EL CONOCIMIENTO Y EL *DATA MINING*

Por Jorge Miller

OBJETIVO

Destacamos la importancia de las herramientas que sirven para dar soporte al proceso de toma de decisiones, entre las cuales se incluyen el *Data Mining* (DM) y el *Knowledge Data Discovery* (KDD). La relevancia de estas tecnologías está íntimamente asociada con la cultura de la organización que las emplea y con una profunda comprensión de la forma en que los decisores convierten los datos en información y la información en conocimiento. Sobre estas bases se asienta el *data mining*, cuyos algoritmos y técnicas principales bosquejamos e ilustramos con varios ejemplos de aplicación. Cerramos el trabajo con algunas reflexiones personales sobre la importancia de estas tecnologías para el crecimiento de la gente, sus organizaciones y sus países.

INTRODUCCIÓN

Las decisiones son actividades que realizamos todos los días. Aquellas que tienen que ver con nuestra vida diaria son rutinarias y están relativamente bien definidas. Sin embargo, las decisiones realmente importantes - las decisiones estratégicas- pueden ser difíciles de tomar, necesitan mucha información y, consecuentemente, herramientas que soporten el proceso de decidir.

Las herramientas de soporte para la toma de decisiones permiten complementar los recursos intelectuales de la gente con la capacidad de las computadoras para mejorar la calidad de las decisiones, especialmente cuando se trata de problemas no estructurados. Los problemas no están estructurados cuando sus objetivos entran en conflicto mutuo, sus alternativas de solución son difíciles de identificar, y la selección de una alternativa acarrea un alto nivel de incertidumbre.

Los decisores no deberían preocuparse por el diseño de las herramientas de soporte, sino que, por el contrario, se deberían poder ocupar de su *aplicación efectiva y estratégica*, para mejorar la calidad de la identificación de los problemas y de su solución. Para ello, las herramientas de soporte deben tener al menos las siguientes características:

- Pueden ser empleadas en contextos de decisión no estructurados.
- Dan soporte al decisor, pero no lo reemplazan.
- Soportan todas las fases del proceso decisor.
- Están enfocadas hacia la efectividad del proceso decisor.
- Están bajo el control del decisor.
- Utilizan datos y modelos.
- Facilitan el aprendizaje del decisor.
- Son interactivos y de uso amigable.
- Emplean un proceso evolutivo o iterativo.
- Proveen soporte desde el nivel ejecutivo más alto hasta los niveles de la línea.
- Soportan múltiples decisiones, independientes o interdependientes.

"Las herramientas de soporte para la toma de decisiones permiten complementar los recursos intelectuales de la gente con la capacidad de las computadoras para mejorar la calidad de las decisiones…"

Lo mínimo que se espera de las herramientas de soporte es que sirvan para expandir las capacidades de las personas para procesar grandes volúmenes de información durante el proceso de toma de decisiones. Aunque muchos de los componentes de una decisión puedan ser estructurados, ellos pueden ser muy complejos y demandar mucho tiempo del decisor. En ese caso, la herramienta de soporte tiene que poder liberarlo, para que utilice sus conocimientos y su tiempo en la parte de la decisión que no se encuentra estructurada. Las herramientas de soporte a las decisiones pueden incorporar sólo el conocimiento de sus diseñadores y servir

únicamente (lo cual no es poco) para procesar un conjunto acotado de habilidades; pero también pueden llegar a "pensar" o a "aprender", como veremos más adelante.

Ahora bien, la creación e incorporación de nuevo conocimiento que permite el *data mining* y el KDD ocurre en el ambiente cultural propio de cada organización. Es esta cultura organizacional la que moldea la forma en que la gente se interroga sobre su trabajo y da lugar a nuevas respuestas. Examinemos el concepto de cultura organizacional y un breve ejemplo de la forma en que afecta a la selección de las herramientas de soporte a las decisiones.

"…el éxito depende del conocimiento sobre el proceso aplicado por el decisor y, muy especialmente, de cómo ese proceso se integra en el marco de la organización."

LA CULTURA DE LAS ORGANIZACIONES

Teniendo en cuenta que la gente trabaja en el marco cultural de su organización, podemos ponderar la cultura organizacional sobre la base de las siguientes características:

- El grado de responsabilidad, libertad e independencia de los individuos que componen la organización.
- El grado en que los empleados son inducidos a ser agresivos, innovadores y tomadores de riesgo.
- El grado en que la organización crea objetivos y expectativas de desarrollo claras.
- El grado en que las distintas unidades de la organización son llevadas a operar de una manera coordinada.
- El grado en que los gerentes proveen información clara, asistencia y soporte a sus empleados.
- El grado de regulaciones para controlar la conducta de los empleados.
- El grado de identificación con la organización como un todo.
- El grado de crecimiento del empleado en relación a su desempeño.
- El grado en que los empleados son inducidos a exponer sus conflictos y manifestar sus críticas.
- El grado en que está restringida la comunicación organizacional.

