View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Servicio de Difusión de la Creación Intelectual

Análisis, interpretación y toma de decisiones estratégicas en la Ciencia de Datos

Mag. María Alejandra Malberti Riveros, Mag. Raúl Oscar Klenzi, Mag. Graciela Elida Beguerí

Instituto de Informática / Departamento de Informática / Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan
Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan,
Teléfonos: 4260353, 4260355 Fax 0264-4234980, Sitio Web: http://www.exactas.unsj.edu.ar
e-mail: {amalberti, rauloscarklenzi, grabeda}@gmail.com

Resumen

Se propone abordar el paradigma de Ciencia de Datos con el objetivo de reconocer, analizar y describir el conjunto de estudios y prácticas inherentes a la misma y aplicable a grandes colecciones de datos provenientes de diferentes áreas tales como Educación, Bibliotecología, Astronomía y redes sociales. Estos datos serán accedidos y analizados por medio de herramientas de software libre licencia AGPL como Knime, Weka, R, Rapidminer y módulos específicos de Python, que se ejecuten en diferentes plataformas de hardware secuenciales, paralelos y distribuidos.

Palabras clave: Ciencia de Datos, Aprendizaje Automático, Minería de Datos, Software Libre

Contexto

La revolución de datos se está extendiendo a diferentes ámbitos, lo que conduce a que en la actualidad las personas que trabajan en muy diversas aéreas estén obligadas a entender cómo usar los datos. En otras palabras, la proliferación de datos ha creado una demanda de conocimientos, y esta demanda está inmersa en muchos aspectos de la cultura moderna similar a lo que ocurría en los años 80 y 90 cuando surgió la necesidad de que todas las personas, independiente de la actividad en que se desempeñaran, tenían que aprender a usar computadoras. El rápido desarrollo de las tecnologías de la Información y Computación en los últimos 20 años ha cambiado muchos campos de la ciencia y la ingeniería. Algunas disciplinas se han visto transformadas por causa de los datos, sea por cantidad como por la dinámica de los mismos, llevando esto al desarrollo de métodos dato-intensivos en el área de la ciencia y la ingeniería. De esta manera se comienzan a acuñar nuevas áreas de conocimiento como la Ciencia de Datos (Data Science -DS-) o Ingeniería Dato-intensivos. (Pierson, L. 2015)

Así como las Ciencias de la Computación, Sistemas de Información, Ingenierías en Sistemas y Computación han sido las principales titulaciones que la academia ha brindado en el estudio del área de la Informática, hoy en día la cantidad de conocimiento necesario, derivado de los diferentes campos involucrados para las aplicaciones de Descubrimiento de Conocimiento en Datos (Knowledge Discovery in Data – KDD-) o Descubrimiento de

Conocimiento en Datos Masivos (Knowledge Discovery in Massive Data – KDDM-) lleva a que a nivel de grado se incorporen conocimientos relativos a Ciencia de Datos, o a nivel de posgrado se propongan titulaciones en Ciencia de Datos.

La presente propuesta se encuentra en instancias de evaluación, por parte del CONSEJO DE INVESTIGACIONES CIENTÍFICAS Y TÉCNICAS Y DE CREACIÓN ARTÍSTICA- CICITCA, y para su desarrollo durante el bienio 2016-2017. En el marco de la misma se pretende dar continuidad al proyecto "Búsqueda de Conocimientos en Datos Masivos" llevado adelante en el bienio 2014-2015. El mencionado proyecto ha permitido a sus integrantes trabajar en una primera aproximación con datos inherentes al área astronómica especialmente brindados por el grupo GAE (Grupo de Astronomía Extragaláctica) de la FCEFN y relevados desde el Sloan Digital Sky Survey22 (SDSS).

Es idea del grupo para la presente propuesta, profundizar el conocimiento y utilización de algoritmos específicos de aprendizaje de máquina que permitan describir y predecir comportamientos no solamente en datos asociados al área de la astronomía sino también a las áreas pertenecientes a las redes sociales, así como determinación de perfiles de alumnos universitarios, utilizando técnicas de caracterización multidimensional.

Respecto al abordaje de problemas actitudinales se integran a la presente propuesta, participantes del proyecto "El comportamiento académico de los alumnos de primer año de las carreras de informática de la FCEFyN de la UNSJ Estrategias para mejorar su rendimiento" desarrollado en el periodo 2011-2013, Cod. 21/E875 res037/11CS, quienes han realizado experiencias teóricas y de campo en la difícil tarea de la caracterización de comportamientos de alumnos, detectando una compleja confluencia de factores que tienen que ver con la personalidad, aspectos vocacionales y su situación social, económica y familiar.

