Sistema Semántico Simplificado para la creación de aplicaciones en la web semántica mediante palabras clave

Norberto Muñoz de la Teja
Escuela de diseño
ingeniería y aqrquitectura
ITESM Campus Ciudad de Mexico
México, D.F. 14380
Email: norberto.delateja@gmail.com

Jose Martin Molina Espinosa
Escuela de diseño
ingeniería y aqrquitectura
ITESM Campus Ciudad de Mexico
México, D.F. 14380
Email: jose.molina@itesm.mx

Rafael Lozano Espinosa
Escuela de diseño
ingeniería y aqrquitectura
ITESM Campus Ciudad de Mexico
México, D.F. 14380
Email: ralozano@itesm.mx

***Resumen*—El "World Wide Web Consortium" (W3C) ha proporcionado un conjunto de estándares sobre lo que se ha desarrollado la *web semántica*. Estos estándares son generalmente de alta complejidad lo cual limita su utilización a desarrolladores, por lo cual se propone bajar la barrera de entrada mediante el uso de una consulta por palabra y contexto clave con el Sistema Semántico Integral de Información (SISIF). El principal interés es la exposición de *información vinculada abierta* (LOD) para ser consultada en dispositivos que proporcionen dicha información.**

I. Introducción

La aparición de la web ha generado una explosión en la cantidad de datos a los que es posible tener acceso, esto ha hecho indispensable el uso de técnicas de representación de la información para organizar los datos disponible. Tim Barners Lee [1] propuso lo que se conocería como la *web semántica* (WS) la cual tiene por objetivo crear una red que este interconectada mediante significados algunos consideran la WS como la web de significado [2]. Su visión fue que los recursos web pudieran representar sistemas del mundo real y mediante un sintaxis uno o mas agentes pudieran evaluar estos sistemas y resolver problemas reales.

La web semántica como fue concebida por la W3C se compone de diferentes tecnologías como el marco de trabajo de descripción de recursos (RDF), RDFS el cual es el esquema de RDF, el lenguaje ontológico de web (OWL), el formato de intercambio de reglas (RIF), estos y otras sintaxis componen la web semántica.

Los estándares anteriormente mencionados están basados en diferentes técnicas de representación de la información por ejemplo RDF esta basado en redes semánticas o OWL esta basado en lógica de descripción. Adicionalmente la sintaxis del lenguaje resulta un poco mas complicada para algunos aunque ya existen herramientas que abstraen toda la serialización del modelo. Sin embargo el lenguaje de consulta de la web semántica es SPARQL el cual no esta en uso general aun. Mediante el uso de técnicas simples como es la consulta por palabra clave se planea reducir la barrera de entrada para desarrolladores principalmente desarrolladores moviles que puedan dar informaciñon de contexto.

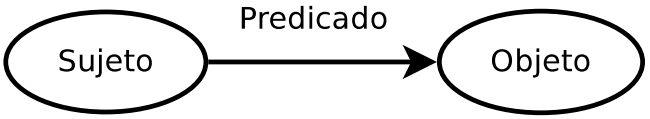

Figura 1. La tripleta con sujeto predicado y objeto

En la siguiente sección se tratará el estado del arte de la web semántica y LOD. En la sección tres se comentarán trabajos relacionados. En la sección cuatro se expondrá la aportación antes descrita.

II. Antecedentes

II-A. Web Semántica

La web semántica utiliza el lenguaje RDF como sintaxis primaría para estructurar la web en una red semántica, adicionalmente tiene un esquema el cual le permite dar una coherencia semántica a la red este es el vocabulario RDFS. El fin de la red es poder modelar sistemas reales y por tanto se apoya en el concepto de ontología, un concepto prestado de filosofía, la cual es definida en [6] como "la especificación formal y explicita de una conceptualización compartida". Mediante el uso de ontologías se extraen las características principales de un sistema para describirlo de una manera simple, para tal objetivo se creo OWL.

II-B. RDF, RDFS, OWL, SPARQL

El marco de trabajo de descripción de recursos o RDF (por sus siglas en inglés) es la sintaxis de la web semántica el cual permite describir conceptos y relaciones en la web, esto se logra mediante enunciados RDF [4] (tripleta RDF). El enunciado esta compuesta por tres sujeto, propiedad y objeto. El sujeto y objeto son conceptos mientras la propiedad es una relación binaria.

