

Creación y manipulación de una estructura RDF sobre investigaciones de Ciencias de la computación usando la Wikidata y Twinkle

BRYAN ASMAT, DANILO BLAS, VICTOR CHAVEZ, WILLIAM GUADALUPE

Universidad Nacional de Ingeniería

Lima, Perú

basmatf@uni.pe, iblass@uni.pe, vchavezb@uni.pe, wguadalupeq@uni.pe

Resumen—En el presente trabajo se implementaron consultas de tipos SparQL que permiten hacer la búsqueda en distintas fuentes de datos como estructuras RDF, la creación de esta estructura, así como un resumen de los pasos que realizamos para la ejecución del proyecto.

Para iniciar el proyecto se usó la base de conocimiento de la Wikidata realizando consultas online <https://query.wikidata.org/> con lo cual seguiremos los siguientes pasos para que se entienda como se ha creado la estructura RDF. Accedemos al enlace mencionado anteriormente. Podemos usar las consultas con extensión .rq que están almacenadas en el repositorio de este proyecto: Si usamos el contenido del archivo `onlineQuerySELECT.rq`

copiaremos la información y la usaremos en el servicio de consultas de la Wikidata nos devuelve la tabla que hemos construido mediante las consultas que hemos hecho. Si usamos el contenido del archivo `onlineQueryCONSTRUCT.rq` copiaremos la información y la usaremos en el servicio de consultas de la Wikidata nos devuelve como serían las tripletas en una nueva estructura RDF que vamos a manejar. Descargamos el archivo TSV detallado que se genera al usar

el archivo `onlineQueryCONSTRUCT.rq` y lo modificamos para que tenga el formato Turtle a usar en el presente proyecto. El archivo RDF en formato turtle una vez generado y se encuentra en el siguiente en el directorio RDF con el siguiente nombre `RDE.ttl`. Hasta ahora los pasos que

hemos hecho es para extraer una estructura RDF nueva de Wikidata realizando consultas SparQL para obtener datos de investigaciones y trabajos científicos, digamos obras, que se basen sobre el tema de Ciencias de la Computación. De estos datos hemos extraído las siguientes características:

Título, Autor, Tema, Idioma, Revista, Fecha de Publicación, URL.

I. INTRODUCCIÓN

La información en la web se puede representar bajo un formato de datos llamado RDF. El consorcio World Wide Web (W3C) propuso una especificación llamada SparQL (lenguaje de consultas) para manipular grafos compuestos por tripletas RDF en la web y en repositorios de datos, así como extraer datos mediante consultas a relaciones desconocidas y relajar búsquedas con combinaciones complejas de base de datos.

La estructura de los resultados de una consulta SPARQL es una tabla donde cada fila es un resultado y cada columna se corresponde con una de las variables de la consulta.

Una consulta SparQL consta de tres partes:

- Coincidencia de patrones: optional, union, nesting, filtering.
- Modificadores de solución: projection, distinct, order, limit, offset.
- Parte de salida: construcción de nuevas tripletas.

Las búsquedas se realizan a través de un SparQL Endpoint, que es la interfaz de entrada y salida de estas bases de datos y que permitirá nuestras consultas. Dicho de otra forma, los Endpoints de SparQL no son más que una dirección web (URL) que acepta peticiones de búsqueda SparQL y devuelve resultados, las formas de consulta SparQL son select, construct, ask y describe.

¿Cuáles son los desafíos de diseño en SparQL?

SparQL debe tener en cuenta las características distintas de RDF:

- Debería poder extraer información de grafos RDF interconectados.
- Debe ser coherente con la semántica de mundo abierto de RDF.
- Debe ofrecer la posibilidad de agregar información opcional si está presente.
- Debería ser capaz de interpretar correctamente grafos RDF con un vocabulario con semántica predefinida.
- Debería ofrecer algunas funcionalidades para navegar en un grafo RDF.

II. ¿QUE ES LA WEB SEMÁNTICA?

La Web Semántica vendría a ser una extensión de la Web actual dotada de significado, esto es, un espacio

donde la información tendría un significado bien definido, de manera que pudiera ser interpretada tanto por agentes humanos como por agentes computarizados.

