Construcción de un grafo de conocimiento para un observatorio inmobiliario

Felipe Dioguardi[1][0000−0001−6039−8653], Diego Torres[1,2][0000−0001−7533−0133], Leandro Antonelli[1][0000−0003−1388−0337], y Juan Pablo del Río[3][0000−0002−4031−3007]

[1] LIFIA, CICPBA-Facultad de Informática, UNLP
{nombre.apellido}@lifia.info.unlp.edu.ar
[2] Departamento de Ciencia y Tecnología, UNQ
[3] LINTA-CICPBA, CONICET, UNLP

Resumen Los observatorios inmobiliarios permiten la producción y sistematización de datos provenientes del mercado inmobiliario. En manos de estadistas y expertos del dominio, resultan herramientas invaluables para el estudio de los valores del suelo en un área geográfica determinada. Crear un observatorio inmobiliario requiere la disponibilidad de una gran cantidad de datos, lo que puede resultar un problema si no se cuenta con información extensa, confiable, actualizada, y pública. Para solucionarlo, este artículo presenta una metodología para la extracción de conocimiento proveniente de páginas web dedicadas a la publicación de avisos inmobiliarios, utilizando tecnologías de *web scraping*. Además, propone el almacenamiento de la información inmobiliaria en un grafo de conocimiento estructurado por una ontología acorde al dominio, que dotará los datos externos de valor semántico. Esto posibilitará la inferencia de nuevo conocimiento, y facilitará su manipulación por parte de máquinas y sistemas automatizados. Por último, este artículo ofrece los resultados preliminares de la implementación de una herramienta que sigue con la metodología propuesta, con relación a la capacidad de relevamiento inherente a un proceso manual. Este trabajo se desarrolla en el marco de la tesina de grado titulada "Evaluación de técnicas de detección de duplicados sobre grafos de conocimiento de avisos inmobiliarios", presentada con el proyecto "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano".

Palabras clave: Observatorio inmobiliario · Grafos de conocimiento · Ontologías · Web semántica

1. Introducción

Los observatorios inmobiliarios son herramientas de información que permiten capturar los precios y características de bienes inmuebles en una zona geográfica particular. Los datos que estos brindan pueden resultar un valioso aporte al Estado, pues son la base de investigaciones y estudios estadísticos realizables en el marco de la creación y mejora de políticas sociales.

En Argentina en general y en la provincia de Buenos Aires en particular, existe una carencia estructural en la disponibilidad pública de información sobre los valores del mercado inmobiliario. Por este motivo, en agosto de 2021 el Ministerio de Ciencia, Tecnología e Innovación aprobó el proyecto denominado "Observatorio de valores del suelo e instrumentos de financiamiento del desarrollo urbano". Este tiene como objetivo principal resolver la falta de información estratégica de valores de mercado inmobiliario para cuantificar las valorizaciones producidas por la acción del Estado, en el marco de la política de Integración social y urbana de barrios populares [6].

En las instancias iniciales del proyecto, los investigadores realizaron un balance de las diversas fuentes de información disponibles, clasificándolas según la cantidad y calidad de los avisos publicados sobre algunos partidos de interés de la provincia de Buenos Aires. A partir de esa evaluación, seleccionaron diferentes sitios de ofertas inmobiliarias, y recabaron manualmente información acerca de cientos de los inmuebles publicitados en ellos.

Para poder efectuar un estudio completo del valor del suelo en las distintas áreas geográficas, es necesario tener la capacidad de extraer grandes volúmenes de datos de manera sistemática. *Web scraping* es una técnica para obtener información de páginas de Internet, y almacenarla en un archivo o base de datos local para su posterior análisis [9]. A su vez, un *web crawler*, *spider*, o araña, es un agente informático utilizado para la descarga masiva de páginas web [7,8]. Ambos conceptos suelen acompañarse, pues es común querer recolectar en una misma base el conocimiento contenido en un gran conjunto de páginas web.

Una herramienta de *web scraping* con la capacidad de acceder a todos los avisos inmobiliarios útiles de los sitios seleccionados se presenta como una alternativa adecuada para resolver la tarea propuesta. La herramienta deberá además normalizar y estructurar los datos obtenidos, para garantizar que el análisis consecuente pueda llevarse a cabo. Esta necesidad surge del carácter heterogéneo inherente a las diversas páginas de Internet.

Una manera de formalizar el conocimiento recuperado es a través del uso de ontologías [1]. Una ontología es una descripción del conocimiento sobre un dominio de interés, cuyo núcleo es una especificación procesable por las máquinas con un significado formalmente definido [5]. Definir una ontología para avisos inmobiliarios no solo permitiría establecer un esquema más riguroso para representar la información, sino que también la dotará de valor semántico que podrá ser aprovechado para su curado.

