ISSN 2007-9737

Visualización en un entorno de minería de datos desde una perspectiva interacción humano computador

Raúl Oscar Klenzi, María Alejandra Malberti, Graciela Elida Beguerí

Universidad Nacional de San Juan,
Departamento de Informática, Facultad de Ciencias Exactas Físicas y Naturales,
Argentina

{rauloscarklenzi, amalberti, grabeda}@gmail.com

Resumen. Con el propósito de apoyar el proceso de diseño, análisis y evaluación de los mecanismos de visualización de resultados que provean información, en entornos de Minería de Datos, este trabajo trata la problemática de la visualización, desde un escenario de Interacción Humano Computador. En él se consideran aspectos relevantes surgidos del estudio de la percepción humana. Se describen tres ejemplos prácticos, sobre datos numéricos, de texto y georreferenciados, valiéndose de la herramienta KNIME Analytics. Asimismo, se expone la utilidad e importancia de los gráficos para una correcta interpretación de la información.

Palabras clave. KNIME, visualización, minería de datos, interacción humano computador.

Visualization in a Data Mining Environment from a Human Computer Interaction Perspective

Abstract. With the aim of providing support to the design, analysis and evaluation of result visualization mechanisms used to supply information in Data Mining Environments, this work analyzes the visualization issue from a Human Computer Interaction setting. Important considerations arisen from the study of human perception are considered. Three practical examples based on numerical, textual and georeferenced data are described by means of the KNIME Analytics tool. In addition, the use and importance of graphs are emphasized for a correct information interpretation.

Keywords. KNIME, visualization, data mining, human computer interaction.

1. Introducción

La búsqueda de conocimiento en datos por medio de la aplicación de Minería de Datos, es un área ampliamente difundida y en expansión a raíz del incremento constante en la generación de datos. Uno de los retos de la visualización radica en presentar una buena información para lograr el interés y análisis de los usuarios. En la actualidad, la disponibilidad de diversos entornos y herramientas de aprendizaje de máquina, análisis de algoritmos y procesamiento de la información que ofrecen diversas funcionalidades, como por ejemplo KNIME, RapidMiner y Weka, entre otras, plantea otra problemática que es la de tener que elegir cuál usar en una situación particular, además de explicitar formas de presentar conclusiones desde las cuales poder extraer conocimiento de una manera más adecuada, rápida y efectiva.

En este punto es pertinente considerar los mecanismos de visualización de resultados provistos, como por ejemplo grafos, gráficos de barras, una matriz de confusión, entre otros. Una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo entre clases [11].

Cuando la cantidad de datos supera la capacidad cognitiva, la información debería ser

volcada en una representación basada en gráficos que permita transformarlos en información útil, visual y accesible para el destinatario. Atento a que es la vista el órgano sensorial que mayor cantidad de información puede captar.

En las secciones 2 y 3, se proporciona una introducción a Visualización y a Interacción Humano Computador. En la sección 4, se presentan tres ejemplos prácticos en los que se aplican diferentes instancias de visualización conforme la tipología de datos utilizada, y particularizando en la utilización de la herramienta de software KNIME Analytics 3.2.1.

2. Visualización

El escritor y diseñador, McCandless en [9], expone que la problemática sobre la sobrecarga de información o exceso de datos que se sufre hoy en día es acarreada por el uso de la tecnología Web. Afirma que existe una solución fácil que radica en la visualización de la información. Por medio de ésta se pueden ver patrones y conexiones, de modo que la información tiene más sentido y permite centrarse únicamente en lo que es importante.

Visualización, para Yuk y Diamond en [12], es el estudio de cómo representar los datos usando un enfoque visual o artístico en lugar del método tradicional de reportes, siendo dos de los tipos más populares el tablero y la infografía. Estos autores consideran que ser útil, usable y deseable son características recomendables de toda visualización, y a ellas les agregan: ser visualmente atractivas, escalables, proveer al usuario de información correcta y ser accesibles en cualquier momento. Es muy importante tener en cuenta ”qué” datos mostrar y ”cómo” mostrarlos. Ignasi Alcalde, [1], señala las diferencias y similitudes entre infografía y visualización. Al respecto aduce que “la visualización de la información es el estudio de la representación visual de los datos abstractos (interactivos), para reforzar la cognición humana, e incluyen tanto datos numéricos y no-numéricos. Como texto e información geográfica. . .

La infografía implica la presentación de información de forma estática, mientras que la visualización de datos tiene como eje principal el componente dinámico y la posibilidad de interacción y exploración de los mismos”.

En la Figura 1, se presenta el pipeline de visualización propuesto en [10].

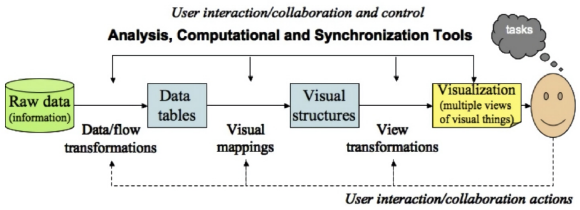

Fig. 1. Pipeline de visualización

El punto de partida es el procesamiento de los datos en bruto en algo utilizable por un sistema de visualización (interpolación por ejemplo para datos faltantes, o en el caso de datos de gran tamaño podría recurrirse a muestreo, filtrado, agregación o particionamiento). Una vez que los datos están preprocesados, se elige una representación visual específica en lo que refiere a geometría, color y sonido, por ejemplo. La etapa final implica el mapeo de datos geométricos a imagen. En [12], se propone el uso de dos medidas: Expresividad y Eficacia, para estimar la información transferida. Una visualización expresiva presenta toda la información, y solo la información, mientras que una visualización es eficaz cuando se puede interpretar con precisión y rapidez.