La web Semántica ha sido impulsada, entre otros, por el creador de la web, Tim Berners Lee en el 2000 para que la información sea reunida de forma que los buscadores puedan "comprender", más allá de limitarse a colocar los documentos en una lista

Figura 1. Transformer XL: Entrenamiento y Evaluación. Fuente: Transformer XL [7]

III. URLs, URIs, IRIs

III-A. URLs

El LRU es una cadena de caracteres con la que se asigna una dirección única a cada uno de los recursos de información disponibles en Internet. Existe un URL único para cada página de cada uno de los documentos de la WWW

III-B. URIs

Un identificador de recursos uniforme es una cadena de caracteres que identifica los recursos de una red de forma unívoca. La diferencia respecto a un localizador de recursos uniforme (URL) es que estos últimos hacen referencia a recursos que, de forma general, pueden variar en el tiempo.

Normalmente estos recursos son accesibles en una red o sistema. Los URI pueden ser localizador de recursos uniforme (URL),(URN)

III-C. IRIs

Es un estándar de protocolo de Internet que se basa en el protocolo Identificador uniforme de recursos (URI) al expandir en gran medida el conjunto de caracteres permitidos.

IV. ESTRUCTURA RDF

RDF es un método general para descomponer cualquier tipo de conocimiento en trozos pequeños, con algunas reglas acerca de la semántica o significado, de esas piezas.

El punto es tener un método tan simple que puede expresar cualquier hecho, y a la vez tan estructurada que las aplicaciones informáticas pueden hacer cosas útiles con él. **Attention Is All You Need** [1]

RDF representa la información en una declaración triple -también llamada tripleta- que sigue la estructura sujeto-predicado-objeto **Aquí citaremos las paginas** [1]

Ejemplo

<La ronda de noche><fue creada por><Rembrandt>

En este ejemplo<La ronda de noche> y el objeto <Rembrandt> se pueden considerar como dos nodos en un

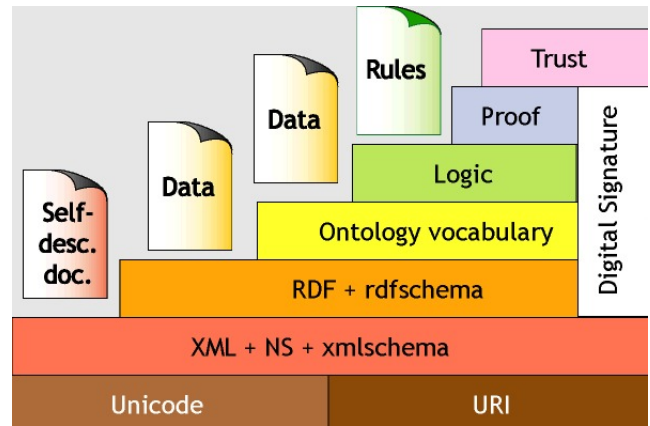


Figura 2. Web semántica [7]

grafo donde el predicado <fue creada por> define una arista o la relación entre estos .

V. BUSCANDO RDF CON SPARQL

SPARQL nos permite traducir datos en grafo, intencionalmente enlazados, en datos normalizados en formato tabular, esto es, distribuidos en filas y columnas, que se pueden abrir en formato como .csv .json .ttl ,etc o también se pueden ver en programas de visualización

Resulta útil pensar las consultas SPARQL como un conjunto de oraciones con espacios en blanco. La base de datos tomará esta consulta y encontrará cada conjunto de oraciones que encaje correctamente en estos espacios en blanco, devolviéndonos los valores coincidentes como una tabla.

VI. CONSULTAS SPARQL

SPARQL, acrónimo de *SPARQL Protocol and RDF Query Language*, es un lenguaje estandarizado para hacer consultas en estructuras RDF. Es una tecnología clave en el desarrollo de la web semántica que se constituyó como recomendación oficial del W3C el 15 de enero de 2008, siendo actualizado a la versión 1.1 en 2013 [1]. En un principio SPARQL únicamente incorpora funciones para la recuperación sentencias RDF. Sin embargo, algunas propuestas también incluyen operaciones para el mantenimiento (creación, modificación y borrado) de datos.

