UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

La Universidad Católica de Loja

ÁREA TÉCNICA

TITULACIONES DE INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

Desarrollo de Servicios Web para el proceso de Enlace y Enriquecimiento de Datos Enlazados

TRABAJO DE FIN DE TITULACIÓN

Autor: Montaño Sozoranga, Wilmer Fabricio

Director: Piedra Pullaguari, Nelson Oswaldo, Ing

Loja - Ecuador
2014

Contenido

CAPITULO I: MARCO TEÓRICO

1.1. Datos Enlazados

1.1.1. Introducción

En sus inicios la web en su primera versión 1.0, donde web era rígida en cuanto a la entrega de información, además de poco actualizada, convertía al visitante de un sitio web un simple lector, restringido de cualquier interacción. Se puede decir que la web no era más que paginas enlazadas mediante hipervínculos.

La web que siempre está creciendo y evolucionado, alcanza su versión conocida como la web 2.0 en donde usuario juega el papel más importante, es quien evalúa, puede calificar, compartir, rectificar, alimentar y subir su propia información a la web. Esto producto de le aparición de nuevas tecnologías y estandarización[1].

Los datos relacionados llegan para dar forma a la siguiente versiona de la web, la web semántica de W3C[2] nos dice : "Linked Data se refiere a la utilización delas mejores prácticas para publicación, estructuración de los datos en la web, de tal forma que puedan ser enlazados entre sí, utilizando tecnología propias de web semántica como RDF, OCW, SPARQL, etc."

Se refiere en si a la estructura de la de la siguiente generación de la web, como es la web semántica, que en si busca que la información que se en publica en internet pueda no solo ser entendida por seres humanos sino también por la maquinas que navegan en la web. En donde a partir de un dato podemos descubrir otros datos por sus relaciones.

1.1.2. Principios de Datos Enlazados

Tim Berners Lee en su publicación Linked Data - Design Issues (Berners-Lee, 2006)describe cuatro reglas base para la publicación de datos enlazados:

1. Usar URIs como nombre de las cosas
2. Usar URIs HTTP par que esas cosas puedan ser referenciadas
3. Representar los datos en RDF y SPARQL como lenguaje de consulta
4. Incluir enlaces hacia otra cosas, para descubrir más cosas

[1] La evolución de la web en cuanto a tecnologías fuete: http://www.evolutionoftheweb.com
[2] Consorcio internación que se dedica a la desarrollo y publicación de recomendaciones para WWW.

La utilizar de estas reglas para la publicación de datos, nos permite que estos por las características propias de las tecnologías sobre las cuales se construyen como:

- Las cosas que nombremos por URIs son inequívocas y estos serán recursos.
- Los detalles o atributos y las relaciones de los datos van a estar descritos y estructurados en formato RDF
- Podremos acceder o realizar consultas sobre estos mediante SPARQL
- La cosas que publiquemos estarán relacionados

1.1.3. Tecnologías

1.1.3.1. URI

El RFC 3986 (W3C, 2005) nos dice lo siguiente: “Un identificador uniforme de recursos (URI) proporciona un medio simple y extensible para la identificación de un recurso.” Es debido un URI es inequívoco por lo tano pueden existir un URI repetido un ejemplo de la sintaxis es la siguiente http://tools.ietf.org/html/rfc3986”

1.1.3.2. HTTP

Protocolo de transferencia de hipertexto (Hypertext Transfer Protocol) se basa en un esquema petición respuesta que se da entre un cliente y un servidor, es el protocolo que dio origen a la web y aun hora es la base para el para la evolución de la web. Es un protocolo de nivel de aplicación para la distribución, colaboración, para sistemas de información hipermedia. Definición de métodos:

- GET recupera información en forma de entidad.
- HEAD es idéntico que el método GET salvo que el servidor no debe devolver en la respuesta el cuerpo.
- POST su utilizara para solicitar que el servidor de origen acepte una nueva entidad.
- PUT solicita que la entidad adjunta sea considerada como una nueva versión de una entidad ya existen en el servidor de origen.
- DELETE que el servidor de origen elimine un recurso identificado.

1.1.3.3. RDF

Marco de Descripción de Recurso (Resource Description Framework) Miller, E. (1998) es una infraestructura que permite la codificación el intercambio y la reutilización de metadatos estructurados, e s una aplicación de XML que impone limitaciones estructurales necesarios para proporcionar métodos inequívocos de proporcionar semántica.