Así como el estilo individual afecta a la estrategia de la decisión y a su calidad, también lo hace el entorno organizacional. Por esta razón, al implementarse las herramientas para el soporte de las decisiones, deben tenerse en cuenta las influencias de la cultura organizacional sobre el tipo de decisiones que se toman y sobre la conveniencia de emplear determinadas estrategias de decisión y herramientas de soporte. Por ejemplo, una organización orientada a la colaboración y el trabajo en equipo que esté comprometida con un esfuerzo de CRM (*Customer Relationship Management*) buscará la implementación de un sistema de *data mining* que no solo ofrezca las características generales de un sistema *standard* sino también herramientas para incorporar en sus proyectos *workgroup management* y otras herramientas para el trabajo en colaboración, que permitan diseñar sistemas de gran escala para toda la organización.

La cultura de las organizaciones, al igual que la cultura en general, evoluciona al compás de los nuevos conocimientos que son, a su vez, causa y efecto de las transformaciones culturales. Antes de seguir avanzando, hagamos un paréntesis para profundizar en el proceso del conocimiento y, más especialmente, en sus limitaciones: porque finalmente, y más allá de cualquier tipo de herramienta de soporte que se utilice, el éxito depende del conocimiento sobre el proceso aplicado por el decisor y, muy especialmente, de cómo ese proceso se integra en el marco de la organización. En la sección siguiente veremos la forma en que el conocimiento combina elementos, como los datos y la información, para dar lugar a una síntesis de conceptos que agregan sentido a los sistemas de apoyo a las decisiones.

DATO, INFORMACIÓN Y CONOCIMIENTO

Vivimos permanentemente bombardeados por informes sobre la importancia del conocimiento y la información para obtener ventajas económicas en un mundo globalizado. Podríamos escribir muchas hojas tratando de definir la palabra "conocimiento", pero aquí solo daremos breves ideas sobre el significado de las palabras *datos, información y conocimiento.*

Datos: son hechos, medidas u observaciones, que pueden presentarse (o no) en un contexto dado. Datos sin contexto son 60, 62, 66, 72. Los mismos datos, ahora con contexto, podrían representar el peso en kilogramos de Laura, Ana, Juan y Pedro, respectivamente. La validez y la efectividad de los datos vienen determinadas principalmente por su exactitud.

Información: Son los datos organizados de cierta manera, de forma tal que sean de utilidad y relevancia para quien tiene que resolver un problema de decisión. El criterio clave para evaluar la información es su utilidad.

"El conocimiento es una interpretación realizada por la mente, que será válida cuando pueda explicar las interacciones de un problema con su contexto"

Conocimiento: Es una combinación de instintos, ideas, reglas, procesos e información que un decisor aplica para guiar sus acciones y decisiones. El conocimiento es una interpretación realizada por la mente, que será válida cuando pueda explicar las interacciones de un problema con su contexto.

De lo anterior, se infieren dos puntos. El primero es que la información es personal; el segundo, que el conocimiento no es estático: es más, debe cambiar cuando cambia el entorno de decisión.

DATA MINING

El *data mining* es un conjunto de actividades utilizadas para encontrar en los datos contextos nuevos, ocultos o inesperados. Utilizando información contenida en un *data warehouse* (o "depósito de datos"), el *data mining* puede responder a preguntas que un decisor no hubiera formulado de no contar con estas herramientas.

Cada vez más se utiliza como sinónimo de *data mining* el término *Knowledge Data Discovery* (KDD). Es un término más descriptivo que *data mining*, y se aplica a todas las actividades y procesos relacionados con la actividad de descubrir conocimiento útil a partir de los datos. Usando una combinación de técnicas que incluyen el análisis estadístico, la lógica neuronal, la lógica difusa, el análisis multidimensional, la visualización de datos y los agentes inteligentes, el KDD puede descubrir patrones útiles para desarrollar modelos predictivos de conductas o de consecuencias, en una amplia variedad de dominios del conocimiento.

Hasta no hace mucho, las herramientas para soporte de las decisiones se basaban en el concepto de la *verificación.* Una base de datos relacional podía accederse para obtener respuestas dinámicas a preguntas bien formuladas (*queries*). La clave estaba en el conocimiento previo que la persona tuviera, que podía ser ampliado y verificado por el resultado de la búsqueda. El concepto de verificación se fue agotando; al sentirse cada vez más la necesidad del *descubrimiento*, está creciendo la demanda de técnicas diseñadas—como el KDD—para encontrar en los datos patrones de conducta nuevos y no clasificados.