En el extremo derecho del pipeline se encuentra el ser humano; es importante entonces, abordar aspectos tratados en Interacción Humano Computador.

3. Interacción humano computador

Cuando las personas y los ordenadores interactúan lo hacen por medio de una interfaz. “La interfaz es el punto en que las personas y los ordenadores se ponen en contacto y se transmiten mutuamente tanto información, órdenes y datos como sensaciones, intuiciones y nuevas formas de ver las cosas” [4].

A la vez se menciona la definición comúnmente aceptada de la disciplina Interacción Persona

ISSN 2007-9737

Computador, según el Grupo SIGCHI (Special Interest Group on Human-Computer Interaction), creado en el ámbito de la ACM (Association for Computer Machinery), como "la disciplina relacionada con el diseño, la evaluación y la implementación de sistemas informáticos interactivos para uso de seres humanos, y con el estudio de los fenómenos más importantes con los que está relacionado" [4].

Se debe también proveer un entendimiento de la forma en que los usuarios trabajan, las tareas que necesitan ejecutar y la forma en que los sistemas computacionales necesitan ser estructurados para facilitar el logro de dichas tareas [2].

Un aspecto importante a ser considerado cuando hay individuos involucrados, es lo relativo a la percepción humana. Los primeros estudios de la percepción se centraron en la visión y sus capacidades, mientras que enfoques posteriores consideraron los problemas cognitivos y de reconocimiento.

Los seres humanos perciben los datos que se les presentan a través de visualizaciones. Se estudia entonces la percepción para mejorar la presentación de los datos.

La mayoría de las definiciones y teorías de la percepción la consideran como el proceso de reconocimiento (ser consciente de), organización (recopilar y almacenar), e interpretación (la construcción de conocimiento), de la información sensorial. La percepción es el proceso mediante el cual el ser humano interpreta el mundo que lo rodea formando una representación mental del entorno, que no es isomorfa con el mundo real [10].

La percepción tiene en cuenta los sentidos que detectan señales del entorno, tales como vista, oído, tacto, olfato y gusto. De ellos la visión y audición son los más tratados. Algunas propiedades perceptivas, tales como color, textura y movimiento, han sido usadas en visualización.

Ward, Grinstein y Keim mencionan también tres tipos de memoria que consideran relevantes en el estudio de la percepción en visualización. La memoria sensorial, este aprendizaje es físico y puede ser aprovechado por acciones repetidas, por ejemplo posiciones de teclas. La memoria de corto plazo, que puede ser aprovechada mediante la agrupación y la repetición. Y la memoria a largo plazo que puede ser aprovechada por el uso de asociación mnemotécnica y fragmentación. En el diseño de visualizaciones es importante considerar diversas características de los seres humanos, relacionadas con su capacidad de percibir, para evitar por ejemplo la generación de imágenes con información ambigua, engañosa o difícil de interpretar [10].

Para poder medir y comparar el rendimiento de la percepción humana sobre diversos fenómenos se necesita una métrica, es decir, una medida para evaluar de forma fiable. Para ello se han realizado numerosos estudios experimentales que han permitido obtener diferentes resultados.

Para cada estímulo primitivo, ya sea visual, auditivo, de gusto, tacto u olfato, se midió el número de niveles distintos que un sujeto promedio puede identificar con un alto grado de precisión. Una reseña de los resultados obtenidos están planteados en [3] y [10].

Para los estímulos estudiados, se observó que los usuarios no podían extraer más de 6 o 7 niveles, de un valor de datos con exactitud. La combinación de más de un estímulo, como por ejemplo punto en un cuadrado, salinidad y dulzura, sonoridad y tono entre otros, según experiencias, permitiría aumentar la cantidad de información que se comunica, pero no a los niveles esperados respecto a los atributos individuales. Esto concuerda con la teoría lingüística que identifica de 8 a 10 dimensiones, cada una de las cuales puede ser clasificada solo en dos o tres categorías.

Como consecuencia de los experimentos de percepción con atributos gráficos, resulta que los gráficos de barras y los diagramas de dispersión son herramientas eficaces para la comunicación de datos cuantitativos ya que ambos dependen de la posición a lo largo de una escala común, y a la vez sugieren que los gráficos circulares no son representaciones convenientes. De igual modo para aplicaciones donde se requiere un juicio absoluto, lo mejor es usar gráficas con un solo atributo que tenga entre 4 y 7 valores.

Para obtener un mayor rango de niveles reconocibles se debería considerar múltiples dimensiones, construir secuencias de decisiones

simples, o llevar a cabo algún tipo de fragmentación [10].

Según Iliinsky y Steele, en [5], es útil pensar en una eficaz visualización de datos consistente en un taburete de tres patas: el diseñador, el lector y los datos. Donde cada una de ellas ejerce una fuerza, o aporta una perspectiva independiente, que debe ser considerada para que una visualización sea estable y exitosa.

