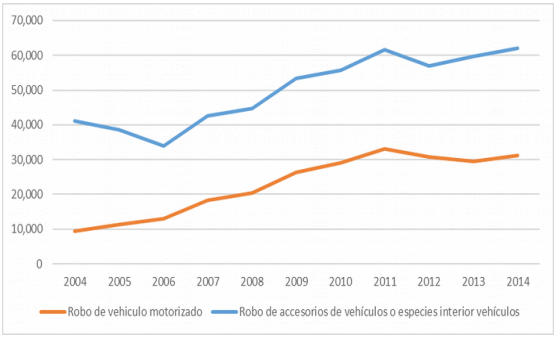
Diego Garrido

Modelamiento y seguimiento de tópicos para detección de *modus operandi* en robos de autos

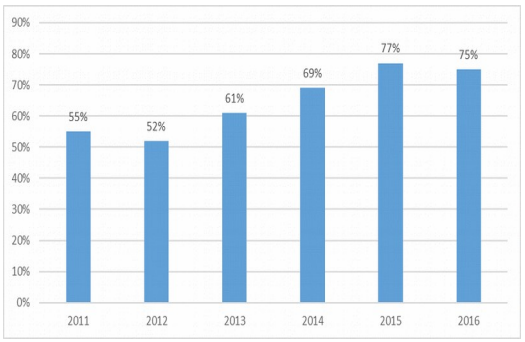
Antecedentes

El robo de vehículos o accesorios de vehículos es un problema que afecta a toda la sociedad en Chile y en el mundo. Este problema se ha vuelto más relevante el último tiempo debido al crecimiento en el robo de vehículo motorizado y de los robos con violencia. Este fenómeno trae consigo un montón de costos para la sociedad, como incremento en al percepción de la seguridad, aumentos en la prima de los seguros de los asegurados, aumento en los costos de las aseguradoras[1] y el incremento de otros tipos de delitos (un cantidad importante de delitos se perpetran con vehículos con encargo por robo).



Fuente: Informe anual Carabineros, 2004-2014. INE.

Cantidad de robos de vehículos y robos de accesorios de vehículos anuales en Chile (2004-2014)



Tasa de robos con violencia del total de robo de autos de lujo (2011- 2016)

Bajo este contexto la Universidad de Chile junto a la Pontificia Universidad Católica de Chile se adjudicó un Fondef el 2017 para desarrollar un proyecto que lleva por nombre “Observatorio Digital de Delincuencia en Chile: Un sistema inteligente de apoyo a la industria automotriz chilena, en el robo de vehículos y accesorios” cuyo director es Richard Weber Haas y la institución beneficiaria es la Asociación de Aseguradores de Chile (AACH).

Para este problema se cuenta con las fuentes de datos de la AACH, lo que corresponde a relatos de las víctimas del robo de sus vehículos desde el 2011 hasta el 2016, lo cual corresponde a 49.015 relatos. Cabe destacar que se estima que un tercio del parque automotriz se encuentra asegurado, por lo que se trabaja con una muestra del parque automotriz.

Objetivo, resultados esperados y alcances

El objetivo del del trabajo de tesis es caracterizar los *modus operandi* de los delincuentes a partir de los relatos de víctimas de robos de vehículos entregados por la AACH.

El resultado esperado es descubrir los *modus operandi* ocultos en los relatos de las víctimas y caracterizarlos a partir de las palabras, como también ver su evolución a través del tiempo, siendo capaz de detectar cuando nacen y mueren, y como cambian en el tiempo.

---

1 Considerando que el costo promedio incurrido en un auto asegurado robado y no recuperado es de $ 5.000.000 de pesos, la pérdida total considerando solo los vehículos no recuperados para el año 2015 es de unos $15.720 millones de pesos.

El trabajo de tesis tiene un propósito más académico puesto que no cuenta con un cliente en específico y tiene por objetivo estudiar técnicas de *clustering* dinámico para detectar patrones en el contexto de robo de vehículos, sin embargo, potenciales beneficiarios del trabajo podrían ser las aseguradoras, los asegurados, carabineros de Chile y la sociedad.

Metodología y marco teórico

El proyecto de tesis se realizará bajo la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) [1], la cual posee la siguientes seis etapas: comprensión del negocio, entendimiento de los datos, preparación de los datos, modelamiento, evaluación e implementación. Dentro de los alcances del trabajo de tesis no está contemplado la puesta en producción de una solución basada en *machine learning,* el alcance es hasta la evaluación e interpretación de los resultados que arroja el modelo.

