Descubrimiento de Conocimiento en Bases de Datos

Lautaro Ramos, Esteban Schab, Ramiro Rivera, Cristhian Richard, Patricia Cristaldo, Juan Pablo Núñez, Giovanni Daián Rottoli, Juan Manuel Ríos, Soledad Retamar, Carlos Casanova, Anabella De Battista
Grupo de Investigación en Bases de Datos, Departamento Ingeniería en Sistemas de Información,
Fac. Reg. Concepción del Uruguay, Universidad Tecnológica Nacional
Entre Ríos, Argentina
{ramosl, schabe, riverar, richardc, cristaldop, nunezjp, rottolig,
riosj, retamars, casanovac,debattistaa}@frcu.utn.edu.ar

Leticia Cagnina, Norma Edith Herrera
Departamento de Informática, Universidad Nacional de San Luis, San Luis, Argentina
{lcagnina,nherrera}@unsl.edu.ar

Resumen

En la actualidad se generan diariamente grandes cantidades de datos de diversos tipos (e.g. textos, imágenes, audios y videos) generando nuevas fuentes de información que pueden ser aprovechadas para agregar valor al trabajo de las organizaciones. Particularmente el análisis automático de textos (análisis de sentimientos, minería de opinión) ha ganado terreno como alternativa o complemento a las fuentes de datos tradicionales de información de las organizaciones, cobrando relevancia las técnicas de Minería de Textos. La mayoría de los algoritmos, herramientas y recursos disponibles para Minería de Textos han sido probados y/o desarrollados para el idioma inglés, y por tanto presentan dificultades al ser empleados sobre textos escritos en otros idiomas como el español. Es por esta razón que es necesario trabajar en la elaboración de recursos específicos y en la adaptación de algoritmos y herramientas que contemplen las particularidades del idioma español con el fin de poder conseguir resultados de mayor calidad.

En este artículo se presentan los tópicos de interés del proyecto *Descubrimiento de Conocimiento en Bases de Datos*, en el que se investigan técnicas de minería de textos aplicables al procesamiento de textos en lenguaje español. En particular, se realizará el estudio, análisis y comparación de algoritmos de minería de textos utilizando corpus de textos en lenguaje español, para posteriormente proponer adaptaciones o mejoras a los mismos. Asimismo, se pretende evaluar el desempeño de técnicas de minería de datos sobre conjuntos de datos tradicionales complementados con información extraída a partir de textos relacionados.

Palabras clave: minería de datos, minería de textos, bases de datos, descubrimiento de conocimiento, idioma español.

Contexto

El presente trabajo se desarrolla en el ámbito del proyecto *Descubrimiento de conocimiento en Bases de Datos (Código 5109)* del Grupo de Investigación en Bases de Datos, perteneciente al Departamento Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay.

1. Introducción

El Descubrimiento de Conocimiento en Bases de Datos consiste en el análisis automático exploratorio y modelado de grandes repositorios de datos e involucra áreas de conocimiento como inteligencia artificial, aprendizaje automático, estadística, sistemas de gestión de base de datos, técnicas de visualización de datos y medios que apoyan toma de decisiones.

La Minería de Datos involucra e integra técnicas de diferentes disciplinas tales como

tecnologías de bases de datos y *data warehouse*, estadística, aprendizaje de máquina, computación de alta performance, computación evolutiva, reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, y análisis de datos espaciales o temporales.

Para aprovechar el conocimiento potencial que puede obtenerse a partir del análisis automático de textos surge la Minería de Textos, consistente en el proceso de extraer patrones relevantes a partir de un gran conjunto de textos con el propósito de obtener conocimiento. Abarca una serie de tareas que se enfocan en distintos aspectos del análisis de textos. Algunas de las más relevantes [1]:

