Descubrimiento de patrones de conocimientos para la reducción de accidentes graves en autopistas

Fontana, Nicolás – Malgor, Rafael – Puñales Donato, Joaquín Ignacio

Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay, Explotación de Información

Abstract

En este trabajo se realiza una breve reseña de los entregables propuestos por la metodología de Explotación de Información CRISP–DM, aplicada en la detección de patrones de conocimiento para la reducción de accidentes graves en autopistas. El proyecto se realizó a partir de datos brindados por AUSA (Autopistas Urbanas Sociedad Anónima), en el marco de la cátedra electiva 'Tecnologías para la Explotación de Datos' del 5º año de Ingeniería en Sistemas de Información, con el fin de mejorar la seguridad vial mediante la implementación de políticas de seguridad no arbitrarias, basada en información histórica objetiva.

Palabras Claves

Explotación de Información, CRISP-DM, Descubrimiento de Reglas, AUSA (Autopistas Urbanas Sociedad Anónima)

Introducción

La Explotación de Información consiste en la extracción de conocimiento no trivial que reside de manera implícita en los datos disponibles en las distintas fuentes de información[1]. Dicho conocimiento es previamente desconocido y puede resultar útil para algún proceso[2]. Para un experto, o para el responsable de un sistema de Información, normalmente no son los datos en sí lo más relevante, sino el conocimiento que se encierra en sus relaciones, fluctuaciones y dependencias[3]. Con las técnicas de Explotación de Información se aborda la solución a problemas de predicción, clasificación y segmentación[4].

En el presente proyecto, se aplican dos procesos de Explotación de Información[5] sobre los datos proporcionados por el organismo administrador de las autopistas urbanas de la ciudad autónoma de Buenos Aires denominado AUSA, referente a accidentes ocurridos en las autopistas concesionados por este organismo y datos del paso vehicular por los puestos de peaje de las rutas concesionadas por el mismo.

Este proyecto nace con la necesidad de descubrir información novedosa, útil e implícita que de soporte a la toma de decisiones para la creación de políticas de seguridad en pos de la disminución de los accidentes de tránsito fatales y/o con pasajeros heridos en las autopistas mencionadas. En función de esta necesidad, se plantearon tres objetivos de requisitos:

El primero plantea encontrar reglas que permitan relacionar el estado del clima y el pavimento con el tipo de accidente. El segundo pretende encontrar reglas que permitan relacionar el tipo de vehículo con el tipo de accidente. Y por último se busca encontrar reglas que permitan relacionar la intensidad y tipo de tráfico con el tipo de accidente.

Elemento de trabajo y metodología

Para llevar a cabo este proyecto se utiliza la metodología CRISP-DM[6]. Ésta propone llevar a cabo proyectos de Explotación de Información mediante la realización de cinco fases, las cuales contienen actividades y entregables definidos.

A continuación se detalla una breve descripción de las actividades realizadas en cada una de las fases así como los entregables generados.

Fase I: Comprensión del Negocio

Las tareas realizadas para esta fase incluye la determinación de objetivos de negocio, valoración de la situación, determinación de los objetivos de explotación de información y la realización del plan de proyecto. Se determina el siguiente objetivo de negocio: *mejorar la seguridad vial mediante la implementación de políticas de seguridad no arbitrarias, basada en información histórica objetiva.*

Para lograr concluir si el corriente proyecto de explotación de información tuvo éxito o no, se debe definir cuál es criterio de éxito del negocio, el cual se considera satisfecho si se logra obtener patrones que permitan describir las condiciones que conllevan a accidentes graves donde personas salen heridas, para luego ser tenida en cuenta al momento de implementar políticas y medidas de seguridad.

En lo que respecta a la valoración de la situación, dentro de esta fase se determina la viabilidad del proyecto, los recursos físicos y lógicos utilizados, las fuentes de información y de conocimiento, limitaciones de los datos y un análisis de riesgos. Se determina que el proyecto es viable a partir del Análisis de Viabilidad[7] realizado a partir de los datos obtenidos en el relevamiento inicial.

Como recursos físicos y lógicos, se utilizaron dos notebooks para desarrollo y pruebas, además de las herramientas de software: Rapid Miner Studio, Ofimática de MS Office, Ofimática de Google Docs y Project in a Box.

La fuente de información utilizada es la base de datos que mantiene AUSA, las cuales son de libre acceso y públicas, que contienen datos sobre los accidentes e información del paso vehicular por los puntos de peaje de las rutas concesionadas por esta entidad.