Este artículo, elaborado en el marco de la tesina de grado titulada "Evaluación de técnicas de detección de duplicados sobre grafos de conocimiento de avisos inmobiliarios", presenta la construcción de una herramienta que permite extraer de conocimiento de páginas web y almacenarlo en un grafo de conocimiento, para su utilización en un observatorio inmobiliario. La sección 2 explica cómo se normalizó el vocabulario hallado en los distintos sitios web, ahondando en la definición de una ontología inmobiliaria con base en la reutilización y alineando estándares preexistentes. La sección 3 introduce la metodología de extracción de conocimiento de los portales web, junto con la descripción de una

herramienta que la implementa, y sus resultados preliminares. Finalmente, la sección 4 muestra las conclusiones y detalla posibles trabajos futuros.

2. Definición de una ontología inmobiliaria

Para llevar a cabo un estudio estadístico del mercado inmobiliario, es necesario seleccionar los datos relevantes de cada aviso disponible. Esto implica realizar un análisis de los distintos portales web a explorar, determinando que campos o variables contienen el conocimiento deseado en cada uno.

En primer lugar, fue necesario definir un vocabulario común para normalizar los nombres con los que cada página se refiere a los conceptos de interés. Por ejemplo, un aviso del *sitio 1* podría referirse al valor al cual se oferta cierta propiedad como *precio*, mientras que un aviso del *sitio 2* como *price*.

Para conseguir estructurar la información proveniente de avisos inmobiliarios es necesario identificar los conceptos principales que los representan y las relaciones que los vinculan.

En este dominio es necesario centrarse primordialmente en dos aspectos: uno referente al inmueble, y otro al aviso que lo oferta. Para lograr este objetivo, se realizó un estudio del estado del arte en lo que respecta a ontologías inmobiliarias y de publicaciones online. Para modelar el conocimiento relativo al inmueble como los espacios físicos, edificios, y sus habitaciones, se utilizó la ontología RealEstateCore. Del mismo modo, para conceptualizar los avisos inmobiliarios y su información (a qué sitio pertenece y quién lo publicita), se aprovechó la ontología SIOC.

RealEstateCore [4] es una ontología modular y libre que rápidamente se convirtió en un estándar en el modelado de conocimiento sobre edificios inteligentes. De todos los conceptos que incluye, son destacables *rec:Real_Estate* y *rec:Space*. *rec:Real_Estate* es el elemento que representa una propiedad inmueble, pudiendo estar conformada por más de un edificio, terreno, o similar. Por otra parte, *rec:Space* representa un área del mundo físico que puede a su vez contener subespacios, por lo que permite representar regiones, terrenos, edificios, y habitaciones.

Además, RealEstateCore hace uso del concepto de *foaf:Agent* definido en FOAF [3] para referirse a los humanos u organizaciones que realizan una acción, particularmente sobre un *rec:Real_Estate*. La clase *foaf:Agent* resulta útil a la hora de vincular una propiedad en el mercado con la persona o entidad que la oferta. Para representar que un *foaf:Agent* publicita un inmueble, detallar las plataformas en las que publica los avisos, y adjuntar el conocimiento que corresponde a cada oferta, se utiliza la ontología SIOC.

SIOC provee los principales conceptos y propiedades requeridos para describir información sobre comunidades online en la Web Semántica [2]. Define clases para modelar sitios web, los items que contienen, el contenido de cada uno, y el tipo de publicaciones que se realizan. La utilidad de SIOC a la hora de modelar avisos inmobiliarios está determinada principalmente por las clases *sioc:Site*, *sioc:Post*, y *foaf:Agent* (que referencia directamente al recurso definido en FOAF). Por un lado, *sioc:Site* representa un sitio web que actúa como

comunidad online. Por el otro, *sioc:Post* es un artículo que se publica en un *sioc:Site*. Y finalmente *foaf:Agent* representa aquellos actores que cumplen un rol en alguna tarea. El uso de SIOC posibilita modelar los avisos inmobiliarios como *sioc:Post*s, dado que suelen permitir visualizar y filtrar conjuntos de publicaciones, solo que limitando las interacciones entre los usuarios. Además, permite indicar que los avisos pertenecen a una plataforma identificada como un *sioc:Site*, y son publicados por *foaf:Agent*s, que podrían ser inmobiliarias o personas particulares. Además, SIOC presenta la propiedad *sioc:about*, que permite vincular a un *Post* con el recurso principal al que hace referencia, sin importar que estuviera definido por otra ontología.

El hecho que tanto RealEstateCore como SIOC hayan diseñado sus clases teniendo en cuenta la definición de *foaf:Agent* permite que ambas ontologías se alineen fácilmente mediante el anunciante de un aviso inmobiliario, y a través de la propiedad específica que se tiene en consideración. A esto pueden sumarse la definición de nuevas clases y propiedades que permitan mejorar la estructura del modelo para el dominio inmobiliario. Por ejemplo, podría agregarse una clase *RealEstateListing* como subclase de *sioc:Post* para representar ofertas inmobiliarias, y una propiedad *price* que indique el precio del inmueble según ese aviso particular.

