Web Mining y Text Mining: enfoques avanzados para analizar el contenido de grandes cantidades de información

José Federico Medrano[1], Valeria Barriento[2]

jfmedrano@fi.unju.edu.ar, barriento.valeria@gmail.com

[1]VRAIn / Visualización y Recuperación Avanzada de Información / Facultad de Ingeniería

Universidad Nacional de Jujuy (UNJu) - Ítalo Palanca 10, +54 (388) 4221587

[2] Escuela Superior de Ciencias Jurídicas y Políticas / Universidad Nacional de Jujuy (UNJu)

RESUMEN

A medida que aumenta la cantidad de información contenida y disponible en la web, analizarla, descubrir patrones y conocimiento informativo demanda grandes cantidades de tiempo. Los buscadores y los motores de bases de datos pueden facilitar en parte la tarea de encontrar contenido adecuado, pero en sitios web grandes, donde los resultados de las búsquedas se cuentan por miles o decenas de miles es necesario aplicar enfoques avanzados que permitan relacionar el contenido buscado de algún modo. Este trabajo plantea la aplicación de técnicas de *Web Mining* y *Text Mining* para procesar grandes cantidades de información de sitios web de noticias para ofrecer contenido relevante y relacionado a partir de una búsqueda inicial. Una de las técnicas a emplear será el modelado temático, que permitirá por un lado conocer los distintos temas o tópicos que tratan estas noticias y por otro lado, una vez identificados los conjuntos de temas, hallar las diversas interrelaciones entre ellos. Esto permitirá describir y analizar de un modo objetivo la información ofrecida por este tipo de portales. Del mismo modo, este trabajo también plantea el estudio y análisis de sitios web de avisos clasificados, de manera de caracterizar por un lado la oferta de inmuebles y por otro la demanda de perfiles para distintos puestos de trabajo.

Palabras clave: *Web Mining; Text Mining; Procesamiento del Lenguaje Natural; Topic Modeling; Recuperación de Información*

CONTEXTO

La línea de investigación aquí presentada se encuadra dentro del Proyecto Bianual 2020-2021 D/B035 denominado "*Agentes Inteligentes para Recuperación de Información y Analítica Visual en Big Data*", aprobado y financiado por la Secretaría de Ciencia y Técnica y Estudios Regionales de la Universidad Nacional de Jujuy (SeCTER – UNJu).

Continuando con lo expuesto en (Medrano, 2020), este proyecto es llevado a cabo por el grupo de investigación Visualización y Recuperación Avanzada de Información

(VRAIn) de la Facultad de Ingeniería de la UNJu.

1. INTRODUCCIÓN

El *Text Mining* o Minería de Texto, un tipo particular de minería de datos, tiene como objetivo extraer conocimientos útiles como relaciones, patrones y tendencias de datos no estructurados o semiestructurados, por ejemplo, documentos de texto (Feldman & Sanger, 2006). El proceso principal en la minería de textos es transformar el texto en datos numéricos utilizando métodos estadísticos para extraer el contenido textual en una matriz organizada documento-término, que abarca las siguientes dos dimensiones: las palabras (o términos, compuestos por *n* palabras) y los documentos (Moro, Pires, Rita, & Cortez, 2019). Estas técnicas aportan una gran ventaja al momento de analizar corpus textuales de miles de registros, puesto que automatizan parte del proceso de extracción de nuevo conocimiento, facilitando los análisis y la obtención de conclusiones de manera más sencilla.

La minería de textos extrae información relevante que se encuentra contenida en los diferentes formatos en los que se puede encontrar un conjunto textual como pueden ser: artículos de revistas, páginas web, entre otros (Ariza-Colpas, Oviedo-Carrascal, & De-la-hoz-Franco, 2019). En este sentido ha sido ampliamente utilizada en diversas áreas del conocimiento como la biomedicina (Kim & Delen, 2018), análisis de sentimiento y opiniones (Liu, 2012), recuperación de información (Meystre, Savova, Kipper-Schuler, & Hurdle, 2008), ciencias sociales (Ignatow & Mihalcea, 2016).

