VISUALIZACIÓN EN UN ENTORNO DE MINERÍA DE TEXTO

Laura Viviana Gutierrez, Graciela Beguerí, María Alejandra Malberti, Raúl Oscar Klenzi, Manuel Ortega

Departamento de Informática – Facultad de Ciencias Exactas Físicas y Naturales. Universidad Nacional de San Juan

Ignacio de la Roza y Meglioli. Complejo Universitario Islas Malvinas CUIM. Rivadavia, San Juan, Argentina

{gutierrez.laura, grabeda, amalberti, rauloscarklenzi, manuel.ortega}@gmail.com

RESUMEN

En el marco de Ciencia de Datos, se propone analizar y caracterizar diferentes estrategias y herramientas de minería de texto, según sus potencialidades de Visualización de Información. Estas se aplicarán a conjuntos de datos obtenidos desde los planes de estudios de las carreras de Informática, disponibles en el Departamento de Informática de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan DI-FCEFN-UNSJ. Se considerarán herramientas de software libre, particularmente RapidMiner y Knime para la visualización de la información obtenida aplicando técnicas de Textmining.

Considerando al usuario como el destinatario del proceso de búsqueda de conocimiento en datos, se investigará sobre aspectos de interpretación y percepción.

Palabras clave: Visualización, Visualización de la Información**,** Visualización de Texto, Minería de Texto, Textmining.

CONTEXTO

La línea de investigación se encuentra en el ámbito de los proyectos "Ciencia de los Datos aplicada a grandes colecciones de datos" ejecutado en el bienio 2016_2017 y la continuidad, buscada por el grupo de investigadores que lo conformaban, en la presentación de un proyecto para el bienio 2018_2019 actualmente en evaluación "Visualización y Deep Learning en Ciencia de los Datos"

En este último se pretende evaluar software libre apropiado al área Ciencia de Datos, indagar en la temática de Deep Learning, investigar sobre aspectos de interpretación y percepción relacionados con mecanismos de visualización de información, así como herramientas para implementar los mismos, en el marco de búsqueda de conocimiento en datos. Además de caracterizar a los usuarios de acuerdo a las potencialidades de las herramientas analizadas.

En este contexto, la presente propuesta de trabajo y línea de investigación se centra en formas de visualizar la información generada a partir de los datos de los planes de estudio de las

carreras de informática del DI-FCEFN-UNSJ, aplicando técnicas de visualización.

1. INTRODUCCIÓN

Según Strecker, Card y otros la Visualización de Información puede ser tratada como un campo de conocimiento bien establecido, asociado al uso de representaciones visuales de datos abstractos que tienen como fin expandir el conocimiento. Para Cairo, A. es una tecnología plural que consiste en transformar datos en información semántica o en la creación de herramientas para dicha transformación, basada en la combinación de señales de naturaleza icónica con otros de naturaleza arbitraria y abstracta (textos, estadísticas, etc.).

La minería de textos es un subcampo de la minería de datos, encontrándose entre sus aplicaciones analizar o comparar textos. Las técnicas de visualización mejoran la comunicación de los resultados obtenidos. En el proceso de análisis se produce la modificación del texto original, también llamado dato no-estructurado, por ejemplo reduciendo un texto a una lista de palabras de acuerdo con su frecuencia. Es así que, la mayoría de las visualizaciones de textos transforman los datos de tipo textual o no estructurados en un nuevo conjunto de datos estructurados, y reducidos, respecto al texto original. Este nuevo conjunto de datos ya no es unidimensional, sino que puede estar ordenado por categorías o con una estructura de red.

En la actualidad se puede contar con varias herramientas libres que ayudan a convertir datos en gráficos. Estas pueden ser usadas desde usuarios principiantes hasta usuarios avezados. En Wikipedia se presentan algunas de ellas, al igual que en la figura 1, situada en https://www.bbva.com/wp-content/uploads/2017/10/ebook-cibbva-visualizacion_de_datos-es.pdf.

Fig.1 – Algunas herramientas para Visualización de datos.

