Catalogación de Recursos Educativos utilizando Vocabulario Controlado a partir de Ontologías

Ana Casali[1,2], Claudia Deco[1,3], Cristina Bender[1,3], Fabricio Mahon[1]

[1] Departamento de Sistemas e Informática – Facultad de Ingeniería – Universidad Nacional de Rosario (UNR) – Av. Pellegrini 250 – Rosario – Argentina

[2] Centro Internacional Franco Argentino de Ciencias de la Información y de Sistemas (CIFASIS) – Argentina

[3] Departamento de Investigación Institucional – Facultad de Química e Ingeniería Campus Rosario – Universidad Católica Argentina (UCA) – Argentina

{acasali, deco, bender}@fceia.unr.edu.ar, fabriciomahon@gmail.com

Abstract. *This paper presents a document indexer system by using controlled vocabulary represented by ontologies. This cataloguer can assist the upload of educational resources in a repository, providing their keywords. In this first proposal, the domain of interest for the classification is Computer Science and ontologies were developed for three of its areas using Methontology methodology. The system architecture is presented and a prototype is implemented in Java, using a keyword extractor and an algorithm that searches for terms that are keywords and are within the ontology, using as parameters the weights of the term in the document and in the ontology. The functionality of this prototype is illustrated by a case study.*

Resumen. *En este trabajo se presenta un sistema catalogador de documentos mediante el uso de vocabulario controlado representado mediante ontologías. Este sistema puede asistir en la carga de recursos educativos en un repositorio, brindando sus palabras claves. En esta primera propuesta, el dominio de interés para la clasificación es el de las Ciencias de la Computación y se desarrollaron ontologías para tres de sus áreas utilizando la metodología Methontology. Se presenta la arquitectura del catalogador y se implementa un prototipo en lenguaje Java, utilizando un extractor de palabras claves del documento y un algoritmo que busca estos términos dentro de la ontología, utilizando como parámetros el peso de los términos en el documento y en la ontología. Se ilustra el funcionamiento de este prototipo mediante un caso de estudio.*

1. Introducción

El objetivo de este trabajo es diseñar un catalogador automático de recursos educativos para la asignación de palabras claves utilizando un vocabulario controlado. La motivación es ayudar a la carga de los recursos en un repositorio y mejorar la recuperación de los mismos frente a una búsqueda. En [Yedid, 2013] se indaga respecto de las formas de indización utilizadas en los repositorios digitales de acceso abierto en Argentina. Como uno de los resultados de este análisis el 100% indicó que utiliza

vocabulario controlado para la indización pero que ésta es realizada manualmente y ninguno utiliza un software que permita la asignación automática de términos pertenecientes a un vocabulario controlado, de esto se desprende la relevancia del desarrollo aquí planteado.

En estos últimos años se vienen desarrollando sistemas catalogadores automáticos de documentos o sitios web. Cada uno presenta características particulares. [Anjewierden y Kabel, 2001] utilizan ontologías como vocabulario fijo (pero no controlado) para indexar documentos con grandes cantidades de datos. [Barrios y Gutiérrez, 2005] propone un modelo para la catalogación semi-automática de un sitio web a partir de la creación de un conjunto de metadatos sobre los contenidos de un sitio. [Wijewickrema y Gamage, 2012] presentan una aplicación para clasificar automáticamente textos utilizando una ontología de dominio. Esta ontología está basada en el esquema de clasificación Dewey utilizada en bibliotecología.

En este trabajo se propone una arquitectura de sistema catalogador de documentos utilizando lenguaje controlado por una ontología de dominio. Se implementa un prototipo en lenguaje Java, utilizando un extractor de palabras claves y la implementación de una función distancia para comparar los términos que son palabras claves del recurso de aprendizaje con los términos que se encuentran dentro de la ontología, utilizando como parámetros el peso de los términos en el documento y en la ontología. En este trabajo, se trabaja en el dominio de las Ciencias de la Computación. A partir de que no fue posible disponer de una ontología de dicho dominio en idioma español abierta, surge la necesidad de diseñar esta ontología, por lo cual, en este trabajo se propone el diseño y la construcción de una ontología en español de algunas áreas de las Ciencias de la Computación. Para el diseño se utiliza la metodología Methontology. Para que la ontología resulte completa y correcta fue necesario explorar distintas fuentes confiables y científicas. En nuestro trabajo, la base de conocimiento está enriquecida por sitios web tales como ACM, CiteSeerX y Springer y bibliografía relevante dentro del dominio de interés.