Introducción

A partir de la década del 2000 se produjeron grandes transformaciones; las transacciones realizadas en distintas organizaciones comenzaron a registrarse lo que causó que en el 2009 una base de datos corporativa promedio contenía alrededor de cinco petabytes, o 5.000.000 gigabytes de datos. (Un petabyte - PB- equivale a 1015 bytes = 1 000 000 000 000 000 de bytes; Un gigabyte -GB- equivalente a 109 =1.000.000.000 -mil millones- de bytes). Así mismo, en 1998 Google ya había comenzado a registrar cada búsqueda realizada. En 2004, Facebook comenzó a registrar cada interacción de sus usuarios, y en 2005 YouTube comenzó a mover en todo el mundo grandes cantidades de datos de video. En esos momentos las tecnologías no eran capaces de manejar las enormes cantidades de datos, pero han evolucionado y en la actualidad requieren de diferentes competencias pues la cantidad de datos generados seguirá creciendo enormemente.

A medida que la demanda de datos crece, también lo hará la demanda de capital humano. Las técnicas cuantitativas (estadísticas, minería de datos, predicción, optimización, etc.) representarán para gestores y analistas el know-how al utilizar el análisis de grandes cantidades de datos para tomar decisiones efectivas; estudiantes en ciencias de datos son parte de ese futuro capital humano.

Si bien la mayoría de las carreras en Ciencia de datos son a nivel de posgrado (datos de EEUU), existe en aumento la creencia de que es necesario la creación de carreras a nivel de grado, con intensa fundamentación en ciencia, tecnología, ingeniería y matemáticas http://www.kdnuggets.com/2015/10/data-science-education-begin.html

En la actualidad habilidades de aprendizaje de máquina son fuertemente requeridas en el mercado laboral, a raíz de que las empresas tienden cada vez más a construir sistemas de decisión automatizados. Esto conlleva a promover la incorporación en las curriculas de las carreras dependientes del Departamento de Informática, de los saberes requeridos para la formación de habilidades inherentes a un científico de datos.

Se han desarrollado métodos de bases de datos que almacenan y administran petabytes de datos on-line, haciendo a estos seguros y accesibles vía Internet o sistemas distribuidos de cómputo y que pueden a la vez ser analizados por potentes herramientas de minería de datos (Data Mining -DM). Es así como DM, que se puede definir como la extracción de conocimiento no trivial, en grandes cantidades de datos y que favorece la toma de decisiones, adquiere centralidad en el área de la ciencia de datos. DM es la confluencia de múltiples áreas interdisciplinarias como la estadística, el aprendizaje de máquina (Machine Learning -ML-), reconocimiento de patrones, sistemas de bases de datos, recuperación de información, la World Wide Web, técnicas de visualización, entre otras, que aplicadas a diferentes dominios han permitido su reciente y creciente progreso. Con el objetivo de asegurar que el DM siga siendo beneficioso para la ciencia y la ingeniería es importante analizar los desafíos en que el DM habrá de ingresar conforme evolucionen la ingeniería y la ciencia dato-intensivas. (Kargupta, H.,,2008)

La tarea esencial de quien hace ciencia de datos es transformar datos en bruto en conocimiento para la toma de decisiones. El análisis de uso de las bases de datos de una biblioteca hace necesario insertar la ciencia de datos en éstas, de modo que permita a los bibliotecarios aplicar herramientas de software con las que logren transformar datos, recuperar, analizar y graficar información. Es decir, adquieran un conjunto de habilidades para unir los extremos entre la necesidad de información de un usuario y el conocimiento almacenado en los datos.

Los equipos de trabajo e investigación en DS han conformado una gama relativamente amplia de habilidades; el lenguaje y/o las herramientas de software requeridas para cualquier aplicación pueden seleccionarse según los conocimientos y experiencia previa. Para algunas aplicaciones - especialmente la creación de prototipos y desarrollo - es más rápido y ágil que se utilicen herramientas y tipo de datos ya conocidos.

En este sentido se presenta el relevamiento estadístico realizado por KDnuggets.com respecto de las estructuras de datos y áreas de aplicación donde se expande la minería de datos, así como los algoritmos y herramientas de software más utilizadas por la comunidad científico-productiva..

Entre otras herramientas están Orange, Weka, R, RapidMiner, Knime, Hadoop, MapReduce y el lenguaje de programación Python. Estas herramientas son entornos de prueba de algoritmos de aprendizaje de máquina y extracción de conocimiento en datos en diferentes formatos, con excelentes tutoriales, y disponibles para plataformas Windows, Linux y Mac. Así mismo muchas

compañías como Google, IBM, Amazon y Microsoft, con el objetivo de procesar grandes cantidades de datos, han incorporado en sus API (Application Programming Interface) tecnología de aprendizaje automático que no requiere elevada experiencia por parte de los potenciales usuarios.