José Mondragón de IBM developerWorks expresa que, en las empresas se generan mucha información estructurada así como datos no estructurados. Opina que combinar los análisis con datamining para ambos tipos de información puede ayudar a incrementar la rentabilidad y la participación en el mercado. En la figura 2 se muestra el diagrama de cómo mejorar el desempeño de un modelo predictivo:

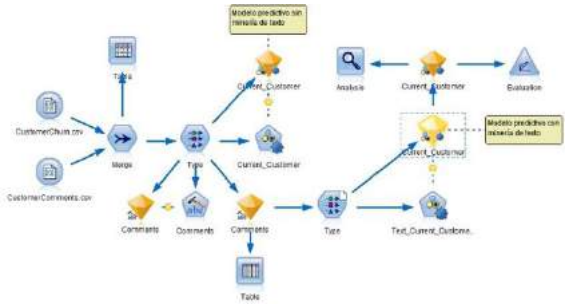

Fig. 2 – Modelo predictivo usando minería de texto.

Es importante tener en cuenta que todo proceso de minería de texto debería culminar con la confección de un

gráfico mediante el cual se visualice la información de manera rápida y clara, tal como lo indica Felipe de Jesús Núñez Cárdena

El punto de partida en el tratamiento de visualizaciones de textos, es reconocer dos grandes categoría de textos: Textos individuales y Colecciones de textos, y a la vez consideran distintos tratamientos de visualizaciones relacionados con cada una de las categorías.

1 - Visualización de textos individuales

Los métodos de visualización de un texto completo suelen utilizar el color como elemento distintivo, variaciones de diagramas de barras, curvas que conectan partes del texto entre otros recursos. Una problemática detectada es la variedad en cuanto a la naturaleza de los textos, por lo que los autores mencionados consideran que es una buena idea trabajar con textos más estructurados y sencillos, con un vocabulario más regular, una longitud del texto estandarizada, con una clara estructura del discurso y corrección en el lenguaje (artículos científicos, textos de patentes, diagnósticos de salud, etc.).

Si la problemática de visualización se traslada a partes de un texto, un método habitual es el llamado bag of words o bolsa de palabras

Asimismo, métodos estadísticos muy simples, como la frecuencia de palabras, pueden tener un resultado fácil de entender. También una lista de varios tamaños de palabras es una forma directa de comunicarse con cualquier usuario, ya sean estos principiantes o expertos.

2 - Visualización de colecciones

Una vez determinadas las colecciones, los datos se pueden considerar como un caso general de visualización de datos. Es así que se emplean métodos utilizados en otros campos como ser visualizaciones de red, líneas temporales, ítems de nubes de palabras -para el caso de comparar textos, y el color para detectar palabras nuevas (Jaume Nualart-Vilaplana, Mario Pérez-Montoro y Mitchell Whitelaw).

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

En el marco del proyecto que contiene la presente línea de investigación se pretende, tomar el trabajo realizado en la tesis de maestría de la Lic. Laura Gutiérrez. Este versa sobre proponer una metodología automática para determinar las pertinencias sintáctico-semáticas entre los contenidos mínimos de carreras de informática con las normas establecidas según las diferentes titulaciones asignadas a las mismas.

El proceso automático agrupa los contenidos mínimos de las carreras estableciendo el grado de similitud sintáctica de estos, con las áreas u objetivos temáticos de las resoluciones o marcos regulatorios correspondientes, favoreciendo así, un análisis inicial de planes de estudio que pueda llevar a la decisión de hacer correcciones sobre los mismos.

La aplicación se llevó adelante mediante la utilización los módulos de modelado y minería de texto (TextMining -TM-) de la herramienta de software libre RapidMiner (RM) versión 5.3.15. El caso de estudio considerado es la carrera Licenciatura en Ciencias de la Computación del DI-FCEFN-UNSJ.

Ese trabajo se ha llevado adelante mediante la utilización de herramientas de software libre del área de Data mining - DM. En este caso los resultados que se presentaron se alcanzaron mediante el uso de de algoritmos de DM que posee la herramienta RapidMiner versión 5.3.15 con la opción de visualizarlos también.