El resto del trabajo se estructura de la forma siguiente: en la Sección 2 se presenta conceptos básicos de tesauros y ontologías; en la Sección 3 se muestra el diseño de la ontología que se utilizará; en la Sección 4 se propone una arquitectura para la catalogación y un caso de uso. Finalmente se presentan las conclusiones.

2. Tesauros y Ontologías

La flexibilidad y variedad del lenguaje natural crea serias dificultades para el manejo automatizado de la información. Para solucionar este problema, surgen los tesauros, que permiten el control del vocabulario para representar en forma unívoca cada concepto. Según la definición de la UNESCO, un tesauro es un instrumento de control terminológico utilizado para traducir a un lenguaje más estricto el idioma natural empleado en los documentos y así asignar palabras claves que describan a cada documento. Por su estructura, es un vocabulario controlado y dinámico de términos relacionados semántica y genéricamente, los cuales cubren un dominio específico del conocimiento. Está estructurado formalmente con el objeto de hacer explícitas las relaciones entre conceptos. Estas interrelaciones pueden ser: jerárquicas, de afinidad, y preferenciales. Las relaciones jerárquicas indican términos más amplios o más específicos de cada concepto. Las relaciones de afinidad muestran términos

relacionados conceptualmente, pero que no están ni jerárquica ni preferencialmente relacionados. Las relaciones preferenciales se utilizan para indicar cuál es el término preferido o descriptor entre un grupo de sinónimos. En las bases de datos documentales se utilizan palabras claves para describir el contenido de un documento. Estas palabras claves o descriptores, pueden estar formadas por términos controlados o permitidos que se eligen de un tesauro. Entonces, los términos del tesauro se clasifican en descriptores, o términos preferidos, y en no descriptores o términos equivalentes no preferidos o prohibidos. Los términos no descriptores no pueden ser utilizados como palabras claves ni como términos de búsqueda. Para cada término no descriptor, el tesauro indica cual es el término permitido correspondiente.

Las ontologías proporcionan una vía para representar el conocimiento y son un enfoque importante para capturar semántica. La definición más consolidada es la que la describe como "una especificación explícita y formal sobre una conceptualización compartida" [Gruber, 1993]. Es decir, definen conceptos y relaciones de algún dominio, de forma compartida y consensuada y esta conceptualización debe ser representada de una manera formal, legible y utilizable por las computadoras. Las ontologías consisten de términos organizados en una taxonomía, sus definiciones y axiomas que los relacionan con otros términos. Tienen los siguientes componentes para representar el conocimiento sobre un dominio: Conceptos (ideas básicas que se intentan formalizar), Relaciones (representan la interacción y enlace entre los conceptos; suelen formar la taxonomía del dominio), Funciones (relación que permite identificar un elemento mediante un cálculo que considera otros elementos), Instancias (representan elementos determinados de un concepto) y Axiomas (teoremas que se declaran sobre relaciones y que permiten inferir conocimiento que no está explicitado en la taxonomía de conceptos). Las ontologías que no utilizan axiomas se denominan ontologías livianas.

En la búsqueda en bases de datos documentales, el uso de tesauros permite obtener un resultado más preciso. Esto se debe a que en el caso de sinónimos el tesauro indica cuál es el término preferido que se utiliza como descriptor en los documentos. La utilización de términos preferidos aumenta la precisión en la búsqueda. Por otro lado, la estructura jerárquica de los tesauros permite que un usuario pueda seleccionar un concepto más específico a su interés de búsqueda y de este modo mejorar la precisión de los resultados.

