Visualización para datos probeta en Buses en la ciudad de Bahía Blanca, Argentina

Martínez, Juan[1] ; Varón Mario[2]
Universidad de los Andes
Departamento de Ingeniería de Sistemas y Computación

Resumen:

La municipalidad de Bahía Blanca, Argentina, utiliza un sistema de AVL (Automatic Vehicle Detection) para la recolección de los datos de posición, velocidad y pasajeros de los buses durante el recorrido de su ruta. Dichos datos se han recopilado desde el año 2010 hasta la fecha para las distintas rutas que existen en la ciudad. Los autores de este documento plantean una visualización para identificar el recorrido geográfico de las distintas rutas a través de la red vial de la ciudad de Bahía Blanca, así como la identificación de la velocidad promedio durante el recorrido. Dichas mediciones pueden ser útiles para la municipalidad con el fin de minimizar los tiempos de viaje, así como labores de control sobre los buses.

[1] E-mail address: js.martinez777@uniandes.edu.co , Candidato a Grado del programa de Ingeniería de Sistemas y Computación de la Universidad de los Andes
[2] E-mail address: ma-varon@uniandes.edu.co, Candidato a Grado de la Maestría de Ingeniería de Sistemas y Computación de la Universidad de los Andes

1. Introducción

La municipalidad de Bahía Blanca, Argentina, utiliza un sistema de AVL (Automatic Vehicle Detection) para la recolección de los datos de posición de los buses durante el recorrido de su ruta. Dichos datos se han recopilado desde el año 2010 hasta la fecha para las distintas rutas que existen en la ciudad. Los dispositivos de medición han sido suministrados e instalados por una empresa privada, no obstante, la municipalidad se ha encargado de la recolección de los datos. Dichos han sido entregados en su totalidad a la Universidad de Nacional del Sur para su análisis y posible visualización. Se ha planteado una alianza entre dicha Universidad y los Andes, con el fin que la visualización se realice dentro del alcance del proyecto de la clase de Visual Analytics del departamento de Ingeniería de Sistemas.

La clasificación de los datos está por secuencia cronológica, incluyendo en orden, el año, el mes, el día y la hora (en formato de 24 horas). Los datos están comprimidos en extensión .kmz, la cual es de tipo de ubicación geográfica para apertura con Google Maps. Su versión descomprimida .kml, contiene una serie de información en lenguaje de etiquetas XML, en donde se incluyen entre otros, la posiciones de latitud y longitud así como el timestamp y otra información de interés.

El dispositivo de AVL, incorpora una unidad de APC (Automatic Passenger Counting System) la cual permite generar un tipo de archivo tipo pass. Este archivo se genera cada vez que un pasajero se sube al bus. Así mismo, la unidad AVL genera cada cierto tiempo un archivo de tipo pcontrol. La frecuencia o la condición por la cual se genera dicho archivo resulta desconocida y se intentará establecer con el estudio de los datos. En este archivo se encuentra entre otros, la placa del vehículo sobre la cual se genera la alerta, la Fecha y la hora, las coordenadas de latitud y longitud, la distancia que ha recorrido el vehículo durante la jornada de operación, el rumbo, el Estado del vehículo y su velocidad. La estructura de los datos se puede visualizar como se muestra a continuación:

Ilustración 1- Estructura de los Datos

Existe un archivo tipo REC que contienes la posición de latitud y longitud para toda la ruta recorrida por el bus y finalmente un archivo de tipo STOP que se genera en el momento de detención de un bus.

En la medida en que se puede obtener de los datos, tanto la ruta, como la cantidad de pasajeros, así como la velocidad media del bus y el tipo de alarmas que se generan, el alcance del proyecto debe estar involucrado directamente con estos datos o con la derivación que se pueda hacer de ellos.

Se debe mencionar que para los fines de este documento los archivos REC y STOP fueron descartados en su totalidad y solo se hizó uso de la información que estaba consignada en los archivos tipo PASS y PCONTROL.

