Minería de Datos Aplicada a Datos Masivos

Anabella De Battista, Patricia Cristaldo, Lautaro Ramos,
Juan Pablo Nuñez, Soledad Retamar, Daniel Bouzenard
Departamento Ingeniería en Sistemas de Información
Fac. Reg. Concepción del Uruguay
Universidad Tecnológica Nacional
Entre Ríos, Argentina
{debattistaa, cristaldop, ramosl, nunezjp, retamars, bouzenardd}@frcu.utn.edu.ar

Norma Edith Herrera
Departamento de Informática
Univ. Nac. de San Luis
San Luis, Argentina
nherrera@unsl.edu.ar

Resumen

Las grandes cantidades de datos que se producen en la actualidad, sumadas a su heterogeneidad, hacen que las herramientas tradicionales de análisis de datos no resulten adecuadas para su recopilación, almacenamiento, gestión y análisis. En este contexto se comienza a hablar del término *Big Data*, haciendo referencia a características como gran volumen, velocidad y variedad de producción de los datos, y a las herramientas que se utilizan para encontrar valor en las mismas. La posibilidad de hallar patrones y tendencias en estas grandes cantidades de datos impacta directamente en la toma de decisiones en áreas tan diversas como salud, genética, agro, predicciones climáticas, redes sociales, marketing, finanzas, educación, entre otras. Otro aspecto de interés en este tipo de análisis, es la aplicación de metodologías de gestión de proyectos de enfoque ágil en los proyectos de minería de datos, en este caso, se aplicarán metodologías específicas con el objetivo de comparar características y restricciones de cada una. En este artículo se presentan los tópicos de interés del proyecto *Minería de Datos: su aplicación a repositorios de datos masivos*.

Palabras Claves: *Big Data*, minería de datos, clustering, agrupamiento, gestión de proyectos, CRISPDM.

1. Contexto

El presente trabajo se desarrolla en el ámbito del proyecto *Minería de Datos: su aplicación a repositorios de datos masivos* (EIUTICU0003781TC) del Grupo de Investigación en Bases de Datos, perteneciente al Departamento Ingeniería en Sistemas de Información de la Universidad Tecnológica Nacional, F. R. Concepción del Uruguay.

2. Introducción

Actualmente se producen diariamente grandes volúmenes de datos de diversos tipos (e.g., textos, imágenes, audio, videos) y desde los más variados orígenes (e.g., web, GPS, redes sociales, sensores). Se prevé que en los próximos años las aplicaciones *Internet de las cosas* aumentarán el volumen de datos a un nivel sin precedentes. En este contexto, surge el término *Big Data* referido a conjuntos de datos cuyo tamaño supera la capacidad de las herramientas tradicionales de bases de datos de recopilar, almacenar, gestionar y analizar la información. En general se habla de *Big Data* o Análisis de *Big Data* como sinónimos, ya que no sólo se desea hacer referencia a la gran cantidad y complejidad de los datos, sino también a las herramientas utilizadas para procesarlos y extraer conocimiento útil de los mismos.

Algunas definiciones indican que *Big Data* puede definirse a partir de las siguientes características [1]:

- Volumen: órdenes superiores a Terabytes de datos
- Variedad: distintos tipos de datos provenientes de diversas fuentes que pueden organizarse tanto en forma estructurada como no estructurada.
- Velocidad: referido a la velocidad de generación de los datos o a la rapidez con la que se generan y procesan los datos.
- Variabilidad: referido a la inconsistencia que pueden presentar los datos en ocasiones, dificultando las tareas de análisis.
- Valor: gracias a la posibilidad de tomar decisiones al responder preguntas que antes no era posible, ofrece a la organización una ventaja estratégica.

Las actividades de la comunidad científica y profesional han cambiado debido al surgimiento de estos grandes repositorios de datos, ya que se requieren nuevas estrategias para su almacenamiento, tratamiento, distribución y análisis, porque no sólo se cuenta con una gran cantidad de datos sino que además su complejidad es creciente. La posibilidad de hallar patrones y tendencias en estas grandes cantidades de datos impacta directamente en la toma de decisiones en áreas tan diversas como salud, genética, agro, predicciones climáticas, redes sociales, marketing, finanzas, educación, entre otras. Es por esta razón que se comienza a trabajar en nuevas herramientas, tecnologías, métodos y sistemas requeridos para manejar grandes conjuntos de datos distribuidos, heterogéneos, no estructurados, diversos y complejos. La detección de agrupamientos en repositorios de datos grandes y complejos es una de las actividades más relevantes en análisis de información. Por ejemplo, dada una base de datos de imágenes satelitales de varias decenas de terabytes: ¿Es posible encontrar regiones con el objetivo de identificar selvas naturales, deforestación o reforestación, o identificar terreno cultivado y el tipo de cultivo? ¿Puede hacerse automáticamente? La respuesta a ambos interrogantes es ”si”. En la actualidad ese análisis puede realizarse en pocos minutos con muy alta precisión [2]. Por otro lado, la clasificación de datos complejos no puede realizarse con las herramientas tradicionales de análisis de datos, por lo que han surgido nuevos algoritmos especialmente diseñados para el análisis de datos masivos. La Minería de Datos involucra e integra técnicas de diferentes disciplinas tales como tecnologías de bases de datos y data warehouse, estadísticas, aprendizaje de máquinas, computación de alta performance, reconocimiento de patrones, redes neuronales, visualización de datos, recuperación de información, procesamiento de imágenes y señales, y análisis de datos espaciales o temporales. En este proyecto se estudiarán procesos de Minería de Datos desde una perspectiva de bases de datos, con enfoque en técnicas eficientes y escalables.

