Leonardo Franco Calpa

Bryan Camilo Grueso

Anderson Ramírez

* **Recopilación de información necesaria**

**Descubrir conocimiento en grandes datos: Allers.**

Allers Group es una empresa colombiana especializada en importación al por mayor y al detal; de equipos médicos, insumos hospitalarios, instrumental quirúrgico y medicamentos. La empresa cuenta con marcas propias y exclusividades de compañías líderes en el mundo.

En la industria farmacéutica la empresa ha logrado la fidelización de las droguerías gracias a que cuentan con el respaldo de grandes compañías para la distribución de sus productos farmacéuticos.

Teniendo en cuenta que la empresa Allers Group ha ido almacenando información durante quince años de sus ventas, clientes, proveedores, entre otros, con la ayuda de un software de gestión empresarial (SAP) y que de igual manera desea sacarle provecho a toda esa información almacenada con el fin de impulsar el crecimiento de la empresa. Es necesario que se tenga en cuenta ciertos puntos. Como por ejemplo el tamaño de la información recaudada; la forma en la que está almacenada la información: formato, estructura, organización. Todo esto con el fin de saber si hay que reestructurarla en un modelamiento posterior.

Como lo dio a conocer Oscar Rendón, CTO de Allers Group, se estima que se hacen aproximadamente entre 300 a 600 registros de ventas por día, sin contar la información generada por otras áreas de la empresa. Con base en lo anterior, podría decirse que la información anualmente puede alcanzar registros del orden de 15x104, y contando desde 15 años atrás, el total de la información puede estar del orden de 2x106 registros, los cuales pueden presentar muchas dimensiones en la base de datos donde están alojados.

Como gran parte de los datos hacen referencia a la información de los clientes, proveedores, fechas, cantidades, es muy probable que se encuentre en formato de texto, por lo cual habría que darle un tratamiento a la información para estructurarla y así poder clasificarla. También, como la información ha sido recolectada mediante el programa de gestión empresarial SAP, puede que no haya información duplicada que aumente innecesariamente el tamaño de la base de datos.

**Problemática similar**

Un caso aplicable de la misma problemática que presenta Allers, se dio en una las compañías y a su vez plataforma virtual más exitosas a nivel mundial: Netflix.

En un momento dado, la compañía observó que la mayoría de sus usuarios se vinculaban a la plataforma por disfrutar alguno que otro contenido en específico, cuando esto sucedía los usuarios completamente satisfechos y sin ganas de observar otro tipo de contenido, ya que la plataforma (en ese entonces) no vinculaba contenido relacionado con el inmediatamente anterior al que habían terminado de disfrutar.

Los altos mandos de la compañía, en especial Todd Yellin (vicepresidente de productos) al ver esto, decidieron implementar un sistema de recomendaciones patentado, donde se usa técnicas de data mining y machine learning.

“Lo que hacemos es buscar personas a las que les gustan contenidos parecidos que los que te gustan a ti y los agrupamos en comunidades. Luego de esto, podemos generar un contenido previamente analizado en el cual podemos mostrar contenido relacionado y con alta de probabilidad de que te guste “

Factores como lo que ven, la velocidad con que ven la serie, en qué horarios, el dispositivo que utilizan, la edad y género; son los que se toman en cuenta.

**Recopilación**

Retomando nuestro caso de grandes de información de Allers Group, para lograr sacar utilidad de toda la información pudimos encontrar que se es necesario llevar a cabo metodologías que permitan transformar simples datos en indicadores, esto con el fin realizar predicciones, explicaciones y clasificaciones, para tal fin se requiere de metodologías de análisis de datos, **KDD (*Knowledge Discovery in Databases*).**

La metodología de KDD consiste principalmente en un proceso para identificar patrones en los datos, que sean de gran utilidad para la finalidad que se desea alcanzar, y de ese modo convertir información de bajo nivel en conocimiento de alto nivel. Como lo explican en el artículo “Minería de Datos: Conceptos y tendencias” el KDD consta de:

**1**. Comprender el dominio de aplicación:

Es decir que incluya el conocimiento relevante previo y las metas de la aplicación.

**2.** Extraer la base de datos objetivo:

Es decir, recoger los datos, evaluar la calidad de los mismos y utilizar análisis exploratorio para familiarizarse con ellos.