Cualquier expresión de RDF corresponde a una colección de tripletas, compuestas por sujeto, predicado y objeto, estos puedes ser graficado como un nodo y un diagrama de arco dirigido como se muestra en gráfico.

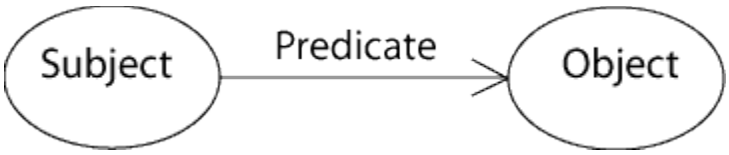

Figura 1: Representación gráfica RDF
Fuente: W3C rdf

1.1.3.4. SPARQL

RDF es un formato de datos para grafos dirigidos y etiquetados para representar la información en la Web. Esta especificación define la sintaxis y la semántica del lenguaje de consulta SPARQL para RDF. SPARQL se puede utilizar para expresar consultas que permiten interrogar diversas fuentes de datos, si los datos se almacenan de forma nativa como RDF o son definidos mediante vistas RDF a través de algún sistema middleware

1.1.4. Acerca de DBpedia

La DBpedia nos da la siguiente definición sobre si misma: “Es un esfuerzo de la comunidad crowd-sourced[3] para extraer información estructurada de Wikipedia y hacer esta información disponible en la web. DBpedia permite que hacer consultas sofisticadas contra Wikipedia.”[4]

La información estructurada que extrae esta publicada cumpliendo con las reglas de Datos Enlazados, es decir, para nombrar los datos utiliza HTTP URI, la se encuentra en formato RDF y por cual también se puede consultar media SPARQL y son datos

[3] Tareas realizada por comunidades en convocatorias abiertas
[4] Fuente: http://dbpedia.org/About

vinculados a otros datos. DBpedia pone a disposición su dataset para ser descargado en diferentes idiomas teniendo en cuenta de que el número de entidades puede cambiar de idioma a idioma puesto que no se trata de una traducción sino de una recopilación de información de Wikipedia la cual se encuentra más extendida en ingles que otros lenguas, esto lo podemos encontrar en la versión 3.9[5] de Dbpedia.

1.2. Procesamiento de Lenguaje Natural (PLN)

1.2.1. Introducción

El procesamiento de lenguaje natural dentro de la ingeniería lingüística comprende la rama se preocupara por entender el lenguaje humano, una tarea que para las personas e inclusive animales es tan natural y cotidiana se vuelve un reto al tratar de interpretarlo mediante procesos computacionales a fin de comprenderlo y poder replicarlo.

La dificultada de la construcción de una aplicación de la ingeniería lingüística variara de acuerdo a objetivo que se persiga, esto explicado por (García, 2005) en donde ejemplifica: “un sistema de generación de cartas personalizado no precisa ningún tratamiento de comprensión, o un sistema de identificación de la lengua (o un detector de errores ortográficos) no necesitan generar lenguaje humano. La mayoría de las aplicaciones incluyen, sin embargo, alguna forma más o menos precisa de comprensión. Así, un sistema de consulta en lenguaje humano a una base de datos precisa un nivel muy alto de comprensión de las expresiones del interlocutor humano para que la respuesta del sistema sea de utilidad. En cambio, en un sistema de traducción o de resumen automáticos se pueden lograr niveles de corrección muy notables con niveles de comprensión bajos. Es decir, no es preciso comprender totalmente una oración para ser capaz de traducirla correctamente.”

1.2.2. Part of Speech Tagger

También denominado POS-tagging Nos permite distinguir la función de una palabra en un determinada contexto mediante la asignación de una etiqueta predefinida. (Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P., 1992) Nos dicen que: “Una part of speech tagger es un sistema que usa el contexto para asignar parte de un discurso a una palabra”.

[5] Sitio de descarga de DBpedia: http://wiki.dbpedia.org/Downloads39

Es te etiquetado de palabra ya permite un primera desambiguación en cuanto a la función de la palabra en un sentencia o contexto. Asi podemos por ejemplo ver que la palabra dado que si bien es nombre en singular también puede ser una foram del verbo dar[5]
http://es.wikipedia.org/wiki/Ambig%C3%BCedad d.