- Orange: es un desarrollo en lenguaje Python de la Universidad de Liubliana, Eslovenia, desde 1996 y factible de descargar de http://orange.biolab.si/download/.
- Weka: Waikato Environment Knowledge Analysis es un desarrollo en Java de la Universidad de Waikato, Hamilton, Nueva Zelanda desde 1993. Se descarga de http://www.cs.waikato.ac.nz/~ml/weka/
- R: (librería Rattle): desarrollado en la Universidad de Auckland, Nueva Zelanda, en 1993 y la librería Rattle, creada por el Dr. Graham Williams. www.r-project.org
- RapidMiner: (originalmente YALE) versión: 5.3.15, desarrollado, en Java (plataforma Eclipse) en la Universidad de Dortmund, Alemania, en lenguaje Java, desde 2001 http://sourceforge.net/projects/rapidminer/
- Knime: desarrollado en la Universidad de Constanza, Alemania. http://www.knime.org/downloads/overview.
- Hadoop: Se caracteriza por su capacidad de acceder y permitir procesar enormes cantidades de datos sobre hardware relativamente económico, además hacer posible el almacenamiento de datos en un sistema de archivos distribuido (Hadoop File System -HDFS-). Particularmente Apache Hadoop es un framework de código libre que soporta aplicaciones distribuidas con datos masivos, bajo licencia apache v2. Puede ser configurado para trabajar con miles de máquinas y petabytes de información. Hadoop está inspirado en los papers de Google sobre MapReduce y Google File System.

Líneas de investigación, Desarrollo e Innovación

- Analizar y describir el conjunto de estudios y prácticas requeridos en Ciencia de Datos.
- Construir y validar instrumentos tendientes a recabar datos inherentes a los estudiantes.
- Estudiar y analizar diferentes conjuntos de datos masivos a procesar.
- Evaluar herramientas de software libre para arquitecturas secuenciales, paralelas y distribuidas.
- Descubrir conocimiento desde diferentes conjuntos de datos.

Resultados y Objetivos

En el marco del proyecto se pretende analizar y estudiar el paradigma de Ciencia de Datos, desde el tratamiento de datos provenientes de diversas áreas por medio de la aplicación de estrategias de aprendizaje de máquina y minería de datos, sobre arquitecturas secuenciales, paralelas y distribuidas.

Algunos de los propósitos, son:

- Sugerir un conjunto de saberes convenientes para la formación de un científico de datos, para que los mismos sean incorporados, según corresponda, en las currículas de las carreras pertenecientes al Departamento de Informática de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan -FCEFN, UNSJ-.

- Realizar aportes a la toma de decisiones con los datos tratados en las diferentes aéreas del saber.

Formación de Recursos Humanos

En el marco de esta propuesta, se están desarrollando dos trabajos finales de licenciatura en las carreras pertenecientes al Departamento de Informática de la -FCEFN, UNSJ-.

También se dirige una tesis de posgrado, que se encuentra en etapa de redacción del informe final, correspondiente a la Maestría en Informática de la Universidad Nacional de La Matanza.

A la vez se prevé la generación de trabajos finales de posgrado, para la Maestría en Informática de la -UNSJ-, así como la tutela de becarios en el área expuesta.

Recientemente se ha creado en dependencias del Instituto de Informática de la -FCEFN, UNSJ- el Laboratorio de Sistemas Inteligentes para la Búsqueda de Conocimientos en Datos Masivos, ámbito de aplicación sustantiva de esta temática.

Referencias

- ANUIES Innovación curricular en instituciones de educación superior. Pautas y procesos para su diseño y gestión. Colección Documentos. Compilación Medina Cuevas, Lourdes; Guzmán Hernández, Laura Leticia. Primera edición, 2011 ISBN 978-607-451-032-4
- Ayllón, Silvia, Merlino, Aldo, Escanés, Gabriel, Variables que influyen en la deserción de estudiantes universitarios de primer año. Construcción de índices de riesgo de abandono Revista Electrónica "Actualidades Investigativas en Educación" [en linea] 2011, Disponible en:<http://www.redalyc.org/articulo.oa?id=44720020005> ISSN
- Felmer, L. R., Pool, G. M., Fisher, I. R., & Fritz, C. G. (2009). Los estilos epistémicos y tipos de personalidad como factores asociados a la elección de carrera Epistemological styles and types of personality as factors associated in choice of study areas. *Revista de Pedagogía*, *30*(86), 115-134.
- Jurney, R. (2013). *Agile Data Science: Building Data Analytics Applications with Hadoop*. " O'Reilly Media, Inc.".
- Kargupta, H., Han, J., Philip, S. Y., Motwani, R., & Kumar, V. (Eds.). (2008).*Next generation of data mining*. CRC Press.
- Pierson, L., Swanstrom, R., & Anderson, C. (2015). *Data Science for Dummies*. John Wiley & Sons.
- Sarro, L. M., Eyer, L., O'Mullane, W., & De Ridder, J. (Eds.). (2012).*Astrostatistics and Data Mining* (Vol. 2). Springer Science & Business Media.
- Van Der Aalst, W. (2011). *Process mining: discovery, conformance and enhancement of business processes*. Springer Science & Business Media.
- Way, M. J., Scargle, J. D., Ali, K. M., & Srivastava, A. N. (Eds.). (2012).*Advances in machine learning and data mining for astronomy*. CRC Press.
- Williams, G. (2011). *Data mining with Rattle and R: The art of excavating data for knowledge discovery*. Springer Science & Business Media.
- Zhao, Y. (2012). *R and data mining: Examples and case studies*. Academic Press.