2. Estado del Arte

La definición más adecuada para los sistemas de información al viajero avanzadas (ATIS) sería "la aplicación sistemática de las tecnologías de información y comunicación para la recopilación de datos relacionados con los viajes y el procesamiento y la entrega de información al viajero" (McQueen, 2002). Este concepto implica claramente tanto a los viajeros en las redes de

transporte público como aquellos que hacen uso del privado. Zhang (Zhang, y otros, 2010), y Adler (Adler & Blue, 1998) clasifican los sistemas de información al viajero (TIS) en dos generaciones: la primera consiste en proveer la guía de ruta al viajero por medio de señales de mensajes variables (VMS), mientras que la segunda generación (ATIS) se puede definir como un proceso dinámico para la guía de ruta mediante el uso sistemático nuevas tecnologías que recojan la información del tráfico en tiempo real. Así mismo, Koppelman (Koppelman & Pas, 1980) y Kanninen (Kanninen, 1996) han argumentado que proveer a los viajeros con información relevante acerca de las opciones sobre su viaje tiene el potencial de cambiar sus comportamientos en formas que son beneficiosas para la mejorar la eficiencia en el uso de la red de transporte. Sin embargo, esta perspectiva es bastante limitada como fundamento principal del uso y despliegue de sistemas ATIS. Sí se incorpora en el análisis la teoría microeconómica del consumidor (Samuelson, 1947), serán las evaluaciones exhaustivas de las alternativas (opciones de ruta), así como la exploración de atributos para cada una y el cálculo de la utilidad del intercambio entre ellas lo que tiene valor para el viajero. Por tanto y de aquí en adelante debe entenderse por un sistema ATIS como : "la aplicación sistemática de las tecnologías de información y comunicación con el fin de proporcionar la mejor guía de ruta a través de la recopilación de datos relacionados con los viajes en tiempo real, así como su procesamiento y la entrega de información al viajero para que puedan evaluar de manera exhaustiva cada uno de los atributos de las alternativas provistas y en consecuencia realizar los respectivos intercambios entre ellas para maximizar su utilidad (por ejemplo, reducir el tiempo de viaje)".

Con el fin de seguir la tesis delineada, hay varios atributos que pueden ser proporcionados por un sistema ATIS para los viajeros públicos en las redes de transporte. Estos atributos se pueden dividir en dos clases diferentes, el primer tipo de atributos de viajes están relacionados con la congestión no recurrente y los segundos, a la congestión recurrente. McGroarty (McGroarty, 2010) ha definido la primera como "demoras de tráfico inesperadas causadas principalmente por accidentes e incidentes, averías de vehículos, actividades de construcción de carreteras, eventos especiales, eventos climáticos extremos, etc" por lo que los atributos relacionados específicamente con este tipo de eventos y que se proporcionarían para el viajero, serán el tipo, la localización, la duración y su reducción de la capacidad en el arco donde ocurrió el incidente, así como su impacto en el tiempo de viaje esperado . El viajero podrá estar interesado en la recoger esta información antes del viaje y durante el momento del viaje. Cabe mencionar que las únicas decisiones posibles ante está información por parte del viajero, es el cambio de la ruta en el transporte público, la modificación de la hora de salida o eventualmente el cambio de modo de transporte (i.e cambiar el bus por el metro).

Por otra parte, la congestión recurrente se asocia con el diagrama fundamental en la modelación macroscópica, que se representa por el flujo, la densidad y la velocidad (Roess, Prassas, & McShane, 2004). Los atributos proporcionados a los viajeros, en este tipo de circunstancias, se basan en cálculos directos o indirectos de estas variables, y se expresan por medio del tiempo de viaje, velocidad media, volumen y costo de viaje. Para que el viajero se familiarice con el comportamiento del tráfico y pueda construir una percepción informada, la recolección de la información debe hacerse en el pre-viaje, durante el momento del viaje y en después de terminado viaje. (Chorus, Molin, & Van Wee, 2005). De la misma forma que para la congestión no recurrente, las opciones quedan delimitadas al cambio de ruta, la modificación en la hora de salida o el cambio en el modo de transporte. No obstante, mientras que para la primera opción

dicha decisión afecta el viaje que se hará a continuación, en este segundo caso, la decisión modifica un hábito de viaje y se perpetúa a lo largo del tiempo, hasta que llegue nueva información que permita evaluar nuevamente, la función de utilidad del viajero.