Este problema que involucra datos se trata de un problema de aprendizaje no supervisado, puesto que no se cuenta con una etiqueta del *modus operandi* a la que pertenece cada relato, esta etiqueta debe descubrirse a partir de los datos. Este problema corresponde a un problema de *clustering dinámico* usando como *input* los relatos y la fecha en que ocurrieron, donde se desea agrupar los relatos en el tiempo en *clusters*, donde cada *cluster* sería un *modus operandi* en particular. Dentro de los métodos de *clustering* que involucran texto el modelamiento de tópicos es el enfoque más prometedor. El modelamiento de tópicos es una herramienta estadística que busca encontrar los temas (tópicos) presentes en un conjunto de documentos (corpus), permitiendo organizar, buscar, indexar, explorar y comprender grandes colecciones de documentos. En *text mining* se asume que los documentos pueden ser representados como una mezcla de tópicos, donde los tópicos son distribuciones sobre las palabras, los tópicos son latentes y la inferencia en modelamiento de tópicos tiene por objetivos descubrir la mezcla de tópicos que originó cada documento y la distribución sobre las palabras de cada tópico. En modelamiento de tópicos las personas son las que le dan un interpretación a los tópicos a partir de las palabras más relevantes y le colocan un etiqueta, por ejemplo, para un tópico, dentro de sus cinco palabras más probables se halla la siguiente secuencia: “llaves”, “domicilio”, “individuos”, “casa”, “portón”, una etiqueta valida para este tópico podría ser “portonazo”.

Algunas de las técnicas de modelamiento de tópicos están basadas en factorización matricial como LSI (Latent Semantic Indexing) [2] o NMF (Non-negative Matrix Factorization) [3], pero en este trabajo se utilizarán técnicas basadas en modelos probabilísticos generativos, como LDA (Latent Dirichlet Allocation) [4] o HDP (Hierarchical Dirichlet Process) [5].

Dentro del trabajo de tesis se busca capturar el dinamismo que puede presentar el fenómeno de robo de vehículos. El aspecto dinámico del problema considera:

1. Nacimiento, muerte, fusión y división de tópicos: En el contexto de robos es natural que en el tiempo aparezcan nuevos *modus operandi* como también que desaparezcan aquellos que ya no parecen tan atractivos.
2. Dinamismo en la mezcla de tópicos: Esto permite modelar tendencia y estacionalidad de los tópicos.
3. Dinamismo en la distribución sobre las palabras que posee un tópico: Esto permite detectar cambios en cómo se comete un mismo tipo de delito, por ejemplo, el “portonazo” en un determinado momento se comete en grupos de 2-3 personas con arma blanca y luego evoluciono de arma blanca a arma de fuego.

Dentro de los modelos de tópicos probabilísticos existen modelos estáticos y dinámicos:

1. Los modelos estáticos con mejores resultados son LDA y HDP. La diferencia en estos dos modelos es que el primero necesita de antemano fijar el número de tópicos a descubrir y el segundo lo infiere a partir del corpus.
2. Dentro de los modelos dinámicos se tiene aquellos que mantienen el número de tópicos fijos durante el tiempo y los que no:

1. Dentro del primer grupo destacan Dynamic Topic Modelling (DTM) [6] y Topic Over Time (TOC) [7], el problema de esto es que si aparece un nuevo tópico en el tiempo este quedará clasificado dentro de un tópico que existía desde el comienzo.
2. Dentro de los modelos que no mantienen el número fijo de tópicos en el tiempo existen de dos tipos, aquellos que modelan todo el problema bajo un modelo único, en este grupo destaca Dynamic Hierarchical Dirichlet Process (DHDP) [8], el cual modela el problema de dinamismo de una forma elegante pero a la vez acompañada de una inferencia bastante complicada, además no es una tecnología madura, ya que no cuenta con una implementación disponible a diferencia de los otros modelos mencionados, los cuales se encuentran disponibles en múltiples lenguajes de programación y cuentan con una amplia adopción de la comunidad científica. El segundo tipo de modelos que no mantienen fijo el número de tópicos son modelos que utilizan modelos de tópicos estáticos de forma iterativa, lo que hacen es dividir el tiempo en intervalos de tiempo discretos, luego entrenan de forma independiente un modelo de tópico para cada período y luego unen los resultados obtenidos, un ejemplo utilizando LDA en [9] y con HDP en [10].

En este trabajo se utilizarán técnicas de modelado dinámico de tópicos como las presentadas en [9-10], debido a que son capaces de modelar los tres puntos mencionados sobre dinamismo y se basan en tecnologías maduras.

Referencias

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, 16.

[2] Dumais, S. T. (2004). Latent semantic analysis. Annual review of information science and technology, 38(1), 188-230.

[3] Xu, W., Liu, X., & Gong, Y. (2003, July). Document clustering based on non-negative matrix factorization. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (pp. 267-273). ACM.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet allocation,” J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003

[5] Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. (2005). Sharing clusters among related groups: Hierarchical Dirichlet processes. In Advances in neural information processing systems (pp. 1385-1392)

[6] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in Proceedings of the 23rd International Conference on Machine Learning, ser. ICML ’06. New York, NY, USA: ACM, 2006, pp. 113–120

[7] Wang, X., & McCallum, A. (2006, August). Topics over time: a non-Markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 424-433). ACM.

[8] Ahmed, A., & Xing, E. P. (2012). Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. arXiv preprint arXiv:1203.3463.

[9] Wilson, A. T., & Robinson, D. G. (2011). Tracking Topic Birth and Death in LDA. Sandia National Laboratories.

[10] Beykikhoshk, A., Arandjelović, O., Venkatesh, S., & Phung, D. (2015, May). Hierarchical Dirichlet process for tracking complex topical structure evolution and its application to autism research literature. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 550-562). Springer, Cham.

Firma profesor guía

Richard Weber

Firma alumno

Diego Garrido