Monitoreo urbano de entidades y eventos geográficos basado en censado social

Juan Carlos Salazar Carrillo[*]
Miguel Jesús Torres Ruiz[**]
Marco Antonio Moreno Ibarra[***]

Recibido el 16 de enero de 2017; aceptado el 13 de junio de 2017

Resumen

La proliferación de las redes sociales y la facilidad que otorgan a sus usuarios para compartir información alienta a compartir lo que ocurre a su alrededor. Las redes sociales ayudan a conocer diferentes acontecimientos, y al estar conscientes de ellos, nos ayuda a tomar decisiones con mayor certeza. Por ejemplo, acerca del tránsito vehicular, el utilizar las redes sociales nos ayuda a saber cuándo están presentes manifestaciones o cuando ha ocurrido un accidente, esta información nos ayuda a eludir un congestionamiento vial desafortunado. Aprovechando la generación de contenido en Twitter y tomando como caso de estudio la Ciudad de México, se recolectó información de usuarios dedicados a publicar eventos viales. Por tanto, se propone una metodología para la *geocodificación* de textos cortos y un método de aprendizaje automático basado en Máquinas de Soporte Vectorial, con el cual se obtiene un modelo capaz de realizar un análisis espacio temporal de eventos viales. Como caso de estudio se consideró a la Ciudad de México.

Palabras clave: *geocodificación, redes sociales, tránsito vehicular, máquinas de soporte vectorial, big data.*

[*] Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO), Liga Periférico-Insurgentes Sur núm. 4903, Col. Parques del Pedregal, 14010 Ciudad de México, México, correo electrónico: jsalazar@conabio.gob.mx.

[**] Centro de Investigación en Computación, Instituto Politécnico Nacional (IPN), Av. Juan de Dios Bátiz esq. Miguel Othón de Mendizábal s/n, Nueva Industrial Vallejo, 07738 Ciudad de México, México, correo electrónico: mtorres@cic.ipn.mx.

[***] Centro de Investigación en Computación, Instituto Politécnico Nacional (IPN), correo electrónico: marcomoreno@cic.ipn.mx.

Resumo

A proliferação das redes sociais e sua facilidade de utilização encorajam os usuários a compartilhar o que ocorre ao seu redor. Elas ajudam a conhecer diferentes fatos e, ao estar cientes deles, as decisões são tomadas com maior certeza. Como exemplo, pode-se citar trânsito de veículos, onde ao se utilizar as redes sociais pode-se saber onde estão ocorrendo manifestações ou acidentes. Esta informação ajuda a evitar entrar num congestionamento indesejado. Aproveitando a geração de conteúdo no Twitter e tomando como caso de estudo a Cidade do México, se coletou informação de usuários dedicados a publicar postagens sobre vias. Para isso, se propõe uma metodologia para a geocodificação de textos curtos e um método de aprendizagem automático baseado em Máquinas de Suporte Vetorial, com a qual se obtém um modelo capaz de realizar uma análise espaço temporal de eventos nas vias.

Palavras chave: *geocodificação, redes sociais, trânsito veicular, máquinas de Suporte Vetorial, Big Data.*

Abstract

Current social networks provide information with high correlation with events that are occurring worldwide. Twitter is a microblogging network of real time messages in which people post about various classes of events. A relevant topic is traffic congestion; user-generated content is useful to assist drivers in avoiding crowded areas. This work proposes a model to predict traffic-related events, based on a set of machine learning methods, in which a spatio-temporal dataset is obtained from Twitter messages. The training stage uses geocoded traffic events, in order to generate possible sites with traffic congestion at a given time. As a case study, partial areas of the Mexico City were taking into consideration.

Key words: *Geocoding, Social networks, traffic jams, Support Vector Machines, Big Data.*

Introducción

Más de la mitad de la población del mundo vive en áreas urbanizadas, es inevitable que poblaciones de gran magnitud lleguen a ser lugares muy complicados y desordenados (Chourabi, H. *et al.,* 2012). Ciudades y megalópolis generan nuevos tipos de problemas, tales como la administración de la basura, la carestía de recursos, la contaminación, problemas de salud en la población, el desplazamiento de sus habitantes, los congestionamientos de tráfico, y el deterioro de las infraestructuras, son los problemas básicos. Para resolver algunos de los aspectos ligados a estos problemas, la detección de accidentes y congestionamientos viales de forma automática puede ser útil. Los problemas de movilidad urbana, pueden ser reducidos si los accidentes y condiciones viales son conocidos por todos, saber en qué áreas se encuentran el tránsito pesado, los accidentes viales, los semáforos descompuestos, y los cortes de circu-