En los repositorios de recursos educativos la terminología no está controlada. Los términos indicados como palabras clave por lo general no están tomados de un vocabulario controlado. Por esto, en este trabajo se propone el uso de una ontología de dominio que es utilizada para incorporar las palabras clave en los documentos. Se utiliza una ontología liviana para representar un tesauro. Se decidió utilizar una ontología por ser reutilizable, por permitir estandarizar y compartir el conocimiento sobre un dominio.

3. Diseño de la ontología

Existen varios enfoques para clasificar las ontologías [Valencia García, 2005]. En este trabajo se utilizan Ontologías de Dominio y en particular, se propone una ontología de Ciencias de la Computación y se conceptualizan algunas subáreas de Inteligencia Artificial (como Aprendizaje automatizado y Agentes) y de Teoría de Bases de Datos.

Actualmente existen muchas metodologías de diseño de ontologías, cada una tiene su particularidad y está orientada a ciertos tipos de contextos y dominios, no sólo

desde la definición de clases, atributos, instancias y relaciones sino en la forma de estudiar sus aplicaciones y el trabajo previo en el dominio dado. En [Valencia García, 2005] podemos encontrar un análisis exhaustivo del estado del arte de la mayoría de las metodologías. A los fines de seleccionar una para este trabajo se han analizado las metodologías Uschold & King [Uschold y King, 1995], TERMINAE [Biébow y Szulman, 1999], Ontology Development 101 [Noy y McGuinness, 2001] y Methontology [Gómez-Pérez, 1997]. Se optó por Methontology dado que toma muchos conceptos de la metodología de Uschold & King (a nuestro entender la más completa de las metodologías) y de las otras metodologías. Además propone un nivel de abstracción más alto que TERMINAE. Methontology permite definir desde un principio el grado de formalidad, el alcance y qué tipo de usuarios tendrán acceso a la ontología. En la especificación debe aclararse las fuentes de conocimiento que se consultan para su construcción y esto es primordial cuando se diseñan ontologías de dominio. Además, esta metodología permite tener una documentación fiable, donde la implementación y el diseño quedan bien distinguidos, algo que en las otras metodologías no queda tan claro. Esta documentación permite que la ontología pueda ser reutilizable.

3.1. Aplicación de Methontology

En una primera etapa se han diseñado tres ontologías con el objetivo de evaluar resultados del prototipo de catalogación. Cada ontología es un aporte a las ontologías en idioma español ya que las existentes en este dominio son escasas y no profundizan tanto en cada subárea de las Ciencias de la Computación. Se consideran las subáreas: Aprendizaje Automatizado, Agentes y Teoría de Bases de Datos. A continuación se resumen los pasos indicados en Methontology.

1) Especificación: En esta etapa se plantean el objetivo y los requerimientos de la ontología. En la Figura 1, se muestra un documento de especificación como el que se propone en [Gómez-Pérez, 1997].

Documento de especificación de requerimientos de la ontología
Dominios: Aprendizaje automático (Machine Learning), Teoría de Bases de Datos, Agentes. Date: 01/04/2015 Conceptualizado por: FM. Implementado por: FM. Propuesta: Ontologías de las disciplinas Aprendizaje automático), Agentes y Teoría de Bases de Datos, perteneciente al dominio de Inteligencia artificial las dos primeras, para utilizarse como base de conocimiento en la clasificación de documentos científicos o texto plano. Nivel de formalidad: Semi-formal. Alcance: - Términos: se investigarán en la adquisición del conocimiento, a priori no son desconocidos. - Conceptos: partes que componen cada disciplinan a nivel conceptual y no como relación que podemos denominar componentes, problemáticas, dificultades, algoritmos, aplicaciones. Fuentes de conocimiento: - Bibliográficas (aprendizaje automático): "Inteligencia Artificial - Un enfoque moderno", de Stuart J. Russell y Peter Norvig; "Inteligencia artificial: modelos, técnicas y áreas de aplicación", de Maria I. A. Galipienso, Miguel A. C. Quevedo, Otto C. Pardo, Francisco Escolano Ruiz, Miguel A. Lozano Ortega. - Bibliográficas (Agentes): "Inteligencia Artificial - Un enfoque moderno", de Stuart J. Russell y Peter Norvig; "Intelligent Agent Architectures and Applications, de Gautam B. Singh. - Bibliográficas (Teoría de Bases de datos): "Sistemas de bases de datos", de Paul Beynon-Davies; "Fundamentos de Bases de datos" de Abraham Silberschatz, Henry Korth y S. Sudarshan. - No bibliográficas (Aprendizaje Automático, Agentes, Teoría de Bases de datos): ACM, CiteSeerX, Springer, Wikipedia.