RDF cuenta con tres elementos la IRI, literales y el nodo en blanco. Al juntar tres de estos elementos en un enunciado RDF nos permite visualizar el contenido en forma gráfica. En

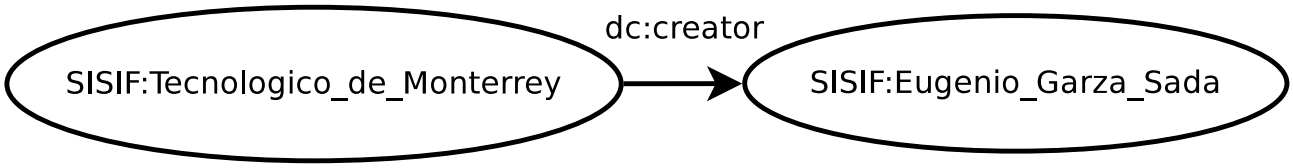

Figura 2. Enunciado RDF, los IRI están escritos en forma relativa siendo SISIF el nombre de la base, los nodos representan conceptos y la arista es una relación.

la figura 1 se puede apreciar un enunciado RDF en formato de grafo.

sujeto ⇒ *SISIF:*`Tecnológico_de_Monterrey`
propiedad ⇒ *dc:creator*
objeto ⇒ *SISIF:*`Eugenio_Garza_Sada`

En la figura 2 se puede formar el enunciado "El Tecnológico de Monterrey fue creado por Eugenio Garza Sada", con la tripleta en el formato de IRI relativa.
RDFS va un paso mas allá y nos permite especificar metadatos sobre los objetos. Esto permite crear jerarquias en el sistema lo cual provee un primer paso hacia la estructuración semántica. Esta clasificación de los objetos nos permite aplicar restricciones e inferir cosas de los mismos [2].

Mientras RDFS permite aplicar inferencias a nivel de miembros de una clase, OWL permite hacerlo entre miembros de la misma clase. Al usarse en conjunto con las propiedades de RDF pueden modelarse relaciones útiles [2].
El lenguaje de consulta de la web semántica es SPARQL. SPARQL tiene una sintaxis parecida a SQL pero utiliza tripletas para expresar condiciones y formatos de salida.

II-C. Datos vinculados abiertos

Los datos vinculados abiertos o "Linked Open Data" (LOD) se refiere a una serie de buenas practicas para publicar y conectar información estructurada en la web utilizando estándares internacionales publicados por la W3C [4]. Se creo un sistema de cinco estrellas para clasificar que tan bien esta vinculada la información :

1. UNA ESTRELLA: los datos se encuentran disponibles en cualquier formato(por ejemplo, imágenes digitales).
2. DOS ESTRELLAS: los datos están disponibles de forma que pueden ser interpretados por una maquina como información estructurada(por ejemplo excel).
3. TRES ESTRELLAS: Los datos están en un formato que no es propietario (por ejemplo csv).
4. CUATRO ESTRELLAS: Los datos están publicados utilizando los estándares del W3C.
5. CINCO ESTRELLAS: Se aplica lo anterior y se tiene vinculada esta información con otros recursos.

La naturaleza de la web es no estar fuertemente conectada [7] con los datos vinculados podemos crear más conexiones y ligar la web. Los datos vinculados nos permiten hacer uso de vocabularios y ontologías ya establecidas en la web y utilizar sus vocabularios y sus clasificaciones [11], algunos ejemplos se pueden ver en la cuadro I.

Cuadro I
EJEMPLO DE VOCABULARIOS RDF

DBpedia	**Una ontología de la información de Wikipedia.**
FAOF	Vocabulario para interacciones en redes sociales.
vCard	Vocabulario para tarjetas de presentación.
Geo	Vocabuario para geo-localización .

III. TRABAJOS RELACIONADO

El trabajo de TODE [10] muestra como se pueden utilizar ontologías de uso general como es el caso de "*Suggested Upper Merged Ontology*" (SUMO), la cual sirve como un paso intermedio para la descripción de la ontología propia para clasificar directorios web, el caso de estudio es organizar directorios web de Yahoo.

En [5] se propone una ontología para organizar el contenido de los datos de Wikipedia. Esto con el fin de apoyar el desarrollo de aplicaciones de la web semántica. En este trabajo se crea una arquitectura para consumir archivos en lote, servicios por OAI-PMH y archivos de actualizaciones de Wikipedia. Dos métodos se proponen para la extracción de datos el genérico y otro basado en un mapeo.

