Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

Métricas para la generación automática de facetas desde grafos RDF

Carlos H. Cordoví García[1], Yusniel Hidalgo Delgado[2]

[1]Departamento de Ciencias Básicas, Facultad 3, Universidad de las Ciencias Informáticas, La Habana, Cuba

[2]Departamento de Técnicas de Programación, Facultad 3, Universidad de las Ciencias Informáticas, La Habana, Cuba

RESUMEN

Los datos abiertos enlazados han incrementado el volumen de información disponible en la Web. Cada día, las organizaciones y empresas publican datos valiosos de su gestión en el Formato de Descripción del Recurso conocido como RDF. Sin embargo, este formato solo es comprensible por los ordenadores, no por las personas. Por tanto, es necesario contar con herramientas fáciles de usar e intuitivas que permitan explorar y descubrir el conocimiento implícito en estos datos. Uno de los paradigmas utilizados actualmente por estas herramientas es el paradigma iterativo exploratorio. Este último establece la navegación a través de dimensiones conceptuales ortogonales llamadas facetas, estas están estrechamente vinculadas a las propiedades del conjunto de datos RDF. Sin embargo, establecer cuán representativo es un concepto entre otros conceptos, es una necesidad en aras de determinar cuáles son imprescindibles para la descripción del conjunto de datos. La presente investigación en progreso propone en uso de 3 métricas determinísticas para la generación automática de facetas desde grafos RDF. Así como para determinar su representatividad dentro del conjunto de datos.

Palabras claves: navegación facetada, datos enlazados, facetas, RDF, métricas

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

Metrics for automatic generation of facets from RDF graphs

ABSTRACT

The linked open data has increased the amount of information available on the web. Each day, organizations and enterprises are publishing important data of their management on Resource Description Format well known as RDF. However, only computers are able to understand this format, but people cannot do it easily. Therefore, the use of user interfaces able to browse and to discover knowledge in RDF data are needed. A search paradigm used by these kind of applications is Interactive/Exploratory paradigm, its establish to browse through orthogonal concepts called facets, they have relationship with RDF dataset resources and properties. But, how representative is a concept above another concept? It is a need how to stablish a representative order among concepts in a RDF dataset. This paper proposes the use of three deterministic metrics for generating and raking representative concepts in a dataset (facets).

Keywords: faceted search, linked data, facets, RDF, metrics

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

1. INTRODUCCIÓN

El surgimiento de la Web representó un salto decisivo en el desarrollo tecnológico de la humanidad. Esta ha reformado significativamente las vías de acceso a la información, haciendo posible la transacción de datos alrededor del mundo y brindando disponibilidad de un gran cúmulo de recursos. Sin embargo, esta presenta algunas limitaciones que obstaculizan su evolución básicamente referidas al formato, integración y recuperación de la información(1).

Para superar estas limitaciones, se hace necesario un estadío superior de la Web, en la cual no solo interactúen los humanos con las máquinas, sino que estas últimas sean capaces de interactuar y comunicarse entre sí realizando diversas tareas. Con esta finalidad en 2001 Tim Berners-Lee enunció el concepto de Web Semántica: “La Web Semántica no pretende sustituir la Web actual, sino que es una extensión de la misma en la que la información tiene un significado bien definido, posibilitando a los humanos y las computadoras trabajar en cooperación”(2).

La necesidad de la transición hacia una fase evolutiva de la Web (Web 3.0), donde se inserta la llamada Web Semántica, propició que en el año 2006, Tim Berners -Lee enunciara el concepto de datos enlazados: ―Los datos enlazados se refieren a un conjunto de buenas prácticas para la publicación y enlazado de datos estructurados en la Web (3). A partir de este momento comienza a implementarse la concepción de publicar datos estructurados siguiendo los principios de los datos enlazados.

La aplicación de estos principios a la publicación de contenido en la Web, ha permitido una considerable mejora en el procesamiento y gestión de la información tanto para humanos como para las máquinas (agentes software). Sin embargo, aún existe una barrera técnica en relación con el uso de datos enlazados para los usuarios que no están familiarizados con las tecnologías de la Web Semántica, principalmente por la necesidad de tener que realizar complejas consultas semánticas sobre grafos RDF (ver definición de grafo RDF en la sección 2).

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

Por tanto, para que la Web Semántica pueda ser explotada tanto por usuarios familiarizados con sus tecnologías como por aquellos que no conocen las mismas es necesario que existan herramientas y sistemas informáticos capaces de abstraer todo el proceso de consumo de estos datos. Es decir, sistemas que tomen el dato primario (alguna serialización de formato RDF (3)) y que lo muestren de manera amena e intuitiva al usuario final, utilizando para ello los paradigmas y componentes de la arquitectura de la información (4).