Figura 1: Documento de especificación

2) Adquisición del conocimiento: En primer lugar se consultó a profesores de las cátedras Inteligencia Artificial y Teoría de Base de Datos de la Licenciatura en Ciencias de la Computación, para determinar conceptos generales y posibles fuentes que puedan aportar a la categorización deseada. Se eligieron libros de estas temáticas y se consultaron fuentes no bibliográficas para identificar posibles clases y jerarquías, así como algunas posibles relaciones como es-sinónimo-de y es-parte-de, entre otras.

3) Conceptualización: Se construyó un glosario de términos agrupándolos en Conceptos y Relaciones y se generó el árbol de clasificación de conceptos mostrado en la Figura 2, donde:

Área: Es el nombre de un área, por ejemplo *Aprendizaje automatizado.*

PropiedadDeArea: Es una propiedad genérica que es trasladable a otros dominios. Por ejemplo, las instancias correspondientes serían *Dificultades* que se presentan en el área, *Algoritmos* conocidos o *Estructuras de datos* utilizadas en general en el área.

SubArea: Es la subdisciplina dentro del dominio que representa el Área. Por ejemplo, dentro de Aprendizaje Automatizado una subárea es *Aprendizaje por refuerzo*. Las subáreas permiten representar términos específicos.

Autor: Representa un Autor reconocido del dominio.

Componente: Parte indivisible de una SubArea. Esto significa que se está en la base de un nivel de abstracción determinado para definir la ontología.
Sinónimo: Contiene el conjunto de sinónimos de un término. Se aplica a Área, SubArea, Autor y Componente. Por ejemplo, en el caso del área Aprendizaje automatizado, éste es el término preferido y los sinónimos (Aprendizaje automático, ML y Machine Learning) son los términos no permitidos.

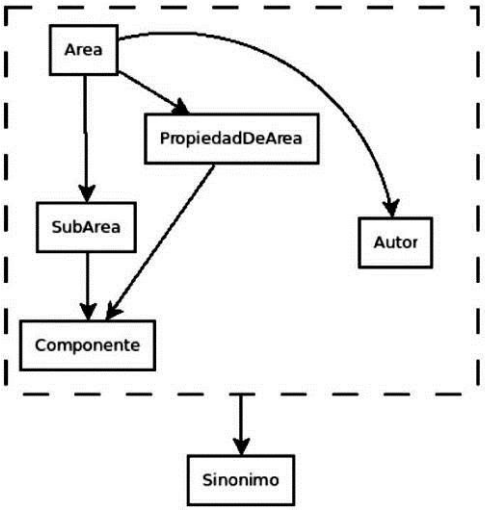

Figura 2: Taxonomía de conceptos

Por ejemplo, para el área Aprendizaje automatizado, las instancias de cada clase pueden ser las indicadas en forma resumida en la Tabla 1.

En la Figura 3, se muestra una parte del árbol de clasificación de atributos para el dominio Aprendizaje Automatizado.

En la Figura 4, se muestra el diagrama de relaciones, donde por ejemplo *contiene_propiedad* expresa qué propiedad general contiene el Área.

Tabla 1: Ejemplo de instancias de cada clase para Aprendizaje automatizado.