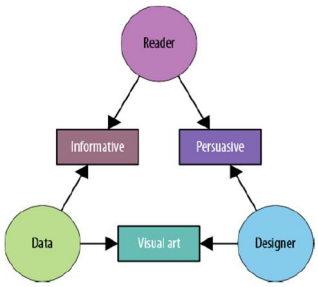

Fig. 2. La naturaleza de la visualización depende de qué relación, entre dos de los tres componentes, es dominante

Cada una de las tres patas tiene una relación única con las otras dos. Si bien es necesario dar cuenta de las necesidades y la perspectiva de los tres componentes en cada proyecto de visualización, la relación dominante determinará en última instancia qué categoría de visualización es necesaria, Figura 2 [5].

4. Aplicaciones

Seguidamente se presentan algunas instancias que la herramienta de minería de datos KNIME Analytics, tiene a modo de visualización interna asociada al preprocesamiento de datos, como así también los módulos para una visualización final de resultados. En este sentido se habrán de considerar tres ejemplos de trabajo con diferentes tipos de datos, destacando en cada uno de ellos las propuestas que consideramos pertinentes para mejorar la interacción humano computador (en este caso la comprensión de las conclusiones a las que la aplicación arriba).

KNIME (Konstanz Information Miner), es un entorno de minería de datos construido bajo la plataforma Eclipse y programado esencialmente en Java, que fue desarrollado originalmente en el departamento de bioinformática y minería de datos de la Universidad de Constanza, Alemania. Es una herramienta gráfica para el desarrollo de modelos en un contexto visual, que dispone de una serie de nodos (que encapsulan distintos tipos de algoritmos), y flechas (que representan flujos de datos), que se despliegan y combinan de manera gráfica e interactiva. Se utiliza KNIME Analytics 3.2.1 bajo licencia GNU versión 3, por sus características y potentes funcionalidades, cuya interfaz de trabajo se puede visualizar en la Figura 3, [7].

Fig. 3. Entorno de software KNIME Analytics 3.2.1

4.1. Reconocimiento automático de dígitos manuscritos

Se necesita un sistema de información en condiciones de reconocer, en forma automática, los dígitos decimales escritos en forma manual sobre un pad de 28*28 pixeles, https://www.kaggle.com/c/digit-recognizer/data.
Para ello se provee un conjunto de datos de entrenamiento (42000 registros), y otro conjunto para posterior testeo (28000 registros), contra los que habrá de validarse el modelo. Cada imagen es de 784 píxeles en total. Cada píxel tiene un único valor asociado que indica la claridad u oscuridad de ese pixel. Este valor de pixel es un número entero entre 0 y 255, inclusive; los números más altos significan color más oscuro.

ISSN 2007-9737

El conjunto de datos de entrenamiento cuenta con 785 atributos. Uno de ellos, llamado ”label”, es el dígito que fue dibujado por el usuario. El resto de las columnas contiene los valores de píxel modificado por el trazo de la imagen asociada, como lo muestra parcialmente la Figura 4.

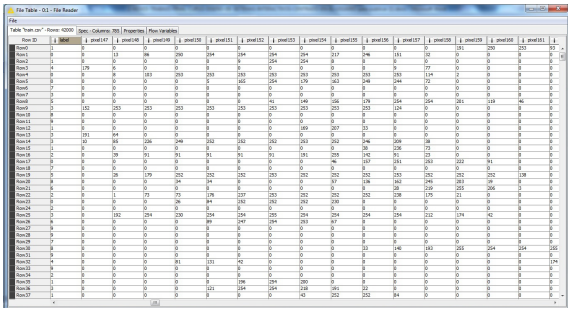

Fig. 4. Datos de entrenamiento

El diagrama en bloques de la Figura 5, esquematiza la tarea a realizar para el reconocimiento de dígitos decimales manuscritos.

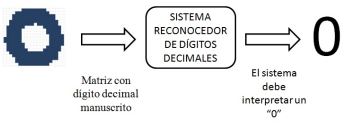

Fig. 5. Bloque de un Sistema Informático en condiciones de reconocer dígitos decimales manuscritos

A los efectos de encontrar un modelo que permita el reconocimiento automático de caracteres manuscritos, primero debe tenerse un conjunto de referencia desde donde realizar la tarea de entrenamiento y simultáneamente testeo de los resultados obtenidos.

Particularmente, atendiendo a que se considera una fuente de datos de 42000 registros etiquetados, se tomará de ellos el 80 % de entrenamiento y el restante 20 % de testeo, aplicando el concepto de validación cruzada para legitimar el algoritmo utilizado para modelar ese conjunto de datos o registros. Así mismo el desafío planteado en esta aplicación es encontrar, tras haber obtenido un modelo que se considera posee una buena aproximación descriptiva de los datos entrenados, las salidas que el modelo propone para los 28000 registros que carecen de etiquetas.

La sucesión de módulos que en la herramienta KNIME Analytics permite realizar esta operación se visualiza en la Figura 6. El algoritmo de aprendizaje utilizado es el de árboles de decisión, basado en ganancia de información relativa. La fracción de la izquierda presenta el conjunto de módulos que hacen la lectura de los datos, tareas de preprocesamiento, y la instancia de cross-validación, ampliada esta en la fracción derecha contiene el conjunto de módulos desde donde se obtiene el modelo a aplicar mediante el algoritmo decision-tree.

