CAPÍTULO 1 Introducción

Desde hace ya mucho tiempo, la información es considerada uno de los recursos más valiosos que alguien pueda poseer, no importa la fuente. Los investigadores, espías, encuestadores, científicos, analistas de mercados, etc., todos son ejemplos de personas que trabajan para obtener ese recurso vital. La capacidad de obtener información ha sido desde tiempos antiguos un factor determinante en el éxito o el fracaso de una entidad particular, ya sea una nación, un ejército, o, en el caso de la sociedad actual, una empresa.

Al ser la información, básicamente, uno de los recursos más importantes para las empresas, cantidades enormes de otros recursos (dinero, personas, etc.) se han enfocado a su obtención, procesamiento y análisis. De ahí se han creado los famosos sistemas de información, que son los sistemas dedicados a la administración y procesamiento de la información, ya sea manual o automático, y que pueden consistir de gente, metodologías o computadoras para su operación.

Cuando se trata de sistemas de la información, las ciencias computacionales han demostrado ser una herramienta muy útil para su éxito y desarrollo adecuado, debido a las capacidades de las computadoras de ejecutar una cantidad enorme de operaciones por segundo y de almacenar grandes cantidades de datos. A tal grado están asociadas las computadoras con la información, que el término sistemas de información más bien ya quiere decir *sistemas computacionales de información*.

Sin embargo, una cosa es poseer la información, y otra muy distinta es saber qué hacer con ella.

1.1 Minería de datos

La gran cantidad de datos almacenados (de cualquier tema, campo y tipo) hoy en día está sobrepasando rápidamente la capacidad de los investigadores y científicos para interpretar esos datos y descubrir información relevante e interesante en ellos [ ]. Podemos tener datos científicos, datos médicos, demográficos, financieros, geográficos, etc., y su creación y almacenamiento han tenido un crecimiento explosivo. Los seres humanos simplemente no tenemos ni el tiempo ni las habilidades necesarias para poder analizar tal cantidad de información. Estos problemas han generado una necesidad urgente de nuevas tecnologías y herramientas automáticas capaces de ayudarnos a transformar esos datos en información y conocimiento útiles o encontrar patrones repetitivos. Una de esas tecnologías es la minería de datos.

Puesto de manera simple, la minería de datos se refiere a la extracción o minado de conocimiento a partir de grandes cantidades de datos ["Data mining"]. También se le conoce como Descubrimiento de Conocimiento en Bases de datos (KDD, por sus siglas en inglés). La minería de datos es un campo multidisciplinario, por lo menos desde el punto de vista computacional, ya que incluye conceptos y temas como la inteligencia artificial, tecnologías de bases de datos, redes neuronales, etc.

La principal razón por la que la minería de datos ha llamado tanto la atención y ha crecido tanto en años recientes es debido a la alta disponibilidad de datos e información y a la necesidad de analizar esos datos para convertirlos en algo útil. Este conocimiento obtenido puede ser utilizado en temas tan diversos como administración de negocios, análisis de mercados, deportes, ingenierías, sistemas geográficos, entre otros, comprobando la gran cantidad de aplicaciones posibles de esta tecnología.

Pero incluso ya dentro de la propia minería de datos, se pueden aplicar otras técnicas para que funcione de manera más eficiente.

1.2 Minería de datos asistida

Existen un sinfín de métodos disponibles para ayudar a una aplicación de minería de datos a realizar su trabajo. Desde parámetros opcionales hasta patrones predeterminados a buscar, las opciones disponibles son todas útiles para refinar la aplicación. Un ejemplo de uno de estos métodos podría ser la eliminación de datos difusos o incompletos antes de realizar una minería sobre de ellos.

Una opción alternativa para la minería asistida es la de poder dar una retro-alimentación en tiempo real acerca de la minería que se está realizando en un momento dado. Esta retro-alimentación debe de proveer información útil para el usuario acerca de lo que está haciendo, en lo que podría considerarse como una guía o ayuda. Esta información puede ser acerca de patrones previamente

encontrados, de parámetros de minería utilizados o de cualquier otra cosa que un usuario pueda considerar útil.

Otra opción disponible dentro de este concepto de minería asistida es la capacidad de realizar anotaciones que queden registradas dentro de los mismos resultados obtenidos (o patrones encontrados), proporcionándole así al usuario una forma de recordar observaciones u opiniones que haya tenido acerca de un resultado en específico.

1.3 Ventajas de la minería asistida

Al hacer uso de las dos opciones descritas anteriormente, se llegan a evitar cierto tipo de problemas de naturaleza recurrente dentro de la minería de datos.