lación. Sin embargo, utilizando dispositivos GPS es muy complicado recolectar información relacionada con eventos viales. En este contexto, redes sociales como Twitter son muy útiles, ya que se puede aprovechar la gran cantidad de eventos que ahí se reportan (Lee *et al.,* 2013). Información relacionada a eventos viales también es muy común en Twitter, cuando las personas están moviéndose alrededor de la ciudad, van publicando información acerca de las condiciones viales utilizando sus dispositivos móviles, de hecho, existen varios usuarios que se dedican exclusivamente a publicar información acerca de este tema. Algunos de estos usuarios son agencias del gobierno, usuarios oficiales de transporte público, estaciones de radio y televisión, además de ciudadanos. Es importante señalar, que la información en Twitter raramente tiene las coordenadas, esto se debe a que los usuarios por cuestiones de seguridad o uso personal, desactivan la "ubicación" en sus dispositivos móviles, lo que origina que el sensor GPS no recupere las coordenadas específicas de cada dispositivo. Por tanto, se requiere geocodificar en la mayoría de los casos los tweets, con el objeto de identificar las ubicaciones donde ocurren los eventos, lo cual es un reto importante. Los métodos de aprendizaje automático como Máquinas con Soporte Vectorial (SVM) han sido de gran utilidad en diferentes áreas, ejemplos como reconocimiento de rostros, minería de datos, análisis de la bolsa de valores, predicciones de tiempo y otras más (Guo *et al.,* 2000), (Huang *et al.,* 2005) y (Akay, 2009).

En este artículo, se propone un enfoque para geocodificar eventos viales publicados en Twitter, incluyendo características espacio-temporales como la hora del día. A partir de esta información se crea un conjunto de prueba para utilizar un método de aprendizaje automático llamado SVM para Regresión. Un modelo de predicción es generado con el fin de mostrar posibles congestionamientos viales en horas específicas del día. Resultados preliminares muestran un precision y un recall del 74% y 70% respectivamente. Los resultados están relacionados a la cantidad de eventos recolectados para generar el conjunto de entrenamiento. En la segunda parte se describe el estado del arte; luego, la metodología, a continuación la creación del conjunto de prueba, y del modelo de predicción, después la evaluación del método y finalmente las conclusiones y trabajo futuro.

Estado del arte

El observatorio de tráfico (Ribeiro Jr *et al.,* 2012) propone la geocodificación de textos cortos en Twitter usando el *gazetteer* GEODICT. Este contiene una colección de segmentos de avenidas y calles, cruces entre ellas, abreviaturas, seudónimos, vecindarios y puntos de referencia. Estos elementos tienen una representación geográfica relacionada. El observatorio de tráfico usa funciones de comparación de cadenas exactas o aproximadas para geolocalizar las calles mencionadas en los tweets. Delboni *et al.* (2007) proponen un método de recolección de información en la web utilizando técnicas de procesamiento de lenguaje natural, de este modo, puede reconocer y posicionar expresiones formadas por puntos de referencia y las

relaciones que existen entre estas. Davis Jr *et al.* (2011) proponen una metodología basada en las relaciones entre usuarios en Twitter para inferir la localización de los tweets. Una red es creada tomando en cuenta las relaciones *siguiendo* y *seguidores.*

Trabajando con datos de Facebook, Backstrom *et al.* (2010) muestra la fuerza de conexión entre relaciones en Facebook y la localización geográfica de los usuarios. Estos interactúan frecuentemente, viven cerca uno de otro, y cada usuario tiene al menos 10 amigos con puntos geográficos compartidos. Con estas suposiciones, esta metodología infiere la probable localización de cada usuario. En el contexto geográfico, las Máquinas de Soporte Vectorial (SVM) han sido utilizadas; siendo una de las principales técnicas de aprendizaje automático. Wang *et al.* (2013), propuso utilizar un modelo de inferencia basado en SVM, junto con un algoritmo de agrupamiento para inferir la localización de nuevas imágenes. En Wu *et al.* (2004) se muestran las predicciones de tiempo de traslado de un punto a otro. Fueron considerados diferentes factores que afectan el movimiento vehicular dentro de una ciudad, por ejemplo, velocidad del vehículo, clima, nivel de tránsito, accidentes, hora del día y día de la semana. La información se recolectó durante cinco semanas en tres rutas distintas y diferentes horas del día. Por tanto, se demostró que las SVM realizan predicciones acertadas, superiores a las predicciones basadas en históricos o métodos de predicción basados en información en tiempo real o bajo un enfoque estadístico.