Nombre del concepto	Instancias	Relaciones
Área	Aprendizaje automatizado	es_area-de contiene_propiedad
PropiedadDeArea	Dificultades Aplicaciones Algoritmos Componentes	contiene_subarea contiene_componente
SubArea	Aprendizaje supervisado Aprendizaje inductivo Aprendizaje no supervisado Aprendizaje por refuerzo Minería de datos	es_compuesta_por
Componente	Overfiting Maldición de la dimensionalidad Red neuronal Proceso gaussiano Desambiguación	
Autor	David S. Touretzi Andrew McCallum Bernhard Schokopf	autor_de
Sinónimo	Aprendizaje automático ML Machine Learning	es-sinonimo_de

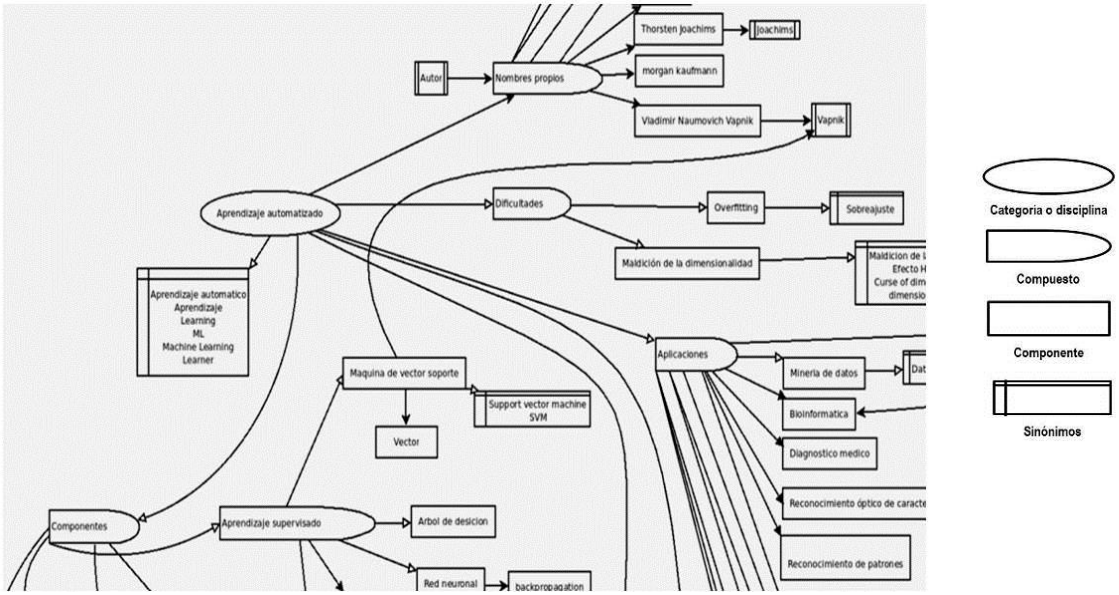

Figura 3: Parte del árbol de clasificación de atributos para el dominio Aprendizaje automatizado

4) Integración: Realizado un relevamiento de ontologías existentes, hay partes del dominio dentro de Inteligencia Artificial que están modelados en una ontología. Sin embargo, estas ontologías son muy poco profundas, describen este dominio sólo en forma general y se encuentran en idioma inglés. En este trabajo se construye una ontología en español, donde sólo se incluyen algunos términos en inglés como instancias de la clase sinónimo.

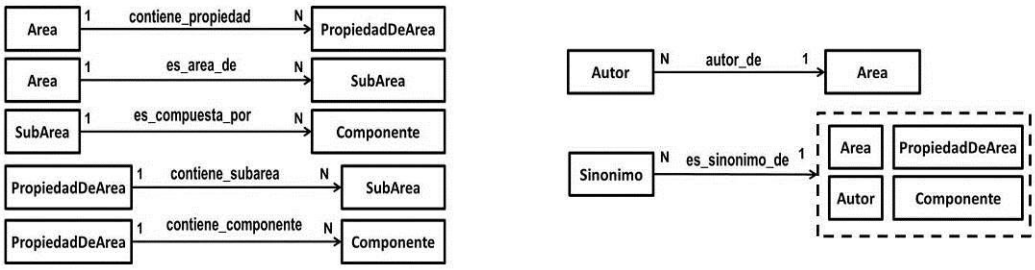

Figura 4: Relaciones consideradas

5) Implementación: En esta etapa se determina la herramienta para implementar la ontología. Entre las herramientas disponibles se pueden nombrar Protege (protege.stanford.edu/) y SWOOP (www.mindswap.org/2004/SWOOP). En [Alatrish, 2012] se presenta una buena comparación de los software existentes.

