Text Mining aplicado a la Base de Datos del Servicio del 911 de la Provincia de Jujuy.

José Federico Medrano, Mario Alberto Tejerina, César Alejandro Castillo, Juan Carlos Rodríguez

jfmedrano@fi.unju.edu.ar, mariotejerina@gmail.com, ce_al_castillo@yahoo.com.ar, jcrodriguez@cegin.com.ar

VRAIn / Visualización y Recuperación Avanzada de Información / Facultad de Ingeniería

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

RESUMEN

La minería de texto o *text mining* la englobamos dentro de las técnicas y modelos de minería de datos o *data mining*, que es el análisis matemático para deducir patrones y tendencias que existen en los datos, patrones que no pueden detectarse mediante una exploración tradicional de los datos porque las relaciones son demasiado complejas o por el volumen de datos que se maneja. A diferencia del *data mining*, el *text mining* trabaja sobre datos no estructurados en grandes colecciones de datos.

La base de datos del Servicio del 911 de la provincia de Jujuy presenta un panorama muy interesante, ya que si bien la información que maneja dicho servicio es almacenada en una base de datos estructurada, la tabla principal que almacena los incidentes que son cargados por los tele-operadores posee un campo llamado descripción donde allí se vuelca la información más importante y los detalles del llamado de emergencia, información que no puede almacenarse en los campos del formulario de carga.

Este campo ofrece una alternativa muy interesante de exploración pues el contenido no estructurado de dicho campo no es analizado ni tenido en cuenta por los informes tradicionales que utilizan en la actualidad.

Palabras clave: *Text Mining; Clasificación de Información; Clustering; Data Mining; PLN*

CONTEXTO

La línea de investigación aquí presentada se encuentra inserta en el proyecto: *"Text Mining aplicado a la Base de Datos del Servicio del 911 de la Provincia de Jujuy."*, ejecutado a partir de enero de 2019 con una duración de 1 año. Dicho proyecto es llevado a cabo por el grupo de investigación *VRAIn* (Visualización y Recuperación Avanzada de Información) de la Facultad de Ingeniería de la Universidad Nacional de Jujuy. El proyecto se encuentra acreditado y financiado parcialmente por la Facultad de Ingeniería de la Universidad Nacional de Jujuy (Resolución FI N° 642/18)

1. INTRODUCCIÓN

Las cantidades de texto que se genera todos los días están aumentando drásticamente. Este tremendo volumen de texto, en su mayoría no estructurado, no puede ser simplemente procesado y percibido por las computadoras. Por lo tanto, se requieren técnicas y algoritmos eficientes y efectivos para descubrir patrones útiles. La minería de texto es la tarea de extraer información significativa del texto, esta especialidad ha ganado atenciones significativas en los últimos años (Allahyari, y otros, 2017)

Los datos de texto son un buen ejemplo de información no estructurada, que es una de las formas más simples de datos que se pueden generar en la mayoría de los escenarios. El texto no estructurado es procesado y percibido fácilmente por los humanos, pero es mucho más difícil de entender para las máquinas (Feldman & Sanger, 2007).

Los servicios públicos, por ejemplo, generan grandes cantidades de datos que pueden ser almacenadas en bases de datos estructuradas, tal es el caso del Servicio del 911 de la provincia de Jujuy. En este servicio, la Base de Datos principal cuenta con más de 5.000.000 de registros. Cada vez que se realiza un llamado a este servicio, es grabado, catalogado y registrado por un operario. El mismo le asigna un tipo de incidente y un color según la severidad del mismo, el color varía entre celeste, amarillo, verde y rojo, yendo de severidad baja a alta según corresponda.

La tabla principal donde se almacenan los llamados, posee una gran cantidad de campos donde el operario puede identificar el origen de la llamada, persona que llama, dirección, tipo de incidente, severidad, la dirección para hacer posible la georeferenciación entre muchos otros.

Un campo llamado *descripción* se emplea para dotar de mayor detalle la llamada recibida, al ser un campo libre y extenso, el operario se explaya con los detalles del llamado, por ejemplo puede introducir cuestiones como el origen o causa del incidente. Es decir, si una persona llama por violencia de género, el operario puede indicar que la causa del llamado se debe a que la persona que cometió la falta estaba en estado de ebriedad, en este caso el llamado queda asentado como "violencia de género", etiqueta de color "roja", pero no queda indicado en otros campos el origen del incidente, no queda registrado de forma paramétrica que la persona que inició el hecho violento se encontraba ebria, salvo el texto en el campo descripción.

