Minería de datos en la detección de desperfectos en el alumbrado público

Sergio Quiroga, Mag. Alejandra Malberti, Mag. Raúl Klenzi

Instituto de Informática / Departamento de Informática /Facultad de Cs. 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 - 4260394 - 4264721 - 4234129, Fax 0264-4234980, http://www.exactas.unsj.edu.ar
{sergiooquiroga,amalberti,rauloscarklenzi}@gmail.com

Resumen

El presente estudio pretende extraer información que permita descubrir la incidencia de factores climáticos en la cantidad y tipología de fallas habituales que se producen en el alumbrado público. Se recurre a datos provenientes de reclamos de usuarios, a datos climáticos inherentes a las condiciones particulares del tiempo registradas para las fechas tratadas, y a datos geográficos relativos a la ciudad Capital de San Juan. Previo al análisis de los datos por medio de Minería de Datos, se consideran aspectos tales como ruido, datos ausentes, y volatilidad, entre otros. A la vez se normalizan los datos correspondientes a las direcciones de los reclamos, para poder realizar tareas de geoposicionamiento, tendientes a facilitar la identificación de las zonas afectadas.. En esta propuesta se aplica la metodología CRISP-DM y se utilizan las herramientas de software libre Rapidminer y Saga GIS.

Palabras clave:

Minería de datos - Minería de texto- Geoposicionamiento- RapidMiner-Saga Gis

Contexto

La línea de investigación que permite elevar la presente propuesta está contenida en el proyecto bianual "Extracción de Conocimiento en Datos Masivos" aprobado por CICITCA-UNSJ sujeto a evaluación externa.

Los datos procesados provienen de "reclamos" generados en el call-center de una empresa local, que se encarga del mantenimiento del alumbrado público en la ciudad de San Juan. Estos reclamos son generados por los vecinos que detectan fallas en el alumbrado público y solicitan, vía telefónica, su reparación. Este reclamo es transcripto en una planilla Excel por el operador receptor del llamado, quien registra la dirección postal o catastral correspondiente a la ubicación de la luminaria, o conjunto de luminarias, que presentan desperfectos. Los datos recepcionados son normalizados por medio de la aplicación de tareas de minería de texto, empleo de métricas de aproximación sintáctica, y uso de una base datos correspondientes a datos catastrales de la ciudad de San Juan suministrados por centro de Fotogrametría, Cartografía y Catastro de la Facultad de Ingeniería de la

Universidad Nacional de San Juan-CEFOCCA-FI-UNSJ.

Introducción

Para las organizaciones en general es muy importante disponer de sistemas eficientes, especialmente en lo que refiere a la gestión de los datos usados para la toma de decisiones tendientes a alcanzar el mayor beneficio con los menores costos posibles. Además, si se trata de una organización que brinda servicios, existe la necesidad de optimizar las operaciones que realiza de tal forma que sus usuarios se encuentren satisfechos. En la actualidad esto se puede lograr gracias a la enorme cantidad de datos digitalizados disponibles, los que pueden ser tratados por medio de técnicas adecuadas, entre las que se encuentra la Minería de Datos (MD).

Para Witten (Witten, 2005) la Minería de Datos es el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

Este proceso conlleva técnicas de análisis de datos destinadas a extraer patrones, describir tendencias y regularidades, predecir comportamientos y, en general, aprovechar un gran volumen de información digitalizada. Perez Lopez (Perez Lopez, 2006) divide al proceso de Minería de Datos en cuatro fases: Selección de Objetivos, Preparación de los Datos, Obtención de Patrones y Análisis e Interpretación de los Resultados.

Es muy importante tener en cuenta que una de las etapas primordiales y que consume más tiempo, en cualquier proyecto en el que se involucran grandes volúmenes de datos, es la preparación de los datos. En este aspecto coinciden la mayoría de los profesionales del área, ya que puede significar la diferencia entre el éxito y el fracaso de un proyecto; en general ocurre que los datos con los que se trabaja suelen no ser homogéneos. Además antes de realizar cualquier análisis, se deben tener en cuenta aspectos tales como ruido, datos ausentes, volatilidad, entre otros. Esto es contemplado en otras definiciones, como la provista por el Instituto SAS[1] que considera a la Minería de Datos como el proceso de Seleccionar, Explorar, Modificar, Modelizar y Valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores.

