Aplicaciones de Ontologías a Problemas Lingüísticos: Bases de Conocimiento Basadas en Texto No Estructurado

Henrión, Guillermo; Azcurra, Diego; Soria, Marcelo, Tabares, Diego

Departamento de Ingeniería
Universidad Nacional de Tres de Febrero
Valentín Gómez 4752 (1678)
Caseros, Tres de Febrero
Buenos Aires, Argentina

Facultad de Agronomía
Universidad de Buenos Aires
Av. San Martín 4453
Ciudad Autónoma
Buenos Aires – Argentina

Dpto. Desarrollo Productivo y Tecnológico
Universidad Nacional de Lanús.
29 de Septiembre 3901 (1826)
Remedios de Escalada, Lanús
Buenos Aires, Argentina.

Resumen

En proyectos anteriores de ontologías biomédicas se realizó un estudio exhaustivo de las ontologías y sus aplicaciones a las ciencias biomédicas, obteniendo además resultados sobre medidas de similitud semántica. Posteriormente se utilizó este conocimiento en problemas lingüísticos sobre textos del ámbito biomédico principalmente.

En este proyecto, y en la misma dirección, se continuará avanzando en la utilización sobre problemas lingüísticos, pero esta vez extendiendo su uso no solo a texto estático sino aplicados a distintas fuentes de conocimiento textuales, como ser RSS, redes sociales, datos abiertos, sitios específicos y datos enlazados.

Se propondrá el diseño de un motor que procese estas fuentes y las integre a una base de conocimiento unificada, utilizando ontologías como forma de representación.

Palabras clave: minería de datos, tecnologías semánticas, ontologías, representación del conocimiento, procesamiento del lenguaje natural, datos enlazados, datos abiertos, minería de grafos.

Contexto

Este proyecto de investigación continúa la línea de trabajo en aplicaciones de Minería de datos y tecnologías semánticas en el marco de la carrera de Ingeniería en Computación de la Universidad Nacional de Tres de Febrero, en colaboración con FCEN y de la Carrera de Licenciatura en Sistemas de la Universidad Nacional de Lanús.

Introducción

Actualmente existe un creciente interés sobre cómo adquirir y estructurar conocimiento obtenido desde distintas fuentes, ya sean estas redes sociales, noticias publicadas mediante RSS e información publicada en forma de datos abiertos, y sobre cómo este conocimiento puede ser integrado de manera de poder sacar conclusiones imposibles de obtener sin la interacción de todas estas fuentes.

Para poder construir esta base de conocimiento integrada es necesario primero establecer una serie de cuestiones

sobre cómo proceder tanto al recuperar la información desde las distintas fuentes, cómo al representar esta información dentro de la base de conocimiento para que pueda ser manipulada como tal, y finalmente cómo los distintos conceptos obtenidos se integran para formar una base consistente.

Todos los conceptos deben quedar representados formalmente, en algún lenguaje de representación adecuado, motivo por el cual se incluye dentro de nuestro proyecto una línea de investigación en tecnologías de representación del conocimiento, principalmente, pero no excluyente, referidas a lógicas para la descripción.

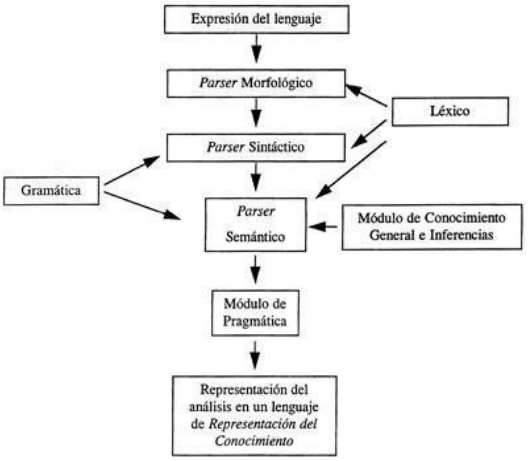
Una forma de representar el conocimiento es mediante el uso de ontologías. Una ontología es una representación formal en donde cada concepto es representado como nodos de un grafo, en donde los arcos indican las relaciones entre los conceptos. La definición formal de acuerdo a Gruber es:

"*Una especificación explícita y formal de una conceptualización compartida*".

Pero para poder llegar a esta representación existe una fase de extracción de conocimiento, en donde son utilizadas técnicas de procesamiento de lenguaje natural [10]

El procesamiento del lenguaje natural (PLN) es el campo que combina las tecnologías de la ciencia computacional (como la inteligencia artificial, el aprendizaje automático o la inferencia estadística) con la lingüística aplicada, con el objetivo de hacer posible la comprensión y el procesamiento asistidos por ordenador [3].