3. Líneas de Investigación, Desarrollo e Innovación

La línea de trabajo principal de nuestro proyecto de investigacion es el estudio de técnicas de Minería de Datos aplicables a repositorios de datos masivos, atendiendo principalmente a su eficiencia y escalabilidad. En particular se realizará el estudio, análisis y comparación del funcionamiento de algoritmos de clustering y de clasificación aplicables a datos masivos, realizando pruebas en distintos modelos de bases de datos (espaciales, temporales, espacios métricos) para posteriormente proponer mejoras a los algoritmos existentes o bien, nuevos algoritmos [3]. Además se pretende desarrollar aplicaciones que implementen algoritmos de clustering o de clasificación, para el tratamiento de repositorios de datos de alta complejidad.

En la gestión de los proyectos específicos de Minería de Datos que se emprendan en este proyecto de investigación se emplearán distintas metodologías de gestión de proyectos de enfoque ágil y se realizará una comparación de características y restricciones de cada metodología, en relación a los factores críticos de éxito y las razones de fracaso en la gestión de proyectos de enfoque ágil [4, 5].

3.1. Técnicas de Minería de Datos Aplicables a Datos Masivos

La gran cantidad de datos que actualmente generan y almacenan aplicaciones de diversas áreas, está en continuo crecimiento, no sólo desde el punto de vista de objetos y atributos, sino en la complejidad de los atributos que describen a cada objeto [2]. La masividad de estos datos ha superado nuestra capacidad de procesar, almacenar adecuadamente, analizar y entender estos grandes repositorios. Es de gran interés para las organizaciones propietarias de datos masivos poder extraer conocimiento de ellos, convirtiéndolos en recursos útiles para la toma de decisiones en lugar de sólo mantenerlos resguardados en discos de computadoras, sin ser accedidos nunca. En este contexto surge el área de investigación conocida como Descubrimiento de Conocimiento en Bases de Datos (KDD por sus siglas en inglés), que se define como el proceso no trivial de identificar patrones válidos, desconocidos, potencialmente útiles y comprensibles en los datos. El proceso de KDD consta de una secuencia iterativa de etapas: integración y recopilación de datos; selección, limpieza y transformación de datos; minería de datos; evaluación; difusión, uso y monitorización de modelos. Actividades comunes dentro del proceso de KDD son clustering, clasificación y etiquetado, identificación de errores de medición, detección de outliers, inferencia de reglas de asociación y datos ausentes, y reducción de la dimensionalidad. KDD es un proceso complejo que tiene un alto costo computacional, producto de explorar varios elementos en distintas combinaciones para lograr el conocimiento deseado [2]. En las bases de datos tradicionales los datos se representan mediante atributos numéricos o categorizados en una tabla, donde cada tupla representa un elemento del conjunto. El desempeño de los algoritmos de análisis de datos en general depende del número de elementos en el conjunto, o de la cantidad de atributos en la tabla, y de las distintas formas en que interactúan tuplas y atributos. En un contexto de *Big Data* las técnicas tradicionales de KDD no son suficientes por la complejidad y magnitud de los repositorios de datos. La fase de Minería de Datos es la más característica del KDD por eso con frecuencia se utiliza esta fase para nombrar todo el proceso. Su objetivo es producir nuevo conocimiento que pueda ser útil al usuario. Esto se realiza construyendo un modelo basado en los datos recopilados a tal fin. El modelo describe los patrones y relaciones entre los datos que pueden utilizarse para: realizar predicciones, entender mejor los datos, explicar comportamientos pasados. Para

ello se deben tomar ciertas decisiones antes de comenzar el proceso: determinar qué tipo de tarea de minería es el más apropiado (por ejemplo, se podría utilizar clasificación para identificar en una universidad que alumnos abandonarán sus estudios); seleccionar el tipo de modelo (para una tarea de clasificación se podría utilizar un árbol de decisión para obtener un modelo en forma de reglas); seleccionar el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo buscado [6].

Existen distintos tipos de tareas dentro de la minería de datos, cada una con sus propios requisitos y que retornan información de distintos tipos. Las tareas se clasifican en:

- Predictivas: a su vez de distinguen las tareas de clasificación (técnicas: árboles de decisión, reglas de asociación, redes bayesianas, redes neuronales, SVM) y regresión.
- Descriptivas: encontramos las siguientes tareas: reglas de asociación, correlaciones, clustering, detección de anomalías.

