Universidad Autónoma Metropolitana

Unidad Azcapotzalco

División de Ciencias Básicas e Ingeniería

Licenciatura en Ingeniería en Computación

Sistemas de almacenamiento semántico de textos de investigación mediante ontologías

Fernando Tébar Martínez

Matrícula: 2123999433

_____

Firma

Trimestre 2012 Primavera

9 de noviembre de 2012

Primera Versión

Asesora

Dra. Maricela Claudia Bravo Contreras

Departamento de Sistemas.

Objetivos

Objetivo general

Diseñar un sistema de reconocimiento, categorización y almacenamiento para los textos de investigación usando ontologías.

Objetivos específicos

- Adecuar el resultado del procesamiento de textos de tal manera que reconozca los atributos importantes en la ontología.
- Sincronizar la captación de los elementos con el almacenamiento inmediato en la ontología y base de datos.
- Relacionar todos los datos útiles del texto con alguna instancia en la ontología, y por tanto con alguna tabla en la base de datos.
- Completar todas las instancias de la ontología con la información de cada texto
- Descartar la información redundante en los textos y duplicada en la base de datos.

Introducción

Podemos considerar el entorno Web como un medio de comunicación y a la vez un sistema de almacenamiento y recuperación de la información. La situación actual de los recursos de información en Internet y el estado de desarrollo de las herramientas de búsqueda dificulta el análisis de la información y crea una incertidumbre sobre si aquello que se recupera es realmente relevante [1].

Cuando se quiere establecer patrones entre publicaciones similares es necesario crear unas relaciones comunes entre sus atributos (autor, universidad, tema, tipo de publicación, etc.).

Para establecer las relaciones, se usa el concepto de ontología. Ésta se define como una representación formal del conocimiento donde los conceptos, las relaciones y las restricciones conceptuales se explicitan mediante formalismos en uno o varios dominios.

Las ontologías, a pesar de ser una definición formal de las relaciones entre varias entidades, no nos acercan a la verdad absoluta, ni son la única solución para modelar un problema. Por ello se usa la palabra “proponer” cuando se presenta una ontología para un sistema de entidades en concreto.

Justificación

Lo que aquí se propone, es la definición de una ontología para un tipo de documentos con un patrón en común, los textos de investigación. Estos textos contienen gran cantidad de información muy diferente pero que comparte atributos asociados a las publicaciones académicas.

El almacenamiento persistente surge ante la necesidad de guardar los datos obtenidos de forma duradera para recuperarlos en otro momento. El término 'persistencia' es sinónimo de 'durabilidad' y 'permanencia'. Obviamente, el almacenamiento persistente en bases de datos supone grandes ventajas sobre el almacenamiento en memoria y el almacenamiento tradicional en el sistema de archivos.

Asimismo, es necesario establecer una ontología firme y completa para que, tras el procesamiento léxico, sintáctico y semántico, podamos catalogar con varios puntos de acceso principales todas las publicaciones para que la información relevante pueda ser accedida más tarde con mayor facilidad y consistencia.

Siguiendo esta dirección, la creación de una base de datos y una ontología, será la base para que el sistema de recuperación sea consistente y pueda tener un apoyo físico para las próximas búsquedas.

A la conclusión de este proyecto se obtendrá un almacenamiento persistente y eficiente para toda la información de cualquier tipo publicada en textos de investigación. Con ayuda de un lenguaje de definición de ontologías (OWL), un procesamiento sencillo de los textos y un módulo de intercambio entre la base de datos y la ontología, conseguiremos construir un sistema robusto de almacenamiento y recuperación de la información.

Trabajos relacionados

En general hay gran número de publicaciones sobre este tema; algunas ponen el foco en la definición consistente y formal de ontologías, otras potencian los procesadores de texto natural de modo que obtengamos información más fiable y se deseche el ruido y otras dando más peso al almacenamiento de esa información en una base de datos.

Internos

*Sistema clasificador de documentos de proyectos terminales usando el concepto de memoria asociativa [3].*Como en el proyecto que se va a desarrollar, en este proyecto las palabras caracterizarán cada uno de los proyectos terminales analizados y serán agrupados y categorizados por cada uno de los términos que emplean. Estas relaciones son una pieza fundamental de los nexos de la ontología, e imprescindibles para poder tratar el texto sobre una base firme de atributos.