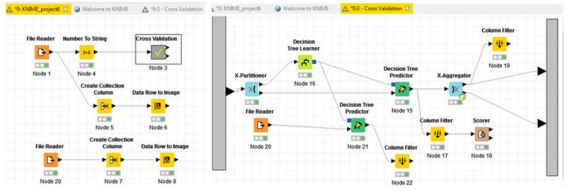

Fig. 6. Módulos de representación gráfica y los utilizados en la instancia de validación cruzada y predicción, por la herramienta KNIME Analytics

Un aspecto importante en el marco de las tareas de preprocesamiento, es la posibilidad de la herramienta de transformar cada registro con los 784 atributos (28*28 pixeles del pad), en la correspondiente imagen. La Figura 7, denota las aproximaciones que cada registro posee con los correspondientes label asociados en el caso de los registros de entrenamiento, y el grado de parecido que pudiera lograrse sobre los registros de testeo para los que no existe etiqueta.

Se puede observar que tras la instancia de aprendizaje o entrenamiento, para el tercer registro de entrenamiento (registro 43), el modelo ha predicho un 2 en lugar del cuatro que originalmente había en ”label”, y desde la visualización de la imagen cualquier usuario interpretaría a ese dibujo a mano alzada como un cuatro. Ahora bien, el modelo ”debe”tratar de modelar en ese universo de los dígitos decimales manuscritos, diferentes formas de escritura. Surge así también un dígito como el dibujado en el registro 73, que desde la visualización de la imagen a una persona le resultaría difícil rotular.

ISSN 2007-9737

Sin embargo el sistema predice adecuadamente el ”label.ºriginal. Esto destaca la fortaleza que debe tener el algoritmo de aprendizaje y modelación utilizado.

Fig. 7. Aproximación entre registros y label asociados

Las fortalezas del algoritmo utilizado, más allá de estas comparaciones de tipo subjetivo como pueden ser el contrastar los resultados de las predicciones con la percepción que de los dígitos dibujados tiene un usuario, se detectan mediante formas estadísticas que hablan del grado de aproximación que el modelo alcanza hacia los registros etiquetados.

En la Figura 8, se observa el error para cada grupo de entrenamiento conforme aquella instancia de cross-validación, que redunda en un 17,4478 % de error promedio sobre la totalidad de los 42000 registros entrenados.

Table "default" - Rows: 5 | Spec - Columns: 3 | Properties | Flo

Row ID	Error in %	Size of Test Set	Error Count
fold 0	17.536	8400	1473
fold 1	17.167	8400	1442
fold 2	17.548	8400	1474
fold 3	17.5	8400	1470
fold 4	17.488	8400	1469

Fig. 8. Error de entrenamiento

De igual manera la herramienta presenta la matriz de confusión que posibilita visualizar y contrastar los ”labelçon sus predicciones, permitiendo al experto calígrafo detectar las bondades del modelo conforme se puede reconocer las aproximaciones para cada dígito en particular.

En la Figura 9, se muestra la matriz de confusión y en la parte inferior las estadísticas de exactitud, ambas resultantes del proceso realizado.

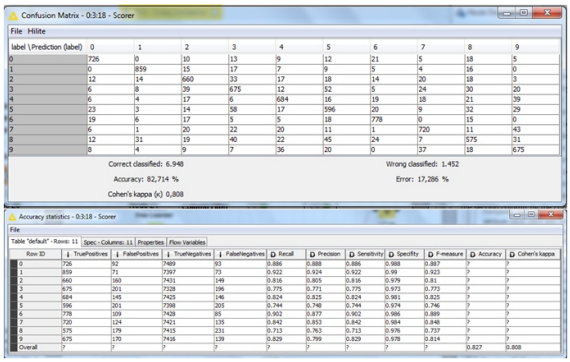

Fig. 9. Matriz de confusión y medidas de exactitud

De la lectura que el experto calígrafo haga de la Figura 9, podría significar no hacer un entrenamiento basado en muestras estratigráficas sino favorecer el aprendizaje de determinados caracteres asociado a los errores relacionados con cada uno de los dígitos.

Se observa entonces que existen errores de predicción de los dígitos escritos a mano alzada. En un comienzo bastaría tener los porcentajes de aciertos y de errores, pero en un análisis más detallado sería interesante conocer por ejemplo los dígitos mejor predichos, así como los que posiblemente generan mayor equivocación en su detección.

Si bien la diagonal de la matriz de confusión contiene los casos correctamente predichos de los diferentes dígitos, no es inmediata la información referente al porcentaje de equivocación en cada caso. Para facilitar la interpretación de los resultados, cada una de las filas podría ser acompañada por un diagrama de barras que en este caso muestre el porcentaje en que el dígito ha sido confundido con alguno de los restantes.

En lo que refiere a los indicadores Expresividad y Eficacia, se podría considerar que la matriz de confusión es expresiva, pues presenta toda la información, pero es discutible su eficacia pues no es inmediata su interpretación.