3.2. Asignación de pesos a términos de las ontologías

Para asignar pesos a cada término de las ontologías se ha utilizado la función propuesta en [Ganzendam, 2010]: $tf \,.\, rr(t) = tf(t) \,.\, rr(t)$, donde $tf(t) = 1 + \log(1+n(t))$ y $rr(t) = 1 + m \,.\, r_1(t) + m^2 \,.\, r_2(t)$. En estas expresiones $n(t)$ es el número de apariciones del término en el conjunto de documentos para el entrenamiento; $r_1(t)$ es el número de relaciones de longitud 1 que tiene el término t en la ontología y $r_2(t)$ es el número de relaciones de longitud 2 que tiene el término t en la ontología. Además, m es un promedio de la cantidad de relaciones que tiene un término en función de las relaciones de longitud 1 ó 2. Para este trabajo se ha utilizado para el cálculo de m las relaciones de longitud 1. En esta primer etapa, el entrenamiento de cada ontología se hizo sobre repositorios de 40 documentos de cada área. A modo de ejemplo se muestra en la Tabla 2 los primeros 5 términos junto con el valor que asume $tf \,.\, rr(t)$ para cada término t en el repositorio R_{ML} correspondiente al área Aprendizaje Automatizado.

Tabla 2: Ejemplo de asignación de pesos a términos del área Aprendizaje automatizado.

Término	Peso
aprendizaje	29,80
aplicación	28,05
aprendizaje supervisado	17,13
clustering	6,02
clasificador	5,18
.....	....

4. Arquitectura propuesta

En la Figura 5 se presenta la arquitectura general de la propuesta para la catalogación de recursos educativos utilizando vocabulario controlado. En esta primera etapa, se plantea trabajar con documentos de texto en formato pdf.

El módulo *Procesamiento del documento* se encarga del análisis léxico, se eliminan stopwords o palabras no significativas y se realiza stemming para llevar cada

término a un término raíz, lo que permite reducir la cantidad de índices por término. Para este procesamiento se evaluaron extractores automáticos de palabras clave que tienen implementadas estas tecnologías y se ha seleccionado Texlexan (texlexan.sourceforge.net/) porque es de código abierto, funciona bajo Linux, es sencillo de utilizar y contiene distintos tipos de análisis que se puede realizar sobre textos. El propósito del uso del extractor es obtener las palabras clave de los documentos que están en formato pdf. El resultado es un vector representativo con las palabras clave de cada documento.