Sistema de reconocimiento del alfabeto dactilológico utilizando procesamiento de imágenes [4]. En este proyecto, el procesamiento del elemento entrante difiere en cuanto que es una imagen la que será analizada. Primero se segmenta, se extraen las características, se reconoce (mediante unos patrones y un autómata celular) y después se presentan los resultados de forma numérica y no a través de búsquedas.

Implementación de un gestor de documentos [5]. Este proyecto comparte ciertos aspectos: implementa un gestor de bases de datos con atributos clave para optimizar las búsquedas. También diseña unos atributos relevantes para la ontología que sirve de pilar a la base de datos y categoriza y distribuye los documentos según hayan sido atribuidos.

Externos

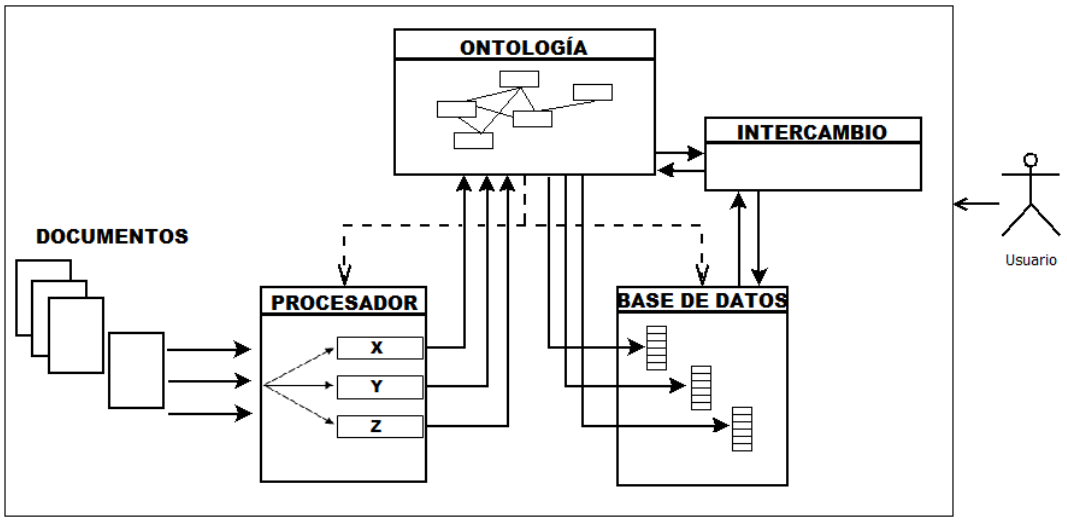
DBOWL [6], es un razonador persistente para ontologías OWL, que es la que se usará en la definición. El objetivo es proporcionar un razonador que permita consultar y razonar sobre ontologías con grandes cantidades de instancias, y que proporcione escalabilidad a dichos razonamientos y consultas. Por tanto, nuestra investigación está más enfocada a la implementación de razonamientos sobre instancias. DBOWL utiliza una base de datos relacional para almacenar las ontologías y para implementar los razonamientos, en contraposición a los razonadores de lógicas de descripciones que utilizan memoria principal. También se usa lógica de descripciones para inferir información sobre la estructura de la ontología y almacenarla en la base de datos.

El modelo de aprendizaje de ontologías que se presenta comprende la importación, extracción, acotamiento y refinamiento, posibilitando la modelación por parte del desarrollador. También se comenta, aparte de los diferentes grados de estructuración de las ontologías, el ciclo de aprendizaje desde texto libre, diccionarios y otros documentos, así como ingeniería inversa para ontologías.

Jena es un framework que nos provee de un entorno de desarrollo (lectura, procesamiento y escritura) para RDF, RDFS, OWL, mediante APIs así como también un motor de inferencia basado en reglas y un motor de búsqueda [7].

Descripción técnica

El diagrama de componentes y de uso de nuestra aplicación es el siguiente:



Ontología

Este módulo comprende la definición de la ontología a través de diagramas y su posterior codificación. En este proceso, tratamos de modelar una ontología consistente para los atributos relevantes de los textos de investigación, para ello se definirán los elementos fundamentales en cualquier texto de investigación o tesis: Título, autor, universidad/empresa, año, tema, etc. Nos basamos en el análisis teórico del formato de los textos de investigación para luego comenzar a tejer las redes de relaciones e instancias que servirán de base para el resto de módulos y procesos. En la imagen, las líneas punteadas corresponden a la dependencia de los otros dos módulos en cuanto a la estructura relacional de instancias.

