Universidad Autónoma Metropolitana Unidad Azcapotzalco
División de Ciencias Básicas e Ingeniería
Licenciatura en Ingeniería en Computación

Propuesta de proyecto terminal:

Sistema de recuperación de información semántico

Selene María de Jesús Ugalde Chávez
Matrícula: 205304493

Trimestre 11O
11 de noviembre de 2011

Versión 1.0

Asesores

Dra. Maricela Claudia Bravo Contreras

Profesor del departamento de sistemas

M. Hugo Pablo Leyva

Profesor del departamento de sistemas

Objetivo general

Diseñar un sistema de recuperación de información semántico para recopilar, minar y descubrir información relevante sobre un conjunto de documentos.

Objetivos específicos

- Aplicar técnicas de Procesamiento del Lenguaje Natural para analizar textos por su contenido semántico
- Construir un sistema que permita generar graficas que representen el significado de una oración
- Convertir la grafica del significado de una oración en información organizada y comprensible para el usuario.
- Diseñar una interfaz de consulta que permita al usuario visualizar la información generada por el sistema.

Introducción

La computación ha revolucionado la forma en que el hombre accede y almacena la información: la digitalización ha hecho posible las bibliotecas virtuales y en tan sólo dos décadas, la Internet ha crecido considerablemente y se ha convertido en una herramienta cotidiana, con la que podemos tener acceso a miles de documentos con un sólo clic. Sin embargo, cantidad no significa calidad; aunque contemos con potentes buscadores que pueden arrojar millones de resultados relacionados para la más simple consulta, eso sólo significa que no se encontró una respuesta exacta.

Este inconveniente obedece a un problema de comunicación. Las maquinas y su poderosa capacidad de procesamiento está construida sobre un lenguaje formal muy distinto al nuestro, mientras que la mayor parte de la información a procesar se encuentra representada en lenguaje humano. Evidentemente, son hombres que poco o nada debieran saber sobre lenguaje máquina los que hacen las consultas y las hacen en su lenguaje. Es así como surge la necesidad de máquinas que entiendan nuestro idioma, en un mundo donde la información oportuna marca la diferencia y la cantidad de datos a procesar es tan grande que necesita ser automatizada.

Como una respuesta entre varias a esta necesidad surge, entre las ramas de la inteligencia artificial y la lingüística computacional, el Procesamiento del lenguaje natural NLP (del idioma inglés *Natural Language Processing),* que se define como una serie de técnicas enfocadas a que las maquinas sean capaces de manejar lenguajes no formales. [1] Su objetivo es automatizar la comprensión del lenguaje natural. Las áreas de aplicación más importantes del NLP son la traducción automática, la recuperación y extracción de información y las interfaces en lenguaje natural.

Con estos antecedentes en mente, este proyecto pretende utilizar técnicas del NLP para analizar el significado del contenido de una serie publicaciones.

Justificación

Uno de los objetivos importantes que se persigue hoy en la vanguardia de la tecnología informática es hacer de la gran magnitud de documentos a nuestra disposición una ventaja y no una desventaja [2]; Si además de tener una colección de documentos pertinentes, contáramos con una herramienta que procesara el contenido de dichos documentos, la posibilidad de extraer información útil crecería significativamente.

Esta propuesta sigue al quehacer computacional en esa dirección, en el sentido de aplicar técnicas innovadoras para la extracción automatizada de información. Ni más ni menos que la tarea de un ingeniero: Aplicar los conocimientos y técnicas científicas disponibles, para la resolución de problemas que afectan a la sociedad en su actividad cotidiana.

El resultado inmediato del proyecto será una herramienta que recupere información relevante sobre las publicaciones de los profesores de la UAM Azcapotzalco. La información generada puede ser útil para delinear de manera automática los perfiles de investigación de nuestros profesores. Dentro de sus aplicaciones a futuro, puede funcionar desde un medio para lograr el acercamiento y colaboración con otras instituciones, hasta un mapa de la investigación desarrollada por la Universidad. También representaría la incursión de nuestra institución en el campo de NLP a nivel licenciatura.

Antecedentes

En la Universidad no existen proyectos registrados con relación directa a este proyecto.