- Recuperación de información (*Information Retrieval*, IR): es la tarea de encontrar material de naturaleza no estructurada (generalmente textos) proveniente de grandes colecciones que satisfagan determinadas necesidades de información [2]. Una tarea crucial para un sistema de IR es indexar la colección de documentos para hacer que sus contenidos sean accesibles de manera eficiente. Generalmente la indexación se realiza sobre una representación lógica del documento, que puede consistir en un conjunto de palabras clave o términos relevantes que aparezcan en el texto [3].
- Procesamiento del Lenguaje Natural: es un campo de las ciencias de la computación que combina Inteligencia Artificial y conceptos lingüísticos con el fin de hacer que oraciones o palabras escritas en lenguaje natural puedan ser interpretados por programas de computadoras [1, 4].
- Resumen de textos (*Text Summarization*): es la tarea de producir un resumen conciso y fluido preservando el contenido clave de la información y el significado general de una colección de textos [6].
- Extracción de Información (*Information Extraction*, IE): es una subdisciplina de la Inteligencia Artificial que se aboca a la identificación, y consecuente clasificación y estructuración en grupos semánticos, de información específica que se encuentran en fuentes de datos no estructurados, como el texto en lenguaje natural, lo que hace que la información sea más adecuada para las tareas de procesamiento de la información [5].
- Métodos de Aprendizaje Supervisado y No Supervisado: los métodos de aprendizaje supervisado son técnicas de aprendizaje automático relacionadas con entrenar un modelo, por ejemplo, de clasificación, utilizando un conjunto de datos de entrenamiento para realizar predicciones sobre datos desconocidos de antemano. Existe una amplia gama de métodos supervisados, como clasificadores de vecinos más cercanos, árboles de decisión, clasificadores basados en reglas y clasificadores probabilísticos.

Los trabajos sobre minería de textos en español que se presentan en la actualidad se enfocan principalmente en Análisis de Sentimientos o Minería de Opinión, en los cuales se evidencian dos enfoques: uno basado en el empleo de lexemas, y otro en técnicas de *Machine Learning*, para identificar los sentimientos expresados en los textos. En la gran mayoría de estos trabajos se utilizan recursos traducidos de forma automática generados para otros idiomas, como el inglés, lo cual manifiesta una escasez de recursos genuinos para el lenguaje español.

Existen también trabajos sobre perfilado de autor, en los que se menciona la dificultad de encontrar colecciones de textos adecuadamente etiquetados y con poco ruido [7]. A partir de eso se han producido trabajos tendientes al desarrollo de conjuntos de textos en español específicos para esta tarea [8,9].

Las técnicas de minería de datos de texto se han propuesto para tareas de estudios bibliográficos, simplificación de textos y etiquetado semántico. Se ha propuesto la aplicación de algoritmos de clasificación [10] para identificar automáticamente el dominio disciplinar de un nuevo texto académico en un repositorio bibliográfico mediante la construcción de lexemas compartidos en cada disciplina.

2. Líneas de Investigación, Desarrollo e Innovación

La línea de trabajo principal de nuestro proyecto de investigación es el estudio,

análisis y comparación de técnicas de minería de datos para el tratamiento de textos en español, el análisis de su desempeño en distintos escenarios y la propuesta de modificaciones o mejoras a las técnicas de minería de textos existentes para incrementar la calidad de los resultados en el tratamiento de textos en español. También está previsto realizar una evaluación del funcionamiento de técnicas de minería de datos sobre conjuntos "tradicionales" enriquecidos con atributos provenientes de textos relacionados.
Se espera poder realizar vinculaciones con empresas u organizaciones que puedan obtener beneficios de la aplicación de técnicas de minería de textos en español.

2.1. Análisis Bibliométrico

Se trabaja en análisis bibliométrico tradicional y alternativo, midiendo el impacto de publicaciones científicas en sus distintas modalidades de difusión. Actualmente se está elaborando un análisis cuantitativo de publicaciones de autores de instituciones argentinas en la bases de datos SCOPUS de Elsevier [11], accedida desde la Biblioteca Electrónica de Ciencia y Tecnología del Ministerio de Ciencia, Tecnología e Innovación Productiva de la Nación y a través de la API proporcionada por SCOPUS [12], utilizando scripts desarrollados en lenguaje R [13]. Para las búsquedas se establece una palabra o frase clave. En algunos casos se aplicó como filtro que las publicaciones correspondiesen a Argentina para identificar y reunir los trabajos en los que al menos uno de los autores incluyera la mención de una institución argentina en los datos de afiliación institucional, a fin de poder comparar con la cantidad de publicaciones del resto del mundo.
Se comenzó trabajando en índices tradicionales de análisis de publicaciones (conocido como Modo 1), que se basan principalmente en analizar las publicaciones realizadas en revistas con referato pagas. En la actualidad, se está trabajando en el denominado Modo 2, estudiando instituciones de pertenencia de los artículos provenientes de Scopus y además de fuentes como Altmetric. En este último caso se busca información publicada en blogs de ONGs, institucionales, etc. Está previsto realizar un análisis del impacto de publicaciones en redes sociales como Facebook o Twitter.