Pero antes de poder etiquetar una palabra por su función es necesario una tokenización del texto que va ha analizar, que consiste en separarlo en palabras individuales reconociendo un token para palabra o carácter extraído.

1.2.3. Chunking

Text Chunking consiste en dividir un texto en frases de tal manera que palabras sintácticamente relacionadas sean miembros de las misma clase. Estas frases no se superponen es decir que una sola palabra puede ser miembro de un chunk. (Tjong Kim Sang, E. F., & Buchholz, S., 2000)

Este proceso es básico al momento de detectar entidades dentro de un texto, este proceso lo podemos observar en la figura 1 en donde la sentencia, *We saw the yellow dog*, está separada en cuadros en los más pequeños observamos etiquetas de POS Tag y las más grades al nivel de chunking. Una vez la frase ha pasado por el proceso de chunking podemos rescatar dos entidades dentro de la sentencia como *We* y *the yellow dog.*

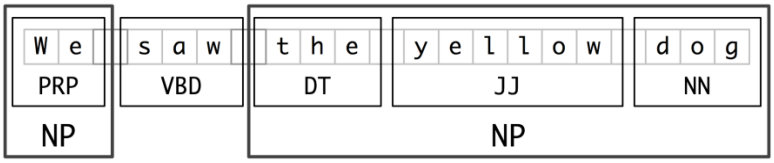

Figura 2: Ejemplo POS Tag y Chunking
Fuente: http://www.nltk.org/book/ch07.html

1.2.4. Desambiguación

Este es un fenómeno muy común conocido polisemia que se refiere a cuando una palabra tiene varios significados la desambiguación busca descifrar que significado es el que está activado en determinado contexto se denomina Word Sense Disambiguation (WSD), este problema propio del procesamiento de lenguaje natural (PLN). El descifrar estos distintos

significados para los seres humanos es muy común, lo resolvemos de forma cotidiana y pasa casi desapercibida.

1.2.4.1. Métodos basados en el conociendo.

El método Lesk (Lesk 1986) es el método basado en el diccionario. Se basa en la hipótesis de que las palabras usadas juntas en el texto están relacionadas entre sí y que la relación se puede observar en las definiciones de las palabras y sus sentidos. Dos (o más) palabras son desambiguadas encontrando el par de sentidos diccionario con la palabra mayor superposición en sus definiciones del diccionario.

Diccionarios externos:

WordNet:

Es una gran base de datos léxica de inglés. Sustantivos, verbos, adjetivos y adverbios se agrupan en conjuntos de sinónimos cognitivos (synsets), cada una expresando un concepto distinto. Synsets están vinculados entre sí por medio de las relaciones conceptuales semántico y léxico.

Corpus:

Corpus es una gran colección de textos. Se trata de un cuerpo de material escrito o hablado sobre la que se basa un análisis lingüístico. Es un conjunto muy amplio de ejemplos reales del uso de la lengua.

WordNet Domains:

WordNet Domains es una extensión de WordNet 1.6, donde cada synset tiene asociado uno o varios dominios (categorías semánticas). Estos dominios, están clasificados en una jerarquía con distintos niveles de especialización, cuanto más profundo es el nivel sobre el que nos movemos, mayor es el grado de especialización.

1.2.4.2. Métodos supervisados

Métodos supervisados se basan en la suposición de que el contexto puede proporcionar evidencia suficiente por sí sola para eliminar la ambigüedad de las palabras. Probablemente cada algoritmo de aprendizaje automático va se ha aplicado a WSD, incluyendo técnicas asociadas tales como la selección de características, la optimización de parámetros, y el aprendizaje conjunto.

1.2.4.3. Métodos semi-supervisados

El enfoque de bootstrapping comienza a partir de una pequeña cantidad de datos de semillas para cada palabra: cualquiera de ejemplos de entrenamiento manualmente etiquetados - o un pequeño número de reglas de decisión de éxito seguro. Las semillas se utilizan para entrenar un clasificador inicial, utilizando cualquier método supervisado. Este clasificador se utiliza en la parte sin etiqueta del corpus para extraer un conjunto de entrenamiento más grande, en el que sólo se incluyen las clasificaciones más seguras. El proceso se repite, cada nuevo clasificador siendo entrenado en un corpus de entrenamiento sucesivamente mayores, hasta que se consume todo el corpus, o hasta que se alcanza un número máximo dado de iteraciones.