Sin embargo, hay información de tráfico que no se derivan ni la congestión recurrente, ni tampoco de la no recurrente, Torres (Torres, 2008), McQueen (McQueen, 2002) y Veneziano (Veneziano, 2010), los han descrito en detalle, e incluyen entre otros, estado meteorológico, puntos de interés, puntos de fiscalización del tráfico vehicular, imágenes de CCTV y otros que son menos habituales. De hecho, existe evidencia que algunos de estos parámetros modifican drásticamente variables como la velocidad o el flujo vehicular, no obstante de una u otra forma se reflejan en las características de los atributos referidos a la congestión recurrente y no recurrente. En este sentido, Wunderlich (Wunderlich, 1996) ha calculado que las condiciones meteorológicas adversas pueden reducir hasta un 75% la capacidad de la red. Sin embargo, ninguno tipo de está información adicional puede considerarse como atributos de viaje, ya que no afecta a la utilidad individual personal y por lo tanto no puede ser objeto de intercambio entre las distintas alternativas provistas por el sistema ATIS. (Innocenti, Lattarulo, & Pazienzia, 2009)

El conjunto de opciones de viaje relacionados con los atributos mencionados anteriormente, incluye, cancelar el viaje, modificar el modo, cambiar las hora de salida y finalmente evaluar las alternativas de elección de ruta y escoger una. (Khattak, Polydoropoulou, & Ben-Akiva, 1995). Cuando el viajero adopta cualquiera de estas opciones, está realizando directamente una evaluación entre los distintos atributos, esperando aumentar su utilidad de su decisión escogida. Teniendo en cuenta este hecho y la definición propuesta para un ATIS en este documento, un ATIS comercial es el único medio que pudiera ofrecer a los viajeros la información para que pueda construir y medir la utilidad objetiva y tomar por lo menos una de las opciones previstas anteriormente. Para todos los efectos de este documento, se entiende que la maximización de la utilidad está en función de reducir su tiempo de conmutación entre su punto de origen y el destino de su viaje.

En la medida en que se obtienen los datos de velocidad y tiempos de viaje para cada una de las rutas, la información provista puede soportar un ATIS de tipo descriptivo para los viajeros, así como permitir la intervención por medio de política públicas que permitan optimizar dichos tiempos, con su consecuente beneficio en la productividad de la ciudad. Este ATIS le permitirá a los pasajeros del transporte público conocer los promedios históricos, referidos a los tiempos de desplazamiento de cada una de las rutas del transporte público en la ciudad de Bahía Blanca.

Finalmente los datos de pasajeros por ruta se utilizan principalmente para verificar la consistencia en la tarifa del transporte público, siempre y cuando, dicha sea fijada por la municipalidad. Así mismo, tales datos son útiles para determinar en el nivel macroscópico la matriz origen-destino, así como la determinación de la factibilidad de ciertas rutas de transporte público, bien cuándo su cantidad de pasajeros esté por debajo de su nivel de factibilidad o en su defecto cuando se encuentren por encima, obligando a la entidad a adoptar más rutas para un mismo viaje con idénticas coordenadas de origen-destino.

En determinadas circunstancias, la cantidad de pasajeros se puede utilizar como variable proxy, para el cálculo del crecimiento de la economía de la ciudad, así como para otro tipo de estudios socioeconómicos que requieran de estos valores.