**3.** Preparar los datos:

Incluye limpieza, transformación, integración y reducción de datos. Se intenta mejorar la calidad de los mismos a la vez que disminuir el tiempo requerido por el algoritmo de aprendizaje aplicado posteriormente.

**4**. Minería de datos:

Como se ha señalado anteriormente, este es la fase fundamental del proceso. Está constituido por una o más de las siguientes funciones: clasificación, regresión, clustering, resumen, recuperación de imágenes, extracción de reglas, etc.

**5.** Interpretación:

Explicar los patrones descubiertos, así como la posibilidad de visualizarlos.

**6.** Utilizar el conocimiento descubierto:

Hacer uso del modelo creado

(Riquelme, Ruiz, Gilbert, 2006)

**Big Data.**

Son un conjunto de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales, tales como bases de datos relacionales y estadísticas convencionales o paquetes de visualización, dentro del tiempo necesario para que sean útiles.

Los datos captados de Big Data pueden provenir de una gran cantidad de fuentes, por ejemplo, de máquinas, sensores, smartphones, correos, videos, redes sociales, tendencias de mercados, etc.

El lugar donde son almacenados los datos en el Big Data es en ficheros distribuidos, además permite el análisis de datos históricos y hasta incluso en tiempo real.

(Molina, 2016)

**Minería de Datos.**

En la minería de datos se intenta hallar patrones con mucha significancia en los conjuntos de datos para poder construir modelos predictivos, descriptivos y clasificadores. Para eso se utilizan técnicas de aprendizaje automático (*Machine Learning*) y programación en bases de datos.

Cuando hablamos de ***ML,*** sabemos que existen dos grandes grupos de algoritmos que se usan para abstraer la información*: aprendizaje supervisado* y *no supervisado*. En ambos se recibe como entrada un data *set* y producen como salida un modelo de los ya mencionados.

De igual manera, para la generación de un modelo basado en ***MD*** puede seguirse una serie de pasos que se describen en el artículo *Introducción a la minería de datos y el data Warehousing* por Coria:

La generación de un modelo basado en MD consiste en los siguientes pasos:

**1**) Definición del objetivo del modelo.

**2**) Selección de datos para análisis y modelación y de sus fuentes.

**3**) Recolección, limpieza y pre-procesamiento de datos.

**4**) Análisis estadísticos básicos.

**5**) Selección y aplicación de algoritmos de aprendizaje automático.

**6**) Reporte y evaluación de hallazgos con el experto de dominio.

**7**) Explotación de los hallazgos.

(Coría, 2016)

Además de la minería de datos se pudo recopilar información de unas técnicas usadas en el campo de Big Data y en menor medida por minería de datos, conocidas como reglas de asociación.

**Reglas de asociación.**

Son un conjunto de reglas y algoritmos que se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos, de igual manera se sabe que existen diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

Algunos de las operaciones que se deben tener en cuenta a la hora de genera reglas de asociación son:

*Fuerza bruta:*

Esta operación enumera todas las reglas de asociación posibles y de paso calcula el soporte y la confianza de cada regla. De igual manera elimina las reglas que no superen los umbrales de soporte y confianza (MinSupp y MinConf).

*A – priori*:

Esta operación sobre bases de datos transaccionales, permite encontrar de forma eficiente "conjuntos de ítems frecuentes", los cuales sirven de base para generar reglas de asociación. Procede identificando los ítems individuales frecuentes en la base y extendiéndolos a conjuntos de mayor tamaño siempre y cuando esos conjuntos de datos aparezcan suficientemente seguidos en dicha base de datos.

**REFERENCIAS**

* José C. Riquelme, Roberto Ruiz, Karina Gilbert. Minería de Datos: Conceptos y Tendencias, 2006
* Molina, J. (2016). *PowerData.* Obtenido de PowerData: <https://www.powerdata.es/> big-data
* Coria, Sergio. Introducción a la Minería de Datos y el Data Warehousing, 2016
* J. Hernández-Orallo, M. J. Ramírez-Quintana and C. Ferri. Introducción a la Minería de Datos. Prentice Hall / Addison-Wesley, 2004.
* Berzal, F. (2015). *Elvex.* Obtenido de Elvex: <https://elvex.ugr.es/>