En este contexto se presentan en la Figura 10 los bloques que constituyen la instancia

ISSN 2007-9737

de aprendizaje y los errores de aproximación, resultantes de aplicar un mismo modelo a dos tipos de datos de entrenamiento. Por un lado píxeles con tipos de datos enteros entre 0 y 255 y por el otro, píxeles con tipos de datos bivaluados, 0 o 255. En la misma figura se presentan, en el extremo izquierdo, los patrones de testeo incorporados (sin etiquetar) y en el extremo derecho los resultados obtenidos para cada una de las instancias de entrenamiento.

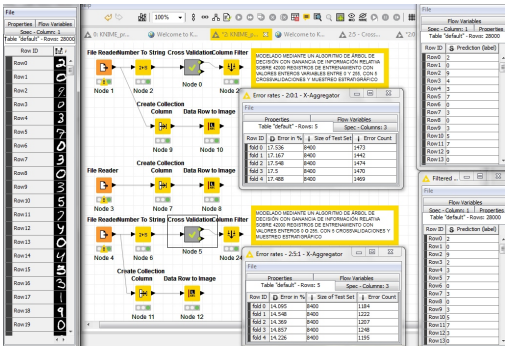

Fig. 10. Módulos de instancias de aprendizaje y testeo

Se aprecia que con solamente filtrar la imagen original, de tal manera que los pixeles no tengan valores enteros entre 0 y 255 y únicamente tengan en cada pixel 0 o 255, y utilizando el mismo algoritmo de aprendizaje, se logra una mejora en la instancia de modelación para los datos entrenados del 3 %.

Atendiendo a la información suministrada por la estadística asociada al ejemplo resuelto y en la instancia de entrenamiento se observa, desde los valores numéricos generados para el valor de precisión y tras haber sido ordenados en forma decreciente, que es el cero el que mejor precisión presenta, seguido por el uno, posteriormente el seis y así consecutivamente el resto de los números (Figura12).

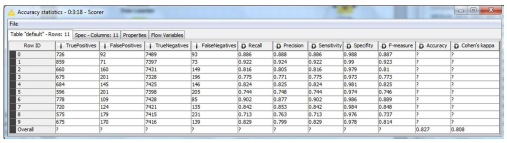

Fig. 11. Estadística de exactitud de aprendizaje

Si bien la información extraída desde los valores numéricos (Figura 11), es fiable, acertada y el procesamiento extra generado por el ordenamiento aporta una mejor posibilidad de evaluar los resultados de la tarea de reconocimiento de los caracteres manuscritos, la Figura 12, permite extraer más rápidamente las mismas conclusiones anteriores.

Allí se muestran diferentes formas de visualizar la información numérica anterior. Se puede observar que la información se presenta en porcentajes, en diferente formatos numéricos, en escala de grises y en barras de colores. Así, si en lugar de representar el dato de precisión en forma numérica lo hacemos en forma relativa mediante un gráfico de barras de colores, no solamente la longitud de la barra expresa el valor de la precisión sino que también el color puede generar un estado de alarma respecto de cuál es el número manuscrito, cuyo modelado para instancias de reconocimiento automático, no funciona adecuadamente.

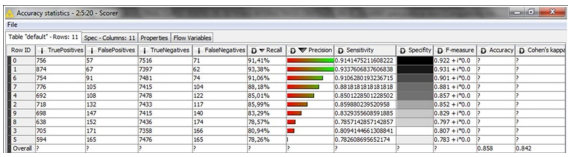

Fig. 12. Diferentes formas de visualizar la información estadística

Una vez más se observa que la imagen de barras asociada a la precisión permite inferir, más eficiente y rápidamente, que de entre todos los números el cinco es el que peor aproximación posee. Es de destacar también que es sobre el que menos ejemplos se han entrenado, lo que lleva a la necesidad de requerir una muestra mayor asociada a este tipo de carácter manuscrito. Esto pone en evidencia que los diagramas de barras son una alternativa valiosa de visualización de datos.

4.1.1. Minería de texto en contenidos mínimos de asignaturas

El análisis o minería de texto es otra de las áreas de minería de datos que cobra relevancia día

a día, dado que más del 80 % de la información de las organizaciones se encuentra en este formato y donde se hacen necesarias diferentes capacidades de visualización con el objeto de presentar las conclusiones a las que se arriba en este proceso.

Sin pretender explayarnos intensamente en los conceptos que hacen a la recuperación de información o minería de texto, sí explicaremos algunos conceptos introductorios a efectos de entender su representación y tratamiento. Para interpretar en detalle los documentos de texto, se puede buscar en ellos palabras claves o categorizarlos según su contenido semántico. Cuando se identifican palabras claves, se observan definiciones o detalles característicos de esos documentos que pueden usarse para buscar relaciones, conexiones o parecidos con otros documentos. En el área de recuperación de información los documentos se han representado tradicionalmente en un modelo de espacio vectorial, generando lo que se conoce como valija de palabras.

Un documento en el modelo de espacio vectorial se representa como un vector de pesos, en el que cada peso de los componentes se calcula sobre la base de alguna variación de la frecuencia con que las diferentes palabras tokens o términos aparecen en los documentos analizados (TermFrecuency). Los documentos se pueden representar como vectores en un espacio n-dimensional. Si un cierto valor t (palabra, token) ocurre n veces en un documento d, entonces la coordenada correspondiente a t del documento d es simplemente n. En la Figura 13, se compara tres documentos.