3. Caracterización

Para atender tanto problemáticas de seguridad vial, como aquellos relacionados con la eficiencia de los sistemas de transporte, se han creado los Sistemas Inteligentes de Transporte (ITS). Estos son aplicaciones avanzadas que combinan sistemas electrónicos, de comunicaciones, de computadores y sensores. Dichos ITS integran vehículos, personas e información de la vía con estrategias de gestión del tráfico para proveer información en tiempo real para aumentar la seguridad, la eficiencia y el confort en los sistemas de tráfico tanto públicos, como privados. (Roess, Prassas, & McShane, 2004).

En especial, el desarrollo de ciertas tecnologías de ITS, tales como las unidades AVL (Detección Automática de Vehículos) ha permitido la recolección de datos de la flota correspondiente al transporte público. Dichos datos, como ya se mencionó, se utilizan para formular políticas públicas que optimicen el uso de está infraestructura. En este caso, la caracterización del problema responde a un problema de transporte agrupado con técnicas de Big Data.

La totalidad del dataset alcanza 80 GB de información con más de 5 millones de Archivos. Los datos corresponden a los puntos de control de cada uno de los buses que compone el sistema del transporte público en Bahía Blanca, así como la cantidad de pasajeros para cada una de las paradas, como también los puntos de latitud y longitud donde se grabó algún tipo de dato. Los datos en promedio, se generan cada minuto. Finalmente se registra los puntos de detención de los buses al finalizar la jornada.

El dataset fue entregado en un FTP, en formato .RAR. Cada carpeta .rar está dividida en sub folderes por años, luego por meses, luego por días y finalmente por horas. Para cada hora hay una cantidad de archivos que corresponde los registros de cada una de las rutas para ese momento del tiempo. El nombre del archivo no tiene ninguna relación con el número de ruta. En la mayoría de los casos, las rutas duran más de una hora en recorrido, con lo cuál debían juntarse ambos archivos para obtener el perfil del tiempo de viaje. En la medida en que había que leer varios archivos y hacer una concatenación de ellos, está tarea resultaba titánica y se decidió omitirla.

Dentro de cada uno de estos archivos KMZ, se encuentra un archivo de posicionamiento geográfico KML. Este archivo es un XML con coordenadas de Latitud y Longitud. Se debe decir que aun cuando no exista información útil dentro de cada uno de estos archivos, el sistema genera la cabecera de los mismos. Esto es, a pesar que el bus este parado, se envía un reporte que no contiene información útil. Para aquellos dónde se genera información útil, la estructura del XML, crea una etiqueta que se llama <extended data>. Aquí se registra la información que se visualiza a continuación -Posición, Distancia, Patente, Rumbo, Velocidad, Fecha, Recorrido y Línea-. Se puede observar la etiqueta a continuación:

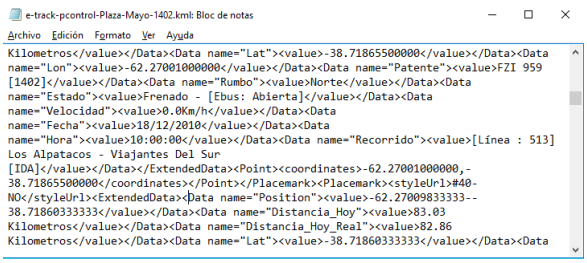

Ilustración 2- Etiqueta de Extended Data con la información útil

La manipulación de estos datos es sumamente compleja, ya que se requirió descomprimirlos dos veces, primero del RAR y después de KMZ a KML. Se aplicaron en ambos casos código de Batch, pero en la medida en que dicho código se ejecuta de forma secuencial, el tiempo para terminar de descomprimir todos los archivos, se acercaba a 30 días de máquina. Por esa razón luego hacer un cambio de extensión de KMZ a ZIP, se borraron todos aquellos archivos que pesaban menos de 1 KB, ya que en su mayoría

solo contenían las cabeceras, sin información útil. Dicho proceso redujó hasta en 60% el tiempo de este proceso.