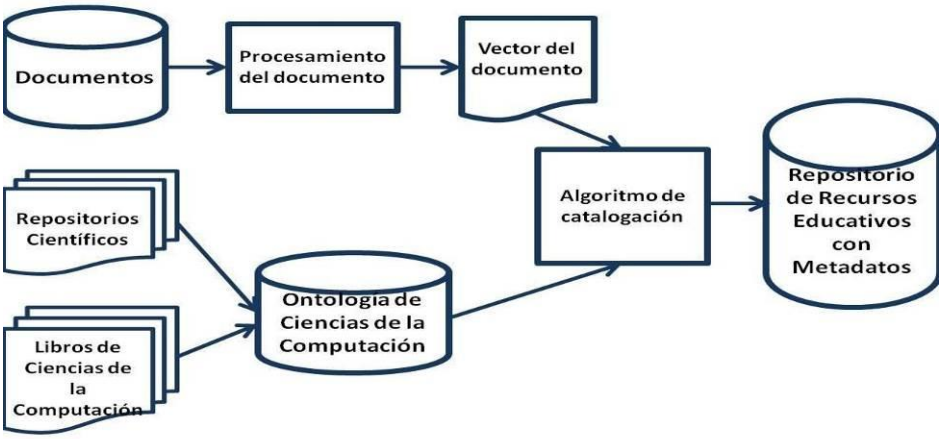

Figura 5: Arquitectura propuesta

El módulo *Algoritmo de catalogación* accede a la ontología que modela el dominio mediante vocabulario controlado. Se utiliza la ontología desarrollada en la Sección anterior. Este módulo calcula la distancia entre el vector que representa al documento y la ontología a fin de catalogar el documento. [Khan et al., 2009] y [Figuerola et al., 2004] describen y comparan algoritmos de clasificación y búsqueda que pueden utilizarse. En esta primera propuesta se utiliza la intersección del vector de palabras claves con la ontología. Los términos del conjunto resultante son ordenados por el peso de las palabras en la ontología. De estos términos se busca su camino a la raíz y de la unión de estos caminos se obtienen los términos para catalogar. El resultado es un documento categorizado y enriquecido con el metadato "Palabras clave" con vocabulario controlado, para ser cargado en el repositorio.

Se implementó un prototipo de esta arquitectura en lenguaje Java y para la extracción de palabras claves se utilizó Texlexan. Para un primer prototipo de la ontología y dado que éstas son ontologías livianas y para facilitar su recorrido mediante el catalogador se las implementó también en Java, utilizando matrices de incidencia.

4.1. Caso de uso

En esta Sección se presenta un caso de uso para la arquitectura propuesta.

1) *Carga de documentos:* se cargan los documentos que se desean catalogar en una carpeta predeterminada. Los documentos deben estar en formato pdf, en español y no hay restricciones para la cantidad de páginas. En este caso de uso se cataloga el siguiente documento: "Algoritmos de aprendizaje automático: aplicación en la solución a problemas medio ambientales". Anabel Vega Calcines. Ingeniera en Ciencias Informáticas, Universidad de Ciencias Informáticas. Cuba.

2) *Preprocesamiento de los documentos:* se convierten los archivos en formato pdf a texto plano txt utilizando pdftotext, se eliminan mayúsculas, se quitan los tildes y caracteres especiales de cada término.

3) *Extracción de palabras claves:* se utiliza Texlexan para obtener una lista de posibles palabras claves del documento que se almacenan en un vector denominado V_{ML}.

4) *Algoritmo de catalogación:* El algoritmo consiste en comparar por igualdad cada término del vector V_{ML} del documento con cada término de la ontología O_{ML}. La ontología se recorre en profundidad. En el caso que la igualdad sea verdadera, se almacena el recorrido del término hasta la raíz y se almacena el peso correspondiente al término en la ontología. Al final del recorrido del vector se tiene un conjunto de caminos (correspondientes a los términos coincidentes) y un peso total asignado para ese documento con respecto a una ontología en particular. Uno de los objetivos es construir estos caminos para retornar una catalogación en forma de árbol. Para este caso el resultado de los términos coincidentes es:

> $V_{ML} \cap O_{ML}$ = {aprendizaje, aplicación, algoritmo, predicción, clasificación, retropropagación, clustering, similitud, clasificador, vector, perceptrón, backpropagation}

Los términos de mayor peso son: {aprendizaje, aplicación, clasificación, clasificador, predicción} con pesos {29,80; 28,05; 5,43; 5,19; 3,07} respectivamente. Recuperando los caminos de estos términos en la ontología se agrega también el término raíz Aprendizaje Automatizado, resultando para su catalogación el siguiente conjunto de palabras clave: {aprendizaje automatizado, aplicación, aprendizaje, clasificación, predicción}.