Sindice.com [9] es un sitio que provee búsqueda en la web semántica por cuatro medios por URL, por URI, por relación RDF y por texto libre, Sindice.com utiliza índices invertidos para buscar en los datos.

Finalmente en [11] se propone que a partir de construcciones básicas de RDFS como es *rdfs:type* el cual ayuda a crear particiones de ontologías como DBpedia para buscar palabras clave en las bases de datos de estas ontologías. De esta manera optimizando el tiempo de respuesta para la búsqueda de datos por palabra clave en datos RDF masivos.

IV. PROPUESTA

La simplificación de la web semántica conlleva muchos trabajos como es el caso de [9], [5], [11] y se pueden resumir a dos campos uno de clarificación de los datos (DBpedia) y otro de extracción de información. El objetivo final es ponmer un servicio simplificado para que aplicaciones puedan nahacer uso de la gran cantidad de datos abiertos que se presenta en LOD [9].

La propuesta doctoral es exponer una plataforma optimizada para la extracción de datos para aplicaciones contextuales. Primeramente podemos ver que a diferencia de DBpedia el objetivo principal no es la creación de una ontología sino la utilización de las mismas. Adicionalmente a diferencia

de Sindice nuestro objetivo principal son las aplicaciones contextuales para lo cual como en [11] se pueden utilizar ciertas técnicas de optimización como lo hizo [11] de dominio para mejorar la eficiencia de la búsqueda.

Primeramente se necesita soportar el sistema para tal motivo se ha utilizado el esquema propuesto por [8] el cual es la arquitectura esencial para el procesamiento y optimización de una ontología. Adoptamos el proceso de extracción de datos que puede apreciarse en la figura 3 con el objetivo de realizar el proceso de jerarquización y exposición de resultados. Adicionalmente tendremos un sistema completo de la web semántica mediante la utilización del software libre Apache Jena[1].

La extracción es un proceso en el cual se consulta fuentes externas y se agregan resultados de indexado al sistema(figura IV). El análisis de los datos se realiza después del almacenamiento en una "*triplestore*" la cual almacena enunciados RDF [13]. Existen diferentes tipos de bases de datos para enunciados RDF en este trabajo se utilizara "TDB". El desempeño para la búsqueda por palabra clave en TDB resulta muy lento ya que la base no esta optimizada para este tipo de consultas. Una consulta no acotada puede ser capaz de traer todas las tripletas de un subconjunto y aplicar expresiones regulares sobre cada elemento del subconjunto lo cual ocasiona que la consulta por medio del motor de SPARQL sea demasiado lenta. Es deseable que la consulta por palabra clave sea rápida por tal motivo se planea crear un índice(figura 6)

El proceso de extracción inicia mediante la identificación de las fuentes de información en LOD mediante el uso de consultas SPARQL se pueden consultar diferentes repositorios disponibles en Internet. Los resultados se indexaran y en caso de que se descubra información adicional se agregara mediante la aplicación de inferencias sobre los datos. Generalmente los puntos SPARQL no contiene toda la información disponible en ya que los conjuntos de datos suelen ser muy grandes. Mediante la indexación a este contenido podemos llegar a tener mayor alcance que si tan solo se consultan los sitios SPARQL normales y esto se logra mediante la consulta de archivos en lotes de RDF.

Posteriormente se tiene un sistema donde se procesan las inferencias de sistemas ya modelados o nuevas ontologías o mediante algoritmos de aprendizaje maquina se descubre nuevo conocimiento y se indexa. El trabajo no se centrara en la minería de datos sino en la extracción de información por lo cual realizaremos un sistema como en [12], [11] creando un índice invertido con elementos de búsqueda por palabra y contexto clave en un índice el cual se generara por tres fuentes de información las inferencias que se obtengan de la información obtenida o un las misma gráfica RDF. En esta parte se pueden añadir ontologías mediante OWL con el fin de crear nuevas vinculaciones entre los datos ya existentes o introducir nuevos vocabularios u ontologías.

Una vez que se obtiene la información se utiliza Apache

[1]https://jena.apache.org/

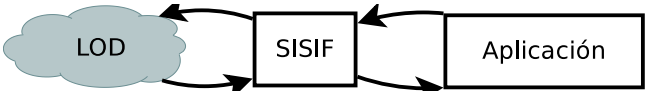

Figura 3. Esquema general de interacción con el SISIF.