La fuente de conocimiento fueron principalmente los artículos científicos provistos por la cátedra y como material bibliográfico se utilizaron manuales de ayuda de las herramientas de software utilizada.

Para dar inicio a este proyecto se da por supuesto que la cantidad de datos disponibles son suficientes para poder obtener reglas que definan el patrón de accidentes de tránsito.

Al igual que se definieron las suposiciones, también se identificaron las limitaciones de los datos disponibles y riesgos del proyecto.

Limitaciones:

- Solo se cuenta con información de las autopistas concesionadas por AUSA.
- Solo el 20% de las muestras de accidente detallan el tipo de vehículo.

Riesgo del proyecto:

- Los datos son insuficientes para obtener el modelo esperado.
- Como plan de contingencia a los riesgos presentados, se planteó

buscar fuentes alternativas de información, o bien comunicarse con AUSA solicitando fuentes de información no pública.

Una vez identificados todos los parámetros del proyecto, se identificaron los mecanismos necesarios para cada uno de los distintos objetivos de requisito:

1. Encontrar reglas que permitan relacionar el estado del clima y el pavimento con el tipo de accidente.
2. Encontrar reglas que permitan relacionar el tipo de vehículo con el tipo de accidente.
3. Encontrar reglas que permitan relacionar la intensidad y tipo de tráfico con el tipo de accidente.

Fase 2: Comprensión de los datos

Dentro de esta fase se realizan las tareas de recolección inicial de los datos, su descripción, exploración y verificación de la calidad de los mismos.

Para cumplir con el objetivo del proyecto de minería de datos, se contó con dos repositorios de datos.

Uno de los archivos contiene información de los accidentes de tránsito transcurridos en las autopistas urbanas de la Ciudad Autónoma de Buenos Aires, llamado de ahora en más RepAccidentes.

El otro contiene información de la circulación de vehículos por los peajes de dichas autopistas, llamado de ahora en más RepTrafico.

En lo que a la calidad de datos respecta, estos provienen de casos reales recolectados desde el año 2004 hasta el presente, lo que debería proveernos de un amplio espectro de casos que no están limitados a cierta época del año o a un año en especial, esto brinda representatividad a los datos y disminuye la arbitrariedad de los mismos.

Fase 3: Preparación de los datos

Durante esta fase se utiliza el conjunto de datos pre-formateados obtenido de la fase anterior, para realizar las tareas de selección, limpieza, construcción, integración y formateo.

Selección de los datos

Para el estudio de la relación del estado del clima, el pavimento y el tipo de vehículo con respecto a los accidentes (Objetivo de requisito 1 y 2) se seleccionaron los atributos del repositorio RepAccidentes que se muestran en la Tabla 1.

Para el estudio de la relación entre el flujo vehicular y los accidentes (Objetivo de requisito 3) se seleccionaron de los repositorios RepAccidentes y RepTrafico los atributos que se muestran en la Tabla 2 y Tabla 3 respectivamente

Tabla 1: Atributos seleccionados del repositorio RepAccidentes para los objetivos de requisito 1 y 2

Fecha: la fecha en la que ocurrió el accidente.	Tipovehiculo: el tipo de vehículo involucrado	Estadopavimento: estado del pavimento al momento del accidente.	Heridos: cantidad de heridos por el accidente.
Causales: motivo del accidente.	Clima: estado del tiempo al momento del accidente.	Fallecidos: cantidad de fallecidos por el accidente.	Autopista: autopista en la que ocurrió el accidente.

Tabla 2: Atributos seleccionados del repositorio RepAccidentes para el objetivo de requisito 3

Fecha: la fecha en la que ocurrió el accidente.	Clima: estado del tiempo al momento del accidente.	Heridos: cantidad de heridos por el accidente.
Causales: motivo del accidente.	Estadopavimento: estado del pavimento al momento del accidente.	Autopista: autopista en la que ocurrió el accidente.
Tipovehiculo: el tipo de vehículo involucrado	Fallecidos: cantidad de fallecidos por el accidente.	Localizacion: lugar, respecto a la autopista, en el que el accidente se produjo.

Tabla 3: Atributos seleccionados del repositorio RepTrafico para el objetivo de requisito 3

Fecha: fecha en la que el vehículo circulo.	Día: día de la semana en el que el vehículo circulo.	Estación: estación de peaje por la cual el vehículo circulo.	Tipovehiculo: tipo de vehículo.

Limpieza de los datos

Del repositorio RepAccidentes se consideraron solo los ejemplos que tienen valores no nulos, los cuales suman un total de 7422 entradas.