El punto fundamental de la ontología será construir una estructura sólida de relaciones y semánticamente congruente. Emplear clases con un significado sólido, asociadas a palabras recurrentes en los textos y tratando de distribuir las entidades en los dominios siguiendo un alto nivel de abstracción evitando redundancias.

Procesamiento de textos

El procesador de textos implementará un analizador sintáctico y semántico mediante un programa en Java, que trate de reconocer un lenguaje de entrada natural de entorno académico, que es el que figura en los textos de investigación. Durante el análisis se comenzarán a crear tablas de símbolos y una estructura de atributos en memoria, modificándose a medida que transcurra el proceso.

Recibiremos una serie de documentos, se leerán, y los datos importantes de cada uno serán transferidos y almacenados primero poblando una ontología y más tarde en una base de datos distribuyéndolos con la misma forma en la que estos se han procesado en la ontología.

Almacenamiento

El diseño de la base de datos seguirá la misma construcción que la ontología, e incluso el diagrama nos podrá ayudar como esquema relacional. La base de datos y las tablas serán creadas y modificadas a medida que avance el proceso de reconocimiento de los textos y recibamos la información. También sus claves ajenas, claves primarias y tablas serán creadas, modificadas y destruidas según reconozcamos patrones nuevos y viejos en el documento.

Módulo de intercambio

Este módulo será implementado para proporcionar una mayor eficiencia a la hora de ejecutar consultas mediante las reglas de inferencia y el razonador lógico de la ontología. El objeto es tratar de aligerar carga física en el almacenamiento de la ontología y servir de caché para ir recuperando y devolviendo a la base de datos la información que no se vaya a usar.

Especificación técnica

Análisis de los textos

Analizar un documento de principio a fin reconociendo todas las palabras, siendo analizados todos los textos minimizando los errores, o al menos controlarlos y subsanarlos de manera retroactiva. Asimismo, se procesarán las palabras asignándolas a las entidades de la ontología, y si es necesario, se crearán. Se podrá permitir cualquier tipo de formato de texto, números, alfabeto, etc.

Para el reconocimiento de textos se usará la herramienta de software libre Gate [8]. Ésta es capaz de procesar textos de temática variada y arrojar resultados dados unos criterios. La herramienta comprende un IDE, un *framework* y una aplicación web.

Ontología

El lenguaje escogido para codificar la ontología es OWL, ayudándonos de la herramienta Protegé.

Almacenamiento de información

Se usará un servidor MySQL para almacenar toda la información tratando de eliminar la redundancia en los campos. Se atribuirán claves primarias y ajenas que sirvan para la correcta navegación e intercambio cuando se ejecute una consulta en la ontología.

Intercambio

Se programará el módulo de intercambio en Java con ayuda del *framework* Jena [9]. Se usará el API proporcionado para establecer reglas y protocolos para el intercambio ordenado e íntegro de la información entre los dos módulos.

Al concluir el proyecto terminal se entregarán tres discos compactos al Coordinador de Estudios de Ingeniería en Computación que incluirán el reporte final del proyecto terminal en un archivo PDF (sin restricciones) y el código fuente de la aplicación en un archivo comprimido (sin restricciones). El reporte final contendrá al menos: portada, resumen, tabla de contenido, objetivos, introducción, desarrollo del proyecto, conclusiones, bibliografía y apéndices. Los apéndices contendrán al menos un listado del código fuente desarrollado.

Calendario de trabajo

Para los dos próximos trimestres éste es el calendario de trabajo, sin embargo, está sujeto a todo tipo de modificaciones que ya se encuentran incluidas en la estimación, ya sea adelantos, o retrasos por enfermedad o dificultades encontradas.

Este esquema está basado en el trabajo diario de 6 ó 7 horas, que son las que el autor podrá dedicarle teniendo en cuenta que únicamente tendré una asignatura durante los dos próximos trimestres.