En instituciones externas, públicas y privadas los proyectos más relevantes con relación a esta propuesta se exponen a continuación:

- *Arnetminer* [3]

  *Arnetmine*r es una herramienta diseñada para realizar operaciones de búsqueda y minado sobre publicaciones en internet. Usa el análisis de redes sociales para identificar conexiones entre investigadores, conferencias y publicaciones. Esto le permite proveer servicios como, búsqueda por asociación, hallazgo de expertos, búsqueda por recorrido, evaluación académica, y modelado por tópico.

  Fue creada como un proyecto de investigación en el análisis sobre influencia social, clasificación y extracción de redes sociales. Ha estado en operación por más de tres años, durante los cuales ha indexado setecientos mil investigadores y más de tres millones de publicaciones. La Investigación fe fundada por el programa nacional de alta tecnología R&D de China. La fundación nacional de China y el laboratorio de investigación IBM China junto con otros patrocinadores.

  Es comúnmente usada por academias para identificar relaciones y dibujar correlaciones estadísticas sobre sus investigadores. El producto es usado en el estudio encaminado a verificar la popular noción de que no más de seis grados de separación conectan a cualesquiera dos personas en el mundo.

Aunque *Arnetminer* utiliza entre otras técnicas de minado y no NLP para extraer información de documentos, es semejante es su propósito de determinación de perfiles académicos, y aunque persigue objetivos distintos como lo es el análisis de redes sociales; si brinda una perspectiva interesante sobe el interés, la importancia y la utilidad de la extracción de información y la definición de perfiles, así como una manera de abordar estas problemáticas y algunas aplicaciones de su solución.

- *Wordnet*[4]

  WordNet es una enorme base de datos léxica del idioma inglés. Agrupa las palabras en conjuntos de sinónimos llamados 'synsets', que proporcionan definiciones cortas y generales y almacenan las relaciones semánticas entre estos

conjuntos de sinónimos. El propósito de este proyecto es doble: por un lado producir una combinación de diccionario y tesauro cuyo uso es más intuitivo, y ayudar al análisis automático de textos y a las aplicaciones de inteligencia artificial. La base de datos y las herramientas se han liberado bajo una licencia BSD y pueden ser descargadas y usadas libremente. Además la base de datos puede consultarse online.

Fue creada y es mantenida por el *Cognitive Science Laboratory* de la Universidad de Princeton bajo la dirección del profesor de psicología George A. Miller. El desarrollo comenzó en 1985. Durante los años el proyecto ha recibido alrededor de tres millones de dólares, principalmente a través de agencias gubernamentales interesadas en traducción automática.

- *Sophia Semantic Engine[*5]

El motor de *Sophia* Semántica es un software comercial de origen Italiano que analiza y comprende el lenguaje natural, creando una capa de interpretación en las aplicaciones que interactúan con los usuarios de una forma lingüística y las aplicaciones que se ocupan de la información no estructurada.

Sus capacidades incluyen:

- extracción de información y anotación de textos no estructurados (identificación de los eventos, personas, nombres, lugares, marcas)
- Clasificación automática y etiquetar los documentos
- Extracción automática y la construcción de un dominio específico de terminología léxicos.

El motor *Sophia* tiene mucha relación con esta propuesta pues uno de sus principales objetivos es la recuperación de la información. Pero esta propuesta se diferencia por integrar todo el proceso de análisis semántico en una sola unidad, así los resultados que devuelva dependerá de los textos que analiza. La salida del motor de *Sophia* en cambio, dependerá de las aplicaciones que se encargan de manejar tanto los documentos y la comunicación con el usuario porque esta herramienta funciona sólo como n intermediario entre ellas.

- Asociación Mexicana de Procesamiento del Lenguaje Natural (AMPLN) [2]

La AMPLN es una organización profesional no lucrativa, cuya misión es fomentar la interacción e intercambio de ideas entre especialistas mexicanos en el procesamiento de lenguaje natural (PLN), así como difundir los logros y la importancia del PLN entre la sociedad nacional.