5. Conclusiones y trabajo futuro

En este trabajo se ha aplicado la metodología Methontology para diseñar y construir ontologías de dominio en algunas áreas de Ciencias de la Computación. Estas ontologías contienen un vocabulario controlado y se utilizan para la asignación de palabras clave a los documentos a catalogar y cargar en un repositorio. Para la catalogación, en primer lugar se realizó un entrenamiento para asignar pesos a cada término de la ontología. Luego se utilizó un algoritmo que busca las palabras clave extraídas del documento a clasificar en la ontología, utiliza los pesos de estos términos y recupera los caminos a la raíz de los términos mejores rankeados. De este modo se obtiene un vector de longitud n de palabras clave del documento con términos controlados obtenidos de la ontología, cada uno con un peso asociado. Como trabajo futuro se plantea mejorar la primera etapa extracción de términos analizando qué partes del documento procesar y qué parte del vector obtenido considerar para obtener resultados óptimos. También se propone utilizar distintas distancias semánticas para buscar términos en la ontología y cómo ampliar la búsqueda a redes de ontologías.

6. Agradecimientos

Este trabajo ha sido parcialmente financiado por la Red CYTED 513RT0471 RIURE: Red Iberoamericana para la Usabilidad de Repositorios Educativos.

Referencias

Alatrish, E. S. (2012) “Comparison of Ontology Editors”. In: eRAF Journal on Computing, Facultad de Matematica, Universidad of Belgrado, Serbia.

Anjewierden, A. and Kabel, S. (2001) “Automatic indexing of PDF documents with ontologies”. In 13th Belgian/Dutch Conference on Artificial Intelligence, Holland. 23:30.

Barrios, J. y Gutiérrez, C. (2005) “Catalogación y búsqueda semántica en un sitio web”. In Proc. XXXI Conferencia Latinoamericana de Informática, Cali, Colombia.

Biébow, B. and Szulman, S. (1999) "TERMINAE: A Method and a Tool to Build a Domain Ontology". En: FENSEL, Dieter y Rudi STUDER (eds): 11th European Workshop on Knowledge Acquisition. London: Springer Verlag, pp. 4966.

Figuerola, C., Alonso Berrocal, J., Zazo Rodríguez, A., y Rodríguez, E. (2004) “Algunas Técnicas de Clasificación Automática de Documentos”, Cuadernos de documentación multimedia, ISSN-e 1575-9733, Nº. 15.

Gazendam, L., Wartena, C. and Brussee, R. (2010) “Thesaurus Based Term Ranking for Keyword Extraction”. In A. M. Tjoa & R. Wagner (eds.), DEXA Workshops (pp. 49-53), IEEE Computer Society. 2010.

Gómez-Pérez, A., Fernández, M. y Juristo, N. (1997) “METHONTOLOGY: From Ontological Art Towards Ontological Engineering”, Spring Symposium on Ontological Engineering of AAAI. Stanford University, California, pp 33–40.

Gruber T. (1993) “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”. Technical Report KSL-93-04, Knowledge Systems Laboratory, Stanford University, CA.

Khan, A., Bahurdin, B., and Khan, K. (2009) “An Overview of EDocuments Classification”. International Conference on Machine Learning and Computing, Singapur.

Noy, N. and McGuinness, D. (2001) “Ontology Development 101: A Guide to Creating Your First Ontology”. Technical Report SMI-2001-0880, Stanford Medical Informatics.

Uschold, M. and King, M. (1995) “Towards a methodology for building ontologies”. Workshop on basic ontological issues in knowledge sharing.

Valencia García, R. (2005) “Un Entorno para la Extracción Incremental de Conocimiento desde Texto en Lenguaje Natural”, Tesis doctoral. Universidad de Murcia, Departamento de Ingeniería de la Información y las Comunicaciones.

Wijewickrema, P. and Gamage, R. (2012) “Automatic Document Classification Using a Domain Ontology”. In National Conference on Library & Information Science (NACLIS 2012), Colombo, Sri Lanka.

Yedid, N. (2014) Modelos de indización temática utilizados en repositorios digitales de acceso abierto de Argentina. En Tendencias en la organización y tratamiento de la información, Comp. Barber E., Grequi C. y Pisano S. Ed. Biblioteca Nacional, Buenos Aires, Argentina.