UNIVERSIDAD AUTÓNOMA METROPOLITANA

Integración Semántica de Recursos en una Memoria Corporativa

Idónea Comunicación de Resultados para obtener el grado de

Maestro en Ciencias
(Ciencias y Tecnologías de la Información)
por

Erik Alarcón Zamora

Asesores:
Dra. Reyna Carolina Medina Ramírez

Dr. Héctor Pérez Urbina

16 de julio de 2013

Resumen

El área de Redes y Telecomunicaciones (RyT) del departamento de Ingeniería Eléctrica (IE) de la Universidad Autónoma Metropolitana (UAM) tiene una amplia y rica variedad (heterogeneidad en formato, contenido y estructura) de recursos de información. Algunos ejemplos de estos recursos de información son: los profesores y alumnos del departamento IE, artículos científicos, notas de curso, bases de datos de los trabajadores del dpto. IE, libros, presentaciones, manuales, inventarios, especificaciones de circuitos eléctricos.

Cada recurso representa el conocimiento sobre investigaciones, colaboraciones, proyectos, cursos y temas de interés de los profesores y alumnos en el dominio RyT. Por ejemplo, los artículos científicos, presentaciones, notas de curso e inclusive el propio profesor autor de estos documentos y multimedia son fuentes de información. Todo el conocimiento de una organización representado a través de los recursos, se conoce como memoria corporativa [1].

Una adecuada gestión del conocimiento en una memoria corporativa (MC) se traduce en varias ventajas a nivel operacional, como: una organización bien informada y con mejores tomas de decisión, una herramienta de aprendizaje para las personas adscritas a la organización, una base de conocimiento persistente y accesible para estas personas, un instrumento para búsqueda, recuperación e intercambio de conocimiento entre personas, por mencionar algunas.

Para llevar a cabo esta gestión de los recursos en una MC, se necesitan dos operaciones: 1) la representación del conocimiento sobre los recursos y 2) la búsqueda sobre esta representación. En las tecnologías de la Información, hay varios enfoques tradicionales de representar/buscar el conocimiento de los recursos, como: motores de búsqueda sintácticos y bases de datos relacionales. Pero, el enfoque que nos llamó la atención, es el de las Tecnologías Semánticas.

Las Tecnologías Semánticas se basan en el uso de tecnologías, herramientas y estándares para: la representación de los recursos en un formato estándar, establecer un vocabulario conceptual, la explotación del conocimiento mediante reglas, la busqueda y recuperación de la información a partir de la representación estándar, el uso de aplicaciones genéricas para la creación, manipulación y visualización de la información sobre los recursos, y para que los expertos en el dominio sean los encargados de suministrar y evaluar la información sobre los recursos.

En esta tesis de maestría, se propone una metodología para la representación, búsqueda, explotación e integración del conocimiento de los recursos de información en una memoria corporativa, mediante el uso de tecnologías semánticas. Esta metodología está guiada por

dos casos de uso base y la memoria corporativa es del área de RyT de la UAM.

- El primer caso de uso (Cartografía de competencias) consiste en la búsqueda de las personas (adscritas o relacionadas al depto. IE) a partir de sus características profesionales. En particular, se buscan a las personas por las competencias de profesionales, lingüísticas y sobre los temas que conocen de Redes y Telecomunicaciones. Por ejemplo, "todos los profesores de la UAM con conocimientos en radios cognitivos y que lean en inglés". Este primer caso también contempla la búsqueda de profesores que pueden impartir un curso, a partir de un conjunto de temas básicos que debe saber para dicho curso.

- El segundo caso de uso (Búsqueda de recursos digitales) consiste en la búsqueda de documentos y archivos multimedia, con base a uno o varios criterios de búsqueda (autor, título, año, temas de RyT, entre otros). Por ejemplo, "todos los artículos de Tim Berners Lee sobre Web Semántica y mayores al 2009".

La metodología para el desarrollo del modelo, la explotación y la integración del conocimiento sobre los recursos en una MC, se ha dividido en varias etapas que concuerdan con cada uno de los objetivos de la tesis. Los objetivos de la tesis son los siguientes:

- Un modelo (representación del conocimiento) de los recursos a partir de los dos casos de uso en un formato estándar.

