UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
ICI-309 – SEMINARIO DE MEMORIA

Análisis de centralidad a base de datos RDF. Caso de uso: Datos biológicos.

Hernán Vargas Leighton – 201073009-3
hernan.vargas@alumnos.usm.cl

18 de abril de 2016

I. DEFINICIÓN DEL PROBLEMA

La web como la conocemos está hecha por y para las personas, por lo que las maquinas actualmente deben emular el comportamiento humano para acceder a gran cantidad de la información publicada en Internet. Esta problemática intenta ser solucionada por la web semántica, que entre sus objetivos tiene crear tecnologías para la publicación y el análisis de los datos en la web por parte de las computadoras, para ello es necesario crear enlaces y relaciones entre los datos existentes y ampliar los mismos a todo ámbito de conocimiento.

En este contexto nace Bio2RDF, proyecto que anexa la información de 35 bases de datos biológicas abiertas a todo publico y crea enlaces entre ellas generando un total de 11.000 millones de datos enlazados. Para lograrlo utiliza las tecnologías de la web semántica.

Pero de la gran cantidad de datos que componen el proyecto Bio2RDF no todos tienen la misma importancia. Actualmente no existen estudios que determinen cuales son los datos más buscados por parte de los usuarios, es decir, no se conoce cual es la información más relevante de la base de datos y por ello, no es posible hacer una optimización teniendo este parámetro en cuenta.

Para determinar la importancia de un dato existen muchas métricas, entre ellas, el análisis de centralidad que es ideal para este estudio pues podemos modelar las consultas hechas al proyecto Bio2RDF como un grafo conectado o una red. La centralidad es uno de los conceptos más estudiados en el análisis de redes, actualmente se utiliza ampliamente en las redes sociales para determinar las personas más influyentes.

Así, un análisis de centralidad a los datos consultados por los usuarios del proyecto Bio2RDF revelará cual es la información más importante para los mismos y con ello se podrá mejorar el soporte que esta tiene.

II. DISCUSIÓN BIBLIOGRÁFICA

Los datos a analizar en esta memoria serán extraídos del proyecto Bio2RDF[1][2], la red de datos enlazados más grande de las ciencias naturales. Esta iniciativa se encuentra en el marco de la web semántica, en un esfuerzo para combinar la información disponible de diferentes bases de datos abiertas al público general. Así, se utilizará el conjunto de tecnologías, estándares y recomendaciones publicadas por la W3C (*World Wide Web Consortium*) para la creación de la web de datos enlazados.

La web semántica es un conjunto de actividades propuestas por el W3C con el fin de publicar datos en la web que sean procesables tanto por los usuarios como por las máquinas. Se basa en la idea de añadir metadatos semánticos y ontológicos que describan el contenido y las relaciones de los datos publicados. Es posible rastrear los orígenes de esta idea hasta una propuesta temprana de la *world wide web* en 1986 [3].

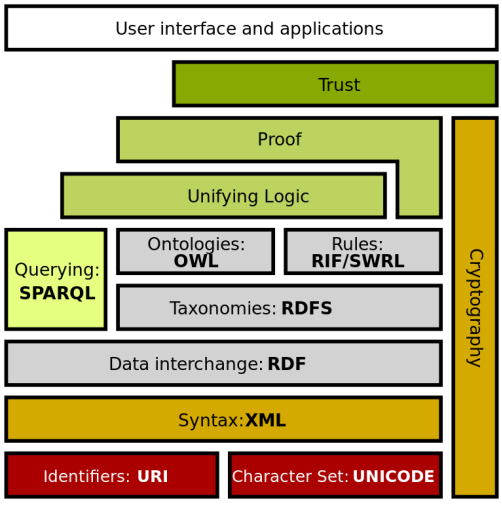

Figura 1: Pila de tecnologías de la web semántica[4].

En la figura 1 se pueden ver las tecnologías involucradas en la creación y publicación de datos semánticos en la web. A continuación se hará una breve revisión de dichas tecnologías.

II-A. Tecnologías del hipertexto

El componente básico para identificar inequívocamente un recurso en la web se llama **URI**[5] (*Uniform Resource Identifier*), una cadena de caracteres con sintaxis fuertemente estructurada. Cabe destacar que una URL es una URI que apunta a un recurso físico en la web, mientras que una URI no necesariamente debe apuntar a una localización que exista realmente.

XML (*Extensible Markup Language*) es un lenguaje de marcas desarrollado por el W3C para almacenar datos de manera legible tanto por personas como por máquinas. El diseño de XML busca la simplicidad, generalidad y usabilidad a través de internet[6] lo que lo hace adecuado para representar la información semántica de los datos. Su extensión XSD (*XML Schema Definition*) especifica formalmente la estructura y restricciones de los contenidos de un fichero XML de manera precisa agregando tipos de datos y sus restricciones[7].