VI-A. Especificaciones

SPARQL 1.0:

- SPARQL como lenguaje de consultas para RDF, cubre la sintaxis de las consultas (QL).
- SPARQL como protocolo para RDF especifica cómo un programa debe pasar consultas SPARQL a un servicio de procesamiento de consultas SPARQL y cómo este servicio las retorna.
- Formato XML de resultados de consultas SPARQL, describe un formato simple XML para que los

- La especificación de consulta federada describe como una simple consulta puede traer data de multiples fuentes, esto hace más fácil construir aplicaciones que toman ventaja de entornos distribuidos.
- La especificación de actualización es la diferencia clave entre SPARQL 1.0 y 1.1 porque lleva a SPARQL de ser un lenguaje de consulta a algo que puede agregar datos a un conjunto de datos, reemplazarlos y eliminarlos también.
- La especificación de Descripción del servicio describe cómo un programa cliente puede preguntarle a un motor SPARQL exactamente qué características admite.
- La especificación del formato JSON de los resultados de la consulta describe un equivalente JSON del formato XML de los resultados de la consulta.
- La especificación de los formatos CSV y TSV de los resultados de la consulta describe los equivalentes de valores separados por comas y tabuladores del formato XML de los resultados de la consulta.
- La especificación del Protocolo HTTP de Graph Store amplía el Protocolo SPARQL con una API similar a REST para la comunicación entre un cliente y un procesador SPARQL sobre gráficos o conjuntos de tripletas.

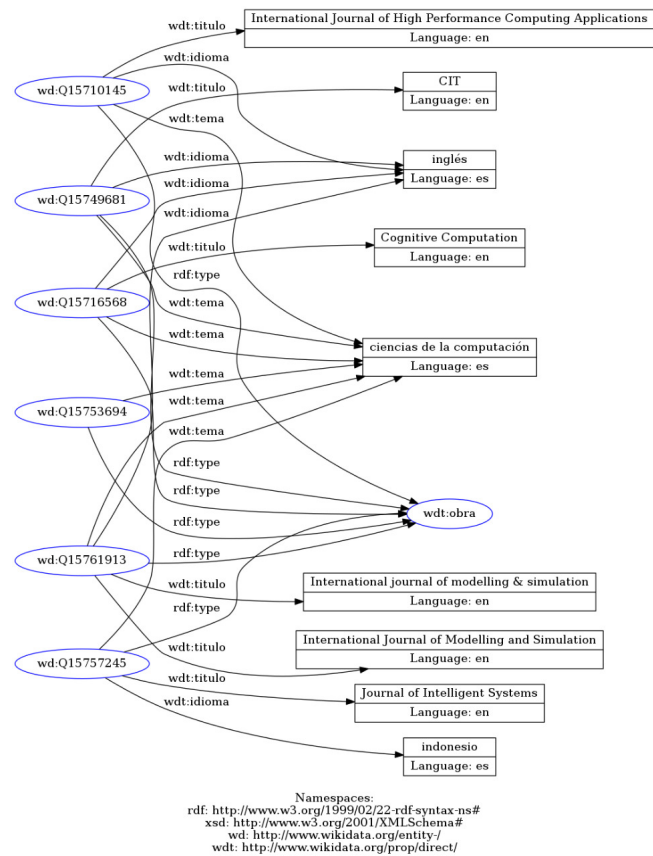


Figura 4. Grafo generado con 30 tripletas de la estructura RDF utilizada [2]

VI-C. Consultas SPARQL hechas en la estructura RDF.