El análisis y procesamiento de grandes cantidades de texto no es una tarea sencilla; encontrar relaciones o nuevo conocimiento, a veces oculto entre miles o millones de registros, se torna impracticable sin recurrir a modernas técnicas y algoritmos de Aprendizaje Automático y Procesamiento del Lenguaje Natural. Al respecto, una de las técnicas muy populares y que ha dado buenos resultados al analizar conjuntos enormes de datos es el modelado de temas o *topic modeling*. Un tipo de modelo estadístico para descubrir los "temas" abstractos que ocurren en una colección de documentos. El modelado de temas es una herramienta de minería de texto (*text mining*) de uso frecuente para el descubrimiento de estructuras semánticas ocultas en un cuerpo de texto.

Como lo indica (García-Marco, Figuerola, & Pinto, 2020), la misión del modelado de temas consiste, en identificar el conjunto de temas de la colección documental y en establecer la proporción de cada tema en cada documento. Estas operaciones se basan en la coocurrencia de palabras en los mismos documentos; y permiten establecer conjuntos de palabras definitorias, con mayor o menor peso, de cada tema. La presencia de unas u otras palabras en cada documento permite también estimar el porcentaje o proporción que cada tema juega en el contenido de ese documento.

La minería web es en realidad un área de minería de datos relacionada con la información disponible en internet. Es un concepto de extracción de datos informativos disponibles en páginas web (Kumar & Singh, 2016; Mughal, 2018).

Para la extracción de datos de una página web se emplean diferentes herramientas y algoritmos, la información a extraer puede ser

la estructura (hipervínculos), el uso (páginas visitadas, uso de datos), o el contenido (documento de texto, páginas), siendo este último el más empleado. Este proceso involucra áreas como la Recuperación de Información (Baeza-Yates & Ribeiro-Neto, 1999) y el Procesamiento del Lenguaje Natural (PLN), para poder recuperar los registros necesarios y luego procesarlos de un modo adecuado de acuerdo a la necesidad del problema; por ello como se mencionó previamente, el modelado temático, una técnica de PLN, permitirá la identificación de patrones y relaciones dentro del conjunto textual.

Uno de los objetivos de este trabajo es la descripción avanzada o mejorada del contenido de grandes sitios web. Por ejemplo, para portales de noticias como La Nación[1], resultados de búsquedas como "violencia de género", "Alberto Fernández", "dólar" o "vacuna covid", por citar algunos ejemplos, ofrecen miles, decenas de miles y en algunos casos centenas de miles de resultados. Para poder procesar esta enorme cantidad de información y caracterizar de manera objetiva el contenido textual de las mismas, es necesario recurrir a las técnicas avanzadas que se mencionaron. Puesto que automatizando gran parte de los procesos de recolección, limpieza y procesamiento de datos, se podrá destinar gran parte del tiempo para el análisis finito de las relaciones identificadas.

Así mismo, este trabajo se plantea caracterizar la demanda de empleo y la oferta de inmuebles a partir de la recolección de avisos clasificados. De este modo se podrán comparar o establecer el precio a un inmueble de acuerdo a las características propias y del conjunto de viviendas circundantes, o se podrá conocer, identificar y monitorear los requerimientos para un perfil de empleo determinado (Karakatsanis, y otros, 2017; Papoutsoglou, Mittas, & Angelis, 2017).

1 https://www.lanacion.com.ar/

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La presente investigación se enfocará en dos aspectos claves:

- Extracción automática de grandes cantidades de información de sitios web.
- Empleo de técnicas de PLN para encontrar información relevante.

Debido al dinamismo de la web, el contenido cambia, se actualiza, se agrega o elimina constantemente, por ello para analizar el contenido de un sitio web sería necesario contar con una instantánea completa que permita tener un vistazo de un momento determinado. Es aquí donde cobra importancia el *web scraping*, una técnica de Recuperación de Información empleada para extraer información de sitios web donde no se cuenta con una API o cuando los datos disponibles son escasos o limitados. De este modo, una vez extraídos los datos relevantes, se conforma un *dataset* para ser procesado y analizado para hallar patrones o tendencias que permitan relacionar distintos conjuntos temáticos o agrupaciones de registros.

3. RESULTADOS OBTENIDOS/ESPERADOS

Se espera en una primera instancia construir un *crawler* específico para un portal para poder extraer y recolectar toda la información necesaria. Puesto que los sitios web objeto de estudio no disponen de mecanismos que

permitan recolectar la información por medio de una API, esta tarea se llevará a cabo mediante *web scraping*.

Por otro lado, se espera caracterizar la información recolectada, empleando un modelado de temas y visualizando los datos con librerías especializadas en la materia.

4. FORMACIÓN DE RECURSOS HUMANOS

El Equipo de Trabajo está conformado por docentes investigadores de la Universidad Nacional de Jujuy. Los mismos llevan adelante esta línea de investigación desde hace años. Cada año se incorporan al proyecto alumnos avanzados de distintas carreras, quienes trabajan en temas relacionados con las temáticas planteadas. Del mismo modo, los integrantes del equipo participan en el dictado de asignaturas/cursos de grado y postgrado de la UNLP, UNJu y UCSEDASS.

5. BIBLIOGRAFÍA

Ariza-Colpas, P., Oviedo-Carrascal, A., & De-la-hoz-Franco, E. (2019). Using K-Means Algorithm for Description Analysis of Text in RSS News Format. *International Conference on Data Mining and Big Data*, (págs. 162-169). Obtenido de https://academic.microsoft.com/paper/2963086233

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval.* Obtenido de https://academic.microsoft.com/paper/1660390307

Feldman, R., & Sanger, J. (2006). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data.* Obtenido de https://academic.microsoft.com/paper/2024228866

García-Marco, F. J., Figuerola, C., & Pinto, M. (2020). Análisis de la evolución temática de la investigación sobre Información y Documentación en español en la base de datos LISA mediante modelado temático (1978-2019). *Profesional de la información, 29*(4).

Ignatow, G., & Mihalcea, R. (2016). *Text Mining: A Guidebook for the Social Sciences.* Obtenido de https://academic.microsoft.com/paper/2748167422

Karakatsanis, I., AlKhader, W., MacCrory, F., Alibasic, A., Omar, M. A., Aung, Z., & Woon, W. L. (2017). Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems, 65*, 1-6.

Kim, Y.-M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal, 24*(4), 432-452. Obtenido de https://academic.microsoft.com/paper/2772572665

Kumar, A., & Singh, R. K. (2016). Web mining overview, techniques, tools and applications: A survey. *International Research Journal of Engineering and Technology (IRJET), 3*(12), 1543-1547.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Obtenido de https://academic.microsoft.com/paper/2108646579

Medrano, J. F. (2020). Agentes inteligentes para recuperación de información y analítica visual en big data. *XXII*

Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz).

Meystre, S., Savova, G., Kipper-Schuler, K., & Hurdle, J. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform, 17*(1), 128-144. Obtenido de https://academic.microsoft.com/paper/2114388055

Moro, S., Pires, G., Rita, P., & Cortez, P. (2019). A text mining and topic modelling perspective of ethnic marketing research. *Journal of Business Research, 103*, 275-285. Obtenido de https://academic.microsoft.com/paper/2736934374

Mughal, M. J. (2018). Data mining: web data mining techniques, tools and algorithms: an overview. *Information Retrieval, 9*(6).

Papoutsoglou, M., Mittas, N., & Angelis, L. (2017). Mining people analytics from stackoverflow job advertisements. *43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, (págs. 108-115).