Desarrollo

El trabajo citado se realizó básicamente en tres pasos:

1_ Preprocesamiento y Análisis de Documentos de Texto (Planes de Estudio)

2_ Visualización de Contenidos Faltantes: El resultado del análisis a los datos, permitió obtener los contenidos faltantes en los documentos de texto utilizados, que con los documentos de texto a simple vista no se observaban (Contenidos mínimos).

3_ Determinación de pertinencias sintácticas (Planes de Estudios-Ordenanzas)

3. RESULTADOS OBTENIDOS/ESPERADOS

Los resultados obtenidos en este trabajo en particular, permitieron visualizar los contenidos faltantes en los documentos de texto y llevaron a sugerir cambios en los mismos obteniendo mejores resultados en un posterior análisis.

Se pretende desde este proyecto aplicar distintas herramientas de visualización en diferentes conjuntos de datos textuales, con el propósito de extraer rápida y sencillamente el conocimiento.

4. FORMACIÓN DE RECURSOS HUMANOS

El equipo de investigación se encuentra conformado por: una directora, un co-director, cuatro docentes investigadores (categorizados en el Programa de Incentivos de la Secretaria de Políticas Universitarias (SPU) perteneciente al Ministerio de Educación de la Nación Argentina), un egresado y cuatro alumnos de los últimos años de las carreras del Departamento de Informática.

En el marco de esta investigación se desarrollarán trabajos finales para la/s carrera/s LCC-LSI del DI.

La ejecución del proyecto incidirá directamente en una formación más profunda de los integrantes del equipo de investigación. Este aspecto beneficiará de manera directa a las carreras del Departamento de Informática, pues las temáticas abordadas están vinculadas con las materias en las cuales se desempeñan los integrantes de este proyecto. También estas líneas de trabajo servirán al medio para proveer nuevas estrategias de administración y presentación del conocimiento.

5. BIBLIOGRAFÍA

- Cairo, A. El arte funcional: infografía y visualización de información. Alamut, 2011. ISBN:9788498890679.
- Card, S. K.; Mackinlay, J. D.; Shneiderman, B. Readings in information visualization: using vision to think. Morgan Kaufmann, 1999. ISBN-10:1558605339.
- De Lucia Castillo, F., & Saibel Santos, C. A. (2016). Nubes de palabras animadas para la visualización de información textual de Publicaciones Académicas.
- Dürsteler, J. C. (2000). Visualizaciónde información. *Gestion*.
- El Cairo, A. (2012). *El arte funcional: una introducción a*

gráficos de información y visualización. Nuevos jinetes.

- Gutiérrez, L., Klenzi, R., Malberti, A., Beguerí, G., Pinto, T., & Araya, J. (2015). Análisis de Planes de Estudio Mediante Determinación Automática de Pertinencias Sintáctico-Temáticas en Carreras de Informática. *Revista EletrôNica Argentina-Brasil De Tecnologias Da InformaçãO E Da ComunicaçãO, 1*(2). doi:10.5281/zenodo.59453
- Hernández Orallo J., Ramirez Quintana, J, Ferri Ramirez, C. (2008) *Introducción a la Minería de Datos*. Pearson-Prentice Hall.
- https://es.wikipedia.org/wiki/Visualizaci%C3%B3n_de_datos
- Kargupta, H., Han, J., Philip, S. Y., Motwani, R., & Kumar, V. (Eds.). Next generation of data mining. CRC Press. (2008).
- Larose, D. (2006) *Data Mining. Methods and Models.* Department of Mathematical Sciences Central Connecticut State University Wiley. A John Wiley & Sons, Inc Publication.
- Larose, D. (2005) Discovering Knowledge In Data - An Introduction to Data Mining. John Wiley & Sons, Inc., Publication.
- Miller, T. W. (2005). *Data and text mining: A business applications approach* (pp. 917-2199). New Jersey: Pearson Prentice Hall.
- Nualart-Vilaplana, J., Pérez-Montoro, M., & Whitelaw, M. (2014). Cómo dibujamos textos: Revisión de propuestas de visualización y exploración textual. *El profesional de la información*, *23*(3), 221-235.
- Strecker, J. Data visualization in review:summary. Tech. Report IDRC, 2012. URL: http://idlbnc.idrc.ca/dspace/bitstream/10625/49286/1/IDL-49286.pdf