Proyecto Terminal 01 (Invierno 2013) **9 créditos**

Semana	Seriación	Descripción de tarea	Producto entregable	Horas
1	D01	Diseño de la ontología.	Diagrama de la ontología.	*9*
2,3,4	C01	Codificación de la ontología.	Código fuente en OWL.	*27*
5,6	D02	Definición del plan de procesamiento y salida del texto.	Algoritmos para el análisis del texto.	*18*
7,8,9	C02	Implementación del procesamiento	Reconocimiento de textos. Contenido útil del artículo clasificado e insertado en estructuras.	*27*
10	D03	Diseño de la base de datos que almacenará los datos.	Esquema relacional de la base de datos diseñada.	9
11	C03	Creación de las tablas e integración de la base de datos en la estructura.	Base de datos creada de forma consistente con la ontología.	*9*
			Total	*99*

Proyecto Terminal 02 (Primavera 2013) **18 créditos**

Semana	Seriación	Descripción de tarea	Producto entregable	Horas
1	C04	Transferencia de información entre el procesador y la ontología	El procesador entrega los datos en estructuras en memoria y pueblan la ontología.	18
2	C05	Transferencia de información entre la ontología y la base de datos	La ontología entrega los datos ordenados y se insertan en las bases de datos en las tablas correspondientes a las entidades.	18
3	D04	Diseño del plan de intercambio entre el razonador de la ontología y la base de datos.	Integración plena e intercambio ordenado de la información requerida y desechada por cada una de las partes	18
4	C06	Implementación del módulo de intercambio.	Codificación usando Jena del módulo de intercambio.	18
5,6,7	P01	Pruebas y corrección de cada una de las tareas por separado.	Cada una de las tareas recibe la información correcta, la procesa y lo entrega al módulo siguiente.	54
8,9,10	P02	Pruebas y corrección de la integración plena e interoperabilidad de los diferentes módulos.	El programa trabaja correctamente. Lee el archivo, la organiza según la ontología y almacena en la base de datos la información útil.	54
1-11	R01	Redacción del reporte final.	Reporte del Proyecto terminal	18
			Total	198

Recursos

Los recursos previstos para la ejecución del proyecto son:

- Protegé
- OWL
- Windows 7
- MySQL
- Eclipse
- Microsoft Word
- Java 7

El alumno proporciona una computadora portátil con las siguientes especificaciones: procesador de 2.2 GHz, 4 GB de RAM y 640 GB de disco duro.

El asesor se responsabiliza de guiar al alumno y de que todos los recursos anteriormente citados estarán disponibles para el alumno, de modo que el proyecto terminal se pueda concluir en tiempo y forma. Asesora: Dra. Maricela Claudia Bravo Contreras

_____

Firma de la asesora

Bibliografía

[1] González Pérez Y. "*Las ontologías en la representación y organización de la información*"**.** Tesis. Bibliotecología y Ciencias de la Información. Universidad de la Habana. Cuba. 2006.

[3] Ugalde Anaya, Juan.L. *Sistema clasificador de documentos de proyectos terminales usando el concepto de memoria asociativa*. Proyecto Terminal. Universidad Autónoma Metropolitana, México DF, México. 2011.

[4] Marín Díaz, L. *Sistema de reconocimiento del alfabeto dactilológico utilizando procesamiento de imágenes.* Proyecto Terminal. Universidad Autónoma Metropolitana, México DF, México. 2010.

[5] Flores Casillas, C.A. *Implementación de un gestor de documentos.* Proyecto Terminal. Universidad Autónoma Metropolitana, México DF, México. 2011.

[6] Roldán-García, M.M. y Aldana-Montes, José F. *DBOWL: Persistencia y Escalabilidad de Consultas y Razonamientos en la Web Semántica.* Artículo. Departamento de Lenguajes y Ciencias de la Computación, Universidad de Málaga, Málaga, España. 2006.

[7] Maedche, A. y Staab, S. (01/11/2012). *Ontology Learning for the Semantic Web.* Artículo. Universidad de Karlsruhe, Alemania. 2008. [En línea] Disponible en: http://icc.mpei.ru/documents/00000833.pdf

[8] GATE (5/12/2012) [En línea] Disponible en: http://gate.ac.uk/

[9] Jena Apache (1/11/2012) [En línea] Disponible en: http://jena.apache.org/