Poder analizar dicho campo, empleando técnicas de minería de texto y de clasificación de información, permitirían obtener muchísimos detalles hasta ahora descuidados, además se ofrecería información que los reportes habituales y estadísticos que se obtienen desde la plataforma que emplean lo pasan por alto totalmente, al ser un campo adicional y al ser de "solo texto", la estadística descriptiva no es capaz de analizarlo.

La minería de texto ha sido ampliamente utilizada para hallar relaciones y patrones entre datos no estructurados (Satyabrata, Mangal, Jinse, Ki-Won, & Hee-Cheol, 2017; Zhong, Li, & Wu, 2012; Kwon, Kim, & Park, 2017), el empleo de técnicas de Procesamiento del Lenguaje Natural es una de las principales tecnologías que favorecen en gran medida el procesamiento de este tipo de información (Kao & Poteet, 2007; Thessen, y otros, 2018; Joshi, Macwan, Mistry, & Mahida, 2018), es por ello que los sistemas basados en este enfoque aportan un valor agregado a los datos que manejan. En ese mismo sentido, técnicas de inteligencia artificial como aprendizaje automático y aprendizaje profundo, han sido utilizadas en combinación a la minería de datos y de textos

para procesar información relacionada con la seguridad y cyber-seguridad (Dua & Du, 2016).

Clasificar los llamados atendiendo a este campo, rico en información, dará acceso a un conjunto de palabras claves, indicadores y datos que complementarán los informes periódicos, esta "nueva información" (que no es nueva porque siempre estuvo presente pero no en un formato accesible) abrirá la posibilidad de manejar informes detallados y específicos, atendiendo no solo a la causa u origen del llamado origen, sino también a las relaciones entre los eventos/sucesos. Este tipo de relaciones tiene que ver con los patrones que se logren identificar (Wang, Rudin, Wagner, & Sevieri, 2015).

Las técnicas de *text mining* implican el uso de técnicas de inteligencia artificial como el PLN (MANNING & SCHÜTZE, 1999) y Aprendizaje Automático (*Machine Learning*) (Witten, Frank, Hall, & Pal, 2016). Estas técnicas permiten por un lado extraer datos relevantes y por el otro encontrar patrones y relaciones entre los datos. Para ello es necesario desarrollar un modelo de predicción, separar los conjuntos de datos en datos de entrenamiento, prueba y validación para asegurar que las clasificaciones y agrupaciones sean correctas.

Una vez que los datos sean procesados, los modelos entrenados y validados, se empelarán técnicas de inteligencia artificial para visualizar las relaciones encontradas. Puesto que se parte de la hipótesis de que estos datos sin analizar, los datos contenidos en el campo descripción, aportarán nueva información a los reportes e informes que se utilicen. Como se mencionó anteriormente, esta nueva información no reemplazará el manejo actual sino que lo complementará y potenciará.

Interfaces basadas en texto requieren esfuerzo cognitivo para entender su contenido informativo. La Visualización de Información (InfoVis) tiene por objeto presentar la información visualmente, en esencia, para reducir la carga de trabajo cognitivo al sistema perceptivo visual humano (Ware, 2004; Ware, 2008).

La Visualización de Información abarca las técnicas de visualización que tienen que ver principalmente con datos abstractos, es decir, los datos para los cuales el usuario no tiene un modelo mental preconcebido. Por esta razón, la interacción es especialmente importante en InfoVis, ya sea para la exploración, análisis y/o presentación de los datos (Kosara, Hauser, & Gresh, 2003). La interacción permite al usuario implícitamente formar modelos mentales de las correlaciones y las relaciones entre los datos, a través del reconocimiento de patrones.