Geocodificación de eventos viales en Twitter

La metodología propuesta involucra procesos automáticos y semi-automáticos con el propósito de ubicar objetos geográficos especificados en tweets; como: calles o sitios. Para esto se utilizó un Gazetteer y un corpus de Tweets. El Gazetteer inicial fue obtenido de GeoNames y el conjunto de datos utilizado está compuesto por 36,236 calles de la Ciudad de México. En algunos casos, podría ser útil una fotografía para llevar el proceso de georreferenciación, siempre y cuando se tuvieran las coordenadas de esta fotografía; entonces se pudiera pensar en agregar estos datos como parámetros de procesamiento al proceso de geocodificación. La Tabla 1 muestra las principales cuentas utilizadas para generar el Corpus.

Tabla 1
Cuentas principales de Twitter para conformar el Corpus

Cuenta Twitter	*Fecha de creación*	*Seguidores*	*Número de tweets*	*Sitio*
SSPDFVIAL	07.14.2010	369,115	154.65	Si
PolloVial	01.31.2013	667	71.91	No
Trafico889	05.14.2009	137,099	90.54	Si
Alertux	10.16.2012	179,574	35.59	Si
072AvialCDMX	10.20.2010	83,535	134.71	Si
RedVial	03.09.2010	63,702	44.81	Si

El proceso de geocodificación consta de las siguientes fases: 1) información del Corpus, 2) diccionarios y ejes equivalentes, 3) división del Gazetteer, 4) estandarización, 5) identificación y localización de eventos viales. Las fases se describen a continuación.

Extracción de información del conjunto de datos

Los tweets se encuentran almacenados en un repositorio denominado conjunto de datos o *corpus* e incluyen las calles, los seudónimos frecuentes, las abreviaturas comunes, lugares populares y monumentos históricos. Con el fin de recolectar esta información se desarrolló un script para obtener los n-gramas[1] más comunes. De cada tweet, fueron obtenidos los n-gramas y se ordenaron frecuencia. Aunque los n-gramas incluyen la combinación de cualquier segmento de palabra, los considerados para este trabajo, son segmentos de frases continuos. La Figura 1 muestra el funcionamiento del script. De la lista de unigramas, bigramas y trigramas se identificaron de forma semi-automática 456 calles principales, 150 tipos de eventos viales, 135 hashtags, 69 seudónimos, 65 edificios, lugares y monumentos, 34 abreviaturas y 26 combinaciones de preposiciones.

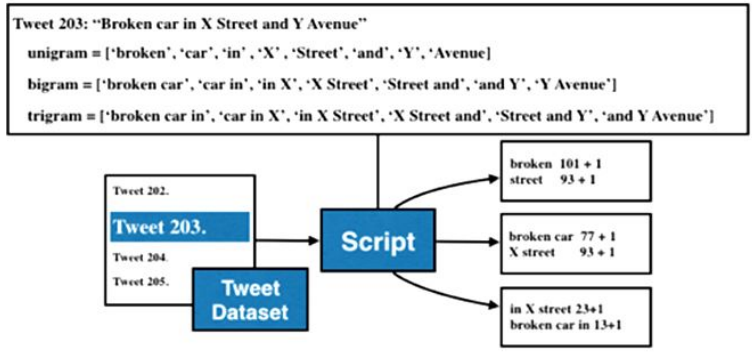

Figura 1. Procesamiento para la obtención de n-gramas comunes.

Creación de diccionarios geográficos y nombre de ejes equivalentes

Del resultado obtenido en el proceso anterior, datos de OpenStreetMap y del Instituto Nacional de Estadística y Geografía (INEGI) fueron utilizados para generar diccionarios que enriquezcan el Gazetteer. Los diccionarios obtenidos son de abreviaturas, de seudónimos, de hashtags, de eventos viales, de estaciones de transporte público, de calles principales (calles que aparecen en los Tweets y en el Gazetteer),

[1] Un n-grama es un segmento de palabra o segmento de frase que pertenece a una más grande (Cavnar *et al.*, 1994).

de lugares, edificios y monumentos y de colonias (Figura 2). Estos diccionarios tienen una componente geográfica relacionada, utilizados para localizar espacialmente eventos y objetos. Es frecuente que las calles tengan más de un nombre, por lo que se usó un Diccionario de ejes equivalentes.