Un primer documento, *Título*, compuesto por la palabra minería repetida tres veces ”minería minería minería”, lo que denota su valor en coordenadas sobre el eje minería. Un segundo documento, *Doc1*, compuesto por la palabra texto repetida dos veces ”texto texto”, y finalmente un tercer documento, *Doc2*, compuesto por ”minería texto minería texto minería datos”. De la representación vectorial se observa que no hay ninguna semejanza sintáctica entre *Título* y *Doc1* (vectores perpendiculares que no tienen ningún término en común) y sí la hay entre *Título* y *Doc2* (vectores que forman un ángulo diferente a 90° y poseen términos en común). Se destaca el ángulo que conforman los vectores, dado que una de las métricas de similitud entre documentos muy usada en el ámbito de la recuperación de documentos, es la del coseno del ángulo que forman los documentos que se están comparando. Documentos de mayor similitud poseen valor de métrica próximo a 1 y aquellos con métrica cercana a 0 son considerados más disímiles.

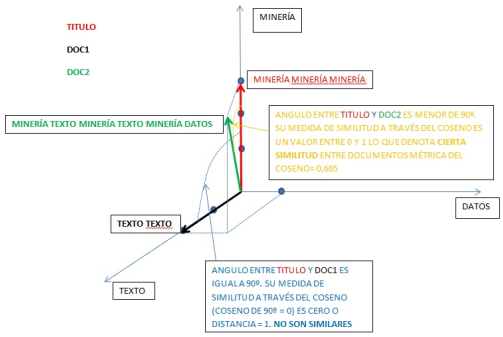

Fig. 13. Ejemplo de representación vectorial de documentos

La aplicación realizada simplemente presenta diferentes formas de visualizar documentos correspondientes a los contenidos mínimos de las asignaturas correspondientes a un plan de Licenciatura en Ciencias de la Computación. La implementación llevada adelante en KNIME Analytics presenta una serie de módulos que permiten realizar la lectura de los distintos documentos, la generación de la valija de palabras plausible ya de representarse vectorialmente, así como todas las tareas inherentes al preprocesamiento de los documentos que consiste en transformar todos los caracteres a mayúsculas, eliminar palabras carentes de significado (stop words), eliminar signos de puntuación y generar la estadística asociada entre otras tareas propias del análisis sintáctico de texto.

La Figura 14, muestra la transformación provocada por el procesamiento de la información conforme se suceden los módulos que generan la valija de palabras, eliminan signos de puntuación, eliminan stop-words, calculan frecuencia relativas de las palabras en los documentos, para finalmente mostrarla además en la nube de

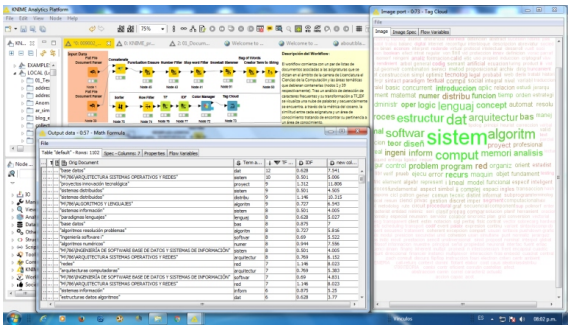

Fig. 14. Procesamiento de documentos de texto en KNIME

palabras o TagsCloud. Ésta permite extraer mayor conocimiento del vistazo que significa la presentación del conjunto de palabras involucradas según su tamaño, conforme la frecuencia relativa de las mismas en los documentos, y color asignado.

La Figura 15, permite visualizar, además, los resultados de la aplicación que busca similitudes sintácticas entre los contenidos mínimos de diferentes asignaturas y las áreas de conocimiento que debieran contenerlas. Esta primera aproximación se presenta a modo de tabla y la distancia (métrica de similitud del coseno), entre cada una. Se destaca que en este caso la distancia refleja la similitud sintáctica entre documentos, así conforme la distancia disminuye mayor similitud sintáctica existe entre documentos.

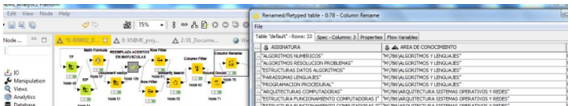

Fig. 15. Secuencia de Módulos y tablas evidenciando similitudes entre asignaturas y áreas de conocimiento

Es interesante destacar que entre las diferentes formas de visualizar la información anterior, la herramienta tiene la posibilidad de presentarla en forma de un grafo, por medio de una serie de plugins que permiten trabajar el área de netmining. Es así como la Figura 16, presenta la secuencia de módulos que permite la visualización del grafo, y la Figura 17, los respectivos nodos (asignaturas y áreas de conocimiento), con las relaciones existentes entre ellos, que en este caso son las correspondientes métricas de similitud.

