrega a la Base de Datos.

Del mismo modo, el módulo de procesamiento se ejecuta mediante una serie de etapas secuenciales, a saber:

1. El Robot 2 toma de la Base de Datos una URL que ya ha sido visitada pero que aun no fue procesada. Esa URL le es pasada al procesador GRDDL.
2. El procesador GRDDL aplica ciertas transformaciones XSLT (mediante hojas de transformaciones) para obtener un documento XML o RDF.
3. El Robot 2 marca la URL como procesada.
4. En caso de ser necesario transforma el fichero XML a RDF.

5. Se buscan los términos en el tesauro definido y se crean instancias de la ontología para poblar el servidor.

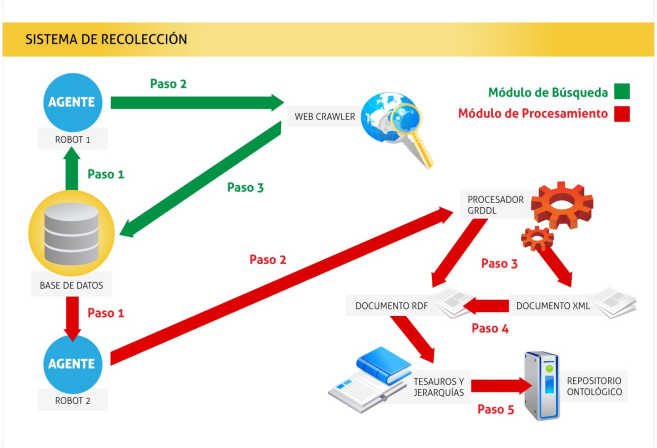

Fig. 3: Pasos del Sistema de Recolección

El sistema de recolección se encarga de las siguientes tareas: a través de un agente [9][10] recorre una lista de páginas WEB que envía a un web crawler [11][12], encargado de recolectar otros enlaces embebidos en la página en cuestión e incorporarlos a la lista previa. Un agente llamado robot 2 toma una URL y la envía a un procesador del tipo GRDDL ("Gleaning Resource Descriptions from Dialects of Languages") [13] que aplica una transformación, en nuestro caso mediante XSLT, de un documento XHTML o XML a XML [14]. Este documento textual sólo contiene ahora las tuplas de interés, lo que permitirá a la aplicación extraer de forma automática información de páginas web estructuradas para integrarla en un repositorio.

El procesador GRDDL detecta microformatos de interés contenidos en las páginas. Con la indicación de la dirección de la localización de la hoja de transformación (XSL) usada para capturar dichos microformatos [15][16] junto a la URL del lugar de extracción, devuelve un XML con la información extraída por la hoja de transformación. El documento XML es a posteriori transformado a un documento RDF (Resource Description Framework) [17] a fin de homogeneizar los datos, clasificarlos y finalmente poblar la ontología. La homogeneización se realiza mediante conversiones (uso de sinónimos) y traducciones para mantener un único idioma. La clasificación de la información se hace mediante una ontología de dominio. Durante esta etapa de clasificación se crean instancias con atributos y relaciones a partir de los documentos RDF. Dichas instancias deben pertenecer a alguna clase definida en la ontología. Además se buscan en el RDF las relaciones de dominio marcadas por la ontología.

En el siguiente paso, se verifica que las instancias satisfagan todas las restricciones impuestas por su clase a través de un módulo de razonamiento. Si la instancia en cuestión pertenece a la clase se le agregan los atributos heredables de la misma.

Una vez que esta etapa ha finalizado, se procede con el almacenamiento de las instancias en un repositorio semántico accesible vía web.

III. Estudio de caso

El número limitado de páginas marcadas en la web actual determinó la elección del caso de muestra. Para probar el sistema de recolección, se realizaron cambios sobre la marcha en la plataforma.

Como primera aproximación, se trabajó sobre los registros de la Biblioteca Digital SeDiCI, que cumple el rol de web heterogénea. Para ello, se adaptó el software de SeDICI al microformato Dublic Core [18][19] para representar los registros [20] del repositorio. A modo de muestra, se realizó una búsqueda por descriptor con el término: “Física del estado sólido”.

IV. Definición de la Ontología

Para la representación de nuestra ontología SediciON se ha utilizado el lenguaje OWL-DL, recomendación del W3C [21]. Las características de OWL-DL aseguran la interoperabilidad con otros sistemas y formatos. Dichas características, tales como su capacidad para la inferencia en sistemas de organización conceptual basados en jerarquías, se utilizarán para proporcionar más funcionalidad al sistema y describir de una forma más rica los recursos involucrados en él.

Nuestra ontología volcada a Protégé [22] posee una clase denominada MATERIAL de la cual es subclase el tipo de documentos que estamos analizando. Para representar los atributos de los materiales, hace reuso de la Dublin-Core Ontology [23].

Las entidades reutilizadas de la Ontología DC aparecen referenciadas muy brevemente arriba, pero se remite al lector a su descripción original en caso de que necesite una aclaración de su significado.