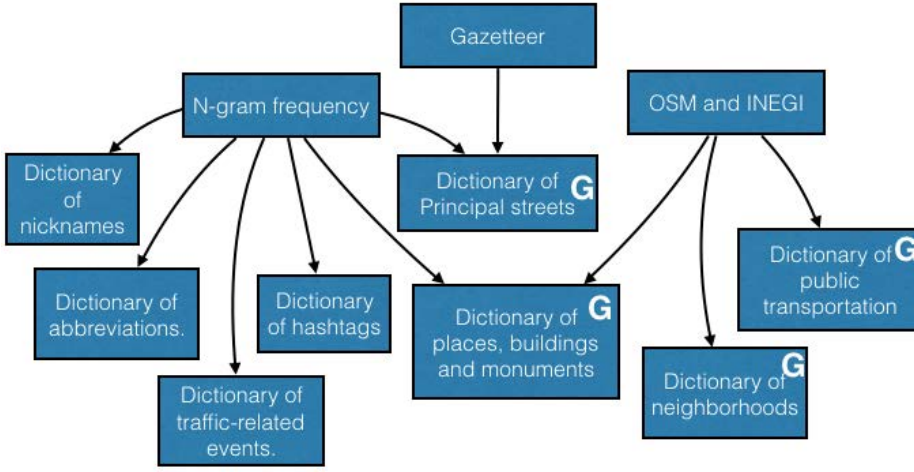

Figura 2. Diccionarios generados.

División del Gazetteer

Existe un grupo reducido de calles donde son concentrados la mayoría de los eventos viales, usando la frecuencia de los n-gramas, se verificó que solo el 19% de las calles del Gazetteer aparece en los tweets. En consecuencia, el Gazetteer es divido en dos partes, la primera, formada por las calles más frecuentes mencionadas en Twitter y la segunda, está compuesta por las calles restantes. Aunque esto no mejora las medidas de validación utilizadas, el rendimiento de la implementación de la metodología incrementa considerablemente en la identificación y localización.

Estandarización de la información

Para mejorar el proceso de estandarización, los Diccionarios de elementos no geográficos son utilizados (de abreviaturas, de seudónimos, de hashtags). Dentro del Gazetteer se encuentra que las calles incluían abreviaturas, calles en mayúsculas, contienen acentos y elementos sin nombre o con valores por defecto. En esta etapa el nombre de las calles es transformado a minúsculas y se eliminaron los acentos,

además, utilizando el Diccionario de abreviaturas, fueron reemplazadas por su nombre completo. Finalmente, las calles sin nombre fueron eliminadas. Otros problemas detectados dentro del Corpus fueron enlaces a otras cuentas, seudónimos, palabras mal escritas, faltas de ortografía y hashtags. Para solventar estos errores se utilizaron los Diccionarios de seudónimos y hashtags. Enlaces y mención a otras cuentas fueron suprimidos de cada tweet, faltas de ortografía no son resueltas en esta metodología. En el Gazetteer como en la colección de tweets se filtraron stop words.[2] No existe una lista universal de stop words, por lo que se utilizó la biblioteca Natural Language Toolkit (Bird, 2006).

Identificación y localización de eventos viales

La identificación de objetos geográficos se llevó a cabo utilizando los Diccionarios de objetos geográficos: de calles principales, de calles no comunes, de colonias, de estaciones de transporte público y de lugares, edificios y monumentos. Los eventos viales reportados por las cuentas de Twitter seleccionadas comúnmente hablan de accidentes, malas y buenas condiciones. Por ejemplo, un evento vial es un accidente cuando se menciona dentro de tweets como **choque, volcadura**, un evento vial considerado como mala condición es descrito como **lento desplazamiento o vuelta de rueda**. Una buena condición es mencionada como **buen desplazamiento o sigue avanzando, sin problemas**.

Con base en la colección de tweets obtenida y el análisis de frecuencia de n-gramas, un accidente es considerado como un evento que ocurre en un punto específico que tiene relación otros objetos geográficos, por ejemplo, **choque en calle X con calle Y**. Una buena o mala condición es considerada como la situación actual de una calle, comúnmente con uno, dos o tres objetos geográficos relacionados, por ejemplo, **asentamiento sobre calle X de calle Y hasta calle W**. El Diccionario de eventos viales está clasificado por accidentes, buenas y malas condiciones. Ya que un tweet está restringido a 140 caracteres, es difícil publicar un enlace, una mención a una cuenta, un evento vial y más de tres objetos geográficos, por tanto, se identifica que el número de objetos geográficos mencionados en los tweets tienen una relación con el tipo de evento vial.

Cada Diccionario de objetos geográficos tiene una primitiva geográfica (punto, línea, polígono). El de transporte público es representado por puntos, el de calles es representado por segmentos de línea y los de colonias y lugares, edificios y monumentos son representados por polígonos. El resultado de la búsqueda de objetos geográficos de los diccionarios y el *Gazetteer* dentro de los tweets se obtiene como una primitiva geométrica (Figura 3).