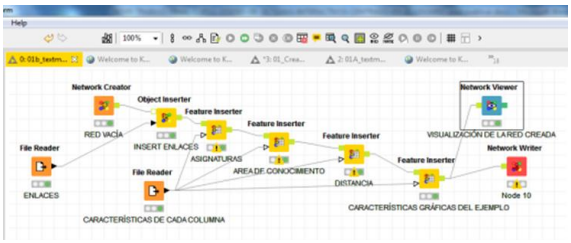

Fig. 16. Secuencia de módulos que permite presentar información en forma de un grafo o red

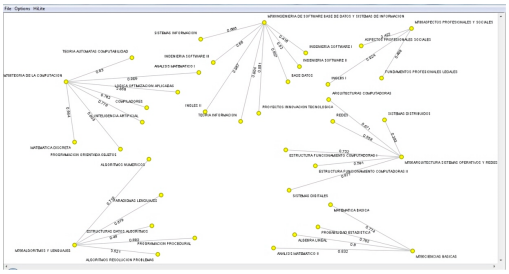

Fig. 17. Nodos y enlaces que visualiza las métricas de similitud sintáctica entre asignaturas y áreas de conocimiento

4.2. Representación de datos georreferenciados

En este caso se presentan diferentes formas de visualizar un conjunto de datos georreferenciados. KNIME permite la traducción de direcciones postales (nombre de calle, numeración, orientación, jurisdicción, provincia, país), a coordenadas latitud y longitud, y desde éstas pasar a una representación en un mapa físico correspondiente a la ciudad o zona asociada a los datos georreferenciados.

Sin duda la representación visual asociada a mapas, es mucho mejor recibida por neófitos

usuarios finales pese a que la información caracterizada por las coordenadas sea exactamente la misma.

En la Figura 18, se observa un conjunto de direcciones postales correspondientes a la ciudad de San Juan, Argentina y su posterior transformación a coordenadas de latitud y longitud, las que finalmente se representan en un mapa geográfico mediante módulos específicos de KNIME.

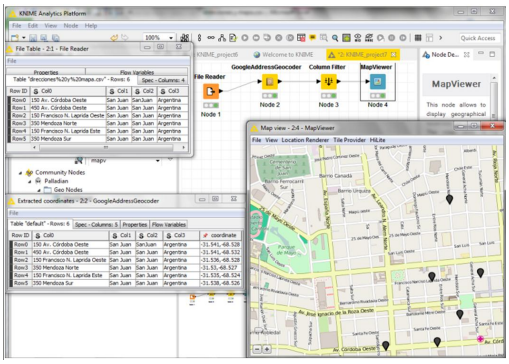

Fig. 18. Visualización de datos georreferenciados en KNIME

En particular, los módulos de georreferenciación de la plataforma KNIME Analytics 3.2.1 han permitido, también, traducir datos correspondientes a las campañas de recolección de semillas del Instituto Semillero Hortícola de la provincia de San Juan (INSEMI), convirtiendo los formatos de latitud y longitud de cada uno de los lugares de recolección, en puntos de coordenadas geográficas sobre un mapa, figura 19; y mediante un rápido análisis visual por parte de sus especialistas, determinar zonas no exploradas y desde allí organizar futuras campañas.

Se puede observar además la posibilidad que brinda la herramienta, a través del módulo mapviewer, de rotular los diferentes puntos de recolección atendiendo al nombre de la especie colectada y colorear la referencia según la altitud en msnm (metros sobre el nivel del mar), variando de la menor altitud (color celeste), a la mayor altitud (azul intenso).

Se destaca además, que el módulo OSM Mapview permite visualizar y analizar zonas según su relieve geográfico y la aridez según su color.

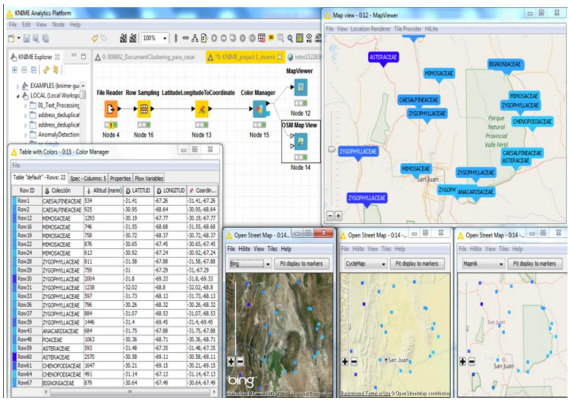

Fig. 19. Georreferenciación de campañas de recolección de semillas

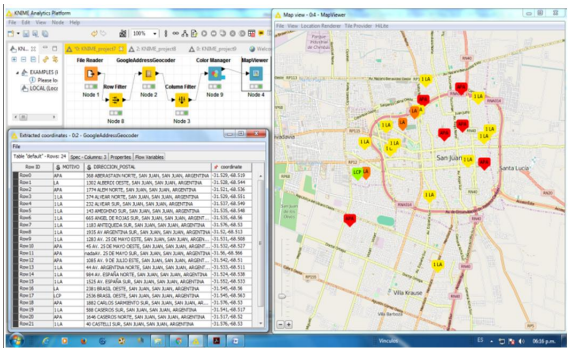

Fig. 20. Visualización georreferenciada de fallas en el alumbrado público

La información asociada a los reclamos y su tipificación en el contexto de la oficina correspondiente al organismo encargado del mantenimiento del alumbrado público, pueden ser presentados como datos georreferenciados identificados por nombre y color, Figura 20.

Esto permite una más rápida toma de decisiones respecto de qué vehículo destinado a reparaciones se encuentra en las inmediaciones de cada reclamo y solucionar el problema con mayor celeridad.

5. Conclusiones

El trabajo propone un análisis de diversas formas de presentar la información o conocimiento

extraído desde los datos, destacando diferentes etapas en la tarea de preprocesamiento.