Las aplicaciones para el consumo de datos enlazados utilizadas en la actualidad se clasifican en dos grandes grupos: aplicaciones genéricas y aplicaciones para un dominio específico. Entre las primeras encontramos los navegadores de datos enlazados (5) y los motores de búsqueda de datos enlazados (3). Por otro lado, entre las aplicaciones dominio específico

encontramos: Integradores de datos enlazados (del inglés *linked data mashups*) (6) y otras aplicaciones de dominio especifico como: DBLP[1] y Rhizomer[2].

Esta última clasificación (otras aplicaciones de dominio específico), se caracteriza por utilizar paradigmas de búsqueda. En la concepción tradicional de la Web (Web de los Documentos) se utilizan paradigmas de búsqueda, cuya aplicación ha sido extendida al contexto de los datos enlazados. Los paradigmas de búsqueda existentes se clasifican en tres categorías: Palabras Clave, Iterativo/Exploratorio, Lenguaje Natural (7). EL objetivo de este trabajo es abordar específicamente el paradigma Iterativo/exploratorio en su variante: navegación facetada.

La navegación facetada constituye una técnica para la exploración de datos estructurados en la teoría de la faceta, esta técnica permite explorar conjuntos de datos a través de dimensiones conceptuales ortogonales, también llamadas facetas (8), las

[1] http://dblp.l3s.de/

[2] http://rhizomik.net/rhizomer/

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

cuales son representaciones de las características importantes de los elementos o recursos de un conjuntos de datos. Sin embargo, la mayoría de estas facetas se generan hoy de manera estática mediante consultas SPARQL.

Es decir, de acuerdo a las propiedades que se quieren establecer como facetas en la interfaz gráfica, se realizan las consultas a los recursos correspondientes del modelo ontológico en el grafo RDF. Sin embargo, esto trae como dificultad que si cambia el grafo RDF, también será necesario volver a realizar consultas SPARQL[3] para las nuevas propiedades de los recursos del nuevo modelo ontológico. De ahí que sea necesario utilizar elementos de las ciencias de la información para abordar esta problemática (9).

Existen varias aproximaciones de aplicaciones de dominio específico que construyen facetas de manera estática y otras tratan de automatizar este proceso mediante la utilización de elementos probabilísticos y de la teoría de la información. Entre ellos: DBLP, Rhizomer[4], Ghent University Academic Bibliography[5] y RACIEN(9). En el caso de DBLP, las facetas son estáticas, siempre se construyen las mismas y se trabaja con el mismo modelo ontológico. Por lo que se dificulta la adaptabilidad de la herramienta, es decir, la capacidad de adaptarse a otro modelo ontológico, incluso dentro del mismo dominio temático.

La Ghent University Academic Bibliography, también sigue el enfoque de DBLP, solo que aquí, siempre es un número fijo de facetas, y estas no varían, incluso al variar el modelo ontológico. Por su lado RACIEN es una herramienta creada en la Universidad de las Ciencias Informáticas (UCI) por el grupo de investigación de Web Semántica, esta utiliza solo cuatro facetas fijas aunque trabaja con un modelo ontológico que soporta la generación de un mayor número de las mismas. Del mismo modo que DBLP y la Ghent University Academic Bibliography, RACIEN no genera facetas para modelos ontológicos

[3] Lenguaje de consultas similar a SQL, pero orientados a tripletas RDF

[4] http://rhizomik.net/rhizomer/

[5] https://biblio.ugent.be/input/home

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

heterogéneos (donde varían las ontologías y clases de ontología en el mismo dominio temático) (ver figura 1).

Por su parte Rhizomer(10) emplea un mecanismo semiautomático empleando elementos de la teoría de la información para generar las facetas del modelo ontológico con que trabajan **(11)**. Para ello emplea 2 métricas: Frecuencia de la propiedad y Balance de la Propiedad. Sin embargo, no basta solo con estas dos métricas para asegurar que se generen automáticamente todas las facetas de acuerdo a las propiedades predominantes del conjunto de datos.