[2] Son palabras que no agregan ningún sentido al enunciado y son más comunes en algunos lenguajes (artículos, pronombres y preposiciones).

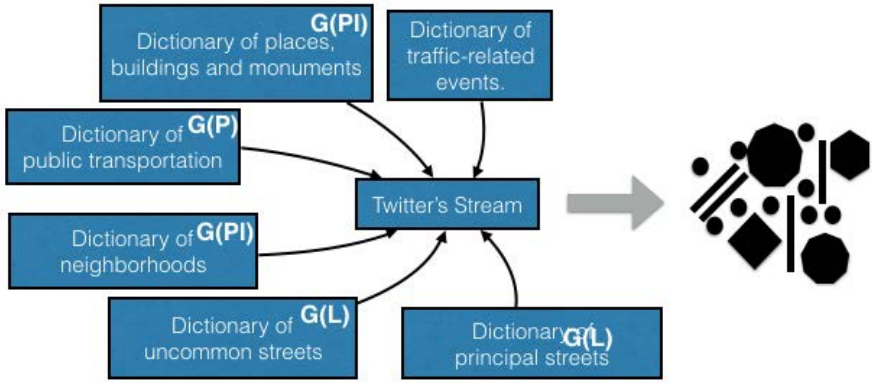

Figura 3. Proceso de identificación de objetos geográficos en el Corpus.

Pueden existir 1, 2 o 3 referencias geográficas, el número de posibles relaciones entre ellas corresponde con la fórmula de combinaciones con reemplazo (Ecuación 1).

$$CR_m^n = \binom{m+n-1}{n} = \frac{(m+n-1)!}{n!(m-1)!} \quad (1)$$

Donde *m* es el número de elementos posibles a seleccionar, en este caso punto, línea o polígono y *n* es el número de elementos encontrados.

Se consideraron las siete relaciones: 1 (punto): representa un accidente en una estación de transporte público; 2 (punto, línea): representa una condición de un segmento de calle frente a una estación de transporte público; 3 (línea, línea): representa un accidente en una intersección; 4 (punto, punto, línea): representa una condición de un segmento de calle delimitado por dos estaciones de transporte público; 5 (punto, línea, línea): representa una condición de un segmento de calle delimitado por otra calle y una estación de transporte público; 6 (línea, línea, línea): representa una condición de un segmento de calle y 7 (línea, línea, polígono): representa un segmento de calle delimitado por una calle y un lugar, edificio o monumento.

Se consideraron tres operaciones geográficas: 1) encontrar la intersección entre calles, 2) encontrar el punto más cercano a otro elemento geográfico y 3) encontrar el *bounding box* o envolvente convexa de un segmento de línea. Las operaciones espaciales fueron ejecutadas utilizando las funciones de *PostGIS* como *ST_Intersection*, *ST_ClosestPoint*, *ST_Envelope* y *ST_ConvexHull*.

Generación del conjunto de entrenamiento

La selección de características determina el éxito o el fracaso de un método de aprendizaje, sin embargo, la mejor forma de seleccionar los atributos más relevantes es de forma manual, basado en el conocimiento del problema de aprendizaje y el conocimiento de cada valor (Witten, I. *et al.*, 2005). El resultado del proceso de geolocalización es la salida que tiene asociado cada vector de características, por tanto, las características que fueron consideradas para conformar el vector están relacionadas al tiempo cuando éste ocurrió. Se contemplan como características temporales, el número del mes (1-12), día del mes (1-28:31), día de la semana (1-7) y hora del día (1-24).

Regresión

El modelo de predicción es generado con la biblioteca Scikit-Learn, la cual contiene diferentes métodos de aprendizaje automático, para hacer uso de las Máquinas de Soporte Vectorial para Regresión (Chang, C. *et al.*, 2011). El conjunto de entrenamiento es seccionado en la entrada compuesta por los vectores de características y la salida por los puntos geocodificados. Con el modelo entrenado se realizan predicciones con la sección de entradas del conjunto de prueba, este es tomado de una sección del conjunto de entrenamiento y no forma parte para el proceso de aprendizaje del modelo de predicción. Los resultados obtenidos se comparan con la sección de salida del conjunto de prueba. La Figura 4 muestra una comparación de la sección de salida del conjunto de prueba (ícono Twitter azul) y los puntos pronosticados (ícono precaución rojo).