Del repositorio RepTrafico se consideraron solo los ejemplos que tienen valores no nulos en el atributo tipovehiculo.

Atributos derivados

Para realizar un mejor estudio consideramos la creación atributos derivados que se muestran en la Tabla 4.

Tabla 4: Atributos derivados

Momento: representa el momento del día en que ocurrió el accidente, es inferido a partir de la hora en que ocurrió, agrupándolo en 3 categorías, "mañana", "tarde" y "noche".	TotalLiviano: es el total de vehículos livianos que circularon por las autopistas administradas por AUSA para un día dado.	TotalPesado: es el total de vehículos pesados que circularon por las autopistas administradas por AUSA para un día dado.	HeridosFallecidos: atributo booleano que indica si al menos hubo un herido o un fallecido en el accidente.

Formateo de los datos

Analizando el repositorio de accidentes notamos que en el atributo "tipovehiculo" se pueden agrupar los valores "AUTOMOVIL" y "MOVIL" con el valor "AUTO PARTICULAR" ya que representan lo mismo. También para este mismo atributo decidimos unir el valor "PICK UP" con el valor "CAMIONETA" y el valor "AMCO VEBA V8164" con el valor "GRUA".

Integración de los datos

Para el estudio de la relación de las condiciones de carretera, clima y tipo de vehículo (Objetivos de requisitos 1 y 2) se utilizará solo el repositorio RepAccidentes, donde los datos ya se encuentran asociados correctamente.

Los elementos del repositorio RepTrafico fueron agrupados por el campo fecha utilizando la sentencia Group By de SQL. También se utilizó la función de agregación Sum para obtener los campos TotalLIviano y TotalPesado. Estos se integraron con los elementos del repositorio RepAccidentes a través de la fecha.

Fase 4: Modelado

En esta fase se llevaron a cabo tareas tales como elección de técnicas a utilizar y construcción de los modelos.

Para el estudio se utilizó una implementación propietaria de RapidMinner que incluye subprocesos de pruning (podado) y prepruning (pre-podado).

Construcción del primer modelo

Como se observa en la Figura 1, se comienza con los datos tal cuales quedaron luego de finalizada la fase de preparación de los datos. Luego ésta es ingresada al operador que se encargará de configurar el atributo "heridosfallecidos" como objetivo. Esto es luego ingresado al operador que creará el árbol de decisión. Luego el modelo es aplicado sobre el mismo repositorio para poder evaluar su rendimiento

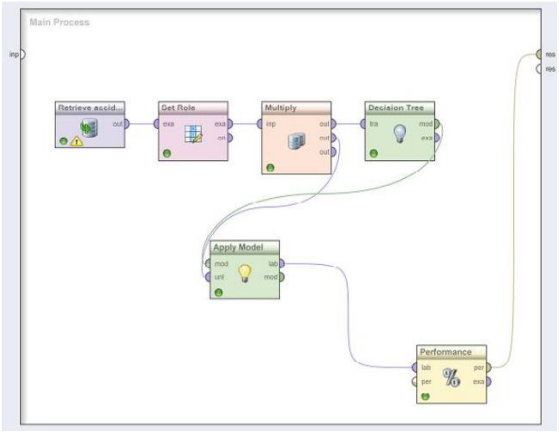
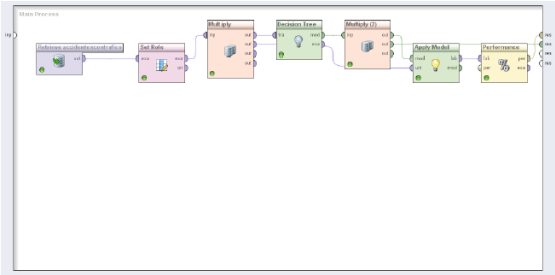


Figura 1: Primer modelo

Construcción del segundo modelo

Hemos decidido que no es necesario presentar resultados de manera detallada ya que estos son muy similares a los del estudio anterior. Luego de variar los parámetros de los operadores no conseguimos que estos utilicen las columnas "TotalLiviano" y "TotalPesado" que contienen información sobre el flujo vehicular. Concluimos entonces que, al menos en la muestra utilizada, el flujo vehicular no es un factor determinante para los resultados de un accidente de tránsito.