II-B. Tecnologías de la web semántica

RDF (*Resource Description Framework*) es una familia de especificaciones del W3C diseñada como un modelo de datos para metadatos. Fue adoptado como una recomendación del W3C en 1999, mientras que la especificación 1.0 fue publicada el 2004 y la 1.1 el 2014[8].

El modelo de datos RDF se basa en la idea de hacer declaraciones sobre recursos web (URIs) en forma de expresiones $\langle sujeto, predicado, objeto \rangle$ que son llamados triples RDF. El $sujeto$ indica el recurso mientras que el $predicado$ denota la relación con el $objeto$.

Llamaremos vocabulario a la definición de conceptos y relaciones (términos) utilizados para describir y representar un área de conocimiento. Otro concepto a tener en cuenta son las ontologías, aunque no existe una clara división entre éstas y los vocabularios, generalmente se les considera más complejas y formales. En la web semántica una colección de triples RDF puede denotar un vocabulario o ontología.

Un conjunto de triples RDF será representado naturalmente por un grafo dirigido. Esta característica faculta la tecnología para ser parte fundamental de la web semántica pues permite relacionar información de diferentes fuentes sin mayor problema y representarla en un esquema fácilmente identificable.

RDF es un modelo abstracto con varios formatos de serialización, por lo que la codificación de un triple varía dependiendo el tipo de archivo en el que se guarde. En esta memoria se trabajará con triples codificados en RDF/XML pues fue fue la primera codificación estándar para serializar RDF (en un archivo XML).

El vocabulario incluido en la especificación RDF es muy básico y por ello fue extendido a *RDF Schema*, por lo que la gran mayoría de los *dataset* RDF contienen ambos vocabularios.

RDFS (*Resource Description Framework Schema*) extiende RDF proveyendo un set de clases y propiedades que mejoran la creación de modelos como son: `Class` para declarar clases, `subClassOf` para denotar herencia, `range` y `domain` para el rango y dominio de cierta propiedad (`rdf:Property`), entre otras.

RDF fue presentado en 1998 e introducido finalmente como recomendación del W3C el 2004[8]. La especificación completa del vocabulario puede encontrarse en [9].

OWL (*Web Ontology Language*) es una familia de lenguajes para la creación de ontologías complejas. Agrega lógica computacional para que las relaciones hechas con este lenguaje puedan ser procesadas con el fin de verificar la consistencia de la información o generar información implicita.

La versión actual de OWL se conoce como "OWL 2" y fue publicada el 2009 como una revisión y extensión de la versión inicial publicada el 2004[8].

OWL 2 tiene tres perfiles dependiendo de la función que cumple.

1. **OWL 2 EL**: Es el fragmento del lenguaje decidible en tiempo polinomial, diseñado para trabajar con grandes volúmenes de propiedades y clases.
2. **OWL 2 QL**: Fue diseñado para facilitar el acceso a *datasets* con un gran número de instancias donde las consultas son más importantes que el razonamiento.
3. **OWL 2 RL**: Está optimizado para el análisis de reglas lógicas, en aplicaciones que requieren un razonamiento escalable sin perder la expresividad del lenguaje.

Se pueden ver las características completas de los perfiles de OWL 2 en [10].

SPARQL (*SPARQL Protocol and RDF Query Language*) es un lenguaje estandarizado para consultar grafos RDF, se constituyó como una recomendación oficial por la W3C en el 2008[8]. La versión actual de SPARQL es la 1.1[11].

SPARQL provee un set completo de operaciones analíticas para sus consultas definidas directamente en la especificación. Particularmente provee 4 formas de consultas:

- **`SELECT`**: Retorna valores en forma de tabla.
- **`CONSTRUCT`**: Retorna valores en forma de triple RDF.
- **`ASK`**: Retorna un resultado binario a la consulta (`True/False`).
- **`DESCRIBE`**: Retorna un grafo RDF con contenido que el administrador del *endpoint* SPARQL considere información útil.

A excepción de `DESCRIBE`, las demás consultas necesitan un bloque `WHERE` para determinar las restricciones de búsqueda.

Una descripción completa del lenguaje puede encontrarse en [12] y en [11].

II-C. Centralidad

La centralidad en un grafo se refiere a una medida de importancia relativa de un nodo dentro de éste.[13]

Conocer la centralidad de un nodo ayuda a determinar, por ejemplo, el impacto de una persona en una red social, la importancia de una carretera en una red urbana, los componentes esenciales de una red de computadoras, entre otros.