2. METODOS

Para la propuesta de solución, se han empleado las dos métricas definidas para la herramienta Rhizomer, más otra métrica elaborada a partir de la fórmula de la entropía de Shannon traída de la teoría de la información. Del estudio preliminar del estado del arte se proponen utilizar las siguientes métricas:

Frecuencia de cada propiedad (F_p): las facetas seleccionadas deben ser las más representativas dentro del conjunto de datos que se encuentra descrito en el grafo RDF seleccionado, lo cual implica que deben encontrarse en la mayor cantidad de recursos posibles. Esta métrica es calculada mediante el número de recursos (nr)que contienen determinada propiedad del conjunto (pi) dividido por la cantidad total de recursos (Tnr).

$$(I) \quad F_p(p_i) = \frac{nr_{(pi)}}{Tnr}$$

2. **Balance de cada propiedad (S):** para discriminar de manera eficiente el conjunto de datos en el cual se realiza la búsqueda, es necesario que las facetas adoptadas posean un rango de valores balanceado, por lo cual no sería útil seleccionar aquellas propiedades cuyas instancias poseen un valor particular, sino que impliquen la mayor cantidad de

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

instancias con igual valor. Para calcular el balance de una propiedad (p)para determinado valor (v_i) se emplea la fórmula de la entropía de Shannon (12).

$$(II)\ S = -\sum_{i=1}^{n}(p_{vi} * \log p_{vi})$$

Una vez determinados cada uno de los elementos mediante los que se efectuará el cálculo del ranking para las facetas (R (pi)), se establece una fórmula en la cual en dependencia de la importancia asociada a cada elemento calculado en las métricas descritas se asigna un determinado peso o ponderación. Para el caso de la frecuencia de cada propiedad se establece un peso (W_f) igual 0.6 y para el balance se establece un peso (W_s) igual a 0.4.

$$(III)\ R_{(pi)=W_f * F_p + W_s * S}$$

Las métricas antes expuestas se implementaron en la versión 2.0 de la herramienta RACIEN. La cual ha sido desarrollada en el Sistema Gestor de Contenidos (CMS, por sus siglas en inglés) Drupal[6] 7.0. Se ha hecho uso además del lenguaje de consultas *SPARQL 1.1* (13), la Biblioteca de Algoritmos *EasyRDF 0.7.2*[7] *y el servidor multipropósito Virtuoso 5.1.6*[8].

3. RESULTADOS

Las métricas propuestas se aplicaron al grafo RDF con que opera RACIEN en el dominio temático de los metadatos bibliográficos (modelo ontológico de RACIEN). A partir del grafo RDF cuyo modelo se presenta en la Figura 1 se generaron las facetas que se muestran en la Tabla 1. La primera columna de dicha tabla representa la propiedad del recurso en el grafo RDF y a manera de ejemplo se ha incorporado una 2da columna basada en la primea métrica propuesta en el presente trabajo

[6] www.drupal.org

[7] http://www.aelius.com/njh/easyrdf/

[8] http://virtuoso.openlinksw.com

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

basada en la entropía de Shannon (11), la frecuencia de la propiedad F_p la cual da una medida de cuan relevante es la propiedad, o cuantas veces se repite la propiedad para un número determinado de recursos dentro del grafo. Esto permite ir elaborando un orden o ranking de acuerdo a la representatividad de la propiedad dentro del conjunto de datos.

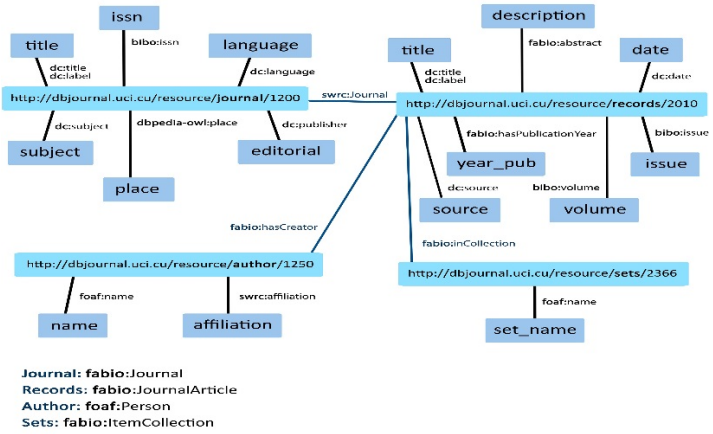

Figura 1. Modelo ontológico de RACIEN

Tabla 1. Relación Faceta Generada-métrica frecuencia de propiedad F_p

Faceta Generada	**Frecuencia de la propiedad F_p**
Autor (propiedad *name*)	7245
Título (propiedad *title*)	3711
Año (propiedad *year_pub*)	71
Fuente (propiedad *source*)	9
Afiliación (propiedad *affiliation*)	32
Editorial (propiedad *editorial*)	15
Idioma(propiedad *language*)	2
País/Ciudad(propiedad *place*)	1

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

La Tabla 1 muestra las 8 propiedades más representativas del conjunto de datos, es decir las que con más frecuencia se repiten en el conjunto de datos. Esto se evidencia a partir de una de las métricas con que más define el ranking automático de facetas (la frecuencia de la propiedad). Esta métrica por sí sola no define el ranking automático, requiere del cálculo de los dos restantes presentadas anteriormente. Combinando estas 3 es posible obtener una representación de los conceptos más importantes del conjunto de datos (facetas), ordenadas de acuerdo a su representatividad dentro del conjunto. Aunque la presente investigación continúa en proceso, sometiendo las métricas a la generación automática de facetas en conjuntos de datos de mayor volumen de recursos, los resultados obtenidos hasta ahora muestran la efectividad determinística de la combinación de estas tres métricas.