No obstante, no fue posible llevar la totalidad de los datos a la memoria de los equipos de cómputo con los que cuentan los estudiantes para realizar analítica sobre ellos. Se intentó hacer uso de WEKA y GRAPHLAB y en ambos casos, la capacidad en memoria resultó insuficiente. De tal forma, se decidió que se solo se trabajarán con los registros del año 2010 y se deja para trabajo futuros la integración con la totalidad de los datos.

4. Método Propuesto

De acuerdo a las tareas identificadas para la solución del problema, se propuso una visualización compuesta por dos conjuntos de "Horizon Charts", el primero permitirá visualizar el comportamiento de las velocidades por cada una de las rutas de buses en Bahía blanca en función del tiempo para un cierto día, mes y año. Otro gráfico de la misma naturaleza estará destinado a presentar la cantidad de pasajeros por cada bus a lo largo del tiempo; nuevamente para un cierto día, mes y año.

Estas visualizaciones se proponen ya que se ha de tratar datos temporales, que presentan cantidades numéricas, es por esto que los "Horizon Charts" presentan los datos sobre una misma escala lineal para un cierto día y utilizan el color como un canal efectivo para presentar rangos en los cuales los datos están presentes. Estos son: velocidades superiores e inferiores a la velocidad promedio de dicho día, y cantidad de personas superiores e inferiores al promedio de personas que transporta el bus en el día. Es por esto que la visualización tiene una alta efectividad a la hora de soportar las tareas de descubrir las distribuciones y "outliers" de velocidades y cantidad de personas durante los días.

Más aún, se propone como interacción la posibilidad de seleccionar puntos específicos de cada uno de los "Horizon Charts" para visualizar, en un mapa de la ciudad, la ruta seleccionada y el punto geográfico correspondiente que corresponde a la velocidad o cantidad de personas que generó interés en el usuario.

Por último, dado que las visualizaciones tienen una granularidad horaria por día, los datos son filtrados por año, mes y día a partir de dos controles circulares y un control lineal. Los primeros dos permiten filtrar el mes y el día de acuerdo a su naturaleza cíclica; el último permite seleccionar el año y es un control lineal debido a que es un dato lineal.

Para las técnicas de analítica se corrió un modelo de regresión logística múltiple siendo la variable dependiente la velocidad instantánea por hora mientras que las independientes se determinaron como la fecha -categórica- y la línea o ruta -categórica-. Se hace uso de Graphlab por su capacidad para trabajar con datos en memoria y en formato XML. Se dividió el dataset del año 2010, con una función aletaoria (seed=0) en 80% como training set y 20% como test Set, sin base de datos de validación. El propósito de la regresión consiste en predecir con los datos del años 2010, la velocidad instantánea promedio para cada hora, en función del día, el mes y la ruta del transporte público. Este método sí bien no resulta del todo confiable para determinar la velocidad media del viaje, se puede extrapolar como un dato probeta para medir la velocidad media del tráfico alrededor del bus.

5. Resultado

Se puede observar, el resultado de la visualización que se construyó en D3 continuación:

Ilustración 3- Resultado de la Visualización con los horizons Charts

Se adjunta así mismo, los resultados de las regresiones para una de las horas para la ruta 1411:

	Coeficiente	Error típico
Intercepción	23,4184006	1,53001217
Variable X 1	-0,0375656	0,21326358
Variable X 2	1,85679	0,19867

La curva de regresión para la velocidad contra la variable con mayor significancia, esto es la ruta, se puede ver así:

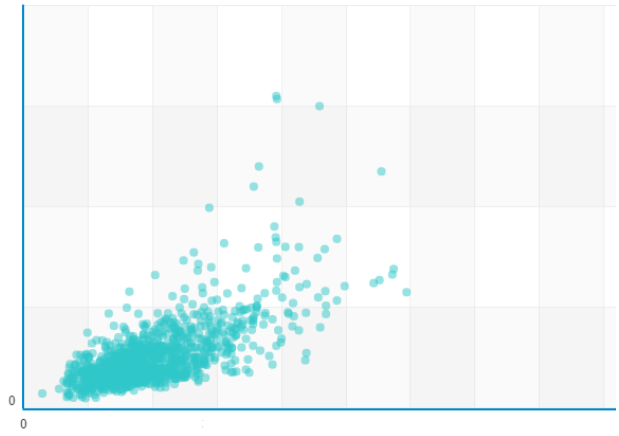

Ilustración 4- Gráfica de regresión sobre velocidad para la línea 411

6. Análisis

Tal y como se puede observar, frente a la analítica de los datos se determina que el principal determinante de la velocidad es la ruta, sobre la fecha. Esto es, la fecha no está correlacionada de ninguna manera con la velocidad instantánea promedio. Se pueden utilizar técnicas de Clustering para verificar en qué sitios se generan congestiones vehiculares, medidas como la reducción de en la velocidad de desplazamiento.

7. Conclusiones

   a. El preprocesamiento de los datos resultó de una magnitud que sobrepasaba el alcance de este curso. Los datos se encuentran mal formateados y el sistema del AVL reporta de una manera poco intuitiva que no permite la gestión de los mismos.

   b. La visualización propuesta se puede implementar sobre un dataset que permita su fácil manipulación. Este visualización permite observar los outliers, así como otras medidas de interés.

   c. La velocidad instantánea para el dataset de prueba, arrojó que la misma se encuentra altamente correlacionada con la ruta, mientras que no lo es tanto con la hora de salida.

   d. Se proponer implementar técnicas de clustering para verificar los puntos donde se genera congestión recurrente, en la medida en que ya se tiene la información de la velocidad relacionada con sus coordenadas de latitud y longitud.

Bibliografía

Adler, J. L., & Blue, V. (1998). Toward the design of intelligent traveler information systems. *Transportation Research Part C: Emerging Technologies*, Vol. 6, No. 3, pp. 157 – 172.

Chorus, G., Molin, E., & Van Wee, B. (2005). Use and Effects of Advanced Traveller Information

Services (ATIS): A Review of the Literature. *Transport Reviews*, Vol. 26, No. 2, 127–149. .

Innocenti, A., Lattarulo, P., & Pazienzia, M. (2009). An Experimental Analysis of Travel Mode Choice. *Società Italiana di Economia dei Trasporti e della Logistica - XI Riunione Scientifica –.* Trieste.

Kanninen, J. B. (1996). Intelligent transportation systems: an economic and environmental policy assessment. *Transportation Research Part A*, 30, pp. 1-10.

Khattak, A., Polydoropoulou, A., & Ben-Akiva, M. (1995). Modeling Revealed and Stated Pretrip Travel Response to ATIS. *Transportation Research Part*, 1537, pp. 46-54.

Koppelman, F., & Pas, E. (1980). Travel choice behaviour: models of perceptions, feelings, preference, and choice,. *Transportation Research Record*, 765, pp. 26-33.

McGroarty, J. (2010). Recurring and Non Recurring Congestion: Causes, Impacts and Solution. *Neihoff Urban Studio*, pp 2-6.

McQueen, B. (2002). *Advanced Traveler Information Systems Intelligent Transportation systems.* Norwood: Artech house.

Roess, R., Prassas, E., & McShane, W. (2004). *Traffic Engineering.* Washington: Pearson.

Samuelson, P. (1947). *Foundations of economic analysis.* Harvard University Press: Cambridge.

Torres, N. (2008). *ATIS Final report.* Hartford: Department of Civil and Enviromental Engineering University of Hartford.

Veneziano, D. (2010). *Rural Traveler Information Needs Assessment and Pilot Study.* Sacramento: California DOT.

Wunderlich, K. (1996). *An Assessment of Pre-Trip and en route ATIS Benefits in a Simulated Regional Urban Network.* Orlando: ITS America.

Zhang, L., Li, J., Zhou, K., Gupta, S., Li, M., Zhang, W.-P., . . . Misener, J. (2010). Design and Implementation of a Traveler Information Tool with Integrated Real-Time transit Information and Multi Modal Trip Planning. *Transport Research Board 91th Meeting.* Washington.