VI-B. Estructura RDF a utilizar

```

>@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
>@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
>@prefix wd: <http://www.wikidata.org/entity/> .
>@prefix wdt: <http://www.wikidata.org/prop/direct/> .

1>wd:Q15710145 rdf:type wdt:obra .
2>wd:Q15710145 wdt:tema "ciencias de la computación"@es .
3>wd:Q15710145 wdt:titulo "International Journal of High Performance Computing Applications"@en .
4>wd:Q15710145 wdt:idioma "inglés"@es .
5>wd:Q15749681 rdf:type wdt:obra .
6>wd:Q15749681 wdt:tema "ciencias de la computación"@es .
7>wd:Q15749681 wdt:titulo "CIT"@en .
8>wd:Q15749681 wdt:idioma "inglés"@es .
9>wd:Q15716568 rdf:type wdt:obra .
10>wd:Q15716568 wdt:tema "ciencias de la computación"@es .
11>wd:Q15716568 wdt:titulo "Cognitive Computing"@en .
12>wd:Q15716568 wdt:idioma "inglés"@es .
13>wd:Q15757245 rdf:type wdt:obra .
14>wd:Q15757245 wdt:tema "ciencias de la computación"@es .
15>wd:Q15757245 wdt:titulo "Journal of Intelligent Systems"@en .
16>wd:Q15757245 wdt:idioma "indonesio"@es .
17>wd:Q15761913 rdf:type wdt:obra .
18>wd:Q15761913 wdt:tema "ciencias de la computación"@es .
19>wd:Q15761913 wdt:titulo "International journal of modelling & simulation"@en .
20>wd:Q15761913 wdt:idioma "inglés"@es .
21>wd:Q15761913 wdt:titulo "International Journal of Modelling and Simulation"@en .
22>wd:Q15753694 rdf:type wdt:obra .
23>wd:Q15753694 wdt:tema "ciencias de la computación"@es .
24>wd:Q15753694 wdt:titulo "AUSIS Australian Journal of information systems"@en .
25>wd:Q15753694 wdt:idioma "inglés"@es .
26>wd:Q15753694 wdt:titulo "Australasian Journal of Information Systems"@en .
27>wd:Q15749681 wdt:titulo "Journal of Computing and Information Technology"@en .
28>wd:Q15757588 rdf:type wdt:obra .
29>wd:Q15757588 wdt:tema "ciencias de la computación"@es .
30>wd:Q15757588 wdt:titulo "Ethics and Information Technology"@en .

```

Figura 3. 30 tripletas de la estructura RDF utilizada (ver archivo RDF.ttl)

Las siguientes imágenes muestran las capturas de algunas de las consultas hechas a la estructura RDF elegida, en el programa Twinkle.

Data URI		File ...
# 1 : Mostre as 50 primeiras tripletas armazenadas na tabela		
SELECT + WHERE: ?subject Predicate ?object .		
LIMIT 50		
subject	predicate	object
http://www.wikidata.org/entity/Q737949	http://www.wikidata.org/wiki/dct:diana	plurilingualism
http://www.wikidata.org/entity/Q737949	http://www.wikidata.org/wiki/dct:título	Kantha pith galati rambura paṇon-gāni
http://www.wikidata.org/entity/Q737949	http://www.wikidata.org/wiki/dct:tema	Servicos de la computación
http://www.wikidata.org/entity/Q737949	http://www.wiki.org/1999-02-22-of-syntaxes#type	Services de la computation
http://www.wikidata.org/entity/Q294249	http://www.wikidata.org/wiki/dct:idioma	Portugies
http://www.wikidata.org/entity/Q294249	http://www.wikidata.org/wiki/dct:tema	Journal d'Automata, Languages and Combinatorics
http://www.wikidata.org/entity/Q294249	http://www.wikidata.org/wiki/dct:tema	Servicos da computação
http://www.wikidata.org/entity/Q294249	http://www.wiki.org/1999-02-22-of-syntaxes#type	Servicos da computacao
http://www.wikidata.org/entity/Q7645405	http://www.wikidata.org/wiki/dct:efecto	http://measurer.tudelft.nl/out/110XMF/42d2+4f89+aeb-dbeet/1897/fc/3e14-C6-EF7D0A0E0C2/?http://www.ai.stanford.edu/LINKS/default.asp
http://www.wikidata.org/entity/Q7645405	http://www.wikidata.org/wiki/dct:efecto	Portugies
http://www.wikidata.org/entity/Q7645405	http://www.wikidata.org/wiki/dct:tema	Komari C.
http://www.wikidata.org/entity/Q7645405	http://www.wikidata.org/wiki/dct:tema	Analyzing the Change-Freeness of APIs and web API-fests
http://www.wikidata.org/entity/Q7645405	http://www.wikidata.org/wiki/dct:tema	Servicos de la computacion
http://www.wikidata.org/entity/Q7645405	http://www.wiki.org/1999-02-22-of-syntaxes#type	Servicos - wikidata.org/wikipedio/cidba
http://www.wikidata.org/entity/Q2841196	http://www.wikidata.org/wiki/dct:idioma	Portugies
http://www.wikidata.org/entity/Q2841196	http://www.wikidata.org/wiki/dct:tema	Aptica transference on speed and information processing

Figura 5. Mostrar 50 primeras tripletas almacenadas en la tabla.