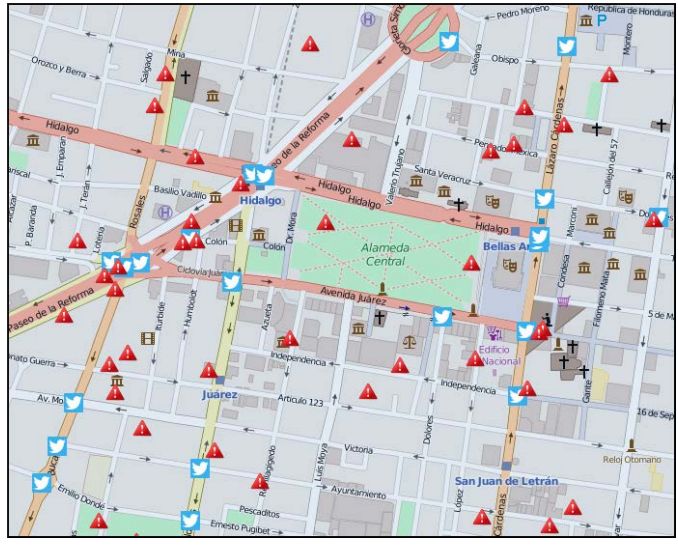

Figura 4. Visualización de puntos pronosticados y los puntos de prueba.

Evaluación, experimentos y resultados

Primeramente, se realiza la evaluación, experimentos y resultados en geocodificación. Para el proceso de predicción de igual forma se describe cuando se considera un verdadero positivo, un falso positivo y un falso negativo para obtener medidas de precision, recall y F. Un verdadero positivo es una muestra encontrada por el sistema que pertenece al conjunto de la solución; un falso positivo es una muestra encontrada por el sistema que NO pertenece al conjunto de la solución; y un falso negativo es una muestra NO encontrada por el sistema que pertenece al conjunto de la solución. El precision es la fracción de instancias que son relevantes con relación al conjunto de instancias recuperadas. El recall es la fracción de instancias que son relevantes con relación al conjunto de instancias que debieron ser recuperadas. La medida F es la precisión de una prueba y refleja la relación entre las medidas precision y recall.

Evaluación para geocodificación

La evaluación considera la calidad de los resultados en los experimentos realizados, para esto se utilizaron las medidas antes mencionadas (precision y recall). Estos parámetros son obtenidos a partir de verdaderos positivos, falsos positivos y falsos negativos. Para evaluar la metodología, 652 tweets fueron geocodificados manualmente. La colección de prueba fue comparada con la metodología descompuesta en etapas, primero por la estandarización, la adición de ejes equivalentes y finalmente con los diccionarios geográficos. Los resultados aparecen en la Tabla 2, donde se observa que tiene una precisión y recall de 85% y 83% respectivamente la cual es superior al 39% y 31% obtenidos de la línea base (el gazetteer).

Tabla 2
Comparación de resultados la geocodificación

	Línea de base	*Estandarización*	*Estandarización + ejes equivalentes*	*Estandarización + ejes equivalentes + diccionarios*	*Colección de prueba*
Todos encontrados	152	152	427	456	652
Al menos uno encontrado	289	388	599	608	652
Errores	363	264	53	44	0
Precisión	0.39	0.43	0.83	0.85	1.0
Recall	0.31	0.39	0.80	0.83	1.0

Evaluación para predicción

Los problemas de regresión trabajan en el conjunto de los números reales, por tanto, los resultados obtenidos por un modelo de regresión cuentan con un rango de distancia entre el valor de la función y el valor real. Los umbrales establecidos en este trabajo son de 50 metros y 100 metros entre el conjunto de elementos pronosticado por el modelo de predicción y el conjunto de los valores de prueba. El proceso de evaluación se lleva a cabo mediante un conjunto de prueba, el cual es una partición del conjunto de entrenamiento y no es utilizado para generar el modelo de predicción. Los vectores del conjunto de prueba son enviados al modelo de predicción y la salida del modelo generado se compara con las coordenadas del conjunto de prueba (véase Figura 5), obteniendo verdaderos positivos, falsos positivos y falsos negativos.