1.2.4.4. Métodos sin supervisión

Aprendizaje no supervisado es el mayor desafío para los investigadores WSD. El supuesto subyacente es que los sentidos similares ocurren en contextos similares, y por lo tanto los sentidos puede ser inducido a partir del texto agrupando las ocurrencias de palabras usando alguna medida de similitud de contexto.

1.3. Rest Web Service

REST no es un protocolo, un formato de archivo, o un marco de desarrollo. Es un conjunto de restricciones de diseño, la hipermedia como el motor de estado de la aplicación. Es posible crear servicios web a partir de REST, teniendo en cuenta lo siguiente:

Utilizar los métodos del protocolo HTTP para la transferencia de los datos a través de World Wide Web, como son PUT, GET, POST y DELETE. Al utilizar protocolo HTTP decimos que cualquier sistema que se implemente con REST es un cliente-servidor y por la naturaleza mismo de HTTP que un sistema sin estado. Los recursos tienen sintaxis universal debido a las URIs que se utilizan para identificarlos.

1.3.1. Hipermedia

(Ruby, 2007)Es una estrategia que nos permite establecer una conexión entre los recursos y describe sus capacidades, La estrategia hipermedia tiene siempre el mismo objetivo. Hipermedia es una manera para que el servidor para decirle al cliente qué HTTP request el cliente podría querer hacer en el futuro. Es un menú, proporcionado por el

servidor, desde el que el cliente es libre de elegir. El servidor sabe lo que podría pasar, pero el cliente decide lo que realmente sucede.

1.3.2. Recursos y representaciones

(Leonard Richardson and Mike Amundsen, 2013)Rest denomina recursos a los datos estructurados que son objetos de las interacciones entre métodos de HTTP, y se dice que cualquier cosa que pueda ser almacenado de un computador puede ser un recurso, como documento electrónico, una fila de una base de datos o el resultado de un algoritmo

No solo las cosas almacenadas en un computador pueden ser llamados recursos también pueden ser recursos artículos tangibles como frutas por ejemplos, y es posible representarlo como recursos a través de la web como por ejemplo como un artículo en de venta o una imagen binaria depende de la aplicación así que por eso decimos sobre las representaciones que puede ser cualquier documento legible que contenga información acerca de un recurso.

1.3.3. URI y relación con los recursos

(Leonard Richardson and Mike Amundsen, 2013)Este tecnología de propia de la web que ya hemos revisa en capítulos anterior y que ahora recordamos teniendo en cuenta de que Rest trabaja sobre recursos o representaciones de los mismos u que la forma estandarizada para la identificación los recursos son las URIs, así como la relación existencial que tiene una sobre la otra.

1.4. Trabajos relacionados

CAPITULO 2: PROBLEMÁTICA

2.1. Estado actual

La información en texto plano, legible para humanos presenta limitaciones en cual a la ampliación de la información y

CAPITULO 3

3. Solución
    - 3.1. Propuesta
    - 3.2. Arquitectura
    - 3.3. Metodología
    - 3.4. Riesgos
    - 3.5. Módulos
        - 3.5.1.Procesado de Texto
        - 3.5.2.Extracción de Entidades y Keywords
        - 3.5.3.Desambiguación y Enlace
        - 3.5.4.Servicio Web
    - 3.6. Implementación

CAPITULO 4

4. Validación
    - 4.1. Comparación con servicios similares
    - 4.2. Pruebas

CAPITULO 5

5. Conclusiones y Recomendaciones

Bibliografía

Berners-Lee, T. (23 de Julio de 2006). Linked Data - Design Issues.

Cutting, D., Kupiec, J., Pedersen, J., & Sibun, P. (Marzo de 1992). A Practical Part of Speech Tagger.

Miller, E. (1998). Wiley Online Library. *Bulletin of the American Society for Information Science and Technology*, 15-19.

Ruby, L. R. (2007). *RESTful Web Services.*

Tjong Kim Sang, E. F., & Buchholz, S. (Septiembre de 2000). Introduction to the CoNLL-2000 shared task: Chunking.

W3C. (01 de 2005). *Uniform Resource Identifier (URI): Generic Syntax.* Recuperado el 22 de 02 de 2014, de January 2005