VI-G. RoBERTa

Es un modelo introducido en Facebook. El enfoque BERT robustamente optimizado RoBERTa [10], es un re-entrenamiento de BERT con una metodología de entrenamiento mejorada, un 1000 % más de datos y potencia de cálculo. Para mejorar el procedimiento de entrenamiento, RoBERTa elimina tareas como Next Sentence Prediction (NSP) del entrenamiento previo de BERT e introduce el enmascaramiento dinámico para que el token enmascarado cambie durante las épocas de entrenamiento.

RoBERTa supera a BERT y XLNet en los resultados de referencia de [GLUE](#).

VI-H. DistilBERT

Este modelo aprende una versión destilada (aproximada) de BERT, que retiene el 95 % de rendimiento pero utiliza solo la mitad del número de parámetros. Específicamente, no tiene embeddings de tipo token, pooler y retiene solo la mitad de las capas del BERT de Google.

DistilBERT [11], utiliza una técnica llamada destilación, que se aproxima a BERT de Google. La idea es que una vez que se ha entrenado una gran red neuronal, sus distribuciones de salidas completas se pueden aproximar usando una red más pequeña. Una de las funciones clave de optimización utilizadas para la aproximación posterior en las estadísticas bayesianas es la divergencia de Kulback-Leiber.

VII. CONCLUSIONES

- Se ha realizado clasificación multiclase utilizando la librería `simpletransformers` y modelos pre-entrenados, basados en BERT, obteniendo buenos resultados en métricas de clasificación, especialmente del modelo `distilBERT`, que tiene un menor número de parámetros que BERT y que utiliza el aprendizaje `teacher-students`, donde se entrena a una red de `students` para imitar la distribución de salida completa de la red del `teacher` (su conocimiento) y retiene el rendimiento de BERT.
- En la tarea de traducción, se realizaron comparaciones entre distintos modelos `seq2seq` y los modelos `transformers`, donde estos últimos obtuvieron mejores resultados. Sin embargo el modelo Transformer XL modificado no logró superar al Transformer original.

REFERENCIAS

- [1] <https://www.w3.org/TR/rdf-sparql-query/>.
- [2] Página web para generar grafos a partir de una estructura RDF: .
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton. Layer Normalization (2016). Disponible en <https://arxiv.org/abs/1607.06450>.
- [4] Kishore Papineni, Salim Roukos, Todd Ward, Wei-jing Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation, pp. 311-318. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).
- [5] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Łukasz Kaiser. Universal Transformers (2018). ICLR 2019. Disponible en <https://arxiv.org/abs/1807.03819>.
- [6] Alex Graves. Adaptive Computation Time for Recurrent Neural Networks (2017). Disponible en <https://arxiv.org/abs/1603.08983>.
- [7] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context (2019). Disponible en <https://arxiv.org/abs/1901.02860>.
- [8] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding (2019). Disponible en <https://arxiv.org/abs/1906.08237>.
- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018). Disponible en <https://arxiv.org/abs/1810.04805>.
- [10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). Disponible en <https://arxiv.org/abs/1907.11692>.
- [11] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter (2019). Disponible en <https://arxiv.org/abs/1910.01108>.
- [12] Pytorch-Transform: Implementación de transformer para NLP. https://pytorch.org/hub/huggingface_pytorch-transformers/.
- [13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Jamie Brew. Transformers: State-of-the-art Natural Language Processing (2019), HuggingFace Inc. Disponible en <https://arxiv.org/abs/1910.03771>.
- [14] SimpleTransformers <https://pypi.org/project/simpletransformers/>.
- [15] Xiang Zhang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification(2015). Disponible en <https://arxiv.org/abs/1509.01626>.
- [16] Desmond Elliott, Stella Frank, Khalil Sima'an, Lucia Specia. Multi30K: Multilingual English-German Image Descriptions(2016). Disponible en <https://arxiv.org/abs/1605.00459>.
- [17] TransformerXL from Scratch [notebook](#).