El más grande de estos problemas consiste en la repetición de trabajo ya realizado. Al utilizar una aplicación de minería de datos, no se puede saber qué patrones o conocimiento ya han sido encontrados, ni tampoco se puede saber de qué tipo de datos fueron obtenidos esos resultados. Esta situación indeseable puede ser evitada mediante el uso de un sistema de guía. También, como una característica adicional, esta guía le puede permitir al usuario realizar una comparación de resultados al modificar de manera mínima alguna de las opciones iniciales.

Otro problema consiste en que, aún cuando se pueda consultar un patrón previamente obtenido, no podemos saber si ese resultado le fue útil o no al usuario que lo obtuvo (a menos, claro, que el usuario sea el mismo), o si tiene alguna recomendación para mejorar esa minería específica. Incluso en el caso

de que el usuario sea el mismo, puede que haya pasado un período de tiempo considerable desde la última vez que realizó una minería y es una posibilidad el que no recuerde para qué la hizo o si sí le sirvió.

1.4 Definición del problema

El estudiante de doctorado Manuel Pech Palacio desarrolló, para propósitos de su tesis, una aplicación de minería de datos, la cual es alimentada con datos espaciales (llamados también datos geográficos). Esta aplicación no cuenta con ningún tipo de minería asistida, por lo que no puede considerarse que esta sea una situación óptima. El desempeño de la aplicación puede incrementarse mediante la implementación de los métodos de la minería asistida descritos anteriormente, proporcionándole también al usuario una productividad más alta y reduciendo el tiempo que pasa el usuario en la aplicación.

1.5 Objetivo general

La implementación de métodos de minería asistida dentro de una aplicación en específico (la creada por Manuel Pech Palacio). Las funciones a agregar son la de guía de minerías previamente realizadas y la de anotaciones sobre los patrones encontrados.

1.6 Objetivos específicos

- Desarrollar este proyecto usando el lenguaje de programación Java.
- Desarrollar el proyecto utilizando la librería de Java *javax.swing* para todos los componentes gráficos.
- Se requiere el almacenamiento en una base de datos de toda la información relevante de un patrón encontrado.
- La creación de dicha base de datos.
- Se quiere que, al momento de estar ingresando los parámetros de la minería de datos, el sistema provea retro-alimentación visual en tiempo real acerca de los patrones que ya se han encontrado usando esos mismos parámetros, para que puedan ser usados como referencia o guía, o simplemente para evitar repetir una minería ya realizada.
- Poder representar de manera gráfica cualquier patrón previamente almacenado, y además con la capacidad de eliminarlo.
- Implementar la funcionalidad de hiper-grafo (descrito posteriormente) a los patrones recuperados de la base datos.
- Permitir la inclusión de parámetros de entrada *equivalentes*; es decir, parámetros que puedan ser sustituidos por otros y que para efectos prácticos signifiquen lo mismo para el usuario (por ejemplo, permitir que la capa geográfica de Terrenos pueda ser equivalente a la capa de Edificios).

- Proporcionar la capacidad de realizar anotaciones sobre los patrones encontrados. Estas anotaciones deberán guardarse en la base de datos junto con los patrones.

1.7 Recursos a utilizar

Recursos de Software:

El lenguaje de programación usado en este trabajo es Java en su versión 1.5.0. La base de datos de patrones se creó utilizando Oracle 10g. El lenguaje SQL se utilizó para toda la creación y administración de la base de datos. De manera muy especial, se hace uso de la librería Grappa (***Grap***h ***pa***ckage) creada en Java por John Mocenigo, de AT&T, para la manipulación y despliegue de grafos. Grappa hace uso de la paquetería *GraphViz*, un programa de dibujo de grafos, que incluye el algoritmo *dot*, el cual es el que se encarga de dar las instrucciones para dibujar un grafo en la pantalla. Finalmente, todo este proyecto se acopla a la aplicación de Manuel Pech Palacio.

1.8 Organización del documento

- **CAPÍTULO 1** – *INTRODUCCIÓN*: En este capítulo se define la problemática a resolver.

- **CAPÍTULO 2** – *MINERÍA DE DATOS Y CONCEPTOS GENERALES*: Se dará una descripción detallada acerca de lo que es minería de datos y de los conceptos, tecnologías y herramientas utilizadas en el proyecto.

- **CAPÍTULO 3** – *TRABAJOS RELACIONADOS*: Se describe el programa usado por Manuel Pech Palacio y la tesis de Víctor Manuel González Carrillo.

- **CAPÍTULO 4** – *ANÁLISIS, DISEÑO E IMPLEMENTACIÓN*: Aquí se definen los requisitos y las características de diseño que tendrá el proyecto, así como la implementación del mismo.

- **CAPÍTULO 5** – *RESULTADOS, CONCLUSIONES Y PERSPECTIVAS*: Conclusiones finales y trabajo a futuro.