- *Association for Computational Linguistics* (*ACL*)[6]

La Asociación de Lingüística Computacional es la sociedad científica internacional y profesional para las personas que trabajan en problemas relacionados con el lenguaje natural y la computación. Su revista *ACL*, Lingüística Computacional, sigue siendo el principal foro para la investigación en lingüística computacional y procesamiento del lenguaje natural. Desde 1988, la revista se ha publicado para la *ACL* por *MIT Press* para ofrecer una base más amplia de distribución. Además organiza una reunión anual que se celebra cada verano en los lugares donde se realiza la investigación significativa sobre la lingüística computacional.

Descripción técnica

Se pretende desarrollar sistema de recuperación de información mediante técnicas de NLP. A continuación se muestra el diagrama del sistema en la figura 1.0 y más adelante la descripción de cada modulo.

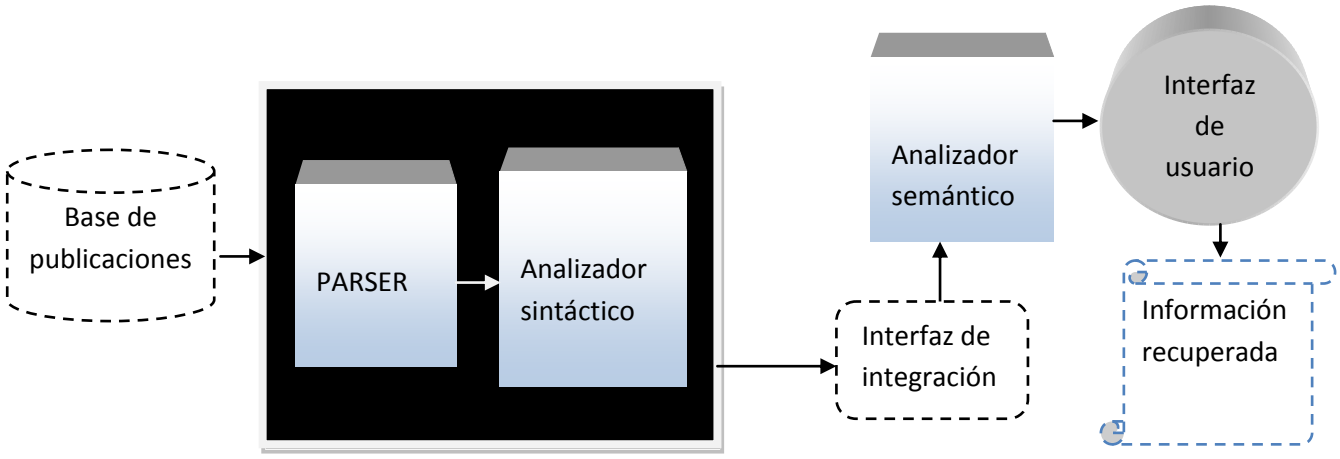

Figura 1.0

Este proyecto sólo desarrollará el analizador semántico, sin embargo para implementarlo es necesario contar con un *parser* y un analizador sintáctico. Es importante aclarar que estos módulos no se programarán, si no que se utilizaran herramientas de terceros para este propósito.

- Base de datos de publicaciones.

La base de datos de publicaciones es el almacén de la colección de publicaciones que servirán como datos de entrada del sistema junto con los parámetros introducidos por el usuario.

- Interfaz de Integración.

La función de este modulo es permitir la integración entre el analizador sintáctico y semántico, es decir hace que los datos arrojados por el analizador semántico sean comprensibles para el analizador semántico.

- Analizador semántico

Este módulo está subdivido en 3 fases según su flujo de funcionamiento:

- Representación del significado: En este periodo el texto se acomoda en graficas que representan el significado de las oraciones mediante las relaciones semánticas entre las palabras que componen dichas oraciones.

- Extracción de la información: En esta etapa se trabajan las representaciones del significado obtenidas en la etapa anterior para descubrir y extraer la información que se busca.

- Generación de resultados: Finalmente la información obtenida se convierte en datos organizados que el usuario pueda entender y manejar fácilmente.

- Interfaz de usuario

Este módulo se encarga de la comunicación entre el usuario y el sistema. Se encarga de presentar las funcionalidades al sistema, recuperar los datos de entrada y devolver los resultados al usuario.