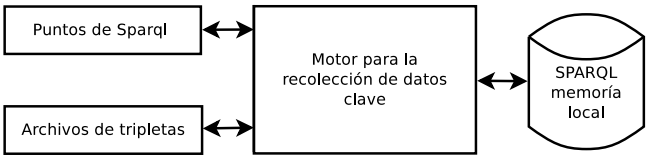

Figura 4. Consulta a la red semántica y creación de índice.

Lucene [2] para construir el indice, Lucene es un proyecto que permite construir indices invertidos, su complemento web es Apache Solr, el cual expone las características de lucen a través de servicios web. Debido a que el índice no estará distribuido es mejor utilizar Lucene en lugar de Solr sin embargo si el indice fuera a estar distribuido la utilización de Solr es conveniente. Con Lucene se creara un índice invertido, para esto se utilizaran las IRI, las etiquetas rdfs *rdfs:label*. Adicionalmente utilizaremos el vocabulario Geo para la latitud y la longitud los predicados son *geo:lat* y *geo:lon* y para todos aquellos que tengan un resumen de los contenidos del recurso.

El proceso de petición (figura 5) se ejecuta mediante una petición GET que manda los contextos claves así como una serie de parámetros para limitar los resultado el cual consulta el índice. A partir de los resultados de una consulta previa se determinan los resultados de no ser suficientes o satisfactorios se utiliza el índice ya que la búsqueda por SPARQL suele ser muy lenta y no tiene ninguna rubrica para implementar

[2]http://lucene.apache.org/

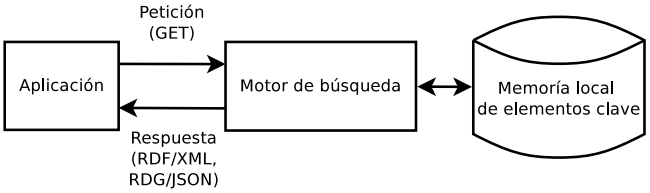

Figura 5. Petición de aplicación.

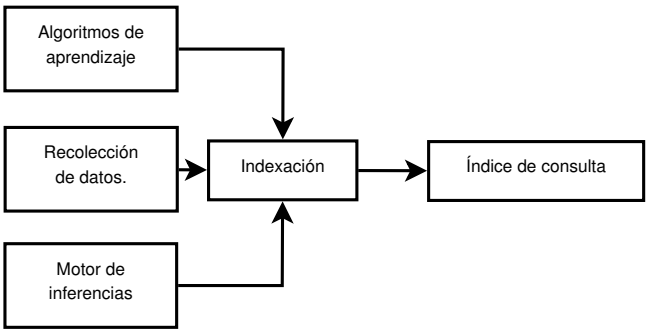

Figura 6. Entradas para la creación del índice.

relaciones entre documentos. Posteriormente se construye una serie de consultas SPARQL que obtiene los resultados en RDF/XML ó RDF/JSON.

Mientras que Sindice provee un API similar el objetivo es optimizarlo para información de contexto. Sindice provee un API el cual se le puede entregar una consulta por palabra clave. Nuestro sistema aportara a los resultados una función que permita tomar en cuenta la información de contexto así como la información de por palabra clave implementando una búsqueda sesgada por estos mismos elementos. El API de Sindice permite especificar ciertas condiciones para utilizar un vocabulario especifico pero esto no puede estar optimizado para el contexto por la naturaleza abierta del API. El SISIF permite exponer este servicio con campos específicos de palabra clave, así como locación clave y tiempo clave.

EL SISIF ofrece un servicio con los siguientes parámetros:

1. qt: el cual es el elemento qu determina la palabra o palabras clave
2. latlon: Dos números flotantes que representen la latitud y la longitud, este elemento es opcional y se utiliza para el contexto.
3. time: Tiempo de la consulta o época relacionada, al igual que latlon este campo es opcional pero utilizado para el contexto.
4. l: tamaño de la pagina para lo n niveles.
5. n: el nivel de profundidad que se busca a partir de los elementos directamente conectados al termino de la búsqueda.
6. lang: El idioma en el que se prefieren los resultados, este campo es opcional.
7. p: La pagina de resultados para la consulta actual.

Este formato de API ayudara a que la información semántica con información de contexto sea mas utilizada. El resultado puede ser utilizado por una aplicación mediante el proceso inverso de publicación de los datos con APIs como Jena.