Consideramos de suma importancia conocer fehacientemente la tipología y área de aplicación de los datos a efectos de sugerir estrategias de visualización que permitan la mayor captura de información por parte de los usuarios. En este contexto la posibilidad de representar la imagen de los números manuscritos, los tagscloud así como formas de representar relaciones entre documentos mediante grafos en las tareas de minería de texto, y transformar un conjunto de puntos georreferenciados a un mapa geográfico, son una valiosa ayuda para los usuarios y sin dudas un complemento a la propia representación interna de matrices, tablas estadísticas y otras formas involucradas para valorar diferentes tareas en el marco de extracción de conocimiento en datos.

En definitiva, y habiendo utilizado una mínima fracción de formas de visualización que la herramienta KNIME posee, consideramos que toda propuesta de visualización del conocimiento extraído que pueda ser transformado de lo numérico a lo gráfico va a ser mejor interpretado por parte de los usuarios. Además, en los tres casos de aplicación se pueden verificar las características de ser útil, usable y deseable. Asimismo resultan ser visualmente atractivas, escalables y desde ellas se puede obtener información correcta y plenamente accesibles por parte de los usuarios.

Así para el caso de los caracteres decimales manuscritos se puede decir que la representación gráfica es una ayuda importante para que un usuario, carente de experticia, infiera si la instancia de modelación alcanzada funciona adecuadamente o no. Aun así hay pocos casos en que la representación lograda difícilmente pueda ser interpretada como un dígito decimal. Del mismo modo en que se modeló un clasificador para el reconocimiento automático de dígitos decimales podría extenderse a cualquier tipo caligráfico.

La instancia de la matriz de confusión presenta información correcta y es fácilmente escalable (con el aumento de registros de datos u otros tipos de caracteres manuscritos), pero requiere de un conocimiento previo, por parte del usuario, para su interpretación.

En lo referente a la aplicación de minería de texto desde la observación de la nube de palabras, éstas, su tamaño, color y cercanía reflejan inmediatamente la relevancia de la temática abordada por los documentos analizados, lo que denota su utilidad. Conforme la cantidad de caracteres aumente y sus frecuencias de aparición tiendan a ser uniformes, se puede tornar difícil el reconocimiento específico de los temas abordados por los documentos.

El trabajo de encontrar métricas de similitud sintáctica entre documentos y con ello poder establecer la pertinencia de una asignatura a un área de conocimiento se explicita mejor en el entorno gráfico basado en grafos. Éste, conforme aumentasen la cantidad de documentos tratados, podría generar una excesiva cantidad de nodos y enlaces pero que, a tales efectos y con el objetivo de permitir escalabilidad, la herramienta dispone de diferentes alternativas de filtros y modos de representación de grafos.

En lo referente al tercer caso de aplicación, la representación cartográfica de los datos es muy útil y deseable como lo ponen en evidencia la exactitud [6, 8], de los posicionadores satelitales actuales montados a bordo de vehículos automotores, y que permiten transitar hacia un destino específico en forma fiable y prácticamente sin tener en cuenta direcciones postales. Así mismo resultan de suma utilidad para los encargados de evaluar la información presentada. La posibilidad de escalabilidad es factible atendiendo a que los módulos que permiten la visualización tan solo presentan como límite, no visualizar más de aproximadamente 10,000 registros por segundo.

Agradecimientos

Agradecemos la colaboración prestada por la Prof. Laura Beatriz Ureta, docente e investigadora de la Universidad Nacional de San Juan, en la elaboración del abstract.

Referencias

1. **Alcalde, I. (2015).** *Visualización de la información: de los datos al conocimiento*. Editorial UOC.
2. **Faulkner, X. & Culwin, F. (2000).** Enter the usability engineer: integrating HCI and software engineering. *ACM SIGCSE Bulletin*, Vol. 32, No. 3, pp. 61–64.
3. **Fayyad, U., Wierse, A., & Grinstein, G. (2002).** *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann.
4. **Granollers i Saltiveri, T., Lorés Vidal, J., & Cañas Delgado, J. J. (2012).** *Diseño de sistemas interactivos centrados en el usuario*. Editorial UOC, Catalunya.
5. **Iliinsky, N. & Steele, J. (2011).** *Designing data visualizations: Representing informational Relationships*. O Reilly Media, Inc.
6. **Karimi, H. (2014).** *Big Data: techniques and technologies in geoinformatics*. CRC Press.
7. **KNIME (2016).** *KNIME Analytics versión 3.2.1*. software.
8. **Mark, S., Resmini, R., Cervone, G., Lin, J., & Waters, N. (2014).** *Data Mining for Geoinformatics Methods and Applications*. Springer, New York.
9. **McCandless, D. (2010).** *The beauty of data visualization*. John Wiley & Sons, TED website.
10. **Ward, M. O., Grinstein, G., & Keim, D. (2010).** *Interactive data visualization: foundations, techniques, and applications*. CRC Press.
11. **Wikipedia (2015).** Matriz de confusión — wikipedia, la enciclopedia libre.
12. **Yuk, M. & Diamond, S. (2014).** *Data visualization for dummies*. John Wiley & Sons.

Article received on 01/08/2016; accepted on 12/10/2016.
Corresponding author is Raúl Oscar Klenzi.