La utilización de representaciones estáticas como imágenes o gráficos sin interacción han quedado atrás, el usuario de hoy en día necesita interactuar con la representación presentada, la herramienta de soporte debe brindar las facilidades para intercambiar formas, colores y modos de representar la misma información, con lo cual la visualización de información en este proyecto no debe ser descuidada puesto que aportará una visión complementaría al análisis que se pretende realizar.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Esta línea de investigación propone estudiar, evaluar y aplicar técnicas de Minería de Textos para analizar la información no estructurada de la Base de Datos del Servicio del 911 de la provincia de Jujuy. Para ello se pretende emplear técnicas de Procesamiento del Lenguaje Natural para hallar relaciones ocultas entre los datos y no visibles con simples informes o consultas SQL. Las relaciones permitirán conocer y formar una nueva imagen a partir de los datos existentes, identificando en muchos casos el origen o la causa de un llamado y no solo la consecuencia del suceso.

3. RESULTADOS OBTENIDOS/ESPERADOS

El presente proyecto tiene como objetivo general analizar, profundizar y rastrear información no estructurada almacenada en la base de datos del Servicio del 911 de la provincia de Jujuy, en busca de datos que aporten mayor detalle y definición a los reportes e informes estadísticos que el sistema actual genera.

Particularmente se espera lograr:

- Analizar la base de datos actual en busca de información no estructurada y que no es tenida en cuenta en los reportes que el sistema entrega.
- Extraer características de datos no estructurados utilizando técnicas de Procesamiento del Lenguaje Natural (PLN).
- Diseñar, testear y calibrar esquemas de PLN en base a las características encontradas.
- Diseñar modelos de clasificación y agrupación de información mediante la aplicación de técnicas de Aprendizaje Automático.
- Testear, calibrar y comparar los distintos modelos de clasificación y agrupación diseñados.
- Representar las distintas relaciones y/o patrones entre los datos procesados mediante el empleo de técnicas de visualización de información.

4. FORMACIÓN DE RECURSOS HUMANOS

El presente proyecto está a cargo del Dr. José Federico Medrano como director, además cuenta con el apoyo de los docentes investigadores: Ing. César Castillo, Ing. Mario Tejerina y el Analista Programador Universitario Juan Carlos Rodríguez.

Este proyecto brinda un marco para que docentes auxiliares y estudiantes lleven a cabo tareas de investigación y se desarrollen en el ámbito académico.

Actualmente, se están dirigiendo dos trabajos finales de la carrera Ingeniería Informática (UNJu y UCSE DASS) relacionadas con la temática propuesta, tres alumnos pasantes colaboran y llevan a cabo tareas a lo largo de todo el proyecto y además participan en un proyecto de vinculación de la SPU, con el objetivo de iniciarse en la investigación.

5. BIBLIOGRAFÍA

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint*(arXiv:1707.02919).

Dua, S., & Du, X. (2016). *Data mining and machine learning in cybersecurity.* Auerbach Publications.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* Cambridge university press.

Joshi, B., Macwan, N., Mistry, T., & Mahida, D. (2018). Text Mining and Natural Language Processing in Web Data Mining. *2 International Conference on Current Research Trends in Engineering and Technology*, *4*, págs. 392-394.

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining.* Springer Science & Business Media.

Kwon, H., Kim, J., & Park, Y. (2017). Applying LSA text mining technique in envisioning social impacts of emerging

technologies: The case of drone technology. *Technovation, 60*, 15-28.

MANNING, C., & SCHÜTZE, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

Satyabrata, A., Mangal, S., Jinse, P., Ki-Won, C., & Hee-Cheol, K. (2017). A text mining approach to identify the relationship between gait-Parkinson's disease (PD) from PD based research articles. *Proceedings of the International Conference on Inventive Computing and Informatics.*

Thessen, A., Preciado, J., Jain, P., Martin, J., Palmer, M., & Bhat, R. (2018). Automated Trait Extraction using ClearEarth, a Natural Language Processing System for Text Mining in Natural Sciences. *Biodiversity Information Science and Standards*.

Wang, T., Rudin, C., Wagner, D., & Sevieri, R. (2015). Finding patterns with a rotten core: Data mining for crime series with cores. *Big Data, 3*(1), 3-21.

Ware, C. (2004). *Information Visualization - Perception for Design.* Morgan-Kaufmann.

Ware, C. (2008). *Visual Thinking for Design.* Morgan Kaufman/Elsevier.

White, H., & McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology, 24*, 119-186.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann.

Zhong, N., Li, Y., & Wu, S. T. (2012). Effective pattern discovery for text mining. *IEEE transactions on knowledge and data engineering, 24*(1), 30-44.