V. Conclusiones

El uso de tecnologías semánticas nos permite tener representaciones de información mas útiles para las personas sin embargo su complejidad inherente resulta una barrera para su implementación. Mediante recursos proyectos que han concentrado la información y la han puesto disponible es posible aumentar el contexto de nuestra aplicaciones mediante le uso de ontologías de dominio o vocabularios como Geo, DBpedia, SKOS, etc. Tecnologías como SPARQL nos pueden ayudar a consultar u crear reglas y crear una vinculación entre ellas para crear una base de conocimiento y automatizar la creación de argumentos respecto a un sistema. Sin embargo no están diseñadas su utilización por el usuario común. Con una consulta por palabra clave y contexto clave el usuario puede tomar ventaja de esta información disponible en una serie de plataformas como pueden ser los dispositivos móviles o el mismo explorador. Aunque el sistema ya cuenta con la base para realizar las operaciones básicas de un sistema de web semántica aun falta hacer la optimización de los índices para que los tiempos de respuesta sean mas eficientes e incluso considerar el uso de un índice HYB [14] para búsqueda interactiva.

Reconocimientos

Los autores desean agradecer al Consejo Nacional de Ciencia y Tecnología y al Instituto Tecnológico y de Estudios Superiores de Monterrey por el apoyo prestado para la realización de esta investigación.

Referencias

[1] T. Barners-Lee, J. Hendler y O. Lassila, *The semantic web*. Scientific American, vol.284, pp. 34-43, May. 2001.

[2] D. Allemang y J. Hendler, *The semantic web for the working ontologist: Efective modeling in rdfs and owl*, segunda edición, Waltham, MA:Morgan Kaufman, pp. 27-258.

[3] B. DuCharme, *Learning sparql querying and updating with sparql 1.1*,segunda edición, Sebastopol, CA: O'Reilly 2013, pp. 19-183.

[4] D. Wood, M. Zaidman y L Ruth, *Linked data : Structured data on the web*, primera edición, Shleter Island, NY: Manning Publications Co., 2014, pp. 3-76.

[5] C. Bizer, J. Lehmann., G. Kabilarov, S. Auer, C. Becker, R. Cyganiak y S. Hellman, *Dbpedia - a crystalization point for the web of data.* , Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7(3), pp. 154-165, Sep.2009.

[6] N. Gaurino,D. Oberle, y S. Staab, *What is an ontology?*(S. Staab y R. Studer, Eds.),Springer, 2009, pp 1-17.

[7] A. Rajaraman y J. D. Ullman(2011). *Mining of massive datasets.* Cambridge, UK:Cambridge University Press,2011, pp. 151-152.

[8] H. O. Nigro, S. E. González Cisaro y D.H. Xodo, D. H. *Data minig with ontologies: Implementations, findings and frameworks*, Hershey, PA: Information Science Reference, 2008. pp. xii-xv.

[9] E. Oren, D. Renaud, M. Catasta, Cyganiak, R, Stenzhorn, H y G. Tumarrello.*Sindice.com: a document-oriented lookup index for open linked data*, JIJMSO. vol. 3,no. 1, pp. 37-52, Ene. 2008.

[10] S. Stamou, A. Ntoulas y D. Christodoulakis.*Tode: An ontology-based model for dynamic population of web directories*, (H. O. Nigro, S. E. González Cisaro, y D. H. Xodo, Eds.). Hershey, PA:Information Science Reference. pp. 1-17.

[11] W. Le, F. Li, A. Kementsietsidis y S. Duan.*Scalable Keyword Search on lare RDF Data*, IEEE Transactions on Knowledge and Data Engineering. 2013.

[12] T. Tran, H. Wang, S.Rudolph, P. Cimiano. *op-k Exploration of Query Candidates for Efficient Keyword Search on Graph-Shaped (RDF) Data*, en IEEE International Conference on Data Engineering, 2009, pp 405-416.

[13] C. Basca y A. Bernstein. *Querying a messy Web of Data with Avalanche* University of Zurich, Zurich, CH, Rep. Tec.Zurich, CH, Rep. Tec.IFI-2013.03, Nov. 2013.

[14] H. Bast, I. Weber.*Type less, find more: fast autocompletion search with a succinct index*, SIGIR,Seattle:WA, pp. 364-371, 2006.