4. CONCLUSIONES

La generación automática de facetas desde conjuntos de datos representados en el formato RDF, constituye una necesidad para la explotación y aprovechamiento de las ventajas que ofrece la web semántica. En este entorno, juega un rol fundamental el uso de métodos determinísticos para determinar cuán representativos son los conceptos que están representados en el conjunto de datos y de esta forma, permitir la navegación por los datos del conjunto a cualquier tipo de usuarios, en base a la utilización de interfaces intuitivas y fáciles de utilizar a nivel productivo.

Si bien existen varias aproximaciones para navegar conjuntos de datos enlazados publicados en RDF, la utilización de las facetas y del paradigma iterativo exploratorio en general facilita el proceso de descubrir el conocimiento implícito en los datos. Por tanto, se hace necesario continuar estudiando el comportamiento de estas métricas cuando se aplican a volúmenes de datos que crecen en el tiempo, y como son capaces de mantener su eficacia a medida que el volumen de los datos crece por el aumento del número de recursos y/o propiedades.

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

5. REFERENCIAS BIBLIOGRÁFICAS

1. Yusniel Hidalgo Delgado and Rafael Rodríguez Puentes. La Web Semántica: Una Breve Revisión. *Revista Cubana de Ciencias Informáticas*. 2013. Vol. 7, no. 1.

2. TIMOTHY BERNERS-LEE. The Semantic Web. *Scientific American Magazine*. 2001. P. 29–37.

3. Tom Heath and Christian Bizer. *LINKED DATA EVOLVING THE WEB INTO A GLOBAL DATA SPACE* [online]. 1st edition. Morgan & Claypool Publishers, 2011. ISBN 9781608454310. Available from: www.morganclaypool.com

4. Carlos Heriberto Cordoví García, Claudia Hernández Rizo, Yusniel Hidalgo Delgado and Liudmila Reyes Álvarez. Using Search Paradigms and Architecture Information Components to Consume Linked Data. In : *1st Cuban Worshop on Semantic Web*. Havana, Cuba, April 2014.

5. Aba-Sah Dadzie and Matthew Rowe. APPROACHES TO VISUALISING LINKED DATA: A SURVEY. *IOS Press*. 2011. Vol. 2, no. 2, p. 89–124. DOI 10.3233/SW-2011-0037.

6. TOM HEATH. HOW WILL WE INTERACT WITH THE WEB OF DATA ? *Journal IEEE Internet Computing*. 2008. Vol. 12, no. 5, p. 81–85. DOI 10.1109/MIC.2008.101.

7. Peter Mika and Thanh Tran. SEMANTIC SEARCH - SYSTEMS, CONCEPTS, METHODS AND THE COMMUNITIES BEHIND IT. In : *Proceedings of Web Science Conference*. Evanston, USA : TKDE journal, 2012. p. 1–21.

8. Shiyali R. Ranganathan. *ELEMENTS OF LIBRARY CLASSIFICATION*. 3rd edition. Bombay : Asia Publishing House, 1962. ISBN 8185273294, 9788185273297.

Revista Publicando, 3(9). 2016, 30-40. ISSN 1390-9304

9. Carlos Heriberto Cordoví García and Claudia Hernández Rizo. *Consumo de datos enlazados mediante búsqueda textual y facetada* [online]. Habana, Cuba : Universidad de Ciencias Informáticas (UCI), 2013. Available from: https://www.researchgate.net/publication/265294470_Using_Search_Paradigms_and_Architecture_Information_Components_to_Consume_Linked_Data

10. Roberto García, Jose Maria Brunetti, Antonio López-Muzás, Juan Manuel Gimeno and Rosa Gil. Publishing and Interacting with linked Data. In : *Proceedings of the International Conference on Web Intelligence*. Sogndal, Norway, 25 May 2011.

11. SRIRAM VAJAPEYAM. Understanding Shannon's Entropy metric for Information. *CoRR*. 2 June 2014. Vol. abs/1405.2061.

12. ROBERT M. GRAY. *Entropy and information theory*. Second Edition. London : Springer Verlang, 2011. ISBN 978-1-4419-7969-8.

13. Eric Prud'Hommeaux. *SPARQL QUERY LANGUAGE FOR RDF* [online]. 2008. W3C. Available from: http://www.w3.org/TR/rdf-sparql-query/.