2.2. Agenda setting

El término Agenda Setting hace referencia a la influencia que tienen los medios de comunicación en la fijación de temas en la opinión pública [14].
Se comenzó a realizar un trabajo de medición de los efectos de la instalación de asuntos en la agenda pública tomando como base artículos escritos sobre diferentes temáticas en medios digitales de relevancia para determinar los tópicos que tratan y luego analizar su difusión en redes sociales empleando técnicas de minería de textos y procesamiento de lenguaje natural [15].

2.3. Transferencia tecnológica a industrias de la zona

En el marco de un convenio con una empresa local de desarrollo de software, se está realizando el desarrollo del prototipo para el procesamiento y posterior análisis de streaming de datos provenientes de las aplicaciones que esta empresa comercializa en bancos, con el objetivo de ofrecer información agregada para la toma de decisiones y que pueda retroalimentar dicha aplicación para automatizar ciertas decisiones a futuro.
En conjunto con el Grupo de Investigación y Desarrollo en Innovación y Competitividad del Departamento Licenciatura en Organización Industrial, de la FRCU-UTN, se encuentra en desarrollo el proyecto “Fortalecimiento de la Gestión productiva integral en PyMEs del sector metalmecánico del Parque Industrial de Concepción del Uruguay, Entre Ríos”. Dicho proyecto ha resultado seleccionado en el marco de la convocatoria “Agregando Valor” (edición 2017) de la Secretaría de Políticas Universitarias de la Nación, orientada a la presentación de proyectos de vinculación tecnológica de alto impacto con la finalidad

de transferir conocimientos y tecnologías innovadoras al sector socio-productivo nacional.

2.4. Visualización de datos

La generación y el almacenamiento de grandes volúmenes de información hacen que el mismo pase desapercibido y muchas veces se pierda la oportunidad de encontrar valor en ella. La visualización de datos es el proceso de representación de datos, en formato gráfico, de una manera clara y eficaz. Se convierte en una herramienta poderosa para el análisis e interpretación de datos grandes y complejos, volviéndose un medio eficiente en la transmisión de conceptos en un formato universal [16, 17].

En este proyecto se trabaja en el análisis de técnicas y herramientas de visualización de datos, para mejorar los procesos de comunicación de resultados de las actividades que desarrolla el grupo. A partir de la generación de visualizaciones de dichos resultados, se logra una mejor comprensión de los datos. Entre las herramientas utilizadas actualmente se encuentran Tableau[18], Gephi[19], D3js[20], React D3[21] y Shiny[22].

2.5. Aplicación de metodología para la gestión de proyectos de Minería de Datos

En la gestión de las actividades de cada una de las líneas de investigación y desarrollo del proyecto se emplean fundamentos de metodologías ágiles. [23, 24] Partiendo de la propuesta metodológica de CRISP-DM [25] se realizó una adaptación empleando dichos fundamentos ágiles. Se formalizó dicha adaptación como una propuesta de metodología ágil para la gestión de proyectos de ciencia de datos. [26, 27]

3. Resultados obtenidos y esperados

Con este proyecto se espera lograr aplicaciones novedosas de técnicas y herramientas de minería de textos para textos en español, en particular en áreas de estudio como bibliometría y la teoría de establecimiento de agenda. Estas iniciativas se desarrollan mediante la aplicación de la metodología ágil para proyectos de ciencias de datos propuesta. [26, 27]

4. Formación de Recursos Humanos

En el marco del proyecto se están desarrollando dos tesis de maestría y dos de sus integrantes están cursando carreras de doctorado. Se cuenta con un becario graduado con beca de iniciación a la investigación y dos becarios alumnos de la carrera Ingeniería en Sistemas de Información que inician su formación en la investigación. Está prevista la realización de al menos dos Prácticas Supervisadas en el marco del proyecto.