- Un modelo coherente y del cual se explote el conocimiento sobre los recursos (ontología), a partir del uso de axiomas y un programa razonador.

- La busqueda y recuperación (integración) de los recursos que satisfagan las necesidades informativas de los usuarios, a partir de un motor de consulta.

- Un prototipo (navegación y consultas específicas) para la interacción fácil y visual de los usuarios con el modelo .

- Evaluar los resultados devueltos y el tiempo de ejecución de las consultas a la ontología.

En las tecnologías de la web semántica, el marco de descripción de recursos (RDF) es la solución para la representación del conocimiento de manera formal sobre los recursos en la MC. La representación se basa en la descripción de las características significativas o relaciones semánticas de/entre los recursos. Por ejemplo, Jorge Aparicio Reyes tiene 29 años, vive en el Estado de México, lee en Inglés, conoce a Erik Alarcón, estudia en la UAM y tiene conocimientos en sistemas operativos, java y flash.

Si bien cada recurso de la MC tiene un nombre propio, en el marco RDF cada persona, documento, multimedia o concepto tiene un identificador único de recurso [2] (URI). Con la finalidad de no tener ambigüedades a la hora de referirse a un recurso. Por ejemplo, el URI de Jorge Aparicio es http://www.mi-ejemplo.com/JorgeAparicio. Para cada recurso

(identificado con URI) se describen las características/relaciones en forma de triples (sujeto-predicado-objeto) y cada elemento de un triple es un URI o en algunos casos el objeto es una Literal.

Esta representación de las características se encuentra en un formato estándar y para almacenar estos triples, se emplea un triplestore. En este trabajo de tesis se empleó el triplestore Apache Jena que proporciona almacenamiento, un motor de consulta y un razonador.

Las descripciones representan la información explícita de los recursos, pero, esta información explícita tiene conocimiento implícito. Por ejemplo, un alumno, niño, profesor, empleado, madre, hijo son personas, pero éstas como tal no tienen un triple que establezca que son personas. Entonces, para explotar este conocimiento implícito de los recursos, se proponen un conjunto de reglas o axiomas que permiten establecer estas relaciones. Aunque, para materializar estos triples a partir de los axiomas, es necesario un programa razonador que infiera estos triples. Este razonador tambien permite encontrar inconsistencias en el modelo. Algunos triplestores integran o permiten importar un razonador, en el caso de Jena permite las dos opciones.

El modelo que captura el conocimiento explícito (descripciones) de los recursos y los axiomas que completan el conocimiento sobre éstos, se denomina ontología. En está tesis se hicieron dos ontologías; una para cada caso de uso, y tambien se modificó una ontología legada que tiene conceptos del área de RyT. Esta última ontología se emplea para vincular a personas, documentos y multimedia con los tópicos de RyT.

La consulta de los triples en el modelo, ya sea únicamente descripciones (triples explícitos) o una ontología con razonador (triples explícitos e inferidos), se hace con un motor de búsqueda (integrado en el triplestore) que compara los triples con un conjunto patrones; aquellos triples que concuerden, se recuperará la información que se solicitó en la consulta.

Un motor de consulta y un razonador que materializa triples en una ontología, son una buena combinación, ya que permiten consultar el conocimiento inferido (triples inferidos) y reducir la complejidad de las consultas. Por ejemplo, se tienen seis individuos que afirman que son alumno, niño, profesor, empleado, madre, hijo respectivamente, tambien se tienen los axiomas que establecen que alumno, niño, profesor, empleado, madre, hijo son personas y se tiene la siguiente pregunta "Quiénes son personas". Si se emplea solamente un motor de búsqueda, entonces no habrá ningún resultado, pero si se emplea la combinación motor y razonador, los seis individuos serán respuesta, porque estos seis individuos tienen el triple que afirma que son personas.