Análisis de los resultados

La primera instancia del estudio según los criterios de éxito fracaso. Sin embargo consideramos luego de haber revisado los repositorios detalladamente, que dichos criterios de éxito fueron planteados de manera sobre-exigente. Si bien el modelo obtenido no es concluyente, no todo es desechable. Por ejemplo hemos confirmado empíricamente que el vehículo más peligroso es la motocicleta. También se puede ver que cuando el causal del accidente es un vuelco las chances de que haya personas heridas y/o víctimas fatales son altas. Esto se podría utilizar como evidencia para justificar un procedimiento de respuesta inmediata en caso de ocurrir un accidente con motocicletas involucradas o en caso de ser un vuelco.

También hemos encontrado que la autopista Dist. 9 de Julio es particularmente peligrosa, dado que clasifica positivo en el modelo, a pesar de que en la ruta de decisión los demás atributos se encuentran en valores que se consideran generalmente favorables para la conducción.

Tabla 5: Riesgo de viajar en moto

Certeza	Condiciones	Resultado
93 %	Sobrepaso (Alcance)	Sufrirá lesiones
93 %	Choque contra defensa	Sufrirá lesiones
76 %	Cualquier causa de choque	Sufrirá lesiones

Tabla 6: Otros casos de riesgo

Certeza	Condición	Resultado
80 %	Autobús colisionó contra defensa	Pasajeros lesionados
60 %	Vuelco	Pasajeros lesionados

Tabla 7: Casos controversiales

Certeza	Condiciones	Resultados
53 %	Choque con vehículo estacionado	Pasajeros lesionados
59 %	Furgón chico en sobrepaso	Pasajeros lesionados
54 %	Taxi colisiona contra defensa	Pasajeros lesionados
100 %	Grúa en cualquier colisión	Pasajeros lesionados
100 %	Coche particular colisiona contra defensa con mal clima	Pasajeros lesionados

Conclusiones

Para concluir se resume los casos de riesgo detectados en las Tablas 5, 6 y 7.

Creemos que se podría mejorar los resultados utilizando como dato el flujo vehicular por ruta en vez de general. Para esto se necesitaría información precisa de la ubicación de los puestos de peaje.

Un factor negativo importante fue el hecho de no contar con el apoyo por parte del AUSA. Sería muy bueno realizar una nueva instancia de explotación de la información contando con el soporte de un experto que brinde al equipo información detallada y actualizada, así como también sirva para despejar dudas que surgieron a lo largo de este proyecto.

Agradecimientos

Agradecemos al M. Ing. Pablo Pytel por sus clases dictadas en la catedra, por su buena disposición al momento de realizarle consultas y por brindarnos las herramientas necesarias para poder llevar a cabo el proyecto de manera exitosa.

Referencias

[1] Schiefer, J., Jeng, J., Kapoor, S. & Chowdhary, P.. Process Information Factory: A Data Management Approach for Enhancing Business Process Intelligence. Proceedings 2004 IEEE International Conference on ECommerce Technology. Pág. 162-169. 2004

[2] Thomsen, E.. BI's Promised Land. Intelligent Enterprise, 6(4): 21-25. 2003.

[3] García-Martínez, R., Britos, P., Pesado, P., Bertone, R., PolloCattaneo, F., Rodríguez, D., Pytel, P., Vanrell. J. Towards an Information Mining Engineering. En Software Engineering, Methods, Modeling and Teaching. Sello Editorial Universidad de Medellín. 2011.

[4] Pytel, P.; Pollo-Cattaneo F.; Rodriguez, D.; Britos, P.; García-Martínez, R. Identificación de Tareas Críticas en una Metodología de Desarrollo de Proyectos de Explotación. Proceedings XVII Congreso argentino de Ciencias de la Computación. 2011.

[5] Britos, P., García-Martínez, R. Propuesta de Procesos de Explotación de Información. Proceedings XV Congreso Argentino de Ciencias de la Computación. Workshop de Base de Datos y Minería de Datos. Págs. 1041-1050. 2009.

[6] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-bystep Data Mining Guide. http://tinyurl.com/crispdm Último acceso Junio 2013.

[6] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000). CRISP-DM 1.0 Step-bystep Data Mining Guide. http://tinyurl.com/crispdm Último acceso Junio 2013.

Datos de Contacto

Fontana Nicolás. Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay. Erausquin 275, Concepción del Uruguay (3260), Entre Ríos. E-mail: nicolasfontanaparis@gmail.com

Puñales Donato, Joaquín Ignacio. Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay. Cochabamba 580, Concepción del Uruguay (3260), Entre Ríos. E-mail: joaquinpunales@gmail.com

Malgor, Rafael. Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay. Mariano López 324, Concepción del Uruguay (3260), Entre Ríos. E-mail: rafaelmalgor@gmail.com