V. Ejemplo de extracción

En nuestro caso se extrajeron los datos de documentos marcados de la biblioteca digital SeDiCI. Durante el proceso de adquisición el módulo de recolección procesa únicamente las páginas que contienen microformatos Dublin Core. El fichero RDF generado contiene los datos de los trabajos (título, autor y descriptores) clasificados junto a su temática. Con los datos RDF el módulo de población de la ontología crea las instancias. En nuestro caso, antes de almacenarlos (en un futuro próximo) podríamos buscar entre los valores del atributo *Description* todos aquellos valores que se correspondan con los términos alternativos de algún tesauro distinto del que se utiliza en SeDiCI.

Tanto la información como los datos sobre los diferentes recursos es almacenada en un repositorio ontológico de SeDiCI: FisSol. Todas las entidades de esta base de datos deben ser mapeadas a instancias RDF, las cuales a su vez se organizan de acuerdo al modelo conceptual de la ontología SediciON.

VI. Conclusiones y Trabajos futuros

Se han presentado los primeros lineamientos de una plataforma de búsqueda semántica en fuentes heterogéneas que combina el uso de ontologías y tesauros**,** y que dará mayor pertinencia a la búsqueda de los usuarios de SeDiCI. Para el grupo de trabajo de PrEBi esta es una primer experiencia que ha servido para conocer el ámbito de trabajo y su aplicabilidad a la biblioteca digital según un objetivo prioritario**:** brindar mejor y más estructurada información. El sistema de recolección deberá analizarse en cuanto a su eficiencia: el uso de dos robots, las herramientas y librerías seleccionadas, especialmente pensando que tras una primer**a** etapa de semantización de la biblioteca, ésta deberá completarse con nuevos contenidos desprendidos de la WEB**.** En este sentido, la búsqueda de páginas marcadas, tal cual la aplicación actual**,** puede resultar una limitación y será necesario pensar en otras técnicas de extracción. Finalmente**,** será necesario pensar en cómo manejar otras relaciones definidas en tesauros y ontologías más complejas.

El sistema de búsqueda deberá realizarse por completo pues la idea de esta plataforma es que los usuarios accedan a los repositorios semánticos creados y puedan realizar búsquedas más pertinentes, para lo cual el módulo de búsquedas deberá permitirles un acceso vía web que brinde una pantalla de búsqueda donde puedan buscar por conceptos, seleccionar atributos y elegir restricciones para que finalmente se les devuelva una lista de resultados que muestren los atributos seleccionados que cumplen con las restricciones establecidas.

VII. Referencias y citas bibliográficas.

1: Servicio de Difusión de la Creación Intelectual, SeDiCI. http://sedici.unlp.edu.ar

2: Universidad Nacional de La Plata. http://www.unlp.edu.ar/

3: Abian, M.A. El futuro de la web. Xml,rdf/rdfs, ontologías y la web semántica. http://www.javahispano.org/contenidos/es/el_futuro_de_la_web/

4: W3C. W3C Semantic Web Activity. http://www.w3.org/2001/sw/grddl-wg/td/grddl-tests#spaces-in-rel/

5: Wikipedia. Web Semántica. http://es.wikipedia.org/wiki/Web_semántica

6: Berners-Lee, T. y Fischetti, M. Weaving the Web: The original Design and Ultimate Destiny of the World Wide Web by its Inventor. San Francisco:Harper.

7: Gruber, T. R. What is an Ontology?. (1992) http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

8: Wikipedia. Tesauro. http://es.wikipedia.org/wiki/Tesauro

9: Khedro, T., Genesereth, M. R. Modeling Multiagent Cooperation as Distributed Constraint Satisfaction Problem Solving, leventh European Conference on Artificial Intelligence, Amsterdam, The Netherlands. (1994)

10: Wikipedia. Agente. http://es.wikipedia.org/wiki/Agente_inteligente_(Inteligencia_Artificial)

11: Wikipedia. Web Crawler. http://en.wikipedia.org/wiki/Web_crawler

12: Sun. Writing a Web Crawler in the Java Programming Language. http://java.sun.com/developer/technicalArticles/ThirdParty/WebCrawler/

13: W3C. Gleaning Resource Descriptions from Dialects of Languages (GRDDL). http://www.w3.org/TR/grddl/

14: W3C. XSL Transformations (XSLT) Version 1.0. http://www.w3.org/TR/xslt

15: Wikipedia. Microformato. http://es.wikipedia.org/wiki/Microformato

16: Microformats. Microformats. http://microformats.org/

17: Wikipedia. Resource Description Framework (RDF). http://es.wikipedia.org/wiki/Resource_Description_Framework

18: Dublin Core Metadata Initiative (DCMI). Expressing

Dublin Core in HTML/XHTML meta and link elements. (2003) http://dublincore.org/documents/dcq-html/

19: Dublin Core Metadata Initiative (DCMI). Dublin Core Metadata Initiative. http://dublincore.org

20: Mendez, E.. DCMF:DC y microformatos, a good marriage". International Conference on Dublin Core and Metadata Applicationes, 22-26 September 2008. http://dc2008.de/wp-content/uploads/2008/09/dc2008_mendezetal.pdf

21: W3C. Web Ontology Language (OWL). http://www.w3.org/2004/OWL/

22: Stanford Center for Biomedical Informatics Research. Welcome to protégé. http://protege.stanford.edu/

23: . Dublin Core Ontology. http://protege.stanford.edu/plugins/owl/dc/protege-dc.owl