Por otro lado en su ensayo introductorio Miller y Han (2009 2ed) observan que el descubrimiento de conocimiento en bases de datos geográficas (Geographic Knowledge Discovery –GKD-) es un caso especial y no trivial de KDD-Knowledge Discovery Database. Los autores señalan que esto se debe en parte al carácter distintivo del marco de medición geográfica en el que existen problemas resultantes de la dependencia y la heterogeneidad espacial, la complejidad de los objetos y las reglas espacio-temporales, como así también de la diversidad de tipos de datos geográficos [Salvador, Resmini 2014]

La tecnología de los SIG-Sistemas de Información Geográficos- busca articular

[1]SAS Institute fue fundado por Anthony Barr, Jim Goodnight , John Sall y Jane Helwig el 1 de julio de 1976. Su nombre es el acrónimo de *statistical analysis systems* (sistemas de análisis estadístico). Es uno de los principales fabricantes de business intelligence software. Fuente: Wikipedia.org

las bases de datos gráficas con las bases de datos alfanuméricas que representan los diferentes rasgos del territorio, tales como caminos, cursos de agua, asentamientos poblacionales, actividades económicas, etc.

El presente trabajo involucra a una empresa que se dedica al mantenimiento del alumbrado público en diversos departamentos de la provincia de San Juan. Esta empresa se ha propuesto mejorar el servicio que brinda a sus usuarios, especialmente en lo que refiere a la detección temprana de problemas en el alumbrado causados por fenómenos climáticos. Para ello se tienen en cuenta los reclamos realizados por los usuarios: alumbrado público apagado (APA), alumbrado público prendido de día (APP), lámparas apagadas (LA), lámpara prendida de día (LPP), lámpara que prende y apaga (LPA), cables cortados (CC), lámparas tapadas por las ramas (LTxR) y globos apagados (GA); así como los datos recopilados sobre condiciones climáticas ocurridas en fechas determinadas – velocidad del viento, temperatura, presión atmosférica a nivel del mar, entre otros. Además, para facilitar la identificación de las zonas afectadas, se consideran también los datos georeferenciados correspondientes a los reclamos. Cuando se habla de datos georreferenciados, se trata de datos referidos a una posición con respecto a un sistema de coordenadas terrestres.

En la Figura 1 se presenta el diagrama en bloques correspondiente a la sucesión de aplicaciones que han permitido marcar en el mapa de la ciudad de San Juan el reclamo realizado por un vecino.

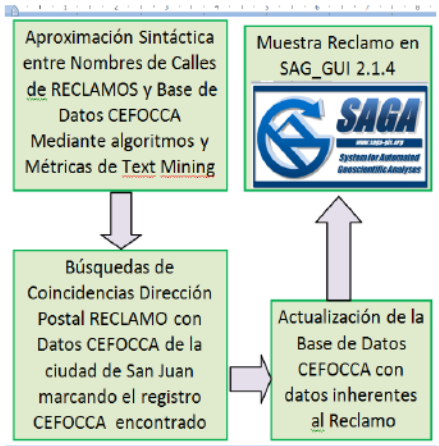

Figura 1 - Diagrama en bloques de procesos realizados en RapidMiner y Saga GIS

La aproximación sintáctica entre Nombres de calles de Reclamos y Base de Datos de CEFOCCA, se realizó por medio del proceso RapidMiner, presentado en la Figura 2, que aplica algoritmos y métricas de text mining.

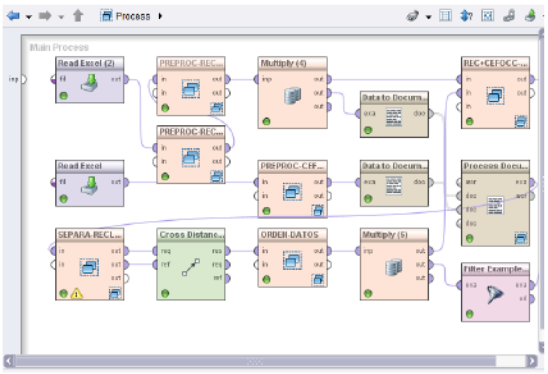

Figura 2 - Proceso RapidMiner de aproximación sintáctica

Como resultado, la tarea de preprocesamiento puesta de manifiesto por la sucesión de bloques presentados en la Figura 1 permitió ubicar los reclamos en el mapa correspondiente a la ciudad de San Juan. La Figura 3 muestra el mapa en el entorno Saga Gis. En él, las líneas más gruesas representan una mayor cantidad de reclamos recibidos, referidos a las cuadras marcadas.

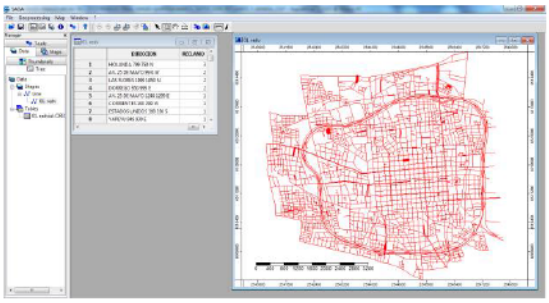

Figura 3 - Ubicación de reclamos en la ciudad de San Juan, en el entorno Saga GIS

Líneas de Investigación, Desarrollo e Innovación

El proyecto "Extraccion de Conocimiento en Datos Masivos", brinda el marco de contención al presente trabajo. Este proyecto continua las líneas de investigación de proyectos anteriores, tales como "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuario", extendiéndolas al tratamiento de grandes cantidades de datos, haciendo uso de hardware paralelo mediante la aplicación de CPU multinúcleos, Unidades de Procesamiento Gráfico GPUs, cluster de computadoras y herramientas de software libre que soporten dicho hardware, tengan implementaciones de algoritmos de minería de datos paralelos, como así también permitan la escritura de nuevas propuestas. Justamente una de estas herramientas de software libre RapidMiner- RM- 5.3.013 bajo licencia AGPL es la que el grupo ha utilizado para el preprocesamiento y posterior análisis sobre los datos de reclamos de usuarios.