La Minería de Datos puede aplicarse sobre cualquier tipo de repositorio de datos (BD Relacionales, Data Warehouses, BD Transaccionales, en la Web y en Sistemas de BD avanzados como BD Objeto-relacionales, BD Espaciales, BD Temporales, BD de Textos, BD Multimedia). Las técnicas a aplicar varían de acuerdo a cada tipo de repositorio [7].

3.2. Metodologías de Gestión de Proyectos de Minería de Datos Masivos

En los últimos años se han propuesto en el área de KDD nuevas técnicas que puedan gestionar Datos Masivos, pero también se han estudiado las metodologías que permiten acceder al nuevo conocimiento que se pretende encontrar. Las metodologías nos permiten llevar a cabo el proceso de minería de datos en forma sistemática y no trivial, definiendo además de las fases del proyecto, las tareas a realizar y cómo llevarlas a cabo. Según la filosofía de desarrollo, tanto las guías como las metodologías, se pueden clasificar según dos enfoques: tradicionales, que se basan en una fuerte planificación durante toda la gestión del proyecto y un ciclo de vida más lineal; y metodologías ágiles, en las que la gestión del proyecto es incremental, cooperativo, ampliamente adaptable y abiertas al cambio. El término KDD fue acuñado en el año 1996 y constituyó el primer modelo aceptado en la comunidad científica que establece las etapas principales de un proyecto de explotación de información. En su versión completa, KDD está formado por nueve etapas [8, 9]. A partir del año 2000, con el gran crecimiento en el área de minería de datos, surgen tres nuevos modelos que plantean un enfoque sistemático para llevar a cabo el proceso [10]: SEMMA, CRISP-DM y Catalyst (conocida como P3TQ). Algunas de las metodologías profundizan en mayor detalle sobre las tareas y actividades a ejecutar en cada etapa del proceso de minería de datos (como CRISP-DM), mientras que otras proveen sólo una guía general del trabajo a realizar en cada fase (como SEMMA). En este proyecto se estudiará la aplicación de metodologías ágiles en proyectos de minería de datos masivos.

4. Resultados y Objetivos

Con este proyecto se espera proponer modificaciones o mejoras a los algoritmos de clustering o de clasificación existentes para datos masivos, o proponer nuevos algoritmos, además del desarrollo de aplicaciones que implementen este tipo de algoritmos en el tratamiento de repositorios de datos masivos. A partir de la aplicación y comparación de metodologías de gestión de proyectos de enfoque ágil a estos proyectos de descubrimiento de conocimiento, se espera poder determinar

cual es la metodología mas adecuada a aplicar en proyectos de estas características.

5. Formación de Recursos Humanos

Este proyecto da inicio a una nueva línea de investigación dentro del Grupo de investigación en Bases de Datos de la Fac. Reg. Concepción del Uruguay de la U.T.N.. La Directora y la codirectora del proyecto, tambien lo son de las carreras de posgrado Especialización y Maestría en Ciencias de la Computación con orientación en Bases de Datos que se dictan en esta Facultad. Tres de los investigadores del proyecto estan desarrollando sus tesis de maestría luego de haber finalizado la cursada de dicho posgrado. Se cuenta con un becario graduado, que está iniciando su camino en la investigación y prevé la realización de un posgrado en el área temática del proyecto. Una de las integrantes del grupo está desarrollando su Tesis Doctoral sobre la temática de indexación en memoria secundaria de bases de datos textuales, tema íntimamente relacionado a las líneas de estudio de este grupo. El grupo cuenta en la actualidad con dos becarios alumnos de la carrera Ingeniería en Sistemas de Información que inician su formación en la investigación.

Referencias

[1] Fan Wei and Bifet Albert. Mining big data: Current status, and forecast to the future. *SIGKDD Explor. Newsl.*, 14(2):1–5, apr 2013.

[2] Robson Leonardo Ferreira Cordeiro, Christos Faloutsos, and Caetano Traina Junior. *Data Mining in Large Sets of Complex Data*. SpringerBriefs in Computer Science. Springer-Verlag London, 2013.

[3] Larose Daniel T. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience, 2004.

[4] Tsun Chow and Dac-Buu Cao. A survey study of critical success factors in agile software projects. *Journal of Systems and Software*, 81(6):961 – 971, 2008. Agile Product Line Engineering.

[5] Gabriella Cserhati and Lajos Szabo. The relationship between success criteria and success factors in organisational event projects. *International Journal of Project Management*, 32(4):613 – 624, 2014.

[6] J.H. Orallo, M.J.R. Quintana, and C.F. Ramirez. *Introduccion a la mineria de datos*. Fuera de coleccion Out of series. Editorial Alhambra S. A. (SP), 2004.

[7] J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2011.

[8] H. Dai, R. Srikant, and C. Zhang. *Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26-28, 2004, Proceedings*. Number v. 8 in Lecture Notes in Artificial Intelligence. Springer, 2004.

[9] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. *Commun. ACM*, 39(11):27–34, nov 1996.

[10] Gonzalo Mariscal, Oscar Marban, and Covadonga Fernandez. A survey of data mining and knowledge discovery process models and methodologies. *The Knowledge Engineering Review*, 25:137–166, 6 2010.