La etapas involucradas en el procesamiento del lenguaje natural involucran los análisis siguientes: léxico, sintáctico, semántico, pragmático y finalmente la representación de los conceptos extraídos en algún lenguaje formal de representación del conocimiento [4].



En el caso del proyecto presente las fuetes posibles a ser analizadas son las siguientes:

RSS: RSS son las siglas de Really Simple Syndication, un formato XML para sindicar o compartir contenido en la web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos. [9]

Datos abiertos: Los datos abiertos son datos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, cuando más, al requerimiento de atribución y de compartirse de la misma manera en que aparecen [2].

Redes sociales: Las redes sociales son sitios de internet que permiten a las

personas conectarse con sus amigos e incluso realizar nuevas amistades, de manera virtual, y compartir contenidos, interactuar, crear comunidades sobre intereses similares: trabajo, lecturas, juegos, amistad, relaciones amorosas, relaciones comerciales, etc [7].

Datos enlazados: Los Datos Enlazados es la forma que tiene la Web Semántica de vincular los distintos datos que están distribuidos en la Web, de forma que se referencien de la misma manera que lo hacen los enlaces de las páginas web [1].

Otras fuentes: Blogs, correos electrónicos, artículos académicos, etc.

Líneas de Investigación, Desarrollo e Innovación

Este proyecto se inscribe en una línea de investigación que busca desarrollar y sistematizar el cuerpo de conocimiento de las tecnologías semánticas, principalmente en lo relacionado a representación del conocimiento, similitud semántica y procesamiento del lenguaje natural.

Así como en proyectos anteriores se estudiaron las medidas de similitud más adecuadas para establecer la similitud semántica dentro de una ontología, en el presente proyecto se estudiarán las distintas fases del procesamiento del lenguaje natural para extraer los conceptos y relaciones que constituirán la ontología que represente el conocimiento de las fuentes originales.

Resultados y Objetivos

El objetivo general de este proyecto es la aplicación de algoritmos extracción de conceptos y relaciones desde fuentes no estructuradas sobre Internet, con el fin de diseñar una base de conocimiento, que utilice ontologías como forma de representación.

En esta línea, como objetivos particulares, identificamos:

1. Proveer herramientas informáticas a investigadores de distintos ámbito (principalmente adecuados para estudioso biológicos y sociales) mediante las cuales puedan organizar su información, vincular los resultados de diferentes experimentos y realizar análisis automáticos.

2. Diseño de un motor que procese fuentes de texto diversas y las integre a una base de conocimiento unificada, utilizando ontologías como forma de representación.

3. Aplicar esta tecnología en otros dominios, en donde estén disponibles grandes volúmenes de información con relaciones complejas.

4. Un objetivo transversal del proyecto es atraer e interesar a los alumnos en temas avanzados en ciencias de la computación y acercarlos a actividades de investigación.

Formación de Recursos Humanos

El equipo de trabajo está conformado por tres investigadores formados y un estudiante avanzado de la carrera de

Ingeniería en Computación. Asimismo, colaboran en el proyecto otros estudiantes de la carrera, quienes han manifestado su interés en desarrollar su trabajo de fin de carrera en la línea de investigación presentada.

Referencias

[1] Guía breve de Linked Data http://www.w3c.es/Divulgacion/GuiasBreves/LinkedData

[2] ¿Qué son los datos abiertos? http://opendatahandbook.org/guide/es/what-is-open-data/

[3] Procesamiento del Lenguaje Natural

http://www.vicomtech.org/t4/e11/procesamiento-del-lenguaje-natural

[4] Eduardo Sosa (1997)

Procesamiento del lenguaje natural: revisión del estado actual, bases teóricas y aplicaciones

[5] Gruber, T. (1995). "Toward Principles for the Design of Ontologies Used for Knowledge Sharing".

International Journal of Human-Computer Studies 43 (5-6): 907–928. doi:10.1006/ijhc.1995.1081.

[6] Chu-ren Huang "Ontology and the Lexicon - A Natural Language Processing Perspective" Cambridge University Press 2010

[7] Matthw Denny (2014)

Social Network Analysis

[8] Alexander Clark, Chris Fox, and Shalom Lappin (2010)

The Handbook of Computational Linguistics and Natural Language Processing

[9] RSS Explicada

http://www.rss.nom.es/

[10] John Davies, Marko Grobelnik (2009)

Semantic Knowledge Managemen