Los usuarios del área de RyT no están familiarizados con las tecnologías semánticas y en particular, al uso de la sintaxis de consulta. Entonces para facilitar a éstos la interacción y consulta del conocimiento de la ontología, se propone un prototipo que medie (interfaz) entre los usuarios y la ontología, específicamente este prototipo tiene los siguientes objetivos:

- Navegación a través de la información de los recursos; guiada por los casos de uso.
- Estructurar la pregunta de un usuario.
- Mapear las preguntas a consultas para el motor de consulta.

- Ejecutar la consulta con el motor de consulta, el razonador y la ontología.
- Publicar la información de los recursos respuesta en un formato visual agradable al usuario.

En esta tesis dos de los aspectos importantes a evaluar son: el desempeño de Apache Jena a la hora de consultar la ontología, así como el número y cuáles resultados responden estas consultas. Para llevar a cabo estas dos evaluaciones se obtuvieron un conjunto básico de preguntas para interrogar el modelo, para cada pregunta se sabe de ante manos el número y los recursos que la responden. En la primer evaluación, para cada consulta básica se calcula 20 veces el tiempo aproximado en milisegundos y se saca un tiempo promedio. Mientras, en la segunda evaluación, para cada consulta se compara el número/recursos que responde el motor con los recursos que previamente se sabe que la responden.

Las contribuciones de esta tesis son:

1. Una metodología para la Integración Semántica de Recursos en la MC de Redes y Telecomunicaciones.
2. Identificación y descripción de los principales escenarios de búsqueda/recuperación de los recursos en la MC de RyT.
3. Ontologías (Triples RDF + axiomas) que capturan el conocimiento de los recursos (apegados a los dos casos de uso) en la memoria corporativa RyT.
4. Prototipo para la consulta interactiva de los usuarios con las ontologías de RyT.
5. Evaluación del desempeño y calidad de resultados del triplestore Jena para la consulta de información.

Agradecimientos

Contenido

Lista de Tablas

Lista de Figuras

Acrónimos

Acrónimo	Descripción	Definición
RyT	Redes y Telecomunicaciones	1.1
UAM	Universidad Autónoma Metropolitana	1.1
IE	Ingeniería Eléctrica	1.1

Capítulo 1
Introducción

Algún texto...

1.1. Problema a tratar

Más texto...
RyT IE UAM

1.2. Motivación tecnologías semánticas

Más texto...

1.3. Objetivos

Más texto...

1.4. Contribuciones

Más texto...

1.5. Estructura del documento

Más texto...

Capítulo 2
Descripción del problema

2.1. Memoria Corporativa

Algún texto...

2.2. Casos de uso

Más texto...

Capítulo 3
Tecnologías Semánticas

3.1. Definiciones y descripciones

Algún texto...

3.2. Marco de Descripción de Recursos (RDF)

Más texto...

3.3. Lenguaje de consulta sobre grafos RDF (SPARQL)

Más texto...

3.4. Reglas de inferencia (RDF(S)/OWL) y razonadores

Más texto...

3.5. Ventajas de las tecnologías Semánticas

Más texto...

Capítulo 4
Estado del arte

4.1. Integración semántica de recursos de información

Algún texto...

4.2. Herramientas para la integración semántica de recursos

Más texto...

Capítulo 5

SIR: Integración semántica de recursos de información de una memoria corporativa

5.1. Representación del conocimiento de los recursos (modelo de datos)

Algún texto...

5.2. Explotación del conocimiento en el modelo (axiomas)

Más texto...

5.3. Búsqueda y recuperación (consulta) del conocimiento de los recursos

Más texto...

Capítulo 6
Evaluación

6.1. Escenarios de experimentación

Algún texto...

6.2. Experimentación

Más texto...

6.3. Resultados

Más texto...

Capítulo 7
Conclusiones y Trabajo Futuro

Algún texto...

Referencias

[1] L. Gandon, Fabien. Ontology Engineering: a Survey and a Return on Experience. Technical Report RR-4396, INRIA, March 2002.

[2] T. Berners-Lee, R. Fielding, and L. Masinter. Uniform resource identifiers (uri): Generic syntax. 1998.