El resultado de este alineamiento es una ontología que permite crear una base de conocimiento capaz de inferir las respuestas a preguntas del estilo *¿cuántas inmobiliarias ofertan el departamento en Libertador al 400?* y *¿a qué precios se ofrece la quinta en Gral. San Martín al 1600?*

3. Recolección de datos

En las etapas iniciales del proyecto los investigadores realizaron un relevamiento manual de información inmobiliaria, seleccionando tres sitios web dedicados a la búsqueda de propiedades e inmuebles, y cinco partidos de prueba dentro la provincia de Buenos Aires. Luego realizaron un muestreo de los sitios, que les permitió formar una base de datos de aproximadamente 2.000 clasificados inmobiliarios en el transcurso de entre 2 y 3 meses.

A fin de conseguir una muestra de mayor tamaño, y de aumentar la significancia estadística del análisis a efectuar, se optó por automatizar el proceso utilizando tecnologías de *web crawling* y *web scraping*. Para eso se diseñó un *scraper web* en Python utilizando como base Scrapy, un framework para la navegación de sitios y extracción de datos estructurados.

Antes de utilizar las arañas, fue necesario configurar períodos de demora personalizados entre cada petición realizada, para evitar sobrecargar los servidores de los portales y no generar problemas en su funcionamiento. Una vez confeccionadas las arañas de cada sitio, se ejecutaron en un entorno paralelo en el que en menos de una semana consiguieron recolectar conocimiento de más de 500.000 avisos inmobiliarios de 40 partidos de la provincia de Buenos Aires.

4. Conclusiones y trabajos futuros

En este artículo se presentó una alternativa para automatizar la extracción y almacenamiento de conocimiento de portales inmobiliarios. Primero, se definió una terminología en inglés con los expertos del dominio, para normalizar el vocabulario variable de las distintas páginas de Internet. Se realizó un estudio del estado del arte de los estándares ontológicos para la representación de inmuebles y publicaciones web. A partir de este, se alinearon las ontologías RealEstateCore y SIOC, consiguiendo un modelo estructurado para la descripción avisos inmobiliarios publicados en la red. Partiendo de este modelo, se diseñó una metodología de extracción de conocimiento basada en técnicas de *web crawling* y *web scraping*. La misma se implementó en una herramienta capaz de almacenar la información de las ofertas inmobiliarias en una estructura que respeta el modelo definido. Finalmente, se comprobó que la herramienta fue capaz de aumentar el tamaño del corpus de datos en más de un 1.000 % en relación a los esfuerzos manuales previos.

Como trabajos futuros se pondrá el foco en la detección automática entidades duplicadas en el grafo de conocimiento, especialmente en lo que refiere a inmuebles referenciados por múltiples avisos. Con este fin será necesario analizar y evaluar diferentes técnicas de deduplicación en grafos de conocimiento, para así determinar la estrategia más eficaz para solucionar el problema.

Referencias

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American p. 4 (May 2001)
2. Breslin, J., Decker, S., Harth, A., Bojars, U.: SIOC: An approach to connect web-based communities. IJWBC **2**, 133–142 (Jan 2006). https://doi.org/10.1504/IJWBC.2006.010305
3. Brickley, D., Miller, L.: FOAF Vocabulary Specification. Namespace Document 2 Sept 2004, FOAF Project (2004), http://xmlns.com/foaf/0.1/
4. Hammar, K., Wallin, E.O., Karlberg, P., Hälleberg, D.: The RealEstateCore Ontology. In: Ghidini, C., Hartig, O., Maleshkova, M., Svátek, V., Cruz, I., Hogan, A., Song, J., Lefrançois, M., Gandon, F. (eds.) The Semantic Web – ISWC 2019, vol. 11779, pp. 130–145. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_9
5. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. Chapman & Hall - CRC Press (Aug 2009). https://doi.org/10.1201/9781420090512, journal Abbreviation: Foundations of Semantic Web Technologies Publication Title: Foundations of Semantic Web Technologies
6. Observatorio de valores del suelo para fortalecer la política de Integración social y urbana de barrios populares (Aug 2021), https://www.argentina.gob.ar/noticias/observatorio-de-valores-del-suelo-para-fortalecer-la-politica-de-integracion-social-y
7. Olston, C., Najork, M.: Web Crawling. Foundations and Trends® in Information Retrieval **4**(3), 175–246 (2010). https://doi.org/10.1561/1500000017, http://www.nowpublishers.com/article/Details/INR-017

8. Schrenk, M.: Webbots, Spiders, and Screen Scrapers: A Guide to Developing Internet Agents with PHP/CURL. No Starch Press, second edition edn. (2012)
9. Zhao, B.: Web Scraping. In: Schintler, L.A., McNeely, C.L. (eds.) Encyclopedia of Big Data, pp. 1–3. Springer International Publishing, Cham (2017). https://doi.org/10.1007/978-3-319-32001-4_483-1