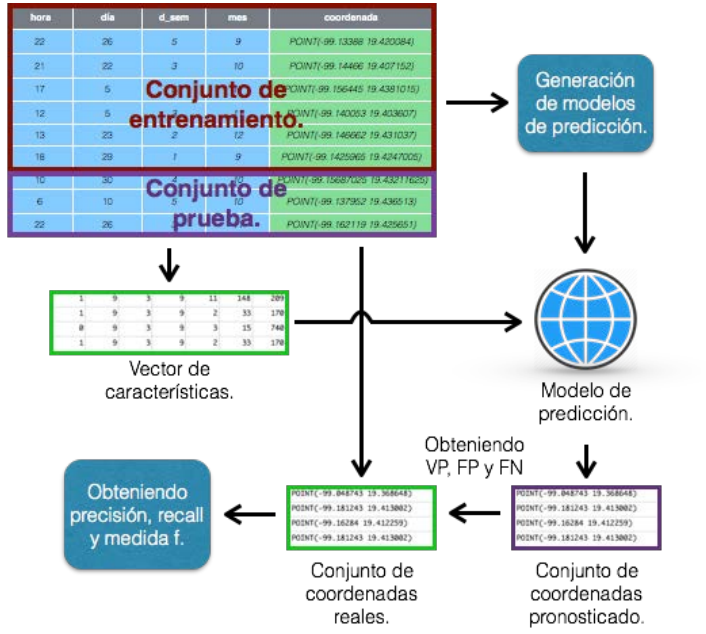

Figura 5. Proceso de evaluación.

Para garantizar que los resultados son independientes del conjunto de entrenamiento y del conjunto de prueba se utiliza validación cruzada. Este método realiza n particiones del conjunto de entrenamiento. Se utilizó la delegación Cuauhtémoc con un umbral de 100 metros, realizando incrementos en el tamaño del conjunto de entrenamiento para cada proceso de evaluación.

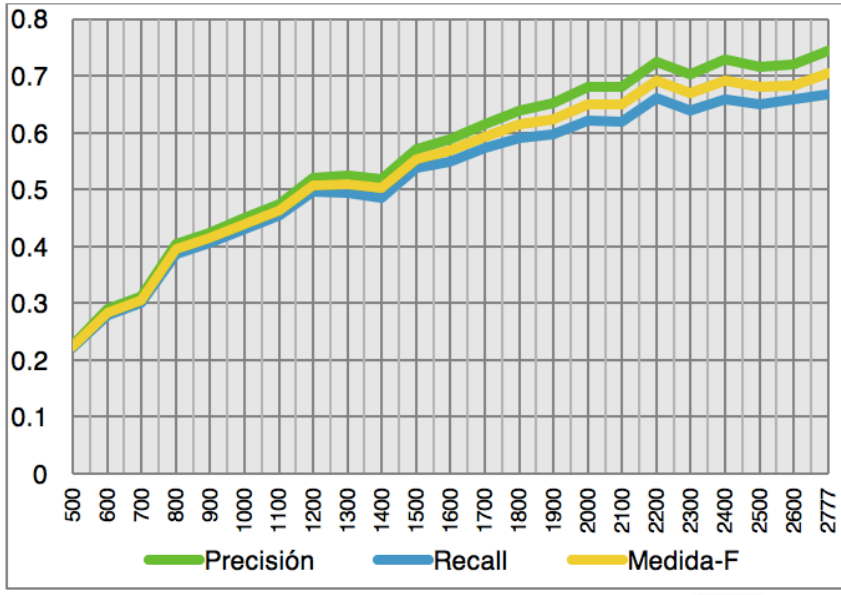

Figura 6. Relación del conjunto de entrenamiento contra la evaluación.

El conjunto de entrenamiento cuenta con las características mencionadas en la sección de Metodología (mes, día, día de la semana y hora) y el método de predicción SVR cuenta con los parámetros kernel rbf, gamma 0.04, constante de penalización de 1e3 y epsilon 1e-4. Los resultados obtenidos se muestran en la Figura 6. El incremento en el tamaño del conjunto de entrenamiento tiene relación con el incremento en precision, recall y medida f obtenidos por los modelos de predicción generados. SVR realiza un proceso de aprendizaje del conjunto de entrenamiento; por tanto, entre más grande sea el conjunto de entrenamiento se obtienen modelos más precisos.

Conclusiones

El trabajo de tesis “Monitoreo urbano de entidades y eventos geográficos basado en sensado social”, fue desarrollado con el propósito de geocodificar eventos viales y hacer predicciones basadas en un espacio-tiempo. Esta actividad se llevó a cabo utilizando métodos de aprendizaje automático, particularmente un enfoque supervisado, en donde los datos fueron entrenados y obtenidos mediante una técnica basada en *Crowdsourcing* y específicamente con Información Geográfica Voluntaria (VGI). La elaboración de este trabajo se realizó en dos etapas: la geocodificación y predicción.