En este caso se aplican secuencialmente las tareas de: 1) ingreso de datos, ajustando los parámetros de los diferentes módulos de RM al formato de los datos de entrada 2) preprocesamiento, filtrados, ajustes y adaptación de formatos de datos, necesarios para la posterior aplicación de Minería de Texto así como de diferentes estrategias, siendo estas segmentación, clasificación entre otras.

Con los resultados obtenidos de las tareas mencionadas, se recurre al sistema libre de información geográfica SAGA GIS- *System for Automated Geoscientific Analyses,* Geographic Information System. Este sistema, que posee capacidades para el procesamiento y análisis de datos geográficos, es usado para ubicar los reclamos de los usuarios en el mapa de la ciudad.

Asimismo se pretende incursionar en el tratamiento de datos georeferenciados en otros entornos ampliamente difundidos, como es el caso de Google Earth.

Objetivos y Resultados

El objetivo que se persigue es detectar factores climáticos que ocasionen fallas en el alumbrado público, a partir de la aplicación de estrategias de minería de datos, análisis de factores climáticos e incorporación y procesamiento de información relativa a geoposicionamiento.

Al momento de la elaboración de este documento, se han obtenido los siguientes resultados**:**

- Profundización en el conocimiento de las tareas de minería de datos y de minería de texto.
- Obtención de bases de datos de distintas fuentes- Reclamos de usuarios, datos meteorológicos y redvial de la ciudad de San Juan.
- Detección y solución de inconsistencias en los datos.
- Normalización automática de las direcciones correspondientes a las bases de datos de Reclamos a partir de las Direcciones provistas por CEFFOCA, por medio la aplicación de Minería de Datos con Rapid Miner.

- Estudio de técnicas y herramientas de geoposicionamiento.
- Interacción efectiva con el sistema Saga Gis.

Formación de Recursos Humanos

Las tareas desarrolladas en el ámbito del proyecto han permitido la conclusión y desarrollo actual de muchos trabajos finales de grado pertenecientes a alumnos de las carreras contenidas en el ámbito del Departamento Informática de la Facultad de Ciencias Exactas Físicas y Naturales. En este contexto, se han desarrollado y defendido recientemente tres trabajos finales de grado a saber:

- **Extracción de Conocimiento desde datos de un Banco de Germoplasma.** Caso de estudio: Instituto de Investigación y Desarrollo Agroindustrial Hortícola Semillero (INSEMI)
- **Descubrimiento de Conocimiento en Portales Institucionales.** Caso de aplicación: Usuarios de la Sociedad Franklin Biblioteca Popular de la Provincia de San Juan.
- **Minería Web de Uso.** -Caso de estudio sitio http://www.kdnuggets.com/

que permitieron obtener no sólo titulaciones en Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información sino también interactuar, como también lo admite el desarrollo de la presente propuesta, con organizaciones externas al ámbito académico donde se imparten las carreras mencionadas.

Referencias

- Mark Salvador, Ron Resmini (auth.), Guido Cervone, Jessica Lin, Nigel Waters (eds.) -**Data Mining for Geoinformatics._ Methods and Applications**- Springer New York (2014)
- Harvey J. Miller, Jiawei Han- **Geographic Data Mining and Knowledge Discovery**, Second Edition (Chapman & Hall CRC Data Mining and Knowledge Discovery Series)-CRC Press (2009)
- Hernández Orallo, José; Ramírez Quintana, Ma José; Ferri Ramírez, César. (2004) **Introducción a la Minería de Datos**. Edit. Pearson educación
- Microsoft (2013) Prueba y validación (minería de datos) http://msdn.microsoft.com/es-AR/library/ms174493.aspx
- Molina L. (2002) **Data mining: torturando a los datos hasta que confiesen.** http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html.
- Pérez López, César; Santín González, Daniel. (2006) **Minería de Datos Técnicas y Herramientas**. Edit. Alfaomega Grupo Editor
- Pyle, Dorian; Kaufmann, Morgan. (1999) **Data Preparation for Data Mining.** Edit. Morgan Kufmann Publishers Inc.
- Witten, Ian H.; Eibe, Frank. (2005) **Data Mining Practical Machine Learning Tools and Techniques**. Edit. Elsevier Inc.