Las categorías básicas de las técnicas de minería de datos actualmente en uso se pueden clasificar en: *clasificación, asociación, secuencia* y *cluster.*

Clasificación: incluye los procesos de minería de datos que buscan reglas para definir si un ítem o un evento pertenecen a un *subset* particular o a una clase de datos. Esta técnica, probablemente la más utilizada, incluye dos subprocesos: la construcción de un modelo y la predicción. En términos generales, los métodos de clasificación desarrollan un modelo compuesto por reglas IF-THEN y se aplican perfectamente, por ejemplo, para encontrar patrones de compra en las bases de datos de los clientes y construir mapas que vinculan los atributos de los clientes con los productos comprados. Con un conjunto apropiado de atributos predictivos, el modelo puede identificar los clientes con mayor propensión a realizar una determinada compra durante el próximo mes. Un caso típico de clasificación es el de dividir una base de datos de compañías en grupos homogéneos respecto a variables como

"posibilidades de crédito" con valores tales como "bueno" y "malo".

Asociación: incluye técnicas conocidas como *linkage analysis*, utilizadas para buscar patrones de las transacciones operacionales que tienen una probabilidad alta de repetición, como ocurre al analizar una canasta en la búsqueda de productos afines. Se desarrolla un algoritmo asociativo que incluye las reglas que van a correlacionar un conjunto de eventos con otro. Por ejemplo, un supermercado podría necesitar información sobre los hábitos de compra de sus clientes, para reubicar los productos que se suelen comprar juntos, para localizar los productos nuevos en el mejor lugar, o para ofrecer promociones.

Secuencia: los métodos de análisis de series de tiempo son usados para relacionar los eventos con el tiempo. A título ilustrativo: como resultado de este tipo de modelo se puede aprender que las personas que alquilan una película de video tienden a adquirir los productos promocionales durante las siguientes dos semanas; o bien, que la adquisición de un horno de microondas se produce frecuentemente luego de determinadas compras previas.

"…el conocimiento no es estático: es más, debe cambiar cuando cambia el entorno de decisión."

Cluster: Muchas veces resulta difícil o imposible definir los parámetros de una clase de datos. En ese caso, los métodos de clustering pueden usarse para crear particiones, de forma tal que los miembros de cada una de ellas resulten similares entre sí, según alguna métrica o conjunto de métricas. El análisis de *clusters* podría utilizarse, entre otras aplicaciones, al estudiar las compras con tarjetas de crédito, para descubrir—digamos—que los alimentos comprados con una tarjeta dorada de uso empresarial son adquiridos durante los días de semana y tienen un valor promedio de *ticket* de 152 pesos, mientras que el mismo tipo de compra, pero realizado con una tarjeta platino personal, ocurre predominantemente durante los fines de semana, por un valor menor, pero incluye una botella de vino más del 65 % de las veces.

Como puede verse, en el mundo del *data mining* las preguntas que se pueden hacer y responder son infinitas y las metodologías utilizadas para responderlas son diversas y cada vez más variadas. Así como existen muchas técnicas para dar soporte al proceso de decidir con la minería de datos, existen variadas tecnologías para construir los modelos, que mencionamos a continuación.

Análisis estadístico: Supongamos que el 70 % de las personas que compran el producto X usando una tarjeta de crédito también compran el producto Y, y que el producto Y nunca se vende independientemente. Resulta entonces relativamente fácil construir un modelo que ayude a predecir la compra del producto Y con una ocurrencia del 70 %. Por supuesto, será de mucho mayor interés poder predecir las compras del producto X. Para ello el *data mining* necesita de técnicas estadísticas capaces de manejar datos no lineales, múltiples *outliers* (datos inusualmente alejados del promedio) y datos no numéricos, como los que se encuentran en un ambiente de *data warehouse.* Las técnicas de regresión lineal, de gran difusión en múltiples aplicaciones, muchas veces no se pueden utilizar en la *data mining* por la complejidad de los patrones de los datos y su falta de linealidad.

Redes neuronales, algoritmos genéticos y lógica difusa: Las redes neuronales son estructuras matemáticas con capacidad de aprender, desarrolladas en el intento de reflejar en ecuaciones la forma en que el cerebro humano reconoce los patrones de datos e información. De esta forma es posible desarrollar modelos predictivos no lineales, que aprenden combinando variables y estudiando la forma en que ellas afectan a los conjuntos de datos. Las técnicas de las máquinas de aprender, entre las cuales incluimos a los algoritmos genéticos y la lógica difusa, tienen la habilidad de encontrarle significado a datos complicados e imprecisos, y pueden extraer patrones y detectar tendencias.