En geocodificación se propone una metodología para asignar coordenadas geográficas a información proveniente de Twitter: mejorar considerablemente la geocodificación enriqueciendo un *Gazetteer* con diccionarios auxiliares; realizar una representación más precisa del tipo de evento vial geocodificado y comprobar que el número de objetos geográficos identificados en cada tweet tiene una relación con el tipo de evento. Además, comprobar que la participación en Twitter tiene una relación directa con las horas pico en la Ciudad de México.

En predicción: proponer una forma de crear un conjunto de entrenamiento a partir de datos geocodificados, con el fin de entrenar un método de aprendizaje automático supervisado para regresión. Se propuso una forma de llevar a cabo un análisis espacio-temporal de eventos viales, así como una forma de evaluar los resultados en un ámbito geográfico. Comprobar que el número de eventos geocodificados utilizados como conjunto de entrenamiento, está relacionado con la precisión del modelo de predicción generado. Entre mayor sea el conjunto de entrenamiento mayor es la precisión y la exactitud del modelo de predicción. Se comprobó que la selección de características es un aspecto fundamental para generar un modelo de predicción acertado. La representación de las características de igual forma, es fundamental para generar un modelo de predicción adecuado, las características deben ser presentadas de forma categorizada con el fin de que no pierdan o agreguen sentido a la característica que se quiere modelar.

La selección de parámetros en la generación del modelo afecta significativamente la precisión, parámetros amplios generan falta de aprendizaje y parámetros justos generan sobre entrenamiento, estos valores pueden ser definidos a prueba y error utilizando validación cruzada.

Por lo tanto, se realizó un aporte a la migración de *Ciudades Inteligentes*, realizando una conexión entre una infraestructura vial, una infraestructura social y una infraestructura de tecnologías de la información.

Bibliografía

Akay, M.F., (2009). “Support vector machines combined with feature selection for breast cancer diagnosis”, Expert systems with applications.

Backstrom, L.; Sun, E. and Marlow, C., (2010). “Find me if you can: improving geographical prediction with social and spatial proximity”, Proceedings of the 19th international conference on World wide web, ACM, pp. 61-70.

Bird, S., (2006). “NLTK: the natural language toolkit”, Proceedings of the COLING/ACL on Interactive presentation sessions, Association for Computational Linguistics, pp. 69-72.

Cavnar, W.B., and Trenkle, J.M., (1994). “N-gram-based text categorization”, Ann Arbor MI, pp. 161-175.

Chang, C.C., and Lin, C.J., (2011). "LIBSVM: a library for support vector machines", ACM Transactions on Intelligent Systems and Technology, pp. 27.

Chourabi, H.; Nam, T.; Walker, S.; Gil-Garcia, J.R.; Mellouli, S.; Nahon, K. and Scholl, H.J., (2012). "Understanding smart cities: An integrative framework", System Science (HICSS), 2012 45th Hawaii International Conference, IEEE, pp. 2289-2297.

Davis Jr, C.A.; Pappa, G.L.; de Oliveira, D.R.R. and L Arcanjo, F., (2011). "Inferring the location of twitter messages based on user relationships", Transactions in GIS, pp. 735-751.

Delboni, T.; Borges, K.A.; Laender, A.H. and Davis Jr, C.A., (2007). "Semantic expansion of geographic web queries based on natural language positioning expressions", Transactions in GIS, pp. 377-397.

Guo, G.; Li, S.Z. and Chan, K., (2000). "Face recognition by support vector machines", Automatic Face and Gesture Recognition, Proceedings. Fourth IEEE International Conference, IEEE, pp. 196-201.

Huang, W.; Nakamori, Y. and Wang, S.Y. (2005). "Forecasting stock market movement direction with support vector machine", Computers and Operations Research.

Lee, R.; Wakamiya, S. and Sumiya, K., (2013). "Urban area characterization based on crowd behavioral lifelogs over Twitter", Personal and Ubiquitous Computing, pp. 605-620.

Ribeiro Jr, S.S.; Davis Jr, C.A.; Oliveira, D.R.R., Meira Jr, W.; Gonçalves, T. S. and Pappa, G.L., (2012) "Traffic observatory: a system to detect and locate traffic events and conditions using Twitter", Proceedings of the 5th International Workshop on Location-Based Social Networks, ACM, pp. 5-11.

Wang, Y. and Cao, L., (2013). "Discovering latent clusters from geotagged beach images", Advances in Multimedia Modeling, Springer, pp. 133-142.

Witten, I.H. and Frank, E., (2005). "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann.

Wu, C.H.; Ho, J.M. and Lee, D.T., (2004). "Travel-time prediction with support vector regression", Intelligent Transportation Systems, IEEE, pp. 276-281.