El concepto fue introducido a fines de los años 1940 por Alex Bavelas[14] Es uno de los conceptos más estudiados en el análisis de redes y desde finales de los años 70 se ha ampliado su uso a las redes sociales[15]

Desde sus inicios se han propuesto diversas medidas para determinar la centralidad de un nodo. Las siguientes son las más utilizadas en análisis de redes:

- La centralidad de grado (*degree centrality*)
- La cercanía (*closeness*)
- La intermediación (*betweenness*)
- La centralidad de vector propio (*eigenvector centrality*).

El conocido algoritmo PageRank de Google para determinar la importancia de una página web es una modificación de un algoritmo de centralidad de vector propio.

III. Objetivos

III-A. Objetivo General

Generar estadísticas de centralidad sobre el subconjunto de datos consultados por los usuarios al proyecto Bio2RDF.

III-B. Objetivos Específicos

Para el logro del objetivo general se plantean los siguientes objetivos específicos:

1. Generar un subgrafo del proyecto Bio2RDF a través del análisis de las consultas SPARQL hechas al servidor por parte de los usuarios.
2. Analizar el grafo generado por medio de métricas de centralidad para grafos.
3. Comparar los resultados del estudio con el proyecto Bio2RDF.

Actividad	Semanas
Elaboración de Estado del Arte	8
Extracción de consultas SPARQL de los usuarios	1
Modificación de consultas para crear triples RDF	1
Creación del grafo consultado por los usuarios	2
Implementación de algoritmos para calcular métricas de centralidad a grafos RDF	2
Calculo de centralidad sobre el grafo de consultas	2
Análisis de resultados	2
Redacción del Informe Final	4
Revisión y Correcciones	2
Total	24

IV. Plan de trabajo

Referencias

[1] A. Callahan, J. Cruz-Toledo, P. Ansell, and M. Dumontier, "Bio2rdf release 2: Improved coverage, interoperability and provenance of life science linked data," in *The semantic web: semantics and big data*, pp. 200–212, Springer, 2013.

[2] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, "Bio2rdf: towards a mashup to build bioinformatics knowledge systems," *Journal of biomedical informatics*, vol. 41, no. 5, pp. 706–716, 2008.

[3] T. Berners-Lee, "The original proposal of the www," *URL http://www.w3.org/History/1989/proposal.html*, 1989. [Revisado el 08/01/2016].

[4] Marobi1, "Semantic web stack." URL https://en.wikipedia.org/wiki/File:Semantic_web_stack.svg, 2014. [Revisado el 08/01/2016].

[5] T. Berners-Lee, R. Fielding, and L. Masinter, "Uniform resource identifier (uri): Generic syntax," *URL http://www.rfc-editor.org/rfc/rfc3986.txt*, 2004.

[6] J. Paoli, C. Sperberg-McQueen, F. Yergeau, E. Maler, and T. Bray, "Extensible markup language (xml) 1.0," *W3C recommendation REC-xml-20040204*, 2004.

[7] P. Biron, A. Malhotra, W. W. W. Consortium, *et al.*, "Xml schema part 2: Datatypes," *W3C Recommendation REC-xmlschema-2-20041028*, 2004.

[8] Bikakis, Tsinaraki, Gioldasis, Stavrakantonakis, and Christodoulakis, "The xml and semantic web worlds: Technologies, interoperability and integration. a survey of the state of the art," in *Semantic Hyper/Multi-media Adaptation: Schemes and Applications*, Springer, 2013.

[9] D. Brickley and R. V. Guha, "Rdf schema 1.1," *W3C Recommendation PER-rdf-schema-20140109*, 2014.

[10] B. Motik, B. C. Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz, "Owl 2 web ontology language profiles," *W3C Recommendation REC-owl2-profiles-20091027*, 2009.

[11] T. W. S. W. Group *et al.*, "Sparql 1.1 overview," *W3C Recommendation REC-sparql11-overview-20130321*, 2013.

[12] E. Prud'hommeaux and A. Seaborne, "Sparql query language for rdf," 2008.

[13] S. P. Borgatti, "Centrality and network flow," *Social networks*, vol. 27, no. 1, pp. 55–71, 2005.

[14] A. Bavelas, "A mathematical model for group structures," *Human organization*, vol. 7, no. 3, pp. 16–30, 1948.

[15] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

Tiempo SCT:

Actividad	Horas
Investigación[1]	26:00
Plan de trabajo	01:00
Creación de informe	04:30
Total	27:30

[1]Incluida parte de la investigación hecha para otros entregables.