Especificación técnica

El dominio de los documentos que procesará el sistema estará limitado a las publicaciones de los profesores del departamento de sistemas y sólo soportará documentos en formato pdf.

El sistema recibirá como parámetros de entrada la base de datos de publicaciones, los procesará mediante técnicas NLP y mínimamente regresará la información extraída de dicho análisis de forma organizada y comprensible.

Los entregables para este proyecto son:

- Código fuente documentado de la aplicación.
- Esquema de la base de datos.
- Diagramas UML de casos de uso, clases y navegación.
- Diccionario de datos.
- Manual de instalación y configuración.
- Manual de usuario.
- Protocolo de pruebas

El proyecto se dará por concluido cuando se entregue la documentación señalada

Calendario de Trabajo

Enseguida se describe el calendario de trabajo para este proyecto dividido en 2 trimestres correspondientes a los 9 créditos (99 horas, 9 por semana) del Proyecto

Terminal de Ingeniería en Computación I y el otro correspondiente a los 18 créditos (198 horas, 18 por semana)

Trimestre 12-I	1	2	3	4	5	6	7	8	9	10	11	Horas
Recopilación de publicaciones												9
Instalación y configuración de herramientas a utilizar												9
Diseño de Base de Datos												9
Desarrollo de Interfaz entre modulo sintáctico y semántico												18
Desarrollo del modulo generador de representaciones del significado												36
Primera revisión												27

Trimestre 12-O	1	2	3	4	5	6	7	8	9	10	11	Horas
Recuperación de la información (Discriminación y Transformación de Representaciones del significado)												54
Automatización de la organización y de la información recuperada												54
Desarrollo de Interfaz de usuario												36
Segunda revisión y pruebas												36
Redacción del manual de usuario y configuración												18

Recursos

- ✓ Software

Los programas que se utilizarán en la implementación del proyecto (gestor de base de datos, Entorno de desarrollo, etc.) serán de carácter libre.

Se cuenta con el acceso a internet para su descarga.

- ✓ Hardware

  No se requiere de equipo especializado para el desarrollo del proyecto.

  Se cuenta con una computadora personal con características suficientes para el desarrollo del mismo.

Bibliografia

[1] B. Coppin, "Understanding Language" in *Artificial intelligence illuminated*, Canada: Jones and Bartlett Publishers, 2004

[2]AMPLN. (2009, noviembre 18). *¿Qué es Procesamiento del Lenguaje Natural?*. [Online]. Available: http://www.ampln.org/pmwiki.php?n=Main.PLN

[3]Arnetminer. (2010, March 10). *Introduction.*. [Online]. Available: http://arnetminer.org/introduction

[4]Princeton University. (2011, June 21). *What is WordNet?*. [Online]. Available: http://wordnet.princeton.edu/

[5]CELI. (2011).Sophia Semantic Engine. [Online]. Available: http://www.celi.it/en/sophia-semantic-engine.shtml

[6] ACL. (2011). *About the ACL*. [Online]. Available: http://www.aclweb.org/index.php?option=com_content&task=view&id=38&Itemid=35

[7] D. Jurafsky and J.H. Martin, "Representing meaning" in *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Marcia Horton, ed. New Jersey: Prentice Hall, 2000.

[8]F. Verdejo *et al.* (1999, June 2). *Information retrieval with NLP techniques*. [Online]. Available: http://nlp.uned.es/~ircourse/

[9]R. Johansson, "*Dependency-based Semantic Analysis of Natural-language Text*" , Ph.D. dissertation, Dept. Comp. Science, Lund Univ., Sweden, 2008

[10]T. Moure y J. Llisterri. (2010, October 10). *Lenguaje y nuevas tecnologías: el campo de la lingüística computacional*. [Online]. Available: http://liceu.uab.es/~joaquim/publicacions/llisterri_moure_96.html

[11]A. Moreno. (2000). *Diseño e implementación de un lexicón computacional para lexicografía y traducción automática*, [Online]. Available: http://elies.rediris.es/elies9/index.htm

[12]A. Gelbukh y G. Sidorov, *Procesamiento automático del español con enfoque en recursos léxicos grandes*, CIC ed. México: Instituto Politécnico Nacional, 2006