Los algoritmos genéticos combinan las nociones de supervivencia del más apto con un intercambio estructurado y aleatorio de características, entre individuos de una población de posibles soluciones, para conformar un algoritmo de búsqueda que puede aplicarse para resolver problemas de optimización en diversos campos. Imitando la mecánica de la evolución biológica en la naturaleza, los algoritmos genéticos operan sobre una población compuesta de posibles soluciones al problema. Cada elemento de la

población se denomina "cromosoma". Un cromosoma es el representante, dentro del algoritmo genético, de una posible solución al problema. La forma en que los cromosomas codifican a la solución se denomina "representación".

La lógica difusa es una forma de lógica en donde las variables pueden tener varios niveles de verdad o falsedad representados por rangos de valores entre el 1 ("verdadero") y el 0 ("falso"). El resultado de una operación se puede expresar tan sólo como una probabilidad y no necesariamente como una certeza. Por ejemplo, además de los valores "verdadero" o "falso", un resultado puede adoptar valores tales como "probablemente verdadero", "posiblemente verdadero", "posiblemente falso" y "probablemente falso". La lógica difusa es un acercamiento matemático para tratar con la naturaleza imprecisa del lenguaje cotidiano y del mundo que nos rodea. Su utilización es, por ejemplo, frecuente en temas relacionados con la selección de recursos humanos en donde se debe seleccionar al aspirante que reúna en mayor grado una cantidad de cualidades requeridas en base al mejor coeficiente de adecuación.

Árboles de decisión: Es una conceptualización matemática simple para seguir el efecto de cada evento o decisión en sucesivos eventos.

CONCLUSIONES

La explosión de datos, información, conocimientos, y, más aún, la explosión de las tecnologías para su tratamiento, han llevado al hombre a buscar nuevas formas de aprovechar tal riqueza, acumulada en forma de bits y bytes, y que, de otra forma, permanecería inexplorada. El data mining primero, y el KDD más recientemente, son respuestas a la pasión del hombre por conocer más.

En la sociedad del conocimiento se compite a velocidades desconocidas por todas las sociedades que la precedieron. El acceso universal al conocimiento pone al alcance de todos la posibilidad de mejorar el rendimiento de las organizaciones. Esto vale también para esas macro-organizaciones que son los países. Ya no habrá países pobres, habrá países ignorantes y, consecuentemente, pobres. Lo mismo ocurrirá con las empresas y las organizaciones. Y este movimiento también nos abarcará a cada uno de nosotros.

En este contexto, aprovechar el potencial de las nuevas herramientas como el *data mining* para enriquecer y nutrir el conocimiento es una oportunidad que no debe desaprovecharse. Debemos difundir las ideas sobre estos sistemas y sus aplicaciones y acercarnos a las nuevas tecnologías para conocer mejor su potencial.

BIBLIOGRAFÍA

- Berry, Michael J. A. y G. Linoff, 1997, *Data Mining Techniques: for Marketing, Sales, and Customer Support*, John Wiley and Sons.
- Bowers, D. G., y S. L. Seashore, 1966, "Predicting Organizational Effectiveness with a Four Factor Theory of Leadership", *Administrative Science Quarterly*, 11 (September).
- Clemen, R. T., 1991, *Making Hard Decisions: An Introduction to Decision Analysis*, Belmont, Duxbury Press.
- Culman, M. J. y B. Gutek, 1989, "Why Organizations Collect and Store Information", en: Stohr, E. A. y B. R. Konsynski, *Information Systems and Decision Processes*, Los Alamitos, IEEE Computer Society Press.
- Fayyad, Usama, Gregory Piatetsky-Shapiro y Padhraic Smyth, 1996, "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM*, pp. 27-34.
- Golub, A. L., 1977, *Decision Analysis: An Integrated Approach*, New York, Wiley.
- Gray, P., 1997, "The New DSS: Data Warehouse, OLAP, MDD and KDD", en: *Proceedings of the Americas Conference on Information Systems*, editado por Carey, Jane M., Phoenix, Association for Information Systems.
- Harrison, E. F., 1995, *The Managerial Decision-Marketing Process*, Boston, Houghton Mifflin Co.
- Inmon, W. H., J. D. Welch y K. L. Glassey, 1997, *Managing the Data Warehouse*, New York, Wiley.

- Watson, H. J., G. Houdeshel y R. K. Rainer, 1997, *Building Executive Information Systems and Other Decision Support Applications*, New York, Wiley.
- Yolis, E., P. Britos, J. Sicre, A. Servetto, R. García-Martinez y G. Perichinsky, 2003, "Algoritmos Genéticos Aplicados a la Categorización Automática de Documentos", ITBA.