5. Referencias

[3] Allahyari, Mehdi et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. 2017. arXiv:1707.02919v2

[4] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to information retrieval. Cambridge University Press, 2008.

[5] S. Ceri, A. Bozzon, M. Brambilla, E. Della Valle, P. Fraternali, and S. Quarteroni, The Information Retrieval Process. In Web Information Retrieval, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–26.

[6] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural Language Processing: State of The Art, Current Trends and Challenges,” Aug. 2017.

[7] Moens, Marie-Francine. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer Netherlands, 2006.

[8] M. Allahyari et al. Text Summarization Techniques: A Brief Survey. Jul. 2017.

[9] M. J. Garciarena Ucelay, M. P. Villegas, L. Cagnina, and M. L. Errecalde. Cross domain author profiling task in spanish language: an experimental study. *JCS&T*, vol.

15, no. 02, pp. p. 122-128. Nov. 2015.

[10] M. P. Villegas, M. J. Garciarena Ucelay, M. L. Errecalde, and L. Cagnina. A Spanish text corpus for the author profiling task. 2014.

[11] M. J. Garciarena Ucelay and M. P. Villegas. Determinación del Perfil del Autor de Documentos en Español. Universidad Nacional de San Luis, 2015.

[12] R. Venegas. Clasificación de textos académicos en función de su contenido léxico-semántico. Rev. signos, vol. 40, no. 63, pp. 239–271, 2007.

[13] SCOPUS. http://www.scopus.com Accedido 03/2018.

[14] Scopus API. https://goo.gl/mqpFpA Accedido 03/2017.

[15] R Project. https://www.r-project.org/ Accedido 03/2018.

[16] M. McCombs and D. Shaw. The agenda-setting function of mass media. Public opinion quarterly, 36(2):176–187, 1972.

[17] Yeooul Kim, Suin Kim, Alejandro Jaimes, and Alice Oh. A computational analysis of agenda setting. In Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion), 323-324, 2014.

[18] Sadiku, Matthew (2016). Data Visualization. International Journal of Engineering Research And Advanced Technology(IJERAT).Volume. 02. Issue.12. p. 11-16.

[19] Finch, Jannette L., and Angela R. Flenner. 2017. “Using Data Visualization to Examine an Academic Library Collection.” College & Research Libraries 77(6). https://goo.gl/fAeW3w (March 18, 2018).

[20] Tableau. https://www.tableau.com/es-es Accedido 03/2018.

[21] Gephi. https://gephi.org/ Accedido 03/2018.

[22] Data-Driven Documents. https://d3js.org/ Accedido 03/2018.

[23] React D3. http://www.reactd3.org/ Accedido 03/2018.

[24] Shiny from R Studio. https://shiny.rstudio.com/ Accedido 03/2018.

[25] Ken Schwaber and Jeff Sutherland. The scrum guide. Scrum Alliance, 2011, vol. 21.

[26] Manifesto for Agile Software Development. Agile Alliance. https://goo.gl/xRFCVL . Accedido 03/2018.

[27] Chapman, Clinton, Kerber, Khabaza, Reinartz, Shearer, & Wirth. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

[28] Cristaldo, Richard, Rivera, Schab, Anabella De Battista. Propuesta Metodológica de Enfoque “Híbrido” para la Gestión de Proyectos de Minería de Datos. SABTIC 2018. ISSN 2237-2970

[29] Cristaldo, Schab, Richard, Rivera, Anabella De Battista, Retamar, Herrera. Adecuación de una Propuesta Metodológica de Enfoque “Híbrido” para la Gestión de Proyectos de Ciencia de Datos. 6to CoNaIISI. 29 y 30 de Noviembre de 2018 – Universidad